CROSS-BATCH SELF-DISTILLATION FOR REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

In language-model post-training with verifiable rewards, successful responses provide evidence for subsequent learning. GRPO-style response weighting, however, is restricted to current-batch gradients and may miss useful directions revealed by earlier successes. We introduce Cross-Batch Self-Distillation (CBSD), which separates rollout generation from evidence retention. An acting policy generates responses, while a non-acting carrier learns from verified successes across batches. The actor reads the carrier on its own sampled prefixes through on-policy distillation, obtaining token-wise guidance without replaying source responses or supplying them as context. A control-variate write and pre-write reading support stable training. Our analysis characterizes the ideal success-conditioned carrier and interprets its log-ratio read as token-level credit. In a binary-verifier family at a fixed actor, sufficient source sampling and reuse yield larger expected first-order gains than fresh-batch response weighting under matched total response and local KL budgets. With Qwen3-8B on SciKnowEval, CBSD achieves 76.22 average maximum Avg@16, surpassing reported SDPO/GRPO by $3.29/5.17$ percentage points and leading in three of four domains. On 27 Physics prompts rarely solved initially, it exceeds GRPO by 29.40 points at step 200.

1 INTRODUCTION

Reinforcement learning and on-policy self-distillation improve the reasoning performance of large language models (LLMs) through post-training (Wen et al., 2026; Zhao et al., 2026). In reinforcement learning with verifiable rewards (RLVR), successful responses discovered through exploration provide evidence for subsequent learning (Trung et al., 2024; Guo et al., 2025). Under binary feedback, Group Relative Policy Optimization (GRPO) converts current-group outcomes into response-level credit (Shao et al., 2024). Such weighting combines only current-batch response gradients, so useful directions revealed by earlier successes may be unavailable on a new batch. Self-Distillation Policy Optimization (SDPO) uses the current group's successful responses as teacher context for token-wise guidance through on-policy distillation (OPD) (Hübotter et al., 2026).

The value of experience reuse depends on how past successes support subsequent learning (Oh et al., 2018). Existing approaches reuse past experience by replaying stored verified responses (Zhang et al., 2025a;b; Wang et al., 2025) or by distilling accumulated knowledge through a context-conditioned teacher (Snell et al., 2022; Ye et al., 2026b;a). We introduce a new approach that reuses verified successes across batches, turning past discoveries into guidance on fresh actor trajectories.

Our central insight is that past experience can define a teaching distribution that remains available across batches. Successful-response self-training can learn this distribution by fitting a policy to the actor's verified samples (Zelikman et al., 2022; Singh et al., 2024; Russo, 2026). Once represented as a policy, this experience can be queried at the prefixes of new responses, including unsuccessful ones (Lin et al., 2020). This separates the responses supplying evidence from those receiving supervision, allowing experience to accumulate across batches while the actor learns on its own current samples.

We propose the *Cross-Batch Self-Distillation* (CBSD), an asymmetric two-policy learner (as shown in Figure 1). The *acting policy* generates every rollout. The verifier outcomes supervise a non-acting *carrier policy* through supervised fine-tuning (SFT) on the actor's successful responses. These updates *write* experience into its parameters across batches. On later batches, the actor receives its task supervision by *reading* the carrier through OPD on its own sampled prefixes and actions (Gu et al.,

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

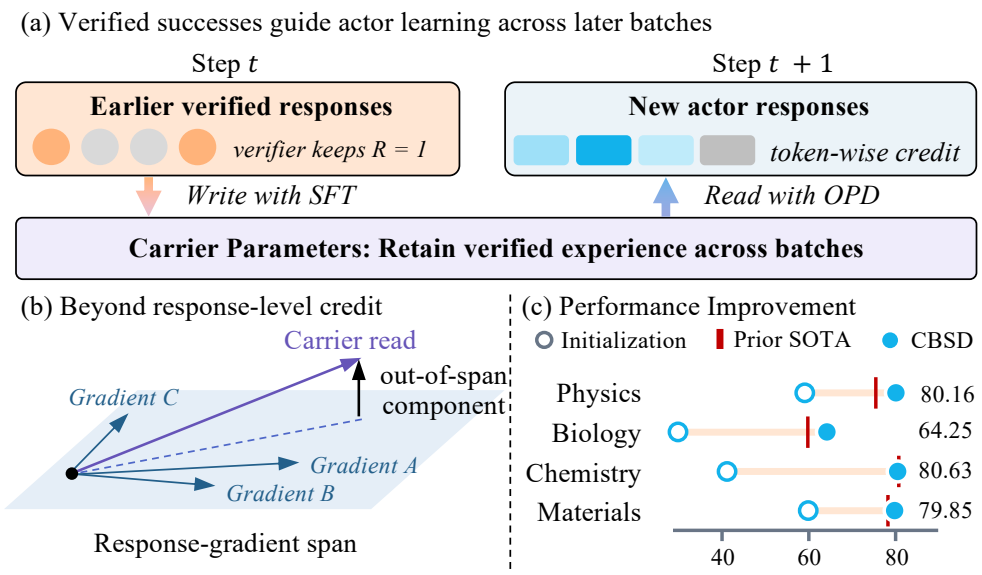

Figure 1: **CBSD reuses verified successes across batches.** (a) A non-acting carrier learns from the actor's verified successes and provides token-wise credit on later actor samples. (b) The read direction can have a component outside the current-batch response-gradient span (schematic). (c) Maximum Avg@16 (%) across four SciKnowEval domains with Qwen3-8B, compared with initialization and the better reported GRPO/SDPO result per domain (Hübotter et al., 2026).

