SELF-DISTILLATION ZERO: SELF-REVISION TURNS BINARY REWARDS INTO DENSE SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Current post-training methods in verifiable settings fall into two categories. Reinforcement learning (RLVR) relies on binary rewards, which are broadly applicable and powerful, but provide only sparse supervision during training. Distillation provides dense token-level supervision, typically obtained from an external teacher or using high-quality demonstrations. Collecting such supervision can be costly or unavailable. We propose **Self-Distillation Zero** (SD-ZERO), a method that is substantially more training sample-efficient than RL and doesn't require an external teacher or high-quality demonstrations. SD-ZERO trains a *single model* to play two roles: a *Generator*, which produces an initial response, and a *Reviser*, which conditions on that response and its binary reward to produce an improved response. We then perform on-policy self-distillation to distill the reviser into the generator, using the reviser's token distributions conditioned on the generator's response and its reward as supervision. In effect, SD-ZERO trains the model to *transform binary rewards into dense token-level self-supervision.*

On math and code reasoning benchmarks with Qwen3-4B-Instruct and Olmo-3-7B-Instruct, SD-ZERO improves performance by at least 10% over the base models and outperforms strong baselines, including Rejection Fine-Tuning (RFT), GRPO, and On-Policy Self-Distillation (OPSD), under the same question set and training sample budget. Extensive ablation studies show two novel characteristics of our proposed algorithm: (a) *token-level self-localization*, where the reviser can identify the key tokens that need to be revised in the generator's response based on reward, and (b) *iterative self-evolution*, where the improving ability to revise answers can be distilled back into generation performance with regular teacher synchronization.

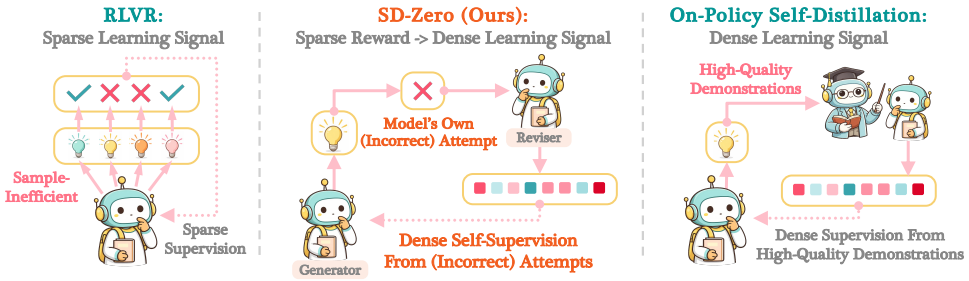


Figure 1: We introduce **SD-ZERO**, an on-policy self-distillation paradigm where the teacher (*reviser*) can condition on incorrect student (*generator*) attempts.

1 INTRODUCTION

Reinforcement learning (RL) has become the dominant approach for post-training language models on reasoning tasks in verifiable settings such as math and coding (Shao et al., 2024; Guo et al., 2025; Chen et al., 2025a; Zheng et al., 2025). These methods require only a binary reward, typically whether the final answer is correct, that makes them broadly applicable. However, a binary reward per response provides no information about which intermediate reasoning steps were sound. This sparse supervision makes training expensive: the model must discover good reasoning by comparing and contrasting many self-generated responses.

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.

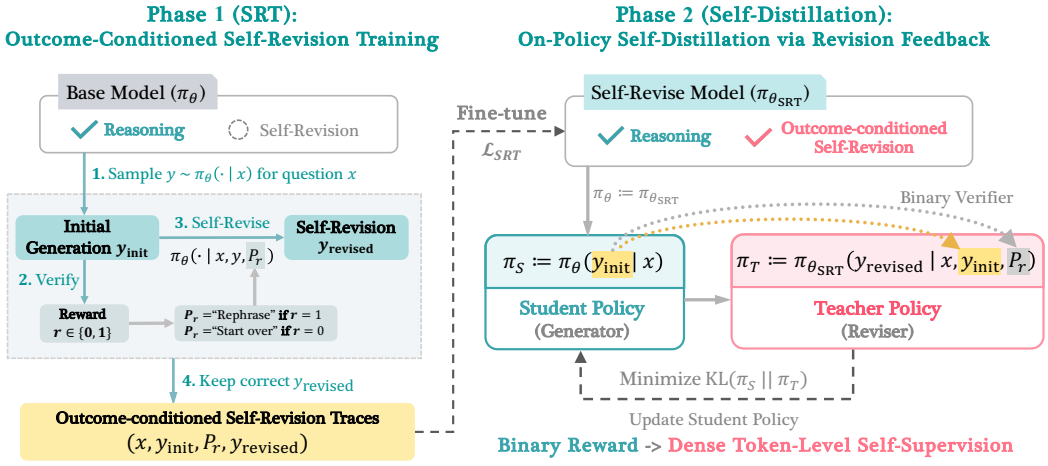


Figure 2: **Overview of SD-ZERO.** In **Phase 1 (SRT)**, we collect 6K outcome-conditioned self-revision traces by sampling an initial response from the base model, prompting the model to self-revise its incorrect response, and keeping the correct self-revision. In **Phase 2 (Self-Distillation)**, we conduct on-policy self-distillation with the self-revise (SRT) model acting as both student and teacher: the *student* generates an on-policy response, and the *teacher* generates a revised version conditioned on that response and its outcome reward. Throughout the Self-Distillation phase, the model bootstraps performance through internalizing self-revision behavior. See Algorithm 1 for details.

Distillation methods provide a more sample-efficient alternative by converting supervision into token-level feedback on student-generated responses. *On-policy distillation* methods (Agarwal et al., 2024; Gu et al., 2024; Boizard et al., 2025; Minixhofer et al., 2025; Lu & Lab, 2025) assume access to an external stronger teacher that can provide token-level feedback on the student's responses. More recent *self-distillation* methods, including OPSD (Zhao et al., 2026), SDFT (Shenfeld et al., 2026), and SDPO (Hübotter et al., 2026), remove the external teacher but still require high-quality demonstrations that are much better than the model's responses. These demonstrations are typically assumed to come from an external teacher (OPSD, SDFT) or repeated generation and filtering from the model itself (SDPO) [1]. Collecting such supervision can be unavailable or prohibitively expensive to collect. This raises the central question of our work:

> Can the model condition its own initial attempts (possibly incorrect) and their sparse rewards, and provide improved dense supervision to itself?

Self-Distillation Zero (SD-ZERO): Our method is built around a single model that plays two roles: a *generator* and a *reviser*. As a generator, the model produces candidate responses to a given question. As a reviser, the same model is conditioned on its own response and reward, and either corrects an *incorrect* response or rephrases a *correct* one.

SD-ZERO proceeds in two phases. In the first phase, we collect self-revision traces by sampling responses, checking for correctness, and prompting the model to revise incorrect ones. We retain only traces where revision succeeds. We fine-tune on these filtered traces, a procedure we call Self-Revision Training (SRT). In the second phase (self-distillation), we use the reviser as a teacher to provide token-level supervision over the generator's responses, thereby transforming outcome-level reward into dense token-level supervision.

Key Findings. Applying SD-ZERO on Olmo and Qwen models on math and code reasoning datasets shows the following:

- **Phase 1 (SRT) alone outperforms all baselines on the same data budget.** Training on 6K self-generated revision traces improves average accuracy by 7.8% for Qwen3-4B-Instruct and 9.2% for Olmo-3-7B-Instruct (Table 1).

[1] See Table 2 for a detailed comparison of post-training methods.

- **Phase 2 (Self-Distillation) is critical for the training and inference efficiency of SD-ZERO.** SRT trained model produces overly long responses. Self-Distillation trains the model to internalize its *revision* behavior and generate stronger answers directly as the *generator*, reducing response length by roughly $2\times$ (Figure 3). It also improves performance beyond SRT, yielding an additional 2.7% gain for Qwen3-4B-Instruct and 1.2% for Olmo-3-7B-Instruct, for total gains of 10.5% and 10.4% over the base models. *Importantly*, Self-Distillation is also what makes SD-ZERO training sample-efficient during training: it requires only one response per question, whereas SRT must sample multiple responses to construct revision traces (Section 3.3)
- **SD-ZERO enables iterative self-evolution through regular teacher synchronization.** Because training also improves the model's revision capability, the updated model can serve as the reviser teacher in subsequent rounds of Self-Distillation. Synchronizing the teacher with the updated student after the first round of Self-Distillation adds a further 2.2 points of training accuracy (Figure 5), and gains continue for two more rounds before saturating.

2 SD-ZERO: TURNING BINARY REWARDS INTO DENSE SELF-SUPERVISION

We consider a dataset $\mathcal{D} = \{(x_i, a_i)\}_{i=1}^{N}$, where x_i denotes an input problem and a_i its corresponding ground-truth final answer. Crucially, we do not assume access to gold solutions. Let π_θ denote the student policy with parameters θ, which generates a reasoning response y for input x according to $\pi_\theta(y \mid x)$. For each example (x, a) and generated response y, we assume access to a binary reward $r(y, a) \in \{0, 1\}$, where $r(y, a) = 1$ if the final answer extracted from y matches a, and $r(y, a) = 0$ otherwise.

The core design of SD-ZERO is around a single model that can play two roles: a *generator* and a *reviser*. We partition the training questions into two disjoint subsets of sizes N_1 and N_2. **Phase 1 (Self-Revision Training)** uses the first N_1 questions to collect self-revision traces and train the model to perform both the generator and reviser roles. **Phase 2 (Self-Distillation)** uses the remaining N_2 questions and leverages the reviser to transform sparse outcome-level supervision into dense token-level supervision over the generator's responses. We refer to the resulting models as the SRT model and the SD-ZERO model.

2.1 PHASE 1 (SRT): SELF-REVISION TRAINING

Our base models exhibit strong *generator* performance but weak *reviser* performance (Figure 3). This phase strengthens the model as a *reviser*, while also improving it as a *generator*.

Outcome-Conditioned Self-Revision. For each input x, we sample multiple initial reasoning responses $y_{\text{init}} \sim \pi_\theta(\cdot \mid x)$. Given a sampled response y_{init}, we construct a revision prompt that conditions on the original problem, the sampled response, and a control phrase indicating whether the response should be revised or simply rephrased:

$$P_r = \begin{cases} \text{“Let me rephrase the above solution.”}, & \text{if } r(y_{\text{init}}, a) = 1, \\ \text{“Wait, this response is not correct, let me start over.”}, & \text{if } r(y_{\text{init}}, a) = 0. \end{cases}$$

We then use the same model to generate a revised response:

$$y_{\text{revised}} \sim \pi_\theta(\cdot \mid x, y_{\text{init}}, P_r).$$

For correct initial responses, this encourages the model to produce a rephrase of a correct solution, for incorrect initial responses, this encourages the model to critique on the previous attempt and regenerate a correct solution. We observe that when asked to rephrase a correct response, the model typically produces a shorter answer, which may additionally encourage more concise responses during the second phase. We keep the traces whose y_{revised} is correct, forming $\mathcal{D}_{\text{REVISION}} = \{(x, y_{\text{init}}, P_r, y_{\text{revised}})\}$.

Self-Revision Objective. We train the model on two tasks simultaneously: given the input, initial attempt, and reward signal, produce a corrected or rephrased response (revision); and given only the input, produce the full correct response from scratch (generation). The corresponding loss $\mathcal{L}_{\text{SRT}}$ consists of two log-likelihood terms $\mathcal{L}_{\text{revision}}$ and $\mathcal{L}_{\text{generation}}$.

