RESCC: CALIBRATING AGENT SELF-DISTILLATION WITH RETRIEVED ACTION OUTCOMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Teacher-guided self-distillation provides long-horizon agents with dense token-level supervision, complementing the sparse and coarse feedback from terminal rewards. Confidence-based gating reflects how strongly a teacher prefers a prediction but does not establish whether the resulting action improves downstream outcomes. We propose RESCC (Retrieval-Simulated Counterfactual Credit), which calibrates teacher guidance using action outcomes from previously collected training trajectories. RESCC retrieves similar state–action histories and compares the current action with well-supported alternatives using behavior-probability correction and a cross-fitted outcome model. It converts this evidence into a bounded signed weight on the distillation loss and falls back to the original update when historical support is insufficient. GRPO and online rollouts remain unchanged. Across ALFWorld, WebShop, and Search-QA with Qwen2.5-3B, Qwen2.5-7B, and Qwen3-1.7B, RESCC achieves improvements of up to 19.5 percentage points over existing baselines. Ablation studies show that outcome-aware retrieval, behavior correction, support-aware abstention, and bidirectional credit all contribute to the performance gains. Executed-alternative evaluations further show that RESCC aligns substantially better with realized action outcomes than confidence-based signals, with reliability improving as historical support increases.

1 INTRODUCTION

Long-horizon agents commonly learn from two complementary signals: sparse environment rewards and dense teacher supervision. Environment rewards directly reflect task success but provide limited guidance for intermediate decisions, whereas teacher-guided self-distillation provides fine-grained token-level supervision without directly accounting for the downstream consequences of an action (Hinton et al., 2015; Lu et al., 2026). This distinction matters for long-horizon agents because each action changes the state faced by later decisions. A locally plausible action may violate a prerequisite or eliminate a better future path, while a useful exception may disagree with a generic teacher strategy (Yao et al., 2022b; Li et al., 2026c; Jiang et al., 2026).

We therefore ask: *when should teacher guidance be trusted for a particular agent action?* SDAR (Lu et al., 2026) couples GRPO (Shao et al., 2024) with self-distillation and gates the auxiliary token loss using a detached teacher–student probability gap. However, this gate measures teacher preference rather than action consequence, so a confident teacher can still endorse an error or reject a useful correction. Other recent approaches explore local on-policy comparisons or extract guidance from completed trajectories to determine when teacher supervision is reliable (Tu et al., 2026; Wu et al., 2026a; Zhang et al., 2026). We ask whether the trajectories already generated during training can provide such evidence directly. Reusing these state–action–outcome records naturally connects teacher supervision to observed consequences, without additional paired online rollouts or a separate trajectory-analysis stage.

Controlled diagnostics show that historical outcomes are informative only when retrieved examples are sufficiently similar and well supported. When evidence is reliable, outcome-based comparisons can disagree with teacher confidence. Weak evidence should leave teacher supervision unchanged. These cases motivate a mechanism that can strengthen, suppress, or preserve teacher supervision. Following this principle, we propose RESCC, which retrieves comparable state–action–outcome histories and contrasts the current action with a well-supported alternative. It corrects for behavior

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

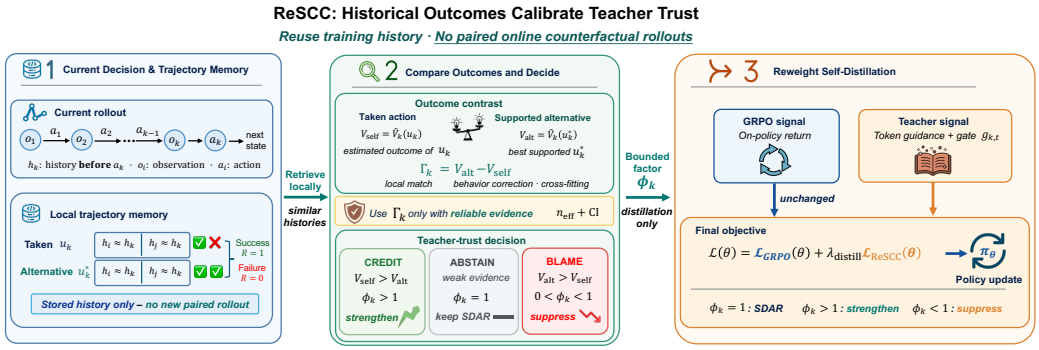


Figure 1: **ReSCC calibrates privileged guidance with retrieved action outcomes.** Retrieved local comparisons produce signed credit that reinforces or suppresses the auxiliary teacher signal; weak support triggers abstention and falls back to the SDAR (Lu et al., 2026) update. The policy-gradient objective and online rollout set remain unchanged.

imbalance and uses cross-fitted outcome prediction to estimate their outcome difference. This difference determines a bounded signed weight on the distillation loss. When historical support is insufficient, ReSCC preserves the original SDAR (Lu et al., 2026) update. The policy-gradient objective and online rollout set remain unchanged. Figure 1 summarizes how ReSCC uses retrieved action outcomes to calibrate teacher supervision while leaving the underlying RL update unchanged.

We evaluate ReSCC on ALFWorld, WebShop, and Search-QA with Qwen2.5-3B, Qwen2.5-7B, and Qwen3-1.7B under a shared 150-update training horizon. On average over the 12 aggregate evaluation metrics, ReSCC improves over GRPO by **12.0 percentage points** and over SDAR by **4.0 percentage points**. The largest gain over SDAR reaches **19.5 points** on WebShop exact-purchase accuracy with Qwen3-1.7B. Controlled ablations further validate the design choices: shuffling historical outcomes reduces ALFWorld by 2.2 points, while removing behavior correction or support-aware abstention reduces it by 2.1 and 2.4 points, respectively. One-sided credit variants also have lower mean performance than the full method. Executed-alternative evaluations show substantially stronger agreement with realized action outcomes, reaching a pairwise AUC of 0.79 compared with 0.58 for teacher confidence. ReSCC also improves a step-aware GIGPO (Feng et al., 2026) backbone by 1.3 points on ALFWorld, suggesting that calibrating teacher supervision is complementary to step-level RL credit.

We make three contributions:

- We identify the mismatch between teacher confidence and downstream action outcomes in long-horizon agent learning, and formulate when to trust teacher supervision as a problem distinct from RL credit assignment.
- We introduce ReSCC, a retrieval-based approach that uses historical state–action–outcome data to calibrate teacher supervision through local outcome comparison, behavior correction, support-aware abstention, and bounded bidirectional weighting, without paired counterfactual rollouts.
- Across three agent benchmarks and three Qwen backbones, we demonstrate average performance gains and validate the proposed mechanism through controlled ablations, executed-action interventions, and support analyses.

