SELECT BEFORE YOU ACT: PRE-COMMITMENT GUIDANCE FOR AGENTIC ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation(OPD) transfers the capabilities of language model agents through teacher supervision on student-generated interactions. In multi-turn tasks, however, student action errors can alter the environment and undermine the conditions for effective supervision at later turns. We introduce PREACT, a pre-commitment teacher-guided approach to multi-turn on-policy distillation. At each turn the student proposes several candidate responses, and a fixed teacher selects one for execution *before* its action reaches the environment Only the selected response is retained for distillation and appended to the interaction history with the resulting observation. Selecting a response before execution shapes not only the current supervision target but also the states in which subsequent learning occurs. PREACT thus extends teacher guidance from supervising individual responses to shaping the interactions that produce future training data, while keeping the student as the sole source of responses. Experiments on ALFWorld, ScienceWorld, and WebShop demonstrate that PREACT can improve student task performance across multiple model scales and under different initialization settings. These results highlight action commitment as an effective point of teacher guidance for training multi-turn agents. Our code is available at this link.

1 INTRODUCTION

Advances in the reasoning (Wei et al., 2022) and tool-use capabilities (Schick et al., 2023) of large language models (LLMs) have enabled agents to solve complex, multi-step tasks through sustained interaction with external environments (Yao et al., 2023b; Zeng et al., 2024). Distilling these capabilities from powerful teacher agents into smaller student models is important for deployment under cost and latency constraints. On-policy distillation (OPD) provides an effective method (Rusu et al., 2015) for this transfer by applying token-level supervision from a stronger teacher to responses (Ma et al., 2026) generated during the student's own interactions with the environment (Li et al., 2026b). Training on student-generated interactions mitigates the distribution mismatch of supervised fine-tuning on fixed demonstrations (Ross et al., 2011) while providing finer-grained guidance than the sparse, delayed outcome rewards used in reinforcement learning (Lightman et al., 2024).

However, sustaining effective teacher guidance along student-generated trajectories becomes challenging in multi-turn agentic tasks (Zhou et al., 2024; Agarwal et al., 2024), where student errors propagate through both the interaction history and the external environment. Executing the action produces observations that, together with the response, form the history for subsequent generation and supervision (Yao et al., 2023a). An early action error can therefore redirect the interaction and change the contexts on which later teacher guidance is conditioned, as illustrated in Figure 1. Irreversible interactions further restrict recovery and can eliminate viable paths to task completion, even when the teacher recognizes the mistake. An early error can therefore compromise the conditions for effective supervision across many subsequent turns (Ross et al., 2025). Agentic OPD thus faces a coupled challenge: the student determines the trajectory on which it learns, while its own action errors can progressively undermine the teacher guidance that learning relies on.

Existing approaches mitigate these challenges through trajectory curricula, selective supervision or the allocation of teacher supervision. Trajectory curricula progressively increase student rollout depth, limiting early exposure to long-horizon errors (Wang et al., 2026). Selective supervision adjusts the contribution of individual turns according to the need for and reliability of teacher feed-

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.

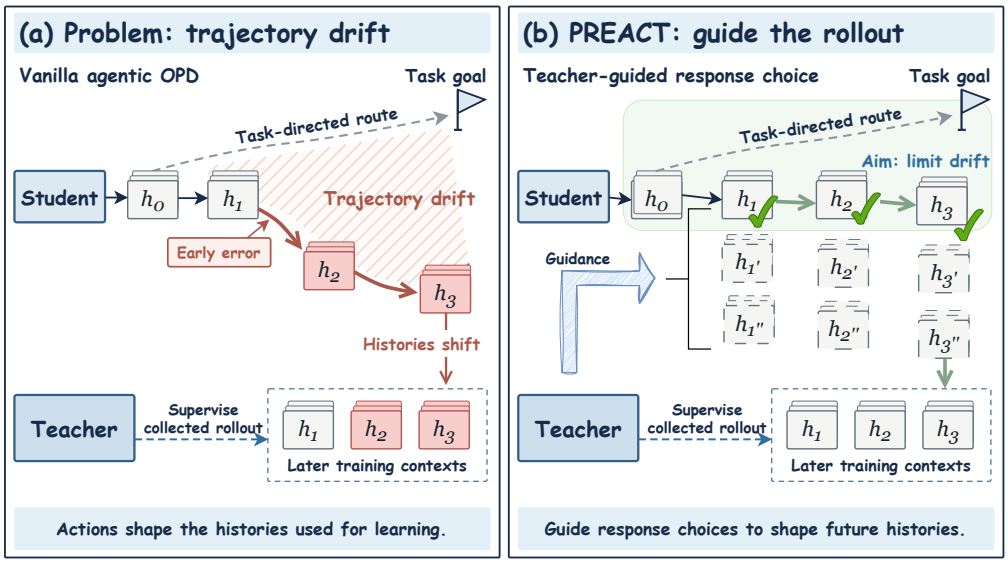


Figure 1: Motivation for PREACT. In multi-turn agentic OPD, action errors can alter the environment and constrain subsequent progress and supervision. Selecting among student-generated responses before execution influences both the current distillation sample and the interaction history on which later teacher guidance depends.

back (Zhou et al., 2026). Guided-OPD (Li et al., 2026a) mixes teacher- and student-generated turns with a decaying intervention probability, whereas ReOPD trains student responses on replayed teacher prefixes (Liao et al., 2026). These strategies regulate how far the student rolls out, which turns contribute to learning, or when the teacher takes over. However, adjusting rollout depth and supervision weights provides only indirect control over the student's environment-facing decisions, while takeover relies on teacher-generated continuations. A complementary challenge is therefore to guide the student's own action choices before execution, when the course of the interaction can still be influenced before environmental consequences constrain subsequent learning.

To address this challenge, we introduce PREACT, a pre-commitment teacher-guided rollout strategy that selects among student-generated turns before their actions are committed to the environment. Our central insight is that selecting a response determines both the current distillation sample and the interaction context on which subsequent supervision is conditioned. At each turn, we select the student-generated candidate with the lowest teacher perplexity (PPL) for execution and distillation, favoring responses that better align with the teacher's distribution at the current history. Each selected action then shapes the history from which subsequent responses are generated, extending the influence of selection beyond the current training sample. Repeated throughout the rollout, this process is designed to limit accumulated trajectory drift and preserve opportunities for effective supervision over long horizons. Across ALFWorld (Shridhar et al., 2021), ScienceWorld (Wang et al., 2022), and WebShop (Yao et al., 2022), our method consistently improves overall task performance over baselines. Additional experiments demonstrate gains across student–teacher model sizes and initialization settings, while ablations examine how the selection criterion and candidate pool size affect distillation performance.

In summary, our contributions are threefold: ✦ **Problem Formulation.** We present a proposal–commit perspective on multi-turn agentic OPD, separating student response generation from action commitment. This perspective clarifies how pre-execution selection determines the current distillation sample and shapes the interaction histories on which subsequent learning depends. ✦ **Proposed Method.** We propose PREACT, a teacher-reranking OPD method that selects student-generated candidate with the lowest teacher perplexity for both execution and distillation. This design guides trajectory construction before environment transitions while retaining the student as the sole source of candidate responses. ✦ **Empirical Evaluation.** Experiments on three benchmarks across different teacher–student pairs show performance gains over vanilla OPD, with additional comparisons against a temporal-curriculum baseline. Ablations compare pre-execution selection with trajectory-level filtering and examine the selection criterion, supervision scope, and candidate budget.

2 Related Work

On-Policy Distillation. On-policy distillation (OPD) addresses the mismatch between expert-provided training sequences and a student's own generations by applying teacher supervision to student-generated data (Agarwal et al., 2024; Chang et al., 2015). This principle of supervising also underlies DAgger, which obtains expert labels at learner-visited states (Ross et al., 2011). Beyond data collection, distillation methods differ in the objective used to align the two models. MiniLLM adopts reverse KL (Gu et al., 2024), while DistiLLM combines skew-KL objectives with adaptive reuse of student-generated outputs (Ko et al., 2024). Recent analyses further characterize when on-policy supervision helps and which mechanisms drive the gain (Li et al., 2026b). Multi-turn interaction introduces (Zhang et al., 2026a) a further dependence: executed actions determine the observations on which subsequent teacher supervision is conditioned.

Agent Learning and Multi-Turn Distillation. Agent learning methods differ in how they construct and use interaction trajectories (Jang et al., 2026). Supervised approaches organize training around collected demonstrations: AgentTuning combines interaction trajectories with general instruction data (Zeng et al., 2024), while ATLAS selects critical steps from expert trajectories for training (Chen et al., 2025). For multi-turn OPD, both the construction of training histories and the allocation of teacher supervision become central design questions (Zhong et al., 2026). Recent work addresses trajectory construction through curricula and teacher participation. TCOD regulates the student-controlled horizon through temporal curricula (Wang et al., 2026). Teacher participation also shapes training histories through mixed teacher-student turns or replayed teacher prefixes (Li et al., 2026a; Liao et al., 2026). SAGE-OPD addresses supervision allocation by weighting the distillation loss using environment feedback, teacher judgments, and predictive confidence (Zhou et al., 2026). Our work focuses on teacher guidance before action execution, when choosing a student-generated response determines both the current distillation sample and the subsequent interaction history.

