SAPD: LEARNING FROM AGENT EXPERIENCE WITHOUT ENVIRONMENT INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling the training of Large Language Model (LLM) agents typically requires a large collection of executable environments, whose construction, maintenance, and operation constrain the expansion of training experience. Existing offline trajectories offer an opportunity to reuse interaction experience, yet effectively refining the current policy without revisiting these environments remains challenging: supervised fine-tuning (SFT) can lead to catastrophic forgetting, while online reinforcement learning requires real-time interaction with the environment. To address this challenge, we propose **S**tep-**A**dvantage **P**olicy **D**istillation (SAPD), a method for learning from offline trajectories using relative step advantages. SAPD decomposes multi-turn trajectories into individual decision contexts for on-policy distillation, allowing the student to generate responses conditioned on recorded histories, and introduces relative step advantage signals to stabilize training. The entire process requires neither new environment observations nor additional multi-turn interactions. Experiments on ALFWorld, WebShop, and Search-based QA show that SAPD improves performance across different teacher configurations, mitigates catastrophic forgetting compared with SFT, and runs at least $4.7\times$ faster per training step than online OPD. We will release our code upon acceptance.

1 INTRODUCTION

Recent advances in large language models (LLMs) have enabled LLM agents to tackle increasingly complex, long-horizon tasks. Scaling agent training across these long-horizon tasks typically requires diverse executable environments in which models can take actions and receive feedback (Liao et al., 2026; Zhou et al., 2026). Recent models such as DeepSeek-V3.2 and Qwen3-Coder-Next further highlight this trend through large-scale construction of executable environments for agent training (DeepSeek-AI et al., 2025; Cao et al., 2026). However, deploying, maintaining, and operating these environments introduces substantial infrastructure demands and engineering complexity, making agent training difficult to scale across heterogeneous tasks (Liao et al., 2026).

Existing offline trajectories offer an opportunity to reuse interaction experience, yet effectively refining the current policy without revisiting the original environments remains challenging. Supervised fine-tuning (SFT) learns from these trajectories by imitating expert actions. However, when trajectories span heterogeneous domains, their supervision signals may conflict, causing interference between domain-specific behaviors. Sequential SFT can further lead to catastrophic forgetting, with gains on new tasks coming at the expense of previously acquired capabilities (Lu & Thinking Machines Lab, 2025). These challenges limit the use of SFT for continually expanding an agent's task coverage. In contrast, online reinforcement learning (RL) and on-policy distillation (OPD) optimize trajectories generated by the current policy and can substantially mitigate forgetting compared with SFT (Shenfeld et al., 2026; Lu & Thinking Machines Lab, 2025). However, their application through online multi-turn rollouts requires continued environment interaction, retaining the associated deployment and execution costs (Liao et al., 2026; Zhou et al., 2026). These limitations raise a central question: *How can we continually improve an agent's policy using offline interaction experience without revisiting the environment, while retaining previously acquired capabilities?*

To address this challenge, we propose **S**tep-**A**dvantage **P**olicy **D**istillation (SAPD), which refines the student policy using offline prefixes and teacher supervision. SAPD consists of two components: offline trajectory decomposition and context-relative step-advantage optimization. **Offline trajectory**

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.

decomposition converts recorded multi-turn trajectories into decision contexts, each retaining the preceding interaction history and current observation as a fixed prefix. For each prefix, the student generates a group of single-turn responses, and a fixed teacher provides token-level supervision on the generated responses. This enables SAPD to reuse offline interaction histories while optimizing the student's own decisions, without executing new actions or obtaining additional environment observations. However, offline prefix replay faces a two-sided mismatch: recorded histories may differ from those the student would encounter through its own interactions and may also be unfamiliar to the teacher. This mismatch can introduce context-dependent errors into teacher supervision, making updates based on absolute teacher–student discrepancies less reliable. To address this challenge, SAPD introduces **context-relative step-advantage optimization**, which aggregates token-level signals into action-level scores and compares candidate responses under the same decision context. Comparing responses under the same prefix yields relative advantages that provide decision-level guidance alongside token-level supervision. We incorporate these advantages into a joint objective to stabilize policy updates, alongside token-level OPD for efficiency and rule rewards for further stability. Experiments on three publicly available agentic benchmarks show that SAPD significantly improves performance across teacher configurations, substantially mitigates catastrophic forgetting compared with SFT, and runs at least $4.7\times$ faster per training step than online OPD in the measured configurations. In summary, our contributions are as follows:

- We introduce a framework that combines offline trajectory decomposition with policy optimization, enabling environment-free learning while mitigating catastrophic forgetting.
- We propose **S**tep-**A**dvantage **P**olicy **D**istillation (SAPD), which constructs relative step advantages to provide more stable decision-level supervision under behavior-prefix shift.
- Experiments on three publicly available agentic benchmarks demonstrate improved performance across teacher configurations, reduced forgetting compared with SFT, and lower per-step training time than online OPD.

2 Related work

LLM agentic learning. LLM-based agents are increasingly trained on long-horizon tasks that require sustained interaction with external environments (Yao et al., 2023). Supervised fine-tuning and behavior cloning enable learning from expert trajectories, but can suffer from gradient interference during joint training across heterogeneous tasks (Yu et al., 2020) and catastrophic forgetting during sequential training (Luo et al., 2025). Reinforcement learning methods such as GRPO optimize policy-generated behavior using task feedback (Shao et al., 2024), while on-policy distillation (OPD) supplies dense teacher supervision on student-generated responses (Agarwal et al., 2024; Gu et al., 2024; Lu & Thinking Machines Lab, 2025). However, these online methods rely on environment interaction throughout training. The deployment, maintenance, and operation of heterogeneous environments incur substantial engineering and resource costs, limiting the scalability of these methods across multiple environments (Liao et al., 2026).

Prefix-based OPD. OPD derives supervision from teacher probabilities rather than environment rewards, potentially enabling environment-free training. Recent work has explored reducing rollout requirements of OPD through prefix-based learning (Zhao et al., 2026; Gui et al., 2026). Reasoning-Prefix OPD shows that teacher-generated prefixes can guide subsequent student reasoning (Zhang et al., 2026). BPO reuses prefixes to explore alternative actions, but still requires environment interaction to evaluate their rollouts (He et al., 2026). Prefix-GRPO replays teacher trajectories and optimizes student using task rewards from environment feedback, whereas ReOPD uses teacher supervision on student rollouts conditioned on the teacher's prefixes without further environment interaction (Wang et al., 2026a; Liao et al., 2026). In practice, these approaches rely on teacher-generated trajectories, whose collection requires deploying and maintaining the corresponding environments, which limits their ability to learn directly from arbitrary offline interaction trajectories.

Reliable OPD under prefix mismatch. However, replaying prefixes from arbitrary offline trajectories faces a two-sided mismatch: recorded histories may differ from those the student would encounter through its own interactions and may also be unfamiliar to the teacher, which can substantially destabilize OPD training. Recent work has investigated OPD instability under prefix mismatch

and explored ways to improve the reliability of teacher supervision (Zhu et al., 2026; Yang et al., 2026). TA-OPD shows that some token-level teacher–student discrepancies are unreliable (Wang et al., 2026b). ReNIO aggregates token signals into sample-level distillation weights, which can be too coarse for credit assignment in long-horizon tasks (Lin et al., 2026). Furthermore, SOD adjusts step-level distillation weights by comparing discrepancies across steps within the same trajectory, but these comparisons span different decision contexts and may therefore be unreliable (Zhong et al., 2026). SAPD instead computes relative advantages from aggregated action-token signals under the same offline prefix, supporting more stable credit assignment.

3 BACKGROUND

Multi-turn agent interaction. Let $x \sim \mathcal{D}$ be a task and $\mathcal{E}$ an environment with observation kernel $P_{\mathcal{E}}$. At turn t, the agent samples a response $a_t \sim \pi_\theta(\cdot \mid h_t)$ from history $h_t = (x, o_1, a_1, \ldots, a_{t-1}, o_t)$. The response includes any reasoning and an executable action; executing the action yields $o_{t+1} \sim P_{\mathcal{E}}(\cdot \mid h_t, a_t)$. A complete trajectory $\tau = (o_1, a_1, \ldots, o_T, a_T, o_{T+1})$, ending at termination or a turn limit, follows

$$p_\theta^{\mathcal{E}}(\tau \mid x) = P_{\mathcal{E}}(o_1 \mid x) \prod_{t=1}^{T} \pi_\theta(a_t \mid h_t) P_{\mathcal{E}}(o_{t+1} \mid h_t, a_t). \tag{1}$$

Index the L_i agent-generated tokens in τ_i by $y_{i,n}$, with context $c_{i,n}$ containing the task, preceding tokens, and intervening environment observations. Environment tokens are excluded from the optimization loss. Both online methods below collect complete multi-turn trajectories under a frozen rollout policy π_{old} and hold sampled contexts fixed during each update.

