AGENTOPSD: RECURSIVE SELF-DISTILLATION FOR AGENTIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning(RL) with verifiable rewards constructs trajectory-level advantage, yet often fails to credit the few pivotal decisions that drive outcomes in long-horizon multi-turn agentic RL. Some recent works introduce privileged self-distillation into credit assignment for RL, offering denser supervision, but it still remains unclear how such a local signal should express *sequential* credit. We therefore propose **AgentOPSD**, a critic-free recursive turn-level credit assignment for agentic reinforcement learning. AgentOPSD aggregates token-level teacher-student log-probability gaps into turn-level evidence, recursively updates a Bayesian belief state in log-odds space. This provides a principled reweighting scheme that transforms sparse outcome supervision into turn-level credit signals and identifies pivotal turns by the marginal revision between consecutive states while remaining fully compatible with standard policy optimization, requiring neither additional rollouts. We evaluate AgentOPSD on ALFWorld, WebShop, and Search-QA with three Qwen model scales (Qwen2.5-3B/7B and Qwen3-1.7B). AgentOPSD improves over GRPO and strong self-distillation baselines, reaching 89.1% success on ALFWorld with Qwen2.5-7B, and ablations attribute the gains to turn-level aggregation and history-dependent recursive belief updates.

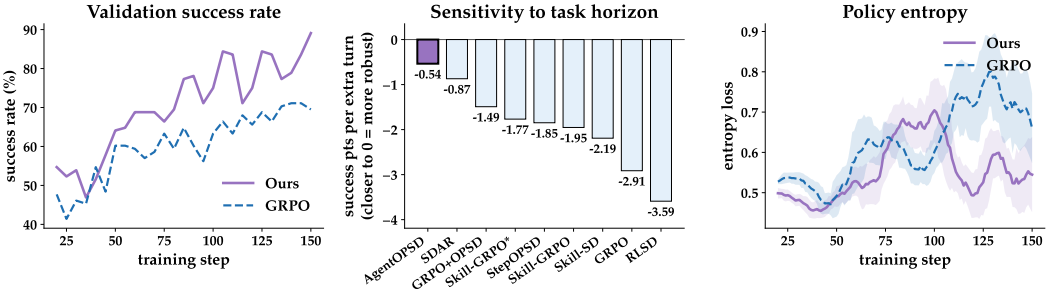

Figure 1: **Training dynamics and horizon-robustness of AgentOPSD** on Qwen2.5-7B-Instruct / ALFWorld. **(a)** Validation success rate over training. **(b)** Sensitivity to task horizon: success points lost per extra turn (OLS slope of per-sub-task success against the measured mean turns of successful episodes). **(c)** Policy entropy over training.

1 INTRODUCTION

Agentic post-training has become an important approach to improving the ability of large language models to solve complex tasks (Guo et al., 2025; Team et al., 2025; Yang et al., 2025; Comanici et al., 2025; Team et al., 2026b). Unlike static, single-turn reasoning, agents must continuously interact with partially observable environments, at each turn acting on the current observation and receiving a new one as the environment transitions (Shen et al., 2023; Shi et al., 2025; Jimenez et al., 2023). However, many interactive environments provide verifiable rewards only after an entire trajectory terminates, forcing training algorithms to infer the contribution of each intermediate decision from a single sparse outcome. The decisions within a trajectory can nevertheless play substantially different

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.

roles: even successful trajectories may contain spurious, redundant, or misleading actions, whereas failed trajectories may still include useful reasoning.

Group-relative policy optimization methods such as GRPO (Shao et al., 2024; Yu et al., 2025) and its agentic variants (Dong et al., 2025; Feng et al., 2025) construct a trajectory-level advantage from outcome rewards and broadcast it uniformly across the trajectory. Such uniform credit cannot distinguish a few pivotal decisions from routine operations. This limitation becomes increasingly pronounced as the interaction horizon grows. Turn-level credit assignment is therefore essential for identifying the decisions that meaningfully influence the outcome and providing more precise supervision throughout long-horizon interactions.

A complementary line of work provides denser, token-level supervision. On-policy distillation (Ye et al., 2026a; Yang et al., 2026c; Team et al., 2026a; Li et al., 2026; Yang et al., 2026b) trains a student on its own rollouts under a teacher, while its *self*-distillation variants (Zhao et al., 2026; He et al., 2026) remove the need for a separate teacher by conditioning the same policy on privileged information available only during training (Lu et al., 2026c). Recent studies further incorporate OPSD signals into reinforcement learning as an auxiliary source of supervision (Lu et al., 2026b; Wang et al., 2026a).

However, applying OPSD to agentic reinforcement learning introduces two mismatches. First, OPSD's token-level signals are not naturally aligned with agentic interaction (Lu et al., 2026b), where multiple tokens jointly form an action and the environment responds only at turn boundaries. Second, even existing step-aware methods consider each turn in isolation (Zhang et al., 2026), without accounting for the evidence accumulated through preceding interactions. The central challenge is therefore to transform local OPSD signals into history-dependent turn-level credit.

Our key insight is that **the credit of a turn should be determined not by its local signal in isolation, but by how much that signal changes the estimated probability of eventual success**. To formalize this intuition, we interpret the per-turn self-distillation gap as new evidence that induces a Bayesian belief update (Åström, 1965; Kaelbling et al., 1998). We define the corresponding belief state as the probability that the trajectory will ultimately succeed given the interaction history.

Based on this insight, we propose AgentOPSD (Recursive Self-Distillation for Agentic Reinforcement Learning), a turn-level credit-assignment method for long-horizon agents. AgentOPSD aggregates token-level teacher–student log-probability gaps into turn-level evidence. Starting from the average group success rate, it then recursively updates Bayesian belief state at each turn in log-odds space without additional rollouts or a learned critic. The outcome verifier determines the global direction of optimization, while the bayesian belief updates redistribute the trajectory-level learning signal across turns. We evaluate AgentOPSD on three interactive environments—ALFWorld (Shridhar et al., 2020), WebShop (Yao et al., 2022), and Search-QA (Jin et al., 2025)—and across three model scales. As shown in Figure 1, AgentOPSD consistently outperforms GRPO and strong self-distillation baselines. Further ablations show that gains from aggregating token-level signals at turn boundaries aligned with environment transitions, and transforming an isolated local gap into a recursive revision of belief.

Our contributions are summarized as follows:

- We formalize turn-level credit as the revision of a success belief induced by each turn. In log-odds space, this connects per-turn evidence to a recursive Bayesian update and reveals that an isolated self-distillation gap is not, by itself, sequential credit.
- We introduce AgentOPSD, which aggregates token-level teacher–student log-probability gaps into environment-aligned turn-level evidence and recursively propagates this evidence through the trajectory-success belief, without additional rollouts or a learned critic.
- Experiments and ablations across three interactive environments and three model scales demonstrate that AgentOPSD consistently outperforms GRPO and strong self-distillation baselines. Ablations further verify the complementary benefits of turn-boundary aggregation and recursive belief revision.

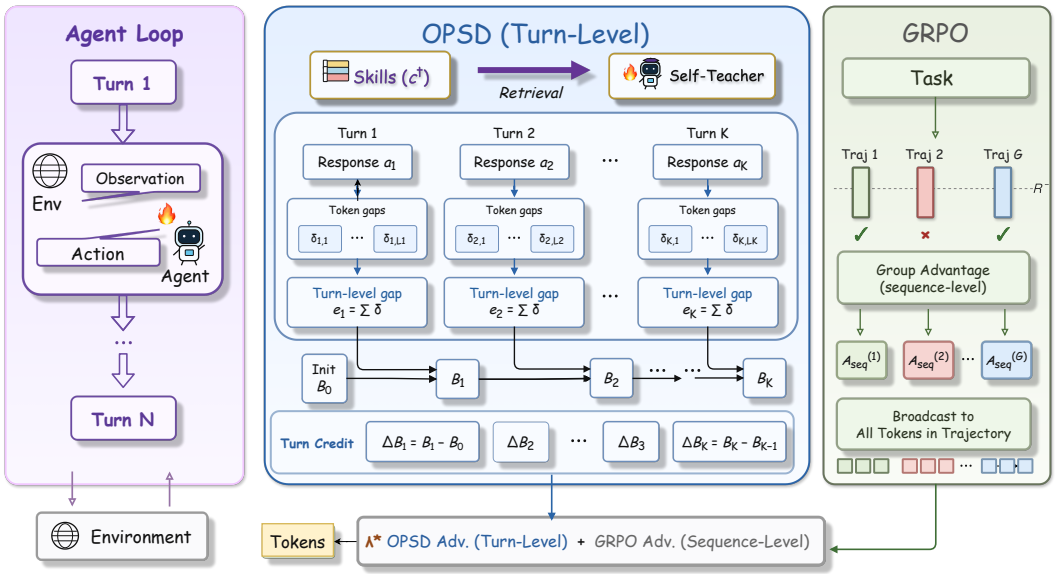

Figure 2: **Overview of AgentOPSD. Left:** the agent loop, interacting with the environment over turns $1, \ldots, K$. **Middle:** AgentOPSD converts GRPO's single sequence-level advantage into turn-level reshaped advantages in three steps: **(1)** aggregate the token-level teacher–student gaps $\delta_{k,t}$ within a turn into a turn-level gap e_k; **(2)** recursively update a belief state B_k (initialized from the group success rate) and read off its marginal revision $\Delta B_k = B_k - B_{k-1}$; **(3)** reshape the sequence-level advantage $A_{seq}^{(i)}$ per turn into $\tilde{A}_k^{(i)}$. **Right:** vanilla GRPO instead broadcasts the same $A_{seq}^{(i)}$ to every token/turn. Each token in turn k inherits $\tilde{A}_k$.