3 Method

3.1 Interactive Distillation Setup

Let π_θ denote the student and π_T a fixed teacher, with both defining next-token distributions over a shared vocabulary. For a task input x, the interaction history at turn t is

$$h_t = (x, o_0, y_1, o_1, \ldots, y_{t-1}, o_{t-1}), \tag{1}$$

where o_0 is the initial observation and y_t is a complete assistant turn, including its reasoning and environment-facing action. An action parser g extracts $a_t = g(y_t)$. Executing this action produces an observation

$$o_t \sim \mathcal{E}(\cdot \mid h_t, a_t), \qquad h_{t+1} = (h_t, y_t, o_t). \tag{2}$$

Here, $\mathcal{E}$ denotes the environment's feedback distribution. The full assistant turn and the returned observation remain in the history used for subsequent generation. A trajectory terminates when the environment reaches a terminal state or the horizon H is reached.

On-policy distillation uses teacher supervision on student-generated responses (Agarwal et al., 2024). In its single-sample agentic form, one student response is generated and executed at each turn. PREACT separates response generation from execution by inserting teacher-guided selection between them. Training alternates between collecting complete trajectory batches and updating the student on the selected responses. We write $p_\theta(\cdot \mid h)$ for the student's complete-response sampling distribution under the decoding configuration used to generate candidates.

3.2 Select Before Execution

Student proposals. At history h_t, we independently sample $K \geq 1$ complete candidate turns from the same student distribution, $y_t^{(1)}, \ldots, y_t^{(K)} \overset{\text{i.i.d.}}{\sim} p_\theta(\cdot \mid h_t)$. Each candidate contains its own reasoning and proposed action. All candidates share the same task instruction, previous turns, and environment observations, so they represent alternative continuations at the same decision point. Their actions remain unexecuted during generation and selection. The complete turn is selected as a unit, preserving the reasoning that precedes the proposed action.

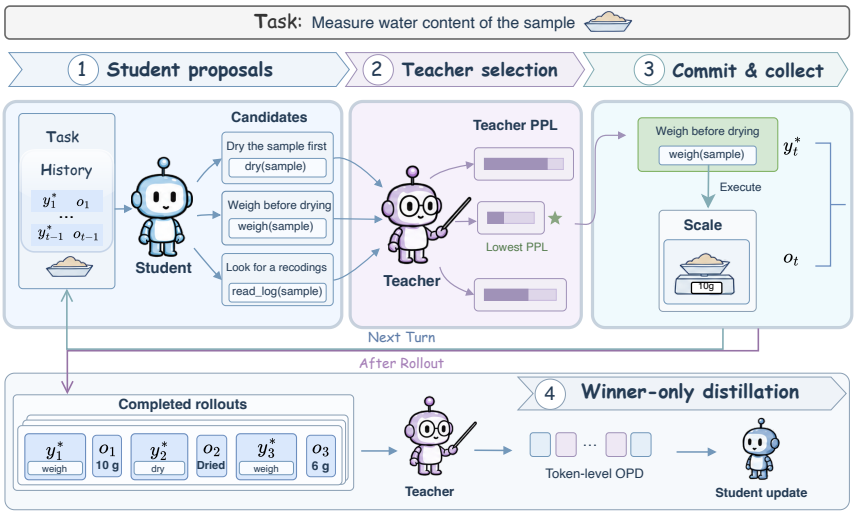


Figure 2: Overview of PREACT. At each turn, the student proposes multiple responses, and the teacher ranks them by perplexity before action execution. Only the action extracted from the selected response is executed in the environment. And the full response and resulting observation extend the interaction history.

Teacher scoring. The teacher assigns each candidate a length-normalized log-likelihood,

$$s_T(y \mid h) = \frac{1}{|y|} \sum_{i=1}^{|y|} \log \pi_T(y_i \mid h, y_{<i}). \tag{3}$$

Here, $|y|$ is the candidate's response-token count, and $y_{<i}$ is its prefix before token i. The teacher evaluates each token given the shared history and that candidate's own preceding tokens. The scored span includes both reasoning and action tokens. The history supplies conditioning context and is excluded from the average. Length normalization expresses the score on a per-token basis across candidates with different response lengths. The corresponding teacher perplexity is

$$\mathrm{PPL}_T(y \mid h) = \exp[-s_T(y \mid h)] = \exp\left[-\frac{1}{|y|} \sum_{i=1}^{|y|} \log \pi_T(y_i \mid h, y_{<i})\right]. \tag{4}$$

Lower perplexity corresponds to higher average teacher log-likelihood for the candidate turn. All candidate scores are available before an environment transition occurs.

Winner commitment. We select the candidate with the lowest teacher perplexity,

$$k_t^\star = \arg\min_{k \in \{1,\dots,K\}} \mathrm{PPL}_T(y_t^{(k)} \mid h_t), \qquad y_t^\star = y_t^{(k_t^\star)}, \tag{5}$$

and execute $a_t^\star = g(y_t^\star)$. Tied scores are resolved by the lowest candidate index. The resulting observation is appended together with the full selected turn,

$$o_t \sim \mathcal{E}(\cdot \mid h_t, a_t^\star), \qquad h_{t+1} = (h_t, y_t^\star, o_t). \tag{6}$$

We repeat this procedure at the updated history. Each turn commits one selected action, while additional candidates increase response-generation and scoring work. The teacher contributes likelihood evaluations, and the executed response is one of the student's proposals.

Selection therefore determines both the response retained at the current turn and, through action execution and feedback, the history from which subsequent responses are generated. We store each selected turn together with its teacher token log-probabilities. These values are reused for distillation under the same recorded history and response prefixes.

3.3 WINNER-ONLY DISTILLATION

Let $\tau = (x, o_0, y_1^\star, o_1, \ldots, y_{T(\tau)}^\star, o_{T(\tau)})$ denote a trajectory of committed turns and observations, where $T(\tau)$ is its number of turns. We collect a batch $\mathcal{D}$ of these trajectories before each student update and apply the baseline token-level OPD rule to the selected responses. Each trajectory is represented as one training sequence. Supervision covers the selected responses, including reasoning and action tokens, while task prompts and environment observations provide conditioning context. Unselected candidates are excluded from both the executed history and the update data.

The distillation term. For a selected token $y_{t,i}^\star$, let $c_{t,i} = (h_t, y_{t,<i}^\star)$ denote its recorded context. The student-to-teacher log-ratio on this token is

$$\Delta_{t,i} = \log \pi_\theta(y_{t,i}^\star \mid c_{t,i}) - \log \pi_T(y_{t,i}^\star \mid c_{t,i}). \tag{7}$$

We evaluate this quantity before optimizing the collected batch and use its negative as the token-level distillation weight,

$$w_{t,i} = -\beta \, \mathrm{sg}[\Delta_{t,i}], \tag{8}$$

where $\beta \geq 0$ controls the distillation strength and sg denotes stop-gradient. These weights remain fixed during the batch update. For $\beta > 0$, a token receives a positive weight when the teacher assigns it a higher probability than the student, and a negative weight when the ordering is reversed. In the policy-gradient update, positive weights encourage higher probabilities for those tokens, while negative weights encourage lower probabilities. Candidate selection uses an average teacher score over a complete turn, while distillation provides separate feedback for its individual tokens.

During optimization, the current-to-pre-update student probability ratio is

$$\rho_{t,i}(\theta) = \frac{\pi_\theta(y_{t,i}^\star \mid c_{t,i})}{\pi_{\theta_{\mathrm{old}}}(y_{t,i}^\star \mid c_{t,i})}. \tag{9}$$

We retain the baseline's clipped policy-gradient loss,

$$\ell_{t,i}^{\mathrm{OPD}}(\theta) = -\min\{\rho_{t,i}(\theta) w_{t,i}, \mathrm{clip}(\rho_{t,i}(\theta), 1-\epsilon, 1+\epsilon)\, w_{t,i}\}, \tag{10}$$

where $\epsilon > 0$ is the clipping threshold. Positive weights encourage higher probabilities for the corresponding tokens, while negative weights encourage lower probabilities. The recorded contexts, pre-update probabilities, and distillation weights remain fixed as θ is optimized.

Let $m_{t,i} \in \{0, 1\}$ mark supervised response tokens and let $N_\tau = \sum_{t=1}^{T(\tau)} \sum_{i=1}^{|y_t^\star|} m_{t,i} > 0$ be their total number in a training trajectory. The training objective on the collected batch is

$$\mathcal{L}_{\mathrm{PREACT}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \frac{1}{N_\tau} \sum_{t=1}^{T(\tau)} \sum_{i=1}^{|y_t^\star|} m_{t,i}\, \ell_{t,i}^{\mathrm{OPD}}(\theta). \tag{11}$$

Token losses are averaged within each trajectory, and trajectories receive equal weight in the batch objective. The denominator includes supervised tokens from all retained turns. This objective applies the baseline distillation rule to the responses selected during interaction.

Relation to on-policy collection. For a given student, let $b_K(\cdot \mid h)$ denote the conditional distribution of the winner in equation 5, and let P_{b_K} denote its induced trajectory distribution. The collected trajectories follow P_{b_K}, with response probabilities determined jointly by student sampling and teacher selection. For $K > 1$, this generally differs from the trajectory distribution P_{p_θ} induced by direct student sampling. Every executed response remains within the support of p_θ because it is selected from the student's proposals. Section D characterizes the distributional effect of selection as a function of K, and Section 4.4 examines candidate budgets empirically.

