RETIREOPD: SELF-RETIRING ON-POLICY DISTILLATION FOR AGENTIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-turn agents trained with reinforcement learning (RL) from verifiable rewards receive one scalar reward per trajectory, which gives little guidance for intermediate decisions. On-policy distillation (OPD) adds dense token-level supervision from a teacher; in agentic training, the teacher is conditioned on privileged information, such as task skills available only during training, so that a skill-free student can internalize them. This recipe rests on two assumptions, and we find that neither holds in agentic tasks. First, privileged information alone does not make a teacher reliable: a skill-conditioned policy that is not itself optimized for the task often fails to outperform the student it supervises. Second, the benefit of teacher supervision is stage-dependent: matching the teacher accelerates early learning but later conflicts with reward optimization and holds the student near the teacher. We therefore propose **RetireOPD** (Self-**Retiring On-Policy Distillation**), which first optimizes a decoupled, skill-conditioned teacher with environment rewards and then trains a skill-free student jointly with RL and OPD. Rather than following a predefined distillation schedule, RetireOPD adopts **Adaptive Retirement**: the student drops the teacher on its own once their discrepancy stops shrinking and it reaches a target fraction of the teacher's success rate, after which training proceeds with RL alone. Across Qwen2.5 models from 1.5B to 7B, RetireOPD improves ALFWorld success rate over GRPO by 14.1% to 18.8% and WebShop accuracy by 11.8% to 19.0%, and it surpasses its own skill-conditioned teacher in every setting. Code is available at https://anonymous.4open.science/r/RetireOPD-7537.

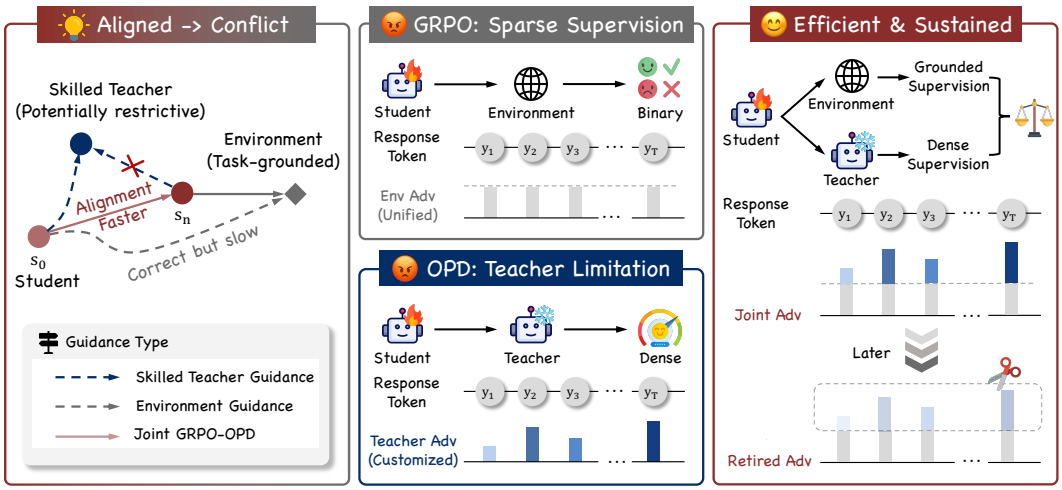

Figure 1: **Left:** GRPO and teacher guidance are aligned early in training but gradually diverge, making continued teacher matching restrictive. **Middle:** GRPO provides a trajectory-level binary reward (0/1) shared across all tokens, whereas OPD provides continuous, token-level supervision. **Right:** RetireOPD combines both signals early and retires teacher supervision after conflict emerges, enabling sustained improvement beyond the teacher.

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) is the standard approach for training large language model (LLM) agents on multi-turn tasks (Singh et al., 2025; Shao et al., 2024), where a single outcome reward per trajectory leaves the intermediate decisions of a long interaction unsupervised. On-Policy Distillation (OPD) complements this sparse reward with dense token-level supervision from a teacher on trajectories sampled by the student (Zeng et al., 2026; Team, 2026; Xu et al., 2026a; Lu & Lab, 2025). In agentic training, the teacher is commonly the same model conditioned on privileged context that is available only during training, such as retrieved task skills (Zhao et al., 2026; Lu et al., 2026a). The student is trained to reproduce the teacher's behavior without access to this context, so that the skills are internalized into its parameters and no additional context is required at inference (Lu et al., 2026b; Wang et al., 2026a).

This paradigm rests on two assumptions. The teacher is assumed to be reliably better than the student because it sees the privileged context, and matching the teacher is assumed to remain useful for as long as training lasts (Zhang et al., 2026d; Zhou et al., 2026). They amount to two questions that current methods answer by assumption (Lu et al., 2026a) or by a fixed schedule (Tan et al., 2026; Ding et al., 2026): *which teacher* the student should learn from, and *for how long*. We examine these assumptions on ALFWorld and find that neither assumption holds (Figure 2).

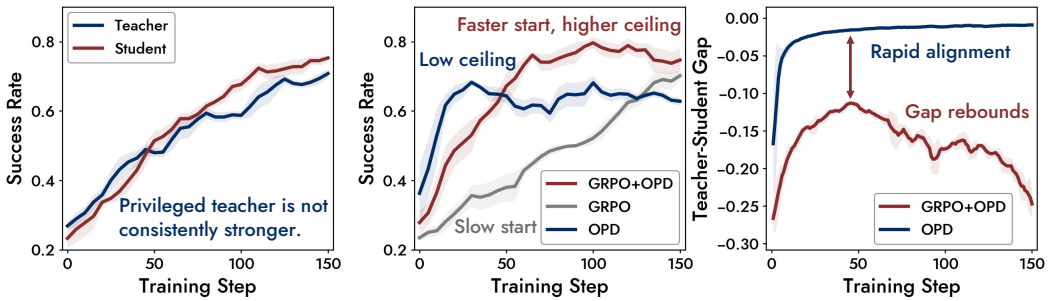

Figure 2: Training dynamics of Qwen2.5-3B-Instruct on ALFWorld. **Left**: Success rates of the privileged teacher and student branches using the same model. **Middle**: Success rate under GRPO, OPD and GRPO+OPD. **Right**: Teacher-student discrepancy under OPD and GRPO+OPD.

Privileged context alone does not make a teacher reliable. Under joint Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and On-Policy Self-Distillation (OPSD) (Zhao et al., 2026), where teacher and student are two branches of one shared policy (Lu et al., 2026a), the skill-conditioned branch does not consistently outperform the student it supervises (Figure 2, left). This suggests that the teacher does not effectively utilize the provided skill: although it receives privileged skill context, it has not been adequately trained to incorporate this information into its decisions. Simply increasing the model size does not resolve this issue either—an unoptimized 7B model prompted with the same skills reaches only 23.4% success on ALFWorld. A teacher must therefore be *optimized to use* its privileged context before it can provide effective supervision.

The benefit of teacher supervision is stage-dependent. Adding OPD to GRPO removes the slow start of pure GRPO and the low ceiling of pure OPD (Figure 2, middle). The teacher discrepancy first narrows and then widens (Figure 2, right). Once the student has internalized the behavior that the skills induce, reward optimization favors actions beyond the teacher, and the two gradients begin to conflict (Figure 1, left). Continuing to match the teacher beyond this point holds the student near the teacher's performance ceiling (Section 4.3). A first-order analysis (Appendix A.1) shows that stagnating discrepancy implies diminishing OPD benefits or increasing gradient conflict. Teacher supervision is therefore *temporary scaffolding*, and the question is when to remove it.

Existing work addresses this transition with predefined schedules, either annealing distillation (Tan et al., 2026; Ding et al., 2026) or switching from distillation to RL at a fixed step (Ye et al., 2026a; Li et al., 2026a). Online decisions exist only at finer granularity, per prompt or per token (Ding, 2026; Lu et al., 2026a), and none of them removes the teacher. In our experiments, the point at which the discrepancy stops decreasing varies from step 50 to step 90 across models and tasks (Appendix A.9),

so a single schedule may withdraw guidance too early in some settings and too late in others. These observations suggest a simple principle: *whether the teacher is still useful can be read from the training signal itself.* The discrepancy stops decreasing when the two objectives conflict, and the student's success rate relative to the teacher indicates whether the skills have been internalized. The transition should therefore be determined online rather than fixed in advance.

Motivated by these findings, we propose **RetireOPD** (Self-Retiring On-Policy Distillation), which internalizes privileged skills into a skill-free student and retires the teacher once its supervision no longer benefits the student. RetireOPD first decouples the teacher from the student and trains it with environment rewards under the skill context, so that distillation starts from a policy that has already learned to exploit the skills. The student is then trained jointly with GRPO and OPD, while the two signals identified above are monitored at fixed intervals. Once the discrepancy stops decreasing and the student's success rate reaches a set fraction of the teacher's, the teacher is retired and training proceeds with GRPO alone. The student is thus no longer constrained by the teacher, and no further teacher forward passes are required.

Across Qwen2.5-1.5B, 3B and 7B on ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022), RetireOPD outperforms RL, distillation and hybrid baselines, improving ALFWorld success rate over GRPO by 14.1% to 18.8% and WebShop accuracy by 11.8% to 19.0%. It also surpasses its own skill-conditioned teacher in every setting. Our contributions are as follows:

- We identify two failure modes of privileged-information distillation for agents: an unoptimized skill-conditioned teacher is unreliable, and teacher matching conflicts with reward optimization once the student has internalized the teacher's knowledge.
- We propose RetireOPD, which trains a same-capacity skill-conditioned teacher with environment rewards and retires it online once the teacher-student discrepancy stops decreasing and the student reaches a set fraction of the teacher's success rate.
- Experiments across three model scales on ALFWorld and WebShop show that RetireOPD consistently outperforms RL, distillation and hybrid baselines while surpassing its own teacher, which validates its effectiveness on agentic tasks.

