FROM IMITATION TO REWARD DISCOVERY: ON-POLICY WARMUP FOR AGENTIC RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with a verifiable reward (RLVR) offers a scalable approach to training language-model agents, yet sparse outcome rewards can leave early training with little signal for policy improvement. We identify an **On-Policy Acceleration Phenomenon**: in our main comparisons, RLVR initialized with on-policy distillation reaches high performance earlier in training and achieves both higher average performance during subsequent RLVR and higher final performance than the alternative baselines. Motivated by this observation, we study **On-Policy Warmup (OPW)**, a teacher-guided stage in which the student trains with teacher supervision on its own interaction trajectories before transitioning to RLVR. Unlike imitation on fixed teacher-generated trajectories, OPW targets states induced by the student's own decisions, including imperfect actions and recovery situations. We provide a theoretical explanation by connecting on-policy reverse-KL distillation to trajectory-level distribution matching. Under a competent teacher and sufficiently small population distillation loss, this connection yields a lower bound on initial verifier success and a corresponding bound on reward-discovery complexity. For group-relative RLVR, we further characterize when increased success probability produces more reward-informative groups. Together, our findings support on-policy distillation as an effective warmup for agentic RLVR and identify initial reward discovery as a mechanism that can contribute to the observed acceleration.

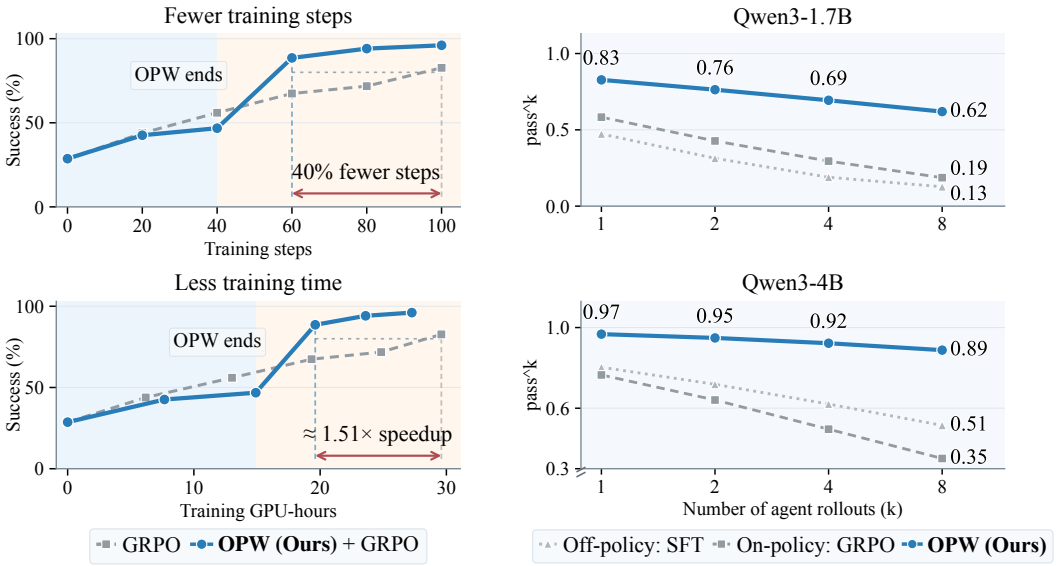

Figure 1: **OPW accelerates GRPO and improves repeated task success in ALFWorld. Left:** Qwen3-4B Seen learning curves for GRPO with and without OPW. GPU-hours include teacher scoring and exclude evaluation. **Right:** Final Unseen repeated task success (pass^k).

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

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves language-model reasoning using automatically checkable outcomes rather than human annotations (Shao et al., 2024; DeepSeek-AI et al., 2025). It is appealing for agents that interleave reasoning with tool use and environment feedback (Yao et al., 2023). Yet outcome-based feedback offers limited guidance about the intermediate decisions required for success. This creates a cold-start challenge: when success requires several coordinated decisions, an initial policy may produce mostly failures with identical rewards. Under binary outcome rewards, all-failure groups provide no within-group task-reward contrast for group-relative methods such as GRPO (Shao et al., 2024). Effective early training requires not just diverse trajectories, but successes frequent enough for the verifier to distinguish useful behavior.

Teacher supervision offers a way to improve this starting point, where knowledge distillation can transfer a denser supervision signal to a student (Hinton et al., 2015). However, imitation on fixed teacher-generated trajectories primarily supervises states visited by the teacher, whereas an agent must act on histories induced by its own decisions. This distribution mismatch is a central concern in imitation learning (Ross et al., 2011) and is particularly consequential in interactive tasks: an imperfect action can alter subsequent tool outputs, available information, and opportunities for recovery. On-policy distillation addresses the mismatch by applying teacher supervision to student-generated trajectories (Agarwal et al., 2024). Whether such supervision provides a useful initialization for subsequent RLVR, rather than merely improving imitation, remains a distinct question.

In this work, we identify an **On-Policy Acceleration Phenomenon**: after on-policy distillation warmup, RLVR reaches high performance earlier and achieves higher final performance than the compared baselines. Motivated by this observation, we propose **On-Policy Warmup (OPW)**, a teacher-guided initialization stage for agentic RLVR. During warmup, the student generates trajectories by interacting with the environment and receives supervision from a fixed teacher. The student learns from this supervision before transitioning to RLVR, where training uses verifiable task outcomes without further teacher queries. Teacher supervision is confined to warmup.

The purpose of warmup is not simply to reduce distillation loss. The downstream objective is task success under sparse verifier feedback, and a student that imitates the teacher more closely need not, in general, learn faster during RLVR. Instead, we ask whether on-policy supervision makes successful trajectories accessible often enough to improve initial reward discovery (Chen et al., 2025). This question connects the distribution of trajectories induced by the warmed-up policy to the training signal available to the subsequent RLVR stage. It also separates the role of teacher guidance during initialization from the role of the verifier during later policy improvement.

We analyze this connection theoretically. Under shared environment dynamics, the cumulative reverse-KL distillation loss on student-visited states equals the KL divergence between student and teacher trajectory distributions. The data-processing inequality then bounds the gap between their success probabilities under the same binary verifier used to assess task completion. A sufficiently successful teacher and sufficiently small student population loss imply a positive lower bound on initial student success probability, and thus an upper bound on the expected rollouts to discover a success. For group-relative RLVR, we also characterize when higher success probability increases the likelihood of sampling a group containing both successes and failures. These results suggest initial reward discovery as a mechanism for the observed acceleration, but neither guarantee convergence nor imply lower total compute once warmup and teacher queries are included.

Our contributions are threefold:

- **We identify the On-Policy Acceleration Phenomenon in agentic RLVR.** In our main comparisons, on-policy distillation warmup leads to earlier high performance, higher average performance during subsequent RLVR, and higher final performance than the compared baselines.
- **We study OPW as a teacher-guided initialization stage for RLVR.** OPW provides teacher supervision along student-generated trajectories, including after unsuccessful actions, and is followed by RLVR without further teacher queries.
- **We connect on-policy reverse-KL distillation to initial reward discovery.** Through trajectory-level distribution matching, we derive a lower bound on initial verifier success and bound reward-discovery complexity under sufficient teacher competence and small population distillation loss.

Table 1: Related work on warmup for RLVR.

Work	Setting		Evidence		Training study	
	On-policy warmup	Environment interaction	Theoretical analysis	Action-level interventions	Training cost	Warmup duration
Sequential Beats Joint (Li et al., 2026a)	✔	✗	✗	✗	✗	✔
RL Starts before RL (Dong et al., 2026)	✔	✗	✗	✗	✗	✗
OPDSearch+ (Ye et al., 2026)	✔	✔	✔	✗	✔	✗
PRISM (Wang et al., 2026a)	✔	✗	✗	✗	✗	✗
PEAR (Zhang et al., 2026)	✗	✗	✔	✗	✗	✗
OPW (Ours)	✔	✔	✔	✔	✔	✔

2 Related Work

Warmup for RLVR. RLVR trains language-model policies using verifiable outcome rewards (Shao et al., 2024). SFT is commonly used as warmup before RLVR in mathematical reasoning and logic tasks (Luong et al., 2024; Shrestha et al., 2025). Stronger SFT performance does not necessarily lead to better performance after RL (Kang et al., 2026; Zhang et al., 2026). PEAR reweights fixed offline supervision using importance sampling to improve subsequent RL (Zhang et al., 2026). TailSFT filters sequences according to their likelihood improvement during SFT to improve coverage and subsequent GRPO performance (Malladi et al., 2026). For on-policy distillation (OPD), Sequential Beats Joint compares sequential and joint OPD–RLVR training and examines when to switch to RL (Li et al., 2026a). RL Starts before RL shows that initial accuracy and pre-RL pass@k do not fully explain subsequent performance, and compares trajectory sources and divergence objectives (Dong et al., 2026). PRISM inserts a black-box, response-level adversarial alignment stage between SFT and multimodal RLVR (Wang et al., 2026a). Closest to our interactive setting, OPDSearch+ applies teacher supervision to student trajectories collected through live search interactions before RL refinement (Ye et al., 2026). Table 1 compares these warmup methods.

On-Policy Supervision for Interactive Agents. Knowledge distillation transfers teacher behavior to a student (Hinton et al., 2015). For language models and agents, a common approach is SFT on teacher-generated responses or interaction trajectories (DeepSeek-AI et al., 2025; Zeng et al., 2023). GKD instead applies teacher feedback to student-generated sequences, addressing the mismatch between training data and the student's own outputs (Agarwal et al., 2024). MiniLLM studies reverse-KL distillation to discourage the student from assigning excessive probability to regions that are unlikely under the teacher (Gu et al., 2024). Guided-OPD extends this line of work to multi-turn agents by mixing teacher- and student-generated turns and gradually reducing the probability of teacher intervention (Li et al., 2026b). Other methods combine ongoing RL with teacher trajectories (Yan et al., 2025), token-level guidance from interaction feedback (Wang et al., 2026b), or self-distillation from completed interactions (Wu et al., 2026). We study teacher supervision on student-generated trajectories as warmup before teacher-free RLVR in multi-turn environments. Under population reverse-KL assumptions, we connect trajectory distributions to initial verifier success and reward-discovery complexity. We remove supervision at repeated invalid actions, evaluate teacher-guided choices through task continuations from common histories, and track the retention of teacher-target preferences during GRPO. We also compare warmup durations and recorded training costs, including teacher supervision, to distinguish faster subsequent learning from lower total cost.

3 Method

We study OPD as warmup for RLVR. Our analysis connects distillation to trajectory-level distribution matching and informative reward observations during early RLVR.

3.1 Theoretical Motivation for On-Policy Warmup

Setup. Fix a task x and a finite-horizon interaction process with horizon H. Let s_t denote the full interaction history, including previous actions and environment observations, and let a_t denote the next policy action. Variable-length trajectories can be padded with an absorbing state and a shared dummy action, which contributes zero KL. For token-level language policies, t indexes policy-

generated tokens, including reasoning and action tokens, and H bounds their number. The student policy π and teacher policy π_T interact with the same environment and initial-state distribution. Let P_π^x and P_T^x denote their induced trajectory distributions. For a binary verifier $R_x(\tau) \in \{0, 1\}$ indicating complete task success, define $p_\pi(x) = \mathbb{E}_{\tau \sim P_\pi^x}[R_x(\tau)]$ and $p_T(x) = \mathbb{E}_{\tau \sim P_T^x}[R_x(\tau)]$.

All theoretical statements below are conditional on x; we suppress this dependence when unambiguous. We define the population on-policy distillation loss as

$$\mathcal{L}_{\mathrm{OPD}}(\pi) = \sum_{t=1}^{H} \mathbb{E}_{s_t \sim d_t^\pi} \left[D_{\mathrm{KL}}\big(\pi(\cdot \mid s_t) \,\|\, \pi_T(\cdot \mid s_t)\big) \right], \tag{1}$$

where d_t^π is the student's state distribution. We assume finite population loss and use natural logs.

Proposition 3.1 (Trajectory-Level Interpretation)**.** *Under the shared-environment assumption,* $D_{\mathrm{KL}}(P_\pi \| P_T) = \mathcal{L}_{\mathrm{OPD}}(\pi)$.

Theorem 3.2 (Success Transfer from Teacher to Student)**.** *If* $\mathcal{L}_{\mathrm{OPD}}(\pi) \le \varepsilon$*, then* $\mathrm{KL}(p_\pi \| p_T) \le \varepsilon$*, where* KL *denotes the KL divergence between Bernoulli distributions. In particular,*

$$|p_\pi - p_T| \le \sqrt{\frac{\varepsilon}{2}}, \qquad p_\pi \ge \left[p_T - \sqrt{\frac{\varepsilon}{2}} \right]_+ . \tag{2}$$

Theorem 3.2 provides a sufficient condition for transferring verifier success from a competent teacher. When the loss is reported as an average over H steps, $\overline{\mathcal{L}}_{\mathrm{OPD}} = \mathcal{L}_{\mathrm{OPD}}/H$, the corresponding deviation bound is $\sqrt{H\overline{\mathcal{L}}_{\mathrm{OPD}}/2}$. A small average token-level loss alone need not give a tight guarantee over long interactions, because loss accumulates across action positions.

