SAGE-OPD: SELECTIVE AGENT-GUIDED INTERVENTION FOR MULTI-TURN ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) improves student models by training them on trajectories induced by their own policy, making it a promising approach for mitigating exposure bias in agent training. However, most OPD studies focus on single-turn settings, while realistic LLM agents interact with environments over multiple turns. In this regime, early errors can alter future observations and compound across the trajectory, and standard dense token-level OPD becomes brittle, as it may over-penalize semantically valid alternatives, reinforce local degeneracies such as repeated actions, and propagate unreliable teacher supervision on off-distribution histories. We propose **SAGE-OPD**, a verifier-free selective intervention framework specifically designed for multi-turn OPD. Instead of applying teacher supervision uniformly, SAGE-OPD uses pre-execution validity checks and teacher judgment to decide whether each student response requires supervision. SAGE-OPD further weights the selected turns by teacher confidence, assigning greater relative weight to turns with more decisive teacher predictions. Finally, normalizing weights within each batch preserves the relative allocation of supervision while maintaining unit average effective token weight. With a Qwen3-1.7B student and Qwen3-8B teacher, SAGE-OPD raises ALFWorld unseen success from 61.94% to 70.15% and ScienceWorld success from 4.70% to 6.71% over standard OPD, while remaining competitive on SearchQA. Component ablations isolate intervention and confidence weighting; three-seed replications, expert-action comparisons, and shared-state continuation evaluations provide complementary evidence of robustness and agent behavior. These findings support feedback-conditioned turn-level supervision allocation for multi-turn OPD.

1 INTRODUCTION

On-policy distillation (OPD) has become a leading approach for post-training open-weight reasoning models (Agarwal et al., 2024; Lu & Lab, 2025; Song & Zheng, 2026; Zhou et al., 2026b). By training the student on samples generated from its own policy, OPD mitigates the exposure-bias limitation of offline supervised fine-tuning (SFT), where the student is optimized only on teacher-generated states (Zhou et al., 2025a; Hao et al., 2026). This on-policy property becomes especially important when moving from single-turn reasoning to multi-turn LLM agents. Unlike single-turn generation, a multi-turn agent repeatedly acts in an environment and conditions on the resulting observations, so each student action can influence the future states encountered during training and evaluation. OPD appears to be a natural fit for agent training, because it allows a strong teacher to supervise the student on histories induced by the student's own behavior, rather than only on expert-generated trajectories.

However, most existing OPD studies primarily evaluate single-turn generation on reasoning tasks (Yang et al., 2026; Jin et al., 2026), where each example is completed in one model rollout and there is no environment state that evolves in response to intermediate actions. This setting abstracts away a central difficulty of realistic agent learning. In multi-turn interaction, errors can compound over time: an early incorrect action may alter the environment state, distort later observations, and make recovery in subsequent turns increasingly difficult.

Despite the advantage of using student-induced trajectories, recent analyses have revealed that the token-level logit signal underlying standard OPD is brittle (Li et al., 2026b; Luo et al., 2026). This brittleness can be amplified in multi-turn agent training. First, standard OPD applies dense

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.

token-level alignment between the student and teacher distributions, even when several continuations are semantically valid (Li et al., 2026b). For example, two actions may follow different surface forms or intermediate reasoning paths while still making equivalent progress toward the task. In such cases, token-level distillation may over-penalize acceptable student behavior simply because it does not match the teacher's preferred wording or local decomposition. Second, dense token-level gradients can reinforce local degeneracies (Luo et al., 2026), such as repeated thoughts, repeated actions, or shallow formatting patterns, when these prefixes accidentally receive favorable teacher likelihood. These local degeneracies are particularly harmful in multi-turn interaction, because a repeated or malformed action does not only affect the current response but also changes the subsequent environment feedback. Third, compounding errors can make later student-induced histories increasingly off-distribution for the teacher (Li et al., 2025). The teacher may still provide a distribution over next tokens, but that distribution can become less reliable when the current history is the result of earlier student mistakes. Therefore, uniformly applying OPD to every token at every turn can both over-correct acceptable behavior and propagate unreliable supervision in corrupted histories.

These observations motivate our central question: *Can response-level assessment and available environment context guide supervision allocation to improve multi-turn OPD?*

We answer this question with **SAGE-OPD**, a selective intervention framework for multi-turn OPD that requires no task-success verifier or reward model. Our key observation is that multi-turn interaction already segments the student trajectory into turn-level units, providing a natural granularity for teacher critique beyond token-level alignment. Before executing each response, SAGE-OPD uses validity checks and teacher judgment to decide whether to apply supervision. Unlike a fixed curriculum over training steps or trajectory depth, this decision depends on the student response and available interaction context. Intervention changes loss weights while the student continues to generate the executed responses, preserving on-policy interaction and reducing supervision on responses judged acceptable.

The need for intervention and the teacher's predictive decisiveness describe different aspects of supervision. A student response may be acceptable even when the teacher strongly prefers another continuation, while a response requiring correction may admit several plausible alternatives. SAGE-OPD therefore combines a response-level intervention weight with the teacher's average top-1 probability, allocating greater relative weight to selected turns with more decisive teacher distributions. Normalization maintains the average effective token weight while preserving this relative allocation.

We evaluate SAGE-OPD on ALFWorld, ScienceWorld, and SearchQA (Shridhar et al., 2020; Wang et al., 2022; Jin et al., 2025). With a Qwen3-1.7B student and Qwen3-8B teacher (Yang et al., 2025a), SAGE-OPD achieves approximate relative improvements of 13.3% in ALFWorld unseen SR and 42.8% in ScienceWorld SR over standard OPD. Across the evaluated distillation methods, SAGE-OPD achieves the best or tied-best unseen SR in three of four ALFWorld configurations and the best SR in three of four ScienceWorld configurations, while remaining competitive on SearchQA. Our contributions are summarized as follows:

- We introduce SAGE-OPD, a response-conditioned supervision allocation method for multi-turn OPD. Validity checks and teacher assessment determine where distillation is applied, predictive decisiveness modulates supervision, and normalization maintains the effective token weight.
- We evaluate SAGE-OPD on ALFWorld, ScienceWorld, and SearchQA, demonstrating gains in embodied interaction and scientific task solving alongside competitive search-based question answering performance.
- We isolate supervision allocation through shuffled-label, confidence-only, and NoConf ablations, and assess intervention judgments against independent expert actions. Three-seed replications, shared-state continuation experiments, and runtime measurements characterize robustness, agent behavior, and training cost.

2 METHODOLOGY

2.1 PRELIMINARIES

Multi-turn Agent Interaction. We consider an agent that interacts with an external environment over multiple turns. At turn t, the agent receives an observation o_t from the environment and

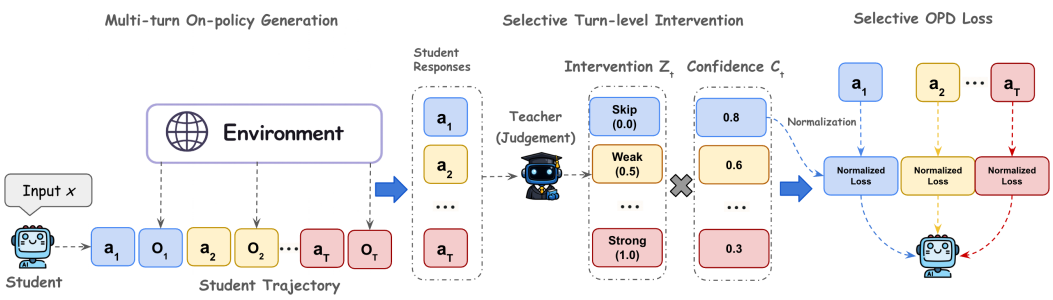

Figure 1: Overview of SAGE-OPD. The student generates responses conditioned on its interaction history. Before each response is executed, SAGE-OPD assigns an intervention weight using deterministic validity checks and teacher judgment, then combines it with teacher predictive confidence. Batch normalization converts these weights into relative contributions to the OPD objective; the student response is executed regardless of its intervention label.

conditions on the full interaction history $h_t = (o_0, a_0, o_1, a_1, \ldots, o_{t-1}, a_{t-1}, o_t)$, where a_i denotes the model response at turn i. The response a_t may include natural language reasoning, tool-use arguments, and an executable action. After the executable part of a_t is submitted to the environment, the environment returns the next observation: $o_{t+1} \sim \mathcal{E}(o_{t+1} \mid h_t, a_t)$.

A complete trajectory containing T turns is therefore $\tau = (h_0, a_0, h_1, a_1, \ldots, h_{T-1}, a_{T-1})$, which terminates either when a termination action is taken or when the maximum horizon T is reached. This dependency distinguishes multi-turn agent learning from single-turn generation: an early mistake can alter the environment state and affect all subsequent observations.

Multi-turn On-Policy Distillation. Let π_θ denote the student policy and π_ϕ denote the teacher policy. In multi-turn OPD, trajectories are sampled from the student policy, i.e., $a_t \sim \pi_\theta(\cdot \mid h_t)$ for $t = 0, \ldots, T-1$. The teacher provides supervision on these student-induced states through a KL-based distillation objective:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\sum_{t=0}^{T-1} \sum_{j=1}^{|a_t|} \mathcal{D}\left(\pi_\theta(\cdot \mid h_t, a_{t,<j}), \pi_\phi(\cdot \mid h_t, a_{t,<j})\right) \right]. \tag{1}$$

Here, $\mathcal{D}$ denotes a KL-based distillation term; our OPD and SAGE-OPD implementations share the same forward-KL token loss. The student is trained on its own state distribution, allowing the teacher to supervise histories arising from the student's behavior rather than only teacher-generated trajectories.

However, applying OPD uniformly to every turn in a multi-turn trajectory can be suboptimal. Early student errors may alter future observations and push later histories away from the teacher's reliable distribution, making subsequent supervision less dependable. Meanwhile, token-level OPD can be brittle when teacher and student policies assign probability mass to different but semantically valid continuations, or when dense alignment reinforces local degeneracies such as repeated actions. These issues motivate a selective mechanism that decides when teacher distillation should be applied and how strongly each selected turn should contribute.

2.2 SAGE-OPD

The limitations of dense multi-turn OPD suggest that teacher supervision should be applied selectively rather than uniformly across all turns. Standard OPD treats a student trajectory as a flat sequence of tokens and applies teacher distillation everywhere, which makes training sensitive to token-level mismatch even when the student response is semantically acceptable. Multi-turn interaction provides a natural way to move beyond this flat token-level view: the trajectory is already segmented by environment feedback, where each turn produces a response a_t, receives an observation, and transitions to a new history h_{t+1}. This turn structure allows the teacher to critique larger units of behavior and decide whether token-level distillation is needed for the current student response. The intervention query evaluates the candidate response using compact task and current-state context,

while token-level targets are conditioned on the student's full interaction history. This separates response assessment from the next-token predictions used for distillation.

Motivated by these observations, and illustrated in Fig. 1, we propose SAGE-OPD, a selective intervention framework for multi-turn OPD. Before executing each student response, SAGE-OPD uses validity checks and teacher assessment of the available interaction context to determine its intervention weight. It then weights the selected turns by teacher predictive decisiveness and normalizes the resulting token weights. The full training procedure is presented in Algorithm 1.

2.3 TURN-LEVEL INTERVENTION

To allocate turn-level supervision without a task verifier, we introduce a discrete intervention variable $z_t \in \{\text{SKIP}, \text{WEAK}, \text{STRONG}\}$, which determines the intervention weight at turn t. Each label is mapped to a scalar weight:

$$i_t(h_t, a_t) = \begin{cases} 0, & z_t = \text{SKIP}, \\ \alpha, & z_t = \text{WEAK}, \\ 1, & z_t = \text{STRONG}, \end{cases} \tag{2}$$

where $\alpha \in (0, 1)$ controls the strength of weak intervention. The SKIP label suppresses OPD at the current turn, while WEAK and STRONG assign partial and unit intervention weights, respectively, before confidence weighting and normalization.

We assign z_t before action execution through deterministic validity checks followed, when needed, by teacher assessment. First, if the student response fails deterministic checks, such as a missing executable action, an unparseable response, or an action-format violation, we assign $z_t = \text{STRONG}$, i.e., $i_t = 1$, without an additional teacher query. These cases bypass teacher judgment because deterministic checks already identify an invalid response.