2 Related Work

2.1 On-Policy Distillation with Privileged Teachers

On-policy distillation (OPD) trains a student on its own samples with token-level feedback from a teacher (Agarwal et al., 2024; Lu & Lab, 2025; Song & Zheng, 2026) and is now a standard post-training stage (Xiao et al., 2026; Zeng et al., 2026; Team et al., 2026). Without a stronger teacher, on-policy self-distillation (OPSD) conditions the same model on privileged information, such as a reference solution or textual feedback, and distills the conditioned branch into the unconditioned one (Zhao et al., 2026; Hübotter et al., 2026; Ye et al., 2026b; Ma et al., 2026). For agents, this privileged information is often a retrieved or hindsight skill (Lu et al., 2026a; Wang et al., 2026a), with teacher supervision controlled through token-level gating or shaping (Lu et al., 2026a; Zhang et al., 2026c). Teacher reliability itself has received less attention, although guidance quality depends on compatibility with the student and on capability beyond it (Li et al., 2026c; Liu et al., 2026; Chen et al., 2026), and only some tokens carry useful signal (Xu et al., 2026b; Zhang et al., 2026b). RetireOPD instead trains the teacher separately with environment rewards before distillation.

2.2 Combining On-Policy Distillation with Reinforcement Learning

RLVR is the dominant approach for multi-turn agents (Shao et al., 2024; Feng et al., 2025; Dong et al., 2026a; Wang et al., 2025; Lu et al., 2025), but its trajectory-level reward is sparse, so hybrid methods add a distillation term or a teacher-shaped advantage to the RL objective (Lu et al., 2026a; Zhang et al., 2026a; Yang et al., 2026). Because the two signals interfere (Li et al., 2026a; Pan et al., 2026) as the student improves, several methods reduce the teacher's influence over training: annealing the distillation weight (Tan et al., 2026; Ding et al., 2026), reducing the trajectory horizon exposed to the student (Wang et al., 2026b; Li et al., 2026b; Xing et al., 2026), or running distillation and RL in sequence (Ye et al., 2026a; Kim & Lee, 2026; Dong et al., 2026b). In all of these the

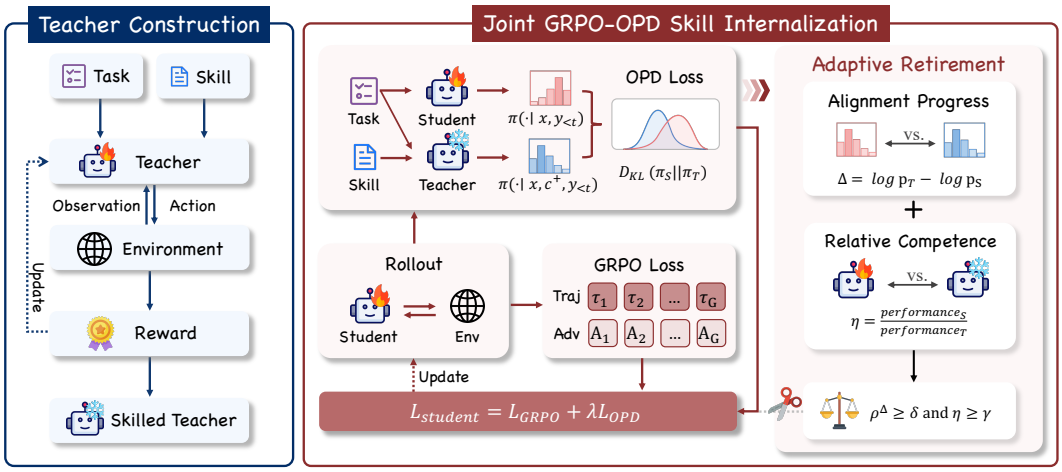

Figure 3: **Overview of RetireOPD. (1) Teacher Construction**: We optimize a skill-conditioned teacher with environmental rewards. **(2) Joint Skill Internalization**: the student learns from both GRPO and OPD training. **(3) Adaptive Retirement**: OPD supervision is removed by current policy.

transition is fixed before training. Signal-dependent decisions exist only at finer granularity, such as distilling on prompts where all rollouts fail (Ding, 2026; Han et al., 2026) or gating the teacher per token (Lu et al., 2026a; Wang et al., 2026c), and none removes the teacher. RetireOPD decides the global transition online from the teacher-student discrepancy and the student's relative competence.

3 Method

As shown in Figure 3, our method consists of three stages: **(1) Teacher construction**, optimizing a skill-conditioned teacher with environment rewards (Section 3.1); **(2) Joint GRPO-OPD training**, where the skilled teacher supervises a skill-free student on student-generated trajectories (Section 3.2) and **(3) Adaptive teacher retirement**, removing OPD once behavioral transfer stagnates and the student reaches sufficient task competence (Section 3.3).

Task Definition. We consider a multi-turn agent interacting with an environment over a sequence of decision steps. Given a task input $x \sim \mathcal{D}$, the agent iteratively generates responses based on the interaction history and receives environment feedback, forming a trajectory $\tau = (x, y)$, where $y = (y_1, \ldots, y_T)$ denotes the flattened sequence of all valid response tokens across interaction turns, and $R(\tau)$ is the scalar environment reward.

3.1 Privileged Teacher Construction

The base model is insufficient to utilize privileged skill context, necessitating teacher training. We therefore train a teacher policy to learn how to leverage skills for decision-making. Specifically, the teacher π_ϕ is provided with the privileged skill context c^+ during training. We optimize the teacher with GRPO (Shao et al., 2024) to obtain a *skilled teacher*, with the following objective:

$$\phi^* = \arg\max_{\phi} \mathbb{E}_{x\sim\mathcal{D},\, \tau\sim\pi_\phi(\cdot|x,c^+)} \left[R(\tau)\right]. \tag{1}$$

Environment reward optimization encourages the teacher to turn privileged information into effective task behavior. After training, we freeze ϕ^* and denote the resulting policy as the skilled teacher π_T. The frozen teacher is subsequently used only to provide dense supervision for the student. Further details on teacher construction are provided in Appendix A.5.

3.2 JOINT GRPO-OPD OPTIMIZATION

With the skilled teacher obtained, we next train the student policy π_θ to internalize the teacher's behaviors. Following SDAR (Lu et al., 2026a), the student is trained without access to the privileged skill context c^+. For each input x, the student π_θ samples a group of G trajectories. Based on the environment rewards, we first optimize the student with GRPO, which is defined as:

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\min\left(r_t^{(i)}(\theta)\hat{A}^{(i)},\ \mathrm{clip}\left(r_t^{(i)}(\theta),1-\epsilon,1+\epsilon\right)\hat{A}^{(i)}\right)\right] + \alpha_{KL}\,\mathbb{E}_{x\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}D_{\mathrm{KL}}\left(\pi_\theta(\cdot\mid x,y_{<t}^{(i)})\,\|\,\pi_{\mathrm{ref}}(\cdot\mid x,y_{<t}^{(i)})\right)\right], \tag{2}$$

where $r_t^{(i)}(\theta) = \pi_\theta(y_t^{(i)}\mid x,y_{<t}^{(i)})/\pi_{\theta_{\mathrm{old}}}(y_t^{(i)}\mid x,y_{<t}^{(i)})$ is the token-level importance ratio.

For each student-generated trajectory, we compare the student and teacher distributions conditioned on the same trajectory prefix $y_{<t}$. The teacher provides supervision directly on states visited by the student. We define the OPD objective using the reverse KL divergence:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}D_{\mathrm{KL}}\left(\pi_\theta(\cdot\mid x,y_{<t})\,\|\,\pi_T(\cdot\mid x,c^+,y_{<t})\right)\right]. \tag{3}$$

To reduce the overhead of exact KL computation, we use a sampled-token approximation on student-generated trajectories, evaluating each y_t under both the student and teacher distributions:

$$\Delta_t = \log\pi_T(y_t\mid x,c^+,y_{<t}) - \log\pi_\theta(y_t\mid x,y_{<t}). \tag{4}$$

Finally, the full optimization objective is formulated as:

$$\mathcal{L}_{\mathrm{student}}(\theta) = \mathcal{L}_{\mathrm{GRPO}}(\theta) + \lambda\mathcal{L}_{\mathrm{OPD}}(\theta), \tag{5}$$

where λ controls the contribution of teacher supervision.

3.3 ADAPTIVE TEACHER RETIREMENT

Teacher supervision is beneficial during the early stages of training, but can eventually constrain further performance gains by imposing a capability ceiling on the student. We therefore introduce an adaptive retirement mechanism that monitors training dynamics and determines when the teacher should be removed from the optimization process.

We partition training into monitoring windows of W steps indexed by m. At step n, let $y_n^{(i)}$ denote the i-th sampled trajectory. Given the teacher-student log-probability gaps $\Delta_t^{(i)}$, we compute the step-level and window-level *alignment progress* K_n and $\bar{K}_m$:

$$K_n = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_n^{(i)}|}\sum_{t=1}^{|y_n^{(i)}|}\Delta_t^{(i)}, \quad \bar{K}_m = \frac{1}{W}\sum_{n=mW}^{mW+W-1}K_n. \tag{6}$$

To capture when teacher-student alignment begins to stagnate or reverse, we define the *rebound indicator* $\rho_m^{(K)}$ and *relative competence* η_m as

$$\rho_m^{(K)} = \frac{\bar{K}_m - \bar{K}_{m-1}}{\bar{K}_m}, \quad \eta_m = \frac{SR_m + SR_{m-1}}{2SR_T}, \tag{7}$$

where SR_m and SR_T denote the student and teacher success rates, respectively. Teacher supervision is retired at the first monitoring window satisfying:

$$m^* = \inf \left\{ m \geq 2 : \rho_m^{(K)} \geq \delta \ \wedge \ \eta_m \geq \gamma \right\}. \tag{8}$$

Here, δ and γ denote the rebound indicator and relative competence thresholds, respectively. Requiring both criteria prevents premature retirement due to transient fluctuations, ensuring that the student has sufficiently benefited from teacher supervision before continuing independently.