The *revision loss* trains the model to produce y_{revised} when conditioned on the input x, the first attempt y_{init}, and the reward prompt P_r:

$$\mathcal{L}_{\text{revision}}(\theta) = \mathbb{E}_{(x, y_{\text{init}}, P_r, y_{\text{revised}}) \sim \mathcal{D}_{\text{REVISION}}} \left[-\sum_{t=1}^{|y'|} \log \pi_\theta\left(y'_t \mid x, y_{\text{init}}, P_r, y'_{<t}\right) \right], \text{where } y' = y_{\text{revised}}. \quad (1)$$

The *generation loss* retains the model's ability to produce the full correct response conditioned only on input x:

$$\mathcal{L}_{\text{generation}}(\theta) = \mathbb{E}_{(x, y_{\text{init}}, P_r, y_{\text{revised}}) \sim \mathcal{D}_{\text{REVISION}}} \left[-\sum_{t=1}^{|y'|} \log \pi_\theta\left(y'_t \mid x, y'_{<t}\right) \right], \text{where } y' = [y_{\text{init}}, P_r, y_{\text{revised}}]. \quad (2)$$

The overall self-revision objective combines these two terms:

$$\mathcal{L}_{\text{SRT}}(\theta) = \mathcal{L}_{\text{revision}}(\theta) + \mathcal{L}_{\text{generation}}(\theta). \quad (3)$$

$\mathcal{L}_{\text{revision}}$ (Eq. 1) trains the model to self-revise, i.e., producing y_{revised} conditioned on the input x, the first attempt y_{init}, and the prompt P_r, while $\mathcal{L}_{\text{generation}}$ (Eq. 2) retains the model's generation capability, i.e., producing correct response y conditioned on input x. Because y' in Eq. 2 contains y_{revised}, $\mathcal{L}_{\text{generation}}$ already supervises the revision tokens, and $\mathcal{L}_{\text{revision}}$ weights them a second time in the input format that the reviser receives in Phase 2; without it, the reviser often inserts meta-phrases such as "let me revise the previous output" (Appendix E). This combination implicitly teaches the model to actively evaluate its current response at inference time and generate a self-revision if necessary. As a result, the trained model often exhibits explicit self-revision behaviors that lead to extremely long responses (see Sections 3.2 and 3.3). In the next phase, we perform on-policy distillation on the SRT model, distilling this learned revision behavior back into the generator to enable sample-efficient self-improvement.

2.2 Phase 2 (Self-Distillation): On-Policy Self-Distillation via Revision Feedback

Self-Distillation phase distills the reviser's behavior back into the generator, so that the model can internalize revision and produce more compact and well-directed responses. Concretely, we leverage the self-revision capability learned in Phase 1 to perform on-policy self-distillation. Let θ_{SRT} denote the model obtained at the end of Phase 1. In Phase 2, we initialize the student parameters as $\theta := \theta_{\text{SRT}}$. The **generator** (student) uses the current parameters θ to produce on-policy responses $\pi_\theta(\cdot \mid x)$; its parameters are updated throughout Phase 2. The **reviser** (teacher) remains frozen at θ_{SRT} and provides token-level supervision by conditioning on the generator's response and its binary reward via $\pi_{\theta_{\text{SRT}}}(\cdot \mid x, y, P_r)$. The generator is trained to match the reviser's distribution with:

$$\mathcal{L}_{\text{Self-Distillation}}(\theta) = \mathbb{E}_{(x,a)\sim\mathcal{D}}\, \mathbb{E}_{y\sim\pi_\theta(\cdot|x)} \sum_{t=1}^{|y|} D_{\text{KL}}(\pi_\theta(\cdot \mid x, y_{<t}) \,\|\, \pi_{\theta_{\text{SRT}}}(\cdot \mid x, y, P_r, y_{<t})). \quad (4)$$

For each training example, a response is sampled from the model and checked against the ground-truth final answer for correctness. The Phase 1 reviser, conditioned on the model's response and whether it was correct, produces a next-token distribution at each token position. The model is trained to match this distribution via KL divergence loss.

3 Experiments

3.1 Experimental Setup

Models. We use Qwen3-4B-Instruct (Yang et al., 2025) and Olmo-3-7B-Instruct (Olmo et al., 2025) as base models. All sampling uses temperature 0.7 with a 16K token limit during training and a 32K limit at evaluation. We report avg@8 throughout.

Datasets. We train on two domains separately: (1) **OpenR1-Math** (Hugging Face, 2025), from which we select 15K competition- and olympiad-level problems with verified solutions, and (2) **Codeforces** (Penedo et al., 2025), from which we select 7.5K samples from the cpp (`solutions`) subset and

	Math						Code		Avg.
	AIME24	AIME25	HMMT25	AMOBench	OpenR1	MATH	Codeforces	LCB	
Qwen3-4B-Instruct									
Base Model	59.6	45.8	26.7	9.8	55.8	91.0	48.0	61.8	49.8
SFT	61.7	46.3	31.3	7.3	56.1	91.3	49.1	57.2	50.0
RFT	64.2	52.1	37.1	11.3	59.3	91.9	50.9	68.0	54.3
GRPO	62.5	50.0	30.4	11.0	**62.9**	93.5	52.2	62.6	53.1
OPSD	63.3	47.9	32.9	9.0	57.0	91.1	49.3	59.2	51.2
SDPO	62.5	47.1	32.1	9.0	56.4	90.6	51.8	61.3	51.4
RLSD	65.0	51.3	37.1	11.8	59.8	93.4	55.0	72.0	55.7
SRPO-Route	62.1	51.3	34.2	11.8	60.3	92.9	51.0	74.0	54.7
SRPO-Reflect	56.7	39.6	27.9	6.8	55.8	91.4	51.0	72.0	50.1
SRT (Ours)	66.7	59.2	40.0	**16.0**	59.8	92.4	52.7	74.4	57.6
SD-ZERO (Ours)	**68.3**	**60.0**	**45.4**	**16.0**	60.4	**93.6**	**56.1**	**82.6**	**60.3**
Olmo-3-7B-Instruct									
Base Model	56.7	42.1	25.0	1.3	48.9	91.1	31.7	32.4	41.1
SFT	51.3	43.8	30.4	1.3	48.5	91.4	31.7	41.0	42.4
RFT	56.7	48.3	35.8	2.3	50.9	91.4	39.2	49.4	46.7
GRPO	54.6	43.8	25.8	4.8	**56.9**	91.8	37.1	43.6	44.8
OPSD	52.9	45.0	32.1	1.3	49.0	91.2	33.5	42.3	43.4
SDPO	52.1	44.2	31.3	1.3	48.5	90.8	32.0	45.5	43.2
RLSD	57.1	43.6	34.4	**7.6**	53.2	90.7	**47.9**	53.0	48.4
SRPO-Route	58.8	44.1	28.8	5.7	55.2	90.4	34.5	44.6	45.3
SRPO-Reflect	52.8	33.4	21.2	2.9	49.7	90.0	32.6	37.2	40.0
SRT (Ours)	59.2	52.9	39.6	3.5	52.4	92.3	42.8	**59.6**	50.3
SD-ZERO (Ours)	**61.7**	**53.8**	**40.4**	5.5	55.3	**94.0**	43.5	57.8	**51.5**

Table 1: **Performance comparison of SD-ZERO and SRT (Self-Revision Training, Phase 1) against standard post-training and on-policy self-distillation baselines** on math and code reasoning benchmarks, reported as avg@8. Across both Qwen3-4B-Instruct and Olmo-3-7B-Instruct, SRT and especially SD-ZERO achieve the strongest overall results; "LCB" denotes LiveCodeBench. All methods are compute equalized (see Appendix C.2 for more details).

7.5K from the Python (`solutions_py`) subset. We include more details on training data curation in Appendix C.1.

Evaluation. We evaluate on eight benchmarks in math and code domains: competition math (AIME24 (Zhang & Math-AI, 2024), AIME25 (Zhang & Math-AI, 2025), HMMT25 (Balunović et al., 2025), MATH (Hendrycks et al., 2021)), olympiad math (AMOBench (An et al., 2025) and OpenR1-Math (Hugging Face, 2025)), and competitive programming (Codeforces (Penedo et al., 2025) and LiveCodeBench (Jain et al., 2024)). For the two in-distribution datasets we hold out 500 test questions each.

Baselines. All baselines train on the same 15K questions. (1) *Standard post-training methods* include **SFT** on DeepSeek-R1 demonstrations (Guo et al., 2025), **RFT** (Yuan et al., 2023), and **GRPO** (Shao et al., 2024)[2]. (2) *On-policy self-distillation methods* use the model as its own teacher, conditioned on extra context: **OPSD** on high-quality demonstrations (Zhao et al., 2026)[3], **SDPO** on successful self-generated responses (Hübotter et al., 2026), **RLSD** on the ground-truth answer (Yang et al., 2026), **SRPO-Route** on successful self-generated responses for failed rollouts only (Li et al., 2026), and **SRPO-Reflect** on the model's reflection on its own rollout (Liu et al., 2026)[4]. Appendix C.2 compares sampling budgets, and Appendix C.3 lists training hyperparameters.

[2]We use DAPO (Yu et al. (2025)), an improved and commonly used variant of GRPO.

[3]SDFT (Shenfeld et al., 2026) is highly similar to OPSD algorithmically, so we only run one of the two.

[4]Li et al. (2026) and Liu et al. (2026) both name their method SRPO; we add suffixes to distinguish them.

3.2 SRT Outperforms Training on Expert Solutions or Filtered Self-Generations

We first compare SRT against training on high-quality demonstrations (SFT) and filtered correct self-generations (RFT) (Table 1). Trained on only 6K self-revision responses, SRT improves average accuracy by 7.8% on Qwen3-4B-Instruct and 9.2% on Olmo-3-7B-Instruct, substantially outperforming both SFT and RFT trained on 15K examples. The baselines fail for different reasons: SFT on DeepSeek-R1 responses degrades Qwen3-4B-Instruct on benchmarks such as AMOBench and LiveCodeBench, while RFT shows minimal gains on harder tasks such as AMOBench.

SRT is effective because it supervises error correction: each example pairs an on-policy incorrect attempt with a self-corrected revision; we find that correctness filtering of revision traces is important (Appendix G). Unlike RFT, which removes incorrect reasoning entirely, SRT preserves the failed attempt as context, allowing the model to learn from its own mistakes. This structure yields especially strong gains on harder benchmarks such as AIME25, HMMT25, and LiveCodeBench. Data quality does not explain the gain: on the same 6K traces, SFT on y_{revised} trails SRT by 2.2 points on Qwen3-4B-Instruct and 5.3 on Olmo-3-7B-Instruct (Appendix D.5). SRT also improves Qwen3-14B (non-thinking) by 3.6 points (Appendix D.7).

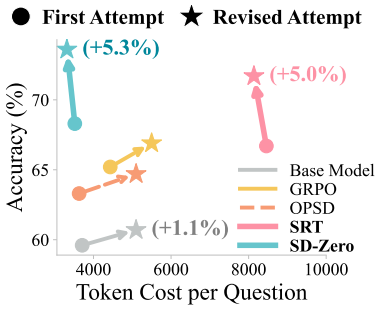


Figure 3: **Comparison of outcome-conditioned self-revision capability** on AIME24, Qwen3-4B-Instruct. SRT unlocks self-revision behaviors, while final SD-Zero model preserves this advantage while improving token efficiency.

SRT Significantly Boosts Self-Revision Capability. To measure directly, we run a Generate-then-Revise evaluation on 1K AIME24 samples using Qwen3-4B-Instruct: (1) *First Attempt*: sample one response per question, (2) *Revised Attempt*: prompt the model to generate a revision conditioned on the first attempt and its binary correctness reward. As shown in Figure 3 (detailed statistics in Table 9), while the base model gains only 1.1% from revision, SRT gains +5.0%, indicating that the SRT model has learned to use outcome reward to correct its first attempt rather than simply resampling a new attempt. Notably, the revision responses produced by SRT are, on average, even shorter than the model's first attempts, suggesting more targeted and proactive reasoning during revision; in contrast, revisions from the base model tend to be longer and likely more redundant. More interestingly, the self-revision capability of the model improves further after the Self-Distillation phase. This observation is central to the self-evolving nature of SD-Zero, which we discuss further in Section 3.4.

3.3 Self-Distillation Distills Revision into Token-Efficient Generation

After SRT, the model as a generator tends to produce longer responses (as shown in Figure 3) with significantly more self-revision behaviors. Self-Distillation phase distills the self-revision behavior of the model back into more proactive responses of the generator.

