Instruction-Conditioned Exploration with Asymmetric Reinforcement Learning and Self-Distillation

Anonymous ACL submission

Abstract

Post-training Large Language Models (LLMs) with Reinforcement Learning (RL) has become an important tool for improving model capabilities, but the LLM action-space structure introduces challenges distinct from classical RL, with implications for inducing exploration. New methods are required that leverage the broad knowledge and flexibility of pre-trained LLMs to deliberately generate diverse experience at training time. We propose Instruction-Conditioned Exploration (ICE), which supplements task prompts during training with one of several distinct instructions, increasing the coverage of behaviours attempted. To facilitate ICE, we propose Asymmetric-RL/SD, a combined Reinforcement Learning and Self-Distillation training objective, to transfer explored behaviours to the unconditioned test-time policy. ICE with the Asymmetric-RL/SD objective improves Qwen3-1.7B held-out pass@1 performance at 4K response length on mathematical reasoning tasks by 5.0% relative to training with DAPO, with improvement persisting at a longer 8K context.

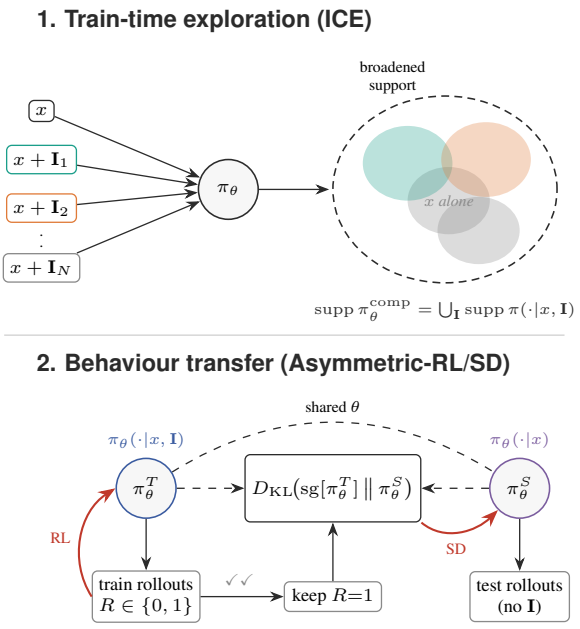

Figure 1: Overview. **(1)** ICE supplements the task prompt x with one of several behavioural instructions $\mathbf{I}$ per rollout, broadening the support available to explore at training time. **(2)** Asymmetric-RL/SD optimises instruction-conditioned teacher rollouts with reinforcement learning, with simultaneous self-distillation of the correct teacher trajectories into the unconditioned student, which is the only policy used at test time.

1 Introduction

Post-training with Reinforcement Learning (RL) is increasingly used to improve and tune Large Language Models (LLMs), including for general-purpose agents (Qi et al., 2025) and deployed real-world systems (Jiang et al., 2025). Underlying these applications is a need for sustained exploration of diverse behaviours during training: on-policy RL provides direct learning signal only for sampled actions; behaviours never produced during training cannot be directly reinforced. Classical RL exploration methods rely on stochasticity in action selection (e.g. ϵ-greedy action selection or action-noise injection), assuming an action space structured such that random perturbation produces useful candidate trajectories (Sutton and Barto, 2018).

In LLMs this assumption fails: "actions" correspond to tokens drawn from vocabularies of tens of thousands, and tokens must be combined in highly specific sequences to be meaningful, so random token-level perturbation lands off the manifold of coherent text rather than producing novel viable trajectories. Nor does restricting sampling to probable tokens resolve this, as superficially diverse text can carry the same meaning (Wang et al., 2026); behavioural diversity requires exploration above the token level. Compounding the problem, RL post-training methods are known to sharpen rather than broaden the model's distribution (Yue et al., 2025), motivating new approaches to increase the

coverage of behaviours experienced during training, what Wu et al. (2026) call *support expansion*.

Existing exploration methods operate at the token or sequence level (Cui et al., 2025; He et al., 2025), with no mechanism for surfacing low-probability behaviours, or use context mediation (Lu et al., 2026; Szot et al., 2026) that either relies on capability-bounded self-feedback or lacks an explicit mechanism for transferring train-time behaviours to test time.

To address both limitations, we introduce **Instruction-Conditioned Exploration** (ICE) and the **Asymmetric-RL/SD** training approach (Figure 1). **ICE** is our novel approach to generating diverse training samples: at training time, we condition text generation on instructions sampled from a set of generic, distinct task strategies, inducing diverse rollouts from this *instruction-conditioned* (teacher) policy. The same parameters, queried without a supplementary instruction, define an *unconditioned* (student) policy.

Asymmetric-RL/SD is our novel self-distillation (SD) training objective combining two components trained on a shared set of teacher rollouts. First, a reward-maximising RL term updates the teacher policy (we use the DAPO (Yu et al., 2025) variant of the GRPO RL algorithm (Shao et al., 2024)). Second, a forward-KL distillation term updates the student policy only, pulling it toward the teacher's correctness-filtered rollouts so that at test time the unconditioned policy imitates useful behaviours that the teacher learns during training.

Empirically, on Qwen3 (Yang et al., 2025) at 1.7B parameters and 4K response length, ICE with Asymmetric-RL/SD improves the macro-mean pass@1 across held-out mathematical reasoning benchmarks by 5.0% relative to DAPO, with 5/5 benchmark wins and a 95% bootstrap CI excluding zero ($n = 5$). The improvement over ICE with RL alone is smaller and less certain. A smaller improvement persists at the longer-context 1.7B/8K setting, while at 4B/4K it does not improve over DAPO. Improvements to small models matter in practice: those small enough to run on consumer phones (Abdin et al., 2024) and laptops (Javat and Kazakov, 2026) are increasingly capable, and raising their performance widens the range of tasks for which they can be deployed.

Our contributions are as follows:

- A novel method of instruction conditioning to induce diverse rollouts from LLMs, driving exploration during RL post-training (ICE).
- Definition of the novel Asymmetric-RL/SD objective combining a shared-parameter teacher policy, updated through RL, with a student policy updated through self-distillation.
- A multi-seed empirical study at various model scales, context lengths and self-distillation settings, demonstrating the improvement of ICE with Asymmetric-RL/SD over a DAPO baseline at the 1.7B parameter scale.

2 Related Work

2.1 Exploration in RL post-training of LLMs

A considerable focus of recent work on LLM post-training is using RL to improve performance on tasks with verifiable outcomes (RLVR). Learning under RL requires actions producing viable trajectories to be sampled with non-negligible probability during training, motivating directed exploration rather than the stochastic methods sufficient in classical RL.

This need is reinforced by recent findings that RL post-training primarily sharpens the distribution over outputs already accessible to the base model, raising pass@1 but reducing pass@k coverage at large k and so narrowing the reasoning boundary (Yue et al., 2025; Zhang et al., 2025). Wu et al. (2026) also find that RLVR reduces the support of the response distribution, excluding previously available correct answers (*support shrinkage*), and call for new methods that instead achieve *support expansion*, up-weighting previously low-probability completions.