Why On-Policy States Matter. The trajectory identity in Proposition 3.1 specifically requires distillation under the student's distribution. To illustrate the role of state coverage, define $k_\pi(s) = D_{\mathrm{KL}}\big(\pi(\cdot \mid s) \| \pi_T(\cdot \mid s)\big)$ and consider evaluating the same local reverse-KL discrepancy under an offline state distribution μ_t: $\mathcal{L}_\mu(\pi) = \sum_{t=1}^{H} \mathbb{E}_{s \sim \mu_t}[k_\pi(s)]$. If d_t^π is absolutely continuous with respect to μ_t and $\frac{d d_t^\pi}{d \mu_t}(s) \le C$ for every t and almost every s, then $\mathcal{L}_{\mathrm{OPD}}(\pi) \le C\, \mathcal{L}_\mu(\pi)$. The success-transfer bound obtained from this offline reverse-KL loss thus depends on an additional state-coverage coefficient C. If $d_t^\pi \not\ll \mu_t$ for any t, no finite C exists. OPD directly targets the state-weighted discrepancy appearing in the trajectory KL, avoiding this additional distribution-transfer requirement at the population level. This is particularly relevant for agentic tasks, where imperfect actions can lead to novel tool outputs, execution failures, and recovery states.

3.2 Implications for Early RLVR Reward Discovery

For the student π_{w} after OPW, Theorem 3.2 gives $p_{\mathrm{w}} = p_{\pi_{\mathrm{w}}} \ge \ell = \left[p_T - \sqrt{\mathcal{L}_{\mathrm{OPD}}(\pi_{\mathrm{w}})/2} \right]_+$.

Corollary 3.3 (Initial Reward-Discovery Complexity)**.** *Suppose* $\ell > 0$*. For independent rollouts from the fixed warm-start policy* π_{w}*, the probability of observing at least one successful trajectory among* N *rollouts satisfies*

$$\Pr(\textit{at least one success}) = 1 - (1 - p_{\mathrm{w}})^N \ge 1 - (1 - \ell)^N \ge 1 - e^{-N\ell}. \tag{3}$$

For any $\delta \in (0, 1)$*,* $N \ge \left\lceil \frac{\log(1/\delta)}{\ell} \right\rceil$ *is sufficient to observe a success with probability at least* $1 - \delta$*. The expected number of rollouts until the first success is at most* $1/\ell$ *under the same fixed policy.*

If the pre-warmup success probability is p_0 and $\ell > p_0$, OPD is certified to improve this initial discovery process. For group-relative methods with binary rewards, the relevant event is observing successes and failures in the same group. For $G \ge 2$ independent rollouts on the same task, its probability is $m_G(p) = 1 - p^G - (1-p)^G$. Such groups contain nonzero empirical reward variance and can support reward-based differentiation among trajectories. The derivative $m_G'(p) = G\big((1-p)^{G-1} - p^{G-1}\big)$ is nonnegative on $[0, 1/2]$. Therefore, in the sparse-success regime $p_0 < p_{\mathrm{w}} \le 1/2$, increasing success probability increases the frequency of reward-informative groups. The lower bound $p_{\mathrm{w}} \ge \ell$ alone does not imply an increase in reward-informative groups. This monotonicity does not hold globally: near $p = 1$, groups can become uninformative because all trajectories succeed. For partial-credit rewards, unsuccessful trajectories can still receive different scores.

Table 2: Final task performance across three training paths after 100 steps (Qwen3-4B).

Split	Task type / Training	ALFWorld							ScienceWorld						
		Pick	Clean	Heat	Cool	Exam	Pick2	All	Find	Life	Gene.	Mix	Prop.	Phys.	All
Seen	Direct OPD	88.6	30.1	9.4	12.5	51.9	57.3	45.9	61.4	46.2	29.6	**30.4**	52.1	44.2	48.9
	Direct GRPO	**100.0**	83.3	78.9	83.5	85.6	95.8	89.4	80.3	39.8	55.3	25.1	61.1	62.1	61.2
	OPW (Ours) + GRPO	98.9	**98.1**	**86.7**	**93.5**	**97.1**	**97.9**	**96.1**	**97.0**	**53.6**	**78.7**	**30.4**	**98.0**	**82.4**	**87.3**
Unseen	Direct OPD	90.6	54.0	12.0	48.2	56.9	59.6	53.5	64.8	45.4	47.7	8.1	60.8	47.1	55.2
	Direct GRPO	88.5	82.3	83.7	90.5	83.3	94.1	86.6	**74.0**	50.7	50.4	13.6	63.8	61.0	61.2
	OPW (Ours) + GRPO	**93.8**	**100.0**	**92.4**	**99.4**	**97.2**	**97.1**	**96.7**	71.8	**62.5**	**52.4**	**15.4**	**75.8**	**74.9**	**68.2**

3.3 On-Policy Warmup for RLVR

Motivated by this connection, we propose *On-Policy Warmup* (OPW): a teacher-guided initialization stage preceding RLVR. Starting from the base model, the pipeline is $\pi_{\theta_0} \xrightarrow{\text{OPD}} \pi_{\theta_{\text{w}}} \xrightarrow{\text{RLVR}} \pi_{\theta_{\text{final}}}$.

Stage I: Teacher Supervision on Student Trajectories. Let $\mathcal{D}_{\text{w}}$ and $\mathcal{D}_{\text{RL}}$ denote the warmup and RLVR task distributions, and π_T a fixed teacher policy. Our experiments use disjoint training task pools. At warmup iteration k, we sample a task $x \sim \mathcal{D}_{\text{w}}$ and generate an interaction trajectory using the current student $\tau \sim P^x_{\pi_{\theta_k}}$. We query the teacher at interaction histories along these trajectories. Unlike offline imitation of teacher trajectories, OPW exposes the teacher to states induced by the student's own decisions, including imperfect intermediate solutions, failed tool calls, and recovery.

The population reverse-KL objective considered in our analysis is $\mathcal{L}_{\text{OPW}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{w}}}\left[\mathcal{L}_{\text{OPD}}(\pi_\theta; x)\right]$, using the task-specific loss in Eq. 1. For this reverse-KL formulation, the corresponding fixed-history surrogate at iteration k is

$$\widetilde{\mathcal{L}}_k(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{w}},\, \tau\sim P^x_{\pi_{\theta_k}}}\left[\sum_{t=1}^{H} D_{\text{KL}}\left(\pi_\theta(\cdot \mid s_t) \,\|\, \pi_T(\cdot \mid s_t)\right)\right]. \tag{4}$$

Sampled histories are held fixed during each step, and trajectories are resampled as the student changes. Teacher supervision is applied to policy-generated tokens, not to environment observations.

Stage II: Verifier-Based Reinforcement Learning. After a prescribed warmup duration, we initialize RLVR with $\pi_{\theta_{\text{w}}}$ and optimize

$$J_{\text{VR}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{RL}},\, \tau\sim P^x_{\pi_\theta}}\left[r_x(\tau)\right], \qquad \theta \leftarrow \theta_{\text{w}}, \tag{5}$$

where $r_x(\tau) \in [0, 1]$ is the training reward. It equals the binary verifier R_x in ALFWorld and includes partial progress in ScienceWorld. We use GRPO's group-relative updates without probability-ratio clipping or additional KL or entropy terms (Table 8). In the basic OPW pipeline, the distillation loss is removed after warmup and no further teacher queries are required.

Scope. The bounds concern initial reward discovery under task-specific population reverse-KL control, with success measured under the same policy distributions. The analysis does not establish that finite warmup meets this condition or that it transfers across tasks or decoding settings.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate agents on ALFWorld (Shridhar et al., 2021) and ScienceWorld (Wang et al., 2022), which require multi-step interaction to complete household and scientific tasks. ALFWorld covers placing, cleaning, heating, cooling, examining objects under a light, and placing two objects, with 3,553 training tasks and 140 Seen and 134 Unseen test tasks. ScienceWorld contains 2,294 training tasks, of which 64 are reserved for Seen evaluation, and 200 Unseen test tasks. We group its tasks into finding, growth/lifespan, genetics, mixing, properties, and mechanics/energy.

Baselines. We use Qwen3-1.7B and Qwen3-4B as students and Qwen3-32B as the teacher (Yang et al., 2025). All RLVR training uses GRPO. OPW uses cross-entropy on teacher top-1 targets at

Table 3: Final task performance after 40 warmup and 60 GRPO steps (Qwen3-4B).

Split	Warmup \ Task type	ALFWorld							ScienceWorld						
		Pick	Clean	Heat	Cool	Exam	Pick2	All	Find	Life	Gene.	Mix	Prop.	Phys.	All
Seen	Off-policy: SFT	97.5	82.4	75.8	74.5	66.3	90.6	83.9	69.9	49.3	67.8	**42.8**	58.2	63.0	60.5
	Off-policy: SFT-RS	98.9	90.7	78.1	78.5	88.5	91.1	89.0	79.7	50.2	58.5	34.6	67.1	62.0	64.6
	On-policy: GRPO	**99.3**	74.1	78.9	69.5	70.2	91.1	82.7	81.5	51.2	61.1	35.3	67.2	73.3	68.8
	OPW (Ours)	98.9	**98.1**	**86.7**	**93.5**	**97.1**	**97.9**	**96.1**	**97.0**	**53.6**	**78.7**	30.4	**98.0**	**82.4**	**87.3**
Unseen	Off-policy: SFT	88.0	74.6	82.1	81.5	72.2	84.6	80.3	71.2	53.0	45.3	14.4	68.6	61.6	61.9
	Off-policy: SFT-RS	88.0	91.1	81.5	92.9	85.4	86.0	87.8	70.2	54.3	43.0	12.4	75.4	59.9	63.6
	On-policy: GRPO	85.4	78.2	71.2	77.4	66.7	77.2	76.5	68.4	54.5	50.0	**15.8**	61.5	60.0	59.6
	OPW (Ours)	**93.8**	**100.0**	**92.4**	**99.4**	**97.2**	**97.1**	**96.7**	**71.8**	**62.5**	**52.4**	15.4	**75.8**	**74.9**	**68.2**

Table 4: Final repeated task success (pass^8, %) for three training paths at 100 steps (Qwen3-4B).

Split	Training \ Task type	ALFWorld							ScienceWorld						
		Pick	Clean	Heat	Cool	Exam	Pick2	All	Find	Life	Gene.	Mix	Prop.	Phys.	All
Seen	Direct OPD	80.0	25.9	0.0	0.0	23.1	16.7	30.0	0.0	0.0	0.0	0.0	0.0	4.8	1.6
	Direct GRPO	**100.0**	48.1	68.8	60.0	38.5	**87.5**	71.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	OPW (Ours) + GRPO	97.1	**92.6**	**81.2**	**80.0**	**84.6**	83.3	**87.9**	**62.5**	**25.0**	**50.0**	0.0	**63.0**	**19.0**	**43.8**
Unseen	Direct OPD	83.3	41.9	0.0	9.5	11.1	5.9	28.4	1.9	2.9	12.0	0.0	0.0	0.0	2.5
	Direct GRPO	79.2	38.7	56.5	61.9	72.2	76.5	61.9	**9.6**	2.9	8.0	0.0	3.1	**21.1**	7.0
	OPW (Ours) + GRPO	**87.5**	**100.0**	**82.6**	**95.2**	**77.8**	**82.4**	**88.8**	3.8	**14.3**	**16.0**	0.0	**4.6**	15.8	**8.5**

response-token positions in student-generated trajectories. We compare OPW with **(1)** off-policy SFT, **(2)** SFT with rejection-sampling (SFT-RS), and **(3)** teacher-free on-policy GRPO warmup as baselines. Each warmup method runs for 40 steps and is followed by 60 GRPO steps without further teacher supervision. We compare OPW + GRPO with direct OPD and direct GRPO over 100 steps.

Evaluation metrics. Task performance is mean task success (pass@1, %) in ALFWorld and normalized task scores (0–100) in ScienceWorld. We measure success coverage by pass@k (at least one success in k rollouts) (Chen et al., 2021), and repeated task success by pass^k (success in every rollout) (Yao et al., 2024; Barres et al., 2025). ScienceWorld scores include partial progress, while success requires full task completion. Final results use eight trajectories per task, and the warmup success profile uses 64. Training and evaluation details are provided in Appendix B.

4.2 Results

Warmup VS. Direct Training. OPW followed by GRPO achieves the best final task performance among the three training paths in both environments (Table 2). On ALFWorld Unseen, it reaches 96.7% success, compared with 86.6% for direct GRPO and 53.5% for direct OPD. The same ordering holds in ScienceWorld, where Unseen task scores reach 68.2, 61.2, and 55.2, respectively. These results support on-policy distillation as an effective warmup for agentic RLVR.