Self-Distillation Enables Stronger, Token-Efficient Generations. Table 1 shows that Self-Distillation phase adds 2.7% on Qwen3-4B-Instruct and 1.2% on Olmo-3-7B-Instruct beyond SRT, for total gains of +10.5% and +10.4% over the base models. For Qwen3-4B-Instruct, the gains concentrate on code benchmarks and HMMT25. For Olmo-3-7B-Instruct, the Self-Distillation phase mainly improves on math benchmarks such as AIME24 and AMOBench. We also report pass@8 results in Table 7. Interestingly, pass@8 improves substantially as well, suggesting that SD-Zero does not merely sharpen the output distribution, a behavior often associated with RL-based methods such as GRPO (Yue et al., 2025).

Comparison to SRT. Since most of the performance gains in SD-Zero arise during the SRT phase, one may question the importance of Self-Distillation. We argue that Self-Distillation is nevertheless a critical component of the SD-Zero pipeline for two reasons. First, Self-Distillation reduces response length at inference by at least 2×, so SD-Zero generates fewer tokens than SRT, the base model, GRPO, and OPSD while achieving the strongest overall performance (Figure 3; Section 4.2). Second, Self-Distillation is highly sample-efficient during training. Constructing the SRT dataset requires collecting multiple sampled responses per question in order to obtain successful revisions,

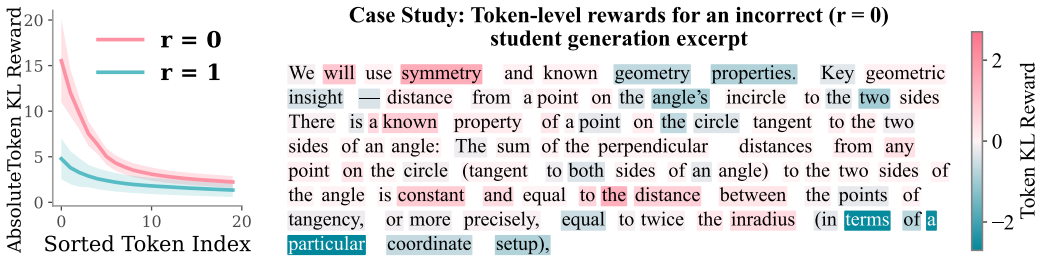


Figure 4: **Reviser converts binary outcome reward into dense token-level reward. Left:** Comparison of token-level KL reward distributions for correct ($\mathbf{r} = \mathbf{1}$) and incorrect ($\mathbf{r} = \mathbf{0}$) student generations. Incorrect trajectories concentrate larger rewards on a small number of tokens, whereas correct trajectories receive a flatter reward distribution that mainly preserves the response. **Right:** Visualization of token-level KL rewards for an incorrect student generation ($\mathbf{r} = \mathbf{0}$). The teacher policy converts binary reward $r = 0$ to dense rewards that localize mistake-relevant tokens and guide targeted correction.

whereas Self-Distillation requires only a single response per question and uses the reviser's token-level feedback to provide dense supervision. We provide a detailed sample budget analysis in Appendix C.2.

Comparison to RL and Self-Distillation Baselines. Under a matched generation budget, SD-ZERO outperforms GRPO by at least 6.7 points on average and every self-distillation baseline by at least 3.1 points (Table 1). OPSD needs gold solutions, and with final answers only it barely improves over the base model (Table 8); SD-ZERO needs only a binary reward. GRPO samples a group of responses per question, whereas the Self-Distillation phase uses one. While we note that GRPO may benefit from additional training epochs, our comparison is under a matched single-epoch budget, where SD-ZERO achieves stronger performance with comparable total sample cost (Appendix C.2). The gap holds under tighter controls: trained only on hard questions, SD-ZERO outperforms OPSD by 2.8 points and SDPO by 3.7 points, and GRPO restricted to the 6K SRT questions trails SRT by 4.4 points (Appendices D.5 and D.6).

Takeaway: SD-ZERO offers the best performance per model generations among the compared methods, while requiring neither an external teacher nor high-quality demonstrations.

3.4 ITERATIVE SELF-EVOLUTION: TEACHER SYNCHRONIZATION ENABLES CONTINUED GAINS

In SD-ZERO, the teacher is fixed as the SRT model throughout Phase 2. Performance eventually plateaus because the supervision signal is bounded by a stale teacher. But Self-Distillation phase also improves the model's revision capability: the SD-ZERO model achieves a 5.3% revision gain on the Generate-then-Revise evaluation, comparable to SRT's 5.0% (Figure 3). The improved student can therefore serve as a stronger teacher, enabling what we call ***iterative self-evolution***. The capability to revise answers is distilled back into generation, and regular teacher synchronization allows the loop to sustain.

We test this by synchronizing the teacher with the student after the first round of 500 steps. On OpenR1-Math with Qwen3-4B-Instruct, the second round raises accuracy on training questions from 55.3 to 57.5 (Figure 5). Two further synchronizations add 0.5 and 0.3 points, so gains saturate by the third round (Appendix D.8). Each round needs only the student's attempt and its binary reward in the teacher's context.

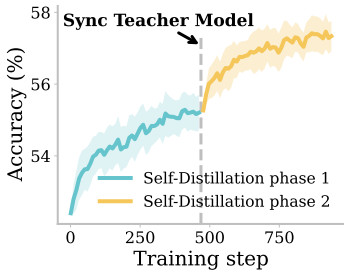


Figure 5: **Self-evolved reasoning through teacher synchronization in SD-ZERO.** SD-ZERO can iteratively improve the model by reusing its own learned self-revision behavior as supervision. The two curves are the first and second rounds of the Self-Distillation phase.

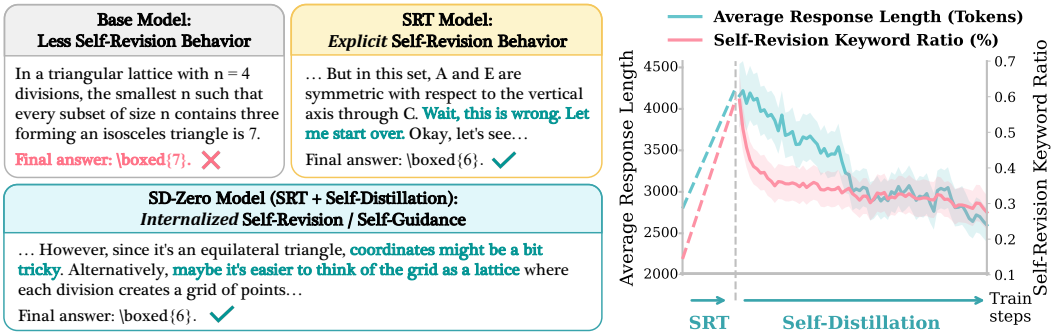


Figure 6: **Evolution of self-revision behavior across training phases** for Qwen3-4B-Instruct on OpenR1-Math. **Left:** Example reasoning traces from different model phases. The base model fails without revision, whereas the SRT model performs explicit self-revision and the SD-ZERO model exhibits more internalized self-guidance, identifying pitfalls and directing itself to the correct answer. **Right:** Training dynamics of self-revision behavior. Response length and explicit self-revision keywords rise during the SRT phase and fall during the Self-Distillation phase, indicating a shift from overt revision to more internalized, token-efficient reasoning.

4 UNDERSTANDING HOW SD-ZERO IMPROVES REASONING

We analyze why SD-ZERO works: the reviser gives localized token-level feedback (Section 4.1), the generator learns shorter reasoning (Section 4.2), and each design choice matters (Section 4.3).

4.1 TOKEN-LEVEL SELF-LOCALIZATION: THE REVISER IDENTIFIES WHICH TOKENS TO CORRECT

Although the reviser receives only a binary outcome $r \in \{0, 1\}$, its feedback concentrates on a small subset of tokens, a property we call *token-level self-localization*. We decompose the Self-Distillation loss into token-wise terms

$$D_{\mathrm{KL}}^{(t)} := D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x, y_{<t}) \,\big\|\, \pi_{\theta_{\mathrm{SRT}}}(\cdot \mid x, y, P_r, y_{<t})\big),$$

which is approximately the log-probability gap between the generator and reviser at token t. We define **Token KL Reward** at token t as: $\log \pi_\theta(y_t \mid x, y_{<t}) - \log \pi_{\theta_{\mathrm{SRT}}}(y_t \mid x, y, P_r, y_{<t})$.

To quantify how this signal is distributed, we sort tokens within each response by $D_{\mathrm{KL}}^{(t)}$, divide them into 20 equal-sized buckets, and average within each bucket across 200 responses, separately for correct ($r = 1$) and incorrect ($r = 0$) generations. As shown in Figure 4, the distributions differ sharply: for incorrect responses, the top 5% of token positions carry about half of the KL mass, whereas for correct responses it is much more uniform.

Figure 4 (right) illustrates this on one incorrect generation. Tokens corresponding to the faulty symmetry-based argument receive large positive KL, while tokens associated with the correct coordinate-based solution receive large negative KL. Thus, the reviser does more than penalize incorrect reasoning: it both localizes the error and redirects the model toward a better alternative, converting a scalar outcome into a two-sided token-level signal.

4.2 EVOLUTION OF REASONING BEHAVIORS: SD-ZERO INTERNALIZES SELF-REVISION

As shown in Figure 3, the SD-ZERO model after Self-Distillation training produces substantially shorter generations, suggesting that it might have internalized some formerly explicit reasoning. Here, we take a closer look at how this generation behavior evolves over SD-ZERO training. Figure 6 (Right) tracks two statistics: average response length and frequency of explicit self-revision keywords (e.g., "Wait" "let me start over"; see Appendix E.1 for the full keyword list). During the SRT phase, both metrics rise substantially. The model learns to solve problems by actively judging its own response and applying self-correction. During Self-Distillation phase, the trend reverses: both response length and self-revision keyword frequency fall steadily, while accuracy continues to improve. The resulting model only uses approximately half as many tokens as the SRT model at better accuracy. The reviser signals where to stop: at the position where its own revision would

end, it places probability 0.42 on the end-of-sequence token, versus 0.03 for the student, and the per-token KL after that position is $3.1\times$ higher than before it (Appendix E.3.2).

We qualitatively illustrate this two-stage pattern via example traces from different models in Figure 6 (Left). The base model answers with a false claim. The SRT model reached the correct answer by backtracking with the exact transition phrase seen in training ("Wait, this is wrong. Let me start over."). By contrast, the SD-ZERO model, without any backtracking, anticipates the same pitfall and directs proactively to the correct answer. Note that Self-Distillation does not reduce the revision capability acquired in SRT. The model still achieves a 5.3% gain in the Generate-then-Revise evaluation (Figure 3), but it has successfully internalized some of that capability into a more directed, pitfall-aware attempt.

4.3 ABLATIONS ON SD-ZERO DESIGN

Loss Terms in SRT Objective. Training with only $\mathcal{L}_{\text{generation}}$ or only $\mathcal{L}_{\text{revision}}$ underperforms full SRT (Table 15). The two terms are complementary: $\mathcal{L}_{\text{revision}}$ elicits self-revision, and $\mathcal{L}_{\text{generation}}$ transfers it to stronger generation.

Self-Distillation Phase without SRT. SD-ZERO requires the model to have certain self-revision ability, which is unlocked only after the SRT phase (Figure 3). To further demonstrate its necessity, we apply the Self-Distillation phase directly to Qwen3-4B-Instruct without prior self-revision training. Table 16 shows that this Phase-2-only variant yields only marginal improvements over the base model (49.8 to 51.4), and does not improve self-revision ability (correction rate 2.7% vs. 2.6%). Therefore, the SRT phase is necessary for SD-ZERO to be effective.

Data Split Between Phases. With the total data fixed, giving more questions to SRT slightly improves the SRT model but does not improve the final performance (Table 17). Once SRT has unlocked self-revision capability, the Self-Distillation phase uses additional data more effectively.

Self-Distillation Objective and Teacher. Reverse KL (57.3 math average) outperforms forward KL with importance sampling (56.2) and SFT on teacher-generated revisions (56.8). A teacher conditioned on a fixed correct self-revision, instead of the student's current attempt, reaches 56.5, and the SD-ZERO teacher stays closer to the student than OPSD's (per-token NLL 0.115 vs. 0.270; Appendices E.2.4 and E.3.1).