Multi-turn on-policy distillation (OPD). OPD trains the student on self-generated sequences using feedback from a fixed teacher π_T (Agarwal et al., 2024). In the standard online multi-turn agent formulation (Zhou et al., 2026), the student generates interaction trajectories through $\mathcal{E}$, executing actions and receiving new observations until termination or a prescribed horizon; the fixed teacher scores the student's response tokens at the resulting histories. Following reverse-KL distillation (Gu et al., 2024), a token-wise objective on these sampled contexts is

$$\mathcal{L}_{\text{OPD}}(\theta; \pi_{\text{old}}) = \mathbb{E}_{x \sim \mathcal{D},\, \tau_i \sim p_{\text{old}}^{\mathcal{E}}(\cdot \mid x)} \left[\frac{1}{L_i} \sum_{n=1}^{L_i} D_{\text{KL}}(\pi_\theta(\cdot \mid c_{i,n}) \,\|\, \pi_T(\cdot \mid c_{i,n})) \right]. \tag{2}$$

Sampled implementations (Lu & Thinking Machines Lab, 2025) use the detached log-probability difference $\log \pi_T(y_{i,n} \mid c_{i,n}) - \log \pi_{\text{old}}(y_{i,n} \mid c_{i,n})$ as token-level advantages. Although OPD does not require task rewards, its later-turn contexts still depend on observations obtained by executing earlier student actions. Thus, both methods require a live environment to refresh their multi-turn rollout distributions, motivating refinement from previously recorded interaction histories.

4 METHOD

SAPD decouples policy refinement from the environment execution required by the online methods in Section 3. As illustrated in Figure 1, we first decompose recorded multi-turn trajectories into a pool of decision prefixes that retain the preceding actions and environment observations. For each prefix, the student generates a group of single-turn responses, which are scored by a fixed teacher without executing their actions. We combine token-level distillation with a step-level relative advantage computed by comparing aggregated teacher–student signals within the group. This provides decision-level supervision alongside fine-grained token feedback. Locally computable rewards, such as repetition penalties and format checks, further support stable training. Together, these components optimize the student's responses using existing interaction histories, requiring neither new environment observations nor additional multi-turn rollouts during training.

4.1 POLICY DISTILLATION OVER OFFLINE PREFIXES

Let $\mathcal{B}$ be a collection of recorded multi-turn trajectories obtained from sources such as human demonstrations, publicly available multi-turn SFT datasets, logged agent interactions, and teacher-generated rollouts. Using the history notation in Section 3, we construct the multiset of decision

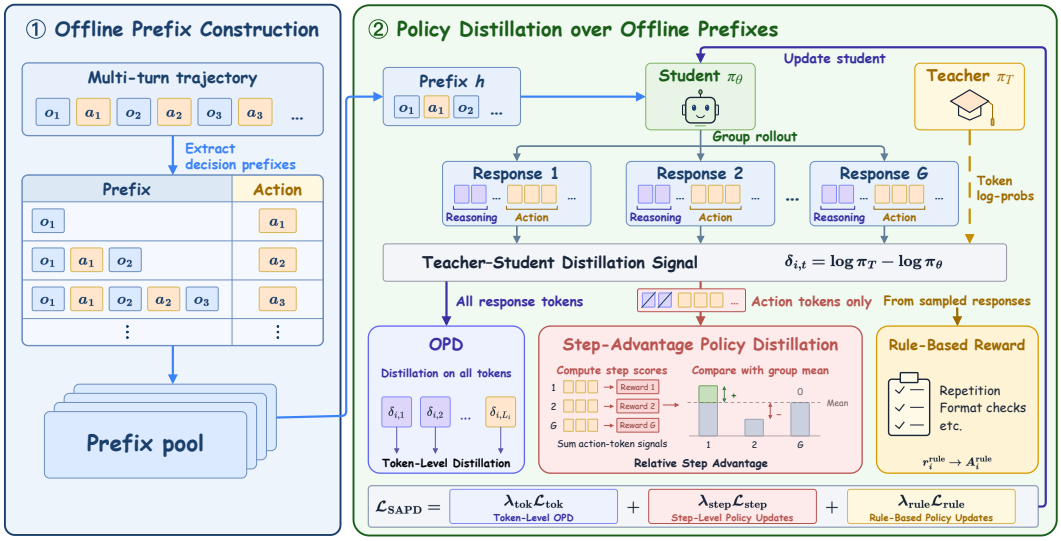


Figure 1: Overview of our environment-free policy refinement framework.

prefixes

$$\mathcal{P}_{\mathcal{B}} = \{h_t^\tau : \tau \in \mathcal{B},\ 1 \le t \le T_\tau\}, \qquad h \sim q_{\mathcal{B}}, \qquad y_i \sim \pi_{\text{old}}(\cdot \mid h), \quad i = 1, \dots, G, \tag{3}$$

where T_τ is the number of turns in τ and $q_{\mathcal{B}}$ is the empirical distribution over prefix entries. Each y_i is a newly generated single-turn response containing reasoning, when applicable, and an action. For example, the recorded interaction (o_1, a_1, o_2, a_2) supplies the prefix (x, o_1, a_1, o_2) for generating a new second response. Its original action a_2 is not used as the token-level distillation target. We use $G \ge 2$ responses per prefix, and none of their actions is executed.

This sampling scheme follows the prefix-replay setting (Liao et al., 2026): responses are sampled from the current rollout policy, while histories remain distributed according to $q_{\mathcal{B}}$. It therefore provides on-policy response generation conditional on offline histories, without claiming on-policy coverage of the student's environment-induced histories. The resulting behavior-prefix shift motivates controlling the learning signals assigned to responses within each recorded context.

For a response $y_i = (y_{i,1}, \dots, y_{i,L_i})$, define $c_{i,k} = (h, y_{i,<k})$. Here L_i counts tokens in one response, rather than a complete multi-turn trajectory. The fixed teacher π_T, assumed to share the student's tokenization, scores these same student-generated tokens. We use the detached teacher–student signal and token importance ratio

$$\begin{aligned} d_{i,k} &= \text{sg}[\log \pi_T(y_{i,k} \mid c_{i,k}) - \log \pi_{\text{old}}(y_{i,k} \mid c_{i,k})], \\ \rho_{i,k}(\theta) &= \frac{\pi_\theta(y_{i,k} \mid c_{i,k})}{\pi_{\text{old}}(y_{i,k} \mid c_{i,k})}. \end{aligned} \tag{4}$$

where sg denotes stop-gradient. The sampled token-level OPD surrogate is

$$\mathcal{L}_{\text{tok}}(\theta) = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{k=1}^{L_i}\rho_{i,k}(\theta)d_{i,k}\right]. \tag{5}$$

All expectations in this section follow Equation 3; sampled responses, teacher scores, and rollout-policy probabilities are held fixed during an update. This token-local surrogate uses the negative sampled reverse-KL signal introduced in the background, covering both reasoning and action tokens. Prefix tokens receive no training loss.

4.2 Step-Advantage Policy Distillation

Token-level distillation supplies local learning directions, but does not explicitly compare complete decisions made under the same history. We introduce a complementary step-level signal by comparing the teacher–student scores of responses within each prefix-conditioned group. In this section, a *step* denotes one agent decision turn, not a token or an optimization iteration.

Action-based step scores. Let $\mathcal{A}_i \subseteq \{1, \ldots, L_i\}$ index the action tokens in response y_i, excluding its reasoning segment. We aggregate the teacher–student signal over this set:

$$r_i^{\text{step}} = \sum_{k \in \mathcal{A}_i} d_{i,k}, \qquad \bar{r}^{\text{step}}(h) = \frac{1}{G} \sum_{j=1}^{G} r_j^{\text{step}}, \qquad A_i^{\text{step}} = \text{sg}\big[r_i^{\text{step}} - \bar{r}^{\text{step}}(h)\big]. \tag{6}$$

Restricting the score to action tokens focuses the comparison on the decision and prevents reasoning length from directly contributing to the aggregate. This choice also limits length-dependent scoring error: as Equation 8 shows, a shared per-token bias leaves a residual proportional to differences in the number of scored tokens. Excluding reasoning tokens avoids an additional source of such variation from potentially long and variable reasoning traces. Reasoning still conditions the teacher and student probabilities and receives token-level supervision through Equation 5. Action-length differences can still contribute residual error; the sum is not length normalized. A positive advantage indicates an above-group-average teacher-to-student log-likelihood ratio, rather than verified environment success.