2 Methodology

2.1 Problem Setup

Given a task x and initial observation o_0, the agent starts from $s_1 = (x, o_0)$. At turn k, it samples where π_θ is current policy, s_k is its visible interaction history, $y_{k,t}$ is the t-th token of action a_k, and L_k is the action length. After observing o_k, the history becomes $s_{k+1} = (s_k, a_k, o_k)$. A K-turn episode forms $\boldsymbol{\tau} = (s_1, a_1, o_1, \ldots, s_K, a_K, o_K)$ and receives a binary outcome reward $R(\boldsymbol{\tau})$.

$$a_k = (y_{k,1}, \ldots, y_{k,L_k}) \sim \pi_\theta(\cdot \mid s_k), \tag{1}$$

For each task, group-relative policy optimization samples G trajectories and computes the sequence-level advantage. Here i indexes one of the G sampled trajectories, while $\bar{R}$ and $\widehat{\sigma}_R$ are the group reward mean and standard deviation, and ϵ_0 is a small positive constant (reused throughout to avoid division by zero or infinite log-odds). GRPO assigns $A_{\text{seq}}^{(i)}$ to every token in trajectory i, leaving turn-level credit unresolved.

$$A_{\text{seq}}^{(i)} = \frac{R^{(i)} - \bar{R}}{\widehat{\sigma}_R + \epsilon_0}, \qquad \bar{R} = \frac{1}{G}\sum_{j=1}^{G} R^{(j)}. \tag{2}$$

2.2 From Outcome Contribution to Bayesian Turn Evidence

Directly measuring the counterfactual contribution of turn k would require marginalizing the outcome reward over all possible continuations following a_k, which is intractable in long-horizon interactions. We therefore adopt a hindsight-based evidential perspective. Let C denote the event that the trajectory eventually succeeds. If a_k supports success, it should be more characteristic of successful behavior than of unsuccessful behavior. Bayes' rule expresses the resulting change in the belief about C as an action-side likelihood ratio (Åström, 1965; Kaelbling et al., 1998):

$$\operatorname{logit} p(C \mid s_k, a_k) - \operatorname{logit} p(C \mid s_k) = \log \frac{p(a_k \mid s_k, C)}{p(a_k \mid s_k, \neg C)}. \tag{3}$$

Here $\mathrm{logit}(u) = \log \frac{u}{1-u}$. The right-hand side is the ideal Bayes factor (Kass & Raftery, 1995) between the success-conditional and failure-conditional likelihoods of a_k. Its sign indicates whether the action increases or decreases support for eventual success.

Because these outcome-conditional behavioral distributions are not directly available, we estimate this Bayesian evidence retrospectively using a computable per-turn self-distillation contrast between a privileged, success-associated self-teacher and the standard student policy. This contrast provides a tractable proxy for the otherwise inaccessible belief update; we characterize the approximation and its conditions below. The student and teacher share parameters θ and score the same student-generated action (Zhao et al., 2026). Their token contexts are

$$h_{k,t} = (s_k, y_{k,<t}), \qquad h^+_{k,t} = (s_k, c^+, y_{k,<t}). \tag{4}$$

where $y_{k,<t}$ is the token prefix within turn k, and c^+ is a training-only retrieved skill describing useful subgoals and action patterns (Xia et al., 2026). The skill-conditioned branch approximates success-associated behavior, while the unconditioned branch provides the background likelihood.

For token $y_{k,t}$, define the detached likelihood contrast

$$\delta_{k,t} = \log \pi_\theta(y_{k,t} \mid h^+_{k,t}) - \log \pi_\theta(y_{k,t} \mid h_{k,t}). \tag{5}$$

Positive $\delta_{k,t}$ means that c^+ increases the likelihood of the generated token. Summing over the L_k tokens gives the turn-level evidence

$$e_k = \sum_{t=1}^{L_k} \delta_{k,t} = \log \frac{\pi_\theta(a_k \mid s_k, c^+)}{\pi_\theta(a_k \mid s_k)}. \tag{6}$$

Accordingly, e_k provides a tractable hindsight approximation to the ideal Bayesian turn evidence in Eq. (3), under the conditions detailed in Appendix A.1. More generally, Bayes' rule gives

$$\log \frac{p(a_k \mid s_k, C)}{p(a_k \mid s_k)} = \log \frac{p(C \mid s_k, a_k)}{p(C \mid s_k)}. \tag{7}$$

Thus, e_k can be interpreted as an evidential score whose sign indicates whether a_k raises or lowers support for eventual success. This sign-consistent Bayesian evidence is precisely the property that AgentOPSD relies on. We therefore treat e_k as a tractable Bayesian-inspired evidence proxy.

2.3 Recursive Belief Update

The local score e_k does not indicate whether the same evidence is pivotal or redundant given earlier turns. We therefore maintain a decaying evidence accumulator and measure each turn by how much it revises the current support state:

$$\begin{aligned} &B_0 = \mathrm{clip}(\bar{R}, \epsilon_0, 1-\epsilon_0), \quad c_0 = 0, \\ &c_k = \gamma\, c_{k-1} + e_k, \qquad \ell_k = \mathrm{logit}(B_0) + c_k = \mathrm{logit}(B_0) + \sum_{j=1}^{k} \gamma^{\,k-j} e_j, \end{aligned} \tag{8}$$

with $B_k = \sigma(\ell_k)$ and $\sigma(u) = (1+e^{-u})^{-1}$. Here $\bar{R} = S/G$ is the fraction of successful trajectories in the group of size G—the standard GRPO group mean (Prop. 7)—and B_0 clips it to $[\epsilon_0, 1-\epsilon_0]$ with $\epsilon_0{=}10^{-4}$ so its log-odds stay finite for all-correct or all-wrong groups. c_k is the accumulated evidence, and $\gamma \in (0, 1]$ is a decay factor that down-weights older turns geometrically; only the evidence c_k decays, while the prior $\mathrm{logit}(B_0)$ is retained at every step. Setting $\gamma{=}1$ recovers the undiscounted accumulation of a log-likelihood ratio familiar from sequential testing (Wald, 1945); $\gamma{<}1$ makes the state recency-weighted, so that evidence from many turns ago no longer pins the support level. Since e_k is estimated by the self-teacher (§2.2), B_k is treated as relative support rather than a calibrated success probability.

The importance of turn k is its marginal support revision:

$$\begin{aligned} \Delta B_k &= B_k - B_{k-1} = \sigma(\ell_k) - \sigma(\ell_{k-1}), \\ \Delta B_k &\approx B_{k-1}(1-B_{k-1})\big(e_k - (1-\gamma)\, c_{k-1}\big). \end{aligned} \tag{9}$$

The increment $\ell_k - \ell_{k-1} = e_k - (1-\gamma)c_{k-1}$ is the new evidence net of the decayed carry-over, and it is weighted by the current state sensitivity $B_{k-1}(1-B_{k-1})$: evidence has greatest effect under uncertainty and is suppressed once support saturates. At $\gamma{=}1$ this reduces to $B_{k-1}(1-B_{k-1})\, e_k$. We update at turn boundaries; a token-level variant is used only as an ablation.

Outcome-aligned recursive credit. We align the revision with the terminal update and read off its magnitude and direction:

$$q_k = \mathrm{sign}(A_{\mathrm{seq}})\,\Delta B_k. \tag{10}$$

The *magnitude* $|\Delta B_k| = |q_k|$ measures how much support the turn revises, while its *sign* $\mathrm{sign}(q_k)$ records whether that revision agrees with the verifier's outcome signal. After the within-trajectory standardization below, turns with above-average q_k are amplified and those below-average are attenuated; since the multiplier stays strictly positive, this never reverses the GRPO update direction.

2.4 Bounded Advantage Reshaping

The raw credit q_k only modulates the magnitude of the verifier-derived advantage. For trajectory i, we normalize its K_i turn credits and apply a bounded multiplier:

$$\begin{aligned}
\mu_q^{(i)} &= K_i^{-1}\sum_{j=1}^{K_i} q_j^{(i)}, & \sigma_q^{(i)} &= \sqrt{K_i^{-1}\sum_{j=1}^{K_i}\big(q_j^{(i)}-\mu_q^{(i)}\big)^2},\\
z_k^{(i)} &= \frac{q_k^{(i)}-\mu_q^{(i)}}{\sigma_q^{(i)}+\epsilon_0}, & w_k^{(i)} &= \mathrm{clip}\Big(1+bz_k^{(i)},\,1-b,\,1+b\Big),\\
\widetilde{A}_k^{(i)} &= A_{\mathrm{seq}}^{(i)}\big[(1-\lambda)+\lambda w_k^{(i)}\big], & b\in(0,1),&\quad \lambda\in[0,1].
\end{aligned} \tag{11}$$

Here $\mu_q^{(i)}$ and $\sigma_q^{(i)}$ are the within-trajectory mean and standard deviation, $z_k^{(i)}$ is the normalized credit (ϵ_0 stabilizes the normalization), b sets $w_k^{(i)} \in [1-b, 1+b]$, and λ controls reshaping strength.