2 Related Work

Long-horizon agent learning and credit assignment. Language agents increasingly operate in long-horizon settings, including embodied environments, web interaction, search, dataset discovery, scientific research, and real-world tasks (Yao et al., 2022b; Shridhar et al., 2020; Yao et al., 2022a; Jin et al., 2025; Li et al., 2025; Wu et al., 2026b; Li et al., 2026c). They are commonly optimized with policy-gradient methods such as PPO and GRPO (Yang et al., 2025; Schulman et al., 2017; Shao et al., 2024). Recent work improves temporal credit assignment with finer-grained signals for intermediate actions, including grouped step-level advantages and step-aware online preferences (Feng et al., 2026;

Zhang et al., 2026). These methods decide which sampled actions RL should reinforce, whereas RESCC controls how strongly teacher supervision should influence a sampled action.

Teacher-guided self-distillation. Knowledge distillation provides dense supervision beyond sparse rewards (Hinton et al., 2015). Recent methods extend self-distillation to on-policy reasoning and RL, where the model produces both the optimized behavior and a stronger teacher view (Zhao et al., 2026; Yang et al., 2026; Lu et al., 2026). Agent-specific approaches further select teacher guidance using local on-policy comparisons or trajectory-derived skills (Tu et al., 2026; Wu et al., 2026a). In contrast, RESCC estimates teacher trust from whether historically similar actions led to better or worse outcomes, without paired on-policy comparisons or skill extraction.

Trajectory reuse, data efficiency, and reliable agent supervision. Recent work improves the use of collected interactions through data-efficient agency, long-horizon experience acquisition, and asynchronous optimization with human guidance (Xiao et al., 2025; Jiang et al., 2026; Fu et al., 2026). At the same time, studies of multi-agent bias amplification and resource-aware evaluation show that additional supervision or interaction does not always improve reliability (Li et al., 2026a;b). RESCC therefore uses historical trajectories not only as training data but also as evidence for when supervision should be trusted, by retrieving comparable state–action–outcome records, correcting behavior imbalance, and abstaining under insufficient support.

3 Background and Problem Setup

Long-horizon agent learning. For a task x, an agent interacts with an environment over K turns. Before choosing action a_k, it observes the interaction history

$$h_k = (x, o_1, a_1, \ldots, o_{k-1}, a_{k-1}, o_k),$$

and produces an action $a_k = (y_{k,1}, \ldots, y_{k,T_k})$. The resulting trajectory is

$$\tau = (x, o_1, a_1, \ldots, o_K, a_K), \qquad R(\tau) \in [0, 1]. \tag{1}$$

GRPO (Shao et al., 2024) samples multiple trajectories for the same task and converts their terminal returns into group-relative trajectory advantages. Our method preserves this RL signal.

Teacher-guided self-distillation. SDAR (Lu et al., 2026) complements the trajectory-level RL signal with dense token-level teacher supervision. For token t of action a_k, let

$$s_{k,t} = (h_k, y_{k,<t}), \qquad s^+_{k,t} = (h_k, c^+, y_{k,<t}),$$

denote the student context and the teacher context with additional guidance c^+, respectively. SDAR weights the auxiliary supervision using the detached teacher–student probability gap

$$\Delta_{k,t} = \mathrm{sg}\Big[\log \pi_T(y_{k,t} \mid s^+_{k,t}) - \log \pi_\theta(y_{k,t} \mid s_{k,t})\Big], \qquad g_{k,t} = \sigma(\beta \Delta_{k,t}). \tag{2}$$

The gate therefore reflects how strongly the teacher prefers the sampled token, but it does not indicate whether the complete action leads to a better downstream outcome.

What is missing? We use an action-level outcome comparison for long-horizon decisions. At turn k, let $V_{\mathrm{self}}(h_k)$ denote the outcome value of the action taken, and let $V_{\mathrm{alt}}(h_k)$ denote the value of the best alternative with sufficient evidence from comparable historical decisions. We define the target contrast

$$\Gamma_k = V_{\mathrm{alt}}(h_k) - V_{\mathrm{self}}(h_k). \tag{3}$$

A positive Γ_k indicates that a better alternative is supported by historical outcomes, whereas a negative Γ_k indicates that the taken action is better supported. Insufficient historical evidence calls for no directional adjustment. The central challenge is therefore to estimate this comparison from non-uniform training histories, determine when the estimate is reliable, and use it to calibrate teacher supervision without changing the underlying RL objective. The next section describes how RESCC implements these three steps.

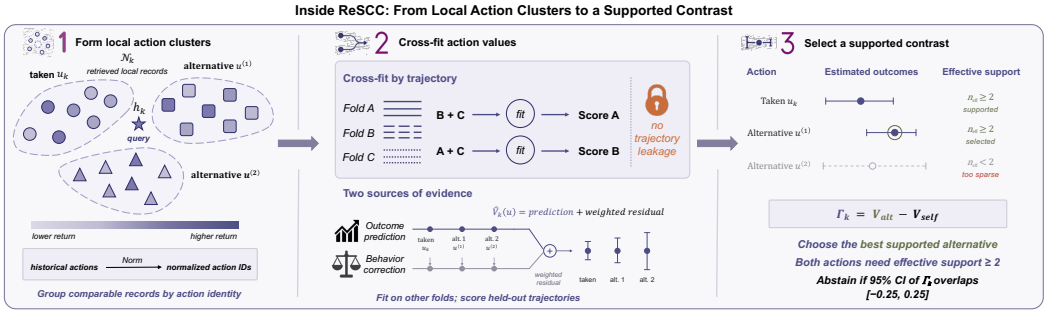


Figure 2: **Local action-outcome comparison with support-aware abstention.** ReSCC estimates the taken action and best supported alternative from comparable historical decisions using behavior correction and cross-fitted prediction. Insufficient support triggers abstention and keeps $\phi = 1$.

4 Retrieval-Simulated Counterfactual Credit

ReSCC estimates the action-level outcome contrast in Equation (3) from trajectories already collected during training and uses it only to reweight the SDAR distillation signal. As shown in Figures 1 and 2, the method has three stages: retrieve locally comparable historical decisions, estimate action outcomes under explicit support checks, and convert the supported contrast into a bounded distillation weight.

4.1 Trajectory Memory and Local Retrieval

We maintain a trajectory memory $\mathcal{D}$ containing decisions collected during training. After each rollout batch, turn i contributes

$$d_i = (z_i, u_i, \ell_i, R_i, \mathrm{id}_i, v_i, t_i), \qquad z_i = f(h_i), \quad u_i = \mathrm{Norm}(a_i), \tag{4}$$

where $R_i = R(\tau_i)$ is the return of the source trajectory, u_i is the normalized action identity used for historical comparison, and ℓ_i is the summed log-probability of the action tokens under the policy that collected the trajectory. id_i identifies the source trajectory, while v_i and t_i record the policy version and collection time. The memory is implemented as a 200,000-record ring buffer.