Token-level and sequence-level diversity interventions (Cui et al., 2025; Wang et al., 2025; Shen, 2025; Chen et al., 2025b; He et al., 2025; Zhang, 2026; Song et al., 2025; Hu et al., 2026; Liang et al., 2025) show promise in preventing entropy collapse but do not directly address the need for support expansion: superficial token diversity can mask semantic similarity (what Wang et al. (2026) call *template collapse*).

2.2 Context-mediated exploration in LLMs

Context-mediated exploration methods modify the context provided to the model during training, aiming to achieve support expansion relative to a singular context, through pre-defined hints, self-feedback, or pre-determined templates.

Pre-defined hint based methods (Zhou et al., 2026; Zhang et al., 2026; Liao et al., 2026; Chen et al., 2025a) selectively provide the LLM with additional information about how to approach a problem, but require privileged information from expert annotators, more powerful models, or knowledge of the ground-truth answer, limiting scalability.

Self-feedback approaches (Madaan et al., 2023; Bai et al., 2022; Shinn et al., 2023; Kim et al., 2023; Kumar et al., 2024; Li et al., 2025; Szot et al., 2026) let the model make multiple attempts, with previous attempts fed back for critique, refinement, or as demonstrations of failure. Although not typically framed as exploration techniques, these repeated attempts under different contexts broaden the strategies the model produces, achieving support expansion. Their common limitation is reliance on attempt-level self-feedback, which increases rollout cost, remains bounded by current capability, and offers no explicit control over the diversity of explored strategies.

Template approaches use a fixed set of predetermined templates, each conditioning the task description with a different context during training and directly driving support expansion relative to the unconditioned policy. This is exemplified by *Prompt Augmentation* (Lu et al., 2026), which conditions each training example on one of a set of response-format templates. However, this introduces a train/test distributional mismatch: the model is trained on templated prompts but evaluated on the raw task prompt.

2.3 Self-Distillation in LLM Post-Training

We build on self-distillation, a form of knowledge distillation (Hinton et al., 2015) in which teacher and student policies share weights but differ in input context, with the teacher typically conditioned on privileged train-time information. Recent variants span on-policy (Zhao et al., 2026; Shenfeld et al., 2026; Hübotter et al., 2026) and off-policy (Bhargava et al., 2024; Zhai et al., 2025; Penaloza et al., 2026) approaches.

We extend π-distill (Penaloza et al., 2026), a flexible off-policy method combining teacher and student RL objectives with reverse and forward KL terms, replacing its privileged information with a behavioural instruction $\mathbf{I}$.

3 Instruction-Conditioned Exploration for LLM RL

We now introduce Instruction-Conditioned Exploration (ICE), our template-based approach to support expansion, beginning with the DAPO formulation it builds on (§3.1) before defining the composite instruction-conditioned policy and its RL objective (§3.2). The test-time transfer problem is addressed separately in §4.

3.1 Preliminaries: RL for LLMs

When applying reinforcement learning to LLMs, we treat the model with parameters θ as a policy π_θ that generates text responses y to input texts x (the task prompt) and receives reward $R(y, x)$. For mathematical reasoning, x contains a question with a verifiable answer; the response y receives reward 1 if the extracted answer is correct and 0 otherwise. The RL objective is to maximise $J_{\text{RL}}(\theta) = \mathbb{E}_{y\sim\pi_\theta}[R(y, x)]$ over the task distribution (Sutton and Barto, 2018).

GRPO (Shao et al., 2024) replaces R with a group-normalised advantage $A_i = (R_i - \mu)/\sigma$ over G responses sampled from $\pi_{\theta_{\text{old}}}$ for the same input x, where μ, σ are the mean and standard deviation of the group rewards. DAPO (Yu et al., 2025), the predominant GRPO variant, adds token-level loss aggregation, asymmetric clipping, dynamic sampling, and overlong reward shaping; we use it as both our baseline and the RL component of our method.

3.2 Instruction-Conditioned Exploration

We introduce a new approach to template-based context-mediated exploration in language models, instruction-conditioned exploration (ICE). ICE bears similarity to Prompt Augmentation (Lu et al., 2026), but supplements task descriptions with opinionated, diverse *behavioural* instructions rather than solution-agnostic formatting instructions. This broadens the support of the generated text distribution relative to a singular task prompt, what we term *support composition* to distinguish it from *support expansion* (defined by Wu et al. (2026), see §2.1), for which the expansion is due only to parameter changes with respect to the base model.

Formally, we define a set $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$ of N instructions that can be selected from during training to supplement the task description. The instructions should contain high-level guidance that pushes the response of the LLM into distinct di-

rections, without over-specifying and confining responses to a small region for each instruction. For example, for mathematical reasoning problems we might use the instruction: "*Ask: what symmetry does this problem have, and how can it reduce the number of cases or simplify the expression?*". If $p(\mathbf{I})$ is the probability that instruction $\mathbf{I}$ will supplement the task definition, the composite ICE policy becomes:

$$\pi_\theta^{\text{comp}}(y|x) = \sum_{\mathbf{I}\in\mathcal{I}} p(\mathbf{I})\pi_\theta(y|x,\mathbf{I}) \tag{1}$$

with the support of this policy being the union of the supports of its components

$$\text{supp}(\pi_\theta^{\text{comp}}(\cdot|x)) = \bigcup_{\mathbf{I}\in\mathcal{I}} \text{supp}(\pi_\theta(\cdot|x,\mathbf{I})) \tag{2}$$

We take $\emptyset \in \mathcal{I}$ to denote the unconditioned case having $\pi_\theta(\cdot|x,\emptyset) = \pi_\theta(\cdot|x)$, which is always part of the instruction set so that $\text{supp}(\pi_\theta^{\text{comp}}(\cdot|x)) \supseteq \text{supp}(\pi_\theta(\cdot|x))$. The RL objective with ICE is then $J_{\text{ICE}}(\theta) = \mathbb{E}_{y\sim\pi_\theta^{\text{comp}}}[R(y,x)]$.

Each rollout receives a single instruction $\mathbf{I}$ (or none) appended to the task prompt. How instructions are distributed across the rollouts of a task is a design choice: they may be sampled per rollout, or allocated so that each group covers several instructions and unconditioned rollouts. The model then generates text conditioned on the augmented prompts. This approach introduces a trade-off: while coverage of different problem-solving strategies may broaden the support of the policy during training, $\mathbf{I}$ could also provide unhelpful instructions, reducing accuracy and therefore positive signal during training.

We combine this with the GRPO RL objective function, allowing each response to a specific input x to be conditioned on different instructions, thereby creating diversity within each group beyond that of sampling alone.

4 Self-Distillation of LLMs

The use of a composite policy at training time requires the transfer of useful behaviours discovered during training to the unconditioned test-time policy. Parameter sharing enables implicit transfer, but an explicit transfer mechanism can strengthen it. We extend the π-distill off-policy self-distillation objective of Penaloza et al. (2026) to provide this mechanism, and define our *Asymmetric-RL/SD* variant as the configuration used in our main experiments.