Token t inherits $\widetilde{A}^{(i)}_{\kappa_i(t)}$, yielding

$$\begin{aligned}
\mathcal{L}_{\mathrm{AgentOPSD}}(\theta) &= -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{\sum_t M_{i,t}}\sum_t M_{i,t}\min\Big(r_{i,t}\widetilde{A}^{(i)}_{\kappa_i(t)},\,\mathrm{clip}(r_{i,t},1-\varepsilon,1+\varepsilon)\widetilde{A}^{(i)}_{\kappa_i(t)}\Big)+\beta\mathcal{L}_{\mathrm{KL}},\\
r_{i,t} &= \frac{\pi_\theta(y_{i,t}\mid h_{i,t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t}\mid h_{i,t})}.
\end{aligned} \tag{12}$$

Here $M_{i,t} \in \{0,1\}$ masks valid response tokens, $\kappa_i(t)$ maps token t to its turn, $r_{i,t}$ is the importance ratio against the rollout policy $\pi_{\theta_{\mathrm{old}}}$, ε is the clipping radius, and β is its coefficient. No separate distillation loss is introduced; the detached self-teacher signal acts only through $\widetilde{A}$.

3 Experiments

3.1 Experimental Setup

Benchmarks. We evaluate on three environments. *ALFWorld* (Shridhar et al., 2020) is a text-embodied benchmark over six household task categories—Pick and Place (Pick), Look at Object in Light (Look), Pick Clean then Place (Clean), Pick Heat then Place (Heat), Pick Cool then Place (Cool), and Pick Two and Place (Pick2). *Search-QA* follows the Search-R1 setup (Jin et al., 2025) and covers single-hop QA (NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023)) and multi-hop QA (HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), Bamboogle (Press et al., 2023)), with NQ and HotpotQA in-domain and the rest held out; retrieval uses E5 (Wang et al., 2022). *WebShop* (Yao et al., 2022) is an interactive online-shopping environment; we evaluate on the 128 fixed validation tasks of Feng et al. (2025).

Implementation. We train Qwen2.5-3B/7B-Instruct and Qwen3-1.7B on 8×H800 GPUs. The privileged skills are retrieved from the `SkillBank` of SkillRL (Xia et al., 2026) by keyword matching and are used only during training; inference uses no external skills. Prior B_0 set to the fraction of successful trajectories in each GRPO group (the standard group mean $\bar{R}$). All other optimization settings are shared with the SDAR baseline. Full training and AgentOPSD hyperparameters are listed in Appendix F (Table 3).

Table 1: **Performance on ALFWorld, Search-QA and WebShop.** We report success rate (%) on ALFWorld, accuracy (%) on Search-QA, and Score/Acc (%) on WebShop. skills are training-only unless marked with $*$ (validation with skills). AgentOPSD uses no skills at inference. **Best** and second-best are highlighted.

	ALFWorld							Search-QA								WebShop	
Method	**Pick**	**Look**	**Clean**	**Heat**	**Cool**	**Pick2**	**Avg**	**NQ**	**Triv**	**Pop**	**Hotp**	**2Wk**	**MuS**	**Bam**	**Avg**	**Score**	**Acc**
Qwen2.5-3B-Instruct																	
Vanilla	44.4	11.1	6.2	15.4	28.6	12.5	21.9	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7	6.7	0.8
Skill-Prompt*	51.7	66.7	48.4	0.0	4.3	10.0	28.9	23.7	46.2	30.6	24.4	22.1	7.5	12.5	23.9	0.2	0.8
OPSD	48.8	41.7	16.7	0.0	15.8	16.7	28.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	11.3	3.1
GRPO	91.2	62.5	96.2	61.9	65.0	47.4	75.0	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	79.8	63.3
Skill-GRPO	88.9	71.4	58.8	70.6	40.7	29.2	60.2	43.5	58.8	43.0	36.8	32.2	11.7	12.5	34.1	77.3	60.9
Skill-GRPO*	94.3	57.1	**100**	66.7	73.1	57.1	80.5	44.3	59.6	44.3	39.0	36.1	14.5	14.9	36.1	76.3	66.4
GRPO+OPSD	**100**	**82.4**	85.7	**75.0**	70.0	60.0	81.2	**44.9**	61.2	45.2	40.4	38.5	16.0	66.1	44.6	77.8	66.4
Skill-SD	88.2	50.0	96.2	52.4	65.0	57.9	73.4	44.4	60.4	44.0	39.5	40.4	15.4	64.9	44.1	75.9	64.0
RLSD	87.9	75.0	90.9	**75.0**	73.1	68.4	79.7	41.5	58.6	42.3	40.4	40.2	**16.8**	66.9	43.8	84.4	66.4
SDAR	97.1	62.5	**100**	61.9	75.0	**84.2**	84.4	44.8	58.1	44.3	38.6	36.2	15.7	66.1	43.4	85.0	68.0
StepOPSD	82.4	66.7	82.6	52.2	73.7	75.0	73.4	43.6	61.2	43.8	39.2	38.1	15.8	64.5	43.7	82.4	66.4
AgentOPSD	**100**	68.8	86.4	66.7	**84.6**	73.7	**84.4**	**44.9**	**61.9**	**48.0**	**40.9**	**41.6**	14.4	**68.1**	**46.7**	**90.4**	**69.5**
Qwen2.5-7B-Instruct																	
Vanilla	36.1	22.2	3.1	0.0	0.0	0.0	12.5	25.2	50.8	29.5	29.0	29.0	10.4	63.7	33.9	5.9	1.6
Skill-Prompt*	51.7	50.0	32.3	5.3	4.3	0.0	23.4	30.9	52.1	32.7	32.7	27.9	12.7	66.1	36.4	1.7	0.8
OPSD	50.0	60.0	22.7	21.4	17.6	9.5	32.8	8.8	8.6	17.5	2.5	4.2	0.5	1.2	6.2	4.5	2.3
GRPO	91.2	87.5	96.2	81.0	65.0	57.9	81.2	45.1	63.7	44.0	43.6	43.2	16.8	37.6	42.0	80.9	72.6
Skill-GRPO	88.5	66.7	65.2	61.1	57.7	73.1	69.5	45.2	63.7	45.7	43.1	43.3	19.6	21.4	40.3	80.4	71.9
Skill-GRPO*	**100**	83.3	96.4	83.3	75.0	78.9	88.3	44.8	63.0	45.1	43.7	43.7	20.5	71.4	47.5	87.0	81.2
GRPO+OPSD	91.4	61.5	**100**	87.5	76.5	52.2	80.4	47.3	64.5	46.9	43.8	39.3	18.0	69.4	47.0	86.8	76.5
Skill-SD	93.9	**93.8**	90.9	**100**	69.2	68.4	85.1	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5
RLSD	**100**	87.5	92.3	58.8	**80.0**	65.2	82.0	46.8	63.0	44.4	45.5	**48.9**	**21.5**	**73.0**	49.0	87.4	77.3
SDAR	94.7	75.0	**100**	86.7	68.2	78.9	85.9	46.3	63.5	**48.2**	43.8	48.4	19.6	**73.0**	49.0	89.4	**82.8**
StepOPSD	98.1	75.0	**100**	90.5	**80.0**	63.2	88.4	45.3	64.6	45.1	44.5	44.4	19.3	69.8	48.2	87.2	78.1
AgentOPSD	91.2	87.5	**100**	90.5	75.0	**84.2**	**89.1**	**47.5**	**64.9**	46.8	**45.8**	45.3	20.9	70.2	**49.2**	**90.2**	79.7
Qwen3-1.7B																	
Vanilla	25.0	22.2	3.1	0.0	21.4	4.2	12.5	29.4	46.9	37.0	23.5	19.6	6.4	10.5	24.8	46.5	4.7
Skill-Prompt*	10.3	50.0	16.1	0.0	0.0	5.0	9.4	29.4	46.5	36.2	22.9	20.8	4.3	10.1	24.3	23.0	2.3
OPSD	26.3	33.3	9.1	0.0	4.5	5.3	14.1	4.2	8.3	4.6	6.6	15.3	0.7	1.2	5.8	47.4	9.3
GRPO	71.1	41.7	36.4	40.0	31.8	31.6	46.1	40.0	58.9	43.5	35.4	30.3	12.0	65.7	40.8	67.3	38.3
Skill-GRPO	27.6	**54.5**	22.7	27.3	0.0	19.2	21.1	39.2	58.6	43.9	35.2	28.2	11.5	66.1	40.4	73.4	46.1
Skill-GRPO*	31.4	42.9	51.9	8.3	11.5	7.1	28.1	38.0	58.4	43.9	36.3	29.0	12.5	**66.9**	40.7	80.4	50.0
GRPO+OPSD	38.2	50.0	30.8	28.6	30.0	21.1	32.0	40.7	58.9	45.0	37.0	34.6	13.3	65.7	42.2	70.7	38.3
Skill-SD	52.9	37.5	69.2	42.9	**60.0**	36.8	52.3	39.1	57.5	45.4	34.8	34.1	10.7	64.1	40.8	81.8	53.9
RLSD	50.0	37.5	61.5	19.0	50.0	21.1	42.2	38.6	57.3	43.0	34.5	34.1	11.5	65.3	40.6	74.0	50.8
SDAR	**73.5**	25.0	**76.9**	33.3	40.0	36.8	53.9	39.7	58.9	45.3	35.9	**35.5**	12.6	65.3	41.9	76.8	58.6
StepOPSD	64.7	44.4	56.5	60.9	42.1	55.0	56.3	40.5	58.6	45.6	34.9	29.8	10.8	64.5	41.4	80.8	57.9
AgentOPSD	61.7	51.4	72.5	**62.9**	45.1	**56.0**	**57.3**	**41.7**	**59.9**	**45.6**	**37.2**	34.9	**14.3**	64.7	**43.2**	**82.3**	**60.9**

Baselines. We compare against three groups. **(1) Training-free:** *Vanilla* (the base model) and *Skill-Prompt*, which prepends retrieved skills at inference. **(2) Group-relative RL:** *GRPO* (Shao et al., 2024) and *Skill-GRPO*, which injects skills into the training prompt (evaluated with, *Skill-GRPO**, or without retrieved skills). **(3) Self-distillation RL:** *OPSD* (Zhao et al., 2026), *GRPO+OPSD*, *Skill-SD* (Wang et al., 2026a), *RLSD* (Yang et al., 2026a), *SDAR* (Lu et al., 2026b) and *StepOPSD* (Zhang et al., 2026), all of which use the teacher–student gap but inject it as a gate, magnitude, or auxiliary loss. All methods share the same backbone, data, and budget. Full algorithm details are in Appendix B.

