DISTILLED REINFORCEMENT LEARNING FOR LLM POST-TRAINING

Chen Wang[1,2]*, **Zhaochun Li**[2,3], **Jionghao Bai**[2,4], **Yining Zhang**[2,5], **Hexuan Deng**[2,6], **Ge Lan**[7]†, **Yue Wang**[2]†

[1] College of Elite Engineers, Nankai University, [2] Zhongguancun Academy
[3] Beijing Institute of Technology, [4] Zhejiang University
[5] Institute of Automation, Chinese Academy of Sciences
[6] Harbin Institute of Technology, [7] College of Software, Nankai University

ABSTRACT

Large language model (LLM) post-training is essential for improving reasoning, adaptation, and alignment. Existing methods mainly follow two paradigms: reinforcement learning (RL) and on-policy distillation (OPD). However, RL relies on coarse-grained outcome supervision, resulting in difficult credit assignment and limited capability to acquire new knowledge. OPD, meanwhile, unconditionally matches teacher logits through KL divergence, which creates a dilemma: similar teachers provide little new knowledge, while substantially different teachers often yield ineffective guidance, largely restricting OPD to within-family distillation. We propose Distilled Reinforcement Learning (Distilled RL), which integrates teacher supervision into the RL objective to provide fine-grained guidance, selectively transfer new knowledge and avoid unconditional imitation. Distilled RL contains three components: reverse importance sampling with clipping, negative sample reset, and sequence-level geometric normalization. Through a concise and interpretable case study, we demonstrate that Distilled RL can effectively transfer previously unavailable knowledge from a teacher model to a student model. Extensive experiments across both within-family and cross-family distillation settings show that Distilled RL substantially outperforms standard RL and OPD in terms of both pass@1 and pass@k. Our code is available at https://github.com/597358816/Distilled-RL.

1 INTRODUCTION

Large language model (LLM) post-training plays a central role in improving reasoning, instruction following, and task adaptation (Schulman et al., 2017; Wei et al., 2022; Touvron et al., 2023; GLM et al., 2024; Rafailov et al., 2023; Zhong et al., 2024; Wang et al., 2024b; Team et al., 2025). Compared with supervised fine-tuning (SFT), reinforcement learning (RL) performs updates on trajectories sampled from the current policy, thereby reducing the distribution shift between training and inference and alleviating catastrophic forgetting (Shenfeld et al.; Chen et al., 2025). This on-policy training paradigm has enabled substantial advances in complex reasoning and generalization (Shao et al., 2024; Liu et al., 2024; Guo et al., 2024; Yu et al., 2025). Inspired by the same principle, on-policy distillation (OPD) reformulates conventional knowledge distillation by replacing teacher- or dataset-generated samples with trajectories sampled from the student policy, and then minimizes the KL divergence between the student and teacher distributions on the states actually visited by the student (Gu et al., 2024; Agarwal et al., 2024). In this way, OPD reduces the mismatch between distillation data and the student's inference-time distribution while retaining dense token-level supervision from a stronger teacher (Yang et al., 2026; Wu et al., 2026).

Despite the effectiveness of RL and OPD, both paradigms exhibit important limitations. RL usually assigns a sequence-level reward to an entire response, even though only a small subset of tokens may

*s-wc25@bza.edu.cn
†Corresponding author

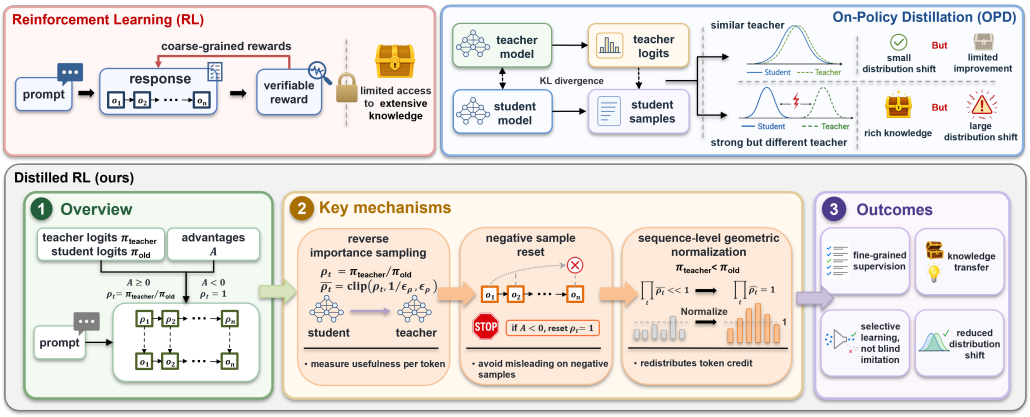

Figure 1: Overview of Distilled RL.

determine whether the final answer is correct (Shao et al., 2024). Such coarse-grained supervision creates a severe credit-assignment problem and provides limited information about which intermediate reasoning steps should be encouraged or corrected (Lightman et al., 2023). Moreover, because the learning signal is derived mainly from task rewards, RL is effective at reinforcing behaviors already accessible to the student but has limited ability to introduce knowledge that is absent from its current policy (Yu et al., 2025; Yue et al., 2025). OPD provides denser token-level feedback, but commonly optimizes a KL-divergence objective that unconditionally encourages the student to imitate the teacher distribution (Wu et al., 2025a; Xu et al., 2025; Jin et al., 2026). This leads to a fundamental dilemma. When the teacher and student are highly similar, their distributions offer limited complementary knowledge; when they differ substantially in architecture, model family, or reasoning pattern, the resulting distribution mismatch makes token-level imitation noisy and difficult to optimize. Recent studies have similarly observed that successful OPD depends strongly on compatible teacher–student reasoning patterns and distributions (Wu et al., 2026; Fu et al., 2026; Li et al., 2026). Beyond this compatibility issue, KL-based imitation may also drive the student to align with the teacher distribution too rapidly, leading to premature convergence and leaving limited room for subsequent reward-driven improvement (see detail in Subsection 3.1).