5 DISCUSSION

Conclusion. SD-ZERO shows that a model's revision ability, elicited with a small amount of supervised training on self-generated revisions, can transform sparse binary outcome reward into dense token-level supervision without requiring an external teacher or high-quality demonstrations. Across two model families and eight benchmarks, this yields consistent $10\%+$ gains over the base models and outperforms all baselines under the same training sample budget (Table 1). We also show experiments on broader tasks and models in Appendix G. Extensive ablations and analyses show that SD-ZERO also leads to improved revision capability which can lead to further gains on accuracy. We hope that by eliminating the need for high-quality demonstrations, we enable future work to apply self-distillation to broader domains.

Limitations and Open Directions. Our study focuses on instruct models that generate short and concise responses, consistent with prior self-distillation works. An important next step is extending self-distillation to *thinking models*, which produce long, exploratory chains of thought. This is challenging because such responses may include false starts and partial corrections that are not mistakes, making it difficult to distinguish productive exploration from genuine errors and assign credit beyond local token decisions. We provide preliminary evidence of this challenge in Section F: applying OPSD to Qwen3-4B with thinking enabled during training hurts performance on competition math benchmarks.

Furthermore, we focus on verifiable domains such as math and coding. Extending SD-ZERO to domains without verifiable rewards remains an open problem. One promising direction is to define rewards using meta-cognitive signals (Didolkar et al., 2024; 2025; Shao et al., 2025; He et al., 2025a;b), such as consistency or self-correction. We leave this to future work.

AI USE STATEMENT

We used generative AI tools for the following tasks: writing and debugging code; searching related literature and identifying candidate baselines; analyzing experimental results and producing plots; rephrasing and polishing text, including turning author-written notes into paper prose; checking numerical consistency across tables and text; and LaTeX formatting. We did not use generative AI tools to develop the research ideas or the SD-ZERO method. All experiments were run by the authors, and every reported number, table, and figure comes from our own experiments. The authors reviewed and verified all AI-assisted content and take full responsibility for this submission, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All authors of this work have read and agree to abide by the ICLR Code of Ethics. This research develops post-training methods for language models in verifiable reasoning domains, specifically math and code, and evaluates them on established public benchmarks and publicly available training data. In particular, our experiments use OpenR1-Math and Codeforces for training, and evaluate on math and code benchmarks including AIME24, AIME25, HMMT25, AMOBench, MATH, OpenR1-Math, Codeforces, and LiveCodeBench. These settings rely on publicly available problem/solution data and automatic correctness verification based on final answers or program outputs, and do not involve human subjects, private user data, or personally identifiable information.

Our method, SD-ZERO, is intended to improve sample-efficient post-training in verifiable domains by converting binary outcome rewards into dense token-level self-supervision, without requiring an external teacher or high-quality demonstrations. The paper does not study deployment in socially sensitive domains, nor does it introduce new human data collection. At the same time, as with language-model post-training more broadly, improvements in reasoning and generation capabilities may have downstream dual-use implications beyond the domains studied here. We therefore emphasize that our experiments are limited to math and coding tasks with automatic verification, and we do not claim that the method is safe or appropriate for high-stakes or non-verifiable applications without further study.

REPRODUCIBILITY STATEMENT

We provide a detailed description of SD-ZERO, including the two-phase training procedure (Self-Revision Training and Self-Distillation), the corresponding objectives, and algorithmic details in the main text and appendix. We also specify the model backbones, sampling settings, data splits, training data curation procedure, compute-equalized baseline comparisons, and optimization hyperparameters used in all experiments. Our training and evaluation datasets are identified explicitly, with held-out test splits for the in-distribution datasets and evaluation across eight math and code benchmarks. Since SD-ZERO assumes only access to ground-truth final answers and a binary verifier rather than gold solutions, these supervision assumptions are also stated precisely. Together, these details are intended to make the method and experimental setup fully reproducible.

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.

Shengnan An, Xunliang Cai, Xuezhi Cao, Xiaoyu Li, Yehao Lin, Junlin Liu, Xinxuan Lv, Dan Ma, Xuanlin Wang, Ziwen Wang, and Shuang Zhou. Amo-bench: Large language models still struggle in high school math competitions, 2025. URL https://arxiv.org/abs/2510.26768.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. URL https://arxiv.org/abs/2112.00861.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, February 2025. URL https://matharena.ai/.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards cross-tokenizer distillation: the universal logit distillation loss for LLMs. *Transactions on Machine Learning Research*, 2025.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025a.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting, 2025b. URL https://arxiv.org/abs/2510.18874.

Aniket Didolkar, Anirudh Goyal, Nan R Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812, 2024.

Aniket Didolkar, Nicolas Ballas, Sanjeev Arora, and Anirudh Goyal. Metacognitive reuse: Turning recurring llm reasoning into concise behaviors. *arXiv preprint arXiv:2509.13237*, 2025.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars, 2025. URL https://arxiv.org/abs/2503.01307.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong,

Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Alex Havrilla, Sharath Chandra Raparthy, Christoforos Nalmpantis, Jane Dwivedi-Yu, Maksym Zhuravinskyi, Eric Hambro, and Roberta Robeyns. GLoRE: When, where, and how to improve LLM reasoning via global and local refinements. In *International Conference on Machine Learning*, 2024.

Yinghui He, Abhishek Panigrahi, Yong Lin, and Sanjeev Arora. Adaptmi: Adaptive skill-based in-context math instruction for small language models, 2025a. URL https://arxiv.org/abs/2505.00147.

Yinghui He, Abhishek Panigrahi, Yong Lin, and Sanjeev Arora. Skill-targeted adaptive training, 2025b. URL https://arxiv.org/abs/2510.10023.

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, 2021. URL https://arxiv.org/abs/2103.03874.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

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

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Ryo Kamoi, Yusen An, Nan Zhang, Jiaqi Gao, and Rui Zhang. When can LLMs actually correct their own mistakes? A critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 2024.

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.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, JD Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

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*, 2026.

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

Jialong Liu, Yuling Shi, Ning Yang, Xiaodong Gu, and Zuchao Li. SRPO: Self-reflective policy optimization for long-horizon reasoning. In *Forty-third International Conference on Machine Learning*, 2026.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Benjamin Minixhofer, Ivan Vulić, and Edoardo Maria Ponti. Universal cross-tokenizer distillation via approximate likelihood matching. In *Advances in Neural Information Processing Systems*, 2025.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Guilherme Penedo, Anton Lozhkov, Hynek Kydlíček, Loubna Ben Allal, Edward Beeching, Agustín Piqueres Lajarín, Quentin Gallouédec, Nathan Habib, Lewis Tunstall, and Leandro von Werra. Codeforces. https://huggingface.co/datasets/open-r1/codeforces, 2025.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. CRISP: Compressed reasoning via iterative self-policy distillation, 2026. URL https://arxiv.org/abs/2603.05433.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

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.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023. URL https://arxiv.org/abs/2308.01825.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation, 2019. URL https://arxiv.org/abs/1905.08094.

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

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

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.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

APPENDIX CONTENT

A Related Work 16

B Method Design 17
- B.1 SD-ZERO Training Algorithm 17
- B.2 Comparison to Existing Methods 17

C Experiment Details 19
- C.1 Curating SD-ZERO training data 19
- C.2 Comparing Sampling Budgets 19
- C.3 Hyperparameters 20
- C.4 Baseline Details 20

D Additional Results 24
- D.1 Pass@8 Performance on Math Benchmarks 24
- D.2 OPSD Under Final-Answer-Only Supervision 24
- D.3 Detailed Self-Revision Statistics 24
- D.4 GRPO Configuration Comparison 24
- D.5 Controlled Baselines on the Same Revision Data 26
- D.6 Self-Distillation on Hard Questions 26
- D.7 Scaling SRT to Qwen3-14B 26
- D.8 Multiple Rounds of Teacher Synchronization 27

E Analysis of SD-ZERO 28
- E.1 Self-Revision Keyword Analysis 28
- E.2 Ablation Studies on SD-ZERO Method Design 28
- E.3 Mechanisms of the Self-Distillation Phase 32

F OPSD with Thinking Models 33

G Self-Revision Design Choices 34
- G.1 Effect of Training Data Source 34
- G.2 Self-Revision as Initialization for RL 35

A Related Work

On-Policy Distillation and Self-Distillation. Knowledge distillation trains a student to match a teacher's output distribution either at the token level via soft targets (Hinton et al., 2015) or at the sequence level (Kim & Rush, 2016). However, such off-policy training introduces a train-inference distribution mismatch that leads to compounding errors at test time (Bengio et al., 2015; Chen et al., 2025b). On-policy distillation (OPD) (Agarwal et al., 2024; Gu et al., 2024; Boizard et al., 2025; Minixhofer et al., 2025) addresses this by using student rollouts with per-token supervision from the teacher.

A key limitation of OPD is the need for an external model during training to score student traces. Recent *self-distillation* methods (Furlanello et al., 2018; Zhang et al., 2019) eliminate this by conditioning the student on privileged information to serve as its own teacher (Askell et al., 2021; Snell et al., 2022), where the privileged context can include reasoning traces (Zhao et al., 2026), in-context demonstrations (Shenfeld et al., 2026), or conciseness prompts (Sang et al., 2026). Still, such works assume access to high-quality signals from external sources that are often costly to obtain, and those that remove the need for an external teacher entirely require rich, dense environment feedback (Hübotter et al., 2026). In contrast, SD-Zero only requires binary correctness signals of the student's response and generates its own supervision through self-revision.

Self-Training and Self-Refinement. Self-training methods improve models by iteratively generating and filtering their own data. Prior work bootstraps reasoning by fine-tuning on self-generated rationales that yield correct answers (Zelikman et al., 2022; Yuan et al., 2023; Singh et al., 2024; Yuan et al., 2024). In particular, STaR (Zelikman et al., 2022) iteratively generates and filters for correct rationales, discarding incorrect reasoning entirely. By contrast, SRT retains the incorrect attempt as context and trains on the full mistake-to-correction trajectory. Consistent with this, recent evidence suggests that productive reasoning behaviors (e.g., verification, backtracking, etc.) matter more for self-improvement than answer correctness (Gandhi et al., 2025). Separately, prompting-based self-refinement methods show that LLMs are capable of critiquing and revise their own outputs at inference time (Madaan et al., 2023; Shinn et al., 2023), though such approaches yield limited gains without external feedback and do not update model weights (Huang et al., 2023; Kamoi et al., 2024). Training-based approaches address this by using RL or supervised learning to internalize correction behavior (Kumar et al., 2024; Havrilla et al., 2024), but still require multi-turn generation at inference time. SD-Zero bridges both avenues by training the model to self-revise and then distilling the revision-informed distribution back into single-pass generation, eliminating multi-turn correction at test time.

Connection to Process Reward Models The per-token KL signal from the reviser is functionally analogous to process reward models (PRMs) (Lightman et al., 2024; Wang et al., 2024), in that both provide localized supervision over intermediate reasoning rather than only the final outcome. PRMs train a separate model to assign step-level scores to individual reasoning steps. In SD-Zero, the token-level divergence $D_{KL}^{(t)}$ similarly concentrates on a sparse subset of tokens associated with errors in the response (Figure 4), so that the log-probability gap at each token both penalizes erroneous reasoning and redirects probability mass towards plausible alternatives. The main difference is cost and supervision: PRMs typically require either step-level correctness annotations (Lightman et al., 2024) or rollout-based estimates from search or sampling (Wang et al., 2024), together with a separately trained reward model, whereas our reviser's signal arises by conditioning the same model on its own response and a binary outcome, without step-level annotation or an auxiliary reward model. Whether this implicit token-level signal can complement or serve as a cheaper alternative to trained PRMs in settings such as guided tree search remains an interesting direction for future work.