In off-policy self-distillation, trajectories sampled from a teacher policy are used to update both the teacher and student policies, with the teacher policy distinguished from the student policy by having additional, privileged information about the task or its solution. In our setting, the privileged information is replaced with a behavioural instruction $\mathbf{I}$.

Let $\pi_\theta^T := \pi_\theta(\cdot \mid x,\mathbf{I})$ and $\pi_\theta^S := \pi_\theta(\cdot \mid x)$ denote the instruction-conditioned (teacher) and unconditioned (student) policies, respectively, sharing parameters θ. The π-distill objective combines reward-maximising terms for both policies, weighted by α, with two KL-divergence terms between the sampled teacher and student policies, weighted by β. Writing $\text{sg}[\cdot]$ for the stop-gradient operator, D_{KL} for the KL divergence, and $R(y,x)$ for the reward of response y on task x, the reward terms are the teacher reward $J_T^R = \mathbb{E}_{y\sim\pi_\theta^T}[R(y,x)]$ and the importance-corrected student reward $J_S^R = \mathbb{E}_{y\sim\pi_\theta^T}[(\pi_\theta^S/\text{sg}[\pi_\theta^T])\,R(y,x)]$, both estimated on teacher rollouts. The two KL terms are a reverse-KL term $J_T^{KL} = -D_{KL}\big(\pi_\theta^T \,\|\, \text{sg}[\pi_\theta^S]\big)$ that pulls the teacher toward the student, and a forward-KL term $J_S^{KL} = -D_{KL}\big(\text{sg}[\pi_\theta^T] \,\|\, \pi_\theta^S\big)$ that pulls the student toward the teacher.

We propose to modify this objective as $J(\theta) = J^R(\theta) + J^{KL}(\theta)$ with components

$$J^R(\theta) = \alpha J_T^R(\theta) + (1-\alpha)J_S^R(\theta) \tag{3}$$

$$J^{KL}(\theta) = \beta_T J_T^{KL}(\theta) + \beta_S J_S^{KL}(\theta) \tag{4}$$

reusing the reward and KL terms defined above. We retain the constraint on reward weighting to avoid adding overall reward scale as an additional degree of freedom (absorbed by the learning rate). The KL terms do not require such a constraint as they trade off against the reward signal, not against each other. The original π-distill objective is recovered by setting $\beta_T = \alpha\beta$ and $\beta_S = (1-\alpha)\beta$.

We define Asymmetric-RL/SD as the case using the student-KL term without student reward ($\alpha = 1$, $\beta_S > 0$), additionally filtering teacher-sampled responses to only include those giving a correct answer, $R(y,x) = 1$. The π-distill motivation for including the complete, unfiltered KL terms is to keep the importance sampling ratio in the student reward term close to 1 for all training samples. In Asymmetric-RL/SD, this motivation no longer applies. This is closer to a standard knowledge distillation approach in which we want the student

policy to learn to replicate good responses generated by the teacher policy.

5 Experimental Setup

To assess our methods, we apply ICE with modified π-distill objectives, including Asymmetric-RL/SD, to mathematical reasoning problems using Qwen3 (Yang et al., 2025) models with thinking mode disabled, consistent with recent post-training and self-distillation work (Lu et al., 2026; Liao et al., 2026; Zhao et al., 2026; Yang et al., 2026). We compare a baseline DAPO configuration to models trained with ICE, comparing variants across selected (α, β_T, β_S) at 1.7B and 4B parameters, and 4K and 8K (1.7B only due to compute budget constraints) token response lengths. Models are trained for up to 7 epochs of the English language subset of the deduplicated DAPO-Math-17K dataset (Yu et al., 2025) (14.1k examples), "DAPO-17K-en". Training was conducted on an HPC with NVIDIA GH200 superchips, using 2 GPUs per seed for 1.7B configurations and 4 GPUs for 4B. Total compute and per-configuration breakdown are reported in Appendix B, and artifact licenses in Appendix C.

We select $N = 5$ instructions via an automated procedure using frontier models (Claude Sonnet 4.6 and Claude Opus 4.6) to propose candidate strategies and the target model (Qwen3 4B) to select among them by coverage on a held-out 250-problem training subset, yielding the instructions in Table 6. Full details are in Appendix A. The selection uses training-set problems, leaving evaluation uncontaminated. In practice, this process could be replaced by human expert instructions.

Algorithm Details and Hyperparameters Hyperparameters are not separately tuned; we adopt standard values commonly used in recent RL post-training work (Yu et al., 2025; Guo et al., 2025; Lu et al., 2026; Liao et al., 2026). Our implementation builds on the OpenRLHF framework (Hu et al., 2024). To train the models, we used the DAPO algorithm described in §3.1 with rollout group size $G = 8$, and clipping bounds $\epsilon_{\text{low}} = 0.2, \epsilon_{\text{high}} = 0.28$. Rollouts are truncated above 4096/8192 tokens (4K/8K) depending on the setting, with overlong penalty above 3072/6144 tokens respectively. Rollout text is generated with temperature $T = 1.0$, top_p $= 1.0$. Learning rate is $1\text{e}{-6}$, constant after 320 warmup steps. Batch sizes are 128 for rollout and training, with micro-batch size 16. For instruction-conditioned training each of the $N = 5$ supplementary instructions is appended to the task prompt for one rollout, and the remaining 3 rollouts use the task prompt alone. Method variants are summarised in Table 1.

Evaluation We evaluate the post-trained models on five held-out mathematical reasoning benchmarks: AIME-24 (Zhang and Math-AI, 2024), AIME-25 (Zhang and Math-AI, 2025), MATH-500 (Hendrycks et al., 2021), HMMT-Feb-25, and HMMT-Nov-25 (Balunović et al., 2026). MATH-500 contains 500 questions; the AIME and HMMT benchmarks each contain 30 more difficult and more recent questions. All evaluation is performed on the unconditioned policy π_θ^S (defined in §4), without any instructions in the context. At test time we have no mechanism for determining which supplementary instructions will be beneficial; this must be learned by the student policy.

We additionally evaluate on DAPO-200, a fixed 200-question subset of DAPO-17K-en, which provides a training-set performance measure when models are evaluated unconditioned.

Evaluation responses are sampled at $T = 0.6$, top_p $= 1.0$, $n = 16$ for AIME/HMMT, and $n = 1$ for MATH-500/DAPO probe. We use pass@1 (averaged for $n = 16$). To identify the best performing evaluation step fairly across runs, we use the DAPO-200 reward to select the best evaluation step of each training run, then report held-out scores at that step. As seed counts are limited, we read results from complementary signals rather than any single statistic: the cross-seed mean$\pm$std, Δ vs DAPO, per-benchmark wins, and, for the five-seed headline, a 95% seed-paired bootstrap interval on Δ (an uncertainty estimate, not a significance test); we treat a result as reliable only where these agree.