After each batch update, we collect new rollouts using the updated student as the proposal model. Setting $K = 1$ recovers single-sample OPD collection and updating under the same baseline settings. At deployment, the trained student generates one response per turn and interacts directly with the environment. Algorithm 1 in Appendix A summarizes the full procedure.

4 EXPERIMENTS

We use Qwen3-32B and Qwen3-8B teachers, each paired with Qwen3-4B and Qwen3-1.7B students (Yang et al., 2025), and a Qwen3.5-27B teacher paired with a Qwen3.5-4B student (Team, 2026). Evaluation covers ALFWorld (Shridhar et al., 2021), ScienceWorld (Wang et al., 2022), and WebShop (Yao et al., 2022). Reported metrics include overall, seen, and unseen success rates (SR) on ALFWorld, SR and mean task score on ScienceWorld, and SR on WebShop. Further experimental details are provided in Appendix B.

4.1 MAIN RESULTS

We report the main results on ALFWorld (Shridhar et al., 2021), ScienceWorld (Wang et al., 2022), and WebShop (Yao et al., 2022). Tables 1–3 and 8 summarize success rates and task scores across model scales, initialization settings, and model generations.

Table 1: Main results with Qwen3-32B as the teacher and Qwen3-1.7B and Qwen3-4B as students. Each block reports standalone evaluations of the starting models followed by student-only results after distillation. Entries with ± summarize three evaluation runs. All metrics use a scale from 0 to 100. Bold identifies the best distillation result within each block, including ties.

	ALFWorld SR (%) ↑			ScienceWorld ↑		WebShop ↑
Method	Overall	Seen	Unseen	SR (%)	Score	SR (%)
	Student: Qwen3-1.7B		*Teacher: Qwen3-32B*			
Initial student	7.66	10.00	5.22	4.70	8.85	9.00
Initial teacher	66.79	63.57	70.15	36.31	59.67	12.00
Vanilla OPD	39.66 (± 2.01)	38.10 (± 2.18)	41.29 (± 2.62)	17.07 (± 1.04)	50.29 (± 0.45)	8.67 (± 1.53)
TCOD-F2B	38.32	39.29	37.31	15.44	48.59	9.00
PREACT (ours)	**45.86 (± 0.84)**	**44.29 (± 0.71)**	**47.51 (± 1.14)**	**18.60 (± 1.15)**	**51.53 (± 0.55)**	**11.33 (± 0.58)**
	Student: Qwen3-4B		*Teacher: Qwen3-32B*			
Initial student	33.21	35.00	31.34	21.64	41.16	10.00
Initial teacher	66.79	63.57	70.15	36.31	59.67	12.00
Vanilla OPD	59.12 (± 1.32)	58.57 (± 1.89)	59.70 (± 1.24)	33.10 (± 1.18)	58.70 (± 0.46)	15.33 (± 1.15)
TCOD-F2B	60.22	59.28	**61.19**	33.63	58.77	11.00
PREACT (ours)	**62.77 (± 0.73)**	**64.76 (± 1.80)**	60.70 (± 1.55)	**37.51 (± 1.16)**	**61.35 (± 0.62)**	**16.33 (± 1.15)**

PREACT improves student performance across all three benchmarks. As shown in Table 1, we first compare PREACT with Vanilla OPD and TCOD-F2B (Wang et al., 2026) using a fixed Qwen3-32B teacher and Qwen3-1.7B and Qwen3-4B students. For both student sizes, PREACT achieves the highest ALFWorld overall success rate, ScienceWorld mean task score, and WebShop success rate among the distillation methods. Relative to Vanilla OPD, the 1.7B and 4B students gain 6.20 and 3.65 percentage points in ALFWorld overall success, respectively. The corresponding ScienceWorld score gains are 1.24 and 2.65 points. The gains for the 1.7B student show that PREACT remains beneficial across a large teacher–student capacity gap.

PREACT improves out-of-distribution (OOD) generalization. Relative to Vanilla OPD, PREACT raises unseen ALFWorld success by 6.22 and 1.00 percentage points for the 1.7B and 4B students, respectively. The gains across both student scales suggest better OOD generalization rather than merely fitting the environment configurations encountered during training. By evaluating teacher compatibility before execution, PREACT favors context-appropriate actions and limits error propagation across later turns. This turn-level guidance supports transfer across layouts and object configurations, which helps explain the gains on unseen environments.

Table 2: WebShop SR (%) with a Qwen3-8B teacher. Entries with ± summarize three evaluation runs; bold identifies the best distillation result for each student, including ties.

	Qwen3-1.7B	Qwen3-4B
Method	SR (%)	SR (%)
Initial student	9.00	10.00
Initial teacher	13.00	13.00
Vanilla OPD	9.67(± 3.06)	11.67(± 1.53)
TCOD-F2B	7.00	11.00
PREACT (ours)	**11.33(± 2.08)**	**12.33(± 1.53)**

Table 3: ALFWorld success rates (%) with task-adapted initialization. The teacher is Qwen3-8B after SFT and GRPO, and the student is Qwen3-1.7B after SFT. Bold identifies the best distillation result in each column.

Model	Overall ↑	Seen ↑	Unseen ↑
Initial student (SFT)	8.76	10.71	6.72
Initial teacher (SFT+GRPO)	94.89	97.14	92.54
Vanilla OPD	70.07	72.14	67.91
PREACT (ours)	**74.09**	**73.57**	**74.63**

PREACT improves performance with a smaller teacher. To examine whether the gains generalize across different teacher scales and teacher–student scale gaps, we replace Qwen3-32B with Qwen3-8B and evaluate both student sizes on WebShop. As illustrated in Table 2, relative to Vanilla OPD, PREACT improves success rates by 1.66 and 0.66 percentage points for the 1.7B and 4B students, respectively. PREACT also outperforms TCOD-F2B in success rate for both student sizes.

PREACT provides additional gains after task adaptation. In practical settings, both teacher and student models are often adapted to the target task before distillation. We therefore evaluate PREACT on ALFWorld using task-adapted checkpoints for both models in Table 3. Specifically, the Qwen3-8B teacher is trained with supervised fine-tuning (SFT) followed by group relative policy optimization (GRPO), achieving 94.89% overall success, while the Qwen3-1.7B student is initialized from an SFT checkpoint with 8.76% overall success. Under this setting, PREACT outperforms Vanilla OPD in overall and unseen success rates.

4.2 Selection and Supervision Ablations

Together, these ablations show that PREACT benefits from coupling teacher-guided selection with trajectory construction and supervision under a matched compute budget.

Teacher-guided selection drives the gains under matched compute. Random-Commit matches the compute of PREACT by using the same candidate budget at each turn, but commits a uniformly sampled response instead of the teacher-preferred one. It achieves 41.61% overall success, close to Vanilla OPD at 41.97%, whereas PREACT reaches 46.35%. The 4.74-point improvement over Random-Commit shows that the additional candidate budget becomes effective when coupled with teacher-guided selection.

Teacher preference provides a more effective selection signal than student preference. Student-PPL selects the candidate with the lowest perplexity under the student itself and achieves 40.51% overall success. Using teacher perplexity, PREACT improves this result by 5.84 percentage points. The difference is larger on unseen environments, where PREACT exceeds Student-PPL by 8.21 points, compared with 3.57 points on the seen split. This result shows that teacher perplexity more effectively distinguishes among student-generated responses.

Turn-level selection is more effective than trajectory-level filtering. Trajectory-PPL uses teacher perplexity to rank completed trajectories and retains the selected trajectory for distillation. In contrast, PREACT applies teacher-guided selection before each action is executed, so the selected response also determines the subsequent environment state and interaction history. Table 4 shows the benefit of incorporating teacher guidance during trajectory construction rather than only after rollout completion.

Aligning supervision with executed responses improves distillation. All-scored follows the same teacher-guided execution rule as PREACT, but supervises all sampled candidates and trails it by 1.99 percentage points in overall success. Although unselected candidates are excluded from execution, their reasoning and action tokens still contribute to distillation updates. Teacher selection therefore shapes the interaction history, while the responses used for supervision remain unfiltered.

Table 4: Selection and supervision ablations on ALFWorld. OPD targets specify which responses or trajectories receive distillation supervision. All results report success rates (%) of the distilled student acting independently. Bold identifies the highest value in each column.