For all remaining turns, we query the teacher with a proactive intervention-evaluation prompt before executing the student's action (see Appendix E). Given the task goal, current interaction context, and candidate student response, the teacher assigns one of $\{\text{SKIP}, \text{WEAK}, \text{STRONG}\}$. The resulting label is converted into i_t using Eq. (2). Intuitively, SKIP is used when the response is acceptable or when multiple strategies are plausible; WEAK is used when the response may be suboptimal but the need for correction is uncertain; and STRONG is used when the response is judged clearly detrimental. These categories represent the assessed need for intervention rather than turn difficulty: a difficult turn may receive SKIP if the response is acceptable, whereas a simple turn may receive STRONG if it violates the required action format.

Although the same teacher supplies intervention judgments and token-level targets, the intervention query assesses the candidate response in the available task and state context, providing evidence beyond next-token prediction. It therefore guides supervision allocation without independently verifying teacher targets; errors in the two outputs may remain correlated. On selected turns, the labels scale the distillation loss while retaining the teacher's distributional targets.

2.4 TEACHER CONFIDENCE WEIGHTING

The intervention label determines the assessed need for teacher supervision at a given turn, but does not characterize the teacher's predictive distribution used for distillation. We introduce a second factor, c_t, to modulate supervision according to the teacher's distributional decisiveness on student-generated prefixes. This factor assigns greater relative weight to selected turns on which the teacher expresses a stronger next-token preference, without assuming that a sharper distribution guarantees a correct target.

Specifically, for each token position i in the student response a_t, let $p^\star_{t,i} = \max_{v \in \mathcal{V}} \pi_\phi(v \mid h_t, a_{t,<i})$ denote the teacher's top-1 probability over the vocabulary $\mathcal{V}$. We define the turn-level teacher confidence score as the average top-1 probability across the student response:

$$c_t(h_t, a_t) = \frac{1}{|a_t|} \sum_{i=1}^{|a_t|} p^\star_{t,i} \in (0, 1]. \tag{3}$$

The intervention decision and confidence score serve distinct roles: i_t captures the assessed need for correction, whereas c_t measures the teacher's average predictive decisiveness. Because c_t averages

over the complete response, it reflects both action-relevant and formatting tokens rather than isolating confidence in the executable action. The NoConf and confidence-only ablations in §4.2 isolate confidence and intervention, respectively, under the same normalization rule. We combine the intervention weight i_t with the teacher confidence score c_t to obtain the final selective OPD weight:

$$s_t = i_t \cdot c_t \tag{4}$$

The same turn-level weight s_t is applied uniformly to all token-level OPD terms within turn t, so confidence modulates the relative contribution of selected turns rather than selecting individual thought, formatting, or action tokens. This scalar weighting leaves the teacher's target distributions unchanged, retaining their soft probability information on supervised turns. However, downweighting reduces the contribution of this information, and skipping a turn removes its direct distillation signal. The objective therefore trades uniform coverage of teacher supervision for selective emphasis on turns assessed as requiring intervention. High teacher confidence alone does not establish this need when the student's response is already acceptable. The normalization introduced below preserves the average effective token weight but does not recover supervision omitted by selection. Since $p^{\star}_{t,i}$ is obtained from the same teacher distribution used for the standard distillation loss, computing c_t requires no additional teacher forward passes. The additional teacher query for intervention labeling is described in §2.3.

2.5 Loss Normalization

The selective weight $s_t \in [0, 1]$ controls the contribution of teacher supervision at each token of the student response in turn t. Applying these weights directly to the OPD loss (Eq. (1)) changes the relative contribution of different turns while reducing the average effective token weight. To separate selective allocation from this reduction, we normalize the weights over each optimizer batch. This preserves their relative values while restoring unit average effective token weight whenever the weighted token count exceeds the numerical stability threshold.

For normalization, we index all turns in the current optimizer batch $\mathcal{B}$ by $t = 0, \ldots, T_{\mathcal{B}} - 1$ and define $N_{\mathcal{B}} = \sum_{t=0}^{T_{\mathcal{B}}-1} |a_t|$ as the total number of student-response tokens included in the loss. Let $\ell^{\mathrm{OPD}}_{t,j}$ denote the distillation loss at token position j of turn t. We define the normalized selective OPD objective as

$$\mathcal{L}_{\text{SAGE-OPD}} = \frac{Z_{\mathcal{B}}}{N_{\mathcal{B}}} \sum_{t=0}^{T_{\mathcal{B}}-1} s_t \sum_{j=1}^{|a_t|} \ell^{\mathrm{OPD}}_{t,j}, \tag{5}$$

where the shared normalization factor is $Z_{\mathcal{B}} = \frac{N_{\mathcal{B}}}{\max\left(\sum_{t=0}^{T_{\mathcal{B}}-1} s_t |a_t|, \epsilon\right)}$. Here, $|a_t|$ counts the student-response tokens included in the loss for turn t, and ϵ is a small constant for numerical stability. When the weighted token count exceeds ϵ, normalization sets the average effective token weight to one while preserving the relative turn-level weighting induced by s_t. Thus, SAGE-OPD preserves the average token weight of standard OPD, while the loss and gradient still depend on which turns receive supervision. The analysis in Appendix B characterizes this relative reweighting and bounds the gap to a normalized oracle objective in terms of intervention and supervision-strength errors. Empirical diagnostics of normalization factors and gradient norms are provided in Appendix D.8.

3 Experiments

3.1 Experimental Setup

Models. Following prior work on multi-turn OPD (Wang et al., 2026a), we use teacher and student models from the same model family to reduce confounding effects from tokenizer, architecture, and instruction-format mismatch. Specifically, we use Qwen3-32B and Qwen3-8B as teacher models, and Qwen3-0.6B and Qwen3-1.7B as student models (Yang et al., 2025a).

Benchmarks and Evaluation Metrics. We evaluate SAGE-OPD on three multi-turn agent benchmarks: ALFWorld (Shridhar et al., 2020), ScienceWorld (Wang et al., 2022), and SearchQA (Jin et al., 2025). These benchmarks cover embodied interaction, scientific task solving, and search-based question answering, respectively, allowing us to assess the effectiveness of selective intervention

across different forms of multi-turn decision making. For ALFWorld, we report success rate (SR) on both the seen and unseen splits. For ScienceWorld, we report success rate (SR) and the mean score on the test set. For SearchQA, we report exact match (EM) across seven question-answering datasets. Detailed definitions of the evaluation metrics are provided in Appendix C.

Baselines. We compare SAGE-OPD with off-policy SFT, standard OPD (Lu & Lab, 2025), TCOD-F2B (Wang et al., 2026a), Entropy-Aware OPD (Jin et al., 2026), Guided-OPD (Li et al., 2026a), and SOD (Zhong et al., 2026). TCOD-F2B progressively expands rollout depth, while Entropy-Aware OPD weights token-level supervision by teacher entropy. Guided-OPD mixes teacher and student turns with decreasing teacher participation. SOD weights step-level distillation by student–teacher divergence and includes reward-based training in our configuration. To assess whether additional rollout sampling accounts for the gains, we also include OPD with two rollouts per training example (OPD rollout 2). This baseline allocates additional computation to standard OPD, while SAGE-OPD uses additional teacher queries to assess the need for intervention. Because rollout generation and intervention assessment incur different costs, this comparison does not imply an exact compute match.

Training Details. The main experiments train each method for one epoch with one rollout per training example, except for the OPD rollout-2 and SOD baselines. For SAGE-OPD, we set the weak-intervention coefficient to $\alpha = 0.5$. We analyze the sensitivity to this choice in §4.3. Additional details are provided in Appendix C. On ALFWorld with a 1.7B student, SAGE-OPD uses 15.5% and 21.5% more GPU-hours than OPD with 8B and 32B teachers, respectively (Appendix D.9).

3.2 Main Results

We present ALFWorld and ScienceWorld results in Table 1, with trajectory-length results and SearchQA results (Table 5) in Appendix D.

Compared with standard OPD, SAGE-OPD achieves higher ALFWorld unseen SR in three of four configurations and higher ScienceWorld mean scores in all four. With the Qwen3-8B teacher, ALFWorld unseen SR increases from 56.72% to 61.94% for Qwen3-0.6B and from 61.94% to 70.15% for Qwen3-1.7B. On ScienceWorld, Qwen3-1.7B SR increases from 4.70% to 6.71% with the Qwen3-8B teacher and from 4.03% to 7.38% with the Qwen3-32B teacher. For Qwen3-0.6B distilled from Qwen3-8B, ScienceWorld mean score increases from -12.88 to $+3.89$, indicating better partial task progress despite unchanged SR. SearchQA gains are smaller: Qwen3-1.7B average EM increases from 30.09% to 30.77% with the Qwen3-8B teacher, a 0.68-point gain.

Among the methods in Table 1, SAGE-OPD achieves the best or tied-best unseen SR in three of four ALFWorld configurations and the best SR in three of four ScienceWorld configurations. With the Qwen3-8B teacher, SAGE-OPD exceeds both Guided-OPD and SOD in ALFWorld unseen SR by 11.19 percentage points for Qwen3-0.6B and 4.48 points for Qwen3-1.7B. With the Qwen3-32B teacher, SAGE-OPD matches SOD on Qwen3-0.6B unseen SR (70.90%), while Guided-OPD and SOD perform better for Qwen3-1.7B. On ScienceWorld, SAGE-OPD exceeds both methods in SR across all four configurations; its mean score exceeds Guided-OPD in three configurations and SOD in the two Qwen3-8B teacher configurations. These comparisons support response-conditioned supervision allocation as a competitive approach alongside scheduled teacher participation and divergence-based weighting.

SAGE-OPD also achieves higher ALFWorld unseen SR than OPD rollout-2 in all four configurations, indicating that additional rollout sampling alone does not match its performance under the evaluated settings. On SearchQA, SAGE-OPD achieves the best average EM among the evaluated methods for Qwen3-1.7B and remains competitive for Qwen3-0.6B. Appendix D.1 summarizes paired task-level statistical evidence and across-seed robustness, with three-seed results detailed in Appendix D.10.

Effects of Student and Teacher Scale. With Qwen3-1.7B, SAGE-OPD achieves the best ScienceWorld SR under both teacher sizes and the best average SearchQA EM among the evaluated methods. For Qwen3-0.6B, it improves ALFWorld unseen SR and ScienceWorld mean score over OPD under both teacher sizes. Larger teachers do not uniformly strengthen its relative advantage: with Qwen3-1.7B and Qwen3-32B, Entropy-Aware OPD, Guided-OPD, and SOD achieve higher ALFWorld unseen SR. The benefits therefore depend on the student–teacher configuration and

Table 1: Main results on ALFWorld and ScienceWorld. We report ALFWorld seen/unseen success rates (SR), and ScienceWorld SR and mean task score. SR is reported in percentages. Best and second-best distinct values within each student–teacher configuration are bold and underlined, respectively, including ties.

Student	Teacher	Method	ALFWorld		ScienceWorld	
			Seen SR ↑	Unseen SR ↑	SR ↑	Score ↑
0.6B	8B	SFT	2.86	4.48	2.01	−31.84
		OPD	55.00	56.72	2.01	−12.88
		OPD (rollout 2)	56.43	59.70	2.01	−17.56
		Entropy-Aware OPD	54.29	54.48	**3.36**	−11.52
		TCOD-F2B	56.43	61.19	**3.36**	−2.07
		Guided-OPD	55.00	50.75	0.67	−19.75
		SOD	46.43	50.75	0.67	−19.70
		SAGE-OPD	**57.86**	**61.94**	2.01	**+3.89**
1.7B	8B	SFT	25.00	26.12	2.01	−17.56
		OPD	60.00	61.94	4.70	−2.78
		OPD (rollout 2)	62.86	64.93	2.01	−7.86
		Entropy-Aware OPD	63.57	65.67	3.36	−3.13
		TCOD-F2B	**64.29**	58.21	4.70	−9.68
		Guided-OPD	60.00	65.67	3.36	−8.93
		SOD	60.00	65.67	3.36	−7.99
		SAGE-OPD	**64.29**	**70.15**	**6.71**	**+0.24**
0.6B	32B	SFT	–	–	2.01	−34.59
		OPD	53.57	68.66	3.36	−28.46
		OPD (rollout 2)	52.14	70.15	3.36	−29.44
		Entropy-Aware OPD	48.57	67.91	4.70	−33.10
		TCOD-F2B	**54.29**	67.91	2.68	−33.42
		Guided-OPD	50.71	70.15	0.00	−21.87
		SOD	52.14	**70.90**	2.01	**−21.23**
		SAGE-OPD	**54.29**	**70.90**	**6.04**	−26.10
1.7B	32B	SFT	–	–	2.01	−15.61
		OPD	58.57	74.63	4.03	−29.89
		OPD (rollout 2)	62.14	72.39	6.04	−28.36
		Entropy-Aware OPD	60.00	**79.10**	3.36	−14.99
		TCOD-F2B	57.14	73.13	6.04	−14.89
		Guided-OPD	**63.57**	78.36	3.36	−30.48
		SOD	57.14	76.12	2.01	**−7.63**
		SAGE-OPD	62.86	73.88	**7.38**	−16.01

task. Three-seed comparisons with the 32B teacher also favor SAGE-OPD in mean ALFWorld unseen SR and ScienceWorld score for both students (Appendix D.10); training dynamics appear in Appendix D.6.