4 Experiments

4.1 Experiment Setup

Benchmarks. We evaluate our method on two widely used interactive agent benchmarks: ALFWorld (Shridhar et al., 2021) and WebShop (Yao et al., 2022). ALFWorld evaluates multi-step household task completion in a text-based embodied environment, while WebShop tests goal-directed search, navigation and decision making in realistic simulated online shopping scenarios.

Baselines. We compare RetireOPD with four groups of baselines using Qwen2.5-1.5B / 3B / 7B-Instruct (Qwen et al., 2025): **(1) Prompting.** Vanilla directly evaluates the instruction-tuned model, while Skill-Prompt additionally provides task-relevant skills at inference time. **(2) RL.** GRPO (Shao et al., 2024), GiGPO (Feng et al., 2025), PPO (Schulman et al., 2017), and RLOO (Ahmadian et al., 2024) are standard RL baselines, while Skill-GRPO (Xia et al., 2026) performs GRPO with privileged skills and serves as our skill-conditioned teacher. **(3) Distillation.** OPD (Ye et al., 2026b) and OPSD (Zhao et al., 2026) use dense teacher supervision. **(4) Hybrid.** GRPO+OPD and GRPO+OPSD (Lu et al., 2026a) combine RL with dense teacher supervision, while RetireOPD further removes teacher supervision adaptively.

Implementation. We apply identically initialized teacher and student models at each scale. We set $W = 5$, $\gamma = 0.9$, and $\delta = 0$. Following SDAR (Lu et al., 2026a), we set $\lambda = 0.01$ and use its Keyword Matching strategy to retrieve privileged skills from the SkillBank of SkillRL (Xia et al., 2026) in both environments. Additional details are provided in Appendix A.3.

4.2 Main Results

Necessity of Skilled Teacher Training. Explicitly training the teacher before distillation brings substantial gains over directly applying OPSD, highlighting the importance of effective skill utilization. As shown in Table 1, using a separately trained teacher for OPD instead of OPSD leads to a 14.8 points improvement on ALFWorld (87.5% vs. 72.7%) and a 3.1 points improvement on WebShop (69.9% vs. 66.8%) with Qwen2.5-1.5B-Instruct. These results demonstrate the necessity of training the teacher to effectively leverage its privileged context before using it for distillation.

Effectiveness beyond Pure RL and Distillation. RetireOPD consistently improves over both pure reinforcement learning and distillation-based approaches across tasks and model scales. On ALFWorld, RetireOPD achieves a 93.8% success rate with Qwen2.5-3B-Instruct, outperforming GRPO (75.0%) and OPD (74.2%). On WebShop, it reaches 75.8% accuracy with Qwen2.5-1.5B-Instruct, compared with 56.8% and 69.1% for GRPO and OPD, respectively. Distillation-based methods remain constrained by fixed teacher supervision: OPD with a skilled teacher achieves strong performance but remains bounded by the fixed distillation process. These results show that RetireOPD effectively combines the complementary strengths of reinforcement learning and distillation while mitigating the limitations of fixed teacher supervision.

Advantages over Naive Hybrid Approaches. Simply combining GRPO and OPD with a fixed objective is insufficient to capture their different roles throughout training. RetireOPD improves over the GRPO+OPD baseline by 11.0 points on ALFWorld and 3.1 points on WebShop with the 3B model. This highlights the importance of adaptive teacher retirement, which dynamically adjusts the reliance on teacher supervision as the student improves.

Table 1: Main results on ALFWorld and WebShop. * means validation with skills. The highlighted suffix of a Distillation method denotes its teacher. **Best** and second-best results are highlighted.

		ALFWorld							WebShop	
Method	**Type**	**Pick**	**Look**	**Clean**	**Heat**	**Cool**	**Pick2**	**Avg**	**Score**	**Acc**
Qwen2.5-1.5B-Instruct										
Vanilla (Qwen et al., 2025)	Prompting	11.1	0.0	6.2	0.0	0.0	4.2	5.5	17.8	5.5
Skill-Prompt* (Xia et al., 2026)	Prompting	3.4	16.7	12.9	5.3	0.0	5.0	6.2	20.8	1.6
GRPO (Shao et al., 2024)	RL	85.3	53.7	84.5	78.2	58.7	53.5	72.8	75.8	56.8
GiGPO (Feng et al., 2025)	RL	94.4	67.5	94.8	94.4	79.8	76.4	86.7	83.1	65.0
PPO (Schulman et al., 2017)	RL	64.8	40.5	57.1	60.6	46.4	47.4	54.4	73.8	51.5
RLOO (Ahmadian et al., 2024)	RL	88.3	52.8	71.0	62.8	66.4	56.9	69.7	73.9	52.1
Skill-GRPO* (Xia et al., 2026)	RL	66.7	63.6	**100.0**	**100.0**	75.0	**100.0**	81.2	85.2	67.9
OPD-Skill-1.5B (Ye et al., 2026b)	Distillation	85.7	46.2	73.1	55.6	76.7	53.3	71.1	84.7	69.1
OPD-3B (Ye et al., 2026b)	Distillation	64.7	7.1	21.9	0.0	9.1	5.6	25.8	32.5	3.9
OPSD (Zhao et al., 2026)	Distillation	26.3	16.7	9.1	6.7	9.1	5.3	14.1	22.3	10.2
GRPO+OPD	Hybrid	**96.4**	88.9	89.3	90.5	73.7	82.6	87.5	82.5	69.9
GRPO+OPSD (Lu et al., 2026a)	Hybrid	87.1	66.7	85.7	58.3	**85.0**	46.2	72.7	79.9	66.8
RetireOPD	Hybrid	94.9	**100.0**	90.0	**100.0**	84.2	77.3	**89.8**	**86.6**	**75.8**
Qwen2.5-3B-Instruct										
Vanilla (Qwen et al., 2025)	Prompting	44.4	11.1	6.2	15.4	28.6	12.5	21.9	6.7	0.8
Skill-Prompt* (Xia et al., 2026)	Prompting	51.7	66.7	48.4	0.0	4.3	10.0	28.9	0.2	0.8
GRPO (Shao et al., 2024)	RL	91.2	62.5	96.2	61.9	65.0	47.4	75.0	79.8	63.3
GiGPO (Feng et al., 2025)	RL	95.8	81.8	96.4	**100.0**	88.9	**86.4**	92.2	83.1	69.5
PPO (Schulman et al., 2017)	RL	93.5	76.5	90.5	50.0	81.2	60.0	81.2	76.0	59.4
RLOO (Ahmadian et al., 2024)	RL	87.5	55.6	71.4	54.5	59.3	80.0	72.7	81.8	68.0
Skill-GRPO* (Xia et al., 2026)	RL	88.9	76.9	**100.0**	77.8	74.3	62.5	79.7	79.7	64.8
OPD-Skill-3B (Ye et al., 2026b)	Distillation	74.3	80.0	76.9	69.2	81.8	58.8	74.2	79.1	66.4
OPD-7B (Ye et al., 2026b)	Distillation	67.5	33.3	37.5	43.8	19.0	7.7	38.3	10.4	2.3
OPSD (Zhao et al., 2026)	Distillation	48.8	41.7	16.7	0.0	15.8	16.7	28.1	11.3	3.1
GRPO+OPD	Hybrid	96.6	**90.0**	83.3	87.5	66.7	78.9	82.8	84.5	74.2
GRPO+OPSD (Lu et al., 2026a)	Hybrid	**97.6**	58.3	91.7	92.3	63.6	31.2	78.1	77.8	66.4
RetireOPD	Hybrid	**97.6**	83.3	**100.0**	85.7	**95.5**	83.3	**93.8**	**85.8**	**77.3**
Qwen2.5-7B-Instruct										
Vanilla (Qwen et al., 2025)	Prompting	36.1	22.2	3.1	0.0	0.0	0.0	12.5	5.9	1.6
Skill-Prompt* (Xia et al., 2026)	Prompting	51.7	50.0	32.3	5.3	4.3	0.0	23.4	1.7	0.8
GRPO (Shao et al., 2024)	RL	91.2	87.5	96.2	81.0	65.0	57.9	81.2	80.9	72.6
GiGPO (Feng et al., 2025)	RL	97.7	82.7	98.8	83.7	89.3	79.2	90.8	84.4	72.8
PPO (Schulman et al., 2017)	RL	92.3	64.0	92.5	89.5	80.3	68.8	80.4	81.4	68.7
RLOO (Ahmadian et al., 2024)	RL	87.6	78.2	87.3	81.3	71.9	48.9	75.5	80.3	65.7
Skill-GRPO* (Xia et al., 2026)	RL	97.1	**100.0**	92.3	**100.0**	85.0	72.7	90.6	88.7	78.9
OPD-Skill-7B (Ye et al., 2026b)	Distillation	**100.0**	77.8	76.0	**100.0**	72.2	90.5	86.7	88.9	78.1
OPD-14B (Ye et al., 2026b)	Distillation	84.2	40.0	77.8	42.1	66.7	47.4	64.8	28.8	4.7
OPSD (Zhao et al., 2026)	Distillation	50.0	60.0	22.7	21.4	17.6	9.5	32.8	4.5	2.3
GRPO+OPD	Hybrid	**100.0**	76.9	95.2	**100.0**	82.6	87.5	92.2	90.3	81.2
GRPO+OPSD (Lu et al., 2026a)	Hybrid	94.3	58.8	81.0	94.4	88.2	50.0	79.7	86.8	76.5
RetireOPD	Hybrid	**100.0**	**100.0**	**100.0**	76.9	**96.4**	**90.9**	**95.3**	**91.7**	**84.4**

4.3 Training Dynamics

Performance Trends. The left panel of Figure 4 illustrates the training dynamics of GRPO+OPD, RetireGRPO, and RetireOPD on Qwen2.5-3B-Instruct. RetireGRPO denotes the variant that adaptively removes GRPO while retaining OPD. GRPO+OPD converges rapidly after around 60 steps, while RetireGRPO suffers a performance drop after removing GRPO, as OPD drives the student toward the teacher and limits further improvement. In contrast, RetireOPD continues to improve after adaptively removing OPD, indicating that timely retirement can release the student from the teacher's capability ceiling and enable further RL-driven improvement.

