LEARNING FROM DECISION CONTRASTS: HISTORY-ALIGNED SELF-DISTILLATION FOR AGENTIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic on-policy self-distillation (OPSD) mitigates coarse credit assignment in outcome-based reinforcement learning (RL) by using privileged information (PI) at the trajectory or step level to guide self-teacher rescoring. However, both levels have limitations: trajectory-level supervision can weaken as student response prefixes grow, while step-level PI may be inappropriate for the target turn. Our key observation is that turns with similar interaction histories exhibit substantial contrasts in actions and outcomes, providing a natural basis for localized and informative PI construction. We propose **H**istory-**A**ligned **S**elf-**D**istillation (HASD), a unified framework that leverages history-aligned decision contrasts to provide fine-grained and contextually appropriate OPSD supervision. HASD first groups turns by their recent observation histories and aggregates empirical action outcomes to identify those with the highest and lowest outcomes. It then constructs two PI views from these action sets with opposite high/low label assignments to produce comparative OPSD signals that mitigate the negative tendency of one-sided signals. Finally, because the comparative signal is not directly verified against task outcomes, an agreement gate retains it only when its turn-level ordering agrees with GRPO advantages. Across 1.7B, 3B, and 7B models, HASD ranks first in seven of nine settings and exceeds the strongest baselines by up to 14.1, 9.3, and 1.7 points on ALFWorld, WebShop, and Search-QA, respectively.

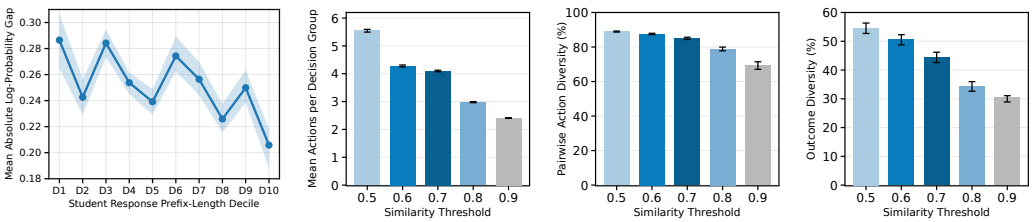
1 INTRODUCTION

Large language model (LLM) agents are increasingly used as interactive systems that observe environments, take actions, and adapt to feedback (Wang et al., 2024; Xi et al., 2025) and have been applied to tasks like tool use, software engineering, and embodied control (Schick et al., 2023; Yang et al., 2024b; Shridhar et al., 2021). When agent interactions yield verifiable outcomes, reinforcement learning methods like GRPO (Shao et al., 2024) provide a practical post-training framework. However, outcome-based RL uses a single trajectory-level signal to supervise every turn and token in a long interaction, making it difficult to identify which intermediate turns actually contributed to success or failure (Feng et al., 2025; Wei et al., 2025a; Tan et al., 2026a).

On-policy self-distillation (OPSD) complements coarse outcome supervision with dense token-level guidance on student-generated responses. Unlike on-policy distillation (OPD), which relies on a separate teacher (Agarwal et al., 2024), OPSD reuses the student policy conditioned on privileged information (PI) as the self-teacher to rescore the student's sampled response (Zhao et al., 2026). In agentic OPSD settings, common forms of PI include general skills, hindsight information, and environment feedback (Wang et al., 2026a; Zhang et al., 2026; Liu et al., 2026b). Across these forms, existing methods mainly construct PI at either the trajectory or step level, but each granularity has distinct limitations for supervising individual turns.

At the trajectory level, PI is used to rescore all student-generated tokens throughout the rollout (Lu et al., 2026; Wang et al., 2026a). Although it provides global supervision, its influence on each token depends not only on the PI content but also on the preceding student response prefix. To quantify this effect, we use the trajectory-level skill PI from SDAR (Lu et al., 2026), group student-generated tokens into prefix-length deciles and compute the mean absolute teacher–student log-probability gap

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.