Relative credit under prefix shift. In online multi-turn OPD, student-generated histories can be unfamiliar to the teacher, making its supervision unreliable; a large teacher–student gap can exacerbate this mismatch (Liao et al., 2026). With heterogeneous offline data, a recorded prefix can additionally be atypical under both policies. We therefore seek tolerance to context-dependent scoring errors, rather than assuming that every teacher likelihood is well calibrated on the offline pool.

We begin with the token-level signal in Equation 4. For a fixed prefix h and its sampled responses, let $d^\star_{i,k}$ denote an ideal calibrated supervisory signal, defined formally in Appendix A.1. On the scored action tokens, consider the structured error model

$$d_{i,k} = d^\star_{i,k} + b_{\text{tok}}(h) + \xi_{i,k}, \qquad k \in \mathcal{A}_i, \tag{7}$$

where $b_{\text{tok}}(h)$ is a prefix-dependent component shared across the scored tokens and $\xi_{i,k}$ captures token- and response-dependent deviations. This is an assumption about sampled scoring errors, not a property guaranteed by OOD inputs. With $N_i = |\mathcal{A}_i|$, $r_i^\star = \sum_{k \in \mathcal{A}_i} d^\star_{i,k}$, and $u_i = \sum_{k \in \mathcal{A}_i} \xi_{i,k}$, aggregation and group centering give

$$\begin{aligned} r_i^{\text{step}} - r_i^\star &= N_i b_{\text{tok}}(h) + u_i, \\ A_i^{\text{step}} - A_i^\star &= b_{\text{tok}}(h)(N_i - \bar{N}) + (u_i - \bar{u}), \end{aligned} \tag{8}$$

where bars denote group means and $A_i^\star = r_i^\star - G^{-1} \sum_j r_j^\star$. For equal action lengths, the shared token offset cancels exactly, leaving only differences between accumulated residual errors. With variable lengths, a length-dependent term remains. Thus, the benefit comes from removing the shared component of the aggregate error, not from assuming that summation makes every token error disappear.

More generally, write the aggregate error as a shared offset plus a residual:

$$r_i^{\text{step}} = r_i^\star + b(h) + e_i, \qquad A_i^{\text{step}} - A_i^\star = e_i - \bar{e}, \tag{9}$$

where $\bar{e} = G^{-1} \sum_j e_j$. For fixed action length N, Equation 7 yields $b(h) = N b_{\text{tok}}(h)$ and $e_i = u_i$; otherwise, e_i must also account for length-dependent error. A useful robustness guarantee requires these residual differences to be small.

Proposition 1 (Robustness to shared score offsets). Let $P = I_G - G^{-1}\mathbf{1}\mathbf{1}^\top$ be the group-centering matrix, with I_G the identity and $\mathbf{1}$ the all-ones vector. For every fixed group, the advantage error is Pe, independent of $b(h)$, and

$$\|A^{\text{step}} - A^\star\|_2^2 = \|Pe\|_2^2 = \sum_{i=1}^{G} (e_i - \bar{e})^2 \le \sum_{i=1}^{G} e_i^2. \tag{10}$$

Moreover, if $|e_i| \le \varepsilon_h$, then $|A_i^{\text{step}} - A_i^\star| \le 2\varepsilon_h$; an ideal advantage with magnitude greater than $2\varepsilon_h$ retains its sign. Here ε_h bounds scoring error and is distinct from the clipping threshold ϵ. Thus, relative step credit can remain informative despite large shared offsets, provided response-dependent errors are small. Appendix A proves the result, derives a local update-error bound, and discusses its limitations. This property does not correct arbitrary teacher errors or restore missing coverage of student-relevant histories.

Relative credit without variance normalization. We retain the magnitude of the centered signal instead of dividing by its group standard deviation. For $r = b\mathbf{1} + \alpha v$, where $\alpha \geq 0$ controls the size of within-group differences, the advantage is $Pr = \alpha Pv$. It vanishes as these differences vanish. By comparison, standardization gives $Pr/(\mathrm{std}(r)+\eta) = \alpha Pv/(\alpha\,\mathrm{std}(v)+\eta)$: when the stabilizer η is negligible, even small nonzero differences can produce an order-one signal. Omitting this division avoids such amplification, although a small score spread alone does not establish teacher reliability.

Step-level policy updates. We broadcast A_i^{step} to the generated tokens of response i and use the clipped surrogate

$$\begin{aligned} \mathcal{C}_\epsilon(\rho, A) &= \min\{\rho A, \mathrm{clip}(\rho, 1-\epsilon, 1+\epsilon)A\}\,, \\ \mathcal{L}_{\mathrm{step}}(\theta) &= -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{k=1}^{L_i}\mathcal{C}_\epsilon\big(\rho_{i,k}(\theta), A_i^{\mathrm{step}}\big)\right]. \end{aligned} \tag{11}$$

The scoring mask and update mask serve different purposes: only action tokens determine r_i^{step}, whereas the response-level advantage updates both reasoning and action tokens. This supplies a shared decision-level signal alongside the distinct per-token signals in $\mathcal{L}_{\mathrm{tok}}$.

4.3 Joint Training Objective

Optimizing likelihood-based supervision can create opportunities for supervision hacking: the student may exploit favorable teacher scores for degenerate responses without improving its decisions. For example, repeatedly issuing the same action can become a spurious local optimum of the training surrogate. Group-relative step advantages compare responses using the same likelihood-based signal and therefore do not explicitly rule out such behavior. We therefore introduce locally computable rule rewards to discourage these failure modes, including penalties for repeated actions detectable from the recorded history and checks on the generated action format. These signals constrain exploitable response patterns independently of teacher likelihoods and require no environment execution. Let $f_m(h, y_i)$ be the score of rule m and $\alpha_m \geq 0$ its weight, with violations represented by negative scores. Define

$$r_i^{\mathrm{rule}} = \sum_{m=1}^{M}\alpha_m f_m(h, y_i), \qquad A_i^{\mathrm{rule}} = \mathrm{sg}\left[\frac{r_i^{\mathrm{rule}} - \bar{r}^{\mathrm{rule}}}{\mathrm{std}(r_1^{\mathrm{rule}}, \ldots, r_G^{\mathrm{rule}}) + \eta}\right], \tag{12}$$

where M is the number of rules and $\bar{r}^{\mathrm{rule}}$ is their group-mean aggregate reward. The rule objective uses the same clipped surrogate:

$$\begin{aligned} \mathcal{L}_{\mathrm{rule}}(\theta) &= -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_i}\sum_{k=1}^{L_i}\mathcal{C}_\epsilon\big(\rho_{i,k}(\theta), A_i^{\mathrm{rule}}\big)\right], \\ \mathcal{L}_{\mathrm{SAPD}} &= \lambda_{\mathrm{tok}}\mathcal{L}_{\mathrm{tok}} + \lambda_{\mathrm{step}}\mathcal{L}_{\mathrm{step}} + \lambda_{\mathrm{rule}}\mathcal{L}_{\mathrm{rule}}. \end{aligned} \tag{13}$$

The nonnegative coefficients λ_{tok}, λ_{step}, and λ_{rule} control the contributions of the three objectives. As shown by the ablation experiments presented later, the performance gains primarily stem from step advantages, while auxiliary token-level OPD improves learning efficiency. Rule rewards further discourage supervision hacking. Rule definitions and coefficient values are specified in the experimental setup. Each iteration samples offline prefixes, generates response groups, computes teacher and rule scores, and minimizes the joint loss with respect to the student alone. The rollout policy is refreshed from the student for subsequent sampling. All supervision is computed from recorded histories and model-generated responses, with no action execution, new environment observation, or environment-return evaluation during policy refinement.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate SAPD on three widely used long-horizon agent tasks: ALFWorld, WebShop, and Search-based QA. **ALFWorld** (Shridhar et al., 2021) is a text-based environment in which agents follow instructions to complete household goals. It comprises six task categories: Pick, Look,

Clean, Heat, Cool, and Pick2. **WebShop** (Yao et al., 2022) is a simulated online shopping environment in which agents search a product catalog, navigate product pages, and select products and options that satisfy user-specified requirements. **Search-based QA** follows the setup of Search-R1 (Jin et al., 2025), requiring agents to interleave reasoning with search, gather relevant evidence, and produce a final answer. It includes three single-hop datasets, Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023), and four multi-hop datasets, HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). Following the setup of SDAR (Lu et al., 2026), we evaluate 128 episodes per evaluation round on each of ALFWorld and WebShop, and all 51,713 questions across the seven Search-based QA evaluation datasets.