To address these limitations, we propose **Distilled Reinforcement Learning (Distilled RL)**, a unified post-training framework that incorporates teacher knowledge directly into the RL gradient (see overview in Fig. 1). Distilled RL evaluates student-generated tokens using a reverse importance ratio between the teacher policy and the old student policy, and uses this ratio to redistribute the RL learning signal at the token level. The framework contains three key components. First, **reverse importance sampling** measures the relative preference of the teacher for each student-generated token and clips the ratio to prevent unstable gradient amplification. Second, **negative sample reset** applies teacher-based reweighting only to responses with positive advantages, while resetting the importance weights of negative samples to one. This prevents the student from imitating teacher preferences along negative trajectories and preserves the original RL penalty for negative responses. Third, **sequence-level geometric normalization** normalizes the clipped token-level ratios within each response such that their geometric mean equals one. This normalization respects the multiplicative factorization of autoregressive sequence probabilities, removes systematic scale differences between teacher and student likelihoods, and retains the teacher's relative preference over tokens. Through these mechanisms, Distilled RL uses the teacher as a selective fine-grained guide rather than an unconditional imitation target. We further demonstrate through a controlled case study that Distilled RL enables the student to acquire new knowledge from the teacher that standard RL cannot effectively learn. Extensive experiments show that Distilled RL consistently outperforms RL and OPD, and, more importantly, achieves substantial gains in cross-family distillation, where conventional OPD suffers from persistent performance degradation.

Our contributions are threefold.

- We propose Distilled RL, which integrates teacher supervision directly into the RL objective through three components: reverse importance sampling, negative sample reset, and sequence-level geometric normalization.
- Through a controlled case study, we demonstrate that Distilled RL enables the student model to acquire new knowledge from the teacher that cannot be learned through standard RL alone.
- Extensive experiments on both within-family and cross-family distillation show that Distilled RL consistently outperforms RL and OPD in terms of both pass@1 and pass@k.

2 Related Work

Reinforcement Learning for LLMs. Reinforcement learning has become a central approach for improving the reasoning capabilities of large language models. Early RLHF methods optimize human-preference reward models using policy-gradient algorithms such as PPO (Ouyang et al., 2022; Schulman et al., 2017), while recent reinforcement learning with verifiable rewards directly evaluates responses using rule-based outcome signals. Given a prompt q and a response o sampled from the old policy, a general policy-optimization objective can be written as

$$\mathcal{J}_{\mathrm{RL}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q),\\ o\sim\pi_{\theta_{\mathrm{old}}}(\cdot|q)}} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} r_{i,t}(\theta) A(q,o) \right], \qquad r_{i,t}(\theta) = \frac{\pi_\theta\left(o_{i,t} \mid q, o_{i,<t}\right)}{\pi_{\theta_{\mathrm{old}}}\left(o_{i,t} \mid q, o_{i,<t}\right)}. \tag{1}$$

GRPO removes the need for a separate critic by estimating $A(q,o)$ from the relative rewards of a group of sampled responses (Shao et al., 2024), and subsequent work demonstrates that large-scale RL can elicit sophisticated reasoning behaviors from pretrained language models (Guo et al., 2025b). Compared with supervised fine-tuning, on-policy RL trains on trajectories generated by the evolving policy, reducing the mismatch between training and inference distributions. Nevertheless, the same sequence-level advantage A is generally assigned to every token in the response, even though different tokens contribute unequally to the final outcome. This coarse-grained supervision creates a challenging credit-assignment problem. Recent RL research has primarily focused on improving exploration through entropy (Cui et al., 2025b; Cheng et al., 2025; Wang et al., 2026b; 2025), mitigating overthinking by controlling response length (Yi et al., 2025; Liu et al., 2026; Wang et al., 2026a), and providing finer-grained credit assignment via process-level rewards (Luo et al., 2024; Cui et al., 2025a; Li et al., 2025; Cheng et al., 2026). However, these methods largely optimize how existing behaviors are reinforced, rather than how new knowledge is acquired. Efficiently transferring knowledge beyond the student's current capability still requires guidance from a stronger teacher model, which is not explicitly addressed by conventional RL.

On-Policy Distillation. Knowledge distillation transfers capabilities from a stronger teacher to a smaller student by matching teacher-generated outputs or predictive distributions (Hinton et al., 2015; Kim & Rush, 2016). Conventional sequence-level distillation is usually performed on fixed teacher-generated data and therefore suffers from exposure bias when the student encounters states not covered by the distillation corpus. On-policy distillation instead samples a response o from the student policy and minimizes the token-level discrepancy between the student and teacher distributions on the states visited by the student:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q),\\ o\sim\pi_\theta(\cdot|q)}} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid q, o_{<t}) \,\|\, \pi_{\text{teacher}}(\cdot \mid q, o_{<t})\right) \right]. \tag{2}$$

MiniLLM adopts reverse KL divergence and optimizes it using student-generated samples (Gu et al., 2024), while generalized knowledge distillation further formalizes token-level teacher supervision along student-generated trajectories (Agarwal et al., 2024). Subsequent studies have explored alternative KL objectives, uncertainty-aware distillation, and combinations of knowledge distillation and reinforcement learning (Wu et al., 2025a; Xu et al., 2025; Jin et al., 2026; Wu et al., 2026; Fu et al., 2026; Li et al., 2026). However, these methods generally retain an explicit KL-based imitation objective that encourages the student to match the teacher distribution at every visited state, regardless of whether the sampled trajectory is beneficial. Such unconditional imitation can become unreliable when the teacher and student belong to different model families and exhibit substantial distributional mismatch.