B METHOD DESIGN

B.1 SD-ZERO TRAINING ALGORITHM

Algorithm 1 summarizes the full two-phase training procedure of SD-ZERO (Section 2). In Phase 1 (SRT), the model first learns outcome-conditioned self-revision: for each sampled response, we use the binary verifier to determine whether the response is correct, construct a control prompt that either triggers rephrasing (r=1) or restarting (r=0), and keep only revised responses that are verified as correct to form the self-revision dataset $\mathcal{D}_{\text{REVISION}}$. The model is then trained on this dataset with the SRT objective, which jointly improves its ability to revise and to generate stronger responses. In Phase 2 (Self-Distillation), we initialize the student from the SRT model and perform on-policy self-distillation: the current student generates a response, the fixed SRT reviser conditions on that response and its binary outcome to define a teacher distribution, and the student is updated to match this distribution via KL minimization. In this way, SD-ZERO first elicits explicit self-revision behaviors and then distills them back into more compact, token-efficient generation.

Algorithm 1 SD-ZERO Training Pipeline

Require: Base model π_θ, dataset $\mathcal{D} = \{(x, a)\}_{i=1}^{N}$, binary verifier $r(y, a) \in \{0, 1\}$

▷ **Phase 1 (SRT): Self-Revision Training**

for each $(x, a) \in \mathcal{D}$ **do**
  Sample initial response $y_{\text{init}} \sim \pi_\theta(\cdot \mid x)$
  Compute reward $r \leftarrow r(y_{\text{init}}, a)$
  Construct control prompt

$$P_r \leftarrow \begin{cases} \text{“Let me rephrase the above solution.”}, & r = 1, \\ \text{“Wait, this response is not correct, let me start over.”}, & r = 0 \end{cases}$$

  Generate revised response $y_{\text{revised}} \sim \pi_\theta(\cdot \mid x, y_{\text{init}}, P_r)$
  Add $(x, y_{\text{init}}, P_r, y_{\text{revised}})$ to $\mathcal{D}_{\text{REVISION}}$ **if** $r(y_{\text{revised}}, a) = 1$
end for
Train π_θ on $\mathcal{D}_{\text{REVISION}}$ with $\mathcal{L}_{\text{SRT}}$ to obtain $\pi_{\theta_{\text{SRT}}}$

▷ **Phase 2 (Self-Distillation): On-Policy Self-Distillation via Revision Feedback**

Initialize $\pi_\theta \leftarrow \pi_{\theta_{\text{SRT}}}$
repeat
  Sample $(x, a) \sim \mathcal{D}$
  Sample student response $y_{\text{init}} \sim \pi_\theta(\cdot \mid x)$
  Compute reward $r \leftarrow r(y_{\text{init}}, a)$
  Construct control prompt P_r as in Phase 1
  Define student policy $\pi_S(\cdot) \leftarrow \pi_\theta(\cdot \mid x)$
  Define teacher policy $\pi_T(\cdot) \leftarrow \pi_{\theta_{\text{SRT}}}(\cdot \mid x, y_{\text{init}}, P_r)$
  Update θ by minimizing

$$\mathcal{L}_{\text{Self-Distillation}}(\theta) = D_{\text{KL}}(\pi_S \,\|\, \text{StopGrad}(\pi_T))$$

until convergence

B.2 Comparison to Existing Methods

Method	Sampling	Signal	Teacher	Teacher can condition on wrong attempt
SFT / Distillation	× off-policy	✓ dense	× external	—
On-Policy Distillation	✓ on-policy	✓ dense	× external	—
RLVR (e.g., GRPO)	✓ on-policy	× sparse	—	—
OPSD / SDFT / SDPO	✓ on-policy	✓ dense	✓ self	×
SD-Zero (ours)	✓ on-policy	✓ dense	✓ self	✓

Table 2: Comparison of post-training methods for LLMs. SD-Zero is the only self-distillation method whose teacher can condition on incorrect student attempts, enabling it to convert sparse binary rewards into dense token-level supervision without requiring gold demonstrations.

Table 2 situates SD-Zero among existing post-training methods. Standard SFT and distillation rely on dense supervision but are off-policy and require external demonstrations, while RLVR methods such as GRPO are on-policy but optimize only sparse outcome rewards. Recent self-distillation methods such as OPSD, SDFT, and SDPO combine on-policy sampling with dense self-supervision, but their teacher does not explicitly condition on the student's failed trajectory. In contrast, SD-Zero is the only method in this comparison that is simultaneously on-policy, self-distilled, and able to condition the teacher on an incorrect attempt, allowing it to transform binary outcome rewards into targeted token-level supervision without requiring gold reasoning traces.

C EXPERIMENT DETAILS

C.1 CURATING SD-ZERO TRAINING DATA

In SD-ZERO, we trained the models on 15K training data from OpenR1-Math and 15K data from Codeforces separately, with 6K in Phase 1 (SRT) and 9K in Phase 2 (Self-Distillation).

In SRT phase, we curate 6K self-revision training data from 10K question-answer pairs in OpenR1-Math (or Codeforces), through the following pipeline:

1. **Initial sampling:** Select the first 10K questions in OpenR1-Math (or Codeforces), and sample 1 initial model responses y_{initial} per question x,
2. **Verification:** Verify the binary reward $r \in \{0, 1\}$ for each y_{initial}, and build self-revision prompt P_r. The 10K initial responses are roughly split into 5K correct and 5K incorrect responses.
3. **Self-Revision:** For each correct initial response, prompt the model to generate 3 rephrased responses y_{revised}; For each incorrect initial response, prompt the model to generate 3 corrected responses y_{revised},
4. **Filtering:** Keep traces $(x, y_{\text{initial}}, P_r, y_{\text{revised}})$ where y_{revised} reaches a correct final answer. The resulting training data contain 6K self-revision traces.

In Self-Distillation phase, we directly sample an additional 9K question-answer pairs from OpenR1-Math (or Codeforces) as training data.

C.2 COMPARING SAMPLING BUDGETS

One feature of SD-ZERO is sample efficiency. In Table 3, we compare the sampling budget of SD-ZERO with baseline on-policy training methods, including rejection fine-tuning (RFT), GRPO, and OPSD. We calculate the sampling budget of each methods as follows:

- **RFT:** total # generations = 15K (questions) $\times$ 4 (response attempts per question) = 60K.
- **GRPO:** total # generations = 15K $\times$ 4 (rollout/question) = 60K.
- **OPSD:** total # generations = 15K $\times$ 4 (rollout/question) = 60K.
- **SRT phase:** total # generations = 10K (questions) $\times$ 1 (initial response attempts per question) + 5K (correct initial responses) $\times$ 3 (rephasing attempts per correct initial response) + 5K (incorrect initial responses) $\times$ 3 (self-correction attempts per incorrect initial response) = 40K.
- **Self-Distillation phase:** total # generations = 9K $\times$ 1 (rollout/question) = 9K.
- **SD-ZERO:** 40K (SRT phase) + 9K (Self-Distillation phase) = 49K.

We trained the GRPO baseline for 1 epoch on 15K training qesutions, with 4 generations per question [5].

We also estimate the total token budget. At approximately 3.7K tokens per response, the sampling budgets for RFT and GRPO are each roughly 222M tokens. For SD-ZERO, the SRT data collection phase generates approximately 148M tokens (10K initial responses plus 30K revisions), and the Self-Distillation phase generates at most 76.5M tokens (9K rollouts at up to 8.5K tokens each, though rollout length decreases during training). The total sampling budget for SD-ZERO is thus at most 225M tokens, comparable to baselines despite achieving substantially stronger performance.

Axis 1: Sampling budget (total completion tokens generated).

- **RFT:** 60K $\times$ 3.7K $\approx$ 222M tokens
- **GRPO:** 60K $\times$ 3.7K $\approx$ 222M tokens

[5] We also compare to the 8-rollout GRPO variants in Table 10.

Method	# Training questions	# Training rollouts per question	# Training epochs	**Total # generations**
RFT	15K	-	1	**60K**
GRPO	15K	4	1	**60K**
OPSD	15K	4	1	**60K**
SRT phase	6K	-	1	**40K**
Self-Distillation phase	9K	1	1	**9K**
SD-ZERO	15K	-	1	**49K**

Table 3: Comparison of training data and generation budgets across different training paradigms. For RFT, GRPO, and OPSD, we assume 15K training questions and 4 sampled rollouts per question, resulting in 60K total generations before any filtering. In our method, the SRT phase starts from 10K seed questions, generates 1 initial response and 3 self-revisions per question (40K generations in total), and retains 6K correct self-revision traces for training. The Self-Distillation phase directly uses 9K additional question-answer pairs, requiring 9K generations. In total, SD-ZERO uses 15K training questions and 49K generations.

- **SD-ZERO Phase 1 (SRT data collection):**
  - 10K initial responses $\times$ 3.7K $=$ 37M
  - 30K revisions $\times$ 3.7K $=$ 111M
  - Subtotal: 148M tokens
- **SD-ZERO Phase 2 (Self-Distillation):**
  - 9K rollouts $\times$ 8.5K $\leq$ 76.5M tokens (upper bound; rollout length decreases during this phase per Figure 6)
- **SD-ZERO total:** $\leq$ 224.5M tokens

Axis 2: Training budget (completion tokens in forward pass for loss computation).

- **GRPO:** 60K $\times$ 3.7K $\approx$ 222M tokens (probably more than this since avg response length increases to 4.4K by the end of GRPO)
- **RFT:** $\leq$ 222M (same sampling budget as GRPO but with incorrect samples dropped)
- **SD-ZERO SRT training** (6K traces, two losses per trace):
  - $\mathcal{L}_{\text{revision}}$: 6K $\times$ 3.7K $=$ 22M
  - $\mathcal{L}_{\text{generation}}$: 6K $\times$ 7.4K $=$ 44M (uses $2 \times 3.7K$ since $y' = [y_{\text{init}}, P_r, y_{\text{revised}}]$)
  - Subtotal: 66M tokens
- **SD-ZERO Self-Distillation** (student + teacher forward pass per step):
  - Student: 9K $\times$ 8.5K $=$ 76.5M
  - Teacher: 9K $\times$ 8.5K $=$ 76.5M
  - Subtotal: $\leq$ 153M tokens (probably less than this since avg response length goes down during self-distillation phase)
- **SD-ZERO total:** $\leq$ 219M tokens

C.3 HYPERPARAMETERS

We attach the hyperparameters used for Qwen3-4B-Instruct in SFT, RFT, and SD-ZERO SRT phase in Table 4, those used in GRPO in Table 5, and those used in OPSD and SD-ZERO Self-Distillation phase in Table 6.

C.4 BASELINE DETAILS

On OpenR1-Math, SFT and OPSD use the DeepSeek-R1 solutions that ship with the dataset (Hugging Face, 2025); no other method sees these solutions. For SDPO, we follow its "Learning without Rich Environment Feedback" setting (Hübotter et al., 2026), the variant that matches our binary-reward setting.

Parameters	SFT
General	
Model	Qwen/Qwen3-4B-Instruct-2507
Thinking	False
Data	
Prompt format	Chat template (system + user)
Completion-only loss	True
Max. sequence length	32768
Max. response length	32768
Batching	
Per-device train batch size	1
Per-device eval batch size	1
Gradient accumulation steps	1
Effective global batch size	4
Optimization / Training	
Optimizer	AdamW
Learning rate	5×10^{-6}
Scheduler	Cosine
Warmup ratio	0.05
Num. epochs	3
Weight decay	1×10^{-4}
Adam β_1	0.9
Adam β_2	0.95
Precision	bfloat16
Gradient checkpointing	True
Gradient accumulation sync each batch	True
Use Liger kernel	True
Parallelism / FSDP	
FSDP mode	`full_shard auto_wrap`
FSDP wrapped layer	`Qwen3DecoderLayer`

Table 4: Hyperparameters used for SFT/ RFT baselines and SD-ZERO SRT phase on Qwen3-4B-Instruct.