Baselines. To evaluate the effectiveness of SAPD, we compare against the following three categories of baselines. **(1) Supervised Fine-Tuning (SFT)** (Zeng et al., 2024) learns to imitate expert actions in fixed demonstration trajectories. As a fully off-policy baseline, SFT allows us to assess SAPD's task performance and resistance to catastrophic forgetting under sequential task training. **(2) Group Relative Policy Optimization (GRPO)** (Shao et al., 2024) collects trajectories by having the student interact with the environment and updates the policy using advantages computed from relative rewards within each rollout group. Its reliance on extensive environment interaction provides a comparison for SAPD in both task performance and training efficiency. **(3) On-Policy Distillation (OPD)** (Agarwal et al., 2024) applies token-level teacher supervision to student-generated responses conditioned on offline SFT prefixes. OPD+Rule further incorporates rule-based rewards to stabilize training. Its objective differs from SAPD only by omitting our step-advantage policy distillation, isolating the contribution of action-level guidance.

Implementation Details. We use Qwen2.5-1.5B-Instruct as the student backbone and SFT- or GRPO-trained Qwen2.5-3B-Instruct models as frozen teachers (Yang et al., 2024). The student samples eight responses per decision prefix extracted from offline SFT trajectories without additional environment interaction. The training framework and experimental hyperparameters follow SDAR (Lu et al., 2026). Full details of implementation are provided in Appendix B.

5.2 Main Results

Table 1 shows the performance of SAPD across two teacher configurations and three long-horizon agentic benchmarks. Based on these experiments, we summarize the following key observations:

SAPD improves upon OPD-based methods. Table 1 demonstrates the effectiveness of SAPD in both benchmark performance and learning dynamics. SAPD outperforms OPD+Rule on all three benchmarks under both teacher configurations. With GRPO teachers, the gains are 6.5, 0.3, and 4.4 percentage points on ALFWorld, Search-based QA, and WebShop, respectively; with SFT teachers, the corresponding improvements are 2.6, 0.2, and 6.3 points. These gains highlight SAPD's ability to mitigate distillation instability in long-horizon tasks, where the teacher may provide unreliable supervision on student-induced trajectories due to distribution mismatch.

Teacher strength alone does not determine distillation quality. On WebShop, the GRPO teacher achieves a success rate of 64.8%, substantially exceeding the SFT teacher's 59.7%. However, the student trained with the SFT teacher reaches 59.4%, compared with 53.4% for the GRPO teacher. This difference highlights the distinction between a teacher's standalone performance and its effectiveness as a supervisor. In our experimental setting, all distillation methods condition student generation on prefixes extracted from offline SFT trajectories, and teachers provide feedback on the resulting student-generated responses. These prefixes reflect SFT behavior and may differ from the histories encountered by the GRPO teacher under its own policy. Consequently, stronger performance during autonomous execution does not guarantee more useful supervision in the distillation setting. The WebShop results therefore show that teacher selection should account for compatibility with the distillation contexts, rather than standalone task performance alone.

Long-horizon tasks benefit more from SAPD. SAPD provides larger gains on ALFWorld and WebShop than on Search-based QA, as shown in Table 1. This difference can be understood in terms of interaction horizon: ALFWorld allows up to 50 steps and WebShop up to 15, compared with four for Search-based QA. Over longer interactions, early decisions shape subsequent observations and actions, allowing errors and mismatches between teacher and student behavior to accumulate. SAPD compares candidate actions under the same prefix, reducing the impact of context mismatch

	ALFWorld							Search-based QA								WebShop
Method	Pick	Look	Clean	Heat	Cool	Pick2	**Avg.**	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	**Avg.**	**Success**
SFT	100.0	100.0	94.1	73.7	88.2	84.2	90.6	32.8	47.4	38.3	29.9	36.5	11.1	28.0	37.0	51.6
GRPO	85.3	53.7	84.5	78.2	59.7	53.5	72.8	34.1	54.5	37.8	32.4	31.9	10.3	26.4	37.7	56.8
GRPO teacher model: Qwen2.5-3B-Instruct																
GRPO	96.8	88.2	92.9	91.7	90.0	70.0	89.1	43.4	60.8	46.2	39.3	35.3	14.8	66.9	44.5	64.8
OPD	83.4	60.7	72.7	49.1	42.0	13.3	57.0	11.1	25.6	16.9	8.6	15.4	1.7	12.8	16.1	29.9
OPD+Rule	83.0	48.0	82.5	72.5	55.1	16.5	62.8	**40.8**	55.4	42.9	36.2	34.3	12.7	**31.2**	41.0	49.0
SAPD	**90.7**	**60.9**	**82.7**	**83.7**	**72.9**	**18.9**	**69.3**	40.0	**56.1**	**43.2**	**36.5**	**34.5**	**13.3**	28.8	**41.3**	**53.4**
SFT teacher model: Qwen2.5-3B-Instruct																
SFT	100.0	100.0	93.5	89.5	91.3	95.0	94.5	40.7	58.4	43.0	36.5	34.6	12.5	63.3	42.0	59.7
OPD	90.3	**93.0**	87.0	84.0	73.2	85.9	86.7	9.0	22.6	12.7	6.3	7.2	2.2	6.4	11.9	47.1
OPD+Rule	97.7	82.9	**95.4**	90.5	96.8	95.5	94.0	**34.1**	49.4	39.9	**35.9**	35.1	**15.0**	32.8	38.6	53.1
SAPD	**99.0**	85.9	94.5	**94.8**	**100.0**	**98.1**	**96.6**	34.0	**49.5**	**40.1**	35.4	**36.1**	14.1	**37.6**	**38.8**	**59.4**

Table 1: Main results on ALFWorld, Search-based QA, and WebShop with SFT and GRPO teachers. We report success rates on ALFWorld and WebShop and accuracy on Search-based QA. The best results are bolded, and the second-best results are underlined.

on the distillation signal. The resulting relative step-level advantages provide more stable guidance for policy optimization, contributing to larger gains on longer-horizon tasks.

	ALFWorld	Search	WebShop
Method	Success	Overall	Success
GRPO teacher			
SAPD	**69.3**	**41.3**	**53.4**
w/o OPD	66.9	41.0	51.6
w/o Rule	62.0	40.7	52.1
SFT teacher			
SAPD	**96.6**	**38.8**	**59.4**
w/o OPD	95.8	38.7	56.0
w/o Rule	95.3	38.3	55.2

Table 2: Results of the ablation study.

	Evaluation after training		
Training stage	**WebShop**	**ALFWorld**	**Search**
SAPD			
Stage 1: WebShop	**52.3**	16.4	16.2
Stage 2: ALFWorld	57.0	**96.1**	13.9
Stage 3: Search	50.8	90.5	**36.6**
SFT			
Stage 1: WebShop	**51.6**	15.3	16.9
Stage 2: ALFWorld	22.7	**92.2**	10.2
Stage 3: Search	2.3	7.8	**37.2**

Table 3: Performance under sequential training.

5.3 Ablation Study

Table 2 presents the ablation results for SAPD, from which we draw two key observations.

Step-level advantages provide the primary learning signal, while token-level OPD plays an auxiliary role. To assess the contribution of step-advantage policy optimization, we remove the token-level OPD branch while retaining step-level advantages and rule-based rewards, yielding a variant denoted as w/o OPD. The w/o OPD variant exhibits only modest performance declines relative to SAPD. Compared with OPD+Rule in Table 1, it outperforms the baseline across all benchmarks and teacher configurations. Specifically, it improves ALFWorld performance by 4.1 points with a GRPO teacher and 1.8 points with an SFT teacher; the corresponding WebShop gains are 2.6 and 2.9 points. These results demonstrate that step-advantage policy optimization accounts for most of gains, while token-level OPD provides dense supervision for further performance improvements.