Teacher–Student Alignment and Exploration. We further analyze the teacher-student gap and entropy (Figure 4 middle and right). GRPO+OPD continuously pulls the student toward the teacher,

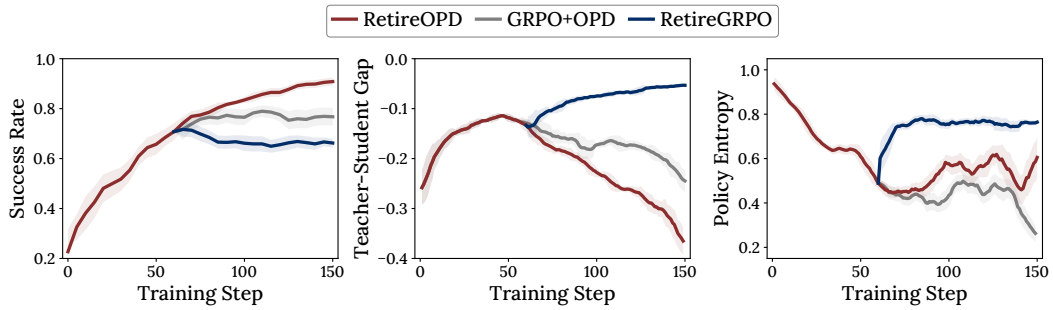

Figure 4: **Left:** Success rate of student model on ALFWorld. **Middle:** Teacher-student gap on ALFWorld. **Right:** Entropy of student model on ALFWorld.

while RetireGRPO allows the student to move even closer to the teacher. In contrast, our method avoids excessive teacher fitting, preserving the student's capacity for improvement. The entropy trends further support this observation: removing GRPO leads to increased entropy and less stable training, whereas GRPO+OPD drives entropy downward, indicating reduced exploration. Our method maintains entropy at a moderate level, balancing training stability and exploration.

4.4 Ablation Studies

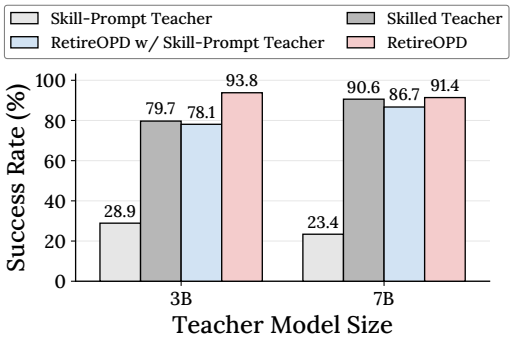

Figure 5: Ablation study of teacher construction on ALFWorld.

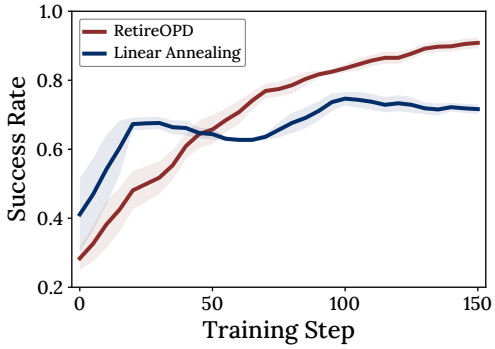

Figure 6: Adaptive Retirement vs. Linear Annealing on ALFWorld.

Teacher Construction. Figure 5 compares teachers with and without skill training, together with the corresponding Qwen2.5-3B-Instruct students trained with our method. Dedicated training substantially improves the teacher itself, with consistent gains across both model sizes. More importantly, the improved teacher provides more effective supervision for the student. For the 3B teacher, the student reaches 93.8% when trained with the dedicated skill-teacher, compared with 78.1% when using the untrained teacher, while the 7B teacher yields a similar improvement from 86.7% to 91.4%. These results demonstrate that dedicated skill training not only enhances the teacher's ability to exploit privileged context, but also enables more effective supervision for the student.

Retirement Criteria. Table 2 presents an ablation of the retirement mechanism on Qwen2.5-3B-Instruct on ALFWorld. Removing either Alignment Stagnation or Relative Competence reduces the success rate from 93.8% to 89.0%, showing that both components contribute to effective retirement. Removing retirement entirely further reduces performance to 82.8%, while removing OPD from the beginning and using GRPO alone achieves only 75.0%. These results show that both early-stage distillation and later-stage independent RL optimization are important, and that RetireOPD benefits from adaptively determining when to retire OPD.

Table 2: Retirement mechanism ablation.

Setting	Retire @	SR
RetireOPD (Ours)	**60**	**93.8**
w/o Rebound Indicator	50	89.0
w/o Relative Competence	100	89.0
w/o Retirement	–	82.8
w/o OPD (GRPO-only)	0	75.0

Table 3: Retirement threshold ablation.

Setting	Retire @	SR
$\delta = 0,\ \gamma = 0.9$ **(Ours)**	**60**	**93.8**
$\delta = 0.03,\ \gamma = 0.9$	65	89.1
$\delta = 0.05,\ \gamma = 0.9$	75	91.4
$\delta = 0,\ \gamma = 0.8$	55	90.6
$\delta = 0,\ \gamma = 1.0$	95	91.4

Sensitivity to Retirement Thresholds. We analyze how the thresholds γ and δ affect the retirement timing. As shown in Table 3, increasing either threshold generally leads to a later retirement point. A larger δ requires a more pronounced stagnation or rebound of the alignment gap, while a larger γ requires a larger relative competence ratio before retirement. Despite these changes in retirement timing, the final performance remains relatively stable across different threshold values, indicating that RetireOPD is not highly sensitive to the choice of thresholds.

Adaptive Retirement vs. Linear Annealing. We further compare our method with a Linear Annealing (Tan et al., 2026) strategy, which replaces adaptive retirement with a predefined linear schedule (Figure 6). RetireOPD exhibits a more stable training trajectory, with performance continuing to improve throughout training, whereas Linear Annealing shows pronounced fluctuations. It also achieves a higher final success rate, demonstrating that adaptively determining when to retire OPD is more effective than relying on a fixed linear schedule.

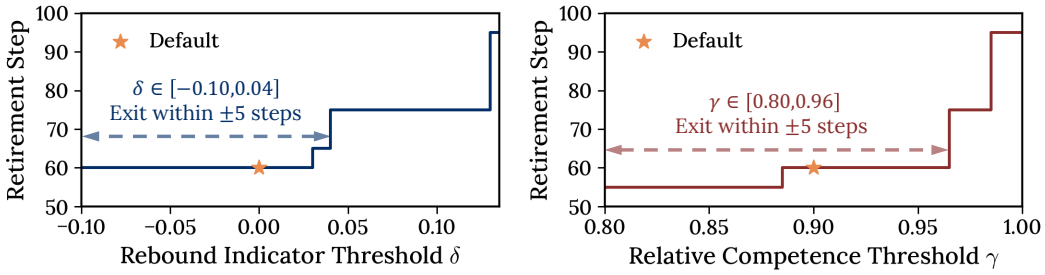

Figure 7: Robustness of the retirement mechanism to threshold variations. **Left:** Sensitivity to the rebound indicator threshold δ. **Right:** Sensitivity to the relative competence threshold γ.

4.5 Robustness to Retirement Thresholds

Figure 7 examines the robustness of the retirement point to variations in δ and γ. Across a relatively wide range of values, the retirement point remains stable: varying δ from -0.1 to 0.04 results in a similarly small variation of around 5 steps, while varying γ from 0.8 to 0.96 changes the retirement step by at most ± 5 steps. This stability indicates that RetireOPD is robust to the choice of retirement thresholds, with the adaptive mechanism consistently identifying a similar transition point despite substantial changes in the threshold values.

5 Conclusion

We propose RetireOPD based on two observations: a skill-conditioned teacher without dedicated training may fail to effectively exploit privileged skills, while continued on-policy distillation can constrain the student's performance by the teacher's capability ceiling. RetireOPD first trains a dedicated skill-conditioned teacher, then uses it to train a skill-free student, with the adaptive retirement mechanism that removes teacher supervision during training, freeing the student from the teacher's capability constraints. Experimental results show that our method consistently outperforms pure reinforcement learning, distillation-based, and hybrid approaches, validating its ability to effectively transfer privileged skills while avoiding the limitations imposed by continued teacher supervision.

AI USE STATEMENT

We used AI assistants in a limited supporting role for language polishing, formatting, coding and debugging, and checking submission materials. AI assistants were not used to develop the research methodology, design experiments, interpret results, or make scientific decisions. All AI-assisted outputs were carefully reviewed and verified by the authors before being incorporated. The authors made the final decisions regarding the research methodology, experiments, analysis, and presentation, and take full responsibility for the content and results of this work.