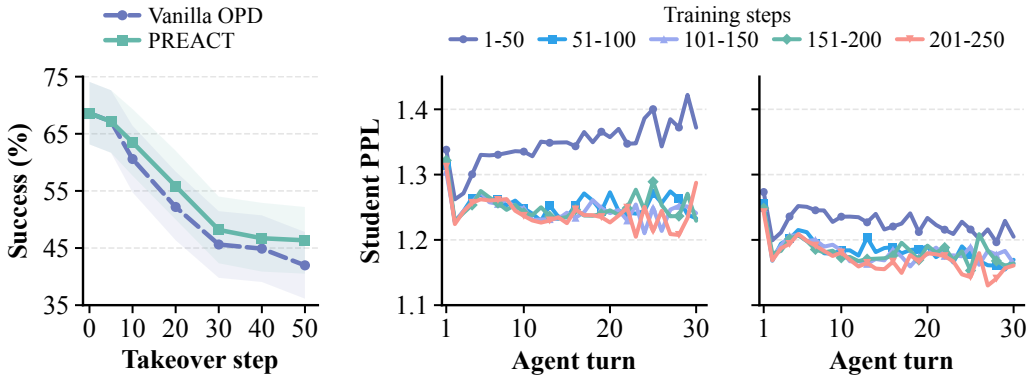
Variant	OPD targets	Overall ↑	Seen ↑	Unseen ↑
	Student: Qwen3-1.7B *Teacher: Qwen3-32B*			
Vanilla OPD	Sampled responses	41.97	40.00	44.03
Random-Commit	Executed responses	41.61	38.57	44.78
Student-PPL	Executed responses	40.51	41.43	39.55
Trajectory-PPL	Selected trajectory	42.70	42.14	43.28
All-scored	All candidates	44.36	41.83	47.01
PREACT (ours)	Executed responses	**46.35**	**45.00**	**47.76**

The performance gap suggests that including these less-preferred alternatives can dilute the benefit of response selection. By distilling only executed responses, PREACT concentrates learning on the teacher-preferred responses that generate the training trajectory.

4.3 Teacher Continuation and Student Likelihood

To examine how student-generated histories affect subsequent teacher intervention, we let trained students act independently for t turns on ALFWorld before a shared teacher continues from the resulting state and history to termination. As shown in Figure 3(a), final success declines with later handoffs for both methods, indicating that the teacher's ability to complete a task is constrained by the preceding student interactions. This motivates teacher guidance during rollout construction, when response selection can influence the histories on which subsequent supervision is conditioned.

Although PREACT ranks candidates by teacher perplexity, its selected responses also have lower pre-update student perplexity than Vanilla OPD samples across training windows and interaction turns on ScienceWorld in Figure 3(b,c). The selected responses are therefore preferred by the teacher while remaining likely under the current student. This suggests that teacher guidance stays close to the student's current response distribution.



(a) Teacher continuation (b) Vanilla OPD (c) PREACT

Figure 3: Teacher continuation and student likelihood. (a) A shared Qwen3-32B teacher continues trajectories from trained Qwen3-1.7B students on 274 ALFWorld tasks. (b,c) Student perplexity on ScienceWorld with Qwen3.5-4B students and a Qwen3.5-27B teacher.

4.4 Effect of Candidate Budget

We vary the number of student candidates per turn, $K \in \{1, 2, 4, 8\}$, on ALFWorld with a Qwen3-1.7B student and Qwen3-8B and Qwen3-32B teachers. The setting $K = 1$ corresponds to Vanilla OPD, and $K = 4$ is our default.

Moderate candidate budgets yield the strongest results in both teacher settings. Table 5 shows that moderate candidate budgets perform best, with overall success peaking at $K = 4$ for the 8B teacher and $K = 2$ for the 32B teacher, while further increasing the budget to $K = 8$ leads to lower performance. The preferred budget also varies across evaluation splits. With the 32B teacher, $K = 2$ achieves the highest unseen success rate, whereas $K = 4$ performs best on seen tasks. Our default $K = 4$ improves both seen and unseen success rates over Vanilla OPD with both teachers.

Table 5: ALFWorld success rates (%) across candidate budgets. Bold identifies the best result for each teacher and metric.

Student: Qwen3-1.7B						
	Teacher: Qwen3-8B			*Teacher: Qwen3-32B*		
Candidate budget	Overall	Seen	Unseen	Overall	Seen	Unseen
$K = 1$ (Vanilla OPD)	34.31	34.29	34.33	41.97	40.00	44.03
$K = 2$	34.67	**37.14**	32.09	**47.08**	40.71	**53.73**
$K = 4$ (default)	**35.77**	36.43	**35.07**	46.35	**45.00**	47.76
$K = 8$	30.29	30.00	30.60	43.07	40.00	46.27

5 Discussion.

PREACT shapes both the responses used for distillation and subsequent interaction histories. At a fixed history h, we hold the student sampling distribution $p_\theta(\cdot \mid h)$ and teacher scoring rule fixed. Let $Y^\star$ be the response with the lowest teacher perplexity among K i.i.d. samples from $p_\theta(\cdot \mid h)$. Let $R_h(u)$ denote the probability that a single candidate's teacher perplexity exceeds u. The selected response satisfies

$$\Pr[\mathrm{PPL}_T(Y^\star \mid h) > u] = R_h(u)^K. \tag{12}$$

Since $R_h(u) \in [0, 1]$, this upper-tail probability is non-increasing in K.

The task quality of the selected response depends on both candidate coverage and ranking reliability. We define acceptable responses using a fixed task-quality criterion specified independently of teacher perplexity. Let $p(h) \in (0, 1)$ denote the probability that a candidate is acceptable. Let $\eta_K(h)$ denote the probability of selecting an unacceptable response given that the pool contains at least one acceptable candidate. The probability of selecting an unacceptable response is

$$\epsilon_K(h) = \underbrace{(1 - p(h))^K}_{\text{coverage failure}} + \underbrace{\eta_K(h)\left[1 - (1 - p(h))^K\right]}_{\text{error despite coverage}}. \tag{13}$$

Increasing K reduces the probability of coverage failure. The total selection error increases exactly when the second term grows by more than the first term decreases. In our ALFWorld experiments, final student success at $K = 8$ is lower than at the best tested budget for each teacher. Appendix D provides further analysis of this pattern.

6 Conclusion

We have presented PREACT, a pre-commitment teacher-guided rollout strategy for multi-turn on-policy distillation. Selecting the student proposal with the lowest teacher perplexity *before* action execution determines both the response used for distillation and the history from which subsequent training samples are generated. A controlled comparison that varies only the timing of selection shows that the second of these channels carries much of the benefit, distinguishing pre-commitment guidance from best-of-K filtering of training data. Experiments across different teacher–student pairs on ALFWorld, ScienceWorld, and WebShop show improvements over vanilla OPD and a temporal curriculum, retained when the trained student acts independently and without teacher generation at deployment. These findings identify action commitment as an effective point for teacher guidance in agentic distillation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and writing, including improving grammar, clarity, phrasing, and readability of the manuscript. The use of language models for training rollout generation and teacher scoring is described in the Method section. Generative AI tools were not used to develop the research ideas or hypotheses, design the methodology or experiments, or conduct the experiments. All AI-assisted text was reviewed, verified, and revised by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript and all associated artifacts.

REPRODUCIBILITY STATEMENT

We describe PREACT's candidate generation, teacher scoring, action commitment, and distillation procedure in Section 3, with pseudocode provided in Appendix A. And Appendix B eports the model configurations, benchmark splits, training and evaluation settings, environment-specific prompts, and interaction protocols. Our code is available at this link.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander Nicholas D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Forty-second International Conference on Machine Learning*, 2025.

Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, and John Langford. Learning to search better than your teacher. In *International Conference on Machine Learning*, pp. 2058–2066. PMLR, 2015.

Zhixun Chen, Ming Li, Yuxuan Huang, Yali Du, Meng Fang, and Tianyi Zhou. Atlas: Agent tuning via learning critical steps. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25334–25349, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

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

Yilun Hao, Yongchao Chen, Yang Zhang, and Chuchu Fan. Large language models can solve real-world planning rigorously with formal verification tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3434–3483, 2025. doi: 10.18653/v1/2025.naacl-long.176.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. Plan-then-execute: An empirical study of user trust and team performance when using llm agents as a daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–22, 2025.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, volume 2025, pp. 21344–21377, 2025.

Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42217–42227, 2026.

Jongwoo Ko, Sungnyun Kim, Tianyi Chen, and Se-Young Yun. Distillm: Towards streamlined distillation for large language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 24872–24895, 2024.

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

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan ang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. In *ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning*, 2026b.

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.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, et al. Mopd: Multi-teacher on-policy distillation for capability integration in llm post-training. *arXiv preprint arXiv:2606.30406*, 2026.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584*, 2024.

Hayley Ross, Ameya Sunil Mahabaleshwarkar, and Yoshi Suhara. When2call: When (not) to call tools. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3391–3409, 2025.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

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*, 2026.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11279–11298, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

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

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.

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, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

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, 2026a.

Ke Zhang, Yunjie Tian, Dongdi Zhao, Yijiang Li, Yuanye Liu, Vishal M Patel, and Di Fu. On-policy distillation with best-of-n teacher rollout selection. *arXiv preprint arXiv:2605.09725*, 2026b.