Rule-based rewards help stabilize the training. To discourage undesirable output behaviors, we supplement teacher-derived supervision with rule-based rewards. We isolate their contribution through w/o Rule, which removes these rewards while retaining step-level advantages and token-level OPD. Removing rule-based rewards consistently degrades performance across all evaluated configurations. For example, ALFWorld success falls from 69.3% to 62.0% with a GRPO teacher, while WebShop success declines from 59.4% to 55.2% with an SFT teacher. These results suggest

that rule-based rewards complement teacher-derived learning signals. Meanwhile, SAPD retains substantial task performance even without rule-based rewards, indicating its effectiveness.

5.4 Sequential Training Compared With SFT

Sequential training requires agents to learn new tasks while maintaining performance on earlier tasks. We investigate whether SAPD mitigates catastrophic forgetting relative to SFT by training both methods sequentially on WebShop, ALFWorld, and Search-based QA. In Table 3, each row corresponds to a completed training stage. Diagonal entries show performance on the newly learned task, and entries below the diagonal track performance on earlier tasks. The two methods begin with similar WebShop success rates of 52.3% for SAPD and 51.6% for SFT, but their performance diverges during subsequent training. Following ALFWorld training, SAPD's WebShop success rises to 57.0%, possibly because the shared output format supports generalization across tasks. In contrast, SFT's WebShop success falls to 22.7%, indicating substantial forgetting of the previously learned task. By the end of Search training, SAPD still achieves 50.8% on WebShop and 90.5% on ALFWorld, whereas SFT reaches only 2.3% and 7.8%, respectively. This retention advantage accompanies comparable accuracy on the final Search task, which demonstrates that SAPD substantially reduces catastrophic forgetting without compromising its ability to learn new tasks.

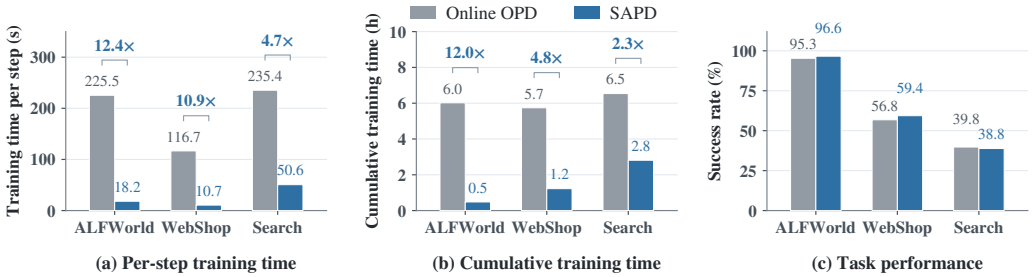


Figure 2: Training efficiency and evaluation performance of SAPD and Online OPD.

5.5 Training Efficiency Compared with Online OPD

Repeated environment interaction is a major source of overhead in online agent training, motivating an efficiency comparison between SAPD and online OPD. Figure 2 compares per-step training time, cumulative training time, and task performance. Compared with online OPD, SAPD achieves per-step speedups of $12.4\times$ on ALFWorld, $10.9\times$ on WebShop, and $4.7\times$ on Search-based QA. These gains also translate into cumulative training-time reductions by factors of approximately 12.0, 4.8, and 2.3, respectively. SAPD achieves these savings with higher performance on ALFWorld and WebShop and slightly lower accuracy on Search-based QA. Furthermore, the observed speedup increases across benchmarks with longer interaction horizons, from Search-based QA with at most four steps to WebShop with 15 and ALFWorld with 50. Longer rollouts require online OPD to repeatedly generate responses and interact with the environment, with repetitive or unsuccessful actions potentially further extending these time costs. SAPD avoids this overhead by reusing offline trajectories, yielding greater efficiency gains on longer-horizon tasks.

6 Conclusion

We presented SAPD, a method for refining LLM agents from recorded interaction histories without further environment interaction. SAPD decomposes multi-turn trajectories into decision prefixes and compares student-generated responses under each shared prefix to construct relative step advantages, providing decision-level guidance alongside token-level distillation. Experiments on ALFWorld, WebShop, and Search-based QA show improved performance over OPD-based baselines across SFT and GRPO teacher configurations, reduced catastrophic forgetting relative to sequential SFT, and at least a $4.7\times$ reduction in per-step training time compared with online OPD under the evaluated configurations. These results demonstrate the potential of reusing recorded interaction experience to expand agent capabilities while reducing reliance on executable training environments.

AI USE STATEMENT

We used generative AI tools to polish the manuscript and improve the readability of the code prepared for submission. The authors developed the core ideas, designed the algorithms, implemented the code, and conducted the experiments. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final manuscript, its claims, and all accompanying artifacts.

REPRODUCIBILITY STATEMENT

To support reproducibility, we use publicly available benchmarks and base models, with references provided in the main text. The methodology section details SAPD's offline prefix replay, relative step-advantage computation, and combined training objective. The appendix provides details on trajectory preparation, training configurations, hyperparameters, and evaluation protocols. Our source code, including scripts for data preprocessing, training, and evaluation, is provided in the supplementary materials to facilitate reproduction of our results.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Ruisheng Cao, Mouxiang Chen, Jiawei Chen, Zeyu Cui, Yunlong Feng, Binyuan Hui, Yuheng Jing, Kaixin Li, Mingze Li, Junyang Lin, et al. Qwen3-coder-next technical report. *arXiv preprint arXiv:2603.00729*, 2026.

DeepSeek-AI et al. DeepSeek-V3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025. URL https://arxiv.org/abs/2512.02556.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Zhiyu Gui, Kexin Huang, Jia Guo, Junkang Wu, Zihao Wang, Zhiqiang Zhang, Jun Zhou, Jiancan Wu, and Xiang Wang. Know when to stop, where to restart: Accelerating multi-turn agentic on-policy distillation. *arXiv preprint arXiv:2609.14636*, 2026.

Bowei He, Yankai Chen, Xiaokun Zhang, and Xue Liu. Branching policy optimization: Sandbox-native language agent reinforcement learning. *arXiv preprint arXiv:2607.14171*, 2026.

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.

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.

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.

Weizhen Li, Jianbo Lin, Zhuosong Jiang, Jingyi Cao, Xinpeng Liu, Jiayu Zhang, Zhenqiang Huang, Qianben Chen, Weichen Sun, Qiexiang Wang, Hongxuan Lu, Tianrui Qin, Chenghao Zhu, Yi Yao, Shuying Fan, Xiaowan Li, Tiannan Wang, Pai Liu, King Zhu, He Zhu, Dingfeng Shi, Piaohong Wang, Yeyi Guan, Xiangru Tang, Minghao Liu, Yuchen Eleanor Jiang, Jian Yang, Jiaheng Liu, Ge Zhang, and Wangchunshu Zhou. Chain-of-agents: End-to-end agent foundation models via multi-agent distillation and agentic rl, 2025. URL https://arxiv.org/abs/2508.13167.

Baohao Liao, Hanze Dong, Christof Monz, Xinxing Xu, Li Dong, and Furu Wei. Multi-turn on-policy distillation with prefix replay. *arXiv preprint arXiv:2607.04763*, 2026. doi: 10.48550/arXiv.2607.04763. URL https://arxiv.org/abs/2607.04763.

Chen Lin, Kedi Chen, and Wei Zhang. Renio: Reweighting negative trajectory importance for llm on-policy distillation. *arXiv preprint arXiv:2606.23104*, 2026.

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

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*, 2026.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3776–3786, 2025. doi: 10.1109/TASLPRO.2025.3606231.

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.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. Rl's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations*, volume 2026, pp. 59839–59864, 2026.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. {ALFW}orld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

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

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*, 2025.

Yihan Wang, Zhong Guan, Haoran Sun, Jiale Huang, Likang Wu, and Hongke Zhao. From trajectories to prefixes: Reusing teacher trajectories via replayed prefixes and online continuation. *arXiv preprint arXiv:2607.19395*, 2026a.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026b.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. SkillRL: Evolving agents via recursive skill-augmented reinforcement learning. In *ICLR 2026 Workshop on Memory for LLM-Based Agentic Systems*, 2026. URL https://openreview.net/forum?id=By7Pj576U3.