(a) Mean absolute OPSD gap (b) Decision-group diversity across similarity thresholds

Figure 1: Empirical motivations for HASD using Qwen2.5-3B-Instruct. (a) Mean absolute teacher–student log-probability gap across deciles of student response-prefix length on 128 WebShop tasks. (b) Decision-group diversity across recent-history similarity thresholds on ALFWorld over 150 training steps. **Left**: mean number of actions per decision group. **Middle**: percentage of within-group turn pairs taking different actions. **Right**: percentage of decision groups containing both success and failure outcomes.

at these tokens on 1,024 rollouts from 128 WebShop tasks using Qwen2.5-3B-Instruct. Figure 1(a) shows an overall decline in the mean absolute gap from lower to higher prefix-length deciles. The trend suggests that token-level guidance is increasingly shaped by the accumulated response prefix, resulting in weaker supervision at longer prefixes.

Step-level OPSD mitigates this effect by restricting teacher rescoring to individual turns, shortening the response prefix and providing more local guidance. However, finer granularity alone does not ensure the PI is appropriate for the target turn. PI constructed from environment feedback is tied to the realized action (Li et al., 2026b; Liu et al., 2026b), but may provide implicit or incomplete evidence of action quality. In contrast, PI derived from successful trajectories provides outcome-supported references (Zhang et al., 2026), but does not ensure that the supporting experience matches the target turn's interaction history. Thus, localizing PI alone is insufficient because effective supervision also depends on PI informativeness and history compatibility with the target turn.

This raises a central question: how can we retain localized OPSD supervision while keeping the PI appropriate for the target turn? We therefore examine whether grouping same-task turns with similar recent histories still yields decision groups with sufficient action and outcome diversity for informative PI construction. Figure 1(b) provides consistent evidence across different string-based similarity thresholds from 0.5 to 0.9: averaged over the tested thresholds, the resulting decision groups contain about 3.9 actions (Left), about 82% of within-group turn pairs take different actions (Middle), and about 35% of groups contain both success and failure outcomes (Right). These results show that different actions can arise under similar interaction histories and be associated with different episode outcomes. We refer to this pattern as decision contrasts, which provide a natural basis for unified OPSD supervision: similar recent histories make shared PI contextually appropriate across turns, while their action–outcome differences provide informative evidence for PI construction.

Building on these motivations, we propose HASD: **H**istory-**A**ligned **S**elf-**D**istillation, a novel framework that leverages history-aligned decision contrasts to provide fine-grained, contextually appropriate OPSD supervision. Specifically, HASD first groups similar turns by recent observation histories and aggregates empirical action outcomes to form two action sets corresponding to the maximum and minimum outcomes in each decision group. Second, to turn these decision contrasts into comparative OPSD supervision while mitigating the negative tendency of one-sided signals, HASD constructs two PI views by assigning the two action sets opposite high/low labels and contrasts their frozen self-teacher scores. Finally, because the comparative signal is not outcome-verified whereas GRPO advantages are grounded in task outcomes, HASD uses GRPO advantages as a reference and applies an agreement gate that retains the comparative signal only when their turn-level orderings agree. Theory establishes the robustness of the agreement gate to shared shifts and bounded perturbations in OPSD scores, limiting the sensitivity of gating decisions to such variation. Across three benchmarks with 1.7B, 3B, and 7B models, HASD ranks first in seven of nine aggregate settings.

Beyond the main results, history-alignment analysis shows that longer histories do not inherently improve alignment, while matching the history window to the policy-visible context best balances decision comparability and PI availability. Theoretical and empirical analyses support the negative tendency of one-sided sampled-token OPSD and show that comparative rescoring mitigates it

while retaining substantial signal magnitude. Despite paired self-teacher rescoring, HASD remains training-efficient, incurring only $1.06\times$ the training cost of GRPO in a representative setting.

2 Related Work

Outcome-based agentic RL and credit assignment. Outcome-based agentic RL commonly uses PPO or GRPO (Schulman et al., 2017; Shao et al., 2024) to learn from environment returns, but trajectory-level rewards provide coarse credit over long interactions. Recent work introduces finer supervision through intermediate rewards, process supervision, turn-level return estimation, or progress rewards (Wei et al., 2025a; Zheng et al., 2025; Ding et al., 2026; Liu et al., 2026a). HASD instead retains the GRPO outcome objective and complements it with token-level self-distillation.