3.2 Main Results

The gain comes from credit construction, not privileged access. Under our unified setup, AgentOPSD and the privileged baselines use the same retrieved skills; they differ primarily in how the skill-induced teacher–student discrepancy enters learning. AgentOPSD outperforms GRPO+OPSD, Skill-SD, and RLSD on all twelve aggregate comparisons across the three model scales, and exceeds SDAR on ten of twelve. This controlled-information comparison isolates the benefit of AgentOPSD: a local teacher-student gap is not yet a reliable credit signal. Accumulating that gap into a belief state and assigning credit according to belief revision more effectively identifies the turns that change the predicted outcome.

The advantage grows with the interaction horizon. AgentOPSD is designed for the regime where uniform credit is most harmful, so we also ask how performance degrades as tasks require more turns. Figure 1(b) regresses per-sub-task success on the measured mean number of turns of successful episodes on ALFWorld (Qwen2.5-7B), reporting the success points lost per additional turn. The uniform-credit methods degrade fastest (-3.59 for RLSD and -2.91 for GRPO points per turn),

Component	Ablation	ALFWorld
AgentOPSD (full)	turn-level, bounded, λ=0.5	89.1
Turn-level granularity	per-token accumulation	85.9
Recursive state revision (8)	raw local gap e_k in place of ΔB_k	82.8
Signed direction (10)	magnitude $\lvert\Delta B_k\rvert$ only (drop outcome sign)	80.5
State prior B_0 anchor	drop empirical-rate initialization	78.9

Table 2: **Component ablation of AgentOPSD** on ALFWorld with Qwen2.5-7B (success rate, %). Each row removes or replaces a single mechanism while holding all other settings fixed. The signed direction and the state prior anchor have the largest impact on performance, while the recursive state revision and turn-level granularity provide smaller but consistent improvements.

whereas AgentOPSD is the flattest at -0.54. This is consistent with the motivation for turn-level credit: the longer the trajectory, the more decisions a single broadcast advantage has to cover, and the more a history-dependent revision helps.

3.3 Mechanism Ablation

Table 2 evaluates each design choice on ALFWorld with Qwen2.5-7B by removing or replacing one component at a time. The full method achieves a success rate of 89.1.

Granularity and recursion. Replacing turn-level belief tracking with per-token accumulation reduces the success rate to 85.9: environment feedback is associated with a complete action rather than an individual token, so token-level accumulation fragments a single decision and weakens the alignment between the gap and outcomes. Replacing the recursive revision ΔB_k with the raw local gap e_k further reduces performance to 82.8. A raw e_k scores each turn in isolation, whereas ΔB_k measures how that gap revises the belief state accumulated over the preceding history—so the same local gap that is decisive while the outcome is open becomes redundant once the accumulated state already points to an outcome. This controlled comparison isolates the value of the recursion and directly confirms our central principle that a local gap is not sequential credit.

Outcome-aligned signed direction. Keeping only the magnitude $|\Delta B_k|$ and dropping the sign (Eq. 10), i.e. standardizing $|\Delta B_k|$ instead of the signed q_k, lowers performance to 80.5. The magnitude identifies where the belief state changes, but cannot determine whether that change agrees with the verifier outcome. For a successful trajectory, an upward belief revision is consistent with the outcome, whereas for a failed trajectory the same revision is inconsistent. The signed direction makes this distinction explicit, allowing outcome-consistent revisions to receive more credit and contradictory revisions to receive less.

State-prior anchoring. Removing the empirical prior $B_0 = \mathrm{clip}(\bar{R}, \epsilon_0, 1-\epsilon_0)$ reduces the success rate to 78.9. The group success rate $\bar{R}$ provides a verifier-grounded estimate of task difficulty before the trajectory-specific gap is accumulated. Moreover, B_0 determines the initial log-odds and thus the operating region of the $B(1-B)$ gate. Without this anchor, trajectories begin from an arbitrary uncertainty level, which can mis-scale early belief revisions and distort which early turns appear pivotal. The ablations therefore separate three roles: belief revision localizes credit, the signed direction aligns it with the final outcome, and prior anchoring stabilizes its reference point.

3.4 Hyperparameter Sensitivity

Whereas the mechanism ablation asks whether each component is necessary, we now examine how sensitive AgentOPSD is to its continuous hyperparameters. We sweep one knob at a time while holding the others at the full-AgentOPSD setting (λ=0.5, γ=0.95, ϵ_{high}=0.24; Appendix F) across three configurations (Figure 3): long-horizon ALFWorld with Qwen2.5-7B (89.1) and Qwen2.5-3B (84.4), and short-horizon Search-QA with Qwen2.5-3B (46.7).

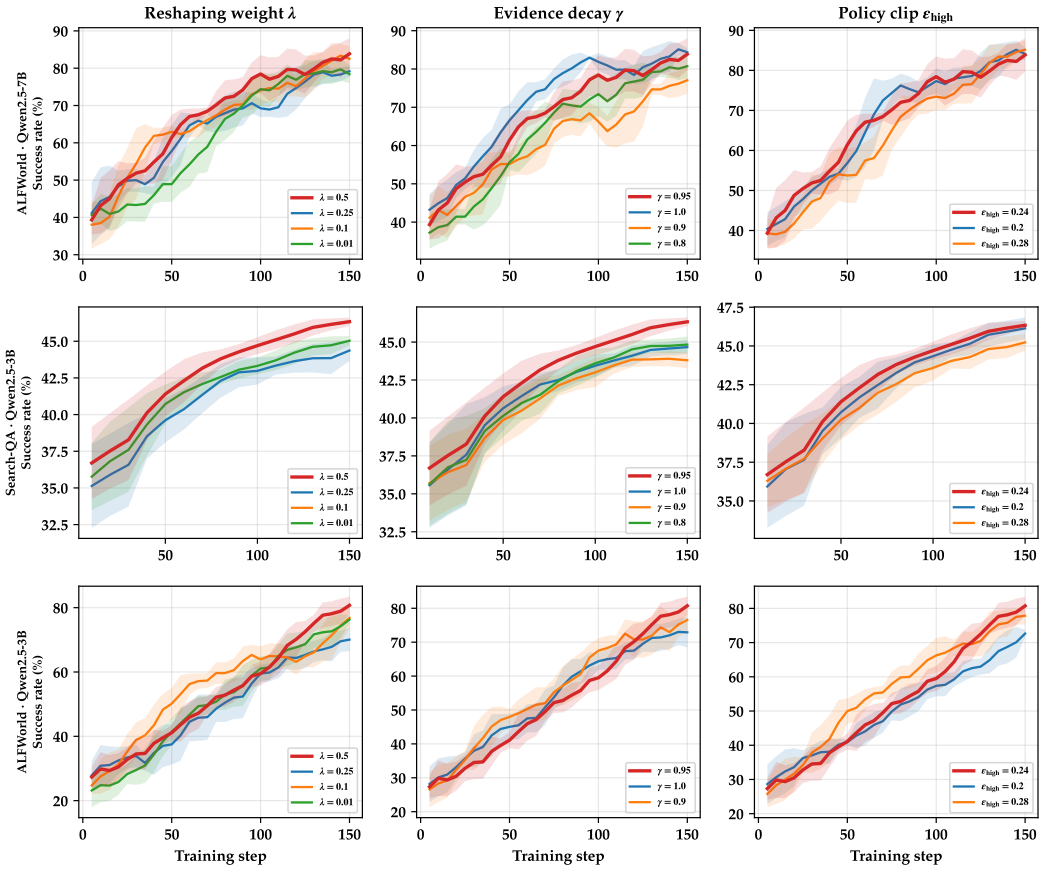

Figure 3: **Hyperparameter sensitivity of AgentOPSD.** Rows: ALFWorld (Qwen2.5-7B), Search-QA (Qwen2.5-3B), and ALFWorld (Qwen2.5-3B). Columns sweep one knob (λ, γ, ϵ_{high}) with the others held at our setting. Curves are rolling means; shaded bands show the ± 1 standard deviation.

Reshaping weight λ. Sweeping $\lambda \in \{0.5, 0.25, 0.1, 0.01\}$ interpolates between pure GRPO (λ=0) and full belief reshaping. This is the knob with the clearest effect: λ=0.5 is best and any smaller value reduces performance (89.1 at λ=0.5 vs. 84.4/85.9/83.6; Search 46.7 vs. 45.1/40.2/45.4), consistent with a smaller λ down-weighting the bounded multiplier and discarding turn-level credit. We use λ=0.5 throughout.