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, 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. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Zhicheng Yang, Zhijiang Guo, Yifan Song, Minrui Xu, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Prune-opd: Efficient and reliable on-policy distillation for long-horizon reasoning. *arXiv preprint arXiv:2605.07804*, 2026.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3fe78a8acf5fda99de95303940a2420c-Paper.pdf.

Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. Agenttuning: Enabling generalized agent abilities for llms. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3053–3077, 2024.

Dongxu Zhang, Zhichao Yang, Sepehr Janghorbani, Jun Han, Andrew Ressler II, Qian Qian, Gregory D Lyng, Sanjit Singh Batra, and Robert E Tillman. Fast and effective on-policy distillation from reasoning prefixes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25553–25569, 2026.

Qingfei Zhao, Huan Song, Shuyu Tian, Jiawei Shao, and Xuelong Li. Prefix-guided on-policy distillation: Mining golden trajectories from rollouts. *arXiv preprint arXiv:2606.21994*, 2026.

Qiyong Zhong, Mao Zheng, Mingyang Song, Xin Lin, Jie Sun, Houcheng Jiang, Xiang Wang, and Junfeng Fang. Sod: Step-wise on-policy distillation for small language model agents. *arXiv preprint arXiv:2605.07725*, 2026.

Yuhang Zhou, Kai Zheng, Haoling Li, Dengyun Peng, Can Xu, and Jingjing Chen. TurnOPD: Making on-policy distillation turn-aware for efficient long-horizon agent training. *arXiv preprint arXiv:2607.05804*, 2026. doi: 10.48550/arXiv.2607.05804. URL https://arxiv.org/abs/2607.05804.

Siqi Zhu, Xuyan Ye, Hongyu Lu, Weiye Shi, and Ge Liu. The many faces of on-policy distillation: Pitfalls, mechanisms, and fixes. *arXiv preprint arXiv:2605.11182*, 2026.

A Analysis of Relative Step Credit under Prefix Shift

This appendix analyzes sensitivity of the step-advantage branch to teacher scoring errors on a fixed set of offline contexts. The results concern the centered supervision and its local policy update; they are not convergence guarantees or guarantees of improved task success under arbitrary distribution shift.

A.1 Error Model and the Role of Sequence Aggregation

Fix a prefix h and G responses sampled according to Equation 3. For analysis only, introduce an ideal supervisory distribution $\pi_\star$ that shares the tokenization and assigns positive probability to the scored tokens. The ideal token signal and action score are

$$d^\star_{i,k} = \log \pi_\star(y_{i,k} \mid c_{i,k}) - \log \pi_{\mathrm{old}}(y_{i,k} \mid c_{i,k}), \qquad r^\star_i = \sum_{k \in \mathcal{A}_i} d^\star_{i,k}. \tag{14}$$

The distribution $\pi_\star$ is a theoretical comparator, not an additional model available to SAPD. No equality between $r^\star_i$ and environment return is assumed. The aggregate teacher error is exactly

$$\Delta_i = r^{\mathrm{step}}_i - r^\star_i = \sum_{k \in \mathcal{A}_i} \log \frac{\pi_T(y_{i,k} \mid c_{i,k})}{\pi_\star(y_{i,k} \mid c_{i,k})}. \tag{15}$$

Our structured-error model is $\Delta_i = b(h) + e_i$. The algebra below holds for any such decomposition; a substantive robustness interpretation requires the residual differences $e_i - \bar{e}$ to be small. In particular, defining $b(h)$ as a mean does not by itself establish that the remaining errors are small. The model permits a teacher to misestimate absolute scores on an unfamiliar prefix while preserving useful relative distinctions among responses.

At the token level, Equation 7 decomposes the log-ratio error $d_{i,k} - d^\star_{i,k} = \log \pi_T(y_{i,k} \mid c_{i,k}) - \log \pi_\star(y_{i,k} \mid c_{i,k})$ into $b_{\mathrm{tok}}(h) + \xi_{i,k}$. Both policies remain normalized distributions; the decomposition concerns errors on sampled tokens and does not posit an arbitrary additive shift to all log-probabilities in the vocabulary. Summing over the action mask gives

$$\Delta_i = N_i b_{\mathrm{tok}}(h) + u_i, \qquad \bar{\Delta} = \bar{N} b_{\mathrm{tok}}(h) + \bar{u}, \qquad \Delta_i - \bar{\Delta} = b_{\mathrm{tok}}(h)(N_i - \bar{N}) + u_i - \bar{u}, \tag{16}$$

which proves Equation 8. This also connects the token model to the aggregate model: for any fixed reference length $N_0(h)$, choose $b(h) = N_0(h) b_{\mathrm{tok}}(h)$ and $e_i = (N_i - N_0(h)) b_{\mathrm{tok}}(h) + u_i$. Proposition 1 then applies without assuming equal lengths, but its residual bound includes length variation.

If $|\xi_{i,k}| \le \varepsilon_{\mathrm{tok},h}$ on every scored token, then $|u_i| \le N_i \varepsilon_{\mathrm{tok},h}$ and $|\bar{u}| \le \bar{N} \varepsilon_{\mathrm{tok},h}$. Hence

$$|A^{\mathrm{step}}_i - A^\star_i| \le |b_{\mathrm{tok}}(h)|\, |N_i - \bar{N}| + (N_i + \bar{N}) \varepsilon_{\mathrm{tok},h}. \tag{17}$$

For equal lengths N, this reduces to $2N\varepsilon_{\mathrm{tok},h}$ independently of the shared token offset. An ideal advantage exceeding the applicable bound in magnitude retains its sign. The bound also makes clear that accumulated residual error can grow with action length under sum aggregation.

Aggregation creates one score per decision to enable these comparisons, but summation alone does not guarantee noise reduction. If token errors $\xi_{i,k}$ are treated as random variables with a fixed action mask, then

$$\mathrm{Var}\left(\sum_{k \in \mathcal{A}_i} \xi_{i,k} \right) = \sum_{k \in \mathcal{A}_i} \mathrm{Var}(\xi_{i,k}) + 2 \sum_{\substack{k < \ell \\ k, \ell \in \mathcal{A}_i}} \mathrm{Cov}(\xi_{i,k}, \xi_{i,\ell}). \tag{18}$$

Consequently, neither independent nor correlated token errors can be assumed to disappear through aggregation. Our cancellation result comes from subsequent group centering.

A.2 Proof of Proposition 1

Write $r = r^\star + b(h)\mathbf{1} + e$. The matrix $P = I_G - G^{-1}\mathbf{1}\mathbf{1}^\top$ satisfies $P\mathbf{1} = 0$, $P^\top = P$, and $P^2 = P$. Therefore,

$$A^{\mathrm{step}} = Pr = Pr^\star + Pe = A^\star + Pe. \tag{19}$$

Since P is an orthogonal projection, it cannot increase Euclidean norm. More explicitly,

$$\|Pe\|_2^2 = \sum_{i=1}^{G} e_i^2 - G\bar{e}^2 \le \sum_{i=1}^{G} e_i^2. \tag{20}$$

If $|e_i| \le \varepsilon_h$, then $|\bar{e}| \le \varepsilon_h$ and $|e_i - \bar{e}| \le 2\varepsilon_h$, proving the claimed sign guarantee. Likewise, pairwise score ordering is preserved whenever $|r_i^\star - r_j^\star| > 2\varepsilon_h$, since the error in that difference is $e_i - e_j$. These conclusions are independent of the magnitude of $b(h)$.

For a conditional stochastic interpretation, suppose $\mathbb{E}[e \mid h] = 0$ and $\mathrm{Cov}(e \mid h) = \Sigma_h$. With randomness taken over response groups and their residual errors,

$$\frac{1}{G}\mathbb{E}[\|A^{\mathrm{step}} - A^\star\|_2^2 \mid h] = \frac{1}{G}\,\mathrm{tr}(P\Sigma_h P) = \frac{1}{G}\,\mathrm{tr}(\Sigma_h) - \frac{1}{G^2}\mathbf{1}^\top \Sigma_h \mathbf{1}. \tag{21}$$

For conditionally uncorrelated residuals with variance σ_h^2, this becomes $(1 - G^{-1})\sigma_h^2$, whereas the uncentered score error has mean squared magnitude $b(h)^2 + \sigma_h^2$. This compares observed and ideal versions of each score representation, not the task performance of two algorithms. It shows that improvement can be substantial when shared offsets dominate, but that centering does not average away arbitrary response-dependent noise.