6 Results

We present results in four parts. §6.1 sweeps method variants at 1.7B/4K, first with a single seed and then with progressively more seeds for the most promising configurations. §6.2 reports the headline five-seed comparison establishing ICE with Asymmetric-RL/SD as the strongest configuration at 1.7B/4K. §6.3 tests whether the gain transfers to longer responses (1.7B/8K) and a larger model (4B/4K). §6.4 separates per-sample accuracy from coverage across sampling budgets for pass@k up to $k = 256$.

Variant	α	β_T	β_S	Description
DAPO	–	–	–	Vanilla DAPO-GRPO; no teacher/student split
$\alpha = 0$	0	0	0	Strip-and-retrain (student-only reward maximisation)
$\alpha = 0.5$	0.5	0	0	Joint blend of teacher and student reward maximisation
$\alpha = 1$	1	0	0	Teacher-only
$\alpha = 0.5, \beta$	0.5	$\{0.05, 0.1\}$	$\{0.05, 0.1\}$	Joint blend + symmetric KL (both directions)
Asymmetric-RL/SD	1	0	$\{0.05, 0.1, 0.5\}$	Teacher-only, Student $\rightarrow$ Teacher KL on filtered samples
unfiltered	1	0	0.1	As above, without the correctness filter

Table 1: Method variants.

6.1 Preliminary Experiments at 1.7B/4K

A broad single-seed sweep over $(\alpha, \beta_T, \beta_S)$ (Table 7, Appendix E) identifies α=0.5 (β=0.1), $\alpha = 1$, and Asymmetric-RL/SD (β_S=0.1) as the promising configurations. It also shows that correctness filtering is essential: the unfiltered forward-KL variant falls 12.9% below DAPO. A student reward term ($\alpha < 1$) instead makes training unstable, with response-length blow-up (Appendix I). These findings motivate the teacher-only, correctness-filtered Asymmetric-RL/SD design. We carry the promising configurations to the multi-seed evaluations below.

Repeating the most promising configurations across three seeds (Table 2) shrinks several of the single-seed effects. The $\alpha = 0.5, \beta_T = \beta_S = 0.1$ setting falls from $+0.024$ to $+0.004$ with high variance (± 0.015). The $\alpha = 1$ result decreases from $+0.029$ to $+0.013$ and 4/5 wins. Asymmetric-RL/SD holds at $+0.016$, 5/5 wins. We carry the two latter configurations forward.

6.2 ICE with Asymmetric-RL/SD is the strongest configuration at 1.7B/4K

Extending $\alpha = 1$ and Asymmetric-RL/SD $\beta_S = 0.1$ to five seeds (Table 3), the $\alpha = 1$ improvement reduces further to $+0.005$ ($+1.9\%$) with a CI spanning zero ($[-0.006, +0.018]$), although 4/5 benchmark wins are retained. The Asymmetric-RL/SD configuration is robust at $+0.012$ ($+5.0\%$, $[+0.006, +0.016]$), 5/5 wins. While inclusion of student policy updates through the forward-KL term improves mean pass@1 by $+0.008$, the 95% CI marginally spans zero ($[-0.003, +0.016]$).

Instruction conditioning does not slow convergence: the probe pass@1 of the leading 1.7B/4K configurations tracks DAPO throughout training (Figure 2; Appendix H).

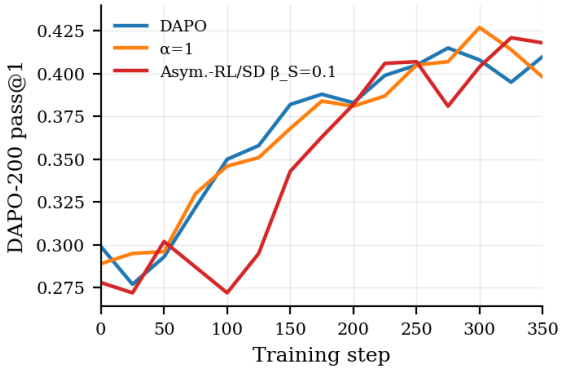

Figure 2: DAPO-200 probe pass@1 during training at 1.7B / 4k. Curves are means across n=5 seeds per method.

6.3 Replication at longer response length and larger model scale

We replicate Asymmetric-RL/SD at 1.7B/8K and 4B/4K, and $\alpha = 1$ at 4B/4K (Table 4), with 3 seeds per cell (2 for Asymmetric-RL/SD at 4B; Table 4). Asymmetric-RL/SD at 1.7B/8K is the only setting to exceed DAPO ($+0.010$, $+3.5\%$). At 4B/4K neither configuration improves on DAPO: $\alpha = 1$ scores -0.003 and Asymmetric-RL/SD -0.012 (-3.3%), though its two completed seeds fall either side of the DAPO mean. The improvement over DAPO observed at 1.7B/4K, and to a lesser extent at 1.7B/8K, does not transfer to the larger model under the 4K-token response limit.

6.4 Accuracy and coverage across sampling budgets

Sampling at increasing budgets separates per-sample accuracy (pass@1) from coverage (pass@k at large k; Figure 3, Table 5). Asymmetric-RL/SD leads at pass@1, but at high k the ordering inverts: RL post-training trades coverage for accuracy, leaving both DAPO and Asymmetric-RL/SD below the base model by pass@256 (0.483 and 0.475 vs 0.513). Asymmetric-RL/SD thus improves pass@1 over DAPO at comparable coverage. Sampling the

Table 2: **Three-seed evaluations:** $\alpha \in \{0, 1\}$ and best-performing β configurations. Δ is vs DAPO for method rows and vs Base for the DAPO row.

Method	Held-out	Δ	$\Delta\%$	Wins
Base (pretrain)	0.201 ± 0.004	—	—	—
DAPO	0.245 ± 0.004	$+0.045$	$+22.4\%$	—
α=0	0.247 ± 0.005	$+0.001$	$+0.5\%$	3/5
α=0.5, $\beta = 0.1$	0.249 ± 0.015	$+0.004$	$+1.5\%$	3/5
α=1	0.259 ± 0.010	$+0.013$	$+5.4\%$	4/5
Asym.-RL/SD, β_S=0.1	0.261 ± 0.003	$+0.016$	$+6.5\%$	5/5

Table 3: **Five-seed evaluations:** α=1 and Asym.-RL/SD (β_S=0.1). Δ is vs DAPO for method rows and vs Base for the DAPO row; CI is a 95% seed-paired percentile bootstrap on Δ (10,000 resamples).