Comparison of Warmup Strategies. OPW yields the highest final task performance under the same GRPO procedure after warmup (Table 3). On ALFWorld Unseen, restricting SFT to successful teacher trajectories raises final success from 80.3% to 87.8%, while OPW reaches 96.7%, compared with 76.5% after GRPO warmup. In ScienceWorld, OPW exceeds the strongest alternative by 18.5 points on Seen tasks and 4.6 points on Unseen tasks. Gains over both SFT variants and teacher-free warmup support teacher supervision on student-generated trajectories during warmup.

Task-Specific Performance. In ALFWorld, Clean, Heat, and Cool require changing an object's state before placement, adding intermediate requirements to the retrieval and placement needed for Pick (Shridhar et al., 2021). OPW improves these tasks in both comparisons (Tables 2 and 3). In the warmup comparison, Seen Pick success already exceeds 97% for all methods, while OPW improves Clean and Cool success over GRPO warmup by 24.0 percentage points each. In ScienceWorld, Mixing requires selecting and combining materials to obtain a target product (Wang et al., 2022).

Table 5: Success coverage and repeated task success after warmup (pass@k/pass^k, %).

Policy \ Split / k	Seen				Unseen			
	$k=8$	$k=16$	$k=32$	$k=64$	$k=8$	$k=16$	$k=32$	$k=64$
ALFWorld								
Teacher (32B)	70.9/27.5	76.4/21.5	81.0/15.2	85.7/10.0	**88.3**/24.2	**93.8**/17.3	**97.1**/11.5	**99.3**/6.7
Base (4B)	51.5/15.8	58.4/13.2	64.6/10.9	69.3/9.3	55.2/10.9	63.1/8.0	70.7/5.9	78.4/4.5
Off-policy: SFT	52.9/12.6	60.0/9.3	66.1/6.7	71.4/4.3	58.7/7.7	67.8/5.2	75.8/3.8	82.8/3.0
Off-policy: SFT-RS	53.2/12.6	60.7/9.1	67.5/6.4	73.6/4.3	59.8/7.0	69.0/3.7	76.8/1.9	84.3/1.5
On-policy: GRPO	**78.6**/**28.4**	**83.6**/23.0	**87.9**/18.6	**91.4**/15.0	80.1/19.8	84.8/13.9	88.1/9.0	90.3/5.2
OPW (Ours)	69.3/28.1	75.2/**24.6**	79.5/**22.4**	82.9/**21.4**	88.1/**27.5**	93.2/**22.9**	95.0/**19.7**	95.5/**17.9**
ScienceWorld								
Teacher (32B)	**75.8**/0.6	**85.0**/0.0	**91.9**/0.0	**95.3**/0.0	**76.3**/3.1	**84.5**/1.1	**88.6**/0.2	**90.5**/0.0
Base (4B)	23.9/0.0	28.9/0.0	32.9/0.0	35.9/0.0	35.9/2.9	44.6/1.7	53.5/0.8	62.0/0.0
Off-policy: SFT	25.6/0.0	31.6/0.0	37.0/0.0	42.2/0.0	37.1/2.1	45.8/0.7	54.4/0.1	63.0/0.0
Off-policy: SFT-RS	22.0/0.0	27.2/0.0	31.9/0.0	35.9/0.0	38.0/2.1	46.6/0.7	55.2/0.1	63.5/0.0
On-policy: GRPO	51.5/1.4	56.3/0.1	59.5/0.0	62.5/0.0	62.9/**5.4**	71.1/**2.3**	77.4/0.8	82.5/**0.5**
OPW (Ours)	73.7/**7.0**	82.3/**3.5**	87.6/**1.7**	90.6/0.0	66.8/4.5	75.9/2.0	82.4/**0.9**	85.5/**0.5**

Each cell: pass@k/pass^k (%). **Best** and second-best nonzero values per metric and column, including Teacher.

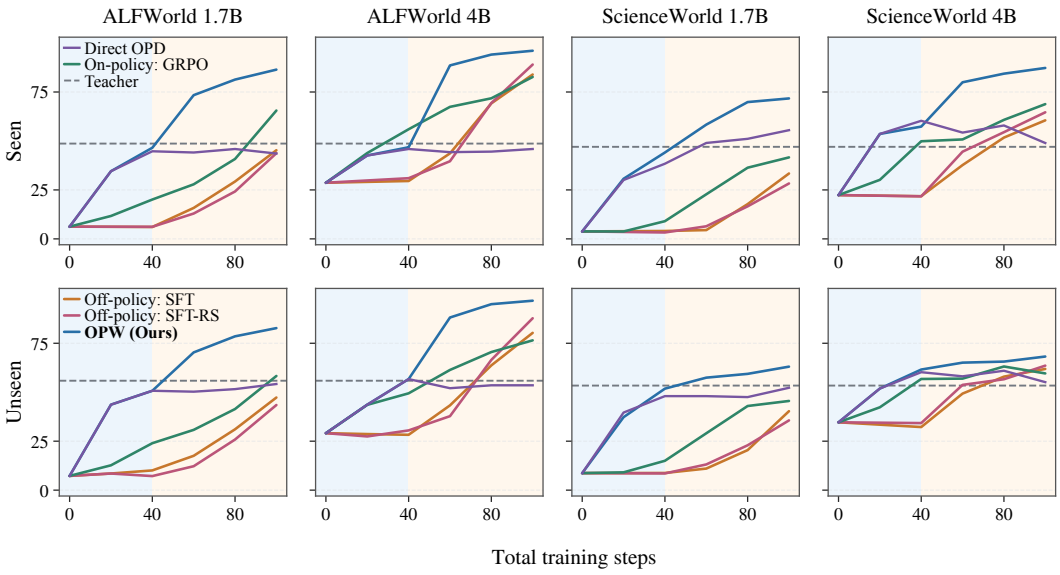

Figure 2: Task performance during warmup and GRPO.

OPW does not lead this category despite its higher scores on Properties and Mechanics/Energy, showing that its gains do not extend uniformly across scientific tasks (Table 3).

Repeated Task Success. OPW improves repeated task success in both environments (Table 4). On ALFWorld Unseen, final pass^8 reaches 88.8%, compared with 61.9% for direct GRPO and 28.4% for direct OPD. In ScienceWorld, OPW followed by GRPO reaches final pass^8 of 43.8% on Seen tasks and 8.5% on Unseen tasks, compared with 0.0% and 7.0% for direct GRPO. Thus, more tasks are completed successfully in every rollout, although repeated task success remains difficult on ScienceWorld Unseen. The comparison after warmup further distinguishes success coverage from repeated task success (Table 5). On ALFWorld Seen, OPW has lower pass@64 than GRPO warmup (82.9% versus 91.4%) but higher pass^64 (21.4% versus 15.0%). Its stronger final task performance

Table 6: Invalid actions after warmup (Qwen3-4B). Values are task-mean shares of all actions (%) across Seen and Unseen.

Warmup \ Type	ALFWorld			ScienceWorld		
	Nonrep.	Repeat	Consec.	Nonrep.	Repeat	Consec.
SFT	3.05	10.10	5.90	16.23	24.91	13.68
SFT-RS	3.12	10.24	5.90	16.11	23.79	12.88
GRPO	3.52	6.78	5.25	29.05	23.15	8.01
OPW (Ours)	3.14	3.19	0.53	30.62	19.17	4.02

Table 7: Task performance for fixed and resampled trajectories (ALFWorld 4B, Seen/Unseen).

Metric	Fixed	OPW (Ours)
Early gain (pp)	38.6 / 29.2	41.8 / 31.4
Average (%)	81.1 / 81.4	84.7 / 86.6
Final (%)	97.4 / 95.1	96.1 / 96.7

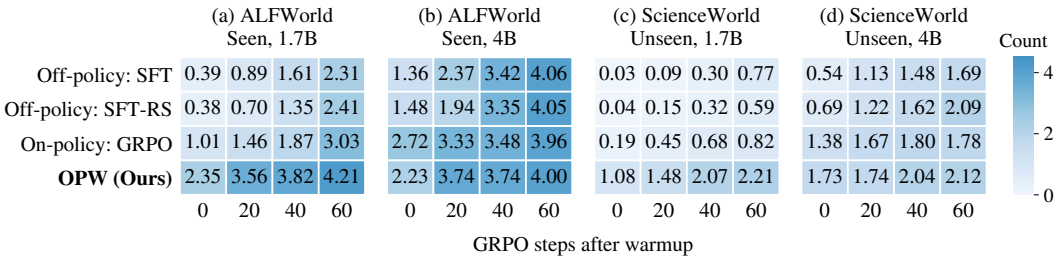

Figure 3: Distinct successful action sequences among eight rollouts per task during GRPO.

therefore does not require the broadest initial success coverage. We next examine how warmup affects subsequent learning and the behavior generated during teacher-free GRPO.

5 More Analysis

Learning Efficiency and Training Cost. OPW improves subsequent learning even from lower warmup success. On ALFWorld 4B Seen, it starts below GRPO warmup (46.8% versus 56.0%) but gains 41.8 versus 11.4 percentage points in the first 20 GRPO steps after warmup. Its average success over all 60 steps is 84.7% versus 69.5% (Fig. 2). OPW has the highest average task performance across environments, model sizes, and splits, even on ScienceWorld 1.7B Unseen, where GRPO warmup yields a larger early gain. Warmup scores alone do not explain subsequent learning.

OPW exceeds 80% Seen success on ALFWorld 4B at 19.6 recorded training GPU-hours including teacher supervision (Appendix C), versus at least 29.6 for other methods. Only OPW exceeds 90% within 100 steps. Over 100 steps, OPW costs less than GRPO warmup at 4B but more at 1.7B.

Repeated Invalid Actions. OPW's main reduction in ALFWorld action errors concerns repeated invalid commands. We classify validity by membership in the current valid-action list and repetition by earlier use in the trajectory. Unparseable responses are invalid. Repeated invalid commands account for 3.19% of actions after OPW, versus 10.10% after SFT and 6.78% after GRPO warmup, while non-repeated invalid actions remain near 3% (Table 6). In ScienceWorld, OPW reduces repeated errors but produces more non-repeated invalid actions than SFT. This list-based measure can also mark accepted navigation aliases as invalid. Similar fractions of SFT and OPW trajectories contain an invalid action on ALFWorld Seen (50.9% and 50.0%), despite different levels of repetition. Appendix H illustrates these errors and valid repetition in two complete action sequences.

To test the contribution of supervision at these errors, we remove teacher supervision at repeated invalid actions. Early GRPO gain on ALFWorld 4B Seen falls to 34.6 percentage points, compared with 42.0 under random removal of a similar number of action tokens (Table 18). Warmup action validity remains similar. Supervision at repeated errors contributes to subsequent learning.

Teacher Supervision on Student-Generated Trajectories. Teacher supervision provides a learning signal on unsuccessful student trajectories. In all-failure groups on ALFWorld 4B, one optimizer update from Base raises mean teacher-target log probability at teacher–student disagreements by 0.183 with OPD, versus 0.128 with SFT and 0.004 with GRPO (Appendix D). Compared with reusing initial student trajectories, resampling improves average success during GRPO by 3.6 per-

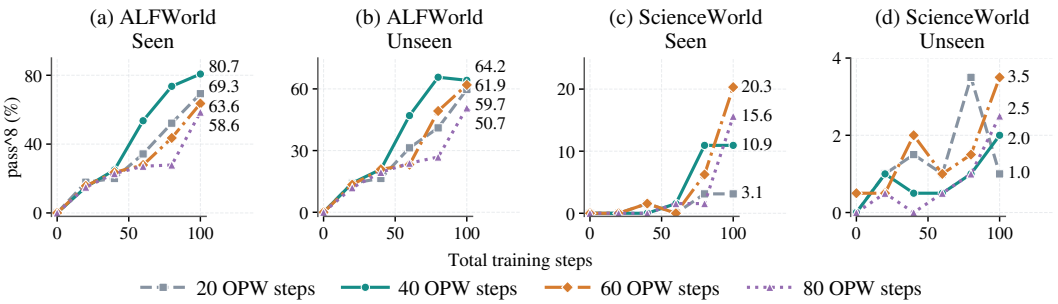

Figure 4: Repeated task success (pass^8) across OPW durations (Qwen3-1.7B).

centage points on Seen and 5.2 on Unseen (Table 7). Fixed trajectories still yield higher final Seen success, so resampling mainly benefits early learning and average success during GRPO.

Task Value and Retention of Teacher Corrections. Teacher substitution raises continuation success from 26.48% to 29.70% at 311 selected action-token disagreements (Appendix E.1). Base continues both branches. Over 40 teacher-free GRPO steps, teacher-target probability rises from 39.3% to 47.2% at these positions but changes little on a broader disagreement set, from 43.8% to 43.2% (Table 19). The broader set shows retention, with further gains at selected action positions.