For the current decision h_k, we embed the complete history as $z_k = f(h_k)$ and retrieve its nearest historical neighbors before inserting the current batch, preventing self-retrieval. We use cosine similarity with `all-MiniLM-L6-v2` (Reimers & Gurevych, 2019), retaining at most 32 neighbors with similarity at least 0.6. The resulting neighborhood is denoted by $\mathcal{N}_k$. We group historical actions inside $\mathcal{N}_k$ by their normalized identity u. Structured environment commands are normalized directly, while free-form actions are grouped semantically. Section B gives further details on action normalization and leakage prevention.

4.2 Local Outcome Estimation and Support

Historical actions are sampled with unequal probabilities. Frequently chosen actions should therefore not dominate the comparison through frequency alone. For a retrieved record i, let

$$q_i = \exp(\ell_i)$$

be its logged behavior probability. For each normalized action u, we use clipped, self-normalized inverse-probability weights

$$\begin{aligned} \widetilde{w}_i(u) &= \mathbb{I}[u_i = u] \min\left(\frac{1}{\max(q_i, 10^{-6})}, 10\right), \\ w_i(u) &= \frac{\widetilde{w}_i(u)}{\sum_{j \in \mathcal{N}_k} \widetilde{w}_j(u)}. \end{aligned} \tag{5}$$

We combine this correction with a cross-fitted outcome model. All records from the same source trajectory remain in the same fold. Let $f(i)$ denote the fold of record i, $\widehat{m}^{-f(i)}$ a return predictor

trained without that fold, and $\overline{m}(z_k, u)$ the average prediction for the current query across held-out-fold models. The self-normalized estimate is

$$\widehat{V}_{\mathrm{SNIPS},k}(u) = \sum_{i \in \mathcal{N}_k} w_i(u) R_i, \tag{6}$$

and the augmented estimate used by ReSCC is

$$\widehat{V}_k(u) = \overline{m}(z_k, u) + \sum_{i \in \mathcal{N}_k} w_i(u) \left[R_i - \widehat{m}^{-f(i)}(z_i, u_i) \right]. \tag{7}$$

We refer to this cross-fitted augmented estimator as ReSCC-DR in the experiments.

We next determine whether an action has enough local evidence to participate in the comparison. Its effective sample size is

$$n_{\mathrm{eff},k}(u) = \frac{\left(\sum_{i \in \mathcal{N}_k} w_i(u)\right)^2}{\sum_{i \in \mathcal{N}_k} w_i(u)^2}. \tag{8}$$

Let

$$\mathcal{U}_k^{\mathrm{sup}} = \{u : n_{\mathrm{eff},k}(u) \geq 2\}$$

be the actions with sufficient local support. If the taken action u_k is unsupported, or no supported alternative exists, ReSCC abstains. Otherwise, we choose the highest-valued *supported* alternative

$$u_k^* = \arg \max_{u \in \mathcal{U}_k^{\mathrm{sup}} \setminus \{u_k\}} \widehat{V}_k(u), \qquad V_{\mathrm{self}} = \widehat{V}_k(u_k), \qquad V_{\mathrm{alt}} = \widehat{V}_k(u_k^*), \qquad \Gamma_k = V_{\mathrm{alt}} - V_{\mathrm{self}}. \tag{9}$$

Figure 2 illustrates this local comparison, from grouping retrieved actions to estimating their outcomes and selecting a supported alternative.

Finally, we use a trajectory-cluster bootstrap with 500 resamples to assess the stability of Γ_k. If its 95% interval overlaps the indifference region $[-0.25, 0.25]$, the comparison is inconclusive and ReSCC abstains. Thus, only locally supported and sufficiently separated action comparisons influence teacher supervision.

4.3 Signed reweighting and optimization

For a supported comparison, the sign of Γ_k determines how teacher supervision should change. If $\Gamma_k < 0$, the taken action has the better estimated outcome and receives positive credit; if $\Gamma_k > 0$, a supported alternative is better and the teacher signal for the taken action is suppressed. We encode this direction with a signed action-level quantity ψ_k. For a supported single-turn comparison, $\psi_k = -\Gamma_k$. When multiple turns in the same failed trajectory receive blame, capped Shapley allocation redistributes their negative magnitudes while preserving their relative ordering and limiting total blame to one (Equation (14) and algorithm 1). Abstention sets

$$\psi_k = 0.$$

We map the signed quantity to a bounded multiplicative factor

$$\phi_k = \exp[\mathrm{clip}(\alpha \psi_k, -0.7, 0.7)], \qquad \alpha = 4, \qquad \phi_k \in [0.497, 2.014]. \tag{10}$$

Every valid token t in action a_k then receives the reweighted self-distillation term

$$\ell_{k,t}^{\mathrm{ReSCC}} = g_{k,t} \phi_k \left[\log \pi_T(y_{k,t} \mid s_{k,t}^+) - \log \pi_\theta(y_{k,t} \mid s_{k,t}) \right]. \tag{11}$$

The final objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{GRPO}}(\theta) + \lambda_{\mathrm{distill}} \mathcal{L}_{\mathrm{ReSCC}}(\theta), \qquad \lambda_{\mathrm{distill}} = 0.01. \tag{12}$$

Teacher probabilities, $g_{k,t}$, ψ_k, and ϕ_k are detached from gradient computation. Therefore, $\phi_k > 1$ strengthens teacher supervision, $\phi_k < 1$ suppresses it, and $\phi_k = 1$ exactly recovers the original SDAR distillation update. The GRPO advantage, clipping, trajectory grouping, and online rollout set remain unchanged.

** indicates evaluation with retrieved skills following the source setup. ALFWorld Avg is overall task success; Search-QA Avg is the macro average over seven subsets; WebShop reports partial-credit Score and exact-purchase accuracy. Trainable rows use 150 policy updates.

Table 1: **End-task performance after 150 policy updates.** All values are percentages; higher is better. Baseline results follow Lu et al. (2026), and RESCC reports three-seed means. In ALFWorld Avg, Search-QA Avg, WebShop Score, and WebShop Acc, **bold** and underline denote the best and second-best results within each backbone, respectively.