Method	Held-out	Δ [95% CI]	$\Delta\%$	Wins
Base (pretrain)	0.201 ± 0.004	—	—	—
DAPO	0.244 ± 0.004	$+0.043$	$+21.5\%$	—
α=1	0.248 ± 0.016	$+0.005$ $[-0.006, +0.018]$	$+1.9\%$	4/5
Asym.-RL/SD, β_S=0.1	0.256 ± 0.008	$+0.012$ $[+0.006, +0.016]$	$+5.0\%$	5/5

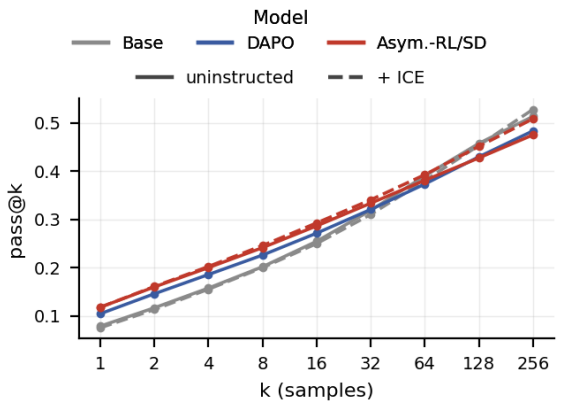

Figure 3: Held-out pass@k at 1.7B (macro-mean over AIME-24/25 and HMMT-Feb/Nov) for Base, DAPO, and Asymmetric-RL/SD, sampled uninstructed (solid) and with the behavioural instructions (+ICE, dashed). RL post-training leads at low k but trails the base model's coverage at high k, which +ICE recovers. Per-benchmark breakdown in Appendix G.

same checkpoint with the behavioural instructions (+ICE) recovers most of this coverage (pass@256 0.508); the instructions lift even the untrained base model (0.528 vs 0.513), indicating that they surface correct solutions independently of post-training.

7 Discussion

We have shown that RL on an instruction-conditioned teacher policy combined with forward-KL distillation to an unconditioned student improves mathematical reasoning over DAPO at 1.7B/4K, with 5/5 benchmark wins and a 95% CI excluding zero. The improvement over teacher-conditioned RL alone is smaller and less certain. A smaller gain persists at the longer-context 1.7B/8K setting (+3.5%, 3/5 wins), and at 4B/4K it does not improve on DAPO (−3.3%, 1/5 wins). Taken together, these results demonstrate that instruction-conditioned exploration, paired with an explicit self-distillation transfer mechanism, can improve a small model's held-out reasoning over the DAPO baseline. The method changes only training, adding a forward pass per update for the student but no extra rollouts. At test time the student is used exactly as a standard model, so its gains come at no extra inference cost, making it well suited to applications where small models are most useful, such as cost- or latency-constrained deployment.

The improvement from adding the forward-KL term to ICE with RL (Asymmetric-RL/SD vs $\alpha = 1$) suggests that parameter sharing alone is insufficient to incorporate teacher-sampled behaviour into the student policy; the explicit distillation signal contributes additional transfer. Another possible effect is that because student and teacher share parameters, the KL-driven updates also affect the teacher's subsequent rollouts, producing a compounding effect beyond what the RL objective alone can produce.

Beyond pass@1, the budgeted sampling in §6.4 ties these gains to exploration. RL post-training raises pass@1 but reduces high-k coverage relative to the base model, the *support shrinkage* of

Table 4: **Replication at 1.7B/8k and 4B/4k.** Held-out is the macro-mean pass@1 over the five benchmarks; Δ is vs DAPO for method rows and vs Base for the DAPO row. Per-seed values in Appendix F. †One 4B Asymmetric-RL/SD seed did not complete the full training schedule and is excluded (n=2).

	1.7B / 8k				4B / 4k			
Method	Held-out	Δ	Δ%	Wins	Held-out	Δ	Δ%	Wins
Base (pretrain)	0.202 ± 0.004	—	—	—	0.280 ± 0.003	—	—	—
DAPO	0.287 ± 0.018	$+0.084$	$+41.6\%$	—	0.361 ± 0.007	$+0.081$	$+28.9\%$	—
α=1	—	—	—	—	0.358 ± 0.014	-0.003	-0.9%	2/5
Asym.-RL/SD	0.297 ± 0.021	$+0.010$	$+3.5\%$	3/5	$0.349 \pm 0.022^{\dagger}$	-0.012	-3.3%	1/5

Table 5: Held-out pass@k at 1.7B over the four AIME/HMMT benchmarks (MATH-500 excluded), unbiased estimator (Chen et al., 2021), T=1.0, 256 samples from a single probe-best checkpoint. "+ICE" samples with the behavioural instructions; this T=1.0/four-benchmark setup is not directly comparable to Table 3.

Model	pass@1	pass@16	pass@256
Base	0.079	0.254	0.513
+ICE	0.075	0.250	0.528
DAPO	0.104	0.271	0.483
Asym.-RL/SD	0.118	0.286	0.475
+ICE	0.118	0.293	0.508

Wu et al. (2026); Asymmetric-RL/SD improves pass@1 over DAPO at comparable coverage. Sampling with the behavioural instructions at inference restores coverage to near the base model, realising the *support composition* the instructions induce and offsetting, on this proxy, the exploration that post-training removes.

At 4B/4K the gain seen at 1.7B does not re-emerge: ICE with Asymmetric-RL/SD does not improve over DAPO (-3.3%, 1/5 wins), even though the same mechanism delivers a clear gain at 1.7B. The point estimate is slightly negative, but with only two completed seeds that fall either side of the DAPO mean, n=2 is too few to tell a real effect from noise. One possible explanation is that instruction conditioning helps most on tasks at the edge of the model's capability, where exploratory strategies can reach correct solutions the unconditioned policy does not; the stronger 4B base leaves less of this headroom. Additionally, we test 4B only at the 4K response length, so its behaviour at longer contexts is unknown.

Because ICE's instructions are fixed and untargeted, a more targeted approach could improve sample efficiency, though Appendix H shows convergence is not slowed by including broad, potentially unhelpful instructions. This motivates generating bespoke, per-task strategies dynamically, as in the self-feedback methods of §2.2 such as Strategy-Guided Exploration (Szot et al., 2026), optionally drawing on privileged information such as the correct answer. Being closely related to π-distill, Asymmetric-RL/SD extends to such privileged-information settings in the same way.

8 Conclusions

We introduced instruction-conditioned exploration (ICE), a template-based context-mediated exploration method that conditions RL post-training on a fixed set of diverse behavioural instructions, and combined it with our Asymmetric-RL/SD off-policy self-distillation objective to transfer instruction-conditioned behaviour into an unconditioned test-time policy.

At 1.7B/4K, RL on the instruction-conditioned teacher combined with forward-KL distillation to the unconditioned student improves held-out pass@1 on mathematical reasoning tasks over DAPO with 5/5 benchmark wins and a 95% CI excluding zero. The gain holds across both the 4K and 8K context lengths at 1.7B (more weakly at 8K), suggesting it is not specific to a single setting, though it does not appear at 4B/4K. These results demonstrate that instruction-conditioning during training, combined with self-distillation, can improve the held-out reasoning of small language models.

We intend the 1.7B/8K and 4B/4K results as first steps toward characterising the approach across a range of small models; future work should build on them with additional seeds, model sizes, and context lengths to map where the gain holds. The success of this self-distillation configuration also motivates its application to other privileged-information methods.