Parameters	GRPO
General	
Model	Qwen/Qwen3-4B-Instruct-2507
Thinking	False
Data	
Max. prompt length	1024
Max. response length	8192
Train batch size	64
Validation batch size	–
Shuffle	True
Truncation	left
Batching	
Rollouts per prompt (n)	8
PPO mini-batch size	64
PPO micro-batch size per GPU	1
Logprob micro-batch size per GPU	1
Max. sequences per rollout batch	64
Max. batched tokens	16384
Rollout / Generation	
Inference engine	vLLM
Rollout mode	sync
Temperature	0.7
Top-p	1.0
Top-k	–
Sampling	True
Max new tokens	8192
Tensor parallel size	2
Enable chunked prefill	True
Actor / Policy Optimization	
Strategy	FSDP
PPO epochs	1
Clip ratio	0.2
Entropy coefficient	0.001
Use KL loss	True
KL loss coefficient	0.001
KL loss type	low-var KL
Gradient clip norm	1.0
Dynamic batch size	False
Reference Model	
Strategy	FSDP
Dynamic batch size	False
Sequence parallel size	1
Parallelism	
Number of GPUs	8
Number of nodes	1
Sequence parallel	1
Actor strategy	FSDP
Reference strategy	FSDP
Gradient checkpointing	True
Training	
Optimizer	AdamW
Learning rate	1×10^{-6}
Scheduler	Constant
Warmup steps ratio	0.0
Weight decay	0.01
Precision	bfloat16
NCCL timeout	1800

Table 5: Hyperparameters used in GRPO baseline on Qwen3-4B-Instruct.

Parameters	SD-ZERO	OPSD
General		
Policy Model	Qwen/Qwen3-4B-Instruct-2507	Qwen/Qwen3-4B-Instruct-2507
Teacher model	Qwen/Qwen3-4B-Instruct-2507	Qwen/Qwen3-4B-Instruct-2507
Thinking	False	False
Data		
Max. teacher prompt length	32768	32768
Max. response length	8192	8192
Batching		
Prompts per step	128	128
Generations per prompt	1	1
Train global batch size	128	128
Train micro batch size	1	1
Generation batch size	64	128
Logprob batch size	1	–
Max rollout turns	1	1
Rollout / Generation		
Inference engine	vLLM	vLLM
Temperature	1.0	1.0
Top-p	1.0	1.0
Top-k	–	–
Max new tokens	8192	8192
vLLM tensor parallel size	4	1
Validation		
Validation batch size	64	–
Validation period	20	–
Distillation loss		
Top-K distillation	64	–
Zero outside top-K	False	–
Parallelism		
Number of GPUs	4	4
Policy tensor parallel size	4	–
Policy context parallel size	1	–
Teacher tensor parallel size	4	–
Teacher context parallel size	1	–
Sequence parallel	False	False
Activation checkpointing	True	True
Training		
Optimizer	AdamW	AdamW
Learning rate	5×10^{-6}	5×10^{-6}
Scheduler	Linear warmup + constant	Cosine (warmup ratio $= 0.1$)
Warmup steps	20	20
Weight decay	0.01	–
Gradient clip norm	1.0	1.0
Precision	bfloat16	bfloat16

Table 6: Hyperparameters used for SD-ZERO Self-Distillation phase and OPSD on Qwen3-4B-Instruct.

D ADDITIONAL RESULTS

D.1 PASS@8 PERFORMANCE ON MATH BENCHMARKS

Table 7 complements the main results in Table 1 by showing that the advantages of SRT and SD-ZERO persist under Pass@8 evaluation. The trend is consistent with the avg@8 results: across both Qwen3-4B-Instruct and Olmo-3-7B-Instruct, SRT already matches or outperforms strong baselines, and SD-ZERO further achieves the best overall average performance. These gains on Pass@8 suggest that our method is not merely sharpening the output distribution around a narrow set of solutions, but genuinely making the model's exploration more well-directed, so that correct reasoning paths are more likely to be discovered across multiple attempts. Overall, the Pass@8 results further reinforce the conclusion that explicit self-revision training improves reasoning quality, and that the second-stage SD-ZERO training strengthens this effect.

	AIME24	AIME25	HMMT25	AMOBench	OpenR1	MATH	**Average**
Qwen3-4B-Instruct	76.7	76.7	50.0	30.0	69.2	97.6	66.7
SFT	83.3	73.3	46.7	18.0	69.2	97.0	64.6
RFT	**86.7**	**80.0**	53.3	30.0	70.2	**98.4**	69.8
GRPO	80.0	70.0	46.7	28.0	70.0	98.0	65.4
OPSD	80.0	70.0	50.0	26.0	70.0	97.0	65.5
SRT (Ours)	**86.7**	**80.0**	60.0	28.0	70.2	97.6	70.4
SD-ZERO (Ours)	**86.7**	**80.0**	**63.3**	**36.0**	**72.0**	97.0	**72.5**
Olmo-3-7B-Instruct	80.0	73.3	46.7	16.0	60.0	**98.0**	62.3
SFT	**83.3**	66.7	46.7	18.0	55.0	96.0	60.9
RFT	80.0	73.3	60.0	22.0	62.0	97.0	65.7
GRPO	80.0	66.7	46.7	18.0	**70.0**	97.0	63.1
OPSD	80.0	73.3	56.7	20.0	58.0	97.0	64.2
SRT (Ours)	**83.3**	76.7	**66.7**	20.0	66.0	**98.0**	68.4
SD-ZERO (Ours)	**83.3**	**80.0**	**66.7**	**30.0**	68.0	97.0	**70.8**

Table 7: **Pass@8 performance comparison of SD-ZERO and SRT (Self-Revision Training, Phase 1) against baseline post-training methods** on math reasoning benchmarks. Across both Qwen3-4B-Instruct and Olmo-3-7B-Instruct, SRT and especially SD-ZERO achieve the strongest overall results.

D.2 OPSD UNDER FINAL-ANSWER-ONLY SUPERVISION

In Table 8, we show that OPSD performs much worse when given access only to ground-truth final answers rather than gold solution traces. Its performance remains close to the base model and falls far behind SD-ZERO across all math benchmarks. These results highlight that existing self-distillation methods do not naturally benefit from final-answer-only supervision, whereas SD-ZERO is specifically designed to convert such sparse outcome feedback into effective token-level learning signals.

D.3 DETAILED SELF-REVISION STATISTICS

Table 9 reports detailed statistics for the Generate-then-Revise evaluation on 1K AIME24 samples with Qwen3-4B-Instruct as the base model. For each method, we show the average token cost of the first attempt and the revision, together with the corresponding accuracies before and after revision. We also report the net wrong-to-correct rate, computed as $(\text{Revised Attempt Accuracy} - \text{First Attempt Accuracy})/(100 - \text{First Attempt Accuracy})$, which measures the fraction of initially incorrect answers that are corrected after revision. These statistics complement Figure 3 by quantifying both the accuracy gains and the token efficiency of revision across methods.

	AIME24	AIME25	HMMT25	AMOBench	OpenR1	MATH	**Avg.**
Base model	59.6	45.8	26.7	9.8	55.8	91.0	48.1
OPSD	63.3	47.9	32.9	9.0	57.0	91.1	50.2
OPSD (final answers only)	62.5	47.1	29.9	9.3	57.0	91.0	49.5
SD-ZERO (Ours)	**68.3**	**60.0**	**45.4**	**16.0**	**60.4**	**93.6**	**57.3**

Table 8: **OPSD with final-answer-only supervision** on Qwen3-4B-Instruct evaluated on math benchmarks. When given access only to final answers rather than gold solutions, OPSD performs only marginally better than the base model and remains far below SD-ZERO, showing that SD-ZERO is much more effective under outcome-only supervision.

Method	**First Attempt Tokens**	**Revised Attempt Tokens**	**First Attempt Accuracy (%)**	**Revised Attempt Accuracy (%)**	**Correction Rate (%)**
Base Model	3708	5098	59.6	60.7	2.7
GRPO	4432	5499	65.2	66.9	4.9
OPSD	3630	5099	63.3	64.7	3.8
SRT	8458	8137	66.7	71.7	**15.0**
SD-ZERO	3518	3314	68.3	73.6	**16.7**

Table 9: Comparison of self-revision effectiveness and token costs across different models on AIME24 with Qwen3-4B-Instruct as the base model. The Correction Rate is computed as $(\text{Revised Attempt Accuracy} - \text{First Attempt Accuracy})/(100 - \text{First Attempt Accuracy})$, i.e., the net fraction of initially incorrect answers corrected after revision. First-attempt accuracies come from the Generate-then-Revise run and can differ from the avg@8 results in Table 1.

D.4 GRPO CONFIGURATION COMPARISON

The GRPO baseline in Table 1 is run under 4 generations per question under a sampling budget matched with SD-ZERO. Here, we vary the number of generations per question and the number of training epochs in GRPO for Qwen3-4B-Instruct on the math domain.

As shown in Table 10, simply increasing the rollout budget in GRPO does not close the gap to our method. Moving from 4 to 8 generations per question (with 0.5 training epochs to match rollout budget) lowers average accuracy from 51.7 to 50.6. Extending training to a full epoch with 8 generations (2.4× the sampling budget of SD-ZERO) recovers to 52.3, still 5.0 points below SD-ZERO.

	AIME24	AIME25	HMMT25	AMOBench	OpenR1	MATH	**Avg.**
GRPO (4 generations, 1 epoch)	62.5	50.0	30.4	11.0	62.9	93.5	51.7
GRPO (8 generations, 0.5 epoch)	58.8	50.0	30.4	9.8	61.5	93.0	50.6
GRPO (8 generations, 1 epoch)	61.7	52.1	31.3	12.0	**63.4**	93.1	52.3
SD-Zero (Ours)	**68.3**	**60.0**	**45.4**	**16.0**	60.4	**93.6**	**57.3**

Table 10: **Comparison with GRPO under larger rollout budgets** on Qwen3-4B-Instruct evaluated on math benchmarks. We compare SD-Zero against GRPO variants that use more sampled generations per prompt and different effective training budgets. Even when GRPO is given substantially more exploration through 8 generations, its gains remain limited, whereas SD-Zero achieves the best average performance and outperforms all GRPO settings on most benchmarks. This suggests that the advantage of our method comes not simply from sampling more trajectories, but from turning binary outcome feedback into more informative self-revision-based supervision.

D.5 Controlled Baselines on the Same Revision Data

Table 1 matches methods by generation budget, so each baseline retains different training examples. To separate the SRT objective from data quality and question selection, we train three baselines on the 6K revision traces that SRT uses (Table 11). SFT on y_{revised} trains on (x, y_{revised}) with the same hyperparameters and number of steps as SRT. SFT on full traces trains on $(x, [y_{\text{init}}, P_r, y_{\text{revised}}])$. GRPO trains on the 6K questions behind these traces.

SFT on y_{revised} trails SRT by 2.2 points on Qwen3-4B-Instruct and 5.3 points on Olmo-3-7B-Instruct. On Olmo-3-7B-Instruct it also falls below RFT, so the revised responses are not better training data than the model's own correct responses. On Qwen3-4B-Instruct, SFT on full traces (54.7) and GRPO on the 6K questions (51.3) also trail SRT (55.7). The gain of SRT therefore comes from the revision-conditioned objective, not from the retained examples or the question subset. The objective also changes behavior: the SRT model's self-revision keyword ratio is 0.57%, compared with 0.31% for SFT on y_{revised} (keyword list in Section E.1).

	AIME24	AIME25	HMMT25	AMOBench	OpenR1	MATH	**Avg.**
Qwen3-4B-Instruct							
Base model	59.6	45.8	26.7	9.8	55.8	91.0	48.1
SFT (DeepSeek-R1 demonstrations)	61.7	46.3	31.3	7.3	56.1	91.3	49.0
RFT	64.2	52.1	37.1	11.3	59.3	91.9	52.6
GRPO (6K SRT questions)	62.1	50.0	30.4	11.0	**61.2**	93.3	51.3
SFT on y_{revised} (6K traces)	63.3	53.3	38.8	13.3	59.3	93.2	53.5
SFT on full traces (6K traces)	64.2	59.2	38.3	**16.0**	58.9	91.8	54.7
SRT (6K traces)	66.7	59.2	40.0	**16.0**	59.8	92.4	55.7
SD-Zero	**68.3**	**60.0**	**45.4**	**16.0**	60.4	**93.6**	**57.3**
Olmo-3-7B-Instruct							
Base model	56.7	42.1	25.0	1.3	48.9	91.1	44.2
SFT (DeepSeek-R1 demonstrations)	51.3	43.8	30.4	1.3	48.5	91.4	44.4
RFT	56.7	48.3	35.8	2.3	50.9	91.4	47.6
SFT on y_{revised} (6K traces)	51.3	45.4	27.5	3.0	50.3	90.9	44.7
SRT (6K traces)	59.2	52.9	39.6	3.5	52.4	92.3	50.0
SD-Zero	**61.7**	**53.8**	**40.4**	**5.5**	**55.3**	**94.0**	**51.8**

Table 11: **Controlled baselines on the SRT revision data.** SFT on y_{revised}, SFT on full traces, and SRT train on the same 6K revision traces and differ only in the objective; GRPO trains on the 6K questions behind these traces. SRT outperforms every variant trained on the same data, so its gain comes from the revision-conditioned objective, not from the retained examples.