4 Ablation Studies and Analysis

4.1 Intervention and Confidence Analysis

Fig. 2 in Appendix D examines intervention patterns across training and turn positions. As training progresses, SKIP becomes more frequent while WEAK and STRONG interventions decrease, indicating that the teacher assesses more student responses as acceptable. Across turn positions, SKIP becomes less frequent and DET-FAIL becomes more common, consistent with increasing interaction difficulty in later turns. The gate therefore retains intervention on detected failures rather than uniformly downweighting later turns.

We assess intervention labels using both episode outcomes and independent per-turn expert recommendations. Across 7,026 ALFWorld training episodes, successful trajectories have a higher SKIP rate than failed trajectories within the same games (66.3% versus 51.7%; Welch's $t = 5.56$,

Table 2: Ablation study on ALFWorld. SR is in percent. Bold and underlined values mark the best and second-best distinct SR within each student–teacher configuration, including ties.

Student	Teacher	Method	Seen SR ↑	Unseen SR ↑
Qwen3-0.6B	Qwen3-8B	OPD	55.00	56.72
		Env-only gate	<u>55.71</u>	60.45
		Intervention-only	**57.86**	**62.69**
		SAGE-OPD-NoNorm	<u>55.71</u>	55.22
		Confidence-only	<u>55.71</u>	57.46
		SAGE-OPD-NoConf	<u>55.71</u>	58.21
		SAGE-OPD-Shuffled	51.43	53.73
		SAGE-OPD	**57.86**	<u>61.94</u>
Qwen3-1.7B	Qwen3-8B	OPD	60.00	61.94
		Env-only gate	60.71	61.19
		Intervention-only	59.29	66.42
		SAGE-OPD-NoNorm	**66.43**	64.93
		Confidence-only	**66.43**	<u>68.66</u>
		SAGE-OPD-NoConf	60.71	63.43
		SAGE-OPD-Shuffled	56.43	63.43
		SAGE-OPD	<u>64.29</u>	**70.15**
Qwen3-0.6B	Qwen3-32B	OPD	53.57	68.66
		Intervention-only	53.57	67.16
		SAGE-OPD-NoNorm	**57.86**	<u>69.40</u>
		Confidence-only	53.57	65.67
		SAGE-OPD-NoConf	53.57	67.91
		SAGE-OPD-Shuffled	52.86	66.42
		SAGE-OPD	<u>54.29</u>	**70.90**
Qwen3-1.7B	Qwen3-32B	OPD	58.57	**74.63**
		Intervention-only	60.71	73.13
		SAGE-OPD-NoNorm	<u>61.43</u>	73.13
		Confidence-only	**62.86**	**74.63**
		SAGE-OPD-NoConf	59.29	72.39
		SAGE-OPD-Shuffled	52.86	70.90
		SAGE-OPD	**62.86**	<u>73.88</u>

$p = 7.2 \times 10^{-8}$). For per-turn validation, we replay 134 unseen episodes and compare student actions with ALFWorld's built-in expert recommendations at the states reached by the student. Among $2{,}208$ turns with usable expert actions, SKIP turns agree with the expert more frequently than STRONG turns (12.7% versus 4.3%; odds ratio 3.28). These complementary analyses associate intervention labels with trajectory quality and expert behavior, while exact expert agreement does not capture every acceptable alternative (Appendix D.2).

Confidence weighting improves performance over the NoConf variant (§4.2); supplementary analysis associates confidence with task-completing actions beyond intervention labels and turn position (Appendix D.4).

4.2 ABLATION STUDIES

Table 2 examines whether supervision benefits from response-specific intervention labels, confidence weighting, and normalization. The shuffled-label control tests the assignment of intervention weights to responses, while SAGE-OPD-NoConf isolates confidence weighting under the same normalization rule. ENV-ONLY supervises deterministic failures, while INTERVENTION-ONLY additionally uses teacher labels without confidence or normalization. SAGE-OPD-NoNorm removes normalization; SAGE-OPD-NoConf removes confidence and recomputes normalization using i_t. Confidence-only sets $i_t = 1$ for every turn, retaining confidence weighting and the same normalization rule. SAGE-OPD-Shuffled redistributes teacher-assigned labels across turns, retaining each turn's confidence and token-level targets, preserving deterministic-failure handling, and recomputing normalization.

Intervention and label assignment. With the Qwen3-8B teacher, intervention-only improves unseen SR over environment-only gating by 2.24 and 5.23 percentage points for the 0.6B and 1.7B students, respectively. The full objective also outperforms SAGE-OPD-Shuffled on both splits in all four configurations, with unseen gains of 2.98–8.21 percentage points. Compared with normalized confidence-only weighting, the full objective improves unseen SR by 4.48, 1.49, and 5.23 points for 0.6B/8B, 1.7B/8B, and 0.6B/32B (student/teacher), respectively, while trailing by 0.75 points for 1.7B/32B. These results support intervention beyond confidence alone in three of four configurations.

Confidence weighting. With normalization retained, the full objective outperforms SAGE-OPD-NoConf on both splits in all four configurations. Unseen gains are 3.73 and 6.72 percentage points for the 0.6B and 1.7B students with the 8B teacher, and 2.99 and 1.49 points with the 32B teacher. These results support confidence weighting as an additional supervision signal in the reported runs.

Normalization and component interactions. The full objective improves unseen SR over SAGE-OPD-NoNorm in all four configurations, including a 5.22 percentage-point gain for Qwen3-1.7B with the Qwen3-8B teacher. It also outperforms intervention-only on unseen SR in three of four configurations, while intervention-only retains a 0.75-point advantage for Qwen3-0.6B with the Qwen3-8B teacher. On ScienceWorld, the full objective improves mean score over intervention-only for both student sizes (Appendix D.5).

Continuation after interaction errors. Using Qwen3-1.7B students distilled from Qwen3-8B, OPD and SAGE-OPD continue from identical states and histories for 60 ALFWorld instances, with four rollouts per instance, method, and condition. The error condition adds an invalid action and "Nothing happens." feedback. SAGE-OPD improves post-handoff success from 35.0% to 40.0% on clean histories and from 30.0% to 33.3% on error histories, reducing subsequent invalid actions in the latter condition from 0.18 to 0.08 per rollout. These results support higher continuation success from shared states, including after unsuccessful interactions (Appendix D.7).

4.3 Sensitivity Analysis

We analyze the sensitivity of SAGE-OPD in terms of weak-intervention weight α (Eq. (2)), intervention-judge temperature, and normalization on ALFWorld with a 1.7B student and 8B teacher. As shown in Table 8 in Appendix D, the default weak-intervention weight $\alpha = 0.5$ achieves the best seen and unseen SR. A smaller value, $\alpha = 0.3$, remains competitive but underperforms on the unseen split, suggesting that weak interventions still provide useful supervision. Larger values degrade performance and increase the average number of turns. As α approaches 1, weak and strong labels receive similar intervention weights, reducing the distinction between uncertain and clear correction needs. This suggests that ambiguous turns benefit from partial rather than full teacher correction, supporting our use of a moderate weak-intervention weight.

Deterministic intervention judgments (temperature 0.0) perform best; higher temperatures reduce SR, consistent with stochastic labels weakening supervision allocation. Counting only active, non-skipped tokens in the normalization numerator remains competitive at 68.66% unseen SR but underperforms the default all-token normalization at 70.15%. Together, these comparisons support moderate weak-intervention weights, deterministic judgments, and all-token normalization.

5 Conclusion

We introduced SAGE-OPD, a framework that uses response-level assessment to allocate token-level supervision in multi-turn OPD. Experiments demonstrate gains over standard OPD on ALFWorld and ScienceWorld, alongside competitive performance on SearchQA. Ablations support response-specific intervention assignment and confidence weighting in the reported configurations, while also revealing interactions with normalization. Independent expert-action comparisons associate intervention labels with expert behavior, and shared-state evaluations show higher continuation success under both clean and injected-error conditions. Together, these findings support supervision allocation as a useful design choice for learning from student-induced interaction trajectories.

AI USE STATEMENT

Generative AI tools were used for language polishing, proofreading, and assistance with coding during the preparation of this work. All AI-assisted text and code were reviewed, verified, and, where necessary, revised by the authors. We take responsibility for the final content of this work, including all text, claims, analyses, code, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies training methods for multi-turn LLM agents using public benchmark environments and does not introduce new datasets containing personal or sensitive information. SAGE-OPD relies on teacher judgments, so biases or unsafe preferences in the teacher model may be propagated to the student. Practical deployment should therefore include appropriate action validation, safety constraints, logging, and human oversight.

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.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, volume 2025, pp. 50466–50494, 2025.

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

Zhaorun Chen, Zhuokai Zhao, Kai Zhang, Bo Liu, Qi Qi, Yifan Wu, Tarun Kalluri, Sara Cao, Yuanhao Xiong, Haibo Tong, et al. Scaling agent learning via experience synthesis. *arXiv preprint arXiv:2511.03773*, 2025.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 52690–52717, 2024.

Guangya Hao, Yitong Shang, Yunbo Long, Zhuokai Zhao, and Hanxue Liang. Self-policy distillation via capability-selective subspace projection. *arXiv preprint arXiv:2605.22675*, 2026.

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

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

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.

Woogyeol Jin, Taywon Min, Yongjin Yang, Swanand Ravindra Kadhe, Yi Zhou, Dennis Wei, Nathalie Baracaldo, and Kimin Lee. Entropy-aware on-policy distillation of language models. *arXiv preprint arXiv:2603.07079*, 2026.

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, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. Beyond single-turn: A survey on multi-turn interactions with large language models. *arXiv preprint arXiv:2504.04717*, 2025.

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.

Jun Liu, Zhenglun Kong, Peiyan Dong, Changdi Yang, Tianqi Li, Hao Tang, Geng Yuan, Wei Niu, Wenbin Zhang, Pu Zhao, et al. Structured agent distillation for large language model. *arXiv preprint arXiv:2505.13820*, 2025.

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

Feng Luo, Yu-Neng Chuang, Guanchu Wang, Zicheng Xu, Xiaotian Han, Tianyi Zhang, and Vladimir Braverman. Demystifying opd: Length inflation and stabilization strategies for large language models. *arXiv preprint arXiv:2604.08527*, 2026.

Michael Luo, Naman Jain, Jaskirat Singh, Sijun Tan, Ameen Patel, Qingyang Wu, Alpay Ariyak, Colin Cai, Tarun Venkat, Shang Zhu, Ben Athiwaratkun, Manan Roongta, Ce Zhang, Li Erran Li, Raluca Ada Popa, Koushik Sen, and Ion Stoica. Deepswe: Training a state-of-the-art coding agent from scratch by scaling rl. `https://pretty-radio-b75.notion.site/DeepSWE-Training-a-Fully-Open-sourced-State-of-the-Art-Coding-Agent-by-Scaling-RL-` 2025. Notion Blog.

Matilda QR Pembury Smith and Graeme D Ruxton. Effective use of the mcnemar test. *Behavioral Ecology and Sociobiology*, 74(11):133, 2020.

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

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

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

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, et al. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025.

Zeyu Teng, Yong Song, Xiaozhou Ye, and Ye Ouyang. Fine-tuning llms for multi-turn dialogues: optimizing cross-entropy loss with kl divergence for all rounds of responses. In *Proceedings of the 2024 16th International Conference on Machine Learning and Computing*, pp. 128–133, 2024.

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*, 2026a.

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.

Yinjie Wang, Xuyang Chen, Xiaolong Jin, Mengdi Wang, and Ling Yang. Openclaw-rl: Train any agent simply by talking. *arXiv preprint arXiv:2603.10165*, 2026b.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, et al. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7920–7939, 2025.

Robert F Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pp. 1–3, 2007.

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*, 2025a.

Chenghao Yang, Lin Gui, Chenxiao Yang, Victor Veitch, Lizhu Zhang, and Zhuokai Zhao. Let it calm: Exploratory annealed decoding for verifiable reinforcement learning. *arXiv preprint arXiv:2510.05251*, 2025b.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652, 2024.

Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. Agentracer: Who is inducing failure in the llm agentic systems? *arXiv preprint arXiv:2509.03312*, 2025.

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

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, Mingrui Zhang, Ke Li, Mingyi Wang, Qiao Liu, Qifei Wang, Jiayi Liu, Fei Liu, Serena Li, Weiwei Li, et al. Mixture-of-minds: Multi-agent reinforcement learning for table understanding. *arXiv preprint arXiv:2510.20176*, 2025a.