A complementary evaluation (Appendix E) samples complete actions from each policy at common histories, with Base taking all subsequent actions. At 447 active positions from 63 tasks, continuation success increases from 45.00% to 46.75% over 40 GRPO steps after OPW, compared with 45.93% to 46.47% after GRPO warmup. OPW again improves more from a lower starting score. With Base fixed, these gains measure action selection at the evaluated histories.

Diversity of Successful Action Sequences. OPW samples more distinct successful action sequences early in GRPO across both environments and model sizes (Fig. 3). After sequence normalization (Appendix F), its mean count among eight rollouts per task rises from 2.23 to 4.00 over 60 GRPO steps on ALFWorld 4B Seen. Failure-only sequences fall from 2.95 to 0.31, accounting for the lower total sequence count. Among four selected successful trajectories on common tasks with enough successes under every method and checkpoint, OPW's expected distinct-sequence count falls from 2.54 to 2.14 on ALFWorld 4B but rises from 1.83 to 2.32 on ScienceWorld 4B. More successful sequences in eight rollouts need not mean more varied successful trajectories.

Choosing the Warmup Duration. Teacher supervision can improve repeated task success after most actions become valid. On ALFWorld 1.7B, extending OPW from 20 to 40 steps raises validity only from 89.4% to 91.8%, while pass^8 more than doubles from 10.8% to 22.0%. Fig. 4 compares 20, 40, 60, and 80 OPW steps followed by GRPO to 100 total steps, with varying seeds. The ALFWorld 20/40 comparison branches from the same warmup run. On Seen tasks, longer warmup raises final pass^8 from 69.3% to 80.7% and moves the first evaluation reaching 80% success from total step 75 to 60. In a separate ScienceWorld comparison from one 40-step OPW checkpoint, switching immediately gives higher Unseen pass^8 than adding 20 OPW steps (4.5% versus 3.5%), despite a lower mean task score. Choose warmup duration by subsequent validation on the required task outcome. Additional duration comparisons and warmup diagnostics are provided in Appendix G.

6 Conclusion

We study On-Policy Warmup (OPW), which uses teacher supervision on student-generated trajectories before RLVR. Our main comparisons in ALFWorld and ScienceWorld show the On-Policy Acceleration Phenomenon, with OPW reaching high performance earlier and achieving higher average and final task performance. Our theory connects on-policy reverse-KL distillation to trajectory-level distribution matching, initial verifier success, and reward-discovery complexity under its stated conditions. Empirical analyses show that teacher-target probability advantages persist during teacher-free GRPO, alongside improved action selection and repeated task success. Together, these findings support OPW as an effective warmup for agentic RLVR.

AI USE STATEMENT

AI assistants were used to refine the language and improve the clarity of the manuscript. All scientific claims, theoretical analyses, experimental results, and references were reviewed and verified by the authors, who take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Section 3 defines OPW and connects population reverse-KL distillation to initial verifier success and reward discovery, with assumptions and complete proofs in Appendix A. Section 4 specifies the models, benchmarks, baselines, and evaluation metrics. Appendix B details task partitions, agent interaction, the teacher top-1 cross-entropy implementation, optimization settings, rollout sampling, and metric estimators. The analyses in Section 5 are supported by protocols for teacher-target probability measurements, trajectory resampling, and supervision removal in Appendix D. Appendix E describes token and complete-action interventions at common histories, including position selection, continuation policies, aggregation, and measurement of teacher-target preferences during teacher-free GRPO. Appendix F defines action-error and successful action-sequence measurements. Appendices C and G document additional training-length and seed comparisons, training-cost accounting including teacher supervision, and warmup-duration allocations. Appendix H provides complete action traces with selected environment feedback. The supplementary code contains the modified verl training framework, OPW configurations, environment interfaces, and task-split manifests.

REFERENCES

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

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. arXiv preprint arXiv:2506.07982, 2025. URL https://arxiv.org/abs/2506.07982.

Fan Chen, Audrey Huang, Noah Golowich, Sadhika Malladi, Adam Block, Jordan T. Ash, Akshay Krishnamurthy, and Dylan J. Foster. The Coverage Principle: How Pre-Training Enables Post-Training. arXiv preprint arXiv:2510.15020, 2025. URL https://arxiv.org/abs/2510.15020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code. arXiv preprint arXiv:2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding,

Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948, 2025. URL https://arxiv.org/abs/2501.12948v1.

Shuai Dong, Yongfu Zhu, Yuqi Xu, Weichu Xie, Liuwenpu, Ziyue Wang, Kaiwen Tuo, Congcong Wang, Siyuan Wang, Wenqi Shao, Shuai Yang, Ji Zhao, Caoyuan Ma, Wenzheng Chang, Taiqiang Wu, Xinlei Yu, Hongrui Wu, Xiaoxuan He, Fangke Chen, Dianyi Wang, Kanghui Tian, Sirry Chen, Xingyu Liu, Xiangnan Wu, Jiawei Guo, Haowen Hou, LingHan Chen, Zhongyu Wei, and Jiaqi Wang. RL Starts before RL: On Policy Distillation for Better Reinforcement Learning. arXiv preprint arXiv:2609.28145, 2026. URL https://arxiv.org/abs/2609.28145.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. Quagmires in SFT-RL Post-Training: When High SFT Scores Mislead and What to Use Instead. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 56876–56918, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/5d6ae8ba43ecb378030753c4408ef9bd-Paper-Conference.pdf.

Boyan Li, Bingsen Chen, Chenghao Yang, Ping Nie, Chen Zhao, and Xi Ye. Sequential Beats Joint: On the Interplay between On-Policy Distillation and RLVR. arXiv preprint arXiv:2609.04108, 2026a. URL https://arxiv.org/abs/2609.04108.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, and Jinqiao Wang. On-Policy Distillation with Curriculum Turn-level Guidance for Multi-turn Agents. arXiv preprint arXiv:2606.15912, 2026b. URL https://arxiv.org/abs/2606.15912.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. ReFT: Reasoning with Reinforced Fine-Tuning. arXiv preprint arXiv:2401.08967, 2024. URL https://arxiv.org/abs/2401.08967.

Sadhika Malladi, Samy Jelassi, Dylan Foster, Jordan T. Ash, and Akshay Krishnamurthy. TailSFT: Filtered Fine-Tuning Improves Post-Training Performance. arXiv preprint arXiv:2608.25756, 2026. URL https://arxiv.org/abs/2608.25756.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/ross11a.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300, 2024. URL https://arxiv.org/abs/2402.03300.

Safal Shrestha, Minwu Kim, Aadim Nepal, Anubhav Shrestha, and Keith W. Ross. Warm Up Before You Train: Unlocking General Reasoning in Resource-Constrained Settings. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14380–14401, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.727. URL https://aclanthology.org/2025.emnlp-main.727/.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.03768.

Tim van Erven and Peter Harremoës. Rényi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014. doi: 10.1109/TIT.2014.2320500. URL https://doi.org/10.1109/TIT.2014.2320500.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your Agent Smarter than a 5th Grader? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, Abu Dhabi, United Arab Emirates, dec 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.775. URL https://aclanthology.org/2022.emnlp-main.775/.

Sudong Wang, Weiquan Huang, Xiaomin Yu, Zuhao Yang, Hehai Lin, Keming Wu, Chaojun Xiao, Chen Chen, Wenxuan Wang, Beier Zhu, Yunjian Zhang, and Chengwei Qin. Beyond SFT-to-RL: Pre-alignment via Black-Box On-Policy Distillation for Multimodal RL. arXiv preprint arXiv:2604.28123, 2026a. URL https://arxiv.org/abs/2604.28123.

Yinjie Wang, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. OpenClaw-RL: Train Any Agent Simply by Talking. arXiv preprint arXiv:2603.10165, 2026b. URL https://arxiv.org/abs/2603.10165.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. SEED: Self-Evolving On-Policy Distillation for Agentic Reinforcement Learning. arXiv preprint arXiv:2607.14777, 2026. URL https://arxiv.org/abs/2607.14777.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to Reason under Off-Policy Guidance. arXiv preprint arXiv:2504.14945, 2025. URL https://arxiv.org/abs/2504.14945.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv preprint arXiv:2406.12045, 2024. URL https://arxiv.org/abs/2406.12045.

Qinglin Ye, Zhiyuan Gu, Jingjie Xia, Yiheng Zhang, Kaiyan Zhao, Shunchao Zheng, Yuhang Mu, Wenchao Du, and Yiming Wang. OPDSearch+: On-Policy Distillation with RL Refinement for Search-Augmented Reasoning. arXiv preprint arXiv:2608.24310, 2026. URL https://arxiv.org/abs/2608.24310.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. Agent-Tuning: Enabling Generalized Agent Abilities for LLMs. arXiv preprint arXiv:2310.12823, 2023. URL https://arxiv.org/abs/2310.12823.

Dylan Zhang, Yufeng Xu, Haojin Wang, Qingzhi Chen, and Hao Peng. Good SFT Optimizes for SFT, Better SFT Prepares for Reinforcement Learning. arXiv preprint arXiv:2602.01058, 2026. URL https://arxiv.org/abs/2602.01058.

APPENDIX CONTENTS

A Proofs and Theoretical Details **14**

A.1 Trajectory KL and Success Transfer 14

A.2 Reward Discovery and Reward-Informative Groups 16

B Training and Evaluation Protocol **16**

B.1 Data, Models, and Agent Interaction 16

B.2 Training Objectives and Optimization 17

B.3 Evaluation and Statistical Summaries 17

C Additional Warmup Comparisons **18**

C.1 Full Warmup Results 18

C.2 Sensitivity to Training Length and Seed 18

C.3 Training Computation and Target-Reaching Cost 20

D Teacher Supervision on Student-Generated Trajectories **21**

D.1 Teacher-Target Probabilities after One Optimizer Step 21

D.2 Resampling Student Trajectories 21

D.3 Removing Supervision at Repeated Invalid Actions 22

E From Teacher Targets to Action Value **23**

E.1 Teacher Alternatives, Task Value, and Retention 23

E.2 Complete-Action Value at Common Histories 23

F Behavior after Warmup and during GRPO **24**

F.1 Execution Errors and Repetition 24

F.2 Successful and Failure-Only Action Sequences 25

G Warmup Duration and Repeated Task Success **25**

G.1 What Continues to Improve during Warmup? 25

G.2 Learning Speed and Repeated Task Success 26

H Long-Trajectory Examples **27**

I Limitations and Scope **29**

A PROOFS AND THEORETICAL DETAILS

A.1 TRAJECTORY KL AND SUCCESS TRANSFER

Fix a task x and suppress its index. The student π and teacher π_T induce distributions P_π and P_T over horizon H, with the same initial distribution ρ and transition kernels $K_t(\cdot \mid s_t, a_t)$. States contain full interaction histories, and d_t^π is the student's state distribution. Padding uses an absorbing

state and a shared deterministic null action with zero KL. We use the cumulative objective in Eq. 1 and natural logarithms, with $0 \log(0/q) = 0$ and $p \log(p/0) = +\infty$ for $p > 0$.

Proof of Proposition 3.1. Write a trajectory as $\tau = (s_1, a_1, \ldots, s_H, a_H, s_{H+1})$. Using probability-mass notation, its probability under the student factorizes as

$$P_\pi(\tau) = \rho(s_1) \prod_{t=1}^{H} \pi(a_t \mid s_t) K_t(s_{t+1} \mid s_t, a_t),$$

and the teacher trajectory probability $P_T(\tau)$ has the same factorization with π replaced by π_T.

First suppose that $P_\pi \ll P_T$. The initial-state and environment-transition factors are identical under the two policies and therefore cancel in the likelihood ratio. Consequently, P_π-almost surely,

$$\log \frac{P_\pi(\tau)}{P_T(\tau)} = \sum_{t=1}^{H} \log \frac{\pi(a_t \mid s_t)}{\pi_T(a_t \mid s_t)}.$$

Taking expectation under P_π and conditioning on s_t yields

$$\begin{aligned} D_{\mathrm{KL}}(P_\pi \| P_T) &= \mathbb{E}_{\tau \sim P_\pi} \left[\sum_{t=1}^{H} \log \frac{\pi(a_t \mid s_t)}{\pi_T(a_t \mid s_t)} \right] \\ &= \sum_{t=1}^{H} \mathbb{E}_{s_t \sim d_t^\pi} \left[\mathbb{E}_{a_t \sim \pi(\cdot \mid s_t)} \left[\log \frac{\pi(a_t \mid s_t)}{\pi_T(a_t \mid s_t)} \right] \right] \\ &= \sum_{t=1}^{H} \mathbb{E}_{s_t \sim d_t^\pi} \left[D_{\mathrm{KL}} \left(\pi(\cdot \mid s_t) \,\|\, \pi_T(\cdot \mid s_t) \right) \right] \\ &= \mathcal{L}_{\mathrm{OPD}}(\pi). \end{aligned}$$

If $P_\pi \not\ll P_T$, the shared initial distribution and transition kernels imply an action-support mismatch on a set of student-visited states with positive probability. Both the expected conditional and trajectory KL divergences are infinite. The trajectory identity therefore holds under these conventions. □

Proof of Theorem 3.2. Let $V(\tau) = R_x(\tau)$ be the shared binary verifier. Its success probabilities are p_π and p_T as defined in the main text. For Bernoulli distributions, $\mathrm{KL}(p\|q) = p \log(p/q) + (1 - p) \log((1-p)/(1-q))$, using the same endpoint conventions.