9 Limitations

Statistics and evaluation. Seed counts are low: headline configurations use n=3–5 seeds and several preliminary variants are single-seed, so our strongest result (α=1, β_S=0.1, filtered) should be replicated at higher n before strong conclusions are drawn. These configurations were also chosen from a single-seed sweep, so the selection seed inflates their headline estimates, which are therefore mildly optimistic. The held-out benchmarks compound this: AIME-24/25 and HMMT-Feb/Nov-25 contain only 30 questions each and carry substantial pass@1 noise, and because these pairs measure overlapping skills the effective number of independent benchmarks is below five, so benchmark-win counts overstate significance when read alone. We therefore treat wins as one signal among several, alongside the mean improvement and its bootstrap interval, rather than as a standalone significance test. Held-out scores are reported at the step maximising a 200-question training-set probe; this is identical across methods, keeping comparisons fair, though absolute numbers may be optimistic. Finally, MATH-500 predates Qwen3's pretraining cut-off and was likely seen during pretraining; we retain it for comparability with prior work but it should be read with this in mind.

Baselines and scope. The only baseline is vanilla DAPO; we do not compare directly against other context-mediated exploration methods, in particular Prompt Augmentation (Lu et al., 2026), the closest in spirit to ICE, so gains are measured over an unconditioned RL baseline rather than alternative exploration approaches. The study is also narrow: a single domain (mathematical reasoning with binary verifiable rewards), one base-model family (Qwen3 with thinking disabled), two model scales and two context lengths, a sparse sweep of $(\alpha, \beta_T, \beta_S)$, and a fixed training budget. Generalisation to other domains, algorithms, and larger models therefore remains open.

Method. ICE requires a small set of describable behavioural instructions, supplied by an expert or a capable model, and samples them without task-specific targeting; adaptive, per-task alternatives are left to future work. Potential risks of the work are discussed in Appendix D.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, and 1 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint*. ArXiv:2404.14219 [cs].

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI Feedback.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2026. MathArena: Evaluating LLMs on Uncontaminated Math Competitions. *arXiv preprint*. ArXiv:2505.23281 [cs].

Aman Bhargava, Cameron Witkowski, Alexander Detkov, and Matt Thomson. 2024. Prompt Baking. *arXiv preprint*. ArXiv:2409.13697 [cs].

Justin Chih-Yao Chen, Becky Xiangyu Peng, Prafulla Kumar Choubey, Kung-Hsiang Huang, Jiaxin Zhang, Mohit Bansal, and Chien-Sheng Wu. 2025a. Nudging the Boundaries of LLM Reasoning. *arXiv preprint*. ArXiv:2509.25666 [cs] version: 1.

Mark Chen, Jerry Tworek, Heewoo Jun, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint*. ArXiv:2107.03374 [cs].

Zhipeng Chen, Xiaobo Qin, Youbin Wu, Yue Ling, Qinghao Ye, Wayne Xin Zhao, and Guang Shi. 2025b. Pass@k Training for Adaptively Balancing Exploration and Exploitation of Large Reasoning Models. *arXiv preprint*. ArXiv:2508.10751 [cs].

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models. *arXiv preprint*. ArXiv:2505.22617 [cs].

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, and 175 others. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638.

Andre Wang He, Daniel Fried, and Sean Welleck. 2025. Rewarding the Unlikely: Lifting GRPO Beyond Distribution Sharpening. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25548–25560, Suzhou, China. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv preprint*. ArXiv:2103.03874 [cs].

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint*. ArXiv:1503.02531 [stat].

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Zilin Zhu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Weikai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. 2024. OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework. *arXiv preprint*.

Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. 2026. Diversity-Incentivized Exploration for Versatile Reasoning. *arXiv preprint*. ArXiv:2509.26209 [cs].

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. 2026. Reinforcement Learning via Self-Distillation. *arXiv preprint*. ArXiv:2601.20802 [cs].

Abdurrahman Javat and Allan Kazakov. 2026. Silicon Showdown: Performance, Efficiency, and Ecosystem Barriers in Consumer-Grade LLM Inference. *arXiv preprint*. ArXiv:2605.00519 [cs.PF].

Daniel R. Jiang, Alex Nikulkov, Yu-Chia Chen, Yang Bai, and Zheqing Zhu. 2025. Improving Generative Ad Text on Facebook using Reinforcement Learning. *arXiv preprint*. ArXiv:2507.21983 [cs].

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 39648–39677, Red Hook, NY, USA. Curran Associates Inc.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D. Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. 2024. Training Language Models to Self-Correct via Reinforcement Learning. *arXiv preprint*. ArXiv:2409.12917 [cs].

Ang Li, Yifei Wang, Zhihang Yuan, Stefanie Jegelka, and Yisen Wang. 2025. LANPO: Bootstrapping Language and Numerical Feedback for Reinforcement Learning in LLMs. *arXiv preprint*. ArXiv:2510.16552 [cs].

Zhenwen Liang, Sidi Lu, Wenhao Yu, Kishan Panaganti, Yujun Zhou, Haitao Mi, and Dong Yu. 2025. Can LLMs Guide Their Own Exploration? Gradient-Guided Reinforcement Learning for LLM Reasoning. *arXiv preprint*. ArXiv:2512.15687 [cs].

Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, and Jiang Bian. 2026. Self-Hinting Language Models Enhance Reinforcement Learning. *arXiv preprint*. ArXiv:2602.03143 [cs].

Wenquan Lu, Hai Huang, and Randall Balestriero. 2026. Prompt Augmentation Scales up GRPO Training on Mathematical Reasoning. *arXiv preprint*. ArXiv:2602.03190 [cs].

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. *arXiv preprint*. ArXiv:2303.17651 [cs].

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. 2026. Privileged Information Distillation for Language Models. *arXiv preprint*. ArXiv:2602.04942 [cs].

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. 2025. WebRL: Training LLM Web Agents via Self-Evolving Online Curriculum Reinforcement Learning. *International Conference on Learning Representations*, 2025:79791–79821.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint*. ArXiv:2402.03300 [cs].

Han Shen. 2025. On Entropy Control in LLM-RL Algorithms. *arXiv preprint*. ArXiv:2509.03493 [cs].

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-Distillation Enables Continual Learning. *arXiv preprint*. ArXiv:2601.19897 [cs].

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. *arXiv preprint*. ArXiv:2303.11366 [cs].

Yuda Song, Julia Kempe, and Remi Munos. 2025. Outcome-based Exploration for LLM Reasoning. *arXiv preprint*. ArXiv:2509.06941 [cs].

Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA.

Andrew Szot, Michael Kirchhof, Omar Attia, and Alexander Toshev. 2026. Expanding LLM Agent Boundaries with Strategy-Guided Exploration. *arXiv preprint*. ArXiv:2603.02045 [cs].

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. *arXiv preprint*. ArXiv:2506.01939 [cs].

Zihan Wang, Chi Gui, Xing Jin, Qineng Wang, Licheng Liu, Kangrui Wang, Shiqi Chen, Linjie Li, Zhengyuan Yang, Pingyue Zhang, Yiping Lu, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. 2026. RAGEN-2: Reasoning Collapse in Agentic RL. *arXiv preprint*. ArXiv:2604.06268 [cs] version: 1.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. 2026. The Invisible Leash: Why RLVR May or May Not Escape Its Origin. *arXiv preprint*. ArXiv:2507.14843 [cs].