Evidence decay γ. Turn-level evidence is accumulated with a geometric decay $c_k = \gamma\, c_{k-1} + e_k$ (equivalently $\ell_k = \text{logit}(B_0) + \sum_{j \le k} \gamma^{k-j} e_j$); sweeping $\gamma \in \{1.0, 0.95, 0.9, 0.8\}$ moves the result within a few points (87.5/82.0/85.2; Search 45.1/44.5/45.5) without a monotone trend, so the recursion is not particularly sensitive to how fast old evidence is discounted. We use the mild setting γ=0.95 for all main results.

Policy clipping ϵ_{high}. Fixing ϵ_{low}=0.2 and varying $\epsilon_{\text{high}} \in \{0.2, 0.24, 0.28\}$ (clip-higher (Yu et al., 2025)), AgentOPSD is largely unaffected (88.3 at both 0.2 and 0.28; Search 46.9 and 45.7), indicating that the reshaped objective inherits the trust-region robustness of GRPO. Overall, only λ produces a systematic effect, and the spread across all knobs shrinks sharply on the four-turn Search-QA task—the settings that matter on long-horizon ALFWorld are largely inert when little history accumulates, which is again consistent with the method acting where long-horizon credit assignment is needed.

4 Related Work

4.1 Agentic Post-Training with Verifiable Rewards

Reinforcement learning with verifiable rewards has advanced from single-turn reasoning (Shao et al., 2024; Guo et al., 2025; Yu et al., 2025; Lu et al., 2026a) to long-horizon agents in embodied text worlds, web shopping, and retrieval-augmented question answering (Shridhar et al., 2020; Yao et al., 2022; Jin et al., 2025; Lu et al., 2025). In these interactive settings, sparse terminal rewards make turn-level credit assignment particularly challenging. Standard GRPO broadcasts a trajectory-level advantage uniformly across turns. GiGPO (Feng et al., 2025) improves reward-side credit assignment by combining episode-level advantages with step-level advantages estimated from repeated anchor states across trajectories. In contrast, AgentOPSD derives turn-level evidence from privileged teacher–student likelihood gaps and assigns credit through recursive belief revision. GiGPO and AgentOPSD therefore operate on complementary signal sources—environment rewards and self-distillation evidence, respectively.

4.2 On-Policy (Self-)Distillation

On-policy distillation trains a policy on its own rollouts under a teacher (Agarwal et al., 2024; Gu et al., 2026; Wen et al., 2023). Its recent self-distillation variants remove the need for a separate teacher while the teacher branch is conditioned on privileged information available only during training (Zhao et al., 2026; He et al., 2026; Lu et al., 2026c). Recent studies incorporate the resulting teacher–student log-probability gap into RLVR by using it to scale or reshape the advantage (Yang et al., 2026b; Wu et al., 2026), as a detached auxiliary objective (Lu et al., 2026b; Wang et al., 2026a), or as reweighted, scheduled, or reward-densifying local supervision (Xu et al., 2026; Wang et al., 2026b; Ye et al., 2026a; He et al., 2026). Existing methods predominantly treat the distillation gap as a local token-level or step-level signal. Token-level signals are not naturally aligned with action turns, and the contribution of a turn depends on the evidence accumulated through preceding interactions. StepOPSD (Zhang et al., 2026) aggregates the teacher–student signal over action-centered step spans but still scores each span by its local log-ratio. In contrast, AgentOPSD first aggregates token-level gaps within each turn and then recursively accumulates the resulting evidence into a support state.

4.3 Long-Horizon Credit Assignment

Assigning credit across a long horizon is a classical problem. PPO learns a value function and, via GAE, derives a per-step temporal-difference signals (Schulman et al., 2017; 2016). When rewards are sparse and delayed, return-decomposition methods such as RUDDER redistribute a terminal reward to the steps responsible for it (Arjona-Medina et al., 2019), while process reward models and Monte-Carlo credit methods such as VinePPO estimate intermediate value by additional rollouts or a learned scorer (Cui et al., 2025; Kazemnejad et al., 2024). These approaches recover per-step structure but reintroduce the cost GRPO removed: a trained critic, a reward model, or many extra rollouts. AgentOPSD restores a per-turn value signal in the critic-free group-relative setting, at the cost of a single teacher forward pass. The belief state plays the role of GAE's value baseline and its per-turn revision the role of the TD signal, but without a learned value network cost.

5 Conclusion

We studied credit assignment for long-horizon language agents. Our key insight is that turn-level credit should depend not only on a local signal, but also on how that signal revises the accumulated belief in eventual trajectory success. Based on this insight, we proposed AgentOPSD, which aggregates token-level self-distillation gaps at environment-aligned turn boundaries and recursively updates a trajectory-success belief in log-odds space. These belief revisions redistribute the trajectory-level advantage across turns without requiring additional rollouts or a learned critic. Experiments across three interactive environments and three model scales show that AgentOPSD consistently outperforms GRPO and strong self-distillation baselines. Ablations further confirm the importance of both turn-level signal aggregation and history-dependent belief revision. Overall, our results suggest that recursive belief updating provides a simple and effective approach to restoring fine-grained temporal credit in critic-free agentic reinforcement learning.

AI USE STATEMENT

Generative AI tools were used only for tasks in the "recommended disclosure" category of the ICLR 2027 AI Policy for Authors: editing the manuscript to improve readability, and assisting with LaTeX formatting and template compliance. We did not use generative AI tools for any task in the "required disclosure" category — in particular, the research idea, the method design, the mathematical claims and their proofs, the experimental design, the implementation, and the interpretation of results are entirely the authors' own work. We have reviewed all AI-assisted edits and take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work studies credit assignment for reinforcement learning on language agents and is evaluated exclusively on publicly available benchmark environments (ALFWorld, WebShop, and Search-QA) using publicly released open-weight models. It involves no human subjects, no personally identifiable information, and no newly collected or released dataset. We are not aware of ethical concerns or foreseeable harmful applications that are specific to the proposed method beyond those already inherent to training capable general-purpose language agents.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes, 2024. URL https://arxiv.org/abs/2306.13649.

Jose A. Arjona-Medina, Michael Gillhofer, Michael Widrich, Thomas Unterthiner, Johannes Brandstetter, and Sepp Hochreiter. Rudder: Return decomposition for delayed rewards, 2019. URL https://arxiv.org/abs/1806.07857.

Karl Johan Åström. Optimal control of markov processes with incomplete state information i. *Journal of Mathematical Analysis and Applications*, 10:174–205, 1965. doi: 10.1016/0022-247X(65)90154-X.

Yuxin Chen, Zhengzhou Cai, Xiangtian Ji, Weixiang Zhao, An Zhang, Xiang Wang, and Tat-Seng Chua. Understanding multilingualism in mixture-of-experts llms: Routing mechanism, expert specialization, and layerwise steering. *arXiv preprint arXiv:2601.14050*, 2026a.

Yuxin Chen, Yu Wang, Yi Zhang, Ziang Ye, Zhengzhou Cai, Yaorui Shi, Qi Gu, Hui Su, Xunliang Cai, Xiang Wang, et al. Learning to self-verify makes language models better reasoners. *arXiv preprint arXiv:2602.07594*, 2026b.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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

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. *arXiv preprint arXiv:2507.19849*, 2025.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: On-policy distillation of large language models, 2026. URL https://arxiv.org/abs/2306.08543.

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

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

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Xiangtian Ji, Yuxin Chen, Zhengzhou Cai, Xiang Wang, An Zhang, and Tat-Seng Chua. Tiny brains, giant impact: Uncovering the keystone neurons of llm with just a few prompts. *arXiv preprint arXiv:2605.24846*, 2026.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998. doi: 10.1016/S0004-3702(98)00023-X.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms, 2024. URL https://arxiv.org/abs/2410.01679.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

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, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Pengxiang Zhao, Guangyi Liu, et al. Ui-r1: Enhancing efficient action prediction of gui agents by reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17608–17616, 2026a.

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*, 2026b.

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*, 2026c.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2016. URL https://arxiv.org/abs/1506.02438.

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

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

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023.

Zhengliang Shi, Shen Gao, Lingyong Yan, Yue Feng, Xiuyi Chen, Zhumin Chen, Dawei Yin, Suzan Verberne, and Zhaochun Ren. Tool learning in the wild: Empowering language models as automatic tool agents. In *Proceedings of the ACM on Web Conference 2025*, pp. 2222–2237, 2025.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.