ETHICS STATEMENT

This work studies reinforcement learning and on-policy distillation for training language-based agents. Our research does not involve human subjects, private data, or personally identifiable information. All datasets and environments used in our experiments are publicly available benchmarks and open-source resources. Our method is developed and evaluated in controlled benchmark environments. While the proposed training techniques may potentially be extended to more general agentic settings, their behavior and safety in real-world applications are beyond the scope of this work. Deployments in such settings may introduce additional risks and therefore require appropriate evaluation, safeguards, and human oversight.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, Section 3 describes the training objective and methodology of RetireOPD. Appendix A.2 provides the pseudocode, while Appendix A.3 details the hyperparameter settings. Appendix A.5 further describes the teacher construction procedure, and Appendix A.8 provides the complete prompt templates. The supplementary materials include the training and evaluation code, corresponding launch scripts, and a version-pinned software environment.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

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

Chishui Chen, Yaoyou Fan, Te Sun, Yi Yang, Chenghao Sun, Delin Mao, Hongbo Qiao, Zuowei Zhang, Junxi Wang, Chenxing Sun, et al. Look ahead before you distill: Future trajectory validation of teacher guidance for agentic on-policy distillation. *arXiv preprint arXiv:2608.01953*, 2026.

Ken Ding. Hdpo: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*, 2026.

Yifan Ding, Xincheng Wei, Yoshua Y Li, Ziheng Li, Yuquan Lu, Siyu Zhang, Dongsheng Ma, Rongxiang Weng, Xunliang Cai, and Yun Chen. Saf-opd: Stable advantage fusion for on-policy distillation. *arXiv preprint arXiv:2607.29209*, 2026.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 16981–17017, 2026a.

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, 2026b. URL https://arxiv.org/abs/2609.28145.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/420c9f777c0b4f78d515e53cf74d58b2-Abstract-Conference.html.

Zhuowen Han, Jinwei Xiao, Zhengxi Lu, Renren Jin, Zhiyuan Yao, Yuxin Liu, Hongyan Hao, Yueqing Sun, Yu Yang, Qi Gu, et al. Distill where you fail: Recovering learning signals of negative rl-groups from adaptive teacher guidance. *arXiv preprint arXiv:2608.00782*, 2026.

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

Jaehoon Kim and Dongha Lee. Opsd compresses what rlvr teaches: A post-rl compaction stage for reasoning models. *arXiv preprint arXiv:2605.06188*, 2026.

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.

Gengsheng Li, Mao Zheng, Mingyang Song, Ruiqi Liu, Tianyu Yang, Jie Sun, Qiyong Zhong, Haiyun Guo, Junfeng Fang, Dan Zhang, et al. On-policy distillation with curriculum turn-level guidance for multi-turn agents. *arXiv preprint arXiv:2606.15912*, 2026b.

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

Junzhuo Liu, Weiwei Li, Jun Ling, and Peng Wang. When privileged guidance misaligns: State-matched routing and contextualized self-distillation for multi-turn agents. *arXiv preprint arXiv:2608.05219*, 2026.

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

Zhengxi Lu, Jiabo Ye, Fei Tang, Yongliang Shen, Haiyang Xu, Ziwei Zheng, Weiming Lu, Ming Yan, Fei Huang, Jun Xiao, et al. Ui-s1: Advancing gui automation via semi-online reinforcement learning. *arXiv preprint arXiv:2509.11543*, 2025.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, et al. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*, 2026a.

Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Chengcheng Han, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Skill0: In-context agentic reinforcement learning for skill internalization. *arXiv preprint arXiv:2604.02268*, 2026b.

Yingzi Ma, Zichen Zhu, Ming Jiang, and Chaowei Xiao. One student, many teachers: Multi-task on-policy distillation via soft-prompt privileged context. *arXiv preprint arXiv:2607.18293*, 2026.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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

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

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Qitai Tan, Zefang Zong, Mo Li, Yipeng Shi, Yang Li, and Peng Chen. Atod: Annealed turn-aware on-policy distillation for multi-turn agentic tasks, 2026. URL https://arxiv.org/abs/2606.27814.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, et al. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents. *arXiv preprint arXiv:2604.10674*, 2026a.

Jiaqi Wang, Wenhao Zhang, Weijie Shi, Yaliang Li, and James Cheng. Tcod: Exploring temporal curriculum in on-policy distillation for multi-turn autonomous agents. *arXiv preprint arXiv:2604.24005*, 2026b.

Zechuan Wang, Siyuan Lu, Hongxuan Zhang, Linjian Mo, Chenyi Zhuang, and Leilei Gan. Teach the magnitude, not the direction: Verifier-bounded credit assignment for multi-turn multi-step llm agents. *arXiv preprint arXiv:2608.13179*, 2026c.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, et al. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*, 2025.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, et al. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning. *arXiv preprint arXiv:2602.08234*, 2026.

Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, et al. Mimo-v2-flash technical report. *arXiv preprint arXiv:2601.02780*, 2026.

Xingrun Xing, Haoqing Wang, Boyan Gao, Ziheng Li, and Yehui Tang. Trust region on-policy distillation. *arXiv preprint arXiv:2606.01249*, 2026.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation. *arXiv preprint arXiv:2604.14084*, 2026b.

Jingxiao Yang, Wangjie Gan, Yingxuan Zhuang, Wenqi Zhang, Jintao Chen, and Xuhong Zhang. Reconciling process supervision with outcome-based credit in agentic policy optimization. *arXiv preprint arXiv:2608.31077*, 2026.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

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*, 2026a.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026b.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, et al. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*, 2026.

Haobo Zhang, Kelong Mao, Sulong Xu, Simiu Gu, and Zhicheng Dou. Learning from online user feedback for shopping agents. *arXiv preprint arXiv:2608.11604*, 2026a.

Jie Zhang, Jingxiao Yang, Zhehao Huang, Yuhang Liu, and Xiaolin Huang. When and where to trust the teacher: Unifying on-policy distillation and grpo through entropy-calibrated credit assignment, 2026b. URL https://arxiv.org/abs/2609.28385.

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference distillation for agent reinforcement learning. *arXiv preprint arXiv:2605.27140*, 2026c.

Zhiwei Zhang, Zechen Sun, Fei Zhao, Kang Peng, Bin Liang, Huayu Deng, Yao Hu, Kam-Fai Wong, and Mu Chuan. Verify before you distill: Prompt-level teacher gating for on-policy distillation. *arXiv preprint arXiv:2609.02998*, 2026d.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Bo Peng, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. Sage-opd: Selective agent-guided intervention for multi-turn on-policy distillation. *arXiv preprint arXiv:2606.19659*, 2026.

A Appendix

A.1 Theoretical Analysis

Setup. We analyze GRPO-OPD interaction through a local teacher-regularized view. Let

$$J(\theta) = \mathbb{E}_{x,\tau\sim\pi_\theta}[R(\tau)],$$

$$D(\theta) = \mathbb{E}_{x,y\sim\pi_\theta}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid x, y_{<t}) \| \pi_T(\cdot \mid x, c^+, y_{<t})\big)\right],$$

$$F_\lambda(\theta) = J(\theta) - \lambda D(\theta), \qquad g = \nabla_\theta J, \qquad h = \nabla_\theta D.$$

Ignoring finite-sample noise and assuming inactive GRPO clipping locally, the expected update is

$$\theta^+ = \theta + \eta(g - \lambda h),$$

where g is the reward-driven direction and $-h$ reduces the teacher–student discrepancy.

Interaction between GRPO and OPD. A first-order expansion gives

$$\begin{aligned} J(\theta^+) - J(\theta) &= \big\langle \nabla_\theta J(\theta), \theta^+ - \theta \big\rangle + O(\eta^2) \\ &= \eta \,\langle g,\, g - \lambda h\rangle + O(\eta^2) \\ &= \eta\left(\|g\|^2 - \lambda\langle g, h\rangle\right) + O(\eta^2), \end{aligned}$$

$$\langle g, h\rangle < 0 \Rightarrow \text{OPD accelerates reward optimization},$$
$$\langle g, h\rangle > 0 \Rightarrow \text{OPD conflicts with reward optimization}.$$

Moreover, if $\lambda\langle g, h\rangle > \|g\|^2$, the joint update decreases the expected return to first order.

Discrepancy as a Conflict Signal. Under joint optimization,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} D(\theta) &= \left\langle \nabla_\theta D(\theta), \frac{\mathrm{d}\theta}{\mathrm{d}t} \right\rangle \\ &= \langle h,\, g - \lambda h\rangle \\ &= \langle g, h\rangle - \lambda\|h\|^2. \end{aligned}$$

Thus, a non-decreasing discrepancy indicates local conflict between the GRPO direction g and the OPD direction $-h$. If $h \approx 0$, OPD already provides little additional first-order guidance. Discrepancy stagnation therefore signals either emerging conflict or a diminishing OPD signal.

Benefit of Teacher Retirement. Suppose GRPO and OPD are locally conflicting, i.e., $\langle g, h\rangle > 0$. Consider

$$\begin{aligned} \theta^+_{\mathrm{joint}} &= \theta + \eta(g - \lambda h), \\ \theta^+_{\mathrm{retirement}} &= \theta + \eta g. \end{aligned}$$

Their expected returns satisfy

$$\begin{aligned} J(\theta^+_{\mathrm{retirement}}) - J(\theta^+_{\mathrm{joint}}) &= \Big[J(\theta) + \eta\|g\|^2\Big] \\ &\quad - \Big[J(\theta) + \eta\big(\|g\|^2 - \lambda\langle g, h\rangle\big)\Big] + O(\eta^2) \\ &= \eta\lambda\langle g, h\rangle + O(\eta^2). \end{aligned}$$

Since $\langle g, h\rangle > 0$, teacher retirement yields a larger first-order reward improvement than continued joint optimization for sufficiently small η.

Implication for Adaptive Retirement. The analysis shows that OPD accelerates learning while teacher guidance aligns with reward optimization, whereas stagnation of the teacher-student discrepancy indicates diminishing OPD benefit or emerging conflict with GRPO. We therefore use stagnation of $\bar{K}_m$ as the retiring signal and additionally require sufficient student competence to avoid premature retirement caused by early training noise.