Applying V maps P_π and P_T to $\mathrm{Bern}(p_\pi)$ and $\mathrm{Bern}(p_T)$. The data-processing inequality (van Erven & Harremoës, 2014, Theorem 9) therefore gives

$$\begin{aligned} \mathrm{KL}(p_\pi \| p_T) &= D_{\mathrm{KL}} \left(\mathrm{Bern}(p_\pi) \,\|\, \mathrm{Bern}(p_T) \right) \\ &\le D_{\mathrm{KL}}(P_\pi \| P_T) \\ &= \mathcal{L}_{\mathrm{OPD}}(\pi) \\ &\le \varepsilon, \end{aligned}$$

where the equality in the third line follows from Proposition 3.1.

For Bernoulli distributions, Pinsker's inequality (van Erven & Harremoës, 2014, Theorem 31) yields

$$2|p_\pi - p_T|^2 \le \mathrm{KL}(p_\pi \| p_T) \le \varepsilon.$$

Taking square roots gives the gap bound in Eq. 2. Rearranging with $p_\pi \ge 0$ gives the lower bound. □

Offline state coverage. Evaluating the same local reverse KL under offline distributions μ_t requires $d_t^\pi \ll \mu_t$ and a finite density-ratio bound $\frac{dd_t^\pi}{d\mu_t} \le C$ to recover the bound. Nonnegativity of local KL then gives $\mathcal{L}_{\mathrm{OPD}}(\pi) \le C \mathcal{L}_\mu(\pi)$. No finite C exists if absolute continuity fails.

Normalization and scope. For a fixed horizon, $\overline{\mathcal{L}}_{\mathrm{OPD}} = \mathcal{L}_{\mathrm{OPD}}/H$ gives $D_{\mathrm{KL}}(P_\pi \| P_T) = H \overline{\mathcal{L}}_{\mathrm{OPD}}$. Thus, $\overline{\mathcal{L}}_{\mathrm{OPD}} \le \delta$ implies $p_\pi \ge [p_T - \sqrt{H\delta/2}]_+$. Normalization by realized trajectory length need not preserve this identity. The guarantee uses the population objective for the specified task, environment, verifier, and sampling policies. Certifying this premise from empirical estimates or under new decoding rules or evaluation distributions requires further error bounds.

A.2 Reward Discovery and Reward-Informative Groups

Proof of Corollary 3.3. For independent rollouts from the fixed policy π_{w}, the first-success time T_{succ} is geometric with parameter p_{w}. When $p_{\mathrm{w}} \geq \ell > 0$,

$$\Pr(T_{\mathrm{succ}} > N) = (1 - p_{\mathrm{w}})^N \leq (1 - \ell)^N \leq e^{-N\ell}, \qquad \mathbb{E}[T_{\mathrm{succ}}] = \frac{1}{p_{\mathrm{w}}} \leq \frac{1}{\ell}.$$

Taking complements gives the success-probability bound. For $\delta \in (0, 1)$, choosing $N \geq \lceil \log(1/\delta)/\ell \rceil$ makes the failure probability at most δ. □

Binary reward groups. Under binary rewards, a group of $G \geq 2$ independent rollouts has all-success probability p^G, all-failure probability $(1-p)^G$, and mixed-group probability

$$m_G(p) = 1 - p^G - (1-p)^G, \qquad m'_G(p) = G\big[(1-p)^{G-1} - p^{G-1}\big]. \tag{6}$$

This probability increases below $p = 0.5$ and decreases above it. The lower bound $p_{\mathrm{w}} \geq \ell$ alone does not imply an increase in mixed groups, since it does not also ensure $p_{\mathrm{w}} \leq 0.5$. All-success groups indicate reliable completion even though their direct GRPO advantages are zero. Empirical group fractions average these per-task quantities over tasks with different success probabilities. In ScienceWorld, partial-credit rewards can differ even without complete success.

B Training and Evaluation Protocol

B.1 Data, Models, and Agent Interaction

The students and teacher use released Qwen3 weights (Yang et al., 2025) without additional agent-task training. Table 8 lists model sizes and shared settings. Chat templates disable thinking, while task prompts request `<reasoning>` and `<action>` fields. ALFWorld supplies the current observation and admissible actions. ScienceWorld supplies the task, observation, action templates, and objects. Both include the two most recent observation–action pairs when available.

Table 8: Training and evaluation settings for the main warmup comparisons.

Parameter	Value
Models and training schedule	
Student	Qwen3-1.7B / Qwen3-4B
Teacher	Qwen3-32B
Environments	**ALFWorld / ScienceWorld**
Main training length	40 warmup + 60 GRPO steps
Shorter warmup comparison	20 warmup + 80 GRPO steps
Teacher during GRPO	Not used
Objectives	
SFT / SFT-RS	Token cross-entropy on all / successful teacher trajectories
OPD	Cross-entropy on teacher top-1 targets along student trajectories
GRPO	Group-relative log-probability objective
GRPO advantage normalization	Within-group mean and sample standard deviation, $\varepsilon = 10^{-6}$
Probability-ratio clipping	Not used
Additional KL / entropy coefficient	0 / 0
Optimization	
Optimizer	AdamW (Loshchilov & Hutter, 2019), $\beta = (0.9, 0.999)$
Learning rate / schedule	10^{-6} / constant
Weight decay / gradient clipping	0.01 / norm 1.0
Precision	bfloat16
Loss within a microbatch	Mean over response tokens

Table 8 (continued)

Parameter	Value
Batching and step counts	
OPD / GRPO batch	16 tasks $\times$ 8 rollouts
OPD / GRPO minibatch	8 tasks $\times$ 8 rollouts
Passes over each collected batch	1
SFT / SFT-RS global batch	128 teacher trajectories
Optimizer steps per training step	SFT: 1, OPD / GRPO: 2
Agent interaction	
Prompt / trajectory response limit	4,096 / 24,576 tokens
Response limit per turn	512 tokens
Maximum interaction length	30 turns
Recent interactions in the prompt	2 observation–action pairs
Chat-template thinking mode	Disabled
Sampling and evaluation	
Training temperature / top-p	1.0 / 1.0
Evaluation temperature / top-p	0.7 / 0.95
Learning-curve rollouts per task	8
Rollouts per task for warmup success profile	64
Compute	
GPUs per training run	8
Student / teacher placement	Shared eight-GPU pool for OPD
Sequence parallel size	4

ALFWorld reserves 64 of its 3,553 training tasks for analysis, leaving disjoint warmup and GRPO pools of 1,744 and 1,745 tasks. ScienceWorld reserves 64 of 2,294 tasks and uses two training pools of 1,115. Both partitions use seed 20260722 and task-instance identities, allowing task types to overlap. ALFWorld Seen and Unseen use separate benchmark instances. ScienceWorld uses its reserved 64 tasks as Seen and the original 200 test tasks as Unseen. Complete-action evaluation uses 32 tasks per benchmark test split, excluding ALFWorld's 64 reserved tasks.

B.2 Training Objectives and Optimization

All warmup methods start from the same Base model. GRPO after warmup uses the second task pool and no teacher supervision. The GRPO warmup baseline also switches pools and loads its warmup weights into a new optimizer, as shown in Figs. 1 and 2. Direct GRPO instead uses the second pool for all 100 steps without restarting, while direct OPD uses the warmup pool throughout. Extended ALFWorld comparisons run 100 or 280 GRPO steps after 40 warmup steps.

In ALFWorld 4B, SFT processes 5,120 teacher trajectories once over 40 steps, using 9,662,640 supervised response tokens. SFT-RS retains 2,375 successful trajectories and repeats them in a fixed order, processing 7.07 million tokens over 40 steps. OPW processes 9.76 million student-generated response tokens. All loss counts exclude prompts, environment observations, and padding.

In OPD and GRPO, each microbatch's response-token mean is weighted by its share of the mini-batch's trajectories before gradient accumulation. In ALFWorld 4B SFT, each data-parallel rank accumulates 64 one-trajectory microbatches. Their token-normalized gradients are summed before clipping, with division by the microbatch count applied only to the logged loss. Axes count training steps, each covering one batch and its optimization. A 40-step SFT warmup followed by 60 GRPO steps contains 160 optimizer steps, compared with 200 for the OPW and GRPO-warmup paths.

B.3 Evaluation and Statistical Summaries

Table 9 lists the trajectory collections used for each analysis. Learning curves average eight rollouts per task and weight tasks equally. All eight rollouts come from the same trained checkpoint. Curves connect recorded evaluations without smoothing. A checkpoint can have several independently sampled evaluations, so each comparison uses its own initial evaluation.

Table 9: Trajectory collections used for evaluation.

Collection	Rollouts/task	Measurements
Original evaluations	8	Learning curves, final task performance and pass^8
New warmup evaluations	64	Success coverage and repeated task success before GRPO
First eight of the new collection	8	Warmup action composition and standard OPW's initial score in supervision removal
ALFWorld warmup diagnostics	32	Validity and repeated task success on 64 reserved tasks
ScienceWorld warmup diagnostics	8	Complete success and partial-credit reward contrast on 64 Seen tasks
Duration comparisons	8	Final performance under different OPW + GRPO allocations

Early gain and average task performance. Let R_t denote the task mean after t GRPO steps following warmup, measured as task success in ALFWorld or task score in ScienceWorld on a 0–100 scale. Early gain is $\Delta R_{20} = R_{20} - R_0$. For recorded evaluations at $0 = t_0 < \cdots < t_m = T$, average task performance during GRPO is

$$\bar{R}_T = \frac{1}{T}\sum_{j=1}^{m}(t_j - t_{j-1})\frac{R_{t_j} + R_{t_{j-1}}}{2}. \tag{7}$$

This averages the recorded curve by the trapezoidal rule, with GRPO steps counted from the end of warmup. Full-training and duration comparisons count both stages. Target-reaching times use the first scheduled evaluation meeting the stated success target, without interpolation.

Task-level success probabilities. For independent rollouts on a task x with success probability p_x, pass@$k(x) = 1-(1-p_x)^k$ measures at least one success, while pass^$k(x) = p_x^k$ measures success in all k rollouts. Given n rollouts with c_x successes, we use (Chen et al., 2021; Yao et al., 2024)

$$\widehat{\text{pass@}k}(x) = 1 - \frac{\binom{n-c_x}{k}}{\binom{n}{k}}, \qquad \widehat{\text{pass\^{}k}}(x) = \frac{\binom{c_x}{k}}{\binom{n}{k}}, \tag{8}$$

then average the estimates over tasks. A binomial coefficient is zero when its upper argument is below k. Table 5 uses $n = 64$ and $k \in \{8, 16, 32, 64\}$. At $k = n$, these are the fractions of tasks solved at least once and in every rollout. ScienceWorld complete success requires a terminal trajectory with normalized task reward one, excluding partial progress.

C Additional Warmup Comparisons

C.1 Full Warmup Results

OPW has the highest overall repeated task success in both environments after 40 warmup and 60 GRPO steps. Table 10 reports pass^8 by task type for the methods in Table 3.

Across environments, model sizes, and splits, OPW has the highest average task performance over 60 GRPO steps (Table 11). GRPO warmup yields a larger early gain on ScienceWorld 1.7B Unseen.

C.2 Sensitivity to Training Length and Seed

Shorter warmup. With 20 warmup and 80 GRPO steps, OPW retains the highest final ALFWorld 4B success and pass^8 on both splits (Table 12). Fig. 5 uses the main comparison's task sets and Base evaluations, counting total steps. Direct OPD is evaluated every 20 steps through step 100.

Longer GRPO training. After 40 warmup and 100 GRPO steps, SFT-RS approaches OPW's final Seen success (95.2% versus 96.2%), but its average GRPO success remains lower (71.0% versus 89.4%). Unseen averages are 68.3% and 90.9%, respectively (Table 13). Fig. 6 extends GRPO to 280 steps and shows sustained high success after OPW.

Table 10: Final repeated task success (pass^8, %) after 40 warmup and 60 GRPO steps (Qwen3-4B).