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, and 41 others. 2025. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. 2026. Learning beyond Teacher: Generalized On-Policy Distillation with Reward Extrapolation. *arXiv preprint*. ArXiv:2602.12125 [cs].

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint*. ArXiv:2503.14476 [cs].

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model? *arXiv preprint*. ArXiv:2504.13837 [cs].

Yunpeng Zhai, Shuchang Tao, Cheng Chen, Anni Zou, Ziqian Chen, Qingxu Fu, Shinji Mai, Li Yu, Jiaji Deng, Zouying Cao, Zhaoyang Liu, Bolin Ding, and Jingren Zhou. 2025. AgentEvolver: Towards Efficient Self-Evolving Agent System. *arXiv preprint*. ArXiv:2511.10395 [cs].

Charlie Zhang, Graham Neubig, and Xiang Yue. 2025. On the Interplay of Pre-Training, Mid-Training, and RL on Reasoning Language Models. *arXiv preprint*. ArXiv:2512.07783 [cs].

Junshuo Zhang. 2026. DPEPO: Diverse Parallel Exploration Policy Optimization for LLM-based Agents. ArXiv:2604.24320.

Xichen Zhang, Sitong Wu, Yinghao Zhu, Haoru Tan, Shaozuo Yu, Ziyi He, and Jiaya Jia. 2026. Scaf-GRPO: Scaffolded Group Relative Policy Optimization for Enhancing LLM Reasoning. *arXiv preprint*. ArXiv:2510.19807 [cs].

Yifan Zhang and Team Math-AI. 2024. American Invitational Mathematics Examination (AIME) 2024.

Yifan Zhang and Team Math-AI. 2025. American Invitational Mathematics Examination (AIME) 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. *arXiv preprint*. ArXiv:2601.18734 [cs].

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, Hengtong Lu, Wei Chen, Yan Xie, and Mingli Song. 2026. Breaking the Exploration Bottleneck: Rubric-Scaffolded Reinforcement Learning for General LLM Reasoning. *arXiv preprint*. ArXiv:2508.16949 [cs].

A Supplementary Instructions

We design an automated and replicable process for selecting behavioural instructions. Claude Sonnet 4.6 solves a 250-problem subset of the training set, with Claude Opus 4.6 retrying any problems it answers incorrectly. The correct solutions are categorised and synthesised into 10 candidate strategies, each with 3 instruction-text variants. The model to be post-trained (Qwen3 4B) then attempts a separate held-out 250-problem subset under every strategy and variant, scored by pass@16. We select the 5-strategy set by greedy set cover on pass@16 coverage: at each step we add the strategy that solves the most as-yet-unsolved problems, breaking ties by per-sample success rate. One variant per selected strategy is used as its instruction, yielding the $N = 5$ instructions in Table 6.

B Compute

Training was conducted on an HPC with NVIDIA GH200 superchips. Each 1.7B configuration used 2 GPUs requiring approximately 100-250 GPU-hours per seed. Each 4B configuration used 4 GPUs requiring approximately 200-250 GPU-hours per seed. The experiments reported in this work total approximately 6,500 GPU-hours. Including all preliminary and exploratory work (pilots, infrastructure development, and runs not included in the final results), total project compute amounted to roughly 10,000 GPU-hours.

ID	Instruction text
work_backwards	"Invert the process: ask what input would produce this output, rather than computing forward."
constructive_building	"After proving a bound, construct a concrete configuration that achieves it to confirm tightness."
symmetry_and_invariance	"Ask: what symmetry does this problem have, and how can it reduce the number of cases or simplify the expression?"
decomposition	"Identify which components of the problem can be handled independently and tackle them one at a time."
reduction	"Reformulate the problem in a different domain where the answer becomes obvious or follows from a standard result."

Table 6: The set of behavioural instructions $\mathcal{I}$ used during training.

C Artifact Licenses

All artifacts used in this work are openly released and were used for research purposes consistent with their licenses. The Qwen3 models (Yang et al., 2025) and the DAPO-Math-17K dataset (Yu et al., 2025) are released under the Apache-2.0 license, as is the OpenRLHF framework used for training. For evaluation, MATH-500 (a subset of the MATH dataset (Hendrycks et al., 2021) distributed through OpenAI's PRM800K release) is under the MIT license, the AIME-24/25 sets (Zhang and Math-AI, 2024, 2025) are under Apache-2.0, and the HMMT-Feb/Nov-25 sets obtained via MathArena (Balunović et al., 2026) are under CC BY-NC-SA 4.0. Our use is non-commercial academic research, consistent with these terms. The behavioural instructions and source code produced are released for research use only.

D Potential Risks

Our method improves the mathematical reasoning of small language models through a training-time-only intervention, and we foresee limited direct risk. As with any technique that raises the capability of compact, cheaply deployable models, the gains are broadly dual-use: more capable small models lower the barrier to both beneficial and potentially harmful downstream applications. Our experiments are confined to mathematical reasoning with binary verifiable rewards, a domain with little scope for harmful content, and we release only behavioural instructions and model checkpoints intended for research use. We introduce no new data collection involving people, and the datasets used contain no personal or sensitive information.

E Preliminary Single-Seed Sweep

We first run a broad preliminary sweep with a single seed per configuration, then repeat the most promising with additional seeds.

Three configurations show promise in the single-seed sweep at 1.7B/4K (Table 7): $\alpha = 0.5, \beta = 0.1$ ($+0.024$, 5/5 wins), $\alpha = 1$ ($+0.029$, 5/5), and Asymmetric-RL/SD $\beta_S = 0.1$ ($+0.017$, 3/5). The $\alpha \in \{0, 0.5\}$ configurations without KL regularisation give no improvement over DAPO; these configurations, which include a student reward term, additionally exhibit response-length blow-up and increased truncation rates during training (Appendix I). Setting Asymmetric-RL/SD β_S to 0.5 degrades performance considerably (-0.034, 0/5) relative to $\beta_S = 0.1$, indicating that the forward-KL weight requires careful tuning: $\beta_S = 0.1$ is the best option tested.

Filtering on correctness is critical for the $\alpha = 1, \beta_S > 0$ setting: the unfiltered variant decreases mean pass@1 by -0.031 (-12.9%) relative to the baseline, with worse performance on all 5 benchmarks. Including the forward-KL objective trains the student to imitate the teacher; incorrect rollouts therefore yield harmful student updates.

F Per-Seed Held-out Results

Table 8 reports the per-seed held-out averages underlying the replication table (§6.3, Table 4), under probe-best selection. The two completed 4B/4k Asymmetric-RL/SD seeds fall either side of DAPO (0.333 and 0.365); the third did not complete the full training schedule and is excluded (n=2).

G Per-Benchmark pass@k

Figure 4 breaks the held-out pass@k curves of Figure 3 down by benchmark, and additionally shows the DAPO-200 training probe (excluded from the held-out macro-mean).

Table 7: **Preliminary results, single-seed.** Δ is vs DAPO for method rows and vs Base for the DAPO row.