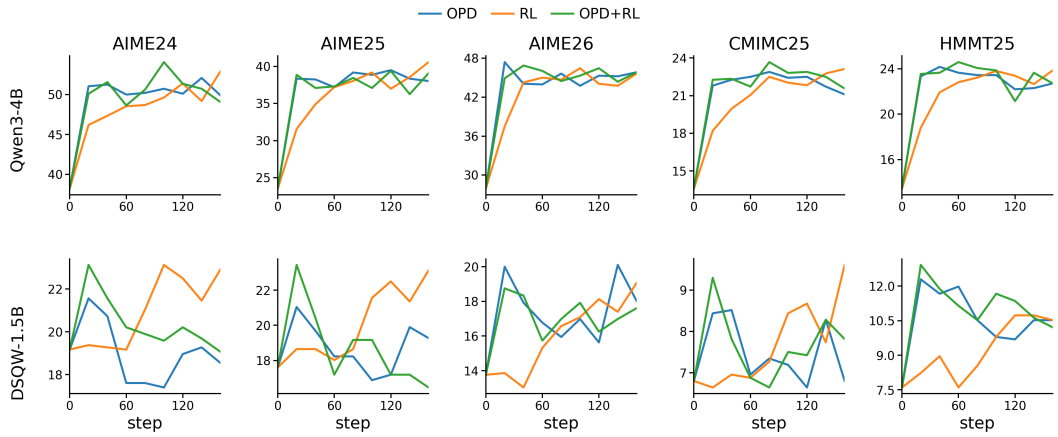

Figure 2: Training dynamics of OPD, RL, and OPD+RL on five competition reasoning benchmarks. The teacher model is Qwen3-8B-GRPO, and the student models are Qwen3-4B and DeepSeek-R1-Distill-Qwen-1.5B (DSQW-1.5B). Models are trained on DAPO-17k with rollout group size $G = 8$.

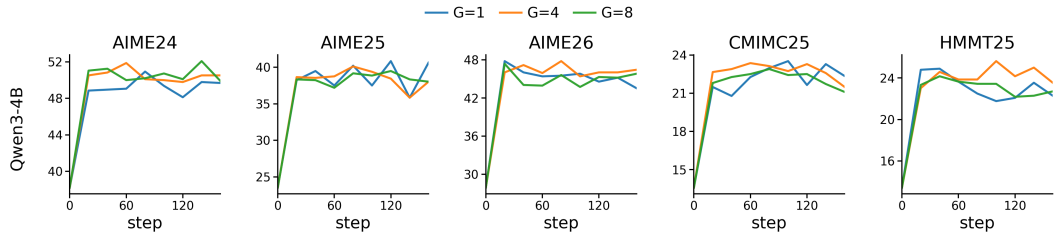

Figure 3: Training dynamics of OPD with different rollout group sizes. The teacher model is Qwen3-8B-GRPO, and the student model is Qwen3-4B.

3 METHOD

3.1 OBSERVATION AND ANALYSIS

Before presenting the formal objective of Distilled RL, we compare OPD and RL on multiple reasoning benchmarks. Figures 2 and 3 show the training dynamics under different base models and rollout group sizes.

OPD converges rapidly but lacks sustained improvement. As shown in Figure 2, OPD improves the student model quickly at the early stage of training, but the performance soon saturates and may even decrease, especially in cross-family distillation. This suggests that KL-based imitation can rapidly align the student with the teacher on visited states, but provides limited task-aware signals for further improvement. In contrast, RL shows a slower but more sustained upward trend, since it directly optimizes task rewards through exploration. However, its sequence-level reward provides coarse-grained supervision, making learning inefficient. A natural remedy is to mix the OPD and RL losses (such as 1:1), but this strategy still exhibits a consistent performance decline in the later training stage. This may be because OPD converges rapidly and constrains the student around a KL-induced local optimum, making it difficult for the RL signal to drive sustained improvement afterward.

Increasing the rollout size brings limited gains to OPD. Figure 3 studies OPD with different rollout group sizes $G \in \{1, 4, 8\}$. Increasing the number of rollouts is a common variance-reduction strategy in on-policy training, as it provides a more stable estimate of relative performance. However, OPD shows similar performance across different G values, and larger rollout groups do not lead to reliable or sustained gains. This indicates that the main bottleneck of OPD is not sampling variance,

but its KL-based imitation objective. OPD quickly reaches a local optimum determined by teacher–student distribution matching, and additional on-policy samples are insufficient to drive continuous improvement.

These observations suggest that teacher supervision should not be introduced as an independent unconditional imitation loss. Instead, it should be coupled with the reward-driven RL objective, so that the teacher provides fine-grained token-level guidance only when the sampled response is beneficial. This motivates Distilled RL, which incorporates teacher preference directly into the RL gradient through selective and normalized importance weighting.