Split	Warmup \ Task type	ALFWorld							ScienceWorld						
		Pick	Clean	Heat	Cool	Exam	Pick2	All	Find	Life	Gene.	Mix	Prop.	Phys.	All
Seen	Off-policy: SFT	82.9	51.9	56.2	36.0	38.5	50.0	55.7	0.0	**25.0**	0.0	0.0	0.0	9.5	4.7
	Off-policy: SFT-RS	91.4	70.4	50.0	44.0	69.2	54.2	65.7	12.5	0.0	0.0	0.0	7.4	4.8	6.2
	On-policy: GRPO	94.3	37.0	43.8	36.0	23.1	58.3	54.3	12.5	0.0	0.0	0.0	0.0	**19.0**	7.8
	OPW (Ours)	**97.1**	**92.6**	**81.2**	**80.0**	**84.6**	**83.3**	**87.9**	**62.5**	**25.0**	**50.0**	0.0	**63.0**	**19.0**	**43.8**
Unseen	Off-policy: SFT	75.0	45.2	52.2	52.4	27.8	52.9	51.5	3.8	2.9	12.0	0.0	1.5	**26.3**	6.0
	Off-policy: SFT-RS	**87.5**	64.5	65.2	76.2	55.6	58.8	68.7	1.9	8.6	4.0	0.0	3.1	10.5	4.5
	On-policy: GRPO	70.8	29.0	34.8	28.6	22.2	17.6	35.1	**7.7**	0.0	12.0	0.0	1.5	15.8	5.5
	OPW (Ours)	**87.5**	**100.0**	**82.6**	**95.2**	**77.8**	**82.4**	**88.8**	3.8	**14.3**	**16.0**	0.0	**4.6**	15.8	**8.5**

Table 11: Early gains and average task performance during 60 GRPO steps after warmup.

Warmup \ Benchmark	ALFWorld				ScienceWorld			
	Seen		Unseen		Seen		Unseen	
	ΔR_{20}	$\bar{R}_{60}$	ΔR_{20}	$\bar{R}_{60}$	ΔR_{20}	$\bar{R}_{60}$	ΔR_{20}	$\bar{R}_{60}$
Qwen3-1.7B								
Off-policy: SFT	+9.7	23.6	+7.5	25.8	+0.5	13.6	+2.2	18.7
Off-policy: SFT-RS	+6.8	20.7	+5.0	21.1	+3.2	12.9	+4.6	19.4
On-policy: GRPO	+7.8	37.2	+6.8	37.8	+13.7	28.1	**+14.0**	34.1
OPW (Ours)	**+26.9**	**73.7**	**+19.6**	**71.9**	**+14.2**	**62.0**	+5.6	**58.1**
Qwen3-4B								
Off-policy: SFT	+14.1	56.5	+15.0	53.8	+15.9	43.5	+17.1	51.4
Off-policy: SFT-RS	+8.7	56.3	+7.3	54.5	**+22.8**	47.4	**+19.5**	53.1
On-policy: GRPO	+11.4	69.5	+11.9	65.0	+0.9	57.0	+0.2	59.4
OPW (Ours)	**+41.8**	**84.7**	**+31.4**	**86.6**	+22.6	**78.9**	+3.5	**65.2**

In Fig. 6, test panels (a,b) use unsmoothed evaluations every 20 steps, with eight rollouts on 140 Seen and 134 Unseen tasks. Training panels (c)–(f) use 16 tasks with eight rollouts per step. Validity pools recorded actions, reward-group fractions average task groups, and interaction length counts action turns. Training curves use exponential moving averages with coefficient 0.2, while interval means weight raw steps equally. Direct OPD remains on the warmup pool over the same total steps.

OPW's all-success group fraction rises from 44.7% to 72.8% between the first two 20-step GRPO intervals. These groups indicate reliable completion but have zero direct GRPO advantages. Over the final 100 steps, OPW averages 97.3% training success with 1.6% all-failure groups, compared with 86.7% and 4.0% after GRPO warmup.

Second training seed. Both ALFWorld 4B seeds use 40 warmup and 100 GRPO steps, with eight rollouts per task on both test splits. Table 14 reports OPW minus GRPO warmup for initial success R_0, early gain ΔR_{20}, and average success $\bar{R}_{100}$. Under both seeds, OPW starts lower on Seen and higher on Unseen, then achieves larger early gains and higher averages on both splits.

Table 12: Final task performance and repeated task success for 20-step warmup (ALFWorld 4B).

Warmup	Seen		Unseen	
	pass@1	pass^8	pass@1	pass^8
SFT	91.2	72.1	83.3	63.4
SFT-RS	89.1	68.6	83.5	59.0
GRPO	72.3	35.0	69.9	26.1
OPW (Ours)	**95.7**	**85.0**	**93.3**	**76.9**

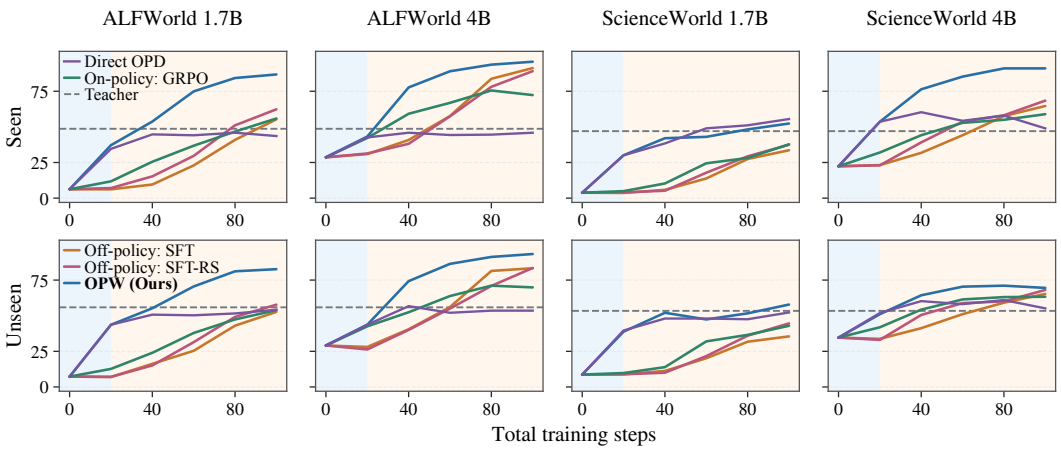

Figure 5: Task performance during 20 warmup and 80 GRPO steps.

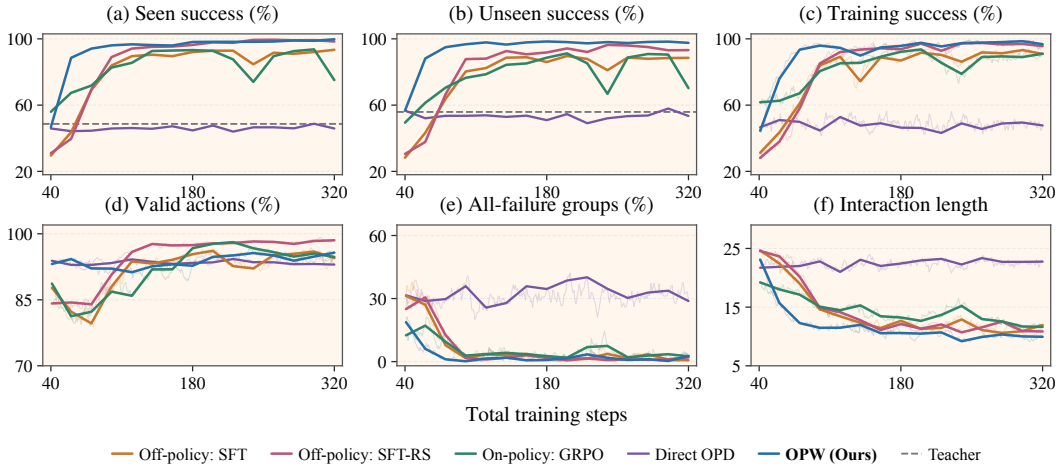

Figure 6: Task performance (a,b) and training dynamics (c–f) during extended training.

C.3 Training Computation and Target-Reaching Cost

We multiply recorded training times by eight GPUs for each path with 40 warmup and 60 GRPO steps in Fig. 2. Timers include trajectory generation, environment interaction, teacher scoring, and student optimization. SFT and SFT-RS include 10.9 GPU-hours for collecting 5,120 shared teacher trajectories before filtering. Separate evaluations, research diagnostics, checkpoint saving, and startup are excluded. Student training and teacher scoring share the eight-GPU allocation, with fully sharded data parallelism, sequence-parallel size four, and data-parallel size two. Supplementary runs use L20X GPUs. Unknown GPU models in earlier runs prevent hardware normalization.

Table 13: Average and final task performance over 100 GRPO steps after warmup (ALFWorld 4B).

Warmup \ Task set	Seen		Unseen	
	$\bar{R}_{100}$	R_{100}	$\bar{R}_{100}$	R_{100}
Off-policy: SFT	69.3	90.4	65.6	88.6
Off-policy: SFT-RS	71.0	95.2	68.3	92.6
On-policy: GRPO	76.3	92.7	70.8	84.3
OPW (Ours)	**89.4**	**96.2**	**90.9**	**96.5**

Table 14: OPW minus GRPO warmup under two training seeds (percentage points).

Seed	Seen			Unseen		
	R_0	ΔR_{20}	$\bar{R}_{100}$	R_0	ΔR_{20}	$\bar{R}_{100}$
1	-9.2	+30.4	+13.0	+7.3	+19.5	+20.1
2	-0.8	+17.2	+6.6	+6.4	+9.0	+9.5

Table 15: Training cost and cost to reach target Seen success in ALFWorld (GPU-hours).

Warmup \ Cost	100 training steps				To reach	
	Data collection	**Warmup**	GRPO	Total	$> 80\%$	$> 90\%$
Qwen3-1.7B						
Off-policy: SFT	10.9	2.5	14.1	27.5	–	–
Off-policy: SFT-RS	10.9	1.9	13.9	26.7	–	–
On-policy: GRPO	0.0	9.3	12.9	**22.2**	–	–
OPW (Ours)	0.0	12.8	11.1	23.9	**20.3**	–
Qwen3-4B						
Off-policy: SFT	10.9	4.8	15.7	31.4	31.4	–
Off-policy: SFT-RS	10.9	3.2	16.8	30.9	30.9	–
On-policy: GRPO	0.0	13.0	16.6	29.6	29.6	–
OPW (Ours)	0.0	14.9	12.4	**27.3**	**19.6**	**23.6**

–: target not reached within 100 steps.

OPW reduces total cost relative to GRPO warmup at 4B (27.3 versus 29.6 GPU-hours) but increases it at 1.7B (23.9 versus 22.2). Table 15 uses the first evaluation at total step 40, 60, 80, or 100 exceeding the target, without interpolation. At 4B, OPW reaches both targets at the same evaluations on Seen and Unseen. GRPO warmup exceeds 80% only on Seen. At 1.7B, only OPW exceeds 80%, at step 80 on Seen and step 100 on Unseen. Only 4B OPW exceeds 90% within 100 steps.

D Teacher Supervision on Student-Generated Trajectories

D.1 Teacher-Target Probabilities after One Optimizer Step

From Base, we take one AdamW step with learning rate 10^{-6}, using 128 student trajectories for OPD/GRPO or 128 teacher trajectories for SFT. These batches contain 329,646 and 257,068 response tokens, respectively. OPD applies $\ell_t = -\log \pi_\theta(v_t \mid h_t)$ at student token histories h_t, with teacher top-1 targets v_t. SFT uses this loss with teacher tokens on teacher trajectories. We score log-probability changes at common histories where teacher targets disagree with student tokens.

The batch has 16 tasks and 22,279 disagreements. Eight groups contain only failures, six have mixed rewards, and two contain only successes. Table 16 averages changes over all disagreements, 11,839 positions in 64 all-failure trajectories, and 4,455 positions with the lowest Base probability for the teacher target. OPD and GRPO use all 16 task groups in one minibatch, compared with eight per minibatch in main training. In all-failure groups, SFT changes probabilities through transfer from teacher trajectories, while OPD supervises the student histories directly. GRPO advantages are zero in these groups, but other groups can change their probabilities through shared parameters. Across the batch, disagreements occupy 6.8% of response positions but account for 97.7% of teacher-target loss and 95.2% of the positive log-probability change after OPD.

D.2 Resampling Student Trajectories

Both ALFWorld Qwen3-4B conditions use teacher top-1 cross-entropy for 40 warmup steps, followed by 60 teacher-free GRPO steps. The fixed version reuses initial-student trajectories while recomputing teacher targets and current student probabilities. OPW resamples trajectories as the

Table 16: Teacher-target log-probability changes after one optimizer step (ALFWorld 4B).

Objective	All disagreements	All-failure groups	Lowest-probability 20%
SFT	$+0.1279$	$+0.128$	$+0.290$
GRPO	-0.0004	$+0.004$	$+0.017$
OPD	$\mathbf{+0.1812}$	$\mathbf{+0.183}$	$\mathbf{+0.392}$

student changes. Evaluations use eight rollouts per task on 140 Seen and 134 Unseen tasks. Table 17 reports the evaluations underlying Table 7. Resampling improves early gains and average success, while fixed trajectories yield higher final Seen success. Resampling jointly changes histories, diversity, and task coverage.

Table 17: Task performance after warmup on fixed or resampled trajectories (ALFWorld 4B).