D.6 Self-Distillation on Hard Questions

Existing self-distillation methods give weaker supervision on questions that the student rarely solves. SDPO needs a successful rollout, which such questions seldom produce. OPSD conditions its teacher on a gold solution, which can differ from the student's attempt and so does not point to the student's own mistakes. The SD-Zero teacher conditions on the student's attempt whether or not it is correct. To test this, we keep only the training questions that the student answers correctly in fewer than 4 of 8 samples and rerun the self-distillation phase of each method on this subset with Qwen3-4B-Instruct. SD-Zero outperforms OPSD by 2.8 points and SDPO by 3.7 points (Table 12).

D.7 Scaling SRT to Qwen3-14B

To test whether SRT transfers to a larger model, we apply it to Qwen3-14B with thinking mode disabled (Table 13). SRT raises the math average from 61.7 to 65.3, improves every benchmark, and outperforms SFT on DeepSeek-R1 demonstrations (63.9). This experiment covers the SRT phase only.

Method	Math Avg.
OPSD	49.6
SDPO	48.7
SD-ZERO	**52.4**

Table 12: **Self-distillation on hard questions only** (Qwen3-4B-Instruct). Each method trains only on questions that the student answers correctly in fewer than 4 of 8 samples. SD-ZERO outperforms OPSD by 2.8 points and SDPO by 3.7 points.

	AIME24	AIME25	HMMT25	AMOBench	OpenR1	MATH	**Avg.**
Qwen3-14B	79.2	72.5	49.6	18.3	55.4	95.5	61.7
SFT	81.7	75.0	51.3	21.5	57.6	96.3	63.9
SRT (Ours)	**82.9**	**76.5**	**53.4**	**23.1**	**59.0**	**96.8**	**65.3**

Table 13: **SRT on Qwen3-14B** (thinking mode disabled) evaluated on math benchmarks. SRT improves the base model by 3.6 points on average and improves every benchmark.

D.8 MULTIPLE ROUNDS OF TEACHER SYNCHRONIZATION

Section 3.4 synchronizes the teacher once. We extend this to four rounds of self-distillation on OpenR1-Math with Qwen3-4B-Instruct and synchronize the teacher with the latest student at the start of each round (Table 14). The first two rounds run 500 steps each, and the third and fourth run 125 steps each. Accuracy on training questions rises by 2.9, 2.2, 0.5, and 0.3 points over the four rounds, so gains saturate by the third round at this scale.

Stage	**Teacher synchronized at**	**Accuracy on training questions**	**Gain**
SRT model	–	52.4	–
Round 1 (steps 1–500)	–	55.3	+2.9
Round 2 (steps 501–1000)	step 500	57.5	+2.2
Round 3 (steps 1001–1125)	steps 500, 1000	58.0	+0.5
Round 4 (steps 1126–1250)	steps 500, 1000, 1125	**58.3**	+0.3

Table 14: **Four rounds of self-distillation with teacher synchronization** on OpenR1-Math with Qwen3-4B-Instruct. Each round synchronizes the teacher with the latest student. Gains shrink after the second round and saturate by the third.

E Analysis of SD-Zero

E.1 Self-Revision Keyword Analysis

Below, we list the self-revision keywords used in our behavioral analysis. These keywords are used in the study described in Figure 6 to quantify the frequency of explicit self-correction language during training, complementing the analysis of response length and downstream performance. We construct the list by collecting common phrases that models frequently use when revising an earlier reasoning step, such as signaling doubt, detecting an error, restarting, rechecking, or correcting a previous claim. Concretely, the list includes expressions associated with hesitation (e.g., "wait," "hold on"), error acknowledgment (e.g., "my mistake," "this is wrong"), and explicit restart or verification behavior (e.g., "let me recheck," "let me try again," "let me start over"). While this keyword set is not intended to exhaustively capture all forms of self-correction, it provides a simple and interpretable proxy for tracking overt self-revision behavior across training stages.

List of Self-Revision Keywords Used in Behavioral Analysis (Section 4.2)

```
REVISION_KEYWORDS = [
    "wait",
    "hold on",
    "actually",
    "on second thought",
    "let me recheck",
    "let's recheck",
    "let me check",
    "let's check",
    "let me recalculate",
    "let's recalculate",
    "let me correct",
    "my mistake",
    "i made a mistake",
    "this is wrong",
    "that is wrong",
    "that's wrong",
    "incorrect",
    "re-evaluate",
    "let's re-evaluate",
    "let me rethink",
    "let's rethink",
    "let me double check",
    "let's double check",
    "wait a minute",
    "let me start over",
    "let me try again",
    "oops"
]
```

E.2 Ablation Studies on SD-Zero Method Design

E.2.1 Ablating Loss terms in SRT

The SRT objective $\mathcal{L}_{\text{SRT}}$ (Section 2) combines two terms: the first term $\mathcal{L}_{\text{revision}}$ trains the model to produce a revision or rephrase conditioned on the outcome; the second term $\mathcal{L}_{\text{generation}}$ trains it to generate better responses by actively evaluating its current response and self-correcting. We ablate this by training with each term alone. If the self-correction term is redundant, removing it should not hurt Phase 2 performance; if the conditioned revision term is redundant, the model should still learn to revise from the self-correction objective alone.

As shown in Table 15, both terms are necessary, and neither alone recovers the full effect of SRT. Using only $\mathcal{L}_{\text{generation}}$ preserves relatively strong first-attempt generation, but its revised-attempt improvement is much smaller, with the correction rate dropping from 15.0% to 7.2%. This suggests that self-correction behavior does not emerge reliably from the generation objective alone. In contrast, using only $\mathcal{L}_{\text{revision}}$ yields a higher correction rate of 12.1%, indicating that the model does learn to revise when explicitly trained to do so, but this comes at a clear cost to generation quality: both average generation accuracy and first-attempt accuracy fall substantially relative to the full SRT objective. Taken together, these results show that the two terms play complementary roles. $\mathcal{L}_{\text{revision}}$ is important for directly eliciting revision behavior, while $\mathcal{L}_{\text{generation}}$ helps transfer that behavior back into stronger standalone generation. Their combination is what enables SRT to simultaneously improve first-attempt performance and unlock effective self-revision. This complementarity is consistent with previous work (Kumar et al., 2024), which similarly finds that self-correction does not emerge from a correction-focused signal alone, but requires preserving the model's underlying generation ability while separately training revision behavior.

The two terms overlap. Because y' in Equation (2) contains y_{revised}, $\mathcal{L}_{\text{generation}}$ already supervises the revision tokens given $(x, y_{\text{init}}, P_r)$, so adding $\mathcal{L}_{\text{revision}}$ weights these tokens twice and the tokens of y_{init} once. $\mathcal{L}_{\text{revision}}$ still matters because it matches the input format of the Phase-2 reviser. Without it, the reviser often inserts meta-phrases such as "let me revise the previous output" instead of writing a fresh solution, consistent with the weaker correction rate of $\mathcal{L}_{\text{generation}}$ alone in Table 15.

Method	Average Generation Accuracy (%)	Generate-then-Revise on AIME24		
		First Attempt Accuracy (%)	Revised Attempt Accuracy (%)	Correction Rate (%)
Base Model	49.8	59.6	60.7	2.7
SRT	57.6	66.7	71.7	15.0
SRT ($\mathcal{L}_{\text{generation}}$ Only)	56.4	65.4	67.9	7.2
SRT ($\mathcal{L}_{\text{revision}}$ Only)	52.2	62.1	66.7	12.1

Table 15: **Ablation of the two loss terms in the SRT objective.** We train Qwen3-4B-Instruct with SRT using only $\mathcal{L}_{\text{generation}}$ or only $\mathcal{L}_{\text{revision}}$, and evaluate both overall generation quality and self-revision ability. Both terms are important: using either term alone underperforms the full SRT objective in average generation accuracy and in generate-then-revise performance. In particular, $\mathcal{L}_{\text{generation}}$ alone yields relatively weak correction after revision, while $\mathcal{L}_{\text{revision}}$ alone improves correction rate but leads to substantially worse first-attempt and overall generation accuracy, showing that the two terms play complementary roles in strengthening both generation and self-revision.

E.2.2 Applying Self-Distillation Phase Directly to the Base Model

Table 16 provides an ablation study showing what happens when we remove the SRT phase and apply the Self-Distillation phase directly to the base model. Although this Phase-2-only variant slightly improves average generation accuracy from 49.8 to 51.4, its Generate-then-Revise (see Section 3.2) behavior remains weak: first-attempt accuracy increases only modestly from 59.6 to 61.2, revised-attempt accuracy rises from 60.7 to 62.2, and the correction rate is only 2.6%, which is nearly identical to the base model's 2.7%. In contrast, models that include the SRT phase achieve much stronger revision gains, with correction rates of 15.0% for SRT and 16.7% for SD-Zero. These results indicate that directly applying preference optimization to a model without prior self-revision training does not meaningfully improve its ability to revise incorrect solutions. Instead, the SRT

phase appears to be a necessary prerequisite that first elicits self-revision behavior, after which the Self-Distillation phase can effectively refine and strengthen it.

Method	Average Generation Accuracy (%)	Generate-then-Revise on AIME24		
		First Attempt Accuracy (%)	Revised Attempt Accuracy (%)	Correction Rate (%)
Base Model	49.8	59.6	60.7	2.7
SRT	57.6	66.7	71.7	15.0
SD-ZERO	60.3	68.3	73.6	16.7
SD-ZERO (Phase 2 Only)	51.4	61.2	62.2	2.6

Table 16: **Ablation of applying SD-ZERO without the SRT phase.** We train a Phase-2-only variant on Qwen3-4B-Instruct by running SD-ZERO directly on the base model, without first inducing self-revision through SRT. This variant shows only marginal improvement over the base model in both overall generation accuracy and generate-then-revise performance, and its correction rate remains nearly unchanged. In contrast, the full two-phase method substantially improves both first-attempt generation and revision effectiveness. These results suggest that the SRT phase is necessary to first elicit self-revision capability, which SD-ZERO can then refine and distill into stronger reasoning.

E.2.3 ABLATING DATA SPLIT BETWEEN PHASES

Our method uses a fixed overall data budget that must be divided between the SRT phase and the Self-Distillation phase. This creates a natural trade-off: allocating more data to SRT may produce a stronger reviser, but leaves fewer examples for Phase 2 distillation; allocating too little data to SRT may instead yield a reviser that is not strong enough to provide useful supervision. To study this trade-off, we vary the split between the two phases while keeping the total amount of training data fixed, and report both the intermediate SRT model accuracy and the final SD-ZERO model accuracy in Table 17.

Data Split		Average Performance (%)	
SRT	**Self-Distillation**	**SRT Model**	**SD-ZERO Model**
6K	9K	57.6	60.3
9K	6K	57.8	59.1
7.5K	7.5K	57.8	59.8

Table 17: **Ablation of data split between SRT and Self-Distillation phases.** On Qwen3-4B-Instruct, we vary how the training data is allocated across the two phases while keeping the total data budget fixed. Allocating more data to SRT slightly improves SRT model performance, but does not lead to the best final SD-ZERO accuracy. The best overall result is achieved by assigning more data to Self-Distillation, suggesting that once SRT has sufficiently elicited self-revision capability, additional data is more effectively used in Self-Distillation phase to distill and refine this behavior into stronger generation.