3.2 DISTILLED RL

Motivated by these observations, we propose **Distilled Reinforcement Learning (Distilled RL)**, which integrates teacher guidance directly into the RL objective instead of using an independent KL-based imitation loss. Given responses sampled from the old student policy, the teacher evaluates the likelihood of each student-generated token, and the resulting token-level importance weights redistribute the RL learning signal. The resulting policy optimization objective is defined as follows:

$$\mathcal{J}_{\text{Distilled RL}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q),\\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\left(r_{i,t}(\theta)\, w_{i,t} A_i,\ \text{clip}(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}})\, w_{i,t} A_i \right) \right]. \tag{3}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta\left(o_{i,t} \mid q, o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid q, o_{i,<t}\right)}, \qquad A_i = \frac{R_i - \text{mean}\left(\{R_j\}_{j=1}^{G}\right)}{\text{std}\left(\{R_j\}_{j=1}^{G}\right)}, \qquad \rho_{i,t} = \frac{\pi_{\text{teacher}}\left(o_{i,t} \mid q, o_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(o_{i,t} \mid q, o_{i,<t}\right)},$$

$$\bar{\rho}_{i,t} = \text{clip}\left(\rho_{i,t}, \frac{1}{\epsilon_\rho}, \epsilon_\rho\right), \qquad \widetilde{\rho}_{i,t} = \frac{\bar{\rho}_{i,t}}{\exp\left(\frac{1}{|o_i|}\sum_{s=1}^{|o_i|}\log\bar{\rho}_{i,s}\right)}, \qquad w_{i,t} = \begin{cases} \widetilde{\rho}_{i,t}, & A_i > 0, \\ 1, & A_i \le 0, \end{cases}.$$

Here, π_{teacher} denotes the teacher policy, $\rho_{i,t}$ is the token-level teacher-to-student importance ratio, $\bar{\rho}_{i,t}$ is the clipped importance ratio, and $\widetilde{\rho}_{i,t}$ is the sequence-level geometrically normalized importance weight. The effective weight $w_{i,t}$ applies the normalized ratio only to positive samples, while resetting the weight of negative samples to one. Distilled RL consists of three core components:

Reverse importance sampling. Importance sampling estimates expectations under a target distribution using samples from another distribution. In autoregressive language-model training, token-level ratios are commonly used as a practical approximation to sequence-level importance sampling, as in PPO, GRPO, and DFT (Schulman et al., 2017; Guo et al., 2025a; Wu et al., 2025b). These methods typically perform off-policy correction by reweighting samples generated by a behavior policy toward the current policy. In Distilled RL, we reverse this direction: rather than correcting a behavior policy toward the current student policy, we reweight trajectories sampled from the current student policy toward the teacher policy. Accordingly, the teacher-to-student ratio $\rho_{i,t}$ measures how much more or less the teacher prefers each student-generated token relative to the old student policy. Although the outer policy-optimization objective already clips the student policy ratio $r_{i,t}(\theta)$, we additionally clip $\rho_{i,t}$ into $[1/\epsilon_\rho, \epsilon_\rho]$ as a redundant stability safeguard against extreme teacher–student likelihood ratios.

Negative sample reset. A key issue of directly applying teacher-based weights to all samples is that the effect of ρ changes with the sign of the advantage. As shown in Table 1, when $A_i > 0$, teacher reweighting behaves as desired: teacher-preferred tokens with $\rho_{i,t} > 1$ receive a stronger reward, while less preferred tokens with $\rho_{i,t} < 1$ are relatively suppressed. This provides fine-grained guidance and enables knowledge transfer on successful trajectories. The problem appears when

Table 1: Effect of importance weighting.

	$\rho_{i,t} > 1$	$\rho_{i,t} < 1$
$A_i > 0$	↑	↓
$A_i < 0$	↓	↑

$A_i < 0$. In this case, a teacher-preferred token receives a larger penalty, pushing the student away from the teacher. If the teacher can solve the problem, the student should imitate rather than avoid the teacher's preference; if the teacher also fails, the distillation signal is not useful. Therefore, for negative-advantage samples, we reset the importance weight to one and reduce the update to the original RL objective. This prevents teacher guidance from acting in the wrong direction and makes distillation selective.

Sequence-level geometric normalization. In on-policy training, responses are sampled from the student policy rather than the teacher policy. Therefore, many sampled tokens may have much higher probability under the student than under the teacher, making $\rho_{i,t}$ frequently smaller than one. Because autoregressive probabilities are multiplicative, directly applying token-level ratios can make $\prod_{t=1}^{|o_i|} \rho_{i,t}$ much smaller than the unweighted case, even for positive responses, thereby over-suppressing successful trajectories. Distilled RL addresses this with sequence-level geometric normalization $\widetilde{\rho}_{i,t} = \bar{\rho}_{i,t} / \exp(\frac{1}{|o_i|} \sum_{s=1}^{|o_i|} \log \bar{\rho}_{i,s})$: each clipped ratio is divided by the geometric mean of all clipped ratios in the same response, ensuring $\left(\prod_{t=1}^{|o_i|} \widetilde{\rho}_{i,t}\right)^{1/|o_i|} = 1$. Thus, teacher guidance does not globally amplify or suppress the whole response, but only redistributes the learning signal across tokens according to the teacher's relative preferences.

4 Interpretable Teacher-Side Information Transfer

Although OPD can transfer teacher knowledge through explicit distribution matching, whether policy-gradient-based RL can effectively incorporate information from the teacher remains unclear. To examine whether Distilled RL can acquire teacher-side information beyond task rewards, we design a controlled and interpretable case study based on entropy. A necessary condition for learning new knowledge from the teacher is the ability to capture certain properties of the teacher's distribution. Entropy is a simple and measurable property: a high-temperature teacher induces a softer distribution with higher entropy, while a low-temperature teacher induces a sharper distribution with lower entropy.

Instead of using a larger model as the teacher, we construct a temperature-controlled teacher from the student model itself. Specifically, when the student entropy is larger than 0.5, we use a low-temperature teacher with temperature 0.8 to provide a sharper distributional signal; otherwise, we use a high-temperature teacher with temperature 1.2 to provide a softer distributional signal. This setting removes confounding factors such as model scale and architecture, allowing us to directly test whether Distilled RL can follow a prescribed distributional property.