A.3 Effect on the Local Step-Advantage Update

Consider the first gradient evaluation at $\theta = \theta_{\mathrm{old}}$, before any clipping is active. Define the length-normalized response score gradient

$$z_i = \frac{1}{L_i}\sum_{k=1}^{L_i} \nabla_\theta \log \pi_\theta(y_{i,k} \mid c_{i,k})\Big|_{\theta=\theta_{\mathrm{old}}}, \qquad \|z_i\|_2 \le M_h. \tag{22}$$

Here M_h is a bound for the fixed group. With detached advantages, the ascent direction associated with $-\mathcal{L}_{\mathrm{step}}$ is $g = G^{-1}\sum_i A_i^{\mathrm{step}} z_i$. Let $g^\star$ use the ideal centered advantages on the same responses. Cauchy–Schwarz gives

$$\begin{aligned} \|g - g^\star\|_2 &= \left\| \frac{1}{G}\sum_{i=1}^{G} (Pe)_i z_i \right\|_2 \\ &\le \frac{1}{G}\|Pe\|_2 \left(\sum_{i=1}^{G} \|z_i\|_2^2 \right)^{1/2} \le \frac{M_h}{\sqrt{G}}\|Pe\|_2. \end{aligned} \tag{23}$$

Thus shared offsets do not perturb this local step update; residual scoring differences control its deviation from the ideal comparator. If $|e_i| \le \varepsilon_h$, projection contraction further implies $\|g - g^\star\|_2 \le M_h\varepsilon_h$. Under the uncorrelated residual model above and a bound M_h uniform over sampled groups, the conditional mean squared update error is at most $M_h^2(1 - G^{-1})\sigma_h^2$. This is a local sensitivity statement, not a global optimization result. We do not assume that the length-normalized, group-centered estimator is an unbiased policy gradient.

Shared-offset invariance also holds exactly for the clipped loss in Equation 11 at any fixed θ, because adding $b(h)$ does not change its advantages. For a bound on loss perturbations, let K bound both $\rho_{i,k}(\theta)$ and $\mathrm{clip}(\rho_{i,k}(\theta), 1 - \epsilon, 1 + \epsilon)$ on the sampled tokens. The function $\mathcal{C}_\epsilon(\rho, A)$ is K-Lipschitz in A, so the group loss obeys

$$|\ell_{\mathrm{step}}(r;\theta) - \ell_{\mathrm{step}}(r^\star;\theta)| \le \frac{K}{G}\sum_{i=1}^{G} |(Pe)_i| \le \frac{K}{\sqrt{G}}\|Pe\|_2, \tag{24}$$

where ℓ_{step} is the per-group summand defining $\mathcal{L}_{\mathrm{step}}$. This bound does not assert gradient smoothness at clipping boundaries.

A.4 Why Omit Standard-Deviation Normalization?

Let $r = b\mathbf{1} + \alpha v$ with $\alpha \ge 0$ and $Pv \ne 0$. Denote the population standard deviation across the G group entries by $s(v) = \|Pv\|_2/\sqrt{G}$. Centering and standardization give, respectively,

$$A = \alpha P v, \qquad A_{\mathrm{std}} = \frac{\alpha P v}{\alpha s(v) + \eta}, \qquad \frac{\|A_{\mathrm{std}}\|_2}{\|A\|_2} = \frac{1}{\alpha s(v) + \eta} \quad (\alpha > 0). \tag{25}$$

Without a stabilizer, $\|A_{\text{std}}\|_2 = \sqrt{G}$ for every $\alpha > 0$, even as the unnormalized signal tends to zero. With a fixed $\eta > 0$, the normalized signal also tends to zero eventually, but its gain relative to A can approach $1/\eta$. SAPD preserves the original scale of within-group differences, avoiding this potentially large gain. Both versions cancel shared additive offsets; the benefit of omitting normalization is magnitude control, not additional offset cancellation. This argument neither proves that small score spread implies uncertainty nor that normalization is universally harmful.

A.5 Scope and Empirical Implications

Three limitations are material. First, the common-offset assumption is at the aggregate action-score level. As Equation 8 shows, a shared token error $b_{\text{tok}}(h)$ leaves the residual $b_{\text{tok}}(h)(N_i - \bar{N})$ when action lengths vary. Thus, action-token summation does not automatically turn a shared token error into a shared sequence error. This provides an additional motivation for the action-only scoring mask: it excludes reasoning-length variation from the aggregate error. It does not guarantee lower error for every response group, since token biases and residuals can also differ between reasoning and action segments. The bound must still account for action-length variation; fixed-length or length-controlled comparisons can isolate it empirically.

Second, centering does not repair incorrect relative teacher preferences. For example, a teacher that reverses the ranking of two responses also reverses their centered advantages. Nor does centering alter $q_{\mathcal{B}}$: histories absent from the offline pool remain unavailable, and familiarity to either policy is not guaranteed. Offline replay can introduce an additional mismatch relative to online OPD, but need not be more out of distribution for every dataset or teacher.

Third, the guarantees apply to the step branch, not automatically to the combined objective. Token-level OPD remains sensitive to its own residual errors, and the weights in Equation 13 determine how the branches interact. Baseline terms can also cancel in expectation in conventional score-function estimators, so the projection analysis alone does not establish a universally less biased or lower-variance update than original token-level OPD.

The analysis motivates testing three separate predictions: exact invariance of the step branch to injected groupwise score offsets; sensitivity to response-dependent perturbations and teacher ranking errors; and the effect of standard-deviation normalization under weak within-group score differences. These controls should be complemented by comparisons across trajectory sources at fixed teacher and student checkpoints, with matched local rewards and action-length statistics. Such experiments are needed to establish whether the structured-error regime explains practical tolerance to unfamiliar prefixes.

B Experimental Details

This appendix describes the offline data, training configuration, and rule-based rewards used in our experiments. We build on the training framework of SDAR (Lu et al., 2026).

B.1 Offline Data Preparation

Trajectory sources. For ALFWorld and WebShop, we reuse the offline expert trajectories released with SPA-RL (Wang et al., 2025), which contain task instructions, agent responses, and environment observations. For WebShop, success filtering yields 909 training trajectories and 3,384 decision prefixes. Search-based QA combines the Search subset of SkillRL (Xia et al., 2026) and the AFM-MHQA (Li et al., 2025) dataset. Table 4 summarizes the offline pools before tokenizer-based length filtering. Each trajectory contributes one prefix per recorded decision.

Prefix construction. For each recorded decision, we assemble a prompt containing the task instruction, preceding action–observation pairs, and current observation. The target decision and subsequent observations are excluded from the prompt. We retain the trajectory identifier, decision index, and recorded action as metadata; the trajectory identifier and decision index jointly identify a prefix. Task-specific templates use `<think>` blocks for reasoning and `<action>` blocks for ALFWorld and WebShop decisions. Search responses use either a `<search>` block or an `<answer>` block, while retrieved passages appear in the recorded history within `<information>` blocks.

	Training pool		Offline validation pool	
Benchmark	Trajectories	Prefixes	Trajectories	Prefixes
ALFWorld	2,382	21,300	123	1,132
WebShop	909	3,384	45	166
Search-based QA	8,146	22,826	428	1,198

Table 4: Statistics of the offline trajectory pools. Offline validation pools are distinct from the environment episodes and QA evaluation sets used to measure task performance.

Data partitioning and batching. We use a nominal 5% offline validation split with random seed 0. Each training batch includes all retained decision prefixes from a sampled set of trajectories. Accordingly, the batch sizes in Table 5 refer to trajectories, while the numbers of decision contexts and generated responses depend on trajectory lengths. Candidate advantages are computed separately within each group sharing the same prefix. The processed offline datasets are provided in the supplementary materials.

Setting	ALFWorld	WebShop	Search-based QA
Student model	Qwen2.5-1.5B-Instruct		
Teacher model	Qwen2.5-3B-Instruct		
Teacher training	SFT or GRPO		
Teacher updates	Frozen throughout distillation		
Learning rate	10^{-6}	10^{-6}	10^{-6}
Trajectories per training batch	16	16	128
Responses per prefix	8	8	8
Maximum prompt tokens	8,192	8,192	8,192
Maximum response tokens	512	512	512
Training temperature	1.0	1.0	1.0
PPO mini-batch size	256	256	256
Micro-batch size per GPU	8	8	8
PPO epochs per sampled batch	1	1	1
Clipping threshold	0.2	0.2	0.2
Gradient clipping norm	1.0	1.0	1.0
Weight decay	0.01	0.01	0.01
Entropy coefficient	0	0	0
Default training updates	300	450	300
Evaluation decoding	Greedy	Greedy	Greedy
Evaluation interval (updates)	5	5	100
Maximum evaluation turns	50	15	4
Episodes/questions per evaluation	128	128	51,713
Retrieved passages per search	—	—	3

Table 5: Model settings and default training and evaluation configurations.