2024; Zhao et al., 2026; Shenfeld et al., 2026). Relative to GRPO, the read provides token-specific guidance; relative to SDPO's use of successful responses as context, it makes retained experience available without supplying the source responses again. Earlier successes can thus guide a new group containing only failures, without replay or additional rollouts.

To stabilize training, we introduce two mechanisms. Building on control-variate methods (Greensmith et al., 2001; Kool et al., 2019; Phan et al., 2023), we use the Write Control Variate (WCV) to reduce carrier write variance while preserving the expected SFT gradient. Pre-write reading uses the carrier's pre-update predictions as the actor's distillation targets. This blocks immediate feedback from the current write, which can amplify actor updates on the same batch and destabilize training.

Our theoretical analysis connects successful-response learning to token-level credit and shows how token-wise reading can supply directions beyond current-batch response weighting. In an idealized setting, sufficient reuse yields greater expected first-order gains than optimal response-level weighting under matched total response and local KL budgets.

We evaluate CBSD with Qwen3-8B on SciKnowEval (Yang et al., 2025; Feng et al., 2024), using Avg@16, the mean accuracy over 16 responses per prompt. CBSD achieves an average maximum Avg@16 of 76.22, surpassing reported SDPO/GRPO by $3.29/5.17$ percentage points. Across four domains, it gains 19.94–39.44 percentage points over initialization and outperforms both reported baselines in three. The carrier supplies credit that varies substantially across tokens, and changes in the actor's mean token log-probabilities closely track those of the carrier. On 27 Physics prompts rarely solved at initialization, CBSD exceeds GRPO by 29.40 pp at step 200.

We make the following contributions.

- We introduce CBSD, which retains verified successes in a non-acting carrier and reuses them as token-specific credit on later actor samples, without replay or additional rollouts.
- We develop WCV to reduce carrier write variance while preserving the expected SFT gradient, and use pre-write reading to prevent immediate feedback from the current write.
- We provide a theoretical account of retained policy credit, connecting successful-response learning to token-level credit, characterizing the read's additional expressivity, and establishing a reuse advantage under matched response budgets.

2 PROBLEM FORMULATION

We consider reinforcement learning with binary verifiable rewards, aiming to increase the probability of generating verifier-accepted responses. At training step t, the acting policy π_t samples a batch of B responses, with G responses per prompt. We index responses by b, with $x_{t,b}$ denoting the corresponding prompt, $\tau_{t,b} \sim \pi_t(\cdot \mid x_{t,b})$, and verifier outcome $R_{t,b} = R(x_{t,b}, \tau_{t,b}) \in \{0, 1\}$. The resulting batch is $\mathcal{B}_t = \{(x_{t,b}, \tau_{t,b}, R_{t,b})\}_{b=1}^{B}$.

GRPO-style RLVR methods assign each response a group-relative advantage shared across its tokens (Shao et al., 2024). We compare task-gradient directions at the sampling policy, before regularization and optimizer transformations. For actor parameters θ_t, we define the response log-probability gradient (Williams, 1992; Sutton et al., 1999) as

$$v_{t,b} = \nabla_{\theta_t} \log \pi_t(\tau_{t,b} \mid x_{t,b}) = \sum_{\ell=1}^{L_{t,b}} \nabla_{\theta_t} \log \pi_t(a_{t,b\ell} \mid h_{t,b\ell}), \tag{1}$$

where $L_{t,b}$ is the number of valid response tokens, $a_{t,b\ell}$ is token ℓ, and $h_{t,b\ell} = (x_{t,b}, a_{t,b,<\ell})$ is the prompt followed by the preceding tokens.

We allow arbitrary response-level coefficients $\widehat{A}_{t,b}$, held fixed during actor differentiation. Response- or batch-level normalization factors can be absorbed into these coefficients. At a fixed actor and realized batch, the resulting gradients lie in the span of the current response gradients:

$$\sum_{b=1}^{B} \widehat{A}_{t,b} v_{t,b} \in \mathcal{S}_{\mathcal{B}_t} := \operatorname{span}\{v_{t,b} : 1 \le b \le B\}. \tag{2}$$

Our CBSD uses a carrier to retain information from earlier verified successes and provide token-wise guidance on new actor samples, enabling update directions beyond $\mathcal{S}_{\mathcal{B}_t}$.

3 METHOD

CBSD introduces an independent *carrier policy* κ_t alongside the acting policy π_t. The two policies are initialized to the same distribution and updated separately. The actor generates every training response, and the carrier is trained to maximize the likelihood of those accepted by the verifier. The actor receives its task supervision solely through on-policy distillation from the carrier. This retains verified evidence for later actor updates after the source responses are discarded.

Carrier write. To retain verified evidence for learning across batches, CBSD trains the carrier by maximizing the likelihood of the actor's successful responses. With carrier parameters ϕ_t and $N_t = \sum_{b=1}^{B} L_{t,b}$ valid response tokens in $\mathcal{B}_t$, the token-normalized write gradient is

$$\widehat{W}_{+} = \frac{1}{N_t} \sum_{b=1}^{B} R_{t,b} \sum_{\ell=1}^{L_{t,b}} \nabla_{\phi_t} \log \kappa_t(a_{t,b\ell} \mid h_{t,b\ell}). \tag{3}$$

To reduce variance without changing the expected SFT write, CBSD introduces the Write Control Variate (WCV), using all responses in each prompt group to construct a zero-mean correction.