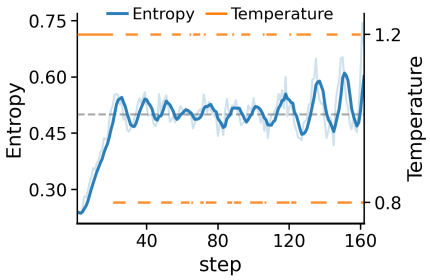

Figure 4: Entropy trajectory of Distilled RL with a temperature-controlled teacher.

As shown in Figure 4, the student entropy rapidly moves from a low initial value toward the target region and then fluctuates around the threshold. When the entropy becomes too high, the low-temperature teacher suppresses excessive uncertainty; when the entropy becomes too low, the high-temperature teacher encourages a softer distribution. This result shows that Distilled RL can capture and transfer the entropy characteristic of the teacher distribution, providing a minimal demonstration that it can acquire information beyond the original RL reward signal.

5 Experiments

To validate Distilled RL, we implement it within the EasyR1 and VeRL frameworks (Yaowei et al., 2025; Sheng et al., 2025) and compare it against representative baselines, including OPD, RL (GRPO) (Shao et al., 2024), OPD+RL. Experiments are conducted on DeepSeek-R1-Distill-Qwen-1.5B (DSQW-1.5B), Qwen3-4B, Qwen3-1.7B(Yang et al., 2024; 2025; Guo et al., 2025b), using DAPO-17K (Yu et al., 2025) for training. Evaluation covers both mathematical reasoning benchmarks

Table 2: Pass@1 results on mathematical reasoning benchmarks. Distilled RL consistently improves performance over OPD, RL, and OPD+RL across different student models.

Method	AIME24	AIME25	AIME26	CMIMC25	HMMT25	AMC23	GSM8K	MATH500	Minerva	Olympiad	Average
DSQW-1.5B	19.17	17.60	13.75	6.80	7.60	52.19	75.06	67.60	23.90	33.33	31.70
OPD	20.72	19.89	20.10	8.51	11.97	57.57	76.64	73.60	26.10	37.62	35.27
RL	21.04	22.70	18.02	8.82	10.93	60.62	80.13	79.80	27.57	38.96	36.86
OPD+RL	21.56	20.41	18.33	8.28	11.67	58.59	81.50	78.40	25.73	40.88	36.54
Distilled RL	**25.31**	**24.58**	**21.56**	**10.23**	**14.27**	**67.26**	**81.57**	**81.00**	**31.25**	**42.96**	**40.00**
Δ vs. OPD	(+4.59)	(+4.69)	(+1.46)	(+1.72)	(+2.30)	(+9.69)	(+4.93)	(+7.40)	(+5.15)	(+5.34)	(+4.73)
Qwen3-4B	38.23	23.54	27.92	13.59	13.44	72.58	94.47	86.20	43.38	49.93	46.33
OPD	52.08	39.47	45.83	22.50	23.64	84.68	94.69	91.80	46.32	58.66	55.97
RL	53.64	42.18	46.45	23.35	27.39	87.50	94.69	91.40	47.42	60.00	57.40
OPD+RL	54.06	39.37	46.45	22.89	23.85	84.06	94.84	92.00	47.05	59.25	56.38
Distilled RL	**56.45**	**42.29**	**49.06**	**26.95**	**28.85**	**87.89**	**94.99**	**93.20**	**48.89**	**61.03**	**58.96**
Δ vs. OPD	(+4.37)	(+2.82)	(+3.23)	(+4.45)	(+5.21)	(+3.21)	(+0.30)	(+1.40)	(+2.57)	(+2.37)	(+2.99)
Qwen3-1.7B	22.60	21.35	18.85	9.14	11.67	64.30	88.78	81.80	35.66	44.44	39.86
OPD	31.56	26.04	27.81	15.15	**15.31**	70.31	90.06	86.20	39.70	49.92	45.21
RL	29.79	**26.87**	27.18	14.21	14.68	69.21	89.76	86.00	40.44	49.48	44.76
OPD+RL	30.72	24.58	26.66	16.09	15.20	71.64	90.14	86.20	38.23	49.48	44.89
Distilled RL	**32.70**	26.45	**28.13**	**17.89**	15.20	**71.79**	**90.21**	**88.80**	**40.44**	**52.14**	**46.37**
Δ vs. OPD	(+1.14)	(+0.41)	(+0.32)	(+2.74)	(-0.11)	(+1.48)	(+0.15)	(+2.60)	(+0.74)	(+2.22)	(+1.16)

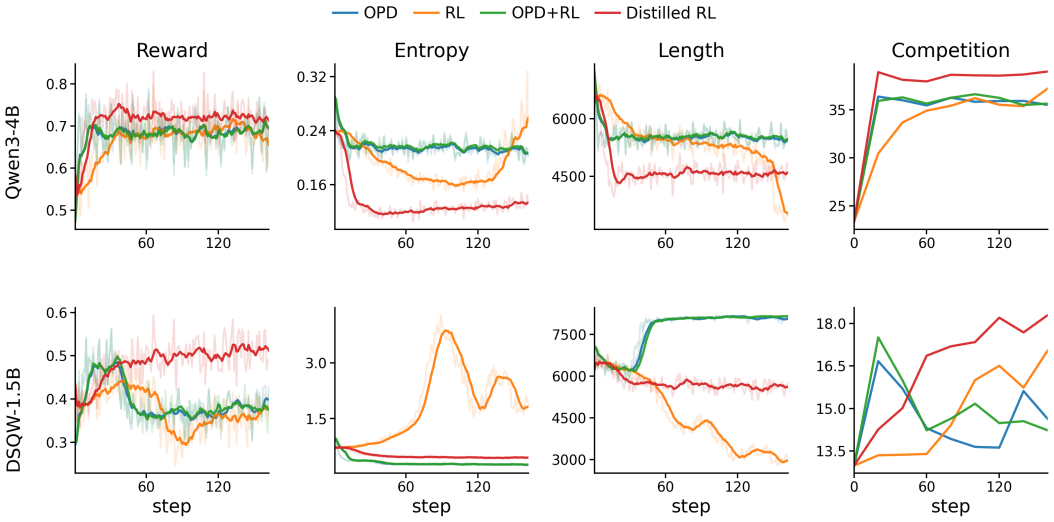

Figure 5: Training dynamics of Distilled RL and baseline methods. The panels show reward, policy entropy, response length, and average Pass@1 on the five competition benchmarks in Fig. 2.