Core Team, Bangjun Xiao, Bingquan Xia, Bo Yang, Bofei Gao, Bowen Shen, Chen Zhang, Chenhong He, Chiheng Lou, Fuli Luo, Gang Wang, Gang Xie, Hailin Zhang, Hanglong Lv, Hanyu Li, Heyu Chen, Hongshen Xu, Houbin Zhang, Huaqiu Liu, Jiangshan Duo, Jianyu Wei, Jiebao Xiao, Jinhao Dong, Jun Shi, Junhao Hu, Kainan Bao, Kang Zhou, Lei Li, Liang Zhao, Linghao Zhang, Peidian Li, Qianli Chen, Shaohui Liu, Shihua Yu, Shijie Cao, Shimao Chen, Shouqiu Yu, Shuo Liu, Tianling Zhou, Weijiang Su, Weikun Wang, Wenhan Ma, Xiangwei Deng, Bohan Mao, Bowen Ye, Can Cai, Chenghua Wang, Chengxuan Zhu, Chong Ma, Chun Chen, Chunan Li, Dawei Zhu, Deshan Xiao, Dong Zhang, Duo Zhang, Fangyue Liu, Feiyu Yang, Fengyuan Shi, Guoan Wang, Hao Tian, Hao Wu, Heng Qu, Hongfei Yi, Hongxu An, Hongyi Guan, Xing Zhang, Yifan Song, Yihan Yan, Yihao Zhao, Yingchun Lai, Yizhao Gao, Yu Cheng, Yuanyuan Tian, Yudong Wang, Zhen Tang, Zhengju Tang, Zhengtao Wen, Zhichao Song, Zhixian Zheng, Zihan Jiang, Jian Wen, Jiarui Sun, Jiawei Li, Jinlong Xue, Jun Xia, Kai Fang, Menghang Zhu, Nuo Chen, Qian Tu, Qihao Zhang, Qiying Wang, Rang Li, Rui Ma, Shaolei Zhang, Shengfan Wang, Shicheng Li, Shuhao Gu, Shuhuai Ren, Sirui Deng, Tao Guo, Tianyang Lu, Weiji Zhuang, Weikang Zhang, Weimin Xiong, Wenshan Huang, Wenyu Yang, Xin Zhang, Xing Yong, Xu Wang, Xueyang Xie, Yilin Jiang, Yixin Yang, Yongzhe He, Yu Tu, Yuanliang Dong, Yuchen Liu, Yue Ma, Yue Yu, Yuxing Xiang, Zhaojun Huang, Zhenru Lin, Zhipeng Xu, Zhiyang Chen, Zhonghua Deng, Zihan Zhang, and Zihao Yue. Mimo-v2-flash technical report, 2026a. URL https://arxiv.org/abs/2601.02780.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chen Gao, Chen Zhang, Chengcheng Han, et al. Longcat-flash-thinking-2601 technical report. *arXiv preprint arXiv:2601.16725*, 2026b.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16 (2):117–186, 1945. doi: 10.1214/aoms/1177731118.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation, 2023. URL https://arxiv.org/abs/2307.15190.

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.

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.

Hongshen Xu, Zichen Zhu, Lei Pan, Zihan Wang, Su Zhu, Da Ma, Ruisheng Cao, Lu Chen, and Kai Yu. Reducing tool hallucination via reliability alignment. *arXiv preprint arXiv:2412.04141*, 2024.

Hongshen Xu, Zihan Wang, Zichen Zhu, Lei Pan, Xingyu Chen, Shuai Fan, Lu Chen, and Kai Yu. Alignment for efficient tool calling of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17787–17803, 2025.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. Tip: Token importance in on-policy distillation, 2026. URL https://arxiv.org/abs/2604.14084.

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

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr, 2026a. URL https://arxiv.org/abs/2604.03128.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026b.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation, 2026c. URL https://arxiv.org/abs/2602.12125.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2369–2380, 2018.

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.

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

Ziang Ye, Wentao Shi, Yuxin Liu, Yu Wang, Zhengzhou Cai, Yaorui Shi, Qi Gu, Xunliang Cai, and Fuli Feng. Look before you leap: Autonomous exploration for llm agents. *arXiv preprint arXiv:2605.16143*, 2026b.

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

Yanfei Zhang, Xu Lin, and Chenglin Wu. Stepopsd: Step-aware online preference distillation for agent reinforcement learning, 2026. URL https://arxiv.org/abs/2605.27140.

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, 2026. URL https://arxiv.org/abs/2601.18734.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, et al. Memento: Fine-tuning llm agents without fine-tuning llms. *arXiv preprint arXiv:2508.16153*, 2025.

TABLE OF CONTENTS

1 Introduction **1**

2 Methodology **3**

2.1 Problem Setup 3

2.2 From Outcome Contribution to Bayesian Turn Evidence 3

2.3 Recursive Belief update 4

2.4 Bounded Advantage Reshaping 5

3 Experiments **5**

3.1 Experimental Setup 5

3.2 Main Results 6

3.3 Mechanism Ablation 7

3.4 Hyperparameter Sensitivity 7

4 Related Work **9**

4.1 Agentic Post-Training with Verifiable Rewards 9

4.2 On-Policy (Self-)Distillation 9

4.3 Long-Horizon Credit Assignment 9

5 Conclusion **9**

A Theoretical Analysis **16**

A.1 From the Bayes factor to the self-teacher contrast 16

A.2 Properties of the reshaping 16

B Algorithm **18**

C Datasets **19**

D Baseline Details **19**

E Evaluation Metrics **20**

F Hyperparameters **20**

G Training Dynamics **20**

H Prompt **21**

A Theoretical Analysis

A.1 From the Bayes factor to the self-teacher contrast

AgentOPSD approximates the ideal per-turn Bayes factor

$$\mathcal{B}_k \;=\; \log \frac{p(a_k \mid s_k, C)}{p(a_k \mid s_k, \neg C)} \;=\; \operatorname{logit} p(C \mid s_k, a_k) - \operatorname{logit} p(C \mid s_k) \tag{13}$$

by the self-teacher contrast

$$e_k \;=\; \log \frac{\pi_\theta(a_k \mid s_k, c^+)}{\pi_\theta(a_k \mid s_k)}, \tag{14}$$

where C denotes eventual success and $\rho_k = p(C \mid s_k)$. We use two assumptions: **(A1)** the skill-conditioned branch is success-conditional, $\pi_\theta(a_k \mid s_k, c^+) \approx p(a_k \mid s_k, C)$; **(A2)** when success is rare (ρ_k small) the marginal is failure-dominated, $\pi_\theta(a_k \mid s_k) \approx p(a_k \mid s_k, \neg C)$.

The marginal action distribution is the success/failure mixture

$$\pi_\theta(a_k \mid s_k) = \rho_k\, p(a_k \mid s_k, C) + (1 - \rho_k)\, p(a_k \mid s_k, \neg C). \tag{15}$$

Substituting (15) into (14) under (A1),

$$e_k \;\approx\; \mathcal{B}_k - \log\!\big(1 - \rho_k + \rho_k\, e^{\mathcal{B}_k}\big) \;\xrightarrow[\rho_k \to 0]{}\; \mathcal{B}_k, \tag{16}$$

so (A2) is the $\rho_k \to 0$ limit in which the contrast recovers the Bayes factor. Under (A1) alone, e_k is the pointwise mutual information

$$e_k \;\approx\; \log \frac{p(a_k \mid s_k, C)}{p(a_k \mid s_k)} \;=\; \log \frac{p(C \mid s_k, a_k)}{p(C \mid s_k)}, \tag{17}$$

positive iff a_k raises the posterior success probability. The correction in (16) is monotone in $\mathcal{B}_k$, hence $\operatorname{sign}(e_k) = \operatorname{sign}(\mathcal{B}_k)$ and e_k preserves the ranking of turns by evidential strength; AgentOPSD uses e_k only through this sign and ranking.

A.2 Properties of the reshaping

Let $A^{(i)}$ be the group-relative advantage, $\Delta B_k = B_k - B_{k-1}$ the per-turn belief revision, z_k its within-trajectory standardization, $m_k = \operatorname{clip}\big(1 + b\,\operatorname{sign}(A^{(i)}) z_k,\, 1-b,\, 1+b\big)$ with $b \in (0,1)$, and $\tilde{A}_k = A^{(i)}\big((1-\lambda) + \lambda m_k\big)$ with $\lambda \in [0,1]$.

Proposition 1 (Boundedness). $\big|\tilde{A}_k - A^{(i)}\big| \le \lambda b\,|A^{(i)}|$, *hence* $(1-\lambda b)|A^{(i)}| \le |\tilde{A}_k| \le (1+\lambda b)|A^{(i)}|$.

Proof. $m_k \in [1-b, 1+b]$ gives $|m_k - 1| \le b$, and $\tilde{A}_k - A^{(i)} = A^{(i)}\lambda(m_k - 1)$. □

Proposition 2 (Sign Preservation). $\operatorname{sign}(\tilde{A}_k) = \operatorname{sign}(A^{(i)})$ *for every turn* k.

Proof. $(1-\lambda) + \lambda m_k \ge 1 - \lambda b > 0$ since $\lambda \le 1$, $b < 1$; a strictly positive factor preserves sign. □

Proposition 3 (Recovery of GRPO). *At* $\lambda = 0$, $\tilde{A}_k = A^{(i)}$ *for every token and the AgentOPSD gradient equals the GRPO gradient.*

Proof. $\lambda = 0$ gives $(1-\lambda) + \lambda m_k = 1$, so $\tilde{A}_k = A^{(i)}$ identically, independent of the belief signal. □

Proposition 4 (First-Order Decomposition of the Belief Revision). *For* $c_k = \gamma c_{k-1} + e_k$, $\ell_k = \operatorname{logit}(B_0) + c_k$, *and* $\Delta\ell_k = e_k - (1-\gamma)c_{k-1}$,

$$\Delta B_k = B_{k-1}(1 - B_{k-1})\,\Delta\ell_k + O\big((\Delta\ell_k)^2\big). \tag{18}$$

Proof. $B_k = \sigma(\ell_k)$ with $\sigma' = \sigma(1-\sigma)$; a first-order expansion around ℓ_{k-1} gives $B_k = B_{k-1} + B_{k-1}(1-B_{k-1})\Delta\ell_k + O((\Delta\ell_k)^2)$. The gate $B(1-B)$ is maximal at $B = \frac{1}{2}$ and vanishes as $B \to \{0,1\}$. □

Proposition 5 (Exact Budget of the Idealized Recursion)**.** *The idealized increments telescope to the endpoint change,* $\sum_k \Delta B_k = B_K - B_0$.