	ALFWorld success							Search-QA accuracy								WebShop	
Method	**Pick**	**Look**	**Clean**	**Heat**	**Cool**	**Pick2**	**Avg**	**NQ**	**Triv**	**Pop**	**Hotp**	**2Wk**	**MuS**	**Bam**	**Avg**	**Score**	**Acc**
Qwen2.5-3B-Instruct																	
Vanilla	44.4	11.1	6.2	15.4	28.6	12.5	21.9	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7	6.7	0.8
Skill-Prompt*	51.7	66.7	48.4	0.0	4.3	10.0	28.9	23.7	46.2	30.6	24.4	22.1	7.5	12.5	23.9	0.2	0.8
OPSD	48.8	41.7	16.7	0.0	15.8	16.7	28.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	11.3	3.1
GRPO	91.2	62.5	96.2	61.9	65.0	47.4	75.0	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	79.8	63.3
Skill-GRPO	88.9	71.4	58.8	70.6	40.7	29.2	60.2	43.5	58.8	43.0	36.8	32.2	11.7	12.5	34.1	77.3	60.9
Skill-GRPO*	94.3	57.1	100	66.7	73.1	57.1	80.5	44.3	59.6	44.3	39.0	36.1	14.5	14.9	36.1	76.3	66.4
GRPO+OPSD	100	82.4	85.7	75.0	70.0	60.0	81.2	44.9	61.2	45.2	40.4	38.5	16.0	66.1	**44.6**	77.8	66.4
Skill-SD	88.2	50.0	96.2	52.4	65.0	57.9	73.4	44.4	60.4	44.0	39.5	40.4	15.4	64.9	44.1	75.9	64.0
RLSD	87.9	75.0	90.9	75.0	73.1	68.4	79.7	41.5	58.6	42.3	40.4	40.2	16.8	66.9	43.8	84.4	66.4
SDAR	97.1	62.5	100	61.9	75.0	84.2	84.4	44.8	58.1	44.3	38.6	36.2	15.7	66.1	43.4	85.0	68.0
RESCC	96.8	90.0	85.2	71.4	87.0	81.2	**85.9**	44.5	60.9	41.7	40.6	38.9	13.3	66.1	43.7	**86.0**	**77.3**
Qwen2.5-7B-Instruct																	
Vanilla	36.1	22.2	3.1	0.0	0.0	0.0	12.5	25.2	50.8	29.5	29.0	29.0	10.4	63.7	33.9	5.9	1.6
Skill-Prompt*	51.7	50.0	32.3	5.3	4.3	0.0	23.4	30.9	52.1	32.7	32.7	27.9	12.7	66.1	36.4	1.7	0.8
OPSD	50.0	60.0	22.7	21.4	17.6	9.5	32.8	8.8	8.6	17.5	2.5	4.2	0.5	1.2	6.2	4.5	2.3
GRPO	91.2	87.5	96.2	81.0	65.0	57.9	81.2	45.1	63.7	44.0	43.6	43.2	16.8	37.6	42.0	80.9	72.6
Skill-GRPO	88.5	66.7	65.2	61.1	57.7	73.1	69.5	45.2	63.7	45.7	43.1	43.3	19.6	21.4	40.3	80.4	71.9
Skill-GRPO*	100	83.3	96.4	83.3	75.0	78.9	**88.3**	44.8	63.0	45.1	43.7	43.7	20.5	71.4	47.5	87.0	81.2
GRPO+OPSD	91.4	61.5	100	87.5	76.5	52.2	80.4	47.3	64.5	46.9	43.8	39.3	18.0	69.4	47.0	86.8	76.5
Skill-SD	93.9	93.8	90.9	100	69.2	68.4	85.1	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5
RLSD	100	87.5	92.3	58.8	80.0	65.2	82.0	46.8	63.0	44.4	45.5	48.9	21.5	73.0	**49.0**	87.4	77.3
SDAR	94.7	75.0	100	86.7	68.2	78.9	85.9	46.3	63.5	48.2	43.8	48.4	19.6	73.0	**49.0**	89.4	**82.8**
RESCC	90.9	73.3	100.0	100.0	66.7	79.2	86.7	48.0	63.4	42.4	44.7	44.4	19.7	72.2	47.8	**90.4**	80.5
Qwen3-1.7B-Instruct																	
Vanilla	25.0	22.2	3.1	0.0	21.4	4.2	12.5	29.4	46.9	37.0	23.5	19.6	6.4	10.5	24.8	46.5	4.7
Skill-Prompt*	10.3	50.0	16.1	0.0	0.0	5.0	9.4	29.4	46.5	36.2	22.9	20.8	4.3	10.1	24.3	23.0	2.3
OPSD	26.3	33.3	9.1	0.0	4.5	5.3	14.1	4.2	8.3	4.6	6.6	15.3	0.7	1.2	5.8	47.4	9.3
GRPO	71.1	41.7	36.4	40.0	31.8	31.6	46.1	40.0	58.9	43.5	35.4	30.3	12.0	65.7	40.8	67.3	38.3
Skill-GRPO	27.6	54.5	22.7	27.3	0.0	19.2	21.1	39.2	58.6	43.9	35.2	28.2	11.5	66.1	40.4	73.4	46.1
Skill-GRPO*	31.4	42.9	51.9	8.3	11.5	7.1	28.1	38.0	58.4	43.9	36.3	29.0	12.5	66.9	40.7	80.4	50.0
GRPO+OPSD	38.2	50.0	30.8	28.6	30.0	21.1	32.0	40.7	58.9	45.0	37.0	34.6	13.3	65.7	**42.2**	70.7	38.3
Skill-SD	52.9	37.5	69.2	42.9	60.0	36.8	52.3	39.1	57.5	45.4	34.8	34.1	10.7	64.1	40.8	81.8	53.9
RLSD	50.0	37.5	61.5	19.0	50.0	21.1	42.2	38.6	57.3	43.0	34.5	34.1	11.5	65.3	40.6	74.0	50.8
SDAR	73.5	25.0	76.9	33.3	40.0	36.8	53.9	39.7	58.9	45.3	35.9	35.5	12.6	65.3	41.9	76.8	58.6
RESCC	66.7	64.3	80.8	40.0	64.7	35.0	**60.9**	42.1	58.0	39.1	35.6	32.5	10.2	65.7	40.5	**89.5**	**78.1**

5 EXPERIMENTS

Experimental Setup. We evaluate RESCC on three long-horizon agent benchmarks. ALFWorld measures task success across six household interaction families (Shridhar et al., 2020). WebShop evaluates web-based product interaction using completion Score and exact-purchase accuracy on a fixed 128-task set (Yao et al., 2022a). Search-QA evaluates search-based reasoning using the macro-average accuracy over NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle (Jin et al., 2025). We use Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and Qwen3-1.7B-Instruct (Yang et al., 2025). All RESCC results report three-seed means from training on eight H200 GPUs. Trainable methods use a shared horizon of 150 policy updates. The main comparison includes GRPO (Shao et al., 2024), SDAR (Lu et al., 2026), and the agent-training baselines reported by Lu et al. (2026). The three-seed ablation suite keeps the training and evaluation protocol fixed while changing one RESCC component at a time; its full-model reference matches the Qwen2.5-3B result in Table 1.