and knowledge-intensive generalization benchmarks. Detailed baselines, datasets, and implementation settings are provided in Appendix B.

5.1 Main Results

Tables 2 and 3 report the main evaluation results of Distilled RL against the base model, OPD, RL, and OPD+RL. We evaluate on DeepSeek-R1-Distill-Qwen-1.5B (DSQW-1.5B), Qwen3-4B, and Qwen3-1.7B. DSQW-1.5B represents a challenging cross-family distillation setting, while Qwen3-4B corresponds to a within-family setting with the Qwen3-8B-GRPO teacher.

Performance. As shown in Table 2, Distilled RL consistently achieves the best Pass@1 performance across all three student models. For DSQW-1.5B, it improves the average score from 31.70 to 40.00, outperforming OPD, RL, and OPD+RL by 4.73, 3.14, and 3.46 points, respectively. On Qwen3-1.7B, Distilled RL achieves an average score of 46.37, exceeding the three baselines by 1.16, 1.61, and 1.48 points. For Qwen3-4B, it further improves the average score from 46.33 to 58.96, surpassing OPD, RL, and OPD+RL by 2.99, 1.56, and 2.58 points, respectively. These results demonstrate that

Table 3: Additional Pass@1 and Pass@16 results. Distilled RL achieves strong performance on both knowledge-intensive and reasoning benchmarks.

Methods	Pass@1		Pass@16				
	MMLU_{pro}	SuperGPQA	AIME24	AIME25	AIME26	CMIMC25	HMMT25
DSQW-1.5B	29.46	16.00	58.33	41.67	40.00	20.00	21.67
OPD	**37.50**	21.00	55.00	**43.33**	43.33	23.75	31.67
RL	32.67	20.60	60.00	38.33	43.33	22.50	26.67
OPD+RL	36.61	22.00	58.33	41.67	48.33	22.50	26.67
Distilled RL	36.25	**22.20**	**70.00**	**43.33**	**50.00**	**27.50**	**31.67**
Qwen3-4B	72.86	32.60	68.33	43.33	45.00	30.00	35.00
OPD	73.03	38.40	73.33	58.33	65.00	40.00	41.67
RL	72.86	38.60	80.00	65.00	68.33	38.75	**50.00**
OPD+RL	73.21	38.00	75.00	56.67	68.33	40.00	41.67
Distilled RL	**74.46**	**39.60**	**80.00**	**66.67**	**70.00**	**47.50**	46.67

incorporating teacher preferences directly into the RL objective consistently yields better optimization than standalone RL or KL-based distillation.

Training dynamics. Figure 5 compares the optimization trajectories of different methods. Distilled RL consistently maintains higher rewards across training. Distilled RL also maintains a moderate and stable policy entropy under both model architectures, avoiding both excessive uncertainty and premature policy collapse. Furthermore, Distilled RL exhibits a stable response length throughout training. Most importantly, the average Pass@1 of Distilled RL continues to improve steadily over the course of training. In contrast, RL improves slowly, OPD converges prematurely after rapid early progress, and OPD+RL provides no clear additional gain. These observations support the motivation of Distilled RL: rather than introducing a separate imitation objective, teacher supervision directly redistributes the RL learning signal, resulting in more stable optimization and sustained performance improvement.

Generalization. Beyond the mathematical reasoning benchmarks used for training evaluation, Table 3 further reports results on knowledge-intensive benchmarks together with Pass@16 evaluation. Distilled RL consistently achieves the best or highly competitive performance on MMLU-Pro and SuperGPQA, indicating that the proposed objective does not overfit to mathematical reasoning and preserves general knowledge capabilities during post-training. Under the Pass@16 evaluation, Distilled RL also obtains the strongest performance on most competition benchmarks for both student models. The larger improvement under multiple-sampling evaluation suggests that Distilled RL not only improves the quality of the highest-probability response but also produces a stronger overall response distribution.

Cross-family distillation. The advantage of Distilled RL is particularly evident in the cross-family setting. For DSQW-1.5B, OPD improves over the base model but remains consistently behind RL, suggesting that direct KL-based imitation becomes less effective when the teacher and student exhibit substantially different reasoning distributions. Simply combining OPD with RL also fails to consistently outperform RL, indicating that unconditional teacher imitation may interfere with reward optimization. In contrast, Distilled RL achieves the largest improvement over all baselines. These results suggest that selectively incorporating teacher preferences into the RL objective is substantially more robust than explicit distribution matching, especially when the teacher and student belong to different model families.

Overall, Distilled RL provides a more effective way to combine reinforcement learning and teacher supervision. By replacing unconditional KL imitation with selective importance-weighted guidance, it consistently improves Pass@1 and Pass@k across different student models, with especially clear gains in cross-family distillation.

Table 4: Ablation results of Distilled RL on mathematical reasoning and knowledge-intensive benchmarks. Values in parentheses denote the performance difference relative to the complete Distilled RL method.