For response b in a prompt group of size $G \ge 2$, the success rate of the other $G - 1$ responses supplies a baseline $\bar{R}_{t,-b}$ (Ahmadian et al., 2024). To account for differences between the actor and carrier, we weight the carrier log-probability gradient by the unclipped ratio $\rho_{t,b\ell} = \kappa_t(a_{t,b\ell} \mid h_{t,b\ell})/\pi_t(a_{t,b\ell} \mid h_{t,b\ell})$ to obtain zero conditional mean at each prefix. The baseline and token normalizer $D_{t,-b}$ exclude response b to preserve zero mean under independent sampling; $D_{t,-b}$ uses all other response lengths in the batch. Full definitions appear in Section B.3. The corrected write is

$$\widehat{W}_{\mathrm{WCV}}(\lambda_{\mathrm{cv}}) = \widehat{W}_{+} - \lambda_{\mathrm{cv}} \sum_{b=1}^{B} \frac{\bar{R}_{t,-b}}{D_{t,-b}} \sum_{\ell=1}^{L_{t,b}} \rho_{t,b\ell} \nabla_{\phi_t} \log \kappa_t(a_{t,b\ell} \mid h_{t,b\ell}). \tag{4}$$

All coefficients are stop-gradient constants, with $\lambda_{\mathrm{cv}} \in [0, 1]$; setting $\lambda_{\mathrm{cv}} = 0$ recovers the SFT write. Proposition 5 establishes expectation preservation and gives sufficient conditions for variance reduction. We update the carrier to κ_{t+1} using $\widehat{W}_{\mathrm{WCV}}$.

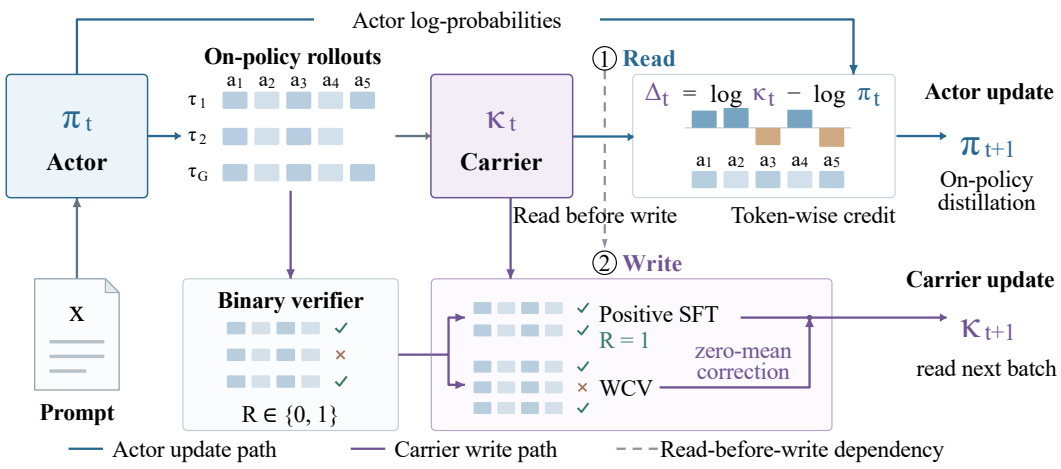

Figure 2: **CBSD's temporal computation graph.** At step t, verifier feedback on actor samples guides the carrier write from κ_t to κ_{t+1}. For on-policy distillation, the actor receives token-wise advantages from cached pre-write log-probabilities of κ_t; κ_{t+1} is first read on the next batch.

Carrier read. CBSD converts retained evidence into token-level guidance for the actor. On the actor's sampled prefixes and actions (h, a), CBSD defines the carrier-derived token advantage as

$$\Delta_t(h, a) = \log \kappa_t(a \mid h) - \log \pi_t(a \mid h). \tag{5}$$

The log-ratio assigns positive credit when the carrier gives a sampled token higher probability than the actor, and negative credit otherwise. We apply the read to all sampled responses, including those rejected by the verifier. Weighting the actor's token gradients by this advantage gives the sampled-token distillation update direction

$$\widehat{G}_t^{\text{read}} = \frac{1}{N_t} \sum_{b=1}^{B} \sum_{\ell=1}^{L_{t,b}} \text{sg}[\Delta_t(h_{t,b\ell}, a_{t,b\ell})] \nabla_{\theta_t} \log \pi_t(a_{t,b\ell} \mid h_{t,b\ell}), \tag{6}$$

where sg holds the advantage fixed during actor differentiation. Unlike the response-level coefficient $\widehat{A}_{t,b}$ in Equation (2), $\Delta_t(h, a)$ can vary across tokens within a response. The read can therefore produce directions outside the current response-gradient span $\mathcal{S}_{\mathcal{B}_t}$, as demonstrated in Theorem 2. The actor update combines this read direction with regularization toward the frozen base policy.

Pre-write reading. At step t, we cache log-probabilities from κ_t before the current batch's carrier write and use them to compute the actor's token advantages. The updated κ_{t+1} is first read on the next batch. This schedule blocks immediate feedback from the current write, which can amplify actor updates on the same samples. Reusing the cached log-probabilities avoids an additional carrier forward pass. Algorithm 1 summarizes the complete training step.

4 Theoretical Analysis

To understand how verified successes guide subsequent actor updates, we analyze the ideal token credit induced by successful-response fitting and the local benefits of reusing information from earlier successes across batches. We measure policy performance by the expected success probability over a prompt distribution $\mathcal{D}$, $J(\pi) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{\tau\sim\pi(\cdot|x)}[R(x, \tau)]$. Proofs and carrier write guarantees appear in Section B.

4.1 From successful responses to token-level credit

Our method fits the carrier to successful responses and reads its log-ratio with the actor as token-wise guidance. To connect these operations, we consider an ideal carrier fitted to a fixed actor's success-conditioned response distribution. Its log-ratio measures the relative change in success probability after choosing a token, giving the read a credit interpretation formalized below.