5.1 RQ1: DOES RETRIEVED OUTCOME EVIDENCE IMPROVE END-TASK PERFORMANCE?

Table 1 summarizes the end-task performance across all three benchmarks and backbones. On average across the four aggregate metrics and all three backbones, RESCC improves over GRPO by **12.0 percentage points** and over the strong self-distillation baseline SDAR by **4.0 points**. The gains are especially clear on Qwen3-1.7B, where RESCC improves ALFWorld by 7.0 points and WebShop

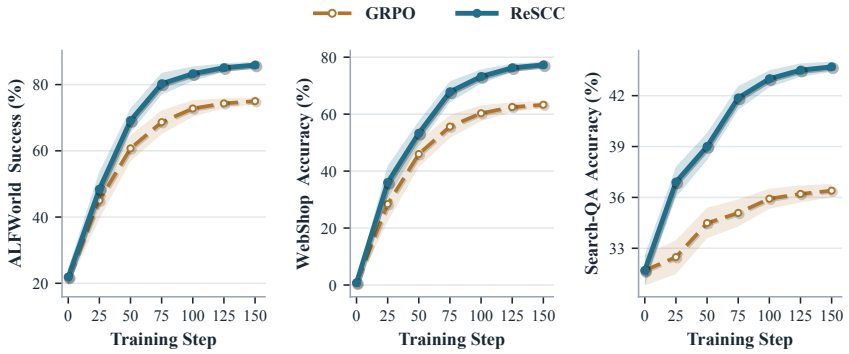


Figure 3: **Learning dynamics on Qwen2.5-3B over 150 policy updates.** Curves show three-seed means for GRPO and RESCC on ALFWorld success, WebShop exact-purchase accuracy, and Search-QA macro accuracy. Shaded regions denote one-standard-deviation uncertainty across the three runs.

Score and exact-purchase accuracy by 12.7 and 19.5 points over SDAR. On Qwen2.5-3B, RESCC improves all four aggregate metrics, including a 9.3-point gain in WebShop exact-purchase accuracy.

The task-level results show where these gains arise. On ALFWorld, improvements are particularly pronounced for Look and Cool, while on WebShop the strongest gains appear in successful final purchases. These differences motivate a closer examination of how local outcome evidence changes teacher supervision. Figure 3 complements the final results by showing how the performance difference develops during optimization, with the clearest separation on WebShop. We next isolate which components of RESCC are responsible for these gains.

5.2 RQ2: WHICH COMPONENTS DRIVE THE PERFORMANCE GAINS?

We evaluate each design choice under a controlled Qwen2.5-3B protocol with 150 policy updates. Each ablation corresponds to a component in Section 4. Retrieval determines which historical decisions are compared. Behavior correction and the outcome model estimate their values. The support rule decides when the comparison is reliable, and signed weighting determines how the resulting evidence changes teacher supervision.

Table 2: **Component ablations on Qwen2.5-3B.** RESCC variants report three-seed mean ± standard deviation. The full-model means match Table 1. Δ gives the percentage-point difference from the full RESCC-DR model under the same protocol.

Method	ALFWorld	Search macro	Δ ALF	Δ Search
Reference objectives				
GRPO	75.0	36.4	−10.9	−7.3
SDAR ($\phi = 1$)	84.4	43.4	−1.5	−0.3
RESCC-DR (full)	85.9 ± 0.6	43.7 ± 0.3	—	—
Estimator and decision rule				
w/o behavior correction	83.8 ± 0.9	43.0 ± 0.5	−2.1	−0.7
SNIPS only (w/o outcome model)	85.0 ± 0.7	43.5 ± 0.4	−0.9	−0.2
w/o support-aware abstention	83.5 ± 1.1	42.9 ± 0.6	−2.4	−0.8
w/o Shapley (uniform blame)	85.2 ± 0.8	43.6 ± 0.4	−0.7	−0.1
Positive credit only	84.6 ± 0.8	43.3 ± 0.4	−1.3	−0.4
Blame suppression only	84.9 ± 0.7	43.5 ± 0.4	−1.0	−0.2
Historical-evidence controls				
Random neighbors	83.2 ± 1.2	42.8 ± 0.6	−2.7	−0.9
Shuffled outcomes	83.7 ± 1.0	43.1 ± 0.5	−2.2	−0.6

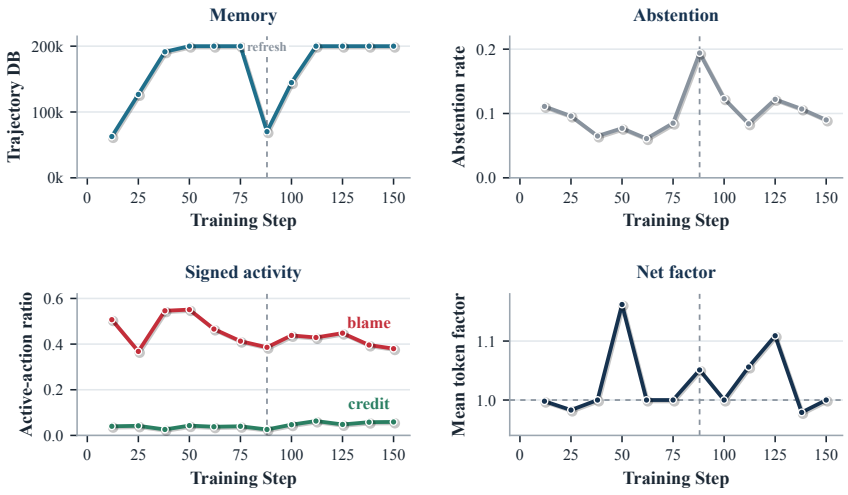


Figure 4: **Training-time dynamics of RESCC on Search-QA.** The trajectory memory grows with training and briefly refreshes at update 88, increasing abstention until coverage recovers. Positive and negative signed weights remain active, while their effects can offset in the mean token factor.

Table 2 shows the contribution of each component. The historical-evidence controls test whether retrieved records contain the information RESCC is designed to use. Random neighbors break the local state matching in Section 4, while shuffled outcomes preserve retrieved actions but break their correspondence with observed results. These cause substantial drops: random neighbors reduce ALFWorld and Search-QA by 2.7 and 0.9 points, respectively, while shuffled outcomes reduce them by 2.2 and 0.6 points. The remaining ablations test how this evidence is used. Removing behavior correction costs 2.1 and 0.7 points, while removing support-aware abstention costs 2.4 and 0.8 points. The SNIPS-only variant removes cross-fitted outcome prediction, and the two one-direction variants prevent either reinforcing useful actions or suppressing harmful ones; all underperform the full estimator on average. Replacing capped multi-turn allocation with uniform blame also lowers mean performance. Overall, the results support the full pipeline of local retrieval, outcome estimation, support checking, and bidirectional reweighting.