Task set	**Trajectories**	**GRPO steps after warmup**			
		0	20	40	60
Seen	Fixed	41.1	79.6	94.4	97.4
	OPW (Ours)	46.8	88.6	94.1	96.1
Unseen	Fixed	50.5	79.7	91.7	95.1
	OPW (Ours)	56.7	88.2	95.0	96.7

D.3 Removing Supervision at Repeated Invalid Actions

Two ALFWorld Qwen3-4B runs share OPW's initialization, tasks, teacher, batch size, seed, and optimizer for 40 warmup and 60 GRPO steps. Targeted removal zeros teacher-loss weights on action-content tokens when the action is invalid and has appeared earlier in the trajectory. Inputs, targets, and loss normalization stay unchanged. Random removal selects complete actions with the same token-length counts as repeated invalid actions in its current batch, allowing overlap. Branches generate their own trajectories, so total removal counts can differ. Targeted removal excludes 34,536 tokens in 6,489 actions (0.354% of response tokens), versus 32,132 tokens in 5,992 actions (0.330%) under random removal. Random removal includes 7,199 tokens in 1,474 repeated invalid actions.

Targeted and random removal yield similar warmup validity on Seen (93.8% versus 93.4%) and Unseen (93.5% versus 93.7%). Yet targeted removal produces smaller early GRPO gains on both splits (Table 18), supporting the contribution of supervision at repeated errors. All conditions use eight rollouts per task on 140 Seen and 134 Unseen tasks at 0, 20, 40, and 60 GRPO steps. OPW's initial evaluation uses the first eight of the new 64-rollout collection, explaining differences from the main summary. Later OPW evaluations use the original collections. The table reports early gain ΔR_{20} in percentage points, average success $\bar{R}_{60}$ from Eq. 7, and final success R_{60} in percent.

Table 18: GRPO learning after targeted and random supervision removal (ALFWorld 4B).

Warmup \ **Task set**	Seen			Unseen		
	ΔR_{20}	$\bar{R}_{60}$	R_{60}	ΔR_{20}	$\bar{R}_{60}$	R_{60}
OPW (Ours)	41.5	84.7	96.1	32.2	**86.5**	**96.7**
Random removal	**42.0**	**84.8**	**97.1**	**35.2**	82.9	94.3
Targeted removal	34.6	80.1	95.4	29.2	80.9	92.9

E From Teacher Targets to Action Value

E.1 Teacher Alternatives, Task Value, and Retention

In ALFWorld Qwen3-4B, we test 311 action-token alternatives from 97 tasks. We inspect eight Base rollouts per task from the first 128 distinct training tasks in trajectory order. Eligible positions have Base probability of at least 0.9 for the original token and a different teacher top-1 token, with both tokens containing letters or digits. We retain 21 positions from an eight-task analysis and add up to four distinct token histories per task, without using continuation outcomes or trained-policy probabilities for selection. At each position, we choose either token and let Base complete the action and remaining trajectory. Four continuations per choice yield 2,488 rollouts. We average continuations within positions, positions within tasks, and tasks equally. Teacher alternatives raise action validity from 67.44% to 82.22% and continuation success from 26.48% to 29.70%. On the 89 added tasks, continuation success still improves by 2.39 percentage points.

We track teacher-target probabilities at these action positions and on a separate set of 39,490 disagreements from 256 student trajectories across 32 tasks (Table 19). All checkpoints use the same token histories and teacher top-1 targets. Probabilities use the full softmax before temperature scaling or top-p truncation, averaged within tasks and then equally across tasks. Over 40 teacher-free GRPO steps, OPW's mean probability rises from 39.3% to 47.2% at the selected action positions. On the broader set, it remains near its warmup level (43.8% to 43.2%) and highest on all 32 tasks. Selected action positions thus show further gains, while the broader set shows retention. The table reports warmup probabilities P_0 in percent and changes $\Delta P_t = P_t - P_0$ in percentage points.

Table 19: Teacher-target probabilities during teacher-free GRPO (ALFWorld 4B).

Warmup \ Positions	All disagreements			Action positions		
	P_0	ΔP_{20}	ΔP_{40}	P_0	ΔP_{20}	ΔP_{40}
Off-policy: SFT	17.7	+0.3	+1.2	3.8	+2.8	+6.2
Off-policy: SFT-RS	17.9	−1.4	+1.0	3.7	+4.7	+14.8
On-policy: GRPO	17.7	+1.6	−0.9	12.5	+6.2	+8.3
OPW (Ours)	**43.8**	−0.2	−0.6	**39.3**	+6.1	+7.9

E.2 Complete-Action Value at Common Histories

We take one trajectory per warmup method from each of 32 Seen and 32 Unseen tasks and inspect positions after four and twelve actions. Checkpoints share 512 records, including 447 active positions from 63 tasks and 65 records from already-ended trajectories. Each policy samples four actions per active position. Base supplies four continuations per distinct position–action pair, reusing continuations for identical actions at their sampled frequencies. We average positions and sources within tasks, then weight tasks equally. Ended records retain their outcomes only in all-position success.

We replay actions in a reset game, checking observations and admissible actions. Base receives only the sampled `<action>` field. Actions and continuations use temperature 0.7, top-$p = 0.95$, a 1,024-token response limit, a 32,768-token context limit, and at most 30 interaction turns. Table 20 reports validity V_t and continuation success C_t (%) after t GRPO steps. Gains are in percentage points, with $\Delta C_{40} = C_{40} - C_0$.

At active positions, OPW starts with the highest validity (94.45%) but lower continuation success than GRPO warmup (45.00% versus 45.93%). Over 40 GRPO steps, its continuation success rises to 46.75%, compared with 46.47% after GRPO warmup. Including already-ended records gives the same ordering of gains, at 1.60 versus 0.49 points, with similar final means. Since Base supplies every continuation, these gains reflect improved action selection at the evaluated histories.

Table 20: Complete-action validity and continuation success under Base (ALFWorld 4B).

Warmup \ Metric	Active positions				All positions: success			
	Validity		Success		Seen		Unseen	
	V_0	V_{40}	C_0	C_{40}	C_0	ΔC_{40}	C_0	ΔC_{40}
Off-policy: SFT	92.57	93.68	44.01	45.16	41.89	+0.61	51.56	+1.20
Off-policy: SFT-RS	92.64	93.74	44.40	45.56	42.33	+0.42	51.78	+1.34
On-policy: GRPO	93.92	93.78	**45.93**	46.47	**43.60**	+0.05	**52.93**	+0.93
OPW (Ours)	**94.45**	**94.61**	45.00	**46.75**	42.26	**+1.68**	52.61	**+1.51**

F Behavior after Warmup and during GRPO

F.1 Execution Errors and Repetition

The Qwen3-4B analysis uses the first eight trajectories per task on 140 Seen and 134 Unseen ALFWorld tasks and 64 Seen and 200 Unseen ScienceWorld tasks. Base and Teacher have one evaluation each. The four warmup methods are evaluated after warmup and after 20, 40, and 60 GRPO steps. Table 6 uses the warmup evaluations from the new 64-rollout collection. Tasks receive equal weight.

We lowercase action text and normalize whitespace. An action repeats if it appeared earlier, whether or not that occurrence was valid. Validity requires membership in the current valid-action list. ScienceWorld navigation aliases absent from that list count as invalid even if execution succeeds. Empty or unparseable actions are invalid format failures. Empty actions are excluded from repetition. Consecutive identical invalid actions require both adjacent actions to be invalid and identical, forming a subset of repeated invalid actions. Event shares divide by all actions in each task's eight trajectories. Error incidence is the fraction of trajectories containing an invalid action.

On ALFWorld Seen, similar fractions of SFT and OPW trajectories contain invalid actions (50.9% and 50.0%), but consecutive identical invalid actions occupy 3.93% and 0.66% of all actions. On ScienceWorld Unseen, the consecutive-error share is 13.83% for SFT and 4.17% for OPW. Appendix H illustrates these patterns with complete action sequences.

Table 21 compares both model sizes using the original eight-rollout evaluations. Its initial means can differ slightly from the new warmup collection. V_0, L_0, and R_0 denote warmup validity, repetition, and task performance, with $\Delta X = X_{60} - X_0$. Repetition includes valid actions, and ScienceWorld task scores include partial progress. For ALFWorld 4B, OPW's validity rises by only 0.9 points during GRPO while success rises by 49.3, showing continued progress after most actions are valid.

Table 21: Execution behavior and task performance before and after 60 GRPO steps.

Warmup	Qwen3-1.7B						Qwen3-4B					
	Action validity		Repeated actions		Task performance		Action validity		Repeated actions		Task performance	
	V_0	ΔV	L_0	ΔL	R_0	ΔR	V_0	ΔV	L_0	ΔL	R_0	ΔR
ALFWorld, Seen												
Off-policy: SFT	76.7	+6.5	60.1	−35.9	6.1	+39.2	87.4	+5.4	29.3	−9.6	29.6	+54.4
Off-policy: SFT-RS	77.1	+4.0	61.7	−29.5	6.2	+37.6	88.0	+7.5	29.7	−16.1	31.0	+58.0
On-policy: GRPO	71.1	+12.7	38.5	−10.6	20.1	+45.4	90.1	−0.5	22.7	−5.1	56.0	+26.7
OPW (Ours)	92.0	+4.5	18.5	+1.6	46.5	+39.9	93.9	+0.9	16.4	−3.5	46.8	+49.3
ScienceWorld, Seen												
Off-policy: SFT	62.8	−39.6	87.8	−39.2	4.0	+29.4	63.5	−23.4	67.5	−31.0	21.8	+38.7
Off-policy: SFT-RS	61.5	−19.2	89.5	−28.4	3.3	+25.0	62.3	−13.1	67.8	−36.0	21.7	+42.9
On-policy: GRPO	35.0	−2.6	79.9	−29.0	9.0	+32.6	49.4	+2.4	41.7	−1.4	49.8	+19.0
OPW (Ours)	43.1	+3.7	53.8	−17.8	44.1	+27.6	51.7	+2.3	38.5	−11.3	57.3	+29.9

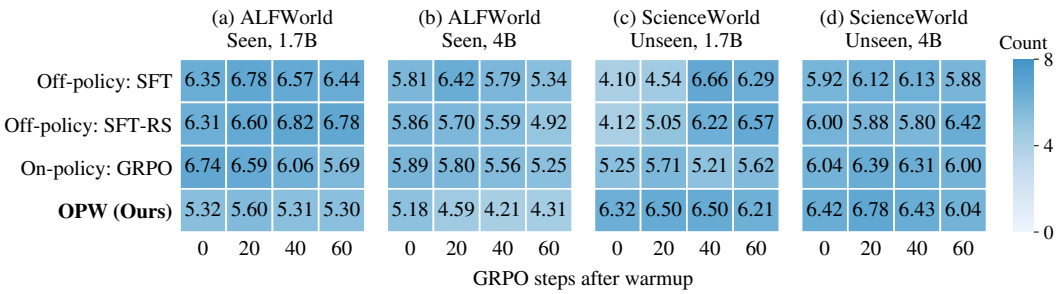

Figure 7: All distinct action sequences among eight rollouts per task during GRPO.

Table 21 (continued)

Warmup	Qwen3-1.7B						Qwen3-4B					
	Action validity		Repeated actions		Task performance		Action validity		Repeated actions		Task performance	
	V_0	ΔV	L_0	ΔL	R_0	ΔR	V_0	ΔV	L_0	ΔL	R_0	ΔR
ScienceWorld, Unseen												
Off-policy: SFT	61.0	−31.4	88.3	−41.4	8.8	+31.6	57.1	−13.3	62.6	−27.0	32.2	+29.7
Off-policy: SFT-RS	59.8	−16.3	86.6	−27.6	8.5	+27.1	56.4	−8.3	59.1	−30.3	34.2	+29.3
On-policy: GRPO	36.8	−0.6	77.1	−24.0	15.0	+30.5	48.5	+4.2	38.5	−2.1	56.8	+2.8
OPW (Ours)	42.5	+3.4	44.8	−8.0	51.8	+11.2	50.3	+3.0	36.6	−3.6	61.6	+6.6

F.2 Successful and Failure-Only Action Sequences

We evaluate both model sizes on 140 ALFWorld Seen and 200 ScienceWorld Unseen tasks, with eight rollouts per task after warmup and after 20, 40, and 60 GRPO steps. After lowercasing and whitespace normalization, we remove invalid actions and `look`, `look around`, and `inventory`, then merge consecutive identical actions. Object identities and order remain. We count each distinct sequence once per task, including the empty sequence. A sequence is successful if any occurrence succeeds and failure-only otherwise, even when filtering gives successful and failed trajectories the same sequence. Task averages include tasks with no successful trajectory.

Total sequence counts in Fig. 7 decrease after OPW, while successful counts in Fig. 3 increase in both environments and model sizes. In ALFWorld 4B, the total falls from 5.18 to 4.31, successful sequences rise from 2.23 to 4.00, and failure-only sequences fall from 2.95 to 0.31.