A.2 ALGORITHM

The complete training procedure of our method, including privileged teacher construction, joint GRPO-OPD training, and adaptive teacher retirement, is summarized in Algorithm 1.

Algorithm 1 RETIREOPD Training Pipeline

Require: Base policy π_0; training set $\mathcal{D} = \{(x, c^+)\}$; validation set $\mathcal{D}_{\text{val}}$; environment reward R; group size G; OPD weight λ; monitoring window W; exit thresholds γ, δ

▷ **Phase 1: Privileged Teacher Construction**

Initialize teacher policy $\pi_\phi \leftarrow \pi_0$
for each teacher training iteration **do**
  Sample $(x, c^+) \sim \mathcal{D}$
  $\{y^{(i)}\}_{i=1}^{G} \sim \pi_\phi(\cdot \mid x, c^+)$; $R^{(i)} \leftarrow R(x, y^{(i)})$
  $\hat{A}^{(i)} \leftarrow \dfrac{R^{(i)} - \mu_G}{\sigma_G}$, $r_t^{(i)} \leftarrow \dfrac{\pi_\phi(y_t^{(i)} \mid x, c^+, y_{<t}^{(i)})}{\pi_{\phi_{\text{old}}}(y_t^{(i)} \mid x, c^+, y_{<t}^{(i)})}$
  Update ϕ by minimizing $\mathcal{L}_{\text{GRPO}}(\phi; \{\hat{A}^{(i)}, r_t^{(i)}\})$
end for
$\pi_T \leftarrow \text{StopGrad}(\pi_\phi)$; $SR_T \leftarrow \text{Success}(\pi_T, \mathcal{D}_{\text{val}}; c^+)$

▷ **Phase 2: Joint GRPO-OPD with Adaptive Retirement**

$\pi_\theta \leftarrow \pi_0$; RETIRED $\leftarrow$ **false**; $m \leftarrow 0$
for each student training iteration n **do**
  Sample $(x, c^+) \sim \mathcal{D}$
  $\{y^{(i)}\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid x)$; $R^{(i)} \leftarrow R(x, y^{(i)})$ ▷ **Skill-free on-policy rollout**
  $\hat{A}^{(i)} \leftarrow \dfrac{R^{(i)} - \mu_G}{\sigma_G}$, $r_t^{(i)} \leftarrow \dfrac{\pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)})}{\pi_{\theta_{\text{old}}}(y_t^{(i)} \mid x, y_{<t}^{(i)})}$ ▷ **GRPO objective**
  if not RETIRED **then**
    $\Delta_t^{(i)} \leftarrow \log \pi_T(y_t^{(i)} \mid x, c^+, y_{<t}^{(i)}) - \log \pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)})$
    $K_n \leftarrow \dfrac{1}{G} \sum_{i=1}^{G} \dfrac{1}{|y^{(i)}|} \sum_{t=1}^{|y^{(i)}|} \Delta_t^{(i)}$; $\mathcal{L}_{\text{OPD}} \leftarrow -K_n$ ▷ **On-policy distillation**
    Update θ by minimizing $\mathcal{L}_{\text{GRPO}} + \lambda \mathcal{L}_{\text{OPD}}$
  else
    Update θ by minimizing $\mathcal{L}_{\text{GRPO}}$
  end if
  if $n \bmod W = 0$ **and not** RETIRED **then**
    $m \leftarrow m + 1$; $\bar{K}_m \leftarrow \dfrac{1}{W} \sum_{n=mW}^{mW+W-1} K_n$
    $SR_m \leftarrow \text{Success}(\pi_\theta, \mathcal{D}_{\text{val}})$
    if $m \geq 2$ **then**
      $\rho_m^{(K)} \leftarrow \dfrac{\bar{K}_m - \bar{K}_{m-1}}{\bar{K}_m}$; $\eta_m \leftarrow \dfrac{SR_m + SR_{m-1}}{2SR_T}$
      if $\rho_m^{(K)} \geq \delta$ **and** $\eta_m \geq \gamma$ **then** ▷ **Adaptive teacher retirement**
        RETIRED $\leftarrow$ **true**
      end if
    end if
  end if
end for
return skill-free student policy π_θ

A.3 HYPERPARAMETER SETTINGS

We implement all methods under a unified training setup to ensure a consistent comparison. Table 4 summarizes the optimization, data, and method-specific hyperparameters used in our experiments. Methods with identical configurations are grouped into the same column for compact presenta-

tion. G denotes the group size, λ is the distillation loss coefficient, and α_{KL} is the KL-divergence penalty coefficient. OP(S)D collectively denotes OPD and OPSD, while GRPO+OP(S)D denotes GRPO+OPD and GRPO+OPSD. GRPO, GiGPO and RLOO are grouped because they use the same hyperparameter configuration.

Table 4: Hyperparameter settings. A dash indicates that the corresponding hyperparameter is not applicable.

	Method				
Hyperparameter	**RetireOPD**	**GRPO / GiGPO / RLOO**	**PPO**	**OP(S)D**	**GRPO+OP(S)D**
Learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Critic learning rate	–	–	1×10^{-5}	–	–
G	8	8	–	8	8
λ	0.01	–	–	0.01	0.01
α_{KL}	0.01	0.01	0.01	0.01	0.01
Train batch size	16	16	16	16	16
Validation data size	128 (ALFWorld) / 256 (WebShop)				
Competence threshold	0.9	–	–	–	–
Window size	5	–	–	–	–
Training steps	150	150	150	150	150

A.4 Analysis of Teacher Restoration after Retirement

Does Teacher Supervision Become Useful Again? We further examine whether the need for teacher supervision may re-emerge after retirement. After the retirement criterion is triggered at step 60, we continue reward-only training and restore teacher supervision at step 90. As shown in Figure 8, restoring the teacher does not yield additional improvement, whereas the model that keeps the teacher retired continues to improve and eventually achieves a higher success rate. This observation suggests that, in this setting, the retirement criterion captures a relatively persistent change in the utility of teacher supervision rather than a short-lived fluctuation. Within the remaining training horizon, reintroducing the same teacher does not recover the benefit observed during the earlier stage of training.

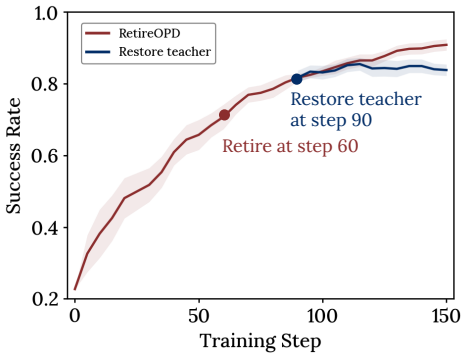

Figure 8: Teacher restoration after retirement.

A.5 Details of Privileged Teacher Construction

We provide additional details of the privileged teacher construction used throughout our experiments.

Teacher Initialization and Skill Conditioning. For each model scale, the teacher is initialized from the same instruction-tuned backbone as the student. The key difference is that the teacher additionally receives task-relevant privileged skill context c^+ throughout interaction. Specifically, at each decision step, the teacher conditions its action on the task input, interaction history and c^+, whereas the student never has access to such privileged information. The same skill format is used throughout teacher training and evaluation.

Teacher Optimization. We train the teacher using GRPO with privileged skills provided as part of its context, corresponding to the Skill-GRPO setting described in Section 3.1. Apart from the additional skill context, teacher training follows the same environment interaction and reward definition used for student optimization. For each sampled group of trajectories, environment rewards are used to compute group-relative advantages and update the skill-conditioned policy. The teacher

uses the same hyperparameter configuration as GRPO reported in Table 4, including the group size, learning rate and generation settings. During evaluation, the teacher is also provided with the privileged skill context c^+, allowing us to directly assess its ability to leverage privileged information for task solving. We report this evaluation setting as Skill-GRPO* in Table 1.

Checkpoint Selection and Downstream Usage. We periodically evaluate the teacher on the validation set and select the checkpoint according to the corresponding task metric. The selected teacher is then frozen and used throughout student training to provide token-level supervision. Its validation performance is also recorded as the teacher reference performance used by the adaptive retirement criterion. No teacher parameters are updated during student optimization.

The effectiveness of explicitly optimizing the skill-conditioned teacher, compared with directly prompting the base model with the same privileged skills, is examined separately in the ablation study in Section 4.4.

A.6 Ablation on OPD Weight

Figure 9 studies the effect of the distillation weight λ. A relatively large weight ($\lambda = 0.1$) leads to the fastest improvement during early training, showing that stronger teacher supervision can effectively facilitate initial policy learning. However, its improvement slows down in later stages and results in a lower final success rate. In contrast, a small weight ($\lambda = 0.001$) provides weaker guidance and therefore learns substantially more slowly at the beginning, although it gradually catches up as training proceeds. The default setting $\lambda = 0.01$ achieves a better balance between these two regimes, maintaining effective early-stage guidance while allowing continued reward-driven improvement, and consequently achieves the best final performance. These dynamics further illustrate that the preferred strength of teacher supervision changes over training, motivating an adaptive mechanism for removing teacher guidance once it becomes less beneficial.

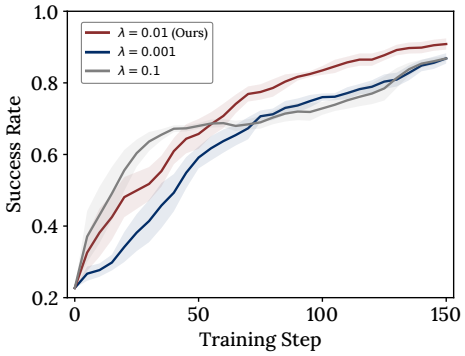

Figure 9: **Sensitivity to the distillation weight λ.** Success rate during training under different fixed distillation weights. Shaded regions indicate variability across runs.