On-policy distillation and self-distillation with PI. On-policy distillation provides dense feedback on student-generated sequences (Agarwal et al., 2024; Gu et al., 2024), while self-distillation reuses the same policy with PI (Penaloza et al., 2026; Zhao et al., 2026). In agentic OPSD, PI has been introduced at different granularities. Trajectory-level methods apply shared PI such as reusable skills across a rollout (Lu et al., 2026; Wang et al., 2026a), whereas step-level methods attach hindsight from environment feedback, reflections, or successful sibling trajectories to local interactions (Li et al., 2026b; Liu et al., 2026b; Zhang et al., 2026). However, localizing PI does not ensure its supporting experience matches the target decision context when drawn from a different interaction history. HASD addresses this gap with history alignment.

Local supervision and alignment in agent training. Fine-grained credit assignment has long been studied through hindsight and counterfactual estimators (Harutyunyan et al., 2019; Mesnard et al., 2021). Recent LLM and agent methods derive local supervision from sibling rollouts, trajectory graphs, marginal contributions, implicit step rewards, or turn-wise reward propagation (Meng & Chen, 2026; Li et al., 2026a;c; Liu et al., 2026c; Zong et al., 2026). Since such signals are most informative when grounded in comparable decision contexts, related methods group or link decisions using repeated states, state-transition graphs, transition dynamics, or compact state-action representations for local credit or advantage estimation (Feng et al., 2025; Cheng et al., 2026; Liu et al., 2026d; Wang et al., 2026b). In contrast, HASD matches recent interaction histories and uses the resulting decision contrasts to construct comparative PI for self-distillation. Context alignment alone does not ensure update compatibility, since auxiliary signals may interfere with the main update objective (Du et al., 2018; Yu et al., 2020; Dery et al., 2021). Recent methods regulate signals using uncertainty reweighting or multi-scale objectives (Tan et al., 2026b; Yang et al., 2026b). HASD further enforces turn-level ordering agreement between local supervision and GRPO advantages.

3 Preliminaries

Long-horizon agentic interaction. For task x, trajectory i contains interaction turns $k = 1, \ldots, K_i$. Following prior agent-training setups that condition the policy at each turn on a bounded window of recent interaction history (Feng et al., 2025; Tang et al., 2025; Zhang & Yang, 2026), let $H_{i,k}$ denote the full policy-visible decision-time context at turn k, including the task instruction, current observation, and available recent observation–action history. At each turn, the policy samples a response $y_{i,k} \sim \pi_\theta(\cdot \mid H_{i,k})$, where $y_{i,k} = (y_{i,k,1}, \ldots, y_{i,k,L_{i,k}})$ and $L_{i,k}$ is the response length. The environment parser extracts an action $a_{i,k}$ from $y_{i,k}$ to produce the next observation $o_{i,k+1}$. At termination, $S_i \in \{0, 1\}$ denotes the binary success signal used by HASD to aggregate empirical action–outcome evidence, while R_i denotes the optimization reward used by GRPO.

Group Relative Policy Optimization. For each task x, GRPO (Shao et al., 2024) samples a rollout group $\mathcal{B}_x$ and converts terminal rewards R_i into detached group-relative advantages A_i. With rollout-policy parameters θ_{old} and likelihood ratio $\rho_{i,k,t}(\theta) = \pi_\theta(y_{i,k,t} \mid H_{i,k}, y_{i,k,<t}) / \pi_{\theta_{\text{old}}}(y_{i,k,t} \mid H_{i,k}, y_{i,k,<t})$, the clipped policy term is

$$\mathcal{L}_{\text{clip}} = -\mathbb{E}_{i,k,t}[\min(\rho_{i,k,t} A_i,\ \text{clip}(\rho_{i,k,t}, 1 - \epsilon_{\text{clip}}, 1 + \epsilon_{\text{clip}}) A_i)]\,. \tag{1}$$

We use $\mathcal{L}_{\text{GRPO}} = \mathcal{L}_{\text{clip}} + \beta_{\text{KL}} \mathcal{L}_{\text{KL}}$ as the base objective with reference-policy KL regularization. Each turn $u = (i, k)$ inherits $A_u = A_i$; we likewise write $H_u = H_{i,k}$, $y_u = y_{i,k}$, and $L_u = L_{i,k}$. The signal S_i is used only to construct empirical action–outcome evidence.