For a fixed actor π and prompts with positive success probability, define the success-conditioned response distribution $\pi^+(\tau \mid x) = \pi(\tau \mid x, R = 1)$ (Štrupl et al., 2022; Russo, 2026). Let $V^\pi(h) = \Pr_\pi(R = 1 \mid h)$ and $Q^\pi(h, a) = \Pr_\pi(R = 1 \mid h, a)$ denote the success probabilities before and after choosing token a at prefix h, continuing under π.

Proposition 1 (Ideal write target and token-level credit)**.** Let π have full support and a finite horizon. The expected negative log-likelihood of verifier-accepted responses sampled from π is minimized over response distributions by π^+. Wherever $Q^\pi(h, a) > 0$, we have

$$\log \frac{\pi^+(a \mid h)}{\pi(a \mid h)} = \log Q^\pi(h, a) - \log V^\pi(h). \tag{7}$$

By Equation (7), the ideal read assigns positive credit to tokens with above-average success probability at their prefix and negative credit to those with below-average success probability. The proof of Proposition 1 appears in Section B.1.

4.2 Cross-batch gains beyond response-level weighting

Our method retains information from earlier successes in a persistent carrier and reads it on newly sampled responses without replaying the source responses. To understand the benefit of this reuse, we construct a setting in which a carrier trained on finitely many successes supplies directions unavailable to current-batch response weighting. With sufficient source sampling and reuse, these directions yield greater expected first-order improvement at a fixed actor under equal response and local KL budgets.

Construction. We construct a task in which successful responses share the decisions required for success but vary in irrelevant decisions. A response $Y \in \{0, 1\}^n$ succeeds if and only if its first m bits are one, $1 \le m < n$. The actor makes independent Bernoulli decisions with one logit per decision, all fixed at zero; tokens are weighted uniformly. Averaging log-probability gradients over successful responses estimates their shared success direction $u = (\mathbf{1}_m, \mathbf{0}_{n-m})$.

Comparison. To assess reuse, we count both source and fresh responses. The carrier starts from the actor and takes one infinitesimal step along the mean log-probability gradient of N_+ successes among N_w source responses, skipping the write if $N_+ = 0$. With both policies then fixed, we evaluate the read in Equation (6) on K independent fresh batches of B responses ($1 \le B < n$), for a total cost of $N_\mathrm{w} + KB$.

The oracle knows u but is restricted to the current response-gradient span. It allocates the same total budget across K batches using past observations, choosing each batch size before sampling and then selecting the best-aligned direction.

In this family, first-order gains in success probability at a common local KL budget are proportional to cosine similarity with u (Kakade, 2001; Schulman et al., 2015). Define $\mathrm{Align}_J(d) = u^\top d / (\sqrt{m}\,\|d\|_2)$ for $d \neq 0$, with $\mathrm{Align}_J(0) = 0$. Let α_car and α_orc denote the alignments of the first-order carrier read and the oracle direction at a uniformly sampled batch, with empty estimates or batches assigned zero.

Theorem 2 (Retention advantage at equal response budget)**.** In this family, in the infinitesimal-write limit, (i) conditional on $N_+ > 0$, the first-order carrier read on a fresh batch $\mathcal{B}$ lies outside $\mathcal{S}_\mathcal{B}$ with probability at least $1 - B/n$. (ii) There exist finite N_w and K such that every oracle allocation described above satisfies

$$\mathbb{E}[\alpha_\mathrm{orc}] \le \sqrt{\frac{B + N_\mathrm{w}/K}{n}} < \mathbb{E}[\alpha_\mathrm{car}]. \tag{8}$$

Averaging over more successes reduces noise from irrelevant decisions. The read recovers this estimate on each fresh batch, amortizing the source cost to N_w/K per read. With sufficient source sampling and reuse, the carrier achieves a larger expected first-order gain in success probability than the oracle at the fixed actor, under equal total response and local KL budgets as stated in Equation (8) and proved in Section B.2.

In this family, earlier successes can still guide the carrier's read when the current group contains no successful response and hence provides no group-relative credit (Corollary 4). This motivates our evaluation on prompts rarely solved at initialization in Section 5.

5 EXPERIMENTS

We evaluate CBSD on four scientific domains of SciKnowEval, focusing on task performance and learning on initially difficult prompts. We then analyze carrier guidance, ablate the write correction and read timing, and compare computational cost.

5.1 EXPERIMENTAL SETUP

We use the public SciKnowEval-L3 split released with SDPO, covering Physics, Biology, Chemistry, and Materials (Feng et al., 2024). We train CBSD on Qwen3-8B (Yang et al., 2025) for 800 steps per domain, using separate LoRA adapters (Hu et al., 2022) for the actor and carrier. For the main comparison, we report the maximum Avg@16 across evaluated checkpoints, where Avg@16 is the mean accuracy over 16 responses per test prompt. We use the GRPO and SDPO results reported by Hübotter et al. (2026) for the four-domain comparison and run both baselines on Physics to compare learning dynamics and computational cost. Implementation details appear in Section C.

5.2 OVERALL PERFORMANCE

As reported in Table 1, CBSD achieves the highest average performance across the four domains, with an Avg@16 of 76.22, compared with 72.93 for SDPO and 71.05 for GRPO. It leads in Physics, Biology, and Materials and ranks second in Chemistry. In Physics, its Avg@16 of 80.16 exceeds SDPO and GRPO by 4.56 and 7.46 percentage points, respectively. The carrier also outperforms both baselines in Physics, Chemistry, and Materials.

Table 1: Maximum Avg@16 (%) on SciKnowEval. Average is the unweighted mean across domains. GRPO and SDPO results are from Hübotter et al. (2026). CBSD uses the best WCV setting per domain. As an auxiliary teacher, the carrier is reported separately at its own best checkpoint and excluded from ranking. Bold and underline indicate the best and second-best acting-policy results.