Yuhang Zhou, Jing Zhu, Shengyi Qian, Zhuokai Zhao, Xiyao Wang, Xiaoyu Liu, Ming Li, Paiheng Xu, Wei Ai, and Furong Huang. Disco balances the scales: Adaptive domain-and difficulty-aware reinforcement learning on imbalanced data. *arXiv preprint arXiv:2505.15074*, 2025b.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Jiayi Liu, Xiangjun Fan, Zhuokai Zhao, and Hong Yan. Synthetic sandbox for training machine learning engineering agents. *arXiv preprint arXiv:2604.04872*, 2026a.

Yuhang Zhou, Lizhu Zhang, Yifan Wu, Mingyi Wang, Peng Bo, Jiayi Liu, Xiangjun Fan, and Zhuokai Zhao. Omniopd: Logit-free on-policy distillation via speculative verification. *arXiv preprint arXiv:2606.01476*, 2026b.

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*, 2026c.

APPENDIX

A RELATED WORK

Multi-turn LLM Agents. LLM agents extend language models from single-turn generation to interactive decision making, repeatedly observing an environment, issuing actions or tool calls, and receiving feedback (Yao et al., 2022; Yang et al., 2024; Zhou et al., 2025a; Chan et al., 2025; Chen et al., 2025; Zhou et al., 2026a). Applications include embodied interaction (Shridhar et al., 2020), web navigation (Gur et al., 2024; Team et al., 2025), tool use (Yao et al., 2024), and search-based question answering (Jin et al., 2025). These settings require reasoning over partially observable states and recovering from earlier mistakes, since an incorrect action can change subsequent observations and cause errors to compound (Zhang et al., 2025).

Common training approaches include supervised fine-tuning on teacher trajectories (Teng et al., 2024; Li et al., 2025; Zhou et al., 2025a) and reinforcement learning with verifiable rewards (Shao et al., 2024; Zhou et al., 2025b; Yang et al., 2025b). Supervised imitation provides demonstrations of reasoning and action selection but generally trains on histories collected before optimization rather than those induced by the current student. Structured Agent Distillation (Liu et al., 2025) further separates reasoning and action spans and applies component-specific supervision to preserve agent behavior. Reward-based approaches directly optimize task outcomes (Luo et al., 2025; Wei et al., 2025), but depend on available reward signals and face credit-assignment challenges when feedback is sparse across long interactions (Wang et al., 2026b). These challenges motivate distillation on student-generated trajectories with dense teacher supervision.

On-Policy Distillation. OPD mitigates the distribution mismatch of offline imitation by obtaining teacher feedback on the student's own generated trajectories (Agarwal et al., 2024; Lu & Lab, 2025; Zhou et al., 2026b). Related self-distillation approaches also use student-generated samples to transfer supervision from stronger or differently conditioned teachers (Zhao et al., 2026; He et al., 2026; Hübotter et al., 2026; Hao et al., 2026). However, recent analyses identify limitations of dense token-level supervision, including sensitivity to teacher–student distribution overlap (Li et al., 2026b) and reinforcement of undesirable local patterns (Luo et al., 2026). In multi-turn interaction, these issues are compounded by student errors that shift subsequent histories away from states where teacher guidance is reliable (Wang et al., 2026a; Li et al., 2025). Consequently, matching the student's training distribution does not by itself ensure that every teacher target provides useful supervision.

Controlling Histories and Teacher Participation. Several methods address multi-turn distillation by controlling the histories encountered during training. TCOD (Wang et al., 2026a) progressively expands rollout depth through a temporal curriculum, while TurnOPD (Zhou et al., 2026c) combines adaptive rollout-depth budgeting with progressive turn-normalized loss weighting. Guided-OPD (Li et al., 2026a) mixes teacher- and student-generated turns within rollouts and gradually reduces teacher participation through a curriculum. ReOPD (Liao et al., 2026) instead reuses pre-collected teacher trajectories as replayed prefixes, employing a step-decaying sampling schedule to balance student relevance and teacher reliability without new environment interaction during student training. FutureBridge-OPD (Chen et al., 2026) executes short teacher-generated bridges at high-disagreement states and evaluates their usefulness through subsequent student continuations. These approaches examine complementary ways to control rollout depth, prefix distributions, and teacher participation.

Selective Supervision for Agent Distillation. Another line of work changes the granularity or strength of distillation supervision. Structured Agent Distillation (Liu et al., 2025) distinguishes reasoning from action spans, whereas SOD (Zhong et al., 2026) adaptively weights supervision using step-level divergence to attenuate potentially misleading teacher signals. Entropy-aware OPD (Jin et al., 2026) uses teacher-distribution entropy to adjust token-level supervision. SAGE-OPD combines deterministic response-validity checks and teacher assessment of the candidate response in its current task context to assign turn-level intervention weights. Teacher predictive decisiveness then modulates the selected supervision, while normalization preserves its relative allocation. Our focus is response-conditioned supervision allocation on student-generated turns, rather than generating replacement teacher turns or selecting replayed prefixes. This distinguishes the intervention decision from distributional uncertainty alone: a decisive teacher prediction need not imply that an acceptable student response requires correction.

B THEORETICAL ANALYSIS

We analyze how selective weighting changes the influence of noisy teacher supervision and relates to an ideal oracle intervention objective. The analysis characterizes the weighting mechanism without assuming that teacher judgments or confidence scores are always correct. Throughout this section, we consider a fixed optimizer batch $\mathcal{B}$ containing $T_{\mathcal{B}}$ nonempty turns and treat all supervision weights as constants during optimization. Let $n_t = |a_t|$ denote the number of student-response tokens included in the loss for turn t, and let

$$\ell_t(\theta) = \frac{1}{n_t} \sum_{j=1}^{n_t} \ell_{t,j}^{\mathrm{OPD}}(\theta) \tag{6}$$

denote its mean token-level OPD loss. Define the selective weight and batch token counts as

$$s_t = i_t c_t \in [0,1], \qquad N_{\mathcal{B}} = \sum_{t=0}^{T_{\mathcal{B}}-1} n_t, \qquad A_s = \sum_{t=0}^{T_{\mathcal{B}}-1} n_t s_t. \tag{7}$$

The normalized objective is

$$\begin{aligned} \mathcal{L}_{\text{SAGE-OPD}}(\theta) &= \frac{Z_{\mathcal{B}}}{N_{\mathcal{B}}} \sum_{t=0}^{T_{\mathcal{B}}-1} n_t s_t \ell_t(\theta) = \sum_{t=0}^{T_{\mathcal{B}}-1} q_t \ell_t(\theta), \\ Z_{\mathcal{B}} &= \frac{N_{\mathcal{B}}}{\max(A_s, \epsilon)}, \qquad q_t = \frac{n_t s_t}{\max(A_s, \epsilon)}. \end{aligned} \tag{8}$$

When $A_s \geq \epsilon$, the coefficients satisfy $\sum_t q_t = 1$, so the objective is a weighted mean of per-turn mean losses. When $A_s = 0$, all coefficients are zero and the distillation objective vanishes.

Selective Reweighting of Supervision Noise. Suppose the observed per-turn mean loss decomposes as

$$\ell_t(\theta) = \ell_t^\star(\theta) + \xi_t, \tag{9}$$

where $\ell_t^\star$ is the ideal supervision loss and ξ_t represents residual supervision noise. Conditioned on the batch context and assigned weights, assume $\mathbb{E}[\xi_t] = 0$ and $\mathrm{Var}[\xi_t] \leq \sigma_t^2$. All expectations and variances below are understood under this conditioning. This is an analytical assumption rather than a claim that teacher errors are generally unbiased.

Proposition B.1 (Relative Reweighting of Supervision Noise)**.** *The variance of the weighted noise term for turn t satisfies*

$$\mathrm{Var}[q_t \xi_t] \leq q_t^2 \sigma_t^2. \tag{10}$$

For two turns p and q with $n_q s_q \sigma_q > 0$, the corresponding variance bounds satisfy

$$\frac{q_p^2 \sigma_p^2}{q_q^2 \sigma_q^2} = \frac{n_p^2 s_p^2 \sigma_p^2}{n_q^2 s_q^2 \sigma_q^2}. \tag{11}$$

Relative to uniform token weighting, selective weighting changes this ratio by $(s_p/s_q)^2$ whenever the reference ratio is positive.

Proof. The noisy component of the objective is

$$\sum_{t=0}^{T_{\mathcal{B}}-1} q_t \xi_t. \tag{12}$$

Because the coefficients are fixed under the stated conditioning,

$$\mathrm{Var}[q_t \xi_t] = q_t^2 \mathrm{Var}[\xi_t] \leq q_t^2 \sigma_t^2. \tag{13}$$

The common denominator $\max(A_s, \epsilon)^2$ cancels when taking the ratio of two bounds. Uniform token weighting assigns coefficient $n_t / N_{\mathcal{B}}$ to turn t, giving the reference ratio $n_p^2 \sigma_p^2 / (n_q^2 \sigma_q^2)$. Comparing the two ratios yields the factor $(s_p/s_q)^2$. □

Remark. The proposition describes individual weighted noise terms rather than guaranteeing a reduction in total objective variance. The total variance also contains cross-turn covariances:

$$\mathrm{Var}\left[\sum_t q_t \xi_t\right] = \sum_t q_t^2 \mathrm{Var}[\xi_t] + 2\sum_{p<q} q_p q_q \mathrm{Cov}(\xi_p, \xi_q). \tag{14}$$

Normalization can increase the coefficients of retained turns relative to uniform token weighting. Nevertheless, it preserves the relative reweighting induced by s_t: assigning a smaller weight to an unreliable turn reduces its noise influence relative to otherwise comparable turns receiving larger weights. The benefit therefore depends on whether the assigned weights distinguish useful supervision from unreliable supervision.

Connection to Oracle Intervention. Let $m_t^\star \in \{0, 1\}$ indicate whether teacher intervention is useful at turn t, and let $r_t^\star \in [0, 1]$ denote the ideal supervision strength. Define the oracle weight, weighted token count, and normalized objective as

$$\begin{aligned} w_t^\star &= m_t^\star r_t^\star, \qquad A_\star = \sum_{t=0}^{T_{\mathcal{B}}-1} n_t w_t^\star, \\ \mathcal{L}_{\mathrm{oracle}}(\theta) &= \frac{\sum_{t=0}^{T_{\mathcal{B}}-1} n_t w_t^\star \ell_t(\theta)}{\max(A_\star, \epsilon)}. \end{aligned} \tag{15}$$

The assigned weighting rule approximates $w_t^\star$ using $s_t = i_t c_t$, where i_t expresses the assessed need for intervention and c_t measures teacher decisiveness.

Proposition B.2 (Normalized Selective-Weight Approximation Error)**.** *Assume* $0 \le \ell_t(\theta) \le L_{\max}$ *and* $A_s, A_\star \ge \epsilon$*. Define*

$$E_{\mathcal{B}} = \sum_{t=0}^{T_{\mathcal{B}}-1} n_t \left[c_t |i_t - m_t^\star| + m_t^\star |c_t - r_t^\star|\right]. \tag{16}$$

Then

$$|\mathcal{L}_{\text{SAGE-OPD}}(\theta) - \mathcal{L}_{\mathrm{oracle}}(\theta)| \le \frac{2L_{\max} E_{\mathcal{B}}}{\max(A_s, A_\star)}. \tag{17}$$

Proof. For each turn,

$$s_t - w_t^\star = c_t(i_t - m_t^\star) + m_t^\star(c_t - r_t^\star). \tag{18}$$

Consequently,

$$D := \sum_t n_t |s_t - w_t^\star| \le E_{\mathcal{B}}, \qquad |A_s - A_\star| \le D. \tag{19}$$

Write $U_s = \sum_t n_t s_t \ell_t(\theta)$ and $U_\star = \sum_t n_t w_t^\star \ell_t(\theta)$. The loss bound implies $|U_s - U_\star| \le L_{\max} D$ and $U_\star \le L_{\max} A_\star$. If $A_s \ge A_\star$, then

$$\begin{aligned} \left|\frac{U_s}{A_s} - \frac{U_\star}{A_\star}\right| &\le \frac{|U_s - U_\star|}{A_s} + \frac{U_\star |A_s - A_\star|}{A_s A_\star} \\ &\le \frac{2L_{\max} D}{A_s}. \end{aligned} \tag{20}$$

The case $A_\star \ge A_s$ follows by interchanging the two objectives. Using $D \le E_{\mathcal{B}}$ gives the stated bound. □

Remark. The bound separates errors in intervention assignment from errors in supervision-strength weighting while accounting for normalization. The term $c_t|i_t - m_t^\star|$ measures disagreement between the intervention weight and the oracle intervention decision. The term $m_t^\star|c_t - r_t^\star|$ measures disagreement between teacher decisiveness and the desired supervision strength on turns where intervention is useful. For deterministic validity failures, SAGE-OPD sets $i_t = 1$; if intervention is useful on such turns, the intervention-assignment error is zero, although the supervision-strength error can remain nonzero. For other turns, teacher judgment supplies an estimate of intervention need, while confidence provides a distributional weighting signal rather than a guarantee of correctness. The intervention-label analysis in §4.1 and the ablations in §4.2 evaluate these choices empirically. The bound identifies the quantities that govern approximation quality but does not establish that the estimated weights match the oracle or improve task success in every configuration.