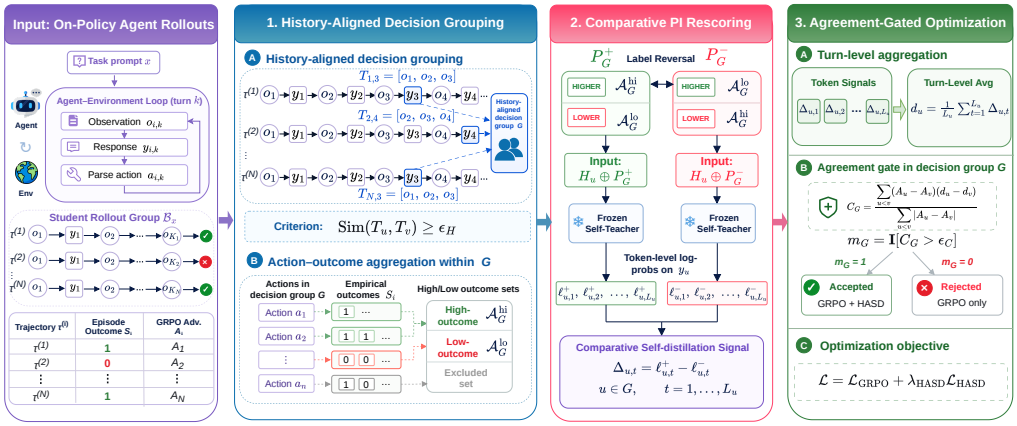


Figure 2: Overview of HASD. Starting from on-policy rollouts, HASD (1) groups turns by bounded Observation Traces and aggregates action–outcome evidence, (2) constructs comparative PI and rescores sampled responses with a frozen self-teacher, and (3) retains the self-distillation signal only when its turn-level ordering agrees with the GRPO-advantage ordering.

On-Policy Self-Distillation. In OPSD, after the student samples a response, its frozen self-teacher conditions on privileged information $P_{i,k}$ and provides token-level distributional supervision along the same response (Zhao et al., 2026). A generic OPSD objective is

$$\mathcal{L}_{\mathrm{OPSD}} = \frac{1}{\sum_{i,k} L_{i,k}} \sum_{i,k,t} D(\mathrm{sg}[\pi_{\theta_{\mathrm{old}}}(\cdot \mid P_{i,k}, H_{i,k}, y_{i,k,<t})] \,\|\, \pi_\theta(\cdot \mid H_{i,k}, y_{i,k,<t})), \tag{2}$$

where $P_{i,k}$ denotes PI available only to the self-teacher for turn k, with trajectory-level PI shared across k. D is a token-level distribution divergence and sg denotes the stop-gradient operation. Thus, $\mathcal{L}_{\mathrm{OPSD}}$ provides dense token-level supervision without introducing off-policy targets.

4 Method: History-Aligned Self-Distillation

Figure 2 summarizes the three-stage HASD pipeline. The key idea is to convert action–outcome contrasts among history-aligned turns into token-level self-distillation signals and to filter them with outcome-grounded GRPO advantages. Algorithm 1 in Appendix A gives the complete pseudocode.