Method	Physics	Biology	Chemistry	Materials	Average
Initialization	59.14	29.88	41.19	59.91	47.53
GRPO	72.70	<u>59.90</u>	74.50	77.10	71.05
SDPO	<u>75.60</u>	56.80	**80.90**	<u>78.40</u>	<u>72.93</u>
CBSD	**80.16**	**64.25**	<u>80.63</u>	**79.85**	**76.22**
CBSD carrier (non-acting)	78.05	59.00	81.10	80.25	74.60

5.3 LEARNING DYNAMICS

CBSD achieves higher Physics accuracy with the same number of responses per step. Its maximum Avg@16 is 80.16, compared with 71.25 for GRPO and 73.05 for our SDPO run (Figure 3(a)).

CBSD improves learning on prompts for which successful responses are initially rare. We select the 27 Physics prompts with the lowest estimated success probability before training, using separate generations to estimate their initial accuracy. At step 200, CBSD exceeds GRPO by $29.40\,\mathrm{pp}$ on this cohort, with a 95% prompt-cluster interval of $[16.44, 42.59]\,\mathrm{pp}$ (Figure 3(b)). This result supports the motivation for retaining successes when fresh groups provide little reward information.

CBSD combines these accuracy gains with stable token entropy and concise successful responses. As shown in Figure 3(c), its token entropy remains relatively steady during training, while GRPO's approaches zero and SDPO's rises sharply. Successful responses remain short, while unsuccessful responses are longer and generally lengthen during training (Figure 3(d)).

5.4 ANALYSIS OF CARRIER-GUIDED LEARNING

We examine the carrier's token-level guidance and how closely the actor follows the carrier.

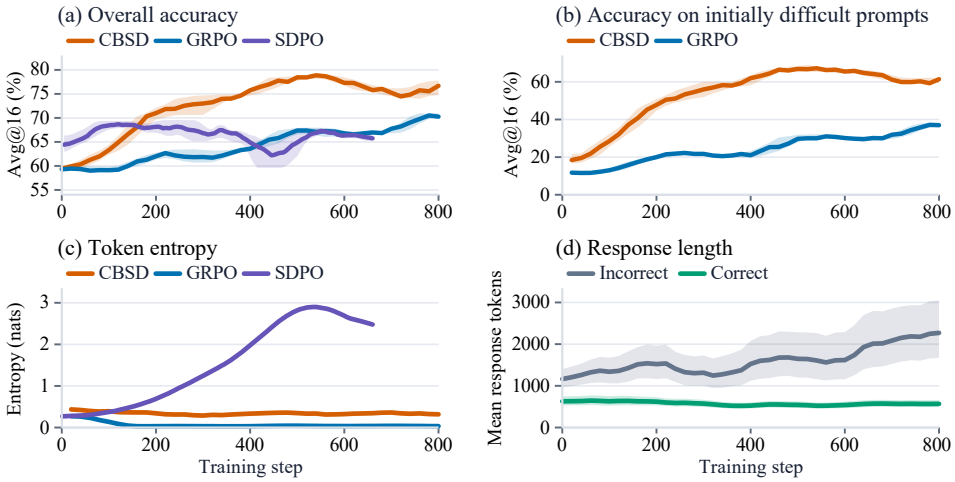

Figure 3: **Physics training dynamics.** (a) Overall Avg@16; (b) Avg@16 on 27 initially difficult prompts; (c) actor token entropy; (d) mean length of correct and incorrect CBSD responses. All curves are from our own runs and are smoothed. Shading indicates within-window interquartile ranges in (a,b) and smoothed 95% prompt-cluster bootstrap confidence intervals in (d).

Token-level guidance. Carrier advantages vary substantially within responses, while their sum becomes more informative about verified success during training. On 256 fixed actor responses, replacing token advantages with their respective response means preserves less than 0.5% of the total sum of squared advantages at steps 100, 400, and 800. Thus, a single weight per response captures little of the token-level variation. At the response level, we use the sum of token advantages to distinguish successes from failures. ROC-AUC rises from 0.527 at step 100 to 0.684 at step 800, indicating that these scores increasingly rank successful responses above unsuccessful ones.

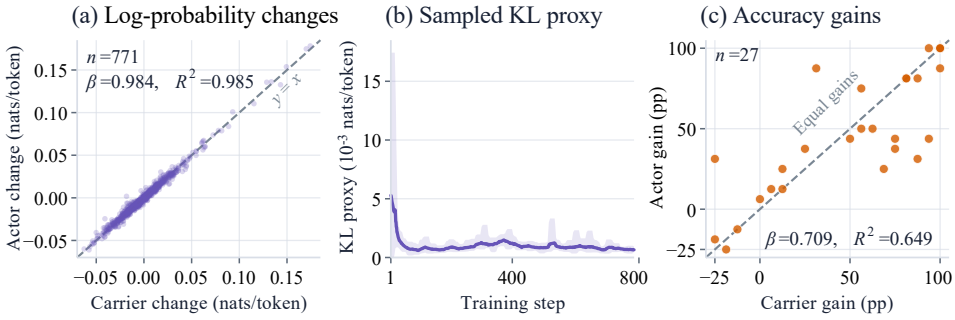

Figure 4: **The actor tracks the carrier.** (a) Adjusted changes in average token log-probability from initialization on the same responses. (b) Sampled KL proxy during training. Shading shows the range within each smoothing window. (c) Accuracy gains over initialization on initially difficult prompts. Dashed lines indicate equal changes in the actor and carrier.

Actor changes track carrier changes. The actor closely tracks the carrier during training. We compare changes in average token log-probability from initialization on shared responses, adjusting for the sampling step, verifier outcome, and response length. At step 400, the adjusted regression yields a slope near one and $R^2 \approx 0.985$, indicating that the actor closely follows both the pattern and magnitude of the carrier's changes (Figure 4(a)). On actor-generated responses, the sampled KL proxy is typically on the order of 10^{-3} nats per token (Figure 4(b)).