C Implementation Details

We implement SAGE-OPD and all baselines on top of the `verl` fully-asynchronous pipeline, using a decoupled rollouter–trainer architecture with a staleness-bounded queue. Vanilla OPD and SAGE-OPD use the same forward-KL token-level distillation loss; SAGE-OPD modifies only its turn-level weighting through intervention, confidence, and normalization. We set `trigger_parameter_sync_step` to 4 and `staleness_threshold` to 0.5. Our original experimental configuration uses AdamW with learning rate 10^{-6}, weight decay 0.1, response length 4096, global batch size 64, and one training epoch. The default rollout count is $n = 1$, except for OPD rollout-2 ($n = 2$) and SOD ($n = 4$). We use the training split of each benchmark; for SearchQA (Jin et al., 2025), we randomly sample $10{,}000$ training examples.

The maximum number of turns is 30 for ALFWorld and ScienceWorld and 16 for SearchQA. The original configuration specifies a maximum generation length of 4096 tokens per turn; the additional-run settings are detailed below. We use student temperature 0.4, top-$p = 1.0$, and top-$k = -1$. Evaluations use the ReAct prompt format with full chat history, no sliding window, thinking mode disabled, and `</action>` as the stop string. For SearchQA, the agent is allowed at most 8 search calls per episode and retrieves the top 3 documents from a Wikipedia-2018 index using dense E5 retrieval with FAISS-Flat. We evaluate on the standard ALFWorld valid-seen and valid-unseen splits, the ScienceWorld test set with 30 task types and 5 variations per task, and capped SearchQA splits with up to 500 questions per split except Bamboogle, which contains 125 questions.

We evaluate the final checkpoint for these additional comparisons, without validation-based checkpoint selection or early stopping. Intermediate checkpoints saved during training are not used to select the reported endpoint.

Baseline configurations. We implement the agentic OPD baselines within the same training backbone, incorporating their respective rollout-generation and supervision rules. All methods use the same benchmark-specific evaluation protocol. TCOD-F2B (Wang et al., 2026a) truncates each rollout after

$$k(u) = \min\left(k_{\text{start}} + \frac{u}{\eta}, T\right) \tag{21}$$

assistant turns, where u denotes the curriculum step. We set $k_{\text{start}} = 1$ and $\eta = 1$, with maximum horizon $T = 30$ for ALFWorld and ScienceWorld and $T = 16$ for SearchQA. Entropy-Aware OPD augments the base OPD objective with a forward-direction distillation term at high-entropy token positions:

$$\ell_{t,j}^{\text{EOPD}} = \ell_{t,j}^{\text{OPD}} + \alpha_e \mathbf{1}[H_{\phi,t,j} > \tau_e]\, \ell_{t,j}^{\text{F}}, \tag{22}$$

where $H_{\phi,t,j}$ is the teacher entropy estimated from its cached top-K distribution and $\ell_{t,j}^{\text{F}}$ is the forward-direction distillation term. In our implementation, the base OPD loss uses the same forward-direction term, so this reduces to scaling it by $1 + \alpha_e \mathbf{1}[H_{\phi,t,j} > \tau_e]$. We use $\tau_e = 0.8$ and $\alpha_e = 1.0$.

Guided-OPD. Guided-OPD (Li et al., 2026a) generates each rollout turn with the teacher with probability β and with the student otherwise. We decrease β from 1 to 0 using a cosine schedule over the first 80% of the rollout-sample budget:

$$\xi(u) = \min\left(\frac{u}{0.8M}, 1\right), \qquad \beta(u) = \frac{1 + \cos(\pi \xi(u))}{2}, \tag{23}$$

where u is the rollout-sample counter and M is the total rollout-sample budget. The final 20% of training uses only student-generated turns. Teacher turns are inserted into the interaction trajectory, with both teacher and student generation capped at 512 tokens per turn in these training runs. Evaluation uses the student alone under the shared evaluation protocol.

At each token position, let p_v and q_v denote the student and teacher probabilities, respectively, and let $\mathcal{S}$ be the teacher's top-K token support. Our implementation uses role-dependent losses:

$$\begin{aligned} \ell_{\text{student}} &= \sum_{v \in \mathcal{S}} p_v (\log p_v - \log q_v), \\ \ell_{\text{teacher}} &= \sum_{v \in \mathcal{S}} q_v (\log q_v - \log p_v). \end{aligned} \tag{24}$$

These approximate reverse- and forward-KL supervision using truncated sums without renormalizing probabilities over $\mathcal{S}$. They are therefore distinct from full-vocabulary KL divergences, and the reverse-direction term can be negative.

SOD. SOD (Zhong et al., 2026) reweights sampled-token distillation according to step-level student–teacher divergence. For ALFWorld and ScienceWorld, each agent response between consecutive environment observations constitutes one step. Let $\mathcal{I}_k$ contain the student-token positions in step k, excluding environment tokens and padding. For a sampled student token y_j with preceding context x_j, we compute

$$\ell_j = \log \pi_\theta(y_j \mid x_j) - \log \pi_\phi(y_j \mid x_j), \qquad d_k = \frac{1}{|\mathcal{I}_k|} \sum_{j \in \mathcal{I}_k} |\ell_j|. \tag{25}$$

The step weights are

$$w_1 = 1, \qquad w_k = \min\left(\prod_{u=1}^{k-1} \frac{d_u + \epsilon}{d_{u+1} + \epsilon}, 1 + \delta \right), \tag{26}$$

with $\epsilon = 10^{-6}$ and $\delta = 0.2$. Weights are detached from the computation graph, and weighted token losses are aggregated using a token-mean reduction. Our SOD configuration adds this distillation term to a GRPO policy loss with coefficient 1.0, using native environment rewards and four rollouts per prompt. A batch of 64 trajectories therefore contains 16 prompt groups. Unlike the distillation-only configurations, SOD additionally uses task-reward supervision during training, while evaluation follows the same protocol. Step weights reuse existing student and teacher log-probabilities and require no additional teacher queries.

Selective supervision. For SAGE-OPD, the selective weight is

$$s_t = i_t \cdot c_t. \tag{27}$$

The default intervention weights are $i_t \in \{0, 0.5, 1\}$, corresponding to $\alpha = 0.5$. Turns that fail deterministic parsing or action-validity checks receive $i_t = 1$ without a teacher query. For the remaining turns, we obtain the intervention label through one teacher query before action execution, using temperature 0 and a maximum of 8 output tokens. Unparseable teacher outputs or failed teacher requests default to $i_t = 1$. The prompt uses benchmark-specific task and current-state information, as detailed in Appendix E.

Confidence is computed from teacher probabilities evaluated on student-token prefixes, reusing the scoring pass that supplies the distillation targets. Zero-token turns are excluded from confidence computation and token-level distillation. Intervention and confidence weights are treated as constants during student optimization. Normalization statistics are aggregated across data-parallel workers over the full optimizer batch before microbatching, so all microbatches share the same normalization factor. The final loss uses a token-mean reduction. An all-skipped batch contributes zero distillation loss and gradient. The confidence-only ablation fixes $i_t = 1$ for every nonempty turn, so $s_t = c_t$, and applies the same optimizer-batch normalization as SAGE-OPD; teacher intervention labels do not affect its loss weights. We apply global gradient-norm clipping with threshold 1.0.

Evaluation Metrics. We use benchmark-specific evaluation metrics. For ALFWorld, we report success rate (SR) on the valid-seen and valid-unseen splits, counting an episode as successful if the environment returns `info["won"] = True` within the turn budget. For ScienceWorld, we report SR and mean score. SR measures the fraction of episodes reaching a score of 100, while mean score averages the environment-provided episode scores. Mean score complements binary success by capturing partial task progress. For SearchQA, we report exact match (EM), counting a final answer as correct if its normalized string matches any reference answer. We follow the SearchR1 normalization protocol, which lowercases text, removes punctuation and articles, and collapses whitespace. We report EM for each evaluation split and macro-average across NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, Musique, and Bamboogle. We also report average turns for ALFWorld and ScienceWorld and average search calls and turns for SearchQA as auxiliary interaction-cost metrics.

Table 3: Main results on ALFWorld. We report success rate (SR) and average number of turns on both seen and unseen splits. Best and second-best distinct SR values among the methods shown within each student–teacher configuration are bold and underlined, respectively, including ties. Reference models are excluded from ranking.

Student Model	Teacher Model	Method	Seen SR ↑	Seen Turns ↓	Unseen SR ↑	Unseen Turns ↓
Reference Models (Base Inference)						
Qwen3-0.6B	–	Base	1.43	29.65	0.75	29.82
Qwen3-1.7B	–	Base	25.71	24.63	26.12	24.72
–	Qwen3-8B	Base	61.43	17.95	67.16	17.24
–	Qwen3-32B	Base	57.14	18.94	69.40	18.21
Student: Qwen3-0.6B, Teacher: Qwen3-8B						
Qwen3-0.6B	Qwen3-8B	SFT	2.86	29.32	4.48	28.87
		OPD	55.00	18.53	56.72	20.30
		OPD (rollout 2)	<u>56.43</u>	18.76	59.70	20.31
		Entropy-Aware OPD	54.29	18.12	54.48	20.75
		TCOD-F2B	<u>56.43</u>	18.51	<u>61.19</u>	18.82
		SAGE-OPD	**57.86**	18.46	**61.94**	18.89
Student: Qwen3-1.7B, Teacher: Qwen3-8B						
Qwen3-1.7B	Qwen3-8B	SFT	25.00	24.58	26.12	24.54
		OPD	60.00	17.80	61.94	17.67
		OPD (rollout 2)	62.86	17.51	64.93	17.87
		Entropy-Aware OPD	<u>63.57</u>	17.28	<u>65.67</u>	18.20
		TCOD-F2B	**64.29**	17.65	58.21	19.16
		SAGE-OPD	**64.29**	18.14	**70.15**	17.86
Student: Qwen3-0.6B, Teacher: Qwen3-32B						
Qwen3-0.6B	Qwen3-32B	OPD	<u>53.57</u>	19.49	68.66	18.65
		OPD (rollout 2)	52.14	20.28	<u>70.15</u>	18.21
		Entropy-Aware OPD	48.57	20.87	67.91	18.28
		TCOD-F2B	**54.29**	20.18	67.91	19.26
		SAGE-OPD	**54.29**	19.85	**70.90**	17.98
Student: Qwen3-1.7B, Teacher: Qwen3-32B						
Qwen3-1.7B	Qwen3-32B	OPD	58.57	18.89	<u>74.63</u>	17.16
		OPD (rollout 2)	<u>62.14</u>	18.17	72.39	18.37
		Entropy-Aware OPD	60.00	18.80	**79.10**	16.75
		TCOD-F2B	57.14	19.42	73.13	17.87
		SAGE-OPD	**62.86**	18.75	73.88	17.57

D Supplementary Results

D.1 Statistical Analysis and Across-Seed Robustness

We assess the reliability of the ALFWorld and ScienceWorld results through paired task-level statistical tests and independent training-seed replications. The paired tests use the exact evaluation instances underlying the main results, while the replications characterize variation across training runs. Because competing methods are evaluated on the same tasks, paired analyses control for instance-level difficulty and are more appropriate than tests that treat aggregate scores as independent samples. All reported tests are two-sided, and the p-values are not corrected for multiple comparisons.

ALFWorld. ALFWorld evaluates each trajectory using a binary success outcome. We therefore apply the exact McNemar test (Pembury Smith & Ruxton, 2020) to paired task-level predictions, considering only discordant instances on which exactly one of the two methods succeeds.