Proof. $\sum_{k=1}^{K}(B_k - B_{k-1}) = B_K - B_0$. □

Proposition 6 (Non-Identifiability of Per-Turn Contribution)**.** *There exist two trajectories with identical outcome reward—hence identical broadcast advantage—whose per-turn contributions differ; per-turn credit is not identifiable from the trajectory return alone.*

Proof (by construction). Take τ_1, τ_2 in one group with $R(\tau_1) = R(\tau_2)$, so $A(\tau_1) = A(\tau_2)$ and GRPO assigns the same scalar to every turn. Let τ_1 succeed through a single decisive turn (ΔB concentrated) and τ_2 through evenly spread progress. The returns coincide but the per-turn contributions differ, so an additional per-turn signal is required to recover them. □

Proposition 7 (B_0 as the Group Success-Rate Estimate)**.** *For a task with success probability* θ_x *and a group of* G *trajectories yielding* S *successes under a binary reward, the maximum-likelihood estimate of* θ_x *is the group success fraction* $\bar{R} = S/G$*, which is the standard GRPO group mean; AgentOPSD sets* $B_0 = \mathrm{clip}(\bar{R}, \epsilon_0, 1 - \epsilon_0)$.

Proof. Under a Binomial(G, θ_x) likelihood the MLE is S/G. The clip (ϵ_0=10^{-4}) only keeps $\mathrm{logit}(B_0)$ finite for all-correct or all-wrong groups, which have zero group-relative advantage and hence do not contribute to the update. □

B ALGORITHM

We give pseudocode for one AgentOPSD training iteration at turn-level granularity in Algorithm 1. The only addition over GRPO is a single teacher forward pass per turn and the per-turn belief reshaping block; everything else is the standard group-relative update.

Algorithm 1 AgentOPSD: Recursive State Updates for Turn-Level Credit

Require: policy π_θ, verifier R, group size G, skill retriever; mixing λ, bound b, evidence decay γ

for each training iteration **do**
  sample a batch of tasks $\{x\}$
  for each task x with retrieved skill c^+ **do**
    sample G trajectories $\{y^{(1)}, \dots, y^{(G)}\} \sim \pi_\theta(\cdot \mid x)$; trajectory i has K_i turns ▷ on-policy rollout
    for $i = 1, \dots, G$ **do**
      obtain reward $R^{(i)} = R(x, y^{(i)}) \in \{0, 1\}$ from the verifier
    end for
    $A^{(i)}_{\text{seq}} \leftarrow (R^{(i)} - \mu_G)/\sigma_G$ ▷ group-relative advantage
    for $i = 1, \dots, G$ **do**
      $B_0 \leftarrow \text{clip}(\bar{R}, \epsilon_B, 1 - \epsilon_B)$; $\ell_0 \leftarrow \text{logit}(B_0)$ ▷ standard GRPO group success rate $\bar{R}{=}S/G$ (Prop. 7)
      $c_0 \leftarrow 0$
      for $k = 1, \dots, K_i$ **do** ▷ per-turn belief state
        $e_k \leftarrow \sum_t \text{sg}[\log \pi_\theta(y_{k,t} \mid s^+_k) - \log \pi_\theta(y_{k,t} \mid s_k)]$ ▷ one extra teacher forward
        $c_k \leftarrow \gamma\, c_{k-1} + e_k$; $\ell_k \leftarrow \ell_0 + c_k$; $B_k \leftarrow \sigma(\ell_k)$; $\Delta B_k \leftarrow B_k - B_{k-1}$
      end for
      $q_k \leftarrow \text{sign}(A^{(i)}_{\text{seq}})\, \Delta B_k$ for all k ▷ outcome-aligned credit
      $z_k \leftarrow (q_k - \text{mean}(q))/(\text{std}(q) + \epsilon)$ ▷ within-trajectory standardization
      $w_k \leftarrow \text{clip}(1 + b\, z_k,\ 1 - b,\ 1 + b)$ ▷ bounded multiplier
      $\widetilde{A}^{(i)}_k \leftarrow A^{(i)}_{\text{seq}}\big[(1 - \lambda) + \lambda w_k\big]$; each token inherits $\widetilde{A}$ of its turn
    end for
  end for
  update θ by maximizing the clipped GRPO objective $\mathcal{L}_{\text{AgentOPSD}}(\theta)$ with $\{\widetilde{A}\}$ ▷ policy update
end for

Token-level variant. Replace the per-turn recursion with the same recursion over the flattened token sequence under a response mask: accumulate δ_t directly, standardize ΔB_t over the trajectory's tokens, and assign $\widetilde{A}_t$ per token. This is the granularity ablation reported in Table 2.

Cost. The overhead over GRPO is one teacher forward pass per trajectory; the belief reshaping block is elementwise and adds no rollouts and no learned parameters.

C DATASETS

Our experiments span three multi-turn agentic environments covering embodied household reasoning, web navigation, and search-augmented question answering.

ALFWorld (Shridhar et al., 2020) is a text-based embodied environment with six task categories—Pick and Place, Look at Object in Light, Pick Clean then Place, Pick Heat then Place, Pick Cool then Place, and Pick Two and Place. Given a language goal and textual observations, the agent selects admissible actions until the goal is satisfied.

WebShop (Yao et al., 2022) is an interactive online-shopping environment. For each user request the agent searches the product catalog, inspects candidate items, selects the required attributes, and attempts a purchase satisfying the specified constraints. We evaluate on the 128 fixed validation tasks of Feng et al. (2025).

Search-QA follows the Search-R1 setup (Jin et al., 2025) over seven datasets: single-hop NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023) and multi-hop HotpotQA (Yang et al., 2018), 2Wiki (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), Bamboogle (Press et al., 2023), with NQ and HotpotQA in-domain and the rest held out. The agent issues search queries, inspects retrieved documents (retrieval via E5 (Wang et al., 2022)), and synthesizes the collected evidence before returning its final answer (Ye et al., 2026b; Chen et al., 2026b;a; Ji et al., 2026; Zhou et al., 2025; Xu et al., 2024; 2025).

D BASELINE DETAILS

We compare against three groups of baselines. Unless a method is marked with $*$, evaluation uses only the standard task prompt and the interaction history returned by the environment; $*$ indicates that a retrieved skill is additionally supplied during validation and testing.

Vanilla. The instruction-tuned backbone evaluated without any post-training.

Skill-Prompt*. The same frozen parameters as Vanilla, but a retrieved task-relevant skill is prepended to the context at validation/test time, measuring the inference-time value of skills without any parameter update.

GRPO (Shao et al., 2024). A critic-free group-relative RL algorithm: it samples a group of trajectories per task, normalizes their terminal rewards into relative advantages, and optimizes a clipped surrogate objective; every token inherits its trajectory's sequence-level advantage.

Skill-GRPO / Skill-GRPO*. GRPO with a retrieved skill injected into the training prompt. The skill is removed at inference for Skill-GRPO (testing whether the guidance has been internalized), and kept at inference for Skill-GRPO*.

OPSD (Zhao et al., 2026). On-policy self-distillation: a teacher branch conditioned on training-only privileged context re-scores the student's sampled tokens and produces dense token-level targets through distribution matching; the teacher outputs are detached and the privileged context is not used at inference.

GRPO+OPSD. Jointly optimizes the trajectory-level GRPO loss and the token-level OPSD objective, a straightforward combination of outcome-based RL and generic self-distillation.

Skill-SD (Wang et al., 2026a). Supplies the retrieved skill only to the teacher branch and trains the student to absorb the skill-conditioned guidance via an importance-weighted distillation loss, without requiring skills at evaluation.

RLSD (Yang et al., 2026a). Converts the teacher–student log-probability gap into a bounded coefficient that scales the magnitude of each token's GRPO update; the sign of the update remains determined by the outcome-derived advantage.

SDAR (Lu et al., 2026b). Adds a separately gated auxiliary self-distillation loss on top of GRPO, leaving the original GRPO advantage unchanged and using a bounded gate to modulate each teacher signal.

StepOPSD (Zhang et al., 2026). Applies the teacher–student self-distillation signal at the turn (step) level rather than per token, but uses each step's local signal in isolation.

All post-training baselines share the same backbone models, environment interfaces, data, and training budget as AgentOPSD; they differ primarily in their optimization objective and in whether skills are available during training or evaluation.

E Evaluation Metrics

ALFWorld. We report the overall success rate over the evaluation tasks (the fraction of episodes that reach the specified goal), which is the sample-weighted average of the six per-category success rates.

Search-QA. We report the overall accuracy over all evaluation questions aggregated across the seven datasets.

WebShop. We report a normalized completion Score (averaged over partial-constraint satisfaction and scaled by 100) and an exact-completion success rate Succ. (the percentage of episodes that satisfy all specified requirements).

F Hyperparameters

Table 3 summarizes the hyperparameters used by AgentOPSD across all our experiments. We deliberately use a *single* setting for every environment and model scale rather than tuning per task: AgentOPSD runs at turn-level granularity with reshaping weight λ=0.5, multiplier band b=0.2, gap accumulation with decay γ=0.95, policy clipping ϵ_{low}=0.2 / ϵ_{high}=0.24, and the empirical group success rate $\bar{R} = S/G$ (clipped) as the state prior B_0. The sensitivity study in §3.4 sweeps λ, γ and ϵ_{high} around this setting and finds no swept value that improves on it by a meaningful margin, which is why one shared configuration is used throughout rather than per-environment tuning.