This correspondence extends to task accuracy, with larger carrier gains associated with larger actor gains across prompts. On the 27 initially difficult Physics prompts, the actor and carrier improve

Avg@16 by 47.69 and 50.93 percentage points, respectively, at step 400 relative to initialization. Regressing actor gains on carrier gains yields a slope of 0.709 and $R^2 = 0.649$ (Figure 4(c)).

5.5 Ablation studies

We examine how write variance reduction and read timing affect training and performance.

Write variance reduction. WCV improves accuracy across four domains, with the best coefficient depending on the domain. $\lambda_{\text{cv}} = 1$ performs best in Physics and Materials, while $\lambda_{\text{cv}} = 0.5$ performs best in Biology and Chemistry. Each selected setting outperforms the uncorrected write (Table 2). On Physics batches, $\lambda_{\text{cv}} = 1$ reduces the estimated variance of the carrier write gradient by a factor of approximately 4–17 relative to the uncorrected write (Section C.3).

Table 2: Effect of the WCV coefficient on maximum Avg@16 (%). Bold values are the best in each domain and are used in Table 1.

WCV coefficient λ_{cv}	Physics	Biology	Chemistry	Materials
0	75.63	62.00	80.03	78.59
0.5	79.14	**64.25**	**80.63**	78.66
1	**80.16**	57.63	80.33	**79.85**

Read timing. Reading the updated carrier immediately can destabilize training. Switching from pre-write to post-write reading drives the training success rate in Physics to zero at step 7, with mean response length subsequently reaching the 8192-token limit (Figure 5). At step 20, actor and carrier Avg@16 are both 0.00, compared with 60.08 for their pre-write counterparts. These results support pre-write reading for stable training.



Figure 5: **Pre-write reading stabilizes training.** (a) Training success rate. (b) Mean response length. Shading marks zero success in the post-write run, and the dotted line marks the length limit.

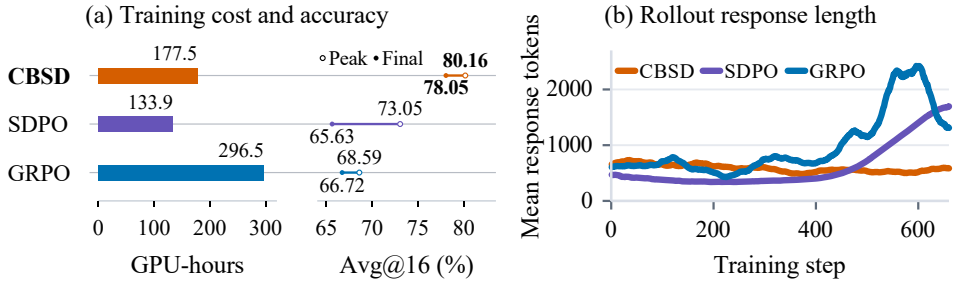

Figure 6: **Training cost and accuracy.** (a) Bars show total GPU-hours, including validation and checkpointing. Open and filled circles show peak and final Avg@16 over the same interval. (b) Mean response length during training (smoothed).

5.6 TRAINING EFFICIENCY

Over the same 660-step interval as our SDPO run, CBSD achieves higher peak and final accuracy than both baselines, using 40% fewer GPU-hours than GRPO and 33% more than SDPO (Figure 6(a)). Shorter responses than GRPO help offset the carrier's computational overhead (Figure 6(b)). Training the carrier adds a backward pass, while reusing its pre-write predictions for actor guidance avoids an additional carrier forward pass.

6 RELATED WORK

GRPO-style RLVR. Building on PPO (Schulman et al., 2017), GRPO uses group-relative advantages to weight response gradients (Shao et al., 2024). Subsequent work studies optimization design (Yu et al., 2025; Liu et al., 2025; Zheng et al., 2025; He et al., 2025) and develops finer-grained token-level credit assignment (Zhang et al., 2026a; Shan et al., 2026; Ding et al., 2026). Although actor updates persist across batches, GRPO's group-relative advantages vanish on all-failure groups under binary rewards. For finer-grained supervision, PRIME derives implicit process rewards from token log-probability ratios of an auxiliary outcome model trained online (Cui et al., 2025). CBSD fits a persistent carrier to verified successes by maximum likelihood and reads it on later actor samples, retaining verified evidence for token-level guidance across batches.

On-policy distillation. On-policy distillation extends knowledge distillation (Hinton et al., 2015) to learner-generated sequences (Agarwal et al., 2024). SDPO forms a self-teacher conditioned on textual feedback or successful rollouts (Hübotter et al., 2026). Related work explores the construction of distillation supervision (Li et al., 2026b; Shenfeld et al., 2026; Yu et al., 2026a; Tan & Hong, 2026; Li et al., 2026c; Wu et al., 2026; Feng et al., 2026; Li et al., 2026d; Cai et al., 2026), variance reduction (Oh et al., 2026), and the integration of verifiable rewards into distillation (Lin et al., 2026; Yang et al., 2026; Akhondzadeh et al., 2026; Xu et al., 2026; He et al., 2026; Zhang et al., 2026b; Li et al., 2026a). Many condition the teacher on information supplied for the current problem (Yu et al., 2026b). CBSD retains evidence from the actor's verified successes in a persistent carrier, which later provides token-level guidance using only newly sampled actor prefixes.