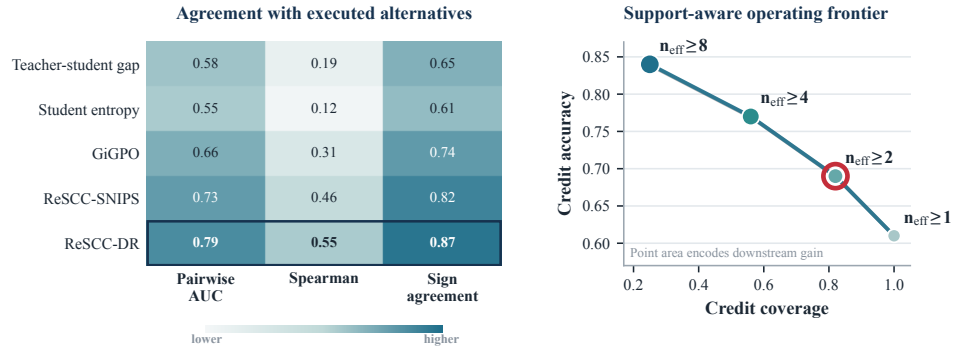


Figure 5: **Agreement with executed alternatives and the effect of historical support.** Left: agreement between different decision signals and realized taken-versus- alternative outcomes. Right: cumulative support-threshold subsets. Coverage is measured over all candidate decisions, and credit accuracy within each retained subset. Table 5 reports the corresponding disjoint bins. Point size denotes downstream gain, and the outlined point uses $n_{eff} \geq 2$.

5.3 RQ3: DOES RETRIEVED OUTCOME EVIDENCE REFLECT REALIZED ACTION OUTCOMES?

Beyond end-task performance, we test whether RESCC's historical comparison predicts which action yields a better outcome. For held-out saved states, we execute both the taken action and the

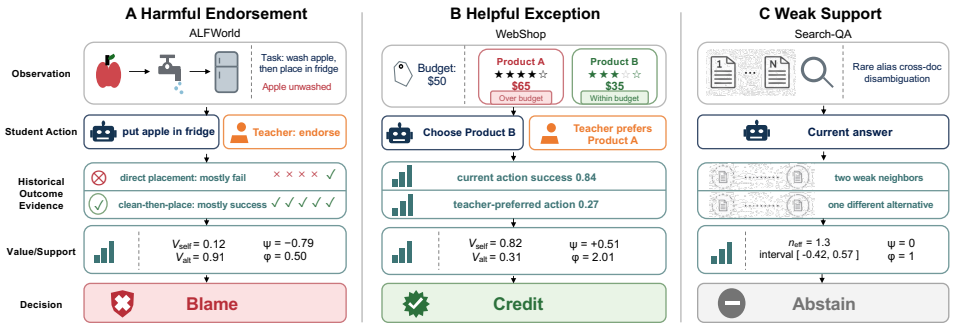


Figure 6: **Representative teacher-supervision decisions.** A harmful teacher endorsement is suppressed when a supported alternative has better outcomes; a useful action is reinforced when its outcome is better; weak historical support leaves $\phi = 1$.

highest-ranked supported alternative, evaluating pairwise AUC, Spearman correlation with signed return differences, and sign agreement. As shown in Figure 5, teacher confidence is weakly aligned with realized outcomes (0.58 AUC). Local outcome comparison with behavior correction raises AUC to 0.73, and adding the cross-fitted outcome model reaches **0.79 AUC**, **0.55 Spearman correlation**, and **0.87 sign agreement**. This supports Section 4: local historical outcomes provide signal beyond teacher confidence. The right panel shows the reliability–coverage trade-off across cumulative support-threshold subsets corresponding to Table 5. Raising minimum support improves credit accuracy from 61% to 84% while reducing coverage from 100% to 25%. The selected threshold retains 82% coverage at 69% credit accuracy with positive downstream gain, indicating that support-aware abstention improves reliability at the cost of coverage.

5.4 RQ4: How does ReSCC behave during training and individual decisions?

Training dynamics. Figure 4 shows how trajectory memory, abstention, and signed reweighting evolve during training. Memory reaches its 200,000-record capacity by update 50; after refresh, it drops to 70,144 and abstention rises to 0.194, then both recover as new trajectories refill the memory. Positive and negative ψ_k remain active throughout training, indicating continued reinforcement and suppression of teacher supervision. Their opposing effects also explain why the mean token factor ϕ can stay near one despite many non-neutral action-level adjustments.

Decision-level behavior. Figure 6 illustrates three cases across ALFWorld, WebShop, and Search-QA. In ALFWorld, ReSCC suppresses supervision when an unwashed-object action is worse than a better-supported sequence. In WebShop, it reinforces the taken action because it satisfies the task constraint and has substantially better historical outcomes than the teacher-preferred alternative. In Search-QA, insufficient evidence triggers abstention, leaving the original SDAR update unchanged. These examples show how bidirectional weighting and abstention operate at the action level. For long-horizon failures, independently applying multiple local negative signals can over-count one terminal failure, motivating the bounded multi-turn allocation introduced in Section 4.

6 Discussion

Conclusion. We introduce ReSCC, which calibrates teacher supervision using historical action outcomes. By comparing actions with supported alternatives, it adjusts distillation signals without changing the GRPO objective or rollouts. Across ALFWorld, WebShop, and Search-QA with three Qwen backbones, ReSCC improves average performance by up to 19.5 points over SDAR. Ablations and executed-alternative evaluations support its key components.

Limitations and future work. ReSCC depends on sufficient and relevant trajectory memory, especially for rare states and diverse free-form actions. Future work can improve retrieval, state representations, memory management, and action matching, and evaluate the approach on larger models, longer horizons, and broader agent environments.

REFERENCES

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

Dayuan Fu, Yunze Wu, Xiaojie Cai, Lyumanshan Ye, Shijie Xia, Zhen Huang, Weiye Si, Tianze Xu, Jie Sun, Keyu Li, et al. Argo: Asynchronous rollout with human guidance for research agent optimization. 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Mohan Jiang, Dayuan Fu, Junhao Shi, Ji Zeng, Weiye Si, Keyu Li, Xuefeng Li, Yang Xiao, Wenjie Li, Dequan Wang, et al. davinci-agency: Unlocking long-horizon agency data-efficiently. *arXiv preprint arXiv:2602.02619*, 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.

Keyu Li, Mohan Jiang, Dayuan Fu, Yunze Wu, Xiangkun Hu, Dequan Wang, and Pengfei Liu. Datasetresearch: Benchmarking agent systems for demand-driven dataset discovery. *arXiv preprint arXiv:2508.06960*, 2025.

Keyu Li, Jin Gao, and Dequan Wang. Aligned agents, biased swarm: Measuring bias amplification in multi-agent systems. *arXiv preprint arXiv:2604.08963*, 2026a.

Keyu Li, Jin Gao, and Dequan Wang. The cost of knowing: A resource-aware protocol for benchmarking hallucination beyond static leaderboards. *arXiv preprint arXiv:2607.24063*, 2026b.

Keyu Li, Junhao Shi, Yang Xiao, Mohan Jiang, Jie Sun, Yunze Wu, Dayuan Fu, Shijie Xia, Xiaojie Cai, Tianze Xu, et al. Agencybench: Benchmarking the frontiers of autonomous agents in 1m-token real-world contexts. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7422–7440, 2026c.

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

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang 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.

Songjun Tu, Chengdong Xu, Qichao Zhang, Yiwen Ma, Yaocheng Zhang, Linjing Li, Dong Li, Xiangyuan Lan, and Dongbin Zhao. Ucob: Learning to utilize and evolve agentic skills via credit-aware on-policy bidirectional self-distillation. *arXiv preprint arXiv:2606.29502*, 2026.