Method	AIME24 AMC23	AIME25 GSM8K	AIME26 MATH500	CMIMC25 Minerva	HMMT25 Olympiad	AVG	MMLU_{pro}	SuperGPQA
DSQW-1.5B								
w/o Negative Reset	19.48	17.71	14.06	7.11	8.85	33.61	34.10	18.80
	54.53	79.30	73.00	27.20	34.81	(-6.39)	(-2.15)	(-3.40)
w/o Geometric Norm.	24.37	23.22	19.79	9.21	13.33	38.76	33.57	21.60
	66.01	80.51	80.80	29.04	41.33	(-1.24)	(-2.68)	(-0.60)
Qwen3-4B								
w/o Negative Reset	38.44	33.85	39.48	17.89	19.38	50.15	71.43	34.80
	79.53	92.49	88.40	41.54	50.52	(-8.81)	(-3.03)	(-4.80)
w/o Geometric Norm.	54.08	41.73	48.54	24.06	26.35	57.49	72.50	37.40
	86.56	94.54	92.60	46.69	59.70	(-1.47)	(-1.96)	(-2.20)

5.2 Ablation Study

Table 4 reports the ablation results of the two key components in Distilled RL.

Negative sample reset. Removing negative sample reset consistently causes the largest performance degradation on both Qwen3-4B and DSQW-1.5B. The average Pass@1 decreases by 8.81 and 6.39 points, respectively, indicating that teacher supervision on negative-advantage trajectories is harmful rather than beneficial. This observation agrees with our analysis in Table 1: when a sampled response receives a negative advantage, teacher-preferred tokens are assigned even larger penalties, forcing the student away from the teacher distribution.

Figure 6 further provides an interpretable demonstration. Under the same entropy-control setting as Fig. 4, removing the negative sample reset prevents the student from following the target entropy specified by the teacher. Instead of approaching the desired entropy region, the student's entropy remains substantially below the target and fails to respond to the teacher's temperature changes. This result suggests that applying teacher guidance to negative trajectories suppresses rather than facilitates knowledge transfer, preventing the student from acquiring new information from the teacher.

Sequence-level geometric normalization. Removing sequence-level geometric normalization also consistently degrades performance, although to a smaller extent. Since the sampled tokens are generated by the student policy rather than the teacher policy, the teacher typically assigns lower probabilities to many student-generated tokens, making the token-level importance ratios frequently smaller than one. Consequently, the product of these ratios decreases rapidly with sequence length, systematically weakening the optimization signal for positive trajectories. Sequence-level geometric normalization removes this global scaling effect while preserving the teacher's relative preference among tokens, enabling stable token-level credit redistribution without altering the overall optimization strength.

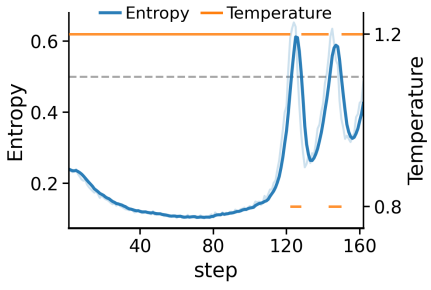

Figure 6: Entropy trajectory without negative sample reset.

6 Conclusion

We presented Distilled Reinforcement Learning (Distilled RL), a unified post-training framework that integrates teacher supervision directly into the reinforcement learning objective. Unlike conventional on-policy distillation, which relies on unconditional KL-based imitation, Distilled RL uses reverse importance sampling to selectively redistribute the RL learning signal according to teacher preferences. The proposed framework consists of three simple yet effective components: reverse importance

sampling, negative sample reset, and sequence-level geometric normalization. Together, these mechanisms enable fine-grained knowledge transfer while preserving the optimization behavior of reinforcement learning. Through an interpretable entropy-control case study, we further demonstrated that Distilled RL can transfer previously unavailable knowledge from the teacher to the student beyond the original reward signal. Extensive experiments on both within-family and cross-family distillation consistently showed that Distilled RL outperforms RL, OPD, and their direct combination across mathematical reasoning, knowledge-intensive benchmarks, and Pass@k evaluation. We hope Distilled RL provides a simple and effective paradigm for combining reinforcement learning and knowledge distillation in future LLM post-training.

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 *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025. doi: 10.48550/arXiv.2505.23281.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. *Advances in Neural Information Processing Systems*, 38:131646–131671, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025a.

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. The entropy mechanism of reinforcement learning for reasoning language models, 2025b. URL https://arxiv.org/abs/2505.22617.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

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

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025b.

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

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

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

Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, et al. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445*, 2025.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Fanfan Liu, Youyang Yin, Peng Shi, Siqi Yang, Zhixiong Zeng, and Haibo Qiu. Length-unbiased sequence policy optimization: Revealing and controlling response length variation in rlvr. *arXiv preprint arXiv:2602.05261*, 2026.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why on-policy reinforcement learning forgets less. In *Non-Euclidean Foundation Models: Advancing AI Beyond Euclidean Frameworks*.

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

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Chen Wang, Zhaochun Li, Jionghao Bai, Yuzhi Zhang, Shisheng Cui, Zhou Zhao, and Yue Wang. Arbitrary entropy policy optimization breaks the exploration bottleneck of reinforcement learning. *arXiv preprint arXiv:2510.08141*, 2025.

Chen Wang, Hexuan Deng, Yining Zhang, Yuchen Zhang, Jionghao Bai, Zhaochun Li, Ge Lan, and Yue Wang. Implicit compression regularization: Concise reasoning via internal shorter distributions in rl post-training. *arXiv preprint arXiv:2605.07316*, 2026a.