For the Qwen3-1.7B student with Qwen3-8B teacher, paired comparisons favor SAGE-OPD over both OPD and TCOD-F2B. Compared with standard OPD, SAGE-OPD succeeds on 16 tasks where OPD fails, while OPD succeeds on only 5 tasks where SAGE-OPD fails. These eleven net task-level improvements correspond to an 8.21 percentage-point increase in unseen SR, from 61.94% to 70.15%, and yield an exact McNemar p-value of 0.027. Compared with TCOD-F2B, SAGE-OPD obtains 22 wins and only 6 losses among the discordant task pairs. The resulting sixteen net improvements correspond to an 11.94 percentage-point SR gain and provide the strongest ALFWorld paired result ($p = 0.004$).

The discordant-pair counts show how the aggregate SR gains arise from changes in individual task outcomes. In particular, for the Qwen3-1.7B/Qwen3-8B configuration, SAGE-OPD improves substantially more individual unseen tasks than it degrades relative to both standard OPD and TCOD-F2B. However, the Qwen3-0.6B comparisons remain statistically inconclusive at the current evaluation-set size.

Table 4: Main results on ScienceWorld. We report success rate (SR), mean task score, and average number of turns. Best and second-best distinct SR and mean-score values among the methods shown within each student–teacher configuration are bold and underlined, respectively, including ties. Reference models are excluded from ranking.

Student Model	Teacher Model	Method	SR ↑	Mean Score ↑	Turns ↓
Reference Models (Base Inference)					
Qwen3-0.6B	–	Base	3.36	-43.51	21.38
Qwen3-1.7B	–	Base	3.36	-24.00	21.38
–	Qwen3-8B	Base	10.74	-16.09	21.67
–	Qwen3-32B	Base	15.44	-17.92	18.67
Student: Qwen3-0.6B, Teacher: Qwen3-8B					
Qwen3-0.6B	Qwen3-8B	SFT	2.01	-31.84	23.94
		OPD	2.01	-12.88	23.94
		Entropy-Aware OPD	**3.36**	-11.52	23.44
		TCOD-F2B	**3.36**	-2.07	24.88
		SAGE-OPD	2.01	**+3.89**	26.94
Student: Qwen3-1.7B, Teacher: Qwen3-8B					
Qwen3-1.7B	Qwen3-8B	SFT	2.01	-17.56	23.28
		OPD	4.70	-2.78	25.12
		Entropy-Aware OPD	3.36	-3.13	25.91
		TCOD-F2B	4.70	-9.68	24.06
		SAGE-OPD	**6.71**	**+0.24**	25.30
Student: Qwen3-0.6B, Teacher: Qwen3-32B					
Qwen3-0.6B	Qwen3-32B	SFT	2.01	-34.59	18.10
		OPD	3.36	-28.46	20.79
		Entropy-Aware OPD	4.70	-33.10	23.40
		TCOD-F2B	2.68	-33.42	23.20
		SAGE-OPD	**6.04**	**-26.10**	20.00
Student: Qwen3-1.7B, Teacher: Qwen3-32B					
Qwen3-1.7B	Qwen3-32B	SFT	2.01	-15.61	23.23
		OPD	4.03	-29.89	19.64
		Entropy-Aware OPD	3.36	-14.99	23.59
		TCOD-F2B	6.04	**-14.89**	22.02
		SAGE-OPD	**7.38**	-16.01	20.91

Table 5: Main results on SearchQA. We report exact match (EM, in percent) across seven question-answering datasets and their macro-average. Best and second-best distinct EM values among the methods shown within each student–teacher configuration are bold and underlined, respectively, including ties. Reference models are excluded from ranking.

Student Model	Teacher Model	Method	NQ ↑	TriviaQA ↑	PopQA ↑	HotpotQA ↑	2WikiMQA ↑	Musique ↑	Bamboogle ↑	Avg. ↑
Reference Models (Base Inference)										
Qwen3-0.6B	–	Base	5.60	15.40	0.40	6.00	8.40	1.20	3.20	5.74
Qwen3-1.7B	–	Base	24.20	36.40	5.20	19.00	14.20	10.80	18.40	18.31
–	Qwen3-8B	Base	34.60	53.00	14.00	38.20	37.40	28.80	45.60	35.94
–	Qwen3-14B	Base	35.00	55.80	15.60	41.20	43.00	30.00	49.60	38.60
Student: Qwen3-0.6B, Teacher: Qwen3-8B										
Qwen3-0.6B	Qwen3-8B	SFT	14.40	23.60	5.80	13.20	11.69	2.23	6.40	11.05
		OPD	25.00	41.40	17.80	28.20	24.80	17.00	24.00	25.46
		Entropy-Aware OPD	**25.60**	40.00	**18.80**	28.80	**25.60**	**17.40**	**24.40**	**25.80**
		TCOD-F2B	25.40	41.00	15.60	29.00	24.60	16.80	23.20	25.09
		SAGE-OPD	25.20	**41.60**	18.40	**29.80**	23.20	15.00	24.00	25.31
Student: Qwen3-1.7B, Teacher: Qwen3-8B										
Qwen3-1.7B	Qwen3-8B	SFT	27.40	39.60	11.60	21.20	8.80	12.00	20.00	20.09
		OPD	30.60	47.20	12.40	33.20	29.60	**24.80**	32.80	30.09
		Entropy-Aware OPD	29.40	48.20	12.60	33.20	30.20	23.80	**36.80**	30.60
		TCOD-F2B	**31.20**	47.60	12.40	**33.80**	28.60	24.00	36.00	30.51
		SAGE-OPD	30.00	**48.40**	**12.80**	33.20	**30.40**	24.60	36.00	**30.77**

ScienceWorld. ScienceWorld provides both a continuous task score and a sparse binary success outcome. Because the success rate is low across methods, we focus the statistical analysis on the more informative continuous score and apply paired tests to the same 149 evaluation tasks.

The strongest ScienceWorld evidence is observed for the Qwen3-0.6B student distilled from the Qwen3-8B teacher. Compared with standard OPD, SAGE-OPD improves the mean score from -12.88 to 3.89, corresponding to an average gain of 16.77 points. The improvement is supported by both a paired t-test ($p = 0.001$) and a Wilcoxon signed-rank test ($p = 0.001$) (Woolson, 2007). At the individual-task level, SAGE-OPD obtains a higher score on 48 tasks, ties on 79, and obtains a lower score on 22. An exact sign test over the 70 non-tied tasks further supports the improvement ($p = 0.003$).

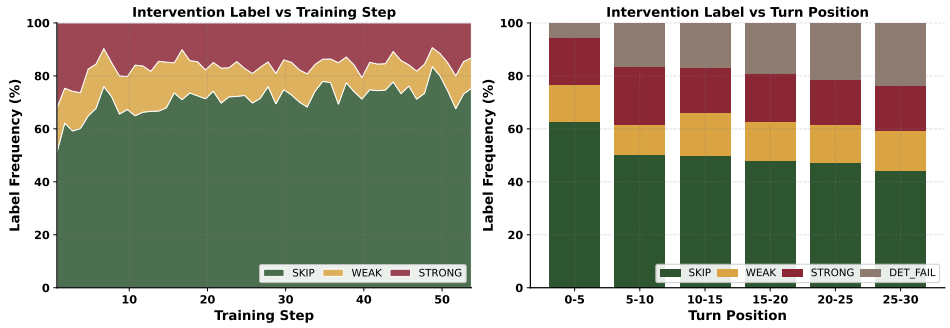

Figure 2: Intervention label distribution of SAGE-OPD on ALFWorld. Left: label frequency across training steps. Right: label frequency across turn positions. DET-FAIL denotes failures identified by deterministic parsing or action-validity checks before teacher judgment.

Relative to Entropy-Aware OPD, SAGE-OPD improves the mean ScienceWorld score by 15.41 points in the same student–teacher configuration. This difference is also supported by a paired t-test ($p = 0.003$) and a Wilcoxon signed-rank test ($p = 0.004$). Across the 149 tasks, SAGE-OPD achieves 44 wins, 79 ties, and 26 losses; the corresponding exact sign test over the non-tied tasks gives $p = 0.041$.

Overall, the paired analyses provide complementary evidence across the two benchmarks. On ALFWorld, SAGE-OPD significantly improves binary task completion in the Qwen3-1.7B/Qwen3-8B unseen setting, particularly relative to standard OPD and TCOD-F2B. On ScienceWorld, SAGE-OPD significantly improves partial task progress for the Qwen3-0.6B/Qwen3-8B setting, even though sparse binary success limits the power of success-rate comparisons. We therefore interpret these results as configuration-specific statistical support for selective intervention, rather than evidence of uniform superiority across every benchmark and student–teacher setting.

Across-seed robustness. Complementing the task-level tests, we repeat OPD and SAGE-OPD with the Qwen3-32B teacher using two additional training seeds for each student size and benchmark. Across the original run and seeds 43 and 44, mean ALFWorld unseen SR increases from 68.66% to 71.89% for Qwen3-0.6B and from 73.88% to 76.86% for Qwen3-1.7B. Mean ScienceWorld scores also improve for both students. Both additional seeds improve ALFWorld unseen SR and ScienceWorld score for each student size; ScienceWorld SR ties for Qwen3-0.6B and improves for Qwen3-1.7B. These replications provide complementary evidence of consistency across training runs, with per-run results and mean $\pm$ sample standard deviation reported in Appendix D.10.

D.2 DETAILED INTERVENTION-LABEL VALIDATION

We validate intervention labels through complementary analyses of training-episode outcomes and agreement with independent expert recommendations on unseen trajectories. The training analysis includes $7{,}026$ episodes comprising $35{,}175$ turns. Rounded label frequencies are 52% SKIP, 16% WEAK, 16% teacher-assigned STRONG, and 15% DET-FAIL. Approximately 47% of turns receive a nonzero intervention signal, including the 15% handled directly by deterministic environment/parser feedback.

Association with episode outcomes. The episode-level SKIP rate is positively associated with success under both Pearson correlation ($r = 0.036$, $p = 0.003$) and Spearman correlation ($\rho = 0.027$, $p = 0.023$). These associations are small; only 2.6% of training episodes succeed, so we additionally examine successful and failed attempts within the same games. Successful trajectories have an average SKIP rate of 0.663 ± 0.316, compared with 0.517 ± 0.394 for failed trajectories, a gap of approximately 14.6 percentage points. The difference is supported by Welch's t-test ($t = 5.56$, $p = 7.2 \times 10^{-8}$; Cohen's $d = 0.407$) and a Mann–Whitney test ($p = 1.1 \times 10^{-5}$). Restricting the comparison to games with both outcomes reduces confounding from task identity and supports an association between teacher-approved turns and successful trajectories.

Table 6: Exact agreement between student actions and expert recommendations, grouped by intervention label. Only turns with usable expert recommendations are included. DET-FAIL labels originate from environment/parser checks rather than teacher judgments.

Label	Turns	Agreements	Agreement (%)
SKIP	1,605	204	12.71
WEAK	175	2	1.14
STRONG	329	14	4.26
DET-FAIL	99	0	0.00

Among teacher-assigned STRONG turns, 97.7% occur in episodes that ultimately fail. Because failures dominate the training data, this statistic is interpreted alongside the within-game comparison rather than as standalone label precision. Episodes with more than 50% STRONG turns achieve 0.9% success, compared with 2.6% for episodes with more than 70% SKIP turns. These outcome-based analyses characterize aggregate trajectory quality rather than correctness at every individual turn.

Agreement with expert recommendations. We additionally use ALFWorld's built-in expert planner as an independent per-turn reference. For each replayed student trajectory, we obtain the expert's recommended next action from `extra.expert_plan`, compare it with the student's action, and then execute the student's action. Each subsequent comparison therefore starts from the state actually reached by the student rather than from an independently generated expert trajectory. The expert is used only for this post-hoc evaluation and does not determine training labels or distillation targets. We retain turns for which an expert recommendation is available and belongs to the admissible action set. Of $2{,}378$ turns from 134 unseen episodes, $2{,}208$ meet this criterion, giving 92.9% reference coverage.

SKIP turns have approximately three times the expert-agreement probability of STRONG turns, corresponding to an odds ratio of 3.28. An unadjusted turn-level Fisher's exact test yields $p < 0.0001$; this test does not account for dependence among turns within an episode. The comparison provides an independent per-turn association between teacher-assessed acceptability and expert behavior. Exact agreement with one recommendation is a restricted measure, since alternative actions may also be acceptable. Moreover, WEAK has lower agreement than STRONG, so the categories do not form a monotonic ordering of expert-action agreement. The zero agreement for DET-FAIL is expected when invalid actions are compared with admissible expert recommendations.

Together, the episode-level comparisons, expert-action analysis, and qualitative examples in §D.3 provide complementary evidence that intervention labels capture meaningful behavioral differences. They do not treat episode success or disagreement with a single expert recommendation as an oracle label for every student action.