Retaining and exchanging generated experience. Prior methods reuse experience through context, replay, or trajectory exchange. OPCD and OEL internalize experience through a context-conditioned teacher (Ye et al., 2026b;a). Replay-based RLVR reuses verified or high-value rollouts (Yoo et al., 2026), self-verified distillation trains on filtered generations (Lee & Liang, 2026), and Mutual RL exchanges trajectories between acting policies (Liu et al., 2026). CBSD stores verified evidence in a non-acting carrier and reads it on fresh actor samples, without retaining source responses as context, replaying them, or adding carrier rollouts.

7 LIMITATIONS AND CONCLUSION

Limitations. Our analysis provides an idealized account of the write–read mechanism, while its connection to learned credit and the full training dynamics warrants further study. Broader evaluation across models, tasks, and random seeds is needed to assess the generality of our findings.

Conclusion. CBSD retains information from verified successes in a separate carrier policy and reuses it across batches through on-policy self-distillation. The actor receives token-level guidance on newly sampled responses without replaying earlier responses or requiring carrier rollouts. WCV preserves the expected write and reduces its variance in our experiments, while pre-write reading supports stable training. Our analysis characterizes the read's additional expressivity and shows that retention can yield greater expected first-order gains than current-batch response-level weighting in a binary-verifier family. The comparison matches response and local KL budgets, including source responses. On SciKnowEval, CBSD achieves higher average accuracy than the reported GRPO and SDPO baselines across four domains. It also improves learning on initially difficult prompts.

AI USE STATEMENT

In this work, we used generative AI tools to assist with developing the conceptual and theoretical account of CBSD's write–read mechanism, formulating mathematical claims and developing proofs concerning token-level credit and cross-batch reuse, proposing and refining research hypotheses, designing and reviewing experiments, implementing the method, supporting qualitative analysis, and interpreting results. We did not use generative AI tools for translation, synthetic dataset generation, or dataset cleaning and reformatting. Additionally, we used these tools to draft and revise manuscript text, search and analyze related literature, organize references, create or modify scientific figures, suggest experimental parameters, and develop auxiliary code. We have reviewed all AI-assisted work. The authors revised the text, checked citations against original sources, examined the theoretical assumptions, derivations, and proofs step by step, reviewed and tested the implementation, inspected data processing outputs, and verified the reported SciKnowEval results and figures against the underlying data and experimental records. We also assessed the originality of research ideas through a literature review. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We describe CBSD's write–read updates and WCV in Section 3 and provide the training algorithm in Section A. The assumptions and complete proofs supporting our theoretical results are given in Section B. Section C documents the data split, model and optimization settings, evaluation and checkpoint-selection procedures, and baseline configurations, with the main training configuration summarized in Table 3. It also specifies the selection of initially difficult prompts, uncertainty estimation, and procedures for measuring carrier guidance and write variance. To facilitate reproducibility, we will open-source our code and release supporting materials after the paper is accepted.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation, 2026. URL https://arxiv.org/abs/2607.04037.

Qiye Cai, Yichuan Ma, Peiji Li, Yongkang Chen, Qipeng Guo, Yicheng Zou, Linyang Li, Xiaocheng Feng, and Bing Qin. H^2SD: Hybrid hindsight self-distillation, 2026. URL https://arxiv.org/abs/2607.18955.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025. URL https://arxiv.org/abs/2502.01456.

Tianyu Ding, Jianhong Xin, and Juan Pablo De la Cruz Weinstein. Keep policy gradient in charge: Sibling-guided credit distillation for long-horizon tool-use agents, 2026. URL https://arxiv.org/abs/2606.12634.

Kehua Feng, Xinyi Shen, Weijie Wang, Xiang Zhuang, Yuqi Tang, Qiang Zhang, and Keyan Ding. SciKnowEval: Evaluating multi-level scientific knowledge of large language models, 2024. URL https://arxiv.org/abs/2406.09098.

Yangyang Feng, Zhuoyan Feng, and Junlan Chen. PAST: Privileged adaptation from complete student trajectories for on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.08726.

Evan Greensmith, Peter Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/584b98aac2dddf59ee2cf19ca4ccb75e-Paper.pdf.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

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 https://doi.org/10.1038/s41586-025-09422-z.

Shenghua He, Tian Xia, Xuan Zhou, and Hui Wei. Response-level rewards are all you need for online reinforcement learning in LLMs: A mathematical perspective, 2025. URL https://arxiv.org/abs/2506.02553.

Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision, 2026. URL https://arxiv.org/abs/2604.12002.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, April 2022. URL https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *International Conference on Machine Learning*, July 2026.

Sham M Kakade. A natural policy gradient. In T. Dietterich, S. Becker, and Z. Ghahramani (eds.), *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/4b86abe48d358ecf194c56c69108433e-Paper.pdf.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning meets Structured Prediction*, 2019.

Tony Lee and Percy Liang. Self-verified distillation: Your language model is secretly its own synthetic data pipeline, 2026. URL https://arxiv.org/abs/2605.26132.

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, 2026a. URL https://arxiv.org/abs/2604.02288.

Yang Li, Erik Nijkamp, Semih Yavuz, and Shafiq Joty. Learning from language feedback via variational policy distillation, 2026b. URL https://arxiv.org/abs/2605.15113.

Yijiang Li, Bingyang Wang, Yijun Liang, Yunjie Tian, Di Fu, and Nuno Vasconcelos. On-policy self-distillation without any supervision, 2026c. URL https://arxiv.org/abs/2608.06296.

Yu Li, Shu Hong, and Tian Lan. Localizing credit at the divergence: Path-conditioned self-distillation for LLM reasoning, 2026d. URL https://arxiv.org/abs/2606.15576.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.494. URL https://aclanthology.org/2020.emnlp-main.494/.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward, 2026. URL https://arxiv.org/abs/2608.24696.