Jinyang Wu, Shuo Yang, Zhengxi Lu, Fan Zhang, Yuhao Shen, Lang Feng, Haoran Luo, Zheng Lian, Shuai Zhang, Zhengqi Wen, et al. Seed: Self-evolving on-policy distillation for agentic reinforcement learning. *arXiv preprint arXiv:2607.14777*, 2026a.

Yunze Wu, Dayuan Fu, Weiye Si, Zhen Huang, Mohan Jiang, Keyu Li, Shijie Xia, Jie Sun, Tianze Xu, Yang Xiao, et al. Innovatorbench: Evaluating agents' ability to conduct innovative ai research. In *International Conference on Learning Representations*, volume 2026, pp. 36144–36185, 2026b.

Yang Xiao, Mohan Jiang, Jie Sun, Keyu Li, Jifan Lin, Yumin Zhuang, Ji Zeng, Shijie Xia, Qishuo Hua, Xuefeng Li, et al. Limi: Less is more for agency. *arXiv preprint arXiv:2509.17567*, 2025.

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

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

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

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*, 2022b.

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

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.

A AI USE STATEMENT

Large language models assist with editorial polishing of the manuscript. The authors determine the scientific claims, supply the experimental results, check the mathematical content, and remain responsible for the final text.

B FULL METHOD DETAILS

This appendix expands the method described in Section 4 and gives the complete update procedure in Algorithm 1. We use the same notation as in the main text: h_k is the interaction history before action a_k, $u_k = \mathrm{Norm}(a_k)$ is its normalized action identity, $z_k = f(h_k)$ is the history embedding, Γ_k is the supported action-outcome difference, ψ_k is the signed action-level adjustment, and ϕ_k is the final multiplicative weight on teacher supervision.

B.1 ACTION REPRESENTATION AND TRAJECTORY-MEMORY SAFEGUARDS

Exact string matching can separate free-form actions that have the same environment-level meaning. Whenever the environment exposes a structured command or tool call, we normalize the action as

$$u_k = (\text{action type}, \text{normalized arguments}). \tag{13}$$

For example, surface forms that execute the same `clean apple with sinkbasin` command share one action identity. For environments without a fixed action grammar, such as Search-QA, we additionally group free-form actions by semantic similarity.

Consistent with Section 4, the retrieval key $z_k = f(h_k)$ encodes the complete interaction history rather than only the latest observation. This helps distinguish similar observations reached through different preceding decisions. Each rollout batch is queried against the trajectory memory before being inserted, preventing self-retrieval. Policy version and collection time stored in Equation (4) are retained for cross-fitting and memory diagnostics.

B.2 BEHAVIOR CORRECTION, OUTCOME PREDICTION, AND SUPPORT

For each retrieved record $i \in \mathcal{N}_k$, the stored action log-probability ℓ_i gives the behavior probability $q_i = \exp(\ell_i)$. The behavior correction in Equation (5) clips inverse-probability weights at 10 and normalizes them within each normalized action group. This reduces the influence of rare actions while keeping comparisons local to the retrieved neighborhood.

The outcome model follows the trajectory-level cross-fitting procedure in Equation (7). All decisions from one source trajectory remain in the same fold, and each residual is evaluated with a model trained without that fold. The prediction for the current query averages predictions from the corresponding held-out models. In the experiments, we use RESCC-DR as shorthand for this augmented estimator, which combines weighted historical outcomes with cross-fitted outcome prediction.

Support is checked before an alternative is selected. For the current decision k, define

$$\mathcal{U}_k^{\mathrm{sup}} = \{u : n_{\mathrm{eff},k}(u) \geq 2\},$$

using the effective sample size in Equation (8). If the taken action $u_k \notin \mathcal{U}_k^{\mathrm{sup}}$, fewer than two distinct actions are available, or no supported alternative exists, RESCC abstains and sets $\psi_k = 0$. Otherwise, the method selects

$$u_k^* = \arg \max_{u \in \mathcal{U}_k^{\mathrm{sup}} \setminus \{u_k\}} \widehat{V}_k(u),$$

and computes the same supported action-outcome difference as in Equation (9),

$$\Gamma_k = \widehat{V}_k(u_k^*) - \widehat{V}_k(u_k).$$

A 500-resample trajectory-cluster bootstrap provides an additional stability check. If the 95% interval for Γ_k overlaps $[-0.25, 0.25]$, the comparison is inconclusive and $\psi_k = 0$. Thus, only a locally supported and sufficiently separated action comparison changes teacher supervision.

B.3 BOUNDED MULTI-STEP ALLOCATION

For an individually supported comparison, the signed adjustment is $\psi_k = -\Gamma_k$. Therefore, $\Gamma_k < 0$ strengthens teacher supervision for the taken action, whereas $\Gamma_k > 0$ suppresses it.

When several decisions in the same failed trajectory receive negative adjustments, applying them independently can over-count one terminal failure. For each such decision, let

$$b_k = \max(0, \Gamma_k).$$

For candidate set C, we define

$$v(S) = \min\left(1, \sum_{k \in S} b_k\right), \qquad S \subseteq C, \tag{14}$$

and distribute $v(C) - v(\varnothing)$ using Shapley allocation. Coalitions are enumerated exactly for at most six candidates and approximated with 256 random permutations otherwise. The resulting negative adjustments preserve the relative contribution of the selected decisions while keeping the total magnitude bounded. Figure 7 gives a worked example.

Require: Policy π_θ, task batch $\mathcal{B}$, trajectory memory $\mathcal{D}$
Ensure: Updated policy π_θ and trajectory memory $\mathcal{D}$
Sample grouped trajectories and obtain environment rewards
Compute standard GRPO (Shao et al., 2024) trajectory advantages
Construct the student and teacher contexts for each sampled action
for each decision (h_k, a_k) **do**
    Compute $z_k = f(h_k)$ and retrieve $\mathcal{N}_k$ from $\mathcal{D}$
    Normalize historical actions into identities u
    Estimate $\widehat{V}_k(u)$ using behavior correction and cross-fitted outcome prediction
    Construct the supported set $\mathcal{U}_k^{\sup} = \{u : n_{\mathrm{eff},k}(u) \geq 2\}$
    if $u_k \notin \mathcal{U}_k^{\sup}$ or no supported alternative exists **then**
        $\psi_k \leftarrow 0$
    else
        Select the best supported alternative $u_k^* = \arg\max_{u \in \mathcal{U}_k^{\sup} \setminus \{u_k\}} \widehat{V}_k(u)$
        $\Gamma_k \leftarrow \widehat{V}_k(u_k^*) - \widehat{V}_k(u_k)$
        Compute the trajectory-cluster bootstrap interval for Γ_k
        if the interval overlaps $[-0.25, 0.25]$ **then**
            $\psi_k \leftarrow 0$
        else
            $\psi_k \leftarrow -\Gamma_k$
        end if
    end if
end for
Apply capped Shapley allocation to negative adjustments within failed trajectories
Map each ψ_k to the bounded factor ϕ_k using Equation (10)
Update θ with Equations (11) and (12)
Insert the completed rollout batch into $\mathcal{D}$

Algorithm 1: Retrieval-simulated counterfactual credit for one policy update

C ADDITIONAL EXPERIMENTAL DETAILS

This section provides the training configuration and additional experiments supporting Section 5. We use ALFWorld (Shridhar et al., 2020), WebShop (Yao et al., 2022a), and Search-QA (Jin et al., 2025), with Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and Qwen3-1.7B-Instruct (Yang et al., 2025). The main RL and self-distillation references are GRPO (Shao et al., 2024) and SDAR (Lu et al., 2026).