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.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

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

A Algorithm

B Experimental Details

Models. We study five teacher–student configurations using models from the Qwen3 (Yang et al., 2025) and Qwen3.5 (Team, 2026) families. Following prior multi-turn OPD studies (Wang et al., 2026; Zhou et al., 2026), teachers and students are paired within the same model family to reduce tokenizer and instruction-format mismatches. For Qwen3, we use Qwen3-32B and Qwen3-8B as

Algorithm 1 PREACT: Teacher-Guided Rollout Construction and Winner-Only Distillation

Require: Student parameters θ, frozen teacher π_T
Require: Training tasks, environment $\mathcal{E}$, candidate count K, horizon H
Require: Distillation coefficient β and clipping threshold ϵ
while the training budget remains **do**
  Fix $\theta_{\text{old}} \leftarrow \theta$ and initialize $\mathcal{D} \leftarrow \emptyset$
  for each task input x in the collection batch **do**
    Initialize the environment and obtain o_0
    Set $h_1 \leftarrow (x, o_0)$ and $\tau \leftarrow (x, o_0)$
    for $t = 1, \ldots, H$ **do**
      if the environment has terminated **then**
        break
      end if
      Sample $y_t^{(1)}, \ldots, y_t^{(K)} \overset{\text{i.i.d.}}{\sim} p_{\theta_{\text{old}}}(\cdot \mid h_t)$
      Compute teacher perplexities using Eq. equation 4
      Select $y_t^\star$ using Eq. equation 5, breaking ties by candidate index
      Cache the teacher token log-probabilities of $y_t^\star$
      Execute $g(y_t^\star)$ and receive o_t
      Append $(y_t^\star, o_t)$ to τ and set $h_{t+1} \leftarrow (h_t, y_t^\star, o_t)$
    end for
    Add τ to $\mathcal{D}$ with its cached teacher log-probabilities
  end for
  Set the supervision masks on selected assistant response tokens in $\mathcal{D}$
  Evaluate $\pi_{\theta_{\text{old}}}$ on these tokens under their recorded contexts
  Compute fixed weights $w_{t,i}$ using Eqs. equation 7 and equation 8
  Minimize Eq. equation 11 using the token loss in Eq. equation 10
end while
return the trained student π_θ

teachers, each paired with Qwen3-4B and Qwen3-1.7B as students, forming a full 2×2 teacher–student grid. This design allows us to compare teacher scales at a fixed student size and student capacities under a fixed teacher. We further include Qwen3.5-27B as a teacher for Qwen3.5-4B, extending the evaluation to another model generation pairs. Together, these configurations allow us to examine how distillation performance varies across teacher scales, student capacities, and teacher–student size gaps.

Benchmarks and Evaluation Metrics. We evaluate our method on three multi-turn agent benchmarks: ALFWorld (Shridhar et al., 2021), ScienceWorld (Wang et al., 2022), and WebShop (Yao et al., 2022), covering household navigation and object manipulation, scientific task solving, and instruction-guided web shopping, respectively. We report success rate (SR), defined as the percentage of tasks successfully completed within the interaction budget, across all three benchmarks. For ALFWorld, we evaluate the seen and unseen splits separately to assess performance in familiar and novel environments. For ScienceWorld, we additionally report the mean task score to capture partial progress and instruction satisfaction, respectively. Agents interact with each environment through a ReAct-style reasoning-and-action interface (Yao et al., 2023b). All students are evaluated independently with a single response, measuring the capabilities retained under teacher-free deployment.

B.1 Benchmarks and Evaluation Metrics

We evaluate agents on ALFWorld, ScienceWorld, and WebShop, which require sequences of actions conditioned on environment feedback. Success rate (SR) is the percentage of evaluation tasks completed successfully within the interaction budget. Each task contributes equally to the aggregate metric.

ALFWorld. ALFWorld (Shridhar et al., 2021) evaluates instruction following in household environments through navigation and object manipulation. Its tasks include placing, examining, cleaning, heating, and cooling objects, as well as placing two objects. We evaluate 140 tasks in seen

Table 6: Training configuration.

Parameter	ALFWorld	ScienceWorld	WebShop
Training instances	3,553	2,064	3,686
Rollout batch size	24	16	16
Global batch size	24	16	16
Completed student updates	147	250	255
Learning rate	10^{-6}	10^{-6}	10^{-6}
Distillation coefficient β	1	1	1
Candidates, Vanilla OPD / PREACT	1 / 4	1 / 4	1 / 4
Training temperature	1.0	1.0	1.0
Maximum agent turns	50	30	30
Context limit	32,768	32,768	32,768
Response limit per turn	4,096	4,096	1,024

environments and 134 tasks in unseen environments. Overall SR is computed over all 274 tasks, and therefore weights the seen and unseen splits by their respective task counts.

ScienceWorld. ScienceWorld (Wang et al., 2022) evaluates scientific task completion in an interactive environment. The reference configuration uses 2,064 training instances and a held-out pool of 230 instances from the TCOD seen partition, with evaluation on 1,308 unseen instances. The task score is the highest environment score attained during an episode, on a scale from 0 to 100. We report its mean across evaluation tasks and count an episode as successful when this score reaches 100.

WebShop. WebShop (Yao et al., 2022) evaluates instruction-guided product search and selection. The reference configuration reserves 410 instances from a pool of 4,096 training instructions, leaving 3,686 training instances, and evaluates on 100 test instructions. Success requires a terminal episode with reward equal to one.

B.2 Training Configuration

Table 6 specifies the training configuration for the Qwen3-32B–Qwen3-1.7B reference runs. Candidate generation uses temperature 1.0, top-p equal to 1.0, and unrestricted top-k sampling. The teacher scores the generated response tokens under their recorded interaction histories. The reported update counts refer to completed student updates, including those completed before a resumed run.

B.3 Prompts and Interaction Protocols

Each benchmark provides a task instruction and observations through its environment-specific interface. The interaction history retains the generated responses and returned observations across turns. Episodes terminate according to the environment status, interaction and token budgets, and the configured response parser (Hu et al., 2025). The reference implementation retains the full history and stops when the history plus the next response allowance exceeds the context limit.

ALFWorld. Responses use a `Thought:` field followed by an `Action:` field containing an environment command. The reference configuration disables the model's chat-template thinking mode while retaining the visible reasoning field. Observations include environment feedback and admissible commands. During evaluation, submitted actions are passed to the environment, which returns its native feedback for invalid or ineffective commands.

ScienceWorld. The prompt presents the task, current observation, available action templates, and relevant objects. The response places reasoning inside `<think>...</think>` and the action inside `<action>...</action>`. The per-turn prompt additionally repeats the two most recent observation–action pairs once available, alongside the retained conversation history. The parser extracts the first complete action span; an absent action span is passed to the service as a no-op.

Table 7: Evaluation settings for the recorded Qwen3 reference runs. All token limits refer to model tokens.

Parameter	ALFWorld	ScienceWorld	WebShop
Evaluation tasks	140 + 134	1,308	100
Maximum agent turns	50	30	30
Context limit	32,768	32,768	32,768
Response limit per turn	4,096	4,096	1,024
Temperature	0.4	0.4	0.4

WebShop. The initial user prompt specifies the shopping objective and the response format. The reference configuration enables private thinking and requires two visible fields, `Thought:` and `Action:`. Actions take the form `search[keywords]` or `click[value]`, with the click value identifying an available interface element. The returned page content supplies the next observation.

B.4 Evaluation and Aggregation

Table 7 gives the evaluation settings for the recorded Qwen3 reference runs. Evaluations are all use temperature 0.4. Evaluation uses a single generated response at each turn. For each run, we first aggregate episode outcomes over the corresponding benchmark split. Entries marked with $\pm$ in the main-result tables summarize three evaluation runs. The selection and candidate-budget ablations(Tabel 3 and Tabel 4) report one reference evaluation per configuration.

B.5 Environment Prompts

We use environment-specific prompts to elicit a reasoning step followed by one executable action. ALFWorld and WebShop use locally adapted ReAct-style instructions (Yao et al., 2023b), while ScienceWorld uses the prompt templates from the official TCOD implementation (Wang et al., 2026). The following boxes reproduce the fixed instructions and indicate environment-dependent fields in braces.

ALFWorld Prompt

System message

You are solving a text-only ALFWorld household task. At every turn, reason only from the task and observations already shown, then choose one exact command from the latest admissible-command list. Never claim an action happened before the environment returns its next observation. Reply with exactly these two fields and no JSON, tool-call markup, or additional fields:

Thought: `<brief, state-grounded reasoning>`
Action: `<one exact current admissible command>`

User observation template

[ALFWorld observation]
{environment_feedback}

Admissible commands:
{admissible_command_list}

Episode status: reward={reward} done={done} won={won}

Example task. A recorded ALFWorld training input contains the following task instruction, accompanied by the room description and admissible commands:

> Your task is to: look at alarmclock under the desklamp.

ScienceWorld Prompt (First Two Steps)

Your ScienceWorld task is: {task_description}
Your current observation is: {current_observation}
Available action commands: [{action_templates}]
Available objects you can interact with: [{objects}]

Now it's your turn to take an action. Combine an action command with appropriate object(s) to form a valid 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 a valid action for the current step and present it within `<action> </action>` tags.

After two actions, the current-turn prompt additionally summarizes the latest two observation–action pairs. The full dialogue history remains in the model context.

ScienceWorld Prompt (Subsequent Steps)

Your ScienceWorld task is: {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}
Available action commands: [{action_templates}]
Available objects you can interact with: [{objects}]

Now it's your turn to take an action. Combine an action command with appropriate object(s) to form a valid 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 a valid action for the current step and present it within `<action> </action>` tags.

Each history entry uses the following format:

[Observation {step_num}: '{obs}', Action {step_num}: '{act}']

WebShop Prompt

Complete the WebShop shopping goal. Think privately before answering. After private thinking, output exactly two visible lines:

Thought: `<brief visible rationale>`
Action: `<one action>`

Use search[keywords] only on a search page or click[value] with an exact available value. Do not output tool-call markup, and do not claim an action happened before the environment returns its result.