4.1 History-Aligned Decision Grouping and Action–Outcome Aggregation

History-aligned decision grouping. In long-horizon agent interaction, the decision-time context $H_{i,k}$ defined in Section 3 includes a bounded window of recent interaction history, which can preserve decision-relevant context without replaying the full history. Following this setup, HASD extracts a bounded, observation-only sequence from $H_{i,k}$ for each turn (i,k), which we call the Observation Trace:

$$T_{i,k} = \left(o_{i,\max(1,k-w)}, \ldots, o_{i,k}\right), \tag{3}$$

where w is the maximum number of prior observations retained before the current observation $o_{i,k}$ and is set by default to match the recent-history window in $H_{i,k}$.

For two same-task decision turns $u = (i,k)$ and $v = (j,\ell)$, let $T_u = T_{i,k}$ and $T_v = T_{j,\ell}$. We compute the similarity between their Observation Traces as

$$\mathrm{Sim}(T_u, T_v) = \frac{1-\alpha_H}{1-\alpha_H^{M_{uv}}} \sum_{r=0}^{K_{uv}-1} \alpha_H^r \, s(\tilde{o}_{i,k-r}, \tilde{o}_{j,\ell-r}), \tag{4}$$

where $K_{uv} = \min(|T_u|, |T_v|)$, $M_{uv} = \max(|T_u|, |T_v|)$, $\alpha_H \in (0,1)$ is the recency decay factor, $\tilde{o}$ is a normalized observation, and $s(\cdot,\cdot) \in [0,1]$ is the `SequenceMatcher` ratio.

Eq. (4) compares traces backward from their most recent observations, weighting recent pairs more heavily and geometrically discounting earlier ones. Identical traces attain $\mathrm{Sim} = 1$, whereas for unequal lengths, normalization penalizes unmatched older observations even when the overlapping suffix matches well. Each parser-valid turn $u = (i, k)$ joins the first same-task decision group G whose representative turn v satisfies $\mathrm{Sim}(T_u, T_v) \geq \epsilon_H$, with ϵ_H denoting the history-matching threshold, or starts a new group otherwise. Let $\mathcal{G}$ denote the resulting collection of decision groups.

Action–outcome aggregation. For a decision group $G \in \mathcal{G}$, let $\mathcal{A}_G = \{a_{i,k} : (i, k) \in G\}$ denote the distinct normalized actions associated with its turns. For each $a \in \mathcal{A}_G$, let $\mathcal{I}_G(a) = \{(i, k) \in G : a_{i,k} = a\}$ denote its occurrences. We estimate its mean empirical outcome as

$$q_G(a) = \frac{1}{|\mathcal{I}_G(a)|} \sum_{(i,k)\in\mathcal{I}_G(a)} S_i. \tag{5}$$

Thus, $q_G(a)$ measures the average episode success associated with action a under similar pre-action decision contexts within decision group G. It is used only as empirical outcome evidence, rather than as a learned or causal action value. Then we retain the actions with the highest and lowest empirical outcomes:

$$\mathcal{A}_G^{\mathrm{hi}} = \arg\max_{a\in\mathcal{A}_G} q_G(a), \qquad \mathcal{A}_G^{\mathrm{lo}} = \arg\min_{a\in\mathcal{A}_G} q_G(a). \tag{6}$$

If multiple actions attain the maximum or minimum $q_G(a)$, all tied actions are retained without tie-breaking, while intermediate actions are excluded. We define $\mathcal{G}' \subseteq \mathcal{G}$ as the groups with at least two distinct actions with different q_G values and whose instantiated PI pair fits the token budget, and use only $\mathcal{G}'$ for comparative rescoring. Appendix E.1 provides an empirical analysis of decision-group structure and of representative matching based on Eq. (4), while Appendix E.2 further examines the roles of cross-trajectory and self-derived evidence in PI construction and training using a leave-one-trajectory-out analysis.

4.2 Comparative PI Construction and Rescoring

Building on Eq. (2), we adopt the sampled-token OPSD formulation commonly used in agentic OPSD (Lu et al., 2026; Yang et al., 2026c), which evaluates PI-induced log-probability changes at student-sampled tokens. For a fixed PI view, these log-probability changes have a non-positive expectation under student sampling, which we call the negative tendency. In finite samples, this tendency can contribute a common negative-mean component to the objective, as analyzed in Appendix D.1. This motivates us to contrast two PI-conditioned views of shared evidence and use their score difference as the token-level signal.