Table 3: **Training and retrieval configuration.** Shared optimization and retrieval settings across the three benchmarks.

Configuration	ALFWorld	WebShop	Search-QA
Training tasks per batch	16	16	128
Rollouts per task	8	8	8
Maximum prompt length	2,048	4,096	4,096
Maximum response length	512	512	512
Maximum interaction turns	50	15	4
Learning rate	10^{-6}	10^{-6}	10^{-6}
Policy clipping parameter	0.2	0.2	0.2
Distillation coefficient λ_{distill}	0.01	0.01	0.01
SDAR gate temperature β	5	5	5
Signed-weight scale α	4	4	4
Retrieval top-K / cosine threshold	32 / 0.6	32 / 0.6	32 / 0.6
Trajectory-memory capacity	200,000	200,000	200,000

Table 3 lists the settings shared across model scales.

C.1 Compatibility with step-level RL credit

Table 4: **Compatibility with step-level RL credit.** ReSCC improves both GRPO and the step-aware GIGPO backbone on ALFWorld. The GRPO + ReSCC mean matches the corresponding entry in Table 1.

RL backbone	Base	+ReSCC	Gain
GRPO	75.0	**85.9**	**+10.9**
GiGPO	85.9	**87.2**	**+1.3**

GiGPO (Feng et al., 2026) already introduces a step-level learning signal among actions sampled from repeated anchor states. Table 4 shows that adding ReSCC improves ALFWorld by 1.3 points. This supports the distinction made in the main text: step-level RL credit determines how sampled behavior contributes to the policy update, whereas ReSCC calibrates how strongly teacher supervision should influence that behavior.

D Additional Mechanism Analysis

This section provides additional evidence for the mechanisms examined in RQ3 and RQ4 of Section 5: training dynamics, historical support, disagreement with teacher confidence, and retrieval quality.

D.1 Historical support and reliability

Table 5: **Outcome-sign reliability by historical support.** Share is the fraction of candidate decisions in each disjoint support bin; credit accuracy is sign agreement with executed alternatives within that bin. Figure 5 reports the corresponding cumulative subsets, with accuracy weighted by candidate counts.

n_{eff}	Share	Credit acc.
< 2	18%	24.6%
$[2, 4)$	26%	51.8%
$[4, 8)$	31%	71.4%
≥ 8	25%	84.0%

Table 5 reports disjoint support bins. Outcome-sign accuracy increases from 24.6% in the lowest-support bin to 84% in the highest-support bin. At $n_{eff} \geq 2$, the retained bins account for 82% of candidates and have approximately 69.0% count-weighted accuracy, consistent with Figure 5. This supports withholding directional teacher adjustments when historical support is weak.

D.2 Disagreement with teacher confidence

Table 6: **Teacher preference and outcome-based adjustment can disagree.** Shares are computed over non-abstained decisions; observed outcome reports the corresponding downstream change in percentage points.

Teacher preference	RESCC adjustment	Share	Observed outcome	Interpretation
Positive	Positive	31%	$+3.4$	Outcome supports the teacher
Positive	Negative	17%	-2.7	Confident teacher, harmful action
Negative	Positive	14%	$+2.2$	Helpful exception
Negative	Negative	38%	-3.1	Rejection agrees with outcome

The two disagreement groups in Table 6 contain 31% of non-abstained decisions: 17% correspond to confident teacher preferences for actions with worse outcomes, while 14% correspond to useful actions that the teacher disfavors. This result complements the one-direction ablations in Table 2 and motivates allowing historical outcomes to adjust teacher supervision in both directions.

D.3 Retrieval quality and fallback

Table 7: **Effect of retrieval quality.** Credit accuracy is measured among candidates retained by the support threshold. Support exclusion is the fraction excluded by that threshold before the additional bootstrap check in Appendix B.2. Default retrieval corresponds to 82% coverage and 69% accuracy in Figure 5.

Neighbors	SDAR	RESCC	Credit acc.	Support exclusion
Default retrieval	84.4	**85.9**	69%	18%
Alternative encoder	84.4	85.1	65%	26%
50% random	84.4	84.2	58%	43%
Fully random	84.4	83.2	52%	71%

Table 7 shows that less informative neighbors reduce credit accuracy and increase support-threshold exclusions. Excluded comparisons retain $\phi_k = 1$, preserving the original SDAR teacher supervision locally. This fallback does not guarantee SDAR-level end-task performance: fully random retrieval yields 83.2%, compared with 84.4% for SDAR.

D.4 Bounded blame analysis

As shown in Figure 7, the three identified decisions receive a total magnitude of 1.65 if treated independently. Capped Shapley assigns them magnitudes of 0.425, 0.325, and 0.250 while preserving their relative ordering and bounding the total at 1.00. This prevents a multi-step failure from over-amplifying the auxiliary teacher adjustment while retaining the relative importance of the contributing decisions.

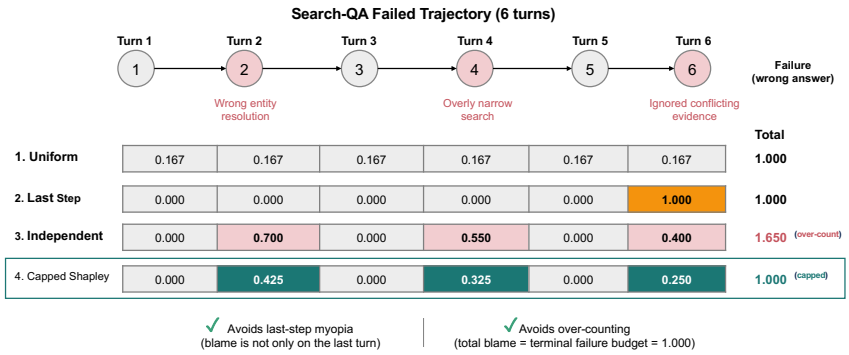


Figure 7: **Bounded blame in a six-decision worked example.** Capped Shapley preserves the ordering of three contributing decisions while reducing total blame from 1.65 to 1.00.