[WebShop observation]
Shopping goal:
{shopping_goal}

Current page:
{current_page}

Available actions:
{available_action_list}

Episode status: task_score={task_score} steps={steps} done={done} won={won}

Example task. The shopping-goal field in a recorded WebShop evaluation instance reads:

Instruction: i am looking for a long sleeve mens hoodie pullover size x-large, and price lower than 40.00 dollars

C Additional Experimental Results

C.1 Cross-Generation Generalization

We evaluate PREACT with a Qwen3.5-27B teacher and a Qwen3.5-4B student on ScienceWorld to test whether its benefits extend across model generations (Team, 2026). Table 8 reports the results.

Table 8: ScienceWorld results with a Qwen3.5-27B teacher and a Qwen3.5-4B student. Bold identifies the best distillation result in each column.

Method	SR (%) ↑	Score ↑
Initial student	28.90	53.38
Initial teacher	47.17	69.90
Vanilla OPD	44.04	68.97
PREACT (ours)	**45.57**	**69.56**

Compared with Vanilla OPD, PREACT improves success rate by 1.53 percentage points and mean task score by 0.59 points. The score gap to the initial teacher decreases from 0.93 to 0.34 points. These results show that the benefits of pre-action selection extend to a different model generation and teacher–student pair.

C.2 Results by Task Type

Table 9 compares Vanilla OPD and PREACT across the six ALFWorld task categories using a Qwen3-32B teacher and a Qwen3-4B student. In this reference evaluation, PREACT improves success rates in four categories and matches Vanilla OPD in the remaining two. The largest gains occur in cooling and placing two objects, with improvements of 13.04 and 9.76 percentage points, respectively. Together with improvements in examination and cleaning, these gains raise the total number of successful tasks from 158 to 174, increasing overall success from 57.66% to 63.50%.

Table 9: ALFWorld success rate (%) by task category for a Qwen3-32B teacher and a Qwen3-4B student. Results pool the 140 seen and 134 unseen tasks from one reference evaluation per method with evaluation seed 1234. Vanilla OPD uses $K = 1$ during training and PREACT uses $K = 4$; both generate one response per evaluation turn. N denotes the number of tasks. Gains are measured in percentage points and computed from unrounded rates.

Task type	N	Vanilla OPD	PREACT	Gain
Simple pick-and-place	59	100.00	100.00	0.00
Examine in light	31	74.19	**83.87**	+9.68
Clean and place	58	44.83	**50.00**	+5.17
Place two objects	41	60.98	**70.73**	+9.76
Cool and place	46	52.17	**65.22**	+13.04
Heat and place	39	2.56	2.56	0.00
Overall	274	57.66	**63.50**	+5.84

C.3 Computational Cost

We assess the practical training cost of PREACT at $K = 4$, using a Qwen3-4B student with Qwen3-32B or Qwen3-8B teachers. Table 10 compares its logged training time with Vanilla OPD ($K = 1$) at matched benchmark-specific update endpoints. Runtime spans the start of logging to the last recorded metric, covering initialization, trajectory collection, and student updates.

At $K = 4$, PREACT expands the per-turn candidate pool fourfold, with logged training time ranging from $1.21\times$ to $2.22\times$ that of Vanilla OPD. The additional computation comes from generating and scoring multiple candidates, which can be processed independently given the same interaction

Table 10: Logged training time with a Qwen3-4B student at matched update endpoints. Vanilla OPD uses $K = 1$ and PREACT uses $K = 4$. Each entry summarizes one successful training chain. The runtime ratio is PREACT divided by Vanilla OPD.

Teacher	Benchmark	Updates	Vanilla (h)	PREACT (h)	Ratio
Qwen3-32B	ALFWorld	147	6.05	13.10	2.16×
Qwen3-32B	ScienceWorld	250	15.34	30.04	1.96×
Qwen3-32B	WebShop	255	7.18	14.65	2.04×
Qwen3-8B	ALFWorld	147	6.59	7.98	1.21×
Qwen3-8B	ScienceWorld	250	15.75	20.77	1.32×
Qwen3-8B	WebShop	255	6.37	14.13	2.22×

history. This candidate-level parallelism means that additional FLOPs need not translate into a proportional increase in wall-clock time. Further improvements in batching, parallel teacher scoring, and scheduling could reduce the runtime overhead. The separate ALFWorld ablation with a Qwen3-1.7B student (Table 4) further demonstrates more effective use of the candidate budget: PREACT improves overall success by 4.74 percentage points over Random-Commit under the same per-turn candidate budget. The additional sampling and scoring are confined to training, while evaluation retains a single student response per turn.

C.4 Action Admissibility

We examine recorded action admissibility for the Qwen3-1.7B and Qwen3-4B reference checkpoints trained with a Qwen3-32B teacher. Both methods generate one student response per evaluation turn (Masterman et al., 2024). The admissibility rate is the fraction of all executed actions explicitly marked admissible by the environment.

ALFWorld checks membership in the current admissible-command list, while ScienceWorld checks membership in the backend's valid-action list. WebShop checks whether a search query is nonempty on a search page or a clicked element is available in the current interface. For each method, we first average evaluation rates across repeated executions within each seed and then average across seeds.

Table 11: Recorded action admissibility (%) in reference evaluations with a Qwen3-32B teacher. ALFWorld uses one evaluation seed; ScienceWorld and WebShop average two seeds. Each evaluation covers 274 ALFWorld tasks, 1,308 ScienceWorld tasks, or 100 WebShop tasks. Gains are measured in percentage points and computed before rounding.

Student	Benchmark	Vanilla OPD	PREACT	Gain
Qwen3-1.7B	ALFWorld	80.74	**82.03**	+1.29
Qwen3-1.7B	ScienceWorld	27.86	**28.98**	+1.11
Qwen3-1.7B	WebShop	50.13	**66.99**	+16.86
Qwen3-4B	ALFWorld	90.14	**91.11**	+0.96
Qwen3-4B	ScienceWorld	44.25	**45.12**	+0.88
Qwen3-4B	WebShop	73.33	**74.75**	+1.43

Table 11 shows higher aggregated action admissibility for PREACT across all six student–benchmark configurations. The largest improvement occurs on WebShop with Qwen3-1.7B, where the rate increases from 50.13% to 66.99%, a gain of 16.86 percentage points.

D Analysis

We analyze selection with the student parameters θ, decoding configuration, and teacher scoring rule held fixed. At a history h, let $p_\theta(\cdot \mid h)$ denote the complete-response sampling distribution (Wang et al., 2023; Yao et al., 2023b; He et al., 2025). The student independently samples K candidates, and the candidate with the lowest teacher perplexity is selected. Ties are resolved by the lowest

candidate index, as in the method. We write $b_K(\cdot \mid h)$ for the selected-response distribution. The finite vocabulary and response-length limit give finite response support. All logarithms below are natural.

D.1 Selection at a Fixed History

Teacher-perplexity tail. Define $R_h(u) = \Pr_{Y \sim p_\theta(\cdot|h)}[\mathrm{PPL}_T(Y \mid h) > u]$. The selected perplexity exceeds u exactly when all K candidate perplexities exceed u. Independence therefore gives

$$\Pr_{Y \sim b_K(\cdot|h)}[\mathrm{PPL}_T(Y \mid h) > u] = R_h(u)^K. \tag{14}$$

Since $0 \le R_h(u) \le 1$, this upper-tail probability is non-increasing in K for every threshold u.

Deviation from student sampling. The best-of-K KL bound of Beirami et al. (2025) takes the form

$$D_{\mathrm{KL}}(b_K(\cdot \mid h) \,\|\, p_\theta(\cdot \mid h)) \le \delta_K, \qquad \delta_K = \log K - \frac{K-1}{K}. \tag{15}$$

The following argument also covers tied scores under our tie-breaking rule. At fixed h, group responses by their score $s = -\mathrm{PPL}_T$. Index the distinct score groups in increasing order. Let $m_j > 0$ be the student probability of group j, and let $F_j = \sum_{\ell \le j} m_\ell$, with $F_0 = 0$. The probability that the selected response belongs to group j is

$$w_j = F_j^K - F_{j-1}^K = \int_{F_{j-1}}^{F_j} K u^{K-1} \, \mathrm{d}u. \tag{16}$$

Conditional on the winning group, selecting its earliest occurrence preserves the original within-group response distribution. Thus, for a response y in group j, $b_K(y \mid h) = p_\theta(y \mid h) w_j / m_j$. Applying Jensen's inequality to the convex function $z \log z$ on each interval yields

$$\begin{aligned} D_{\mathrm{KL}}(b_K \| p_\theta) &= \sum_j w_j \log \frac{w_j}{m_j} \\ &\le \int_0^1 K u^{K-1} \log\big(K u^{K-1}\big) \, \mathrm{d}u = \log K - \frac{K-1}{K}. \end{aligned} \tag{17}$$

The bound is zero at $K = 1$ and increases with K. The actual divergence additionally depends on the student distribution and teacher ranking.