Comparative PI construction. For each $G \in \mathcal{G}'$, we construct a comparative PI pair by reversing the high/low labels:

$$\begin{aligned} P_G^+ :&\quad \texttt{HIGHER} \leftarrow \mathcal{A}_G^{\mathrm{hi}}, \quad \texttt{LOWER} \leftarrow \mathcal{A}_G^{\mathrm{lo}}, \\ P_G^- :&\quad \texttt{HIGHER} \leftarrow \mathcal{A}_G^{\mathrm{lo}}, \quad \texttt{LOWER} \leftarrow \mathcal{A}_G^{\mathrm{hi}}. \end{aligned} \tag{7}$$

We call P_G^+ the original-label view and P_G^- the label-reversed view. They contain the same action evidence and differ only in the HIGHER/LOWER label assignment. Appendix C provides the shared prompt template, the P_G^+/P_G^- field assignments, and a case study of comparative PI construction.

Comparative self-teacher rescoring. With P_G^+ and P_G^- as the two PI views, the same frozen rollout policy $\pi_{\theta_{\mathrm{old}}}$ serves as the self-teacher and rescores each sampled response. For $\sigma \in \{+, -\}$, let $\ell_{u,t}^{\sigma} = \log \pi_{\theta_{\mathrm{old}}}(y_{u,t} \mid P_G^{\sigma}, H_u, y_{u,<t})$ denote the PI-conditioned teacher score, and let $\ell_{u,t}^{0} = \log \pi_{\theta_{\mathrm{old}}}(y_{u,t} \mid H_u, y_{u,<t})$ be the shared PI-free score. The corresponding one-sided signal is $g_{u,t}^{\sigma} = \ell_{u,t}^{\sigma} - \ell_{u,t}^{0}$. We define the comparative token-level signal as

$$\Delta_{u,t} = g_{u,t}^{+} - g_{u,t}^{-} = \ell_{u,t}^{+} - \ell_{u,t}^{-} = \log \frac{\pi_{\theta_{\mathrm{old}}}(y_{u,t} \mid P_G^{+}, H_u, y_{u,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{u,t} \mid P_G^{-}, H_u, y_{u,<t})}. \tag{8}$$

Thus, the shared PI-free reference term cancels exactly, while the value of $\Delta_{u,t}$ reflects the self-teacher's relative preference for $y_{u,t}$ between the original-label and label-reversed views.

4.3 AGREEMENT-GATED SELF-DISTILLATION

Turn-level agreement gate. Although the comparative PI is constructed from empirical outcome evidence, the resulting signal is not directly verified against task outcomes. We therefore use detached GRPO advantages as an outcome-grounded reference and gate the comparative signal by the agreement of its turn-level scores with the advantages within each $G \in \mathcal{G}'$. For each turn u, let $d_u = \frac{1}{L_u}\sum_{t=1}^{L_u} \Delta_{u,t}$ be its mean comparative score, and let $W_G = \sum_{u<v} |A_u - A_v|$ for the group. For $W_G > 0$, define the agreement score as

$$C_G = \frac{1}{W_G} \sum_{u<v} (A_u - A_v)(d_u - d_v). \tag{9}$$

Here, W_G normalizes pairwise advantage contrast in G. Positive C_G indicates aggregate ordering agreement between GRPO advantages and comparative scores. Using the agreement threshold ϵ_C, we set the binary gate $m_G = \mathbb{I}[C_G > \epsilon_C]$ and $m_G = 0$ when $W_G = 0$.