B.2 Models and Training Configuration

We use Qwen2.5-1.5B-Instruct as the student and frozen SFT- or GRPO-trained Qwen2.5-3B-Instruct models as teachers. The SFT teachers are trained on the corresponding offline datasets described in Appendix B.1, which are provided in the supplementary materials. We train the student on 8 NVIDIA H200 GPUs using the SDAR framework (Lu et al., 2026), with `vLLM` for generation. The student samples eight responses per offline prefix, and only generated response tokens contribute to the training loss. In Equation 13, we set $\lambda_{\mathrm{tok}} = 0.1$, $\lambda_{\mathrm{step}} = 0.5$, and $\lambda_{\mathrm{rule}} = 0.1$, with the rule coefficient following SDAR (Lu et al., 2026). Token-level feedback is additionally scaled by 0.15. Table 5 summarizes the default configurations. The efficiency analysis in Figure 2 reports training time only, excluding evaluation and checkpoint saving for both methods.

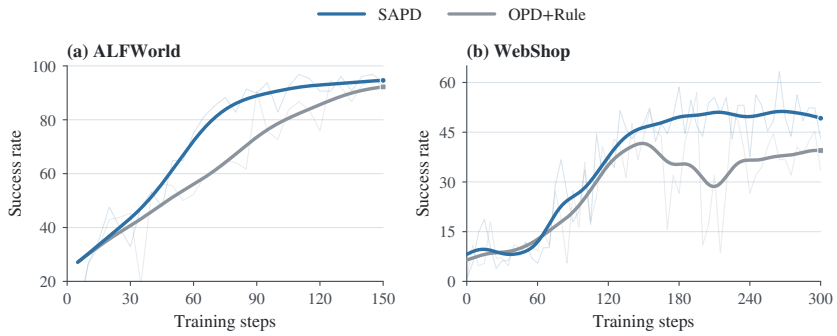


Figure 3: Evaluation success rates of SAPD and OPD+Rule during training with SFT teachers.

B.3 Rule-Based Rewards

To discourage malformed responses and repetitive actions, we use two locally computable rule checks. For format checking, we verify the required response tags and complete decision blocks: `<action>` blocks for ALFWorld and WebShop, and `<search>` or `<answer>` blocks for Search-based QA. Missing or incomplete required tags receive a negative format score. For repetition checking, we extract the generated action and assign a negative score if it exactly matches the immediately preceding action in the recorded prefix. Rule scores are aggregated and standardized within each prefix group following Equation 12. The resulting advantages supervise all generated response tokens through Equation 13, with $\lambda_{\text{rule}} = 0.1$ following SDAR (Lu et al., 2026). All checks use generated text and recorded histories without additional environment interaction.

C Additional Experiments

C.1 Learning Dynamics

To understand how SAPD's benefits develop over training, we compare its learning dynamics with OPD+Rule on ALFWorld and WebShop. Both methods train a Qwen2.5-1.5B-Instruct student using offline SFT prefixes and a frozen, SFT-trained Qwen2.5-3B-Instruct teacher. Figure 3 shows checkpoint evaluation results together with smoothed trends. On ALFWorld, SAPD improves more rapidly during training, while SAPD maintains a higher success-rate trend during later training on WebShop. These observations suggest that SAPD supports faster learning progress on ALFWorld and stronger sustained performance on WebShop, complementing the results in Table 1.

C.2 Distillation with Teacher-Generated Trajectories

Our main experiments focus on reusing existing offline trajectories, avoiding the need to collect teacher-specific rollouts through environment interaction. As a complementary analysis, we follow ReOPD (Liao et al., 2026) in reusing teacher-generated trajectories to construct decision prefixes, which are expected to better align with the teacher's interaction distribution and reduce teacher–prefix mismatch. Although obtaining these trajectories requires environment interaction, subsequent student training remains environment-free. We compare SAPD and ReOPD on ALFWorld and WebShop using a 1.5B student and a GRPO-trained 3B teacher. As shown in Table 6, SAPD achieves 76.6% on ALFWorld and 70.8% on WebShop, compared with 75.0% and 66.7% for ReOPD, yielding gains of 1.6 and 4.1 percentage points, respectively. These gains are smaller than those on SFT prefixes under the GRPO teacher in Table 1. These results suggest that SAPD remains effective on teacher-generated prefixes and may offer larger gains under stronger teacher–prefix mismatch.

C.3 Distillation with a 3B Student

Our main experiments focus on 1.5B students; we further investigate whether step-level guidance remains beneficial for a larger student under teacher–prefix mismatch. We evaluate a Qwen2.5-3B-Instruct student on ALFWorld using a GRPO-trained Qwen2.5-3B-Instruct teacher and decision

	ALFWorld							WebShop
Method	Pick	Look	Clean	Heat	Cool	Pick2	**Avg.**	**Success**
ReOPD	92.2	**61.1**	**93.0**	88.0	72.9	**25.4**	75.0	66.7
SAPD	**94.2**	60.0	85.5	**94.0**	**74.7**	20.4	**76.6**	**70.8**

Table 6: Performance between SAPD and ReOPD. Avg. reports overall success across all evaluation episodes, giving greater weight to categories with more episodes. Best results are bold.

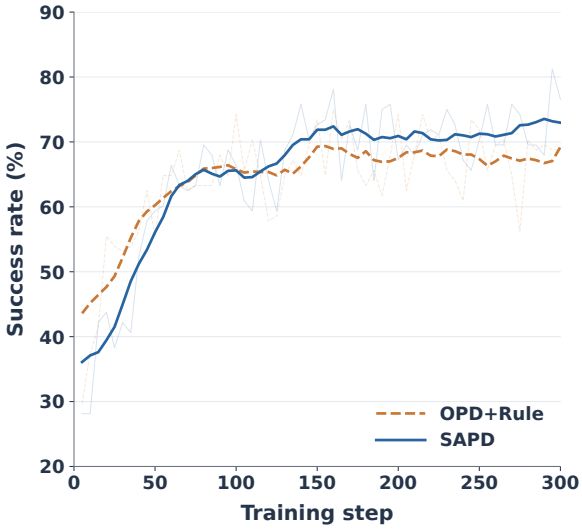


Figure 4: ALFWorld evaluation performance with a 3B student and a GRPO-trained 3B teacher using offline SFT prefixes.

prefixes extracted from offline SFT trajectories. Figure 4 compares OPD+Rule with SAPD, which computes relative advantages among candidate actions sharing the same prefix. Within 300 training steps, SAPD achieves the performance of 78.6%, compared with 74.5% for OPD+Rule. These results show that SAPD improves the performance of a 3B student under prefix mismatch, supporting its effectiveness in this setting.

	SFT teacher			GRPO teacher		
Scoring scope	ALFWorld	WebShop	Search	ALFWorld	WebShop	Search
Action-only	**96.6**	**59.4**	**38.8**	**69.3**	**53.4**	41.3
Full-token	95.9	50.0	38.6	60.8	52.3	**41.6**

Table 7: Comparison of action-only and full-token scoring for step advantages. Both variants use offline SFT prefixes and update the complete response. Results are reported in percentages, with the better result in each setting shown in bold.

D Ablation on Step-Advantage Scoring

A generated response contains both reasoning text and an executable action, raising the question of whether step advantages should be computed from action tokens alone or from all response tokens. We compare these two scoring scopes using offline SFT trajectories under both SFT and GRPO teachers. Both variants apply policy updates to the complete response; only the tokens used to compute step scores differ. As shown in Table 7, action-only scoring performs better on ALFWorld and WebShop under both teacher configurations. It improves ALFWorld success from 60.8% to 69.3% with the GRPO teacher and WebShop success from 50.0% to 59.4% with the SFT teacher.

On Search-based QA, the differences are small. These results support action-focused scoring and suggest that excluding thinking tokens may provide more reliable step-level guidance when prefix mismatch makes teacher supervision less reliable.