D.2 Propagation Through Interaction

The selected response enters both the distillation data and the next interaction history:

$$O_t \sim E(\cdot \mid h_t, g(Y_t^\star)), \qquad h_{t+1} = (h_t, Y_t^\star, O_t). \tag{18}$$

Here, g extracts the executed action and E is the environment feedback kernel. For a fixed student, compare trajectory distributions P_K and P_1 induced by b_K and p_θ, respectively, with the same initial-history distribution, environment, and stopping rule. Assume at most H turns, padding terminated trajectories with common deterministic absorbing transitions. Selection ensures $\operatorname{supp} b_K(\cdot \mid h) \subseteq \operatorname{supp} p_\theta(\cdot \mid h)$. The common initial and environment factors cancel in the trajectory likelihood ratio (Du et al., 2023; Hao et al., 2025). Writing d_t^K for the distribution of h_t under P_K, the KL chain rule gives

$$\begin{aligned} D_{\mathrm{KL}}(P_K \| P_1) &= \sum_{t=1}^{H} \mathbb{E}_{h_t \sim d_t^K}\left[D_{\mathrm{KL}}(b_K(\cdot \mid h_t) \| p_\theta(\cdot \mid h_t))\right] \\ &\le H \delta_K. \end{aligned} \tag{19}$$

At absorbing histories, both response distributions are deterministic and their KL contribution is zero. This calculation connects response selection to the distribution of collected trajectories.

D.3 Candidate Coverage and Selection Errors

Fix h and a binary task-quality criterion that labels each candidate response as acceptable or unacceptable. The criterion is fixed across K and defined independently of teacher perplexity (Gao et al., 2023). Let $p(h)$ be the probability that a response sampled from $p_\theta(\cdot \mid h)$ is acceptable. Let A_K denote the event that at least one of the K candidates is acceptable. For $p(h) > 0$, define $\eta_K(h) = \Pr(Y^\star \text{ is unacceptable} \mid A_K)$. For $0 < p(h) < 1$, independence and the law of total probability give

$$\epsilon_K(h) = \underbrace{(1-p(h))^K}_{\text{coverage failure}} + \underbrace{\eta_K(h)\left[1-(1-p(h))^K\right]}_{\text{error despite coverage}}, \tag{20}$$

where $\epsilon_K(h)$ is the probability of selecting an unacceptable response. If $p(h) = 0$, then $\epsilon_K(h) = 1$; if $p(h) = 1$, then $\epsilon_K(h) = 0$. For $K = 1$ and $p(h) > 0$, $\eta_1(h) = 0$.

To compare budgets $L > K$, suppress the dependence on h and set $a = 1 - p(h)$. Rearranging Equation equation 20 gives

$$\epsilon_L - \epsilon_K = (1-a^L)(\eta_L - \eta_K) - (a^K - a^L)(1-\eta_K). \tag{21}$$

Hence, for $0 < p(h) < 1$, total selection error increases precisely when

$$\eta_L - \eta_K > \frac{(a^K - a^L)(1-\eta_K)}{1-a^L}. \tag{22}$$

Additional candidates reduce coverage failure. An increase in conditional selection error raises total error only when it exceeds this threshold. The behavior of $\eta_K(h)$ depends on how teacher ranking relates to the chosen task-quality criterion.

D.4 Hypotheses for the Non-monotonic Budget Effect

The ALFWorld budget study reports lower overall success at $K = 8$ than at the best tested budget for each teacher. Two mechanisms could contribute to this pattern (Zhang et al., 2026b). First, larger candidate pools may expose responses that receive low teacher perplexity despite poor task quality, increasing $\eta_K(h)$ (Gao et al., 2023). Second, teacher selection may reduce exposure to histories that the student frequently encounters when acting without selection, providing fewer training examples at those histories. Both mechanisms concern the learning effects of the collected data. The preceding derivations hold for a fixed student, whereas the budget study evaluates students trained separately at each K. We therefore treat these mechanisms as hypotheses for the observed performance pattern; their contributions require direct measurements of selection errors and interaction-history coverage.

E Training Dynamics

We report teacher perplexity on the responses used for distillation throughout training on ALFWorld, ScienceWorld, and WebShop (Figures 4–6). We compare Vanilla OPD and PREACT using a fixed Qwen3-32B teacher and a Qwen3-1.7B student, with $K = 4$ candidates per turn for PREACT. The horizontal axis denotes training steps, and the vertical axis denotes teacher perplexity. Faint lines show raw per-batch values, while dark lines show smoothed perplexity computed over the supervised response tokens in the most recent 20 batches.

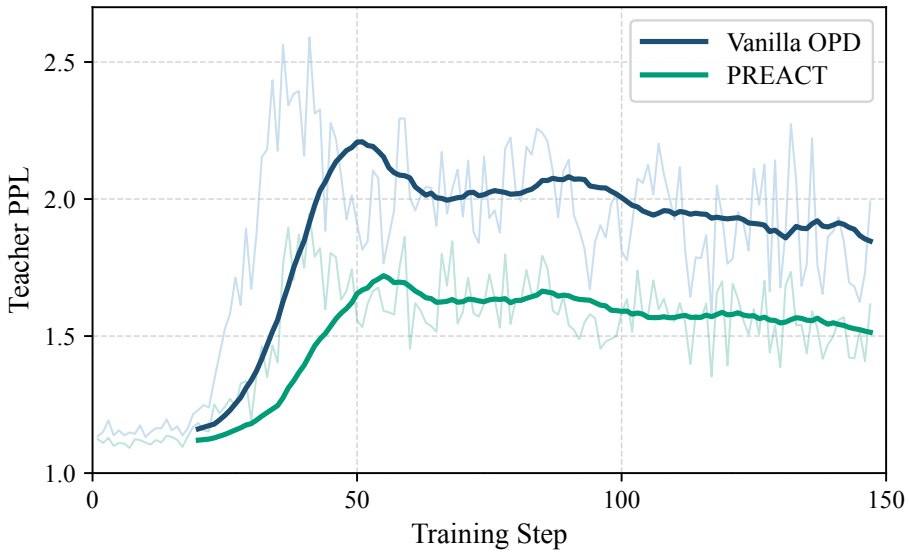


Figure 4: Teacher perplexity during training on ALFWorld.

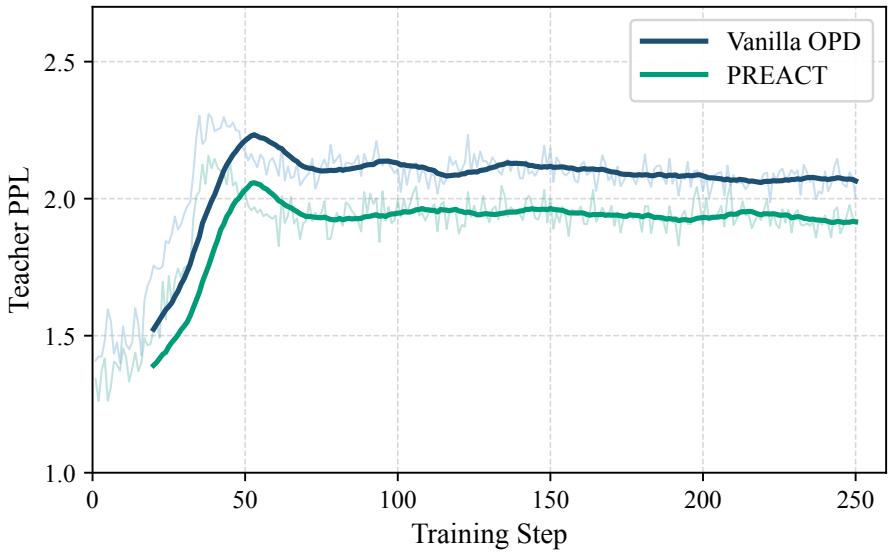


Figure 5: Teacher perplexity during training on ScienceWorld.

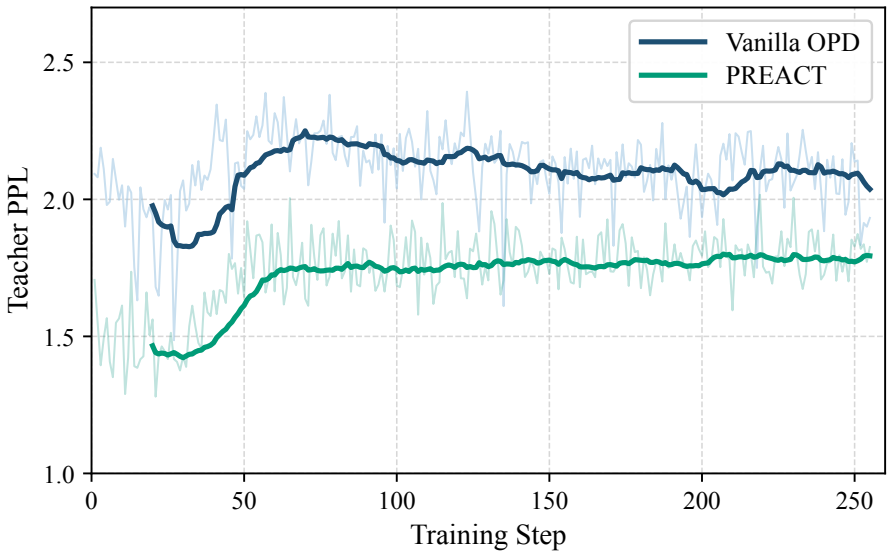


Figure 6: Teacher perplexity during training on WebShop.

F QUALITATIVE CASE STUDIES

We present one evaluation case from each benchmark. For each case, the students trained with PREACT and Vanilla OPD execute separate rollouts on the same task. The cases illustrate three aspects of carrying task requirements across turns: completing an intermediate subgoal in ALFWorld, retaining the selected object in ScienceWorld, and satisfying a product attribute constraint in WebShop.