Training objective. Let $Z_{\mathrm{HASD}} = \sum_{G\in\mathcal{G}'} m_G \sum_{u\in G} L_u$ be the number of tokens retained by the gate. For $Z_{\mathrm{HASD}} > 0$, the HASD self-distillation objective is

$$\mathcal{L}_{\mathrm{HASD}} = -\frac{1}{Z_{\mathrm{HASD}}} \sum_{G\in\mathcal{G}'} m_G \sum_{u\in G} \sum_{t=1}^{L_u} \mathrm{sg}(\Delta_{u,t}) \log \pi_\theta(y_{u,t} \mid H_u, y_{u,<t}). \tag{10}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{GRPO}} + \lambda_{\mathrm{HASD}} \mathcal{L}_{\mathrm{HASD}}. \tag{11}$$

All self-teacher-derived scores and gating quantities are detached during optimization; at inference, only π_θ is retained. Appendix F further analyzes HASD's training efficiency by reporting its operation breakdowns and wall-clock cost per training step.

Theoretical properties of agreement gating. We next characterize the agreement score defined in Eq. (9) and establish the robustness of the resulting gate. Complete derivations are provided in Appendix D.2. For $G \in \mathcal{G}'$, let $A = (A_u)_{u\in G}$, $d = (d_u)_{u\in G}$, and $n_G = |G|$. We use C_G as shorthand for $C_G(A, d)$. With group means $\bar{A}_G$ and $\bar{d}_G$, define the within-group covariance as

$$\mathrm{Cov}(A, d) = \frac{1}{n_G} \sum_{u\in G} (A_u - \bar{A}_G)(d_u - \bar{d}_G).$$

Lemma 1 (Pairwise ordering characterization)**.** *For any decision group $G \in \mathcal{G}'$ with $W_G > 0$,*

$$C_G = \sum_{u<v} \frac{|A_u - A_v|}{W_G} \mathrm{sgn}(A_u - A_v)(d_u - d_v) = \frac{n_G^2}{W_G} \mathrm{Cov}(A, d). \tag{12}$$

Eq. (12) expresses C_G as both a weighted average of pairwise margins and a scaled covariance between GRPO advantages and comparative scores.

Proposition 1 (Robustness of the agreement score)**.** *For any additive shift $b \in \mathbb{R}$ shared across a decision group,*

$$C_G(A, d + b\mathbf{1}) = C_G(A, d), \tag{13}$$

where $\mathbf{1}$ is the all-ones vector. For an arbitrary token-level additive perturbation $\varepsilon_{u,t} \in \mathbb{R}$, let

$$\widetilde{\Delta}_{u,t} = \Delta_{u,t} + \varepsilon_{u,t}, \qquad \widetilde{d}_u = \frac{1}{L_u}\sum_{t=1}^{L_u} \widetilde{\Delta}_{u,t} = d_u + \xi_u, \quad \xi_u = \frac{1}{L_u}\sum_{t=1}^{L_u} \varepsilon_{u,t}. \tag{14}$$

Then, for $\xi = (\xi_u)_{u\in G}$,

$$|C_G(A, d + \xi) - C_G(A, d)| \le 2\|\xi\|_\infty. \tag{15}$$

Prior work shows that PI may induce systematic shifts in OPSD signals (He et al., 2026). Eq. (13) shows that any additive shift shared within a decision group under the same PI cancels exactly in C_G. Such signals may also be unreliable (Lu et al., 2026; Nguyen et al., 2026). Eq. (15) bounds the effect of local perturbations, with the gate decision unchanged when $|C_G - \epsilon_C| > 2\|\xi\|_\infty$.

Table 1: Main results across three backbones. We report Search-QA accuracy (%), WebShop success rate (%) with task score included for reference, and ALFWorld success rate (%). Bold and underlined values are the best and second-best distinct results in shaded columns; dashes denote unavailable results; methods with no available results for an entire backbone are omitted.