E.2.4 PHASE-2 OBJECTIVE AND TEACHER CONTEXT

We vary the Phase-2 objective while keeping the SRT initialization, the teacher, and the training questions fixed (Table 18). Forward KL samples $y = (y_1, \ldots, y_T)$ from the student and minimizes $\sum_{t=1}^{T} w_t \, D_{\mathrm{KL}}(\pi_T(\cdot \mid y_{<t}) \,\|\, \pi_S(\cdot \mid y_{<t}))$, where π_T and π_S condition on the same inputs as in Equation (4) and the importance weight $w_t = \pi_T(y_t \mid y_{<t})/\pi_S(y_t \mid y_{<t})$, clipped to $[0.1, 1.0]$, corrects for sampling from the student. SFT on teacher-generated revisions samples revisions from the teacher and trains the student on them with next-token prediction. Reverse KL, our default, outperforms forward KL by 1.1 points and SFT on teacher-generated revisions by 0.5 points, consistent with prior work on on-policy distillation (Gu et al., 2024; Agarwal et al., 2024).

We also test an off-policy teacher. We sample an initial response to each training question from the SRT model, sample revisions of it, keep the correct revisions, and condition the teacher on this fixed

revision during Phase 2, as OPSD conditions on a gold solution. This variant improves over SRT (56.5 vs. 55.7) but trails SD-ZERO (57.3). A teacher conditioned on the student's current attempt corrects the student's own tokens, whereas a fixed revision pulls the student toward one correct answer.

Phase-2 variant	AIME24	AIME25	HMMT25	AMOBench	OpenR1	MATH	**Avg.**
None (SRT model)	66.7	59.2	40.0	16.0	59.8	92.4	55.7
Forward KL, importance-weighted	–	–	–	–	–	–	56.2
Teacher on a fixed self-revision	67.5	59.6	42.5	**16.3**	60.1	93.0	56.5
SFT on teacher revisions	–	–	–	–	–	–	56.8
Reverse KL, on-policy (SD-ZERO)	**68.3**	**60.0**	**45.4**	16.0	**60.4**	**93.6**	**57.3**

Table 18: **Phase-2 objective and teacher context** on Qwen3-4B-Instruct (math benchmarks). All variants start from the SRT model and use the same teacher and training questions. Reverse KL on the student's own attempts gives the best average; a teacher conditioned on a fixed self-revision (off-policy) improves over SRT but trails SD-ZERO.

E.3 Mechanisms of the Self-Distillation Phase

E.3.1 The SD-Zero Teacher Stays Close to the Student

The OPSD teacher conditions on a gold solution that the student did not write, so its outputs can lie far from the student's distribution. For each method, we sample a response y_{teacher} from its teacher, e.g., $\pi_{\text{teacher}}(\cdot \mid x, y_{\text{student}}, P_r)$ for SD-Zero, and compute the student's per-token negative log-likelihood $-\log \pi_{\text{student}}(y_{\text{teacher}} \mid x)$ (Table 19). The student assigns SD-Zero's teacher responses less than half the NLL of OPSD's (0.115 vs. 0.270 on average), so SD-Zero supervises the student with targets that stay close to its own distribution.

Teacher	Mean	Median	Std.
OPSD	0.2704	0.2654	0.0967
SD-Zero	**0.1150**	**0.1158**	**0.0271**

Table 19: **Per-token negative log-likelihood of teacher responses under the student** (lower is closer to the student's own distribution). The student assigns SD-Zero's teacher responses less than half the NLL of OPSD's.

E.3.2 The Reviser Signals Where to Stop

The self-distillation phase halves response length (Figure 3), even though the teacher only scores the student's tokens and never generates its own sequence. To see how, we study 10 OpenR1-Math questions where the SRT teacher's own revision is at least $2\times$ shorter than the student's response (1.1K vs. 2.8K tokens on average). At the position t^* where the teacher's revision ends, the teacher assigns probability 0.42 to the end-of-sequence token on average, while the student assigns 0.03. The per-token KL on positions after t^* is $3.1\times$ higher than on positions before it, so the loss tells the student to stop where the reviser would stop.

F OPSD WITH THINKING MODELS

In Section 5 we noted that extending self-distillation to *thinking models*—where the student generates long, exploratory chains of thought before reaching an answer—is a natural but non-trivial next step. Below, we provide concrete evidence for this claim by applying OPSD to Qwen3-4B with thinking enabled during the student's on-policy rollouts.

Setup. We use the `Ashkchamp/Openthoughts_math_filtered_30K` dataset, taking the `solution` column as the privileged information supplied to the teacher. Training runs for one epoch with learning rate 5×10^{-6} and batch size 64. The sole experimental variable is the `enable_thinking` flag, which controls whether the student generates with Qwen3's native thinking mode turned on or off *during OPSD training*. At evaluation, thinking is always enabled for all methods; the model may generate up to 38K completion tokens, and we sample with temperature 0.6 and top-$p = 0.95$, reporting avg@16.

Results. Table 20 shows that enabling thinking during training degrades the base model on every benchmark we tested: -9.8 points on AIME24, -10.8 on AIME25, and -8.3 on HMMT25. Disabling thinking during training largely preserves (and on HMMT25 slightly improves) the base model's accuracy.

Method	**AIME24**	**AIME25**	**HMMT25**
Qwen3-4B	0.735	0.647	0.458
+ OPSD (`enable_thinking=True`)	0.637	0.539	0.375
+ OPSD (`enable_thinking=False`)	0.733	0.616	0.466

Table 20: Effect of the student's thinking mode on OPSD training (Qwen3-4B, avg@16). `enable_thinking` refers to whether thinking is active during training; all models are evaluated with thinking enabled. Red entries denote degradation relative to the base model.

G SELF-REVISION DESIGN CHOICES

We present complementary experiments on Countdown, a constrained arithmetic reasoning task, using Qwen2.5-7B. Although the setup differs from the main experiments (different base model, task domain, and pipeline configuration), three findings align with and corroborate the design principles of SD-ZERO: (1) on-policy self-revision data outperforms off-policy teacher data, (2) correctness filtering of revision traces is essential, and (3) self-revision training provides a stronger foundation for subsequent training than standard SFT does.

G.1 EFFECT OF TRAINING DATA SOURCE

We compare several training data sources for supervised fine-tuning on Countdown. For each of the 8K original SFT questions, we sample N solutions per question and optionally filter for correctness. Table 21 reports the results.

- **LLaMA-70B**: $N = 16$, pass@16 $= 81\%$, filtered $\rightarrow$ 6.5K pairs.
- **Qwen2.5-7B** (self-generated): $N = 16$, pass@16 $= 78.7\%$, filtered $\rightarrow$ 6.2K pairs.
- **GPT-4o**: $N = 5$, pass@5 $= 67.5\%$, filtered $\rightarrow$ 5.4K pairs.
- **SFT on GRPO-Qwen2.5-7B data** (on-policy): GRPO reaches pass@1 $= 91\%$ after RL; we distill its generations into 6.5–8K SFT pairs without filtering.
- **Self-Revision**: Qwen2.5-7B critiques and refines the previous response attempt; unfiltered (6.5K) and filtered (5.9K) variants.

Training Source	pass@1	pass@2	pass@4
No training	0.410	0.538	0.639
LLaMA-70B (6.5K, filtered)	0.605	0.725	0.814
Qwen2.5-7B (6.2K, filtered)	0.552	0.676	0.768
GPT-4o (5.4K, filtered)	0.617	0.711	0.775
SFT on GRPO data (8K, unfiltered)	0.878	0.904	0.922
SFT on GRPO data (6.5K, unfiltered)	0.883	0.910	0.919
Self-Revision (6.5K, unfiltered)	0.529	0.639	0.717
Self-Revision (5.9K, filtered)	0.630	0.737	0.810

Table 21: Countdown pass@k after fine-tuning Qwen2.5-7B on different data sources. On-policy sources (SFT on GRPO data, self-revision) consistently outperform off-policy ones (LLaMA-70B, GPT-4o). Among methods that do not require a prior RL stage, filtered self-revision achieves the highest pass@1.

Findings. Three patterns emerge. First, **on-policy data dominates off-policy data**: SFT on data generated by the GRPO-trained checkpoint ($\sim$0.88) substantially outperforms SFT on LLaMA-70B (0.605) or GPT-4o (0.617) data, despite comparable data sizes and correctness filtering. Note that GRPO itself reaches $\sim$0.91 via RL alone; the $\sim$0.88 figure reflects SFT distillation of its outputs. Distribution match matters more than teacher quality—the on-policy 6.5K subset performs comparably to the 8K variant. SD-ZERO also uses the student's own reviser as the teacher.

Second, **self-revision outperforms first-attempt sampling**: filtered self-revision (pass@1=0.630) exceeds the model's own correctness-filtered first attempts (pass@1=0.552). Both are on-policy, so the gain isolates the value of the revision step itself.

Third, **correctness filtering is essential**: unfiltered self-revision (0.529) underperforms fine-tuning on the model's own filtered first attempts (0.552), while filtering raises it to 0.630—a 19% relative gain. Noisy revision traces actively hurt. SD-ZERO's Phase 1 also retains only successful revisions (Section 2.1).

G.2 SELF-REVISION AS INITIALIZATION FOR RL

We evaluate whether self-revised targets (A′) improve OOD generalization, especially when combined with RL. We fine-tune Qwen2.5-7B on 8K Countdown questions with either LLaMA-70B–generated answers (A) or self-revised answers (A′), optionally followed by GRPO. Neither A nor A′ are filtered for correctness. Note that A uses off-policy teacher answers; Table 21 shows this is a conservative baseline, since off-policy data underperforms on-policy data even in-distribution. All evaluations use a 4096-token generation budget.

	AIME24		**AIME25**		**Countdown**		**IFBench**		**IFEval**		**MMLU**		**MMLU-Pro**	
Method	@1	@128	@1	@128	@1	@128	@1	@128	@1	@128	@1	@128	@1	@128
Base	.07	.27	.07	.43	.41	.90	.27	.58	.71	.90	.74	.91	.56	.82
GRPO (Countdown)	.08	.30	.08	.40	.91	.98	.27	.61	.72	.91	.74	.92	.56	.84
Countdown (A)	.05	.27	.05	.43	.39	.91	.23	.54	.52	.85	.69	.96	.46	.91
Countdown (A) + GRPO	.06	.23	.07	.40	.92	.98	.24	.55	.53	.84	.71	.97	.50	.93
Countdown (A′)	.07	.30	.06	.47	.40	.88	.19	.62	.60	.91	.71	.97	.50	.94
Countdown (A′) + GRPO	.07	**.40**	.07	**.47**	.91	.98	.20	.58	.62	.91	.72	.97	.52	.93

Table 22: OOD evaluation (pass@1 / pass@128) after training on Countdown with LLaMA-70B answers (A) or self-revised answers (A′), optionally followed by GRPO. A′+GRPO achieves the highest OOD scores.

The key result is that **self-revised targets provide a better initialization for RL on OOD benchmarks.** Countdown (A′)+GRPO yields pass@128 of 0.40 on AIME24 and 0.47 on AIME25, compared to 0.23 and 0.40 for Countdown (A)+GRPO, and 0.30 and 0.40 for standalone GRPO. In-distribution Countdown performance is nearly identical across all GRPO variants (0.91–0.92 pass@1, 0.98 pass@128).

Notably, this comparison is conservative: neither A nor A' were filtered for correctness, and A uses off-policy answers rather than the model's own attempts. Despite this, self-revised targets still yield a meaningfully stronger RL initialization on OOD benchmarks (Table 22). This suggests that the revision process itself — even without correctness filtering — carries useful signal, and that the gains would likely be larger with filtering (Table 21 shows a 19% relative gain from filtering in-distribution). Moreover, these experiments use GRPO rather than SD-ZERO's distillation objective as the second stage, suggesting the value of self-revision as initialization is not specific to a particular second-stage method.

More broadly, the fact that unfiltered self-revision provides useful signal raises the question of whether self-revision training can extend to settings without verifiable rewards, where correctness filtering is unavailable. We leave this to future work.