ALFWorld: completing the cooling subgoal. The task in Figure 7 requires a cool mug to be placed in a coffee machine. After matching actions and observations for the first five steps, the agents differ in their next subgoal: PREACT cools the mug in the fridge before placement, whereas Vanilla OPD proceeds directly to placement and receives zero reward. PREACT completes the task in 10 steps; Vanilla OPD later enters a repeated `look` loop and exhausts the 50-step budget.

ScienceWorld: retaining the selected object. Figure 8 shows a task in which the agent must focus on a non-living object and move it into a red box. Both agents initially focus on a picture and enter the kitchen. PREACT returns to collect that picture and places it in the box, reaching 100/100 in six environment actions. Vanilla OPD switches to a painting in the kitchen. Although the environment confirms that the painting is moved into the box, the score remains at 75/100; the agent ends with a completion claim after its fifth action.

WebShop: satisfying the requested size. The request in Figure 9 combines product retrieval with a size constraint: an x-large men's hoodie priced below $40. PREACT finds a matching item, explicitly selects `x-large`, and completes the purchase in six steps, with the selected size recorded in the purchase result. Vanilla OPD returns to the Search page at step 12, then repeats an unavailable back-to-search action for the remaining 18 steps.

G LIMITATION

Our evaluation focuses on text-based interaction in ALFWorld, ScienceWorld, and WebShop. The applicability of PREACT to environments with visual observations remains to be evaluated. Extending the study to multimodal interaction is a direction for future work.

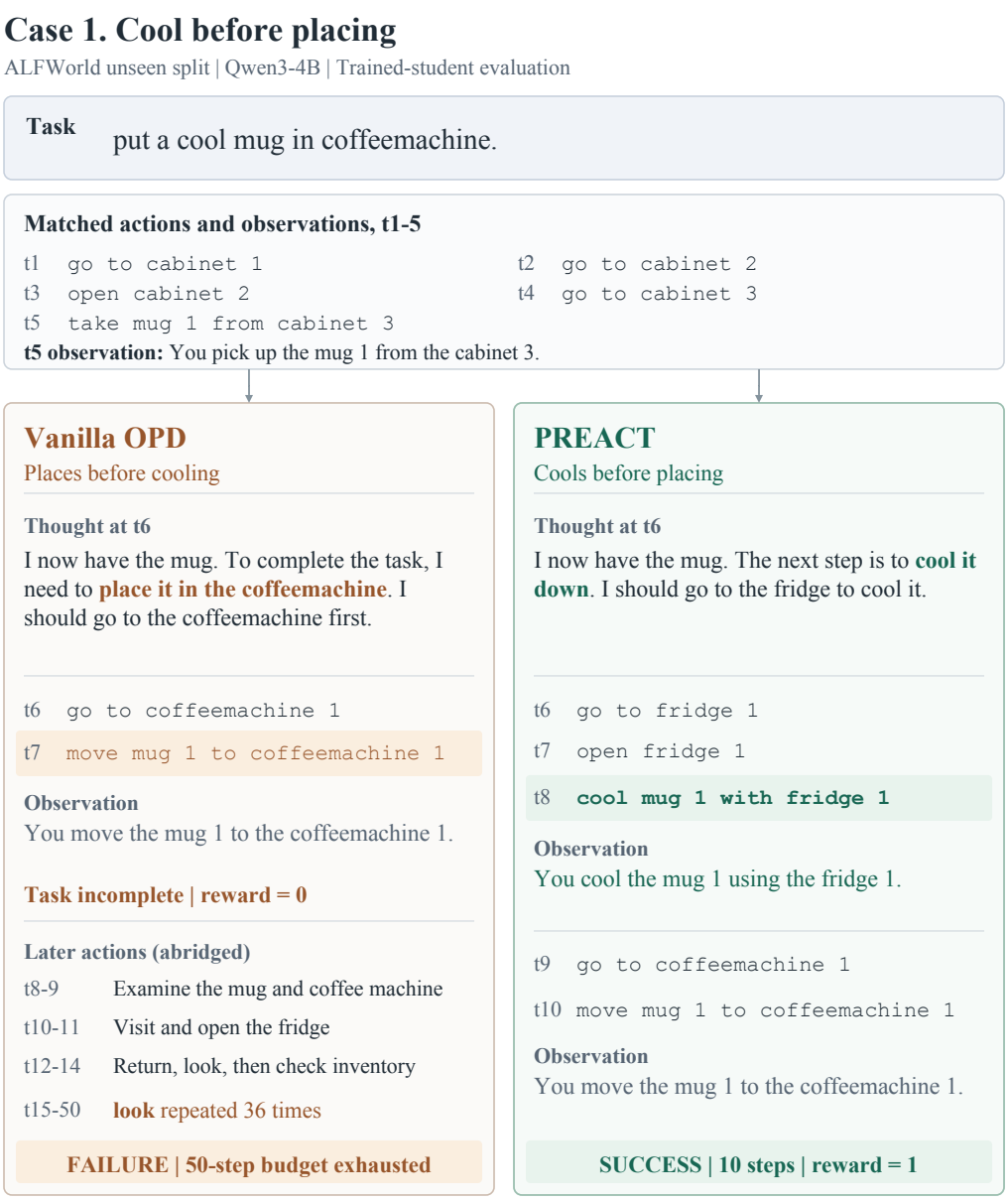


Figure 7: **ALFWorld: completing the cooling subgoal.** PREACT cools the mug before placement and succeeds in 10 steps; Vanilla OPD places the mug first and exhausts the 50-step budget. The Qwen3-4B students are evaluated independently on the same task and have matching actions and observations for the first five steps. Thoughts at step 6 are verbatim; environment feedback is excerpted and the remaining Vanilla trajectory is summarized.

Case 2. Return for the selected object

ScienceWorld TCOD unseen | Qwen3.5-4B | Trained-student evaluation

Task

Your task is to find a(n) non-living thing. First, focus on the thing. Then, move it to the red box in the kitchen.

Matched actions and observations, t1-2

t1 `focus on picture` t2 `go kitchen`

t2 observation: You move to the kitchen.

Score after t2: 75/100

Vanilla OPD

Switches to a nearby painting

Thought at t3 (excerpt)

Let me **focus on the painting** in the kitchen, since that's in the same room as the red box.

t3 **`focus on painting`**

t4 `move painting to red box`

Observation

You move the painting to the red box.

Score remains 75/100

t5 `task`

Observation (summary)

Displays the task description again.

Final answer (excerpt)

The task has been completed successfully!

INCOMPLETE | 5 steps | score = 75/100

PREACT

Returns for the selected picture

Thought at t3 (excerpt)

I need to go back to the hallway to **pick up the picture**, then bring it to the kitchen and move it to the red box.

t3 `go hallway`

t4 **`pick up picture`**

Observation

You move the picture to the inventory.

Score increases to 83/100

t5 `go kitchen`

t6 **`move picture to red box`**

Observation

You move the picture to the red box.

Score reaches 100/100

SUCCESS | 6 steps | score = 100/100

Figure 8: **ScienceWorld: preserving the selected object.** PREACT returns for the initially focused picture and places it in the red box, reaching 100/100 in six actions. Vanilla OPD switches to a painting and declares completion after five actions while scoring 75/100. The Qwen3.5-4B students (iteration 250) are evaluated independently; actions and observations match for the first two steps. Thought, final-answer, and feedback excerpts are verbatim, with the repeated task description summarized. Scores are 100 times the recorded rewards.

Case 3. Select the requested size

WebShop test split | Qwen3-1.7B | Trained-student evaluation

Task

i am looking for a long sleeve mens hoodie pullover size x-large, and price lower than 40.00 dollars

Vanilla OPD

Repeats unavailable actions

Thought at t13

No items match the criteria. The assistant must conclude the task as no matching item found.

t1-11 Earlier search (abridged)

Search, return to Search, and re-query.
t2: free-form output without an action.
t9: search issued from a results page.

t12 `click[back to search]`

Returned to the Search page

Available action

`search[<your query>]`

t13-30 Repeated 18 times

`click[back to search]`

Unavailable on the Search page.
All 18 actions are inadmissible.

20 of 30 actions/outputs inadmissible

FAILURE | 30-step budget exhausted

PREACT

Sets size before purchase

Thought at t5

The item is a men's hoodie under $40. **The size X-large is available.** Selecting it now will complete the purchase.

t1 `search[long sleeve men's hoodie pullover X-large under 40 dollars]`

t2 `click[next >]`

t3 `click[next >]`

t4 `click[b09p3mp6s7]`

Product page (excerpts)

Mens Pullover Hoodies ...
... Long Sleeve ...
Price: $9.99 to $14.99
Size option: **x-large**

t5 **`click[x-large]`**

t6 `click[buy now]`

Purchase record (excerpt)

options
{"size": "x-large"}

SUCCESS | 6 steps | reward = 1

Figure 9: **WebShop: satisfying the size constraint.** PREACT browses to a matching hoodie, selects `x-large`, and completes the purchase in six steps with reward 1. The purchase record confirms the selected size. Vanilla OPD exhausts the 30-step budget, including 18 consecutive unavailable back-to-search actions at steps 13–30. These are independent Qwen3-1.7B evaluations on `goal_4107`. Thoughts are verbatim; product details are excerpted and the early Vanilla trajectory is summarized.