D.3 CASE STUDY: TURN-LEVEL INTERVENTION EXAMPLES

To qualitatively examine how SAGE-OPD allocates teacher supervision, we present representative ALFWorld cases for different intervention outcomes. These examples illustrate that the intervention mechanism does not simply supervise every student response uniformly. Instead, it distinguishes between responses judged acceptable, mildly suboptimal, or clearly harmful, and responses that fail deterministic pre-execution validity checks. All intervention labels are assigned before the current response is executed, using the candidate response and the available pre-action context. Execution outcomes mentioned below describe subsequent observations; they are not inputs to the intervention decision for that response.

Skip: acceptable student behavior. When the student response is reasonable under the current observation, SAGE-OPD assigns SKIP and suppresses the OPD loss for that turn. For example, in a `pick_two_obj_and_place` task, the student observes the following environment state:

```
You arrive at coffeetable 1.  On the coffeetable 1, you see
a box 2, a creditcard 2, a creditcard 1, a houseplant 1, a
plate 2, a plate 1, a statue 1, and a wateringcan 1.
```

The student responds:

```
Thought: I need to find two vases and place them on the
coffeetable.
Action: look
```

The parsed action is `look`, and the teacher assigns SKIP with intervention weight $i_t = 0$. Although the action does not immediately complete the task, it is a valid exploratory step under the current observation. In another `pick_clean_then_place_in_recep` task, the student similarly takes an exploratory action:

```
Thought: I need to examine the cabinet to find a knife.
Action: examine cabinet 1
```

The teacher assigns SKIP before execution because the proposed action is judged executable and plausible in the available context. The subsequent observation that the cabinet contains nothing is an execution outcome, not the basis for this label. These cases show that SAGE-OPD avoids unnecessary token-level correction when the student is making plausible progress.

Weak: executable but suboptimal behavior. When the student action is valid but potentially inefficient or only partially aligned with the task, SAGE-OPD assigns WEAK. For instance, in a `pick_clean_then_place_in_recep` task, the student sees the same coffeetable observation and produces:

```
Thought: I need to find a clean kettle in the cabinet.
Action: examine cabinet 1
```

The action `examine cabinet 1` is executable and may still provide useful information, but it is not clearly the best next step for the task. The teacher therefore assigns WEAK with intervention weight $i_t = \alpha$. As another example, in a `pick_and_place_simple` task, the student reasons:

```
Thought: The cloth is not on the countertop. I'll check
the cabinets to find a cloth.
Action: go to cabinet 1
```

Before execution, the teacher again assigns WEAK, judging the proposed move uncertain rather than clearly optimal. The environment subsequently accepts the action and moves the agent to the closed cabinet. These examples demonstrate that SAGE-OPD can provide partial supervision for ambiguous turns rather than forcing a binary choice between full imitation and no correction.

Strong: clearly harmful or inconsistent behavior. When the student response is clearly inconsistent with the task state or likely to push the trajectory in an unhelpful direction, SAGE-OPD assigns STRONG. For example, in a `pick_cool_then_place_in_recep` task, the student states that it should cool a plate in the fridge, but then navigates to a cabinet:

```
Thought: I need to find a plate and cool it in the fridge
before putting it in the cabinet.
Action: go to cabinet 1
```

The teacher judges this cabinet-navigation action insufficiently aligned with the cooling subgoal in the recorded pre-action context. The teacher assigns STRONG with intervention weight $i_t = 1$. A related failure occurs when the student repeats the same high-level intent but selects another cabinet-navigation action:

```
Thought: I need to find a plate and cool it in the fridge
before putting it in the cabinet.
Action: go to cabinet 10
```

The teacher again assigns strong intervention to an executable action judged unhelpful in context; the action string alone does not establish whether cabinet exploration is warranted. These cases illustrate teacher-assigned strong intervention based on perceived task relevance.

Deterministic failure: pre-execution validity checks. In addition to teacher-judged intervention, SAGE-OPD assigns $z_t = \text{STRONG}$ and $i_t = 1$ when the candidate response fails a deterministic validity check before execution, as described in §2.3. These checks identify missing executable actions, unparseable responses, and action-format violations without querying the teacher. We report this pathway as DET-FAIL to distinguish it from teacher-assigned STRONG labels. If the response passes these checks, the teacher assigns its intervention label from the candidate response and available pre-action context. Subsequent feedback such as `Nothing happens.` does not determine or revise the current turn's label; it becomes part of the interaction context available at later turns. An unsuccessful execution therefore does not, by itself, establish that a response belongs to the DET-FAIL category.

Discussion. Overall, these case studies support the motivation for selective multi-turn OPD. The teacher assigns different supervision strengths to student turns according to their assessed acceptability, ambiguity, and task relevance. By combining pre-execution validity checks with teacher assessment of the candidate response and available interaction context, SAGE-OPD allocates supervision at the turn level and avoids applying dense token-level OPD uniformly across the entire trajectory. This behavior is especially important in multi-turn environments, where over-correcting acceptable exploration can reduce behavioral diversity, while failing to intervene on invalid or harmful actions can cause errors to compound in later turns.

D.4 SUPPLEMENTARY ANALYSIS OF CONFIDENCE AND TASK COMPLETION

As a supplementary behavioral analysis, we examine whether the turn-level confidence score c_t is associated with task-completing actions beyond intervention labels and turn position. Because ALFWorld provides only a sparse terminal reward, we define a task-completion turn as the final action that successfully terminates an episode. Among $2{,}514$ analyzed turns, 82 are task-completion turns (3.26%). The confidence score is computed from the history through the pre-action observation and the student's current response; it does not observe the environment feedback produced after executing that response.

We first fit an episode-clustered generalized estimating equation predicting task completion from standardized confidence, intervention labels, and turn index. We exclude 48 DET-FAIL turns because none completes the task, causing complete separation. After controlling for intervention category and turn position, a one-standard-deviation increase in c_t remains strongly associated with task completion (adjusted OR $= 5.61$, 95% CI $[3.34, 9.44]$, robust $p = 7.8 \times 10^{-11}$).

Confidence is particularly informative where it affects the training objective. Across all turns, c_t achieves an AUROC of 0.682 and an average precision of 0.048. On the 567 actively supervised turns with $i_t > 0$, these values increase to 0.875 and 0.402, respectively; within STRONG turns, they reach 0.902 and 0.551. Thus, confidence most clearly distinguishes task-completion actions on turns where it directly modulates the loss through $s_t = i_t c_t$.

We further evaluate the incremental predictive value of the two components underlying the selective weight $s_t = i_t c_t$ using episode-grouped five-fold cross-validation, with turn index included in every model. A turn-only model obtains an AUROC of 0.520 and an average precision of 0.032. Adding the intervention category z_t, which determines i_t, increases these values to 0.846 and 0.238. Jointly including teacher confidence c_t, thereby evaluating both factors used to construct s_t, further improves AUROC to 0.880 and average precision to 0.425. These results show that c_t provides additional predictive information about task-completing actions beyond intervention labels and trajectory position.

This association complements the NoConf ablation in §4.2, which directly evaluates the contribution of confidence weighting to training. The prediction target here is student task completion, not teacher-target correctness.

D.5 CROSS-BENCHMARK COMPONENT ABLATION

To examine whether these component-level findings extend beyond ALFWorld, we additionally compare OPD, intervention-only, and full SAGE-OPD on ScienceWorld with Qwen3-8B as teacher, as reported in Table 7.

Table 7: Component ablation on ScienceWorld with Qwen3-8B as teacher. Best and second-best distinct values within each student configuration are bold and underlined, respectively, including ties.

Student	Method	Mean Score ↑	SR ↑
Qwen3-0.6B	OPD	-12.88	2.01
	Intervention-only	-10.90	**3.36**
	SAGE-OPD	**+3.89**	2.01
Qwen3-1.7B	OPD	-2.78	4.70
	Intervention-only	-6.80	5.37
	SAGE-OPD	**+0.24**	**6.71**

Table 8: Sensitivity analysis on ALFWorld using Qwen3-8B as teacher and Qwen3-1.7B as student. The default configuration uses $\alpha = 0.5$, teacher temperature 0.0, and standard loss normalization over all student tokens. Best and second-best distinct SR values within each sensitivity block are bold and underlined, respectively, including ties.

Type	Setting	Seen SR ↑	Seen Turns ↓	Unseen SR ↑	Unseen Turns ↓
α	0.3	63.57	17.34	67.16	16.78
	0.5 (*default*)	**64.29**	18.14	**70.15**	17.86
	0.7	57.86	18.86	66.42	18.47
	1.0	58.57	19.31	61.94	19.02
Intervention-judge temp.	0.0 (*default*)	**64.29**	18.14	**70.15**	17.86
	0.3	60.00	18.61	62.69	18.35
	0.7	60.71	18.77	65.67	18.34
	1.0	60.71	18.24	63.43	18.11
	1.5	57.86	19.24	63.43	18.74
Normalization	All-token norm. (*default*)	**64.29**	18.14	**70.15**	17.86
	Active-token norm.	62.14	18.41	68.66	17.46

As shown in Table 7, the full objective improves mean score over intervention-only for both students and improves SR from 5.37 to 6.71 for Qwen3-1.7B, although it does not improve Qwen3-0.6B SR. Together with the ALFWorld ablations, these comparisons support the combined use of confidence weighting and normalization, with benefits that depend on the metric and student configuration.

D.6 Training Dynamics

Fig. 3 compares SAGE-OPD and OPD (Lu & Lab, 2025) on ALFWorld over the reported 55-step training budget. Both methods improve rapidly early in training and largely plateau between steps 30 and 50, with SAGE-OPD improving at the final evaluated checkpoint to 70.15% unseen SR, compared with 61.94% for OPD. Training-batch success rates follow similar trends, while the distillation loss of SAGE-OPD decreases over training. Because each method generates its own trajectories, these training-batch rates reflect performance on their respective on-policy distributions. These observations show higher final-checkpoint unseen success under the evaluated budget, without establishing a sustained advantage over longer training horizons.

D.7 Continuation from Shared States

We evaluate continuation from shared states on ALFWorld using a Qwen3-1.7B student distilled from a Qwen3-8B teacher. For each of 60 instances, we replay a fixed prefix and evaluate vanilla OPD and SAGE-OPD under two conditions: a clean continuation and a continuation following an injected invalid action. Both methods receive identical replayed states and histories within each condition, with four continuation rollouts per instance and method. Before continuation begins, the injected action produces "Nothing happens." feedback without advancing the task, and both methods receive this feedback in their shared history. This setup tests continuation after an unsuccessful interaction rather than recovery from an action that irreversibly changes the environment.



Figure 3: Training dynamics on ALFWorld comparing SAGE-OPD with vanilla OPD. Left: unseen success rate during training. Middle: training-batch success rate. Right: distillation loss of SAGE-OPD.

Table 9: Continuation performance from identical replayed states. Success is the percentage of post-handoff rollouts completing the task. Invalid actions, repetitions, and turns are mean counts per continuation.

Method	Condition	Success (%)	Invalid	Repeats	Turns
Vanilla OPD	Clean	35.0	0.12	0.10	4.7
SAGE-OPD	Clean	40.0	0.13	0.08	4.8
Vanilla OPD	Error	30.0	0.18	0.05	4.4
SAGE-OPD	Error	33.3	0.08	0.10	4.4

Metrics. We report post-handoff task completion, mean subsequent invalid-action count, mean consecutive-action repetition count, and mean continuation length. Task completion is recorded when the environment reports `won=True` or terminates with a positive reward. This measures completion from the replayed state rather than success over a full independently generated episode. Invalid actions are model-generated turns for which the parser cannot map the output to an admissible action and invokes the fallback behavior. Repetitions count consecutive identical snapped actions and do not capture longer cycles such as alternating between two actions. All action counts and rollout lengths exclude the replayed prefix and the injected error itself. The four rollouts are repeated evaluations of each instance, yielding 240 continuations per method and condition over 60 distinct instances.

As shown in Table 9, SAGE-OPD achieves higher continuation success under both clean and error conditions, by 5.0 and approximately 3.3 percentage points, respectively. Following the injected error, SAGE-OPD produces fewer subsequent invalid actions (0.08 versus 0.18) with the same average continuation length of 4.4 turns. Consecutive repetitions increase from 0.05 to 0.10 in this condition, so the behavioral improvement does not extend uniformly across all measures. The clean-to-error success decrease is approximately 6.7 percentage points for SAGE-OPD and 5.0 points for vanilla OPD. These descriptive results therefore support higher continuation success from shared starting states, including histories containing error feedback, but do not establish reduced sensitivity to the injected error.