Table 3: **Hyperparameters**. η: learning rate; G: group size; $\epsilon_{\text{low}}/\epsilon_{\text{high}}$: PPO clip range; α_{KL}: KL penalty coefficient toward the reference policy; SRS: skill retrieval strategy (KM = keyword matching). AgentOPSD-specific reshaping knobs: λ (`mult_lambda`, reshaping weight), b (`mult_band`, multiplier band), and γ (`gap_decay_gamma`, gap-accumulation decay). A single shared setting is used across all environments and model scales.

Method	η	G	$\epsilon_{\text{low}}/\epsilon_{\text{high}}$	α_{KL}	λ	b	γ	SRS
AgentOPSD	10^{-6}	8	0.2/0.24	0.01	0.5	0.2	0.95	KM

We use a single shared optimization recipe across environments (learning rate, group size, PPO clip range, and KL coefficient in Table 3; dual-clip constant c=3.0, gradient clipping 1.0, entropy coefficient 0.001, one PPO epoch per update, and FSDP on a single node). The remaining settings are environment-specific and summarized in Table 4.

G Training Dynamics

We present the full training dynamics of AgentOPSD across all model scales and environments in Figures 4–5, tracking the teacher–student gap and the reward throughout training.

Table 4: **Per-environment training configuration.** Optimization settings shared across all environments are listed in Table 3; the environment-specific settings are below.

	ALFWorld	**WebShop**	**Search-QA**
Training steps	150	150	150
Train batch size	16	16	128
Rollout group size G	8	8	8
Max prompt length	2048	4096	4096
Max response length	512	512	512
Max interaction turns	50	15	4
Rollout temperature (train / val)	1.0 / 0.4	1.0 / 0.4	1.0 / 0.4
GPUs (tensor-parallel size)	8 (2)	2 (2)	4 (1)

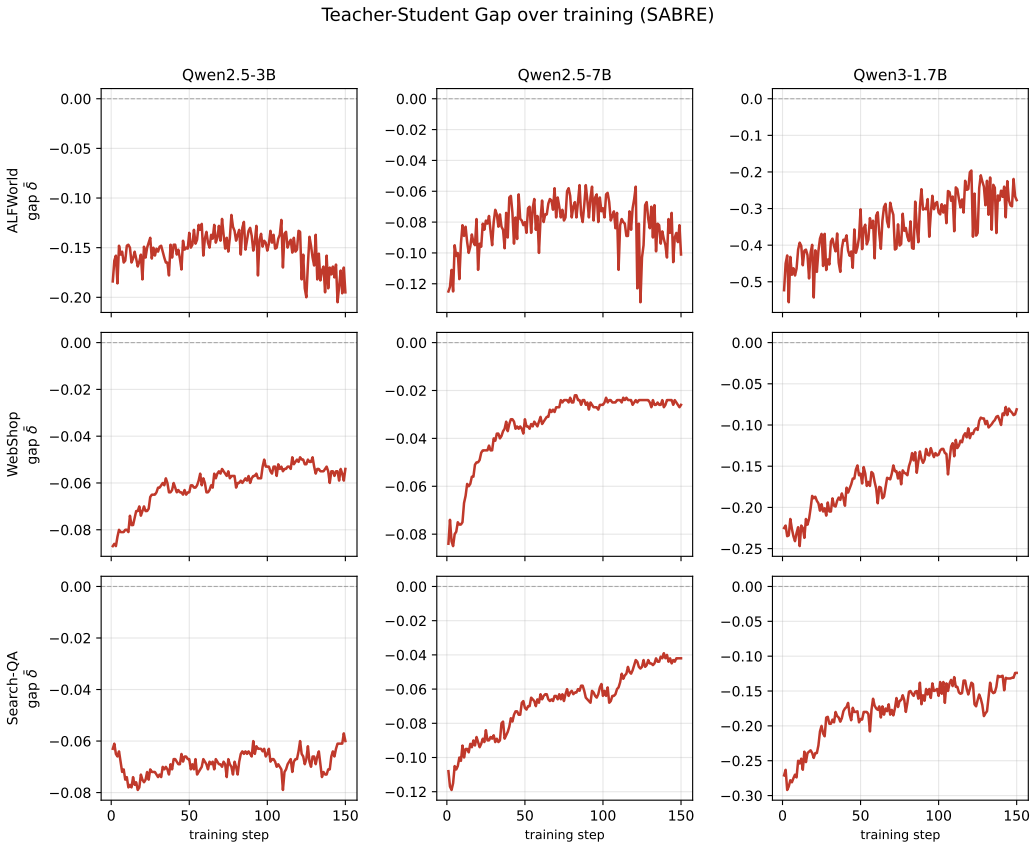

Figure 4: **Teacher–Student Gap** $\bar{\delta}$ when training Qwen2.5-3B, Qwen2.5-7B and Qwen3-1.7B on ALFWorld, WebShop and Search-QA.

H PROMPT

Figures 6–8 present the full prompt templates used by AgentOPSD for the three evaluation environments, where {`skill_context`} is populated with the retrieved skill during training and left empty at inference time.

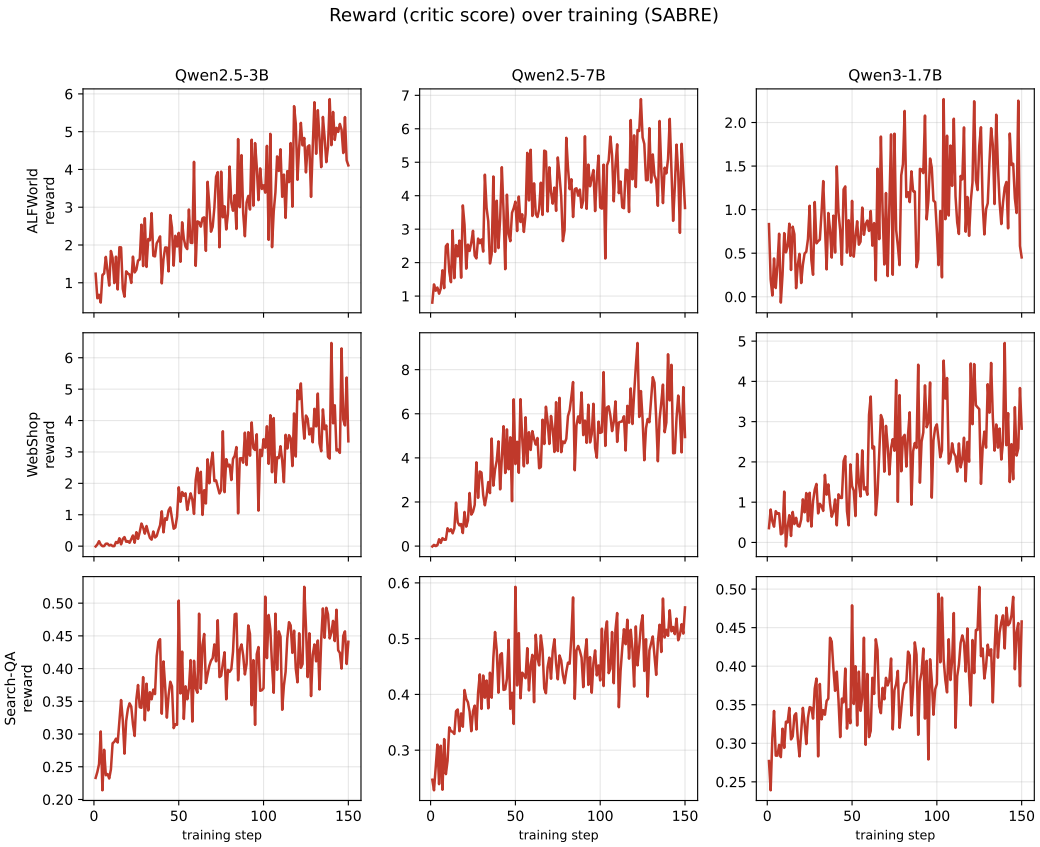

Figure 5: **Reward Curve** when training Qwen2.5-3B, Qwen2.5-7B and Qwen3-1.7B on ALFWorld, WebShop and Search-QA.

Prompt of AgentOPSD on ALFWorld

You are an expert agent operating in the ALFRED Embodied Environment. Your task is to: {task_description}.
{skill_context}
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 current step and present it within `<action> </action>` tags.

Figure 6: Prompt template used by AgentOPSD for the ALFWorld task environment.

Prompt of AgentOPSD on Search-based QA

You are an expert agent tasked with answering the given question step-by-step.
{skill_context}
Your question: {task_description}.
Prior to this step, you have already taken {step_count} step(s). Below is the interaction history where <search> </search> wrapped your past search queries and <information> </information> wrapped the corresponding search results returned by the external search engine. History:
{memory_context}
Now it's your turn to respond for the current step. You should first conduct a reasoning process. This process MUST be enclosed within <think> </think> tags. After completing your reasoning, choose only one of the following actions (do not perform both):

1. If you find you lack some knowledge, you **MUST** call a search engine to get more external information using format: <search> your query </search>.
2. If you have enough knowledge to answer the question confidently, provide your final answer within <answer> </answer> tags, without detailed illustrations. For example, <answer>Beijing</answer>.

Figure 7: Prompt template used by AgentOPSD for the Search-based QA task environment.

Prompt of AgentOPSD on WebShop

You are an expert autonomous agent operating in the WebShop e-commerce environment.
{skill_context}
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 current step and present it within <action> </action> tags.

Figure 8: Prompt template used by AgentOPSD for the WebShop task environment.