Chen Wang, Lai Wei, Yanzhi Zhang, Chenyang Shao, Zedong Dan, Weiran Huang, Ge Lan, and Yue Wang. Targeted exploration via unified entropy control for reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 16789–16802, 2026b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024a.

Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5737–5755, 2025a.

Yecheng Wu, Song Han, and Hai Cai. Lightning opd: Efficient post-training for large reasoning models with offline on-policy distillation. *arXiv preprint arXiv:2604.13010*, 2026.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*, 2025b.

Hongling Xu, Qi Zhu, Heyuan Deng, Jinpeng Li, Lu Hou, Yasheng Wang, Lifeng Shang, Ruifeng Xu, and Fei Mi. Kdrl: Post-training reasoning llms via unified knowledge distillation and reinforcement learning. *arXiv preprint arXiv:2506.02208*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Zheng Yaowei, Lu Junting, Wang Shenzhi, Feng Zhangchi, Kuang Dongdong, and Xiong Yuwen. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

Jingyang Yi, Jiazheng Wang, and Sida Li. Shorterbetter: Guiding reasoning models to find optimal inference length for efficient reasoning. *arXiv preprint arXiv:2504.21370*, 2025.

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

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

Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*, 2024.

Appendix

A Limitation

A fundamental limitation of on-policy distillation is that the teacher is only queried to evaluate student-generated trajectories, without sampling complete responses from the teacher. Consequently, the training procedure cannot directly determine whether the teacher is capable of solving a given problem. Distilled RL mitigates much of the resulting risk through negative sample reset, which removes teacher-based reweighting from negative student trajectories. Nevertheless, an implicit assumption remains: for problems that the student solves correctly, the teacher should also possess sufficient competence such that the token-level preferences provide useful guidance. This assumption is theoretically reasonable because the teacher is generally expected to be stronger than the student. In practice, however, teacher superiority is neither uniform across different prompts nor guaranteed throughout the dynamic training process. As the student improves, it may become comparable to or even locally outperform the fixed teacher on certain problems, in which case teacher-based reweighting may introduce suboptimal or misleading preferences into otherwise successful trajectories. Future work could alleviate this limitation by explicitly estimating teacher competence, for example, through occasional teacher-side rollouts, verifier-based teacher correctness checks, confidence-aware gating, or adaptive attenuation of the distillation weights when the teacher provides uncertain or inconsistent guidance.

B Implementation Details

Experimental setup. We implement Distilled RL using the EasyR1 and VeRL frameworks (Yaowei et al., 2025; Sheng et al., 2025). Unless otherwise specified, all methods are trained on DAPO-17K (Yu et al., 2025) using the same prompt distribution, rollout budget, optimization schedule, and evaluation protocol. We use Qwen3-8B-GRPO as the teacher model and evaluate two student models: Qwen3-4B and DeepSeek-R1-Distill-Qwen-1.5B, abbreviated as DSQW-1.5B (Yang et al., 2024; 2025; Guo et al., 2025b). Qwen3-4B provides a relatively compatible teacher–student setting, whereas DSQW-1.5B presents a more challenging setting with a larger mismatch in model scale, initialization, and learned reasoning distribution. The complete training configuration is summarized in Table 5.

Table 5: Detailed configurations in the experiments.

Configuration	Setting
Training Settings	
Global batch size $\|B\|$	128
Mini-batch size	64
Micro-batch size per GPU	2
Rollout group size G	8
Maximum prompt length	2048
Maximum response length	8192
Sampling temperature	1
Top-p	0.99
Learning rate	1e-6
Number of training steps	160
Policy Optimization	
Advantage estimation	Group-normalized outcome reward
Reward type	Binary correctness reward
Lower policy clip ϵ_{low}	0.2
Upper policy clip ϵ_{high}	0.2
Distilled RL Settings	
Teacher–student ratio	$\rho_{i,t} = \pi_{\text{teacher}}/\pi_{\theta_{\text{old}}}$
Ratio clipping range	$[1/\epsilon_\rho,\ \epsilon_\rho]$
Ratio clipping threshold ϵ_ρ	3
Negative sample reset	$w_{i,t} = 1$ when $A_i \le 0$
Geometric normalization	Sequence-level geometric mean normalization
Baseline Settings	
RL	GRPO
KL coefficient	1e-2
OPD divergence	Reverse KL divergence
OPD+RL mixing coefficient	1.0
Evaluation Settings	
Maximum response length	8192
Temperature	0.1
Top-p	0.95
Number of samples for Pass@1	32
Number of samples for Pass@16	32

For each training prompt q, the old student policy $\pi_{\theta_{\text{old}}}$ generates a group of $G = 8$ responses. Each response is evaluated using a binary correctness reward. Following GRPO (Shao et al., 2024), we normalize the rewards within each rollout group to obtain the response-level advantage.

Evaluation benchmarks. We evaluate mathematical reasoning on AIME24, AIME25, AIME26, CMIMC25, HMMT25, AMC23, GSM8K, MATH500, Minerva Math, and Olympiad (Balunović et al., 2025; Lightman et al., 2023; Cobbe et al., 2021; Lewkowycz et al., 2022). The first five datasets are recent competition benchmarks and are also used to report the averaged competition score in the training-dynamics figures. The remaining datasets cover a broader range of grade-school, competition-level, and formal mathematical reasoning problems.

To examine whether post-training preserves or improves capabilities beyond the primary mathematical evaluation, we further assess knowledge-intensive generalization on MMLU-Pro (Wang et al., 2024a) and SuperGPQA (Du et al., 2025). Since the full SuperGPQA benchmark is relatively large and computationally expensive to evaluate, we construct a fixed subset of 500 examples for evaluation. No examples from either benchmark are used as supervised training targets.