Xiaoze Liu, Dhananjay Ram, Yuting Zhang, Zhaoyang Zhang, Wei Xia, and Stefano Soatto. Experience sharing in mutual reinforcement learning for heterogeneous language models, 2026. URL https://arxiv.org/abs/2605.07244.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025.

Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. 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. 3878–3887. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/oh18b.html.

Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. KL for a KL: On-policy distillation with control variate baseline, 2026. URL https://arxiv.org/abs/2605.07865.

Du Phan, Matthew Douglas Hoffman, David Dohan, Sholto Douglas, Tuan Anh Le, Aaron Parisi, Pavel Sountsov, Charles Sutton, Sharad Vikram, and Rif A. Saurous. Training chain-of-thought via latent-variable inference. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 72819–72841. Curran Associates, Inc., 2023. doi: 10.52202/075280-3184. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e69a9560c450ca76584d9eb37e7f5ae8-Paper-Conference.pdf.

Daniel Russo. Success-conditioning as policy improvement: The optimization problem solved by imitating success. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65267.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Yingyu Shan, Yuhang Guo, Zihao Cheng, Zeming Liu, Xiangrong Zhu, Xinyi Wang, Jiashu Yao, Wei Lin, Hongru Wang, and Heyan Huang. Learning from own solutions: Self-conditioned credit assignment for reinforcement learning with verifiable rewards, 2026. URL https://arxiv.org/abs/2606.18810.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *International Conference on Machine Learning*, 2026.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron Parisi, Abhishek Kumar, Alex Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models, 2024. URL https://arxiv.org/abs/2312.06585.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context, 2022. URL https://arxiv.org/abs/2209.15189.

Miroslav Štrupl, Francesco Faccio, Dylan R. Ashley, Rupesh Kumar Srivastava, and Jürgen Schmidhuber. Reward-weighted regression converges to a global optimum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8361–8369, June 2022. doi: 10.1609/aaai.v36i8.20811. URL https://doi.org/10.1609/aaai.v36i8.20811.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf.

Zhiquan Tan and Yinrong Hong. Self-supervised on-policy distillation for reasoning language models, 2026. URL https://arxiv.org/abs/2605.17497.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with reinforced fine-tuning. 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. 7601–7614, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.410. URL https://aclanthology.org/2024.acl-long.410/.

Chen Wang, Lai Wei, Yanzhi Zhang, Chenyang Shao, Zedong Dan, Weiran Huang, Yuzhi Zhang, and Yue Wang. EFRame: Deeper reasoning via exploration-filter-replay reinforcement learning framework, 2025. URL https://arxiv.org/abs/2506.22200.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In *International Conference on Learning Representations*, volume 2026, pp. 49450–49483, 2026.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Jianyu Wu, Yizhou Wang, Encheng Su, Chen Tang, and Shixiang Tang. DAPD: Dual-anchored policy distillation, 2026. URL https://arxiv.org/abs/2608.01735.

Haoran Xu, Hongyu Wang, Yifei Gao, Jiaze Li, Xiaofeng Zhang, and Xiaosong Yuan. SG-OPD: Sign-gated on-policy distillation via sign-consistency gating and phased teacher sampling, 2026. URL https://arxiv.org/abs/2606.09304.

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.

Yuxiao Yang, Xiaoyun Wang, and Weitong Zhang. OGLS-SD: On-policy self-distillation with outcome-guided logit steering for LLM reasoning, 2026. URL https://arxiv.org/abs/2605.12400.

Tianzhu Ye, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Online experiential learning for language models, 2026a. URL https://arxiv.org/abs/2603.16856.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models, 2026b. URL https://arxiv.org/abs/2602.12275.

Gyeongtae Yoo, Sanghyeok Park, Soohyuk Jang, Ik hwan Kim, and Sungroh Yoon. Rollout-level advantage-prioritized experience replay for GRPO, 2026. URL https://arxiv.org/abs/2606.04560.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures, 2026a. URL https://arxiv.org/abs/2605.12652.

Xinlei Yu, Gen Li, Qingyi Si, Guibin Zhang, Yuqi Xu, Congcong Wang, Shuai Dong, Kaiwen Tuo, Xiangyu Zeng, Kaituo Feng, Qunzhong Wang, Yang Shi, Xiaobin Hu, Xiangyu Yue, Jiaqi Wang, and Shuicheng Yan. DOPD: Dual on-policy distillation, 2026b. URL https://arxiv.org/abs/2606.30626.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping reasoning with reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/639a9a172c044fbb64175b5fad42e9a5-Paper-Conference.pdf.

Hongzhi Zhang, Jia Fu, Jingyuan Zhang, Kai Fu, Qi Wang, Fuzheng Zhang, and Guorui Zhou. RLEP: Reinforcement learning with experience replay for LLM reasoning, 2025a. URL https://arxiv.org/abs/2507.07451.

Kaiyi Zhang, Wei Wu, and Yankai Lin. DelTA: Discriminative token credit assignment for reinforcement learning from verifiable rewards, 2026a. URL https://arxiv.org/abs/2605.21467.

Yuheng Zhang, Wenlin Yao, Changlong Yu, Yao Liu, Qingyu Yin, Bing Yin, Hyokun Yun, and Lihong Li. Improving sampling efficiency in RLVR through adaptive rollout and response reuse, 2025b. URL https://arxiv.org/abs/2509.25808.

Zhu Zhang, Jixun Wang, Xiaoang Xu, Xiaorong Wang, Zihan Zhou, Zhiyuan Wang, Shuo Wang, Chaojun Xiao, and Yuezhi Zhou. Beyond teacher likelihood: Group-calibrated on-policy distillation for long-context reasoning, 2026b. URL https://arxiv.org/abs/2608.19181.

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. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64784.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.