A.7 Case Study

To provide a more fine-grained view of how teacher supervision affects policy learning, we examine token-level teacher scores on representative ALFWorld trajectories. We focus on two complementary cases: how privileged skills improve supervision during early exploration, and how the same teacher supervision can become restrictive after the student has acquired sufficient task competence.

Skill-informed supervision during early exploration. Figure 10 compares our skilled teacher with the OPSD teacher on the same student trajectory. While both assign similar scores to routine actions, clear differences emerge at decision-critical states requiring task-specific knowledge. At Step 8, the agent reaches the sink holding the target lettuce, and the retrieved skill specifies cleaning before placement. Our teacher assigns much higher likelihood to `clean` (-0.826) than the OPSD teacher (-16.625), showing that explicit teacher optimization better translates privileged skills into effective token-level supervision.

Teacher conflict after competence improves. Figure 11 shows a successful trajectory in which the student takes a task-effective action that is strongly disfavored by the teacher. In particular, the transition toward the microwave receives a low teacher score despite leading to successful task completion. This illustrates that, once the student becomes sufficiently competent, continued teacher matching can conflict with reward-driven optimization, motivating adaptive teacher exit.

Step	State summary	Retrieved skill	Skilled teacher score	OPSD teacher score
0	initial search	Clean before placing	`go to drawer 1`	`go to drawer 1`
1	drawer search	Search likely containers	`open drawer 1`	`open drawer 1`
2	cabinet search after empty drawer	Search likely containers	`open cabinet 10`	`open cabinet 10`
3	moving to likely food surface	Search likely food surfaces	`go to diningtable 1`	`go to diningtable 1`
4	lettuce found on dining table	Pick target object	`take lettuce 2 from diningtable 1`	`take lettuce 2 from diningtable 1`
5	checking object state	Verify object state	`examine lettuce 2`	`examine lettuce 2`
6	lettuce held; cleaning required	Use sink for cleaning	`go to sinkbasin 1`	`go to sinkbasin 1`
7	before cleaning	Verify object state	`examine lettuce 2`	`examine lettuce 2`
8	at sink; cleaning should happen	Use sink for cleaning Clean before placing	`clean lettuce 2 with sinkbasin 1`	`clean lettuce 2 with sinkbasin 1`
9	after cleaning; place target object	Place after cleaning	`move lettuce 2 to countertop 1`	`move lettuce 2 to countertop 1`

Figure 10: **Early-stage teacher comparison on ALFWorld.** Both teachers score the same student-generated action tokens. The retrieved skill provides privileged guidance available only to the skilled teacher. At the critical cleaning step, the skilled teacher assigns substantially higher likelihood to the key token `clean`.

Step	State summary	Student action colored by teacher score	Env. return
1	searching for tomato	`take tomato 2`	1
2	holding tomato; needs cooling	`go to fridge 1`	1
3	holding cooled tomato; microwave needed	`go to microwave 1`	1
4	near microwave	`put tomato in microwave`	1

Figure 11: **Token-level teacher conflict on a successful ALFWorld trajectory.** Token scores indicate the teacher's preference for the student's generated actions. Despite receiving positive environment feedback, the student's correct transition toward the microwave receives strong negative teacher supervision.

A.8 PROMPT TEMPLATES

We use a unified prompt format for agent interaction across environments. The student receives task observations and action history, while the teacher additionally accesses privileged skill information during OPD. The teacher receives the same student trajectory prefix, with privileged skill information prepended to the prompt.

Prompt template on ALFWorld

You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: {task_description}
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is: {current_observation}
Your admissible actions of the current situation are: [{admissible_actions}].
Now it's your turn to take an action. You should first reason step-by-step about the current situation. This reasoning process MUST be enclosed within `<think> </think>` tags. Once you've finished your reasoning, you should choose an admissible action for the current step and present it within `<action> </action>` tags.

Figure 12: Student prompt template for the ALFWorld task environment.

Prompt template on WebShop

You are an expert autonomous agent operating in the WebShop e-commerce environment.
Your task is to: {task_description}.
Prior to this step, you have already taken {step_count} step(s). Below are the most recent {history_length} observations and the corresponding actions you took: {action_history}
You are now at step {current_step} and your current observation is: {current_observation}.
Your admissible actions of the current situation are: [{available_actions}].
Now it's your turn to take one action for the current step. You should first reason step-by-step about the current situation, then think carefully which admissible action best advances the shopping goal. This reasoning process MUST be enclosed within `<think> </think>` tags. Once you've finished your reasoning, you should choose an admissible action for the current step and present it within `<action> </action>` tags.

Figure 13: Student prompt template for the WebShop task environment.

Privileged teacher input template

[Privileged Skill Information]
{skill_text}
{student_prompt}
{student_response_prefix}

Figure 14: Teacher-side input used for OPD.

A.9 Training Dynamics

We present the full training dynamics of RetireOPD across all model scales and environments in Figures 15–18.

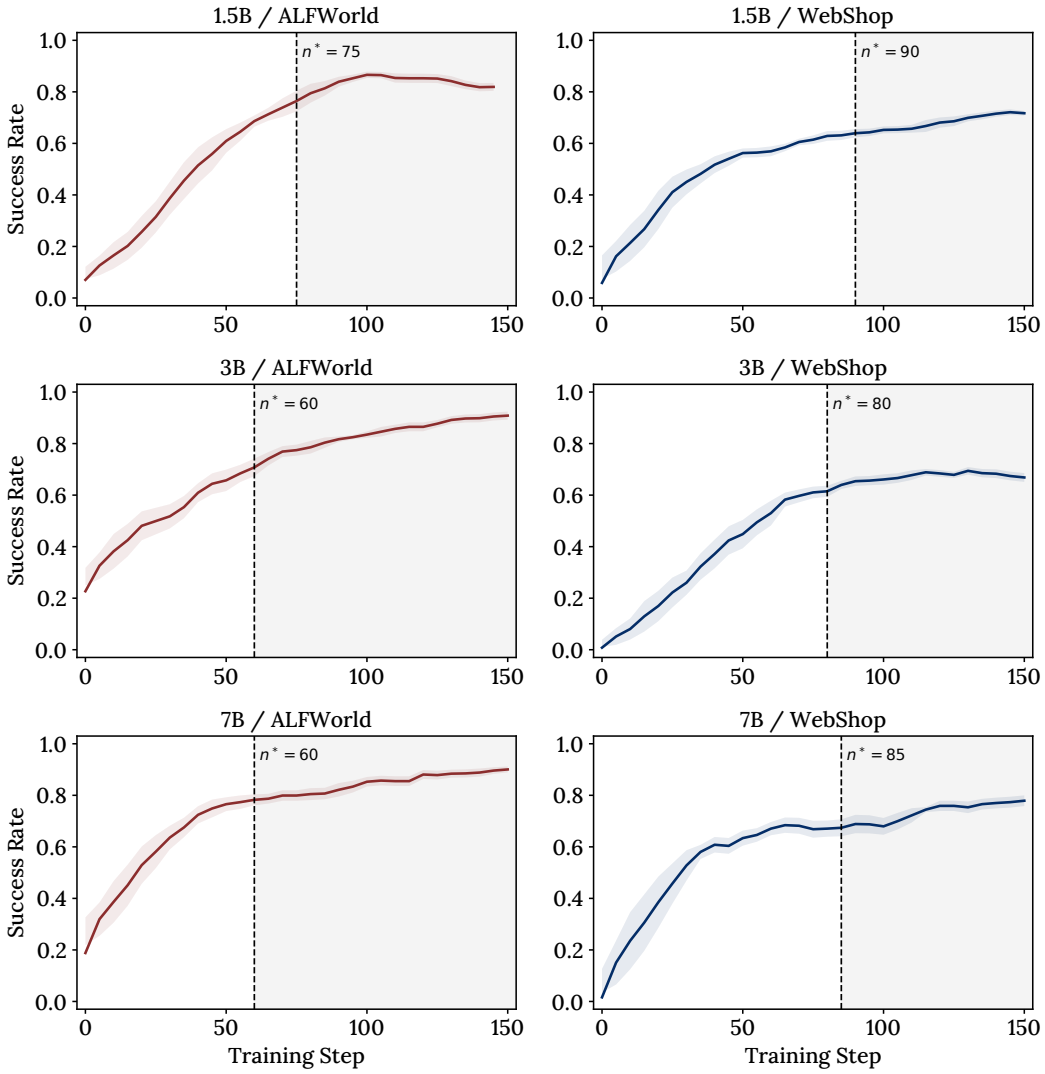

Figure 15: **Validation Score** when training with Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct on ALFWorld and WebShop. n^* denotes the retirement step.