To compare equal numbers of successes, we retain tasks with at least $m \in \{2, 4\}$ successful trajectories under every method at all four evaluations. If a task has S successes and sequence j occurs c_j times, its expected distinct count in m successes selected without replacement is $\sum_j [1 - \binom{S-c_j}{m} / \binom{S}{m}]$. Table 22 averages over the N retained tasks, at warmup end and step 60.

For OPW 4B with $m = 4$, the expected count falls from 2.54 to 2.14 in ALFWorld but rises from 1.83 to 2.32 in ScienceWorld. More successful sequences within eight rollouts can therefore accompany either less or more variation among equally many successes. All 1.7B cells except ALFWorld with $m = 2$ describe a single task.

G Warmup Duration and Repeated Task Success

G.1 What Continues to Improve during Warmup?

Table 23 evaluates Qwen3-1.7B OPW with data-order seed 4 every five steps on 64 reserved ALFWorld analysis tasks, using 32 rollouts per task. These differ from the 140 Seen test tasks in the full-training comparison. Pass@8 and pass^8 use Eq. 8 with $n = 32$. From steps 20 to 40, valid-

Table 22: Distinct action sequences among equal numbers of successful trajectories.

Environment	Size	m	N	Off-policy: SFT	Off-policy: SFT-RS	On-policy: GRPO	**OPW (Ours)**
ALFWorld	1.7B	2	10	1.72 → 1.64	1.81 → 1.54	1.52 → 1.48	1.51 → 1.34
		4	1	3.00 → 3.21	2.00 → 2.60	1.97 → 2.41	1.00 → 1.93
	4B	2	49	1.69 → 1.60	1.70 → 1.50	1.62 → 1.58	1.67 → 1.52
		4	39	2.53 → 2.46	2.65 → 2.08	2.48 → 2.36	2.54 → 2.14
ScienceWorld	1.7B	2	1	1.67 → 1.00	1.73 → 1.46	1.68 → 1.00	1.29 → 1.00
		4	1	2.43 → 1.00	2.60 → 2.00	2.41 → 1.00	1.57 → 1.00
	4B	2	17	1.47 → 1.57	1.56 → 1.59	1.50 → 1.60	1.43 → 1.57
		4	7	1.98 → 2.02	2.03 → 2.31	1.52 → 2.07	1.83 → 2.32

ity gains slow while repeated task success continues to improve. Tasks solved at least once in 32 rollouts increase from 52 to 53, while those solved in every rollout increase from 2 to 11.

Table 23: Action validity and repeated task success during OPW (ALFWorld 1.7B, $n = 32$, %).

OPW steps	Valid actions	pass@8	pass^8
10	72.9	39.4	0.4
15	85.1	52.4	1.3
20	89.4	69.7	10.8
25	90.4	71.4	13.4
30	91.1	71.2	14.5
35	91.6	73.7	20.7
40	91.8	74.8	22.0

For ScienceWorld Qwen3-1.7B with data-order seed 3, we evaluate 64 Seen tasks every five OPW steps from 10 to 40, with eight rollouts per task. These samples differ from initial GRPO evaluations. Fig. 8 compares partial-credit reward contrast with complete-task success. At both steps 15 and 40, 92.2% of groups contain different scores, while pass@8 rises from 18.8% to 56.2%. Success coverage thus continues to improve when partial-credit reward differences are already common.

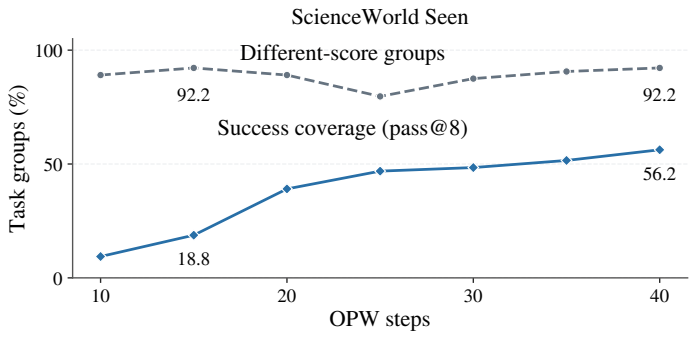

Figure 8: Success coverage and partial-credit reward contrast during OPW (ScienceWorld 1.7B).

G.2 Learning Speed and Repeated Task Success

Table 24 compares Qwen3-1.7B paths sharing a warmup run and ending at 100 total steps, evaluated on both benchmark splits with eight rollouts per task. ALFWorld branches after 20 or 40 OPW steps from the same seed-4 run. Longer warmup raises final Seen pass@1 from 88.3% to 94.9% and pass^8 from 69.3% to 80.7%. With evaluations every five steps, the first reaching 80% success moves from total step 75 to 60. Total steps are fixed, but GPU-hours differ.

ScienceWorld branches from one seed-1 checkpoint after 40 OPW steps. Switching to GRPO immediately yields higher Unseen pass@1 (36.6% versus 32.2%) and pass^8 (4.5% versus 3.5%) than 20 more OPW steps. Mean score favors longer warmup (63.84 versus 63.06). Longer warmup improves Seen mean score (71.72 to 76.65) and pass@8 (84.4% to 85.9%). The preferred allocation depends on the task set and whether evaluation rewards partial progress or complete success.

Table 24: Repeated task success after 100 total steps (Qwen3-1.7B, $n = 8$, %).

Environment / task set	**OPW** steps	pass^1	pass^2	pass^4	pass^8
ALFWorld / Seen	20	88.3	83.5	77.2	69.3
	40	**94.9**	**91.8**	**86.9**	**80.7**
ScienceWorld / Unseen	40	**36.6**	**20.1**	**8.7**	**4.5**
	60	32.2	17.8	8.5	3.5
ScienceWorld / Seen	40	51.8	37.4	24.4	10.9
	60	**56.1**	**42.7**	**30.7**	**20.3**

Four-duration comparison across training runs. Fig. 4 compares 20, 40, 60, and 80 OPW steps followed by GRPO to 100 steps, evaluated every 20 steps. ALFWorld uses seed 4 for 20/40, seed 3 for 60, and seed 1 for 80 OPW steps. ScienceWorld uses seed 3 for 20/40 and seed 1 for 60/80. Rankings include training-run differences. ScienceWorld's 40-step curve ends at 2.0% Unseen pass^8, while the seed-1 branch in Table 24 ends at 4.5%. The 60-step branch is shared.

Within the ScienceWorld seed-3 run, extending warmup from 20 to 40 steps raises Unseen pass@1 from 28.3% to 34.1% and pass^8 from 1.0% to 2.0%, while mean score changes little (60.94 to 60.64). Seen pass^8 also rises, from 3.1% to 10.9%. These comparisons support choosing duration by subsequent validation on the required outcome.

H Long-Trajectory Examples

Two ALFWorld Unseen tasks illustrate long interactions and failure patterns. Comparisons use Qwen3-4B on the same task and the sixth evaluation trajectory. Base uses released weights, while SFT and OPW use their 40-step warmup models. Table 6 reports task-wide error frequencies.

The tables retain every action and quote selected environment responses, omitting reasoning and other observations. Orange marks invalid actions, red marks repeated invalid actions, and R marks any repetition. Other actions are valid. Highlights and turn labels are our annotations.

OPW completes the requested task while retaining useful repeated actions. Table 25 follows a task that requires cleaning a pan and placing it on a countertop. Base cleans and places a pot instead, then repeats `finish` without completing the task. OPW locates the pan and completes the task in 26 actions. Its returns to the sink and countertop serve cleaning and placement.

Table 25: Complete action sequences before and after OPW on the pan-cleaning task.

Task: "clean some pan and put it in countertop." ALFWorld Unseen, Qwen3-4B.

	Base		**OPW (Ours)**
1	go to cabinet 1	1	go to cabinet 1
2	go to cabinet 2	2	go to cabinet 2
3	open cabinet 2	3	open cabinet 2
4	go to cabinet 3	4	go to cabinet 3
5	go to cabinet 4	5	go to cabinet 4
6	go to cabinet 5	6	go to cabinet 5
7	open cabinet 5	7	open cabinet 5
8	go to cabinet 6	8	go to cabinet 6
9	open cabinet 6	9	open cabinet 6
10	go to drawer 1	10	go to drawer 1

Table 25 (continued)

Task: "clean some pan and put it in countertop." ALFWorld Unseen, Qwen3-4B.

	Base			OPW (Ours)	
11	open drawer 1		11	open drawer 1	
12	go to drawer 2		12	go to drawer 2	
13	open drawer 2		13	open drawer 2	
14	go to drawer 3		14	go to drawer 3	
15	open drawer 3		15	open drawer 3	
16	go to shelf 1		16	go to countertop 1	
17	go to shelf 2		17	go to countertop 2	
18	go to shelf 3		18	go to countertop 3	
19	go to sinkbasin 1		19	go to sinkbasin 1	
20	go to stoveburner 1		20	go to stoveburner 1	
21	take pot 1 from stoveburner 1		21	go to stoveburner 2	
22	go to sinkbasin 1	R	22	take pan 1 from stoveburner 2	
23	clean pot 1 with sinkbasin 1		23	go to sinkbasin 1	R
24	go to countertop 1		24	clean pan 1 with sinkbasin 1	
25	move pot 1 to countertop 1		25	go to countertop 1	R
26	finish		26	move pan 1 to countertop 1	
27	finish	R			
28	finish	R			
29	finish	R			
30	finish	R			
Environment feedback			**Environment feedback**		
Turn 21: "You pick up the pot 1 from the stoveburner 1."			Turn 22: "You pick up the pan 1 from the stoveburner 2."		
Turns 26–30: "Nothing happens."			Turn 24: "You clean the pan 1 using the sinkbasin 1."		
Outcome: Task not completed.			Turn 26: "You move the pan 1 to the countertop 1."		
25/30 actions valid, 4 repeated invalid actions.			**Outcome:** Task completed.		
			26/26 actions valid, 0 repeated invalid actions.		

OPW completes the task requirements after a longer search. In Table 26, SFT finds the egg earlier and places it in the microwave without cleaning it. It issues `end` 11 times with feedback "Nothing happens." OPW searches longer but cleans and places the egg in 29 actions.

Table 26: Complete action sequences after SFT and OPW on the egg-cleaning task.

Task: "put a clean egg in microwave." ALFWorld Unseen, Qwen3-4B.

	Off-policy: SFT		OPW (Ours)
1	go to fridge 1	1	go to fridge 1
2	open fridge 1	2	open fridge 1
3	go to cabinet 1	3	go to cabinet 1
4	go to cabinet 2	4	go to cabinet 2
5	open cabinet 2	5	open cabinet 2
6	go to drawer 1	6	go to cabinet 3
7	open drawer 1	7	go to cabinet 4
8	go to shelf 1	8	go to cabinet 5
9	go to shelf 2	9	open cabinet 5
10	go to shelf 3	10	go to cabinet 6
11	go to countertop 1	11	open cabinet 6
12	go to countertop 2	12	go to drawer 1
13	go to countertop 3	13	open drawer 1
14	take egg 1 from countertop 3	14	go to drawer 2
15	examine egg 1	15	open drawer 2
16	go to microwave 1	16	go to drawer 3
17	open microwave 1	17	open drawer 3
18	move egg 1 to microwave 1	18	go to shelf 1
19	close microwave 1	19	go to shelf 2

Table 26 (continued)

Task: "put a clean egg in microwave." ALFWorld Unseen, Qwen3-4B.

Off-policy: SFT		OPW (Ours)	
20 end		20 go to shelf 3	
21 end	R	21 go to countertop 1	
22 end	R	22 go to countertop 2	
23 end	R	23 go to countertop 3	
24 end	R	24 take egg 1 from countertop 3	
25 end	R	25 go to sinkbasin 1	
26 end	R	26 clean egg 1 with sinkbasin 1	
27 end	R	27 go to microwave 1	
28 end	R	28 open microwave 1	
29 end	R	29 move egg 1 to microwave 1	
30 end	R		
Environment feedback Turn 18: "You move the egg 1 to the microwave 1." Turns 20–30: "Nothing happens."		**Environment feedback** Turn 26: "You clean the egg 1 using the sinkbasin 1." Turn 29: "You move the egg 1 to the microwave 1."	
Outcome: Task not completed. 19/30 actions valid, 10 repeated invalid actions.		**Outcome:** Task completed. 29/29 actions valid, 0 repeated invalid actions.	

I LIMITATIONS AND SCOPE

Our experiments use ALFWorld and ScienceWorld as controlled testbeds for studying how OPW affects subsequent RLVR in multi-step interactive tasks. Extending this investigation to more diverse interaction settings is a natural direction for future work. Our theoretical analysis characterizes initial reward discovery under the stated assumptions, complementing the empirical evaluation of subsequent learning dynamics. In practice, the appropriate warmup duration and end-to-end training efficiency depend on the task, model configuration, and cost of teacher supervision. We examine multiple warmup durations and include teacher scoring in the recorded training cost, distinguishing faster subsequent learning from overall computational savings. Adaptive criteria for transitioning from warmup to RLVR could further improve the allocation of training resources.