D.8 Normalization Diagnostics

We inspect normalization factors and pre-clipping gradient norms over 55-step ALFWorld runs with a Qwen3-1.7B student. For the Qwen3-8B teacher, the normalization factor has median 4.86, 95th percentile 7.01, and maximum 9.25; the corresponding pre-clipping gradient norms are 3.28, 11.85, and 17.87. For the Qwen3-32B teacher, the normalization-factor statistics are 2.70, 3.09, and 3.29, while the gradient-norm statistics are 1.67, 10.86, and 22.33. Neither run records an all-zero-weight batch or a nonfinite loss or gradient, and both use global gradient-norm clipping at 1.0. These measurements document finite training behavior under the implemented normalization and clipping scheme.

Table 10: Training cost on ALFWorld with a Qwen3-1.7B student. All runs use eight GPUs of the same node type and 55 training steps. Step time is the median over logged steps; GPU-hours are totals.

Teacher	Method	s/step	GPU-h
Qwen3-8B	OPD	224.0	29.1
	Intervention-only	260.6	33.6
	SAGE-OPD-NoConf	263.1	33.5
	SAGE-OPD-NoNorm	268.6	34.1
	SAGE-OPD	261.8	33.6
Qwen3-32B	OPD	233.3	29.8
	SAGE-OPD	281.5	36.2

D.9 Training Cost

We measure training cost in 55-step ALFWorld runs using a Qwen3-1.7B student and eight GPUs of the same node type. GPU-hours are calculated from allocated GPU count and job elapsed time, including initialization, rollout generation, teacher queries, optimization, checkpoint writing, and waiting within the allocated job. Evaluation jobs and queue waiting time are excluded. Each comparison between SAGE-OPD and OPD uses the same teacher size and hardware configuration. Table 10 reports median time per logged training step and total GPU-hours.

With the Qwen3-8B teacher, SAGE-OPD increases median step time from 224.0 to 261.8 seconds (16.9%) and total GPU-hours from 29.1 to 33.6 (15.5%). With the Qwen3-32B teacher, median step time increases from 233.3 to 281.5 seconds (20.7%), while total GPU-hours increase from 29.8 to 36.2 (21.5%).

For the Qwen3-8B teacher, intervention-enabled variants have median step times between 260.6 and 268.6 seconds. Their similar timings are consistent with intervention querying being the principal additional cost, with no clearly distinguishable incremental latency from confidence weighting or normalization in these measurements. Computing c_t reuses the teacher distributions obtained for distillation, whereas intervention labeling requires an additional teacher-generation query on turns not handled by deterministic failure detection. These measurements reflect overall training runtime under each method's own on-policy trajectories and do not isolate intervention-query latency.

D.10 Three-Seed Replication

We evaluate training-seed variability with Qwen3-32B as teacher and Qwen3-0.6B and Qwen3-1.7B as students on ALFWorld and ScienceWorld. For each method and configuration, we combine the original run reported in Table 1 with two independent training reruns using seeds 43 and 44. The reruns use the same training settings, evaluation splits, and checkpoint-selection rule as the original experiment. Table 11 reports the mean and sample standard deviation across the three training runs; Table 12 gives the individual results.

Table 11: Three-seed results with Qwen3-32B as teacher. Values are mean ± sample standard deviation across the original run and seeds 43 and 44. SR is in percent; bold entries indicate the higher mean within each student configuration.

Student	Method	ALFWorld		ScienceWorld	
		Seen SR ↑	Unseen SR ↑	SR ↑	Score ↑
0.6B	OPD	54.76 ± 1.49	68.66 ± 0.75	2.91 ± 0.78	-25.64 ± 3.23
0.6B	SAGE-OPD	$\mathbf{57.14 \pm 2.47}$	$\mathbf{71.89 \pm 1.72}$	$\mathbf{3.80 \pm 2.05}$	$\mathbf{-23.62 \pm 2.23}$
1.7B	OPD	60.00 ± 1.89	73.88 ± 0.75	3.80 ± 1.03	-21.98 ± 6.87
1.7B	SAGE-OPD	$\mathbf{61.67 \pm 1.49}$	$\mathbf{76.86 \pm 2.69}$	$\mathbf{5.82 \pm 1.69}$	$\mathbf{-16.31 \pm 0.81}$

Across the three runs, SAGE-OPD improves mean ALFWorld unseen SR from 68.66% to 71.89% for Qwen3-0.6B and from 73.88% to 76.86% for Qwen3-1.7B. Both additional seeds favor SAGE-

OPD for both student sizes, with unseen SR gains of 1.50–5.97 percentage points. On ScienceWorld, SAGE-OPD improves the three-run mean score from -25.64 to -23.62 for Qwen3-0.6B and from -21.98 to -16.31 for Qwen3-1.7B. Each additional seed improves mean score for both students; success rates tie for Qwen3-0.6B and improve for Qwen3-1.7B. These replications support the gains in ALFWorld unseen success and ScienceWorld task score across training seeds under the 32B-teacher setting.

Table 12: Individual runs underlying Table 11. All runs use Qwen3-32B as teacher. Original denotes the run already reported in Table 1; seeds 43 and 44 are additional independent training runs. SR is in percent.

Student	Run	Method	ALFWorld		ScienceWorld	
			Seen SR ↑	Unseen SR ↑	SR ↑	Score ↑
0.6B	Original	OPD	53.57	68.66	3.36	-28.46
0.6B	Original	SAGE-OPD	54.29	70.90	6.04	-26.10
0.6B	43	OPD	54.29	69.40	2.01	-22.11
0.6B	43	SAGE-OPD	58.57	70.90	2.01	-21.77
0.6B	44	OPD	56.43	67.91	3.36	-26.34
0.6B	44	SAGE-OPD	58.57	73.88	3.36	-23.00
1.7B	Original	OPD	58.57	74.63	4.03	-29.89
1.7B	Original	SAGE-OPD	62.86	73.88	7.38	-16.01
1.7B	43	OPD	59.29	73.88	2.68	-18.62
1.7B	43	SAGE-OPD	60.00	79.10	4.03	-15.69
1.7B	44	OPD	62.14	73.13	4.70	-17.44
1.7B	44	SAGE-OPD	62.14	77.61	6.04	-17.23

E PROMPT TEMPLATES

We use task-specific ReAct-style prompts for each benchmark. All prompts enforce a fixed output protocol so that the executable action can be reliably parsed and sent to the environment. We present the prompt details in Table 13, Table 14 and Table 15.

We use a shared intervention prompt across benchmarks, adapting its context fields to each environment. For ALFWorld, the prompt includes the task goal, scene description, latest observation, and admissible actions; for ScienceWorld, the inventory replaces the scene description. For SearchQA, it includes the question, search-budget usage, latest observation, and the action schemas `search[<query>]` and `answer[<text>]`, which allow free-form queries and answers. The judge receives these fields and the candidate student response in a single user message, providing compact interaction context rather than the full dialogue history. We include at most 20 admissible actions and truncate the latest observation and candidate response to $1{,}500$ and $2{,}000$ characters, respectively. Intervention labels are generated with temperature 0 and a maximum of 8 output tokens; unparseable outputs or failed requests default to $i_t = 1$. The prompt is shown in Table 16.

Table 13: Prompt template for ALFWorld. We use a demonstration before the user turn.

Field	Prompt Content
System prompt	`You are an embodied agent solving a household task in a text-based simulator (AlfWorld). You receive an Observation describing what you can see and a list of Admissible commands you may execute. On every turn you must reply with exactly two lines: Thought: <one short sentence of reasoning> Action: <one command, copied verbatim from the Admissible list when possible> Plan briefly, then issue the next action. Only the line beginning with 'Action:' is used to step the environment. Do not add any other text after the Action line.`
One-shot demonstration	`[One-shot demonstration omitted for brevity.]`

User turn	`Task: {task_goal}`
	`Observation: {scene_description}`
	`Admissible: cmd1; cmd2; ...; cmd30 (+N more)`

Assistant turn	`Thought: <reasoning>` `Action: <command>`

Table 14: Prompt template for ScienceWorld. We use a single generic one-shot ReAct trajectory.

Field	Prompt Content
System prompt	`You are an embodied agent solving a science task in the ScienceWorld text-based simulator. You receive an Observation describing what you can see and an Inventory listing what you carry. On every turn you must reply with exactly two lines: Thought: <one short sentence of reasoning> Action: <one valid action> Plan briefly, then issue the next action. Only the line beginning with 'Action:' is used to step the environment.`
One-shot demonstration	`[Generic one-shot ReAct trajectory omitted for brevity.]`

User turn	`Task: {task_description}`
	`Observation: {observation}`
	`Inventory: {inventory or "(empty)"}`

Assistant turn	`Thought: <reasoning>` `Action: <valid action>`

Table 15: Prompt template for SearchQA. The agent repeatedly issues Wikipedia search queries and terminates with a concise final answer.

Field	Prompt Content
System prompt	`You are an expert open-domain question-answering agent. You answer questions by repeatedly issuing wikipedia search queries, reading the returned snippets, and finally producing a concise short answer.` `Output protocol -- on every turn you must reply with exactly two blocks: <thought>... brief reasoning (1-2 sentences) ...</thought> <action>VERB[PAYLOAD]</action> where VERB[PAYLOAD] is exactly one of: search[your concise wikipedia query, 3-8 words] to fetch top-3 snippets answer[your final short answer] to terminate the episode Use search[...] until you have enough evidence, then answer[...]. Final answers should be short and direct (a single entity, date, or short phrase). Do not output anything after the </action> tag.`
One-shot demonstration	`[Fixed one-shot demonstration omitted for brevity.]`
User turn	`Question: {question}?` `<information>1) ... 2) ... 3) ...</information>`

Assistant turn	`<thought>...</thought>`
	`<action>search[query]</action>`
	or
	`<thought>...</thought>`
	`<action>answer[final]</action>`

Algorithm 1 SAGE-OPD Training Procedure

Require: Student π_θ, teacher π_ϕ, environment $\mathcal{E}$, horizon T, weak weight α, constant ϵ

Collect student trajectories for optimizer batch $\mathcal{B}$, labeling each response before execution:
for each nonempty student response $a_t \sim \pi_\theta(\cdot \mid h_t)$ during collection **do**
  if a_t fails deterministic parsing or action-validity checks **then**
    Set $i_t = 1$
  else
    Query the teacher using the candidate response and pre-action context
    Map SKIP, WEAK, and STRONG to $i_t = 0, \alpha, 1$, respectively
    Use $i_t = 1$ if the request fails or the label cannot be parsed
  end if
  Execute the student response and retain (h_t, a_t, i_t) in $\mathcal{B}$
end for
for each nonempty turn (h_t, a_t, i_t) in $\mathcal{B}$ **do**
  Compute $c_t = |a_t|^{-1} \sum_j \max_v \pi_\phi(v \mid h_t, a_{t,<j})$ and set $s_t = i_t c_t$
end for
Aggregate $N_\mathcal{B} = \sum_{t \in \mathcal{B}} |a_t|$ and $A_\mathcal{B} = \sum_{t \in \mathcal{B}} s_t |a_t|$ across workers
Set $Z_\mathcal{B} = N_\mathcal{B} / \max(A_\mathcal{B}, \epsilon)$; treat all weights as constants
Update the student with $\mathcal{L}_{\text{SAGE-OPD}} = \frac{Z_\mathcal{B}}{N_\mathcal{B}} \sum_{t \in \mathcal{B}} s_t \sum_j \ell^{\text{OPD}}_{t,j}$

Table 16: Prompt template for turn-level intervention. The teacher acts as a supervision gatekeeper and assigns one intervention label for each student response.

Field	Prompt Content
Instruction	`You are a SUPERVISION GATEKEEPER for distillation training of a smaller agent operating in a multi-turn text-based environment. Your job is to decide whether the student's proposed next action deserves teacher distillation supervision.`

Episode goal	`GOAL OF THE EPISODE:`
	`Task type: {task_type}`
	`Task goal: {task_goal}`

Current environment state	`CURRENT STATE OF THE ENVIRONMENT (what the student is reacting to):`
	`Scene description: {scene_description}`
	`Last observation: {last_obs}`
	`Currently admissible actions (top {admissible_count}):`
	`{admissible}`

Student response	`STUDENT'S CANDIDATE NEXT ACTION (what we are evaluating):` `{llm_text}`

Decision rubric	`DECISION RUBRIC:`
	`- SKIP: The student's action is reasonable, strategically equivalent to your preferred action, or the situation is too ambiguous for confident correction. Distillation here would just over-regularize the student.`
	`- WEAK: The student's action may be suboptimal but the issue is uncertain or low-impact. A mild distillation signal is appropriate but the action is not clearly wrong.`
	`- STRONG: The student's action is clearly problematic, off-task, redundant, or likely to fail the goal. The student needs to learn the correct behavior here.`

Output format	`Respond with EXACTLY ONE WORD: SKIP, WEAK, or STRONG.`