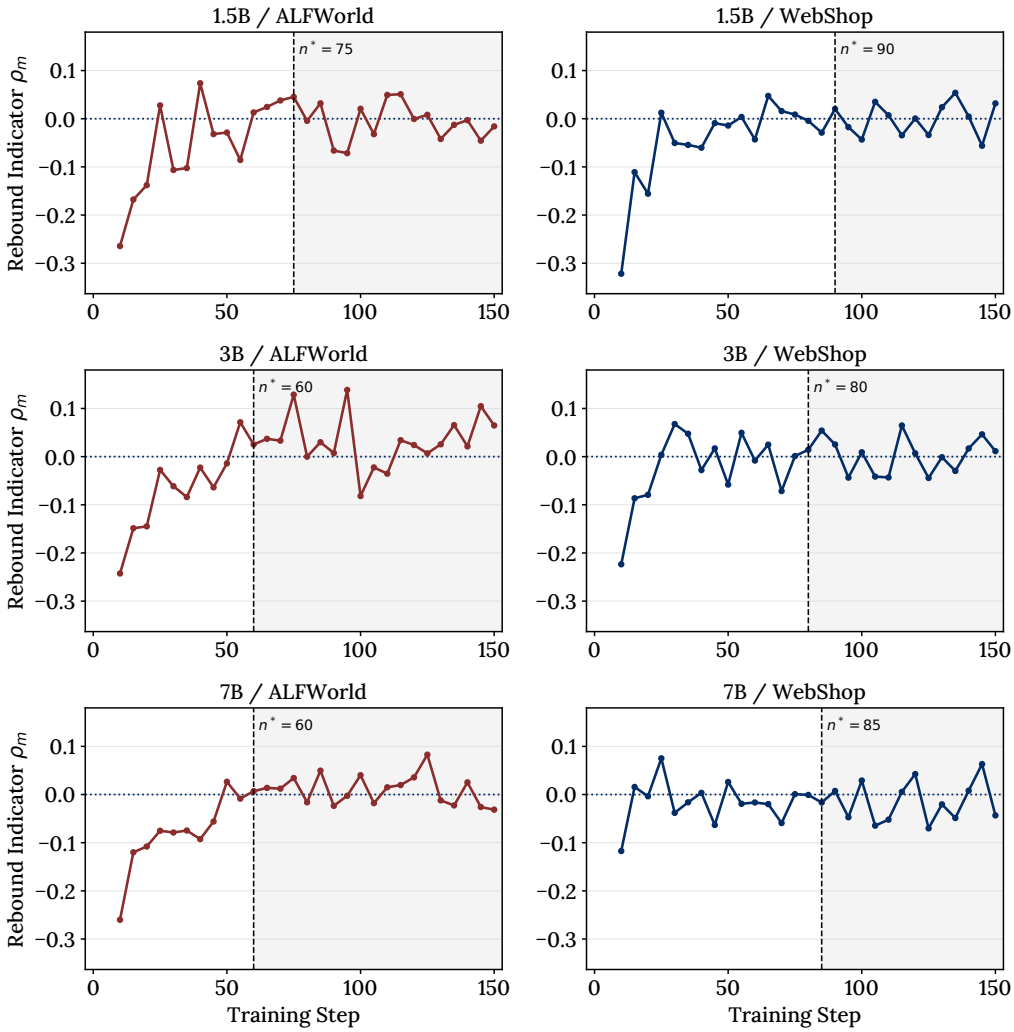

Figure 16: **Rebound Indicator Curve** when training with Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct on ALFWorld and WebShop. n^* denotes the retirement step, and Rebound Indicator threshold δ is 0.

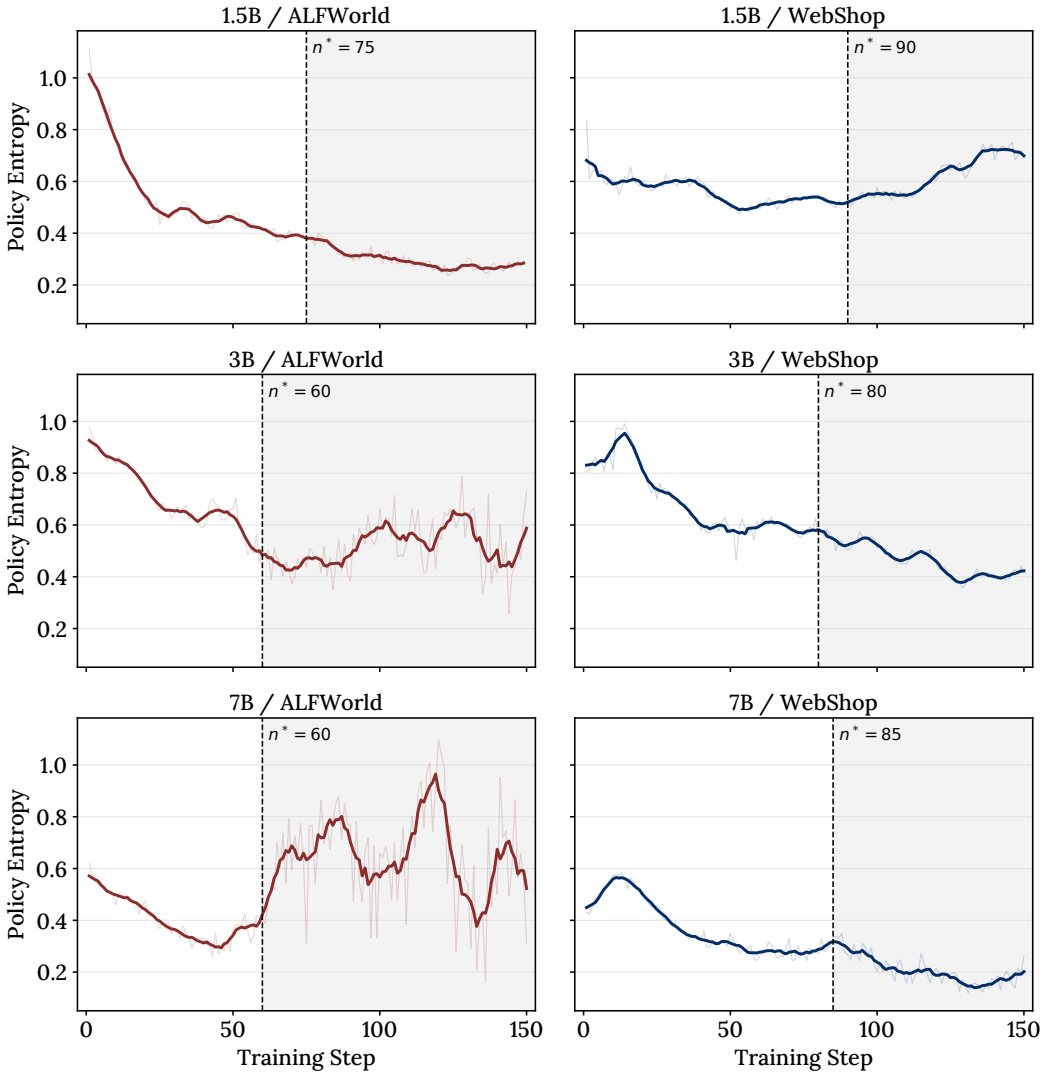

Figure 17: **Entropy Curve** when training with Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct on ALFWorld and WebShop. n^* denotes the retirement step.

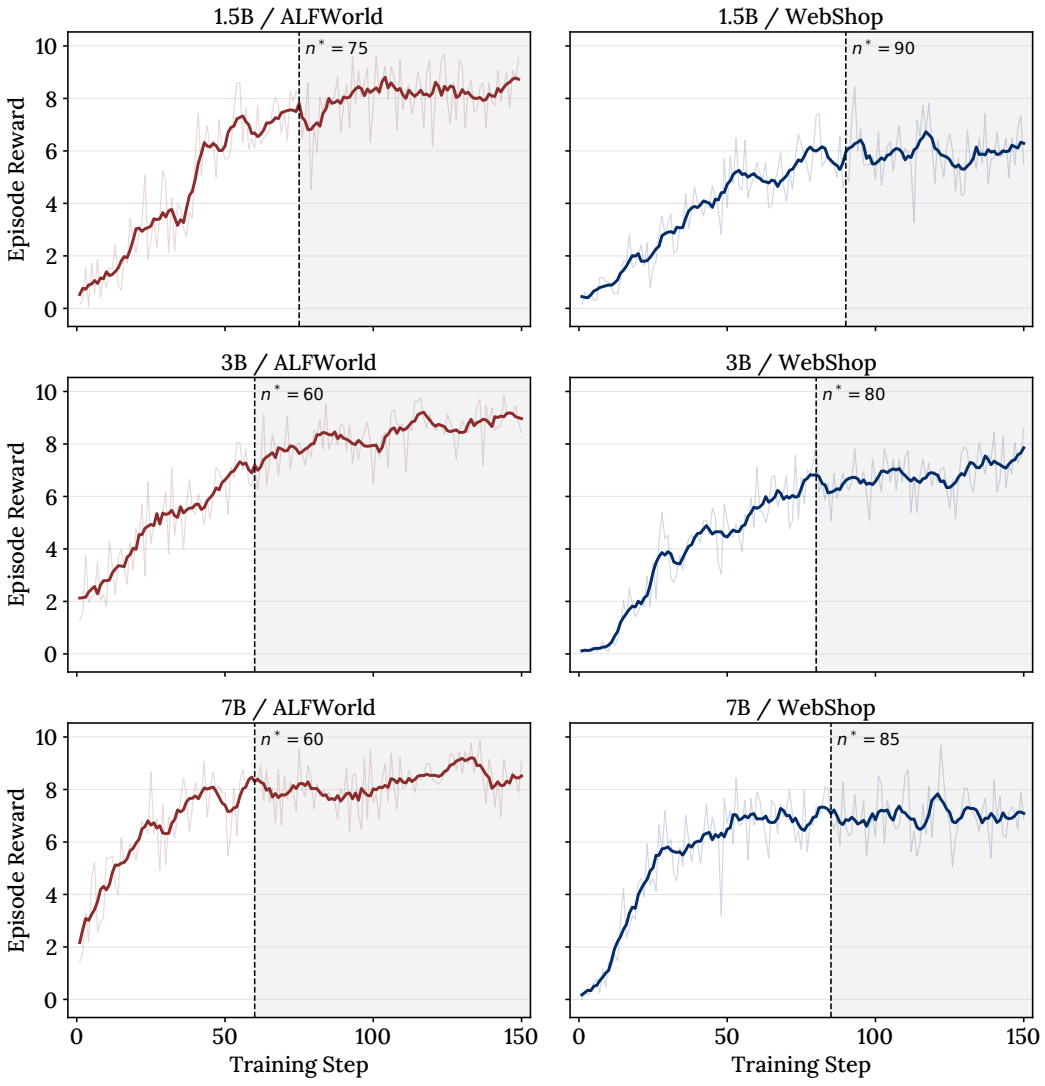

Figure 18: **Reward Curve** when training with Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct on ALFWorld and WebShop. n^* denotes the retirement step.