Method	Held-out	Δ	$\Delta\%$	Wins
Base (pretrain)	0.201 ± 0.004	—	—	—
DAPO	0.241	+0.041	+20.4%	—
α=0	0.244	+0.003	+1.1%	3/5
α=0.5	0.238	−0.004	−1.5%	2/5
α=0.5, β=0.05	0.242	+0.001	+0.2%	2/5
α=0.5, β=0.1	0.266	+0.024	+10.1%	5/5
α=1	0.270	+0.029	+12.0%	5/5
α=1, β_S=0.1 (unfilt.)	0.210	−0.031	−12.9%	0/5
Asym.-RL/SD, β_S=0.05	0.243	+0.001	+0.5%	4/5
Asym.-RL/SD, β_S=0.1	0.258	+0.017	+7.0%	3/5
Asym.-RL/SD, β_S=0.5	0.208	−0.034	−14.0%	0/5

Table 8: Per-seed held-out pass@1 (macro-mean over MATH-500, AIME 24/25, HMMT Feb/Nov), probe-best selection. "DAPO" is the RL-only baseline.

Method	Seed 1	Seed 2	Seed 3	Mean±std
1.7B / 8k				
DAPO	0.308	0.276	0.277	0.287 ± 0.018
Asym.-RL/SD	0.304	0.312	0.273	0.297 ± 0.021
4B / 4k				
DAPO	0.370	0.357	0.357	0.361 ± 0.007
α=1	0.373	0.344	0.357	0.358 ± 0.014
Asym.-RL/SD	0.333	0.365	—	0.349 ± 0.022

H Instruction Conditioning Does Not Slow Learning

Despite replacing some neutral prompts with potentially unhelpful instruction-conditioned ones, instruction conditioning does not slow convergence: the $\alpha = 1$ probe pass@1 tracks DAPO throughout training, pulling clearly ahead only late in training (around 250–300 steps; Figure 2). Asymmetric-RL/SD (β_S=0.1) lags at early steps but catches up to $\alpha = 1$ by step 200.

I Training Dynamics of the α-Sweep

Configurations with $\alpha \in \{0, 0.5\}$, which include a student reward term with importance-sampling correction, exhibit response length blow-up and increased truncation rates during training. The truncation rate for both DAPO and $\alpha = 1$ remains low throughout training (Figure 5), ending at 0.023 and 0.020 respectively (Table 9). In contrast $\alpha = 0$ and $\alpha = 0.5$ increase to 0.616 and 0.093 respectively at step 400. In particular, the response length of $\alpha = 0$ reaches 3033 tokens, 1.40 times that of DAPO (2160 tokens).

Table 9: Endpoint summary for the α-sweep at 1.7B / 4k (DAPO-200 at step 400, response length (number of tokens) and truncation rate are the mean of steps 376-400). Each row is from a single training seed.

Method	DAPO-200	Resp. length	Truncation rate
DAPO	0.415	2160	0.023
α=0	0.305	3033	0.616
α=0.5	0.445	2231	0.093
α=1	0.390	2114	0.020

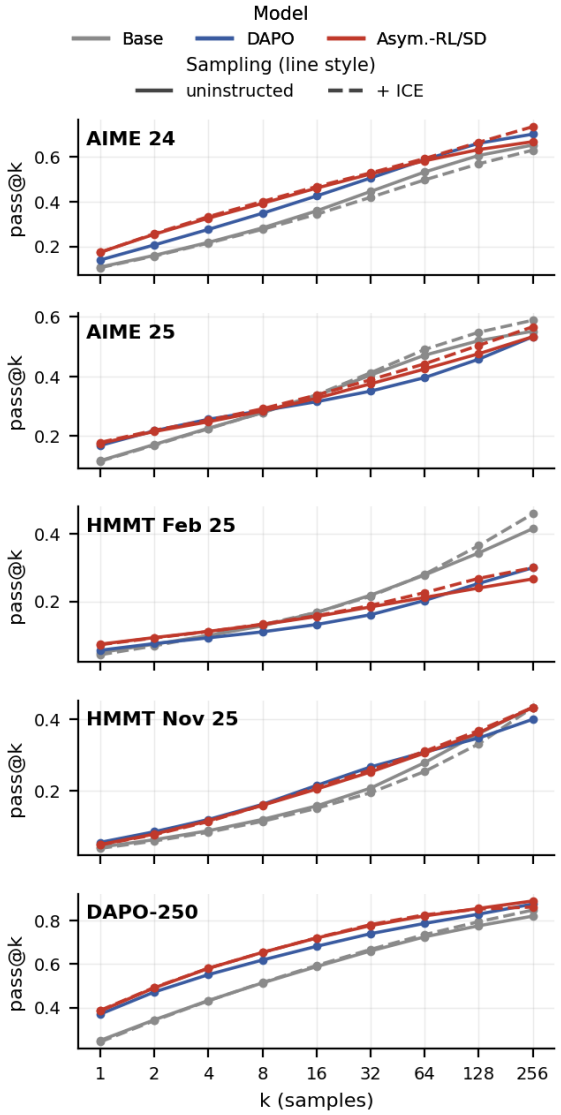

Figure 4: Per-benchmark pass@k at 1.7B for the held-out benchmarks (AIME-24/25, HMMT-Feb/Nov) and the DAPO-200 training probe, for Base, DAPO, and Asymmetric-RL/SD, sampled uninstructed (solid) and with the behavioural instructions (+ICE, dashed).

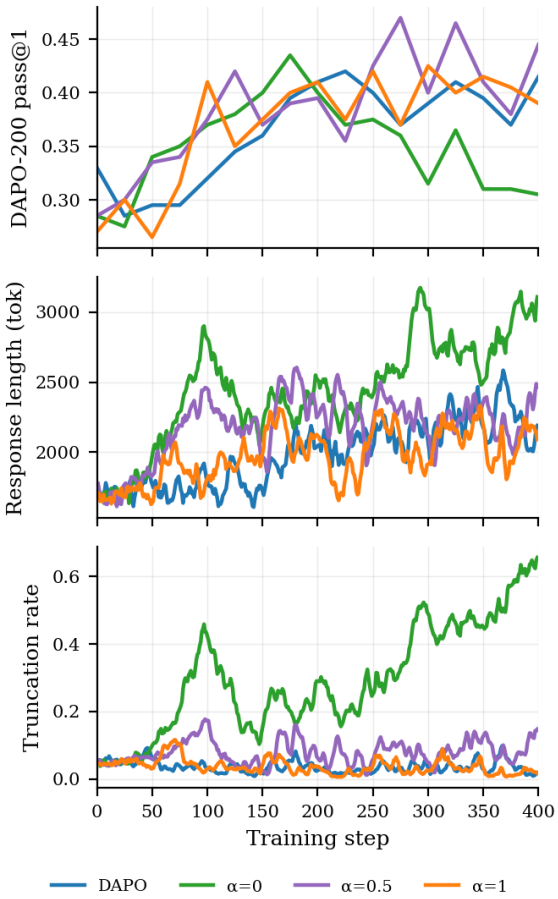

Figure 5: α-sweep training dynamics at 1.7B / 4k from a single training seed per method.