	Search-QA								WebShop		ALFWorld						
Method	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg	Score	Succ	Pick	Look	Clean	Heat	Cool	Pick2	Avg
Qwen3-1.7B																	
Vanilla	29.4	46.9	37.0	23.5	19.6	6.4	10.5	24.8	46.5	4.7	25.0	22.2	3.1	0.0	21.4	4.2	12.5
Skill-Prompt	29.4	46.5	36.2	22.9	20.8	4.3	10.1	24.3	23.0	2.3	10.3	50.0	16.1	0.0	0.0	5.0	9.4
GRPO	39.9	57.3	43.9	34.8	32.1	11.3	64.9	40.6	76.1	57.0	64.5	36.4	65.4	53.3	41.2	60.7	57.0
Skill-SD	39.1	57.5	45.4	34.8	34.1	10.7	64.1	40.8	81.8	53.9	52.9	37.5	69.2	42.9	60.0	36.8	52.3
RLSD	38.6	57.3	43.0	34.5	34.1	11.5	65.3	40.6	74.0	50.8	50.0	37.5	61.5	19.0	50.0	21.1	42.2
SDAR	39.7	58.9	45.3	35.9	35.5	12.6	65.3	<u>41.9</u>	76.8	58.6	73.5	25.0	76.9	33.3	40.0	36.8	53.9
SERL	–	–	–	–	–	–	–	–	81.5	<u>67.2</u>	70.0	37.5	40.0	93.8	76.9	21.7	<u>57.8</u>
StepOPSD	40.6	59.4	44.4	37.1	32.0	11.6	64.1	41.3	–	–	64.7	44.4	56.5	60.9	42.1	55.0	56.3
HASD	41.4	59.5	45.9	38.5	38.8	14.1	67.0	**43.6**	83.7	**71.1**	74.2	36.4	92.3	86.7	70.6	57.1	**71.9**
Qwen2.5-3B-Instruct																	
Vanilla	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7	6.7	0.8	44.4	11.1	6.2	15.4	28.6	12.5	21.9
Skill-Prompt	23.7	46.2	30.6	24.4	22.1	7.5	12.5	23.9	0.2	0.8	51.7	66.7	48.4	0.0	4.3	10.0	28.9
GRPO	44.7	61.1	46.6	39.4	40.9	15.4	29.6	39.7	70.0	60.2	93.5	90.9	92.3	53.3	82.4	71.4	82.0
Skill-SD	44.4	60.4	44.0	39.5	40.4	15.4	64.9	44.1	75.9	64.0	88.2	50.0	96.2	52.4	65.0	57.9	73.4
RLSD	41.5	58.6	42.3	40.4	40.2	16.8	66.9	43.8	84.4	66.4	87.9	75.0	90.9	75.0	73.1	68.4	79.7
SDAR	44.8	58.1	44.3	38.6	36.2	15.7	66.1	43.4	85.0	<u>68.0</u>	97.1	62.5	100.0	61.9	75.0	84.2	<u>84.4</u>
SERL	–	–	–	–	–	–	–	–	85.9	63.3	100.0	75.0	84.0	93.8	69.2	43.5	78.1
StepOPSD	45.0	61.6	46.2	39.5	39.5	14.4	65.3	<u>44.5</u>	–	–	97.1	66.7	87.0	79.1	78.9	95.0	83.6
HASD	44.5	61.9	45.5	40.6	41.3	15.0	66.5	**45.0**	88.7	**77.3**	97.0	87.5	96.2	91.7	84.6	73.9	**89.1**
Qwen2.5-7B-Instruct																	
Vanilla	25.2	50.8	29.5	29.0	29.0	10.4	63.7	33.9	5.9	1.6	36.1	22.2	3.1	0.0	0.0	0.0	12.5
Skill-Prompt	30.9	52.1	32.7	32.7	27.9	12.7	66.1	36.4	1.7	0.8	51.7	50.0	32.3	5.3	4.3	0.0	23.4
GRPO	45.0	64.6	46.0	45.0	45.0	19.0	47.2	44.5	85.6	74.2	91.9	58.8	100.0	84.6	76.2	78.9	83.6
Skill-SD	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5	93.9	93.8	90.9	100.0	69.2	68.4	85.1
RLSD	46.8	63.0	44.4	45.5	48.9	21.5	73.0	**49.0**	87.4	77.3	100.0	87.5	92.3	58.8	80.0	65.2	82.0
SDAR	46.3	63.5	48.2	43.8	48.4	19.6	73.0	**49.0**	89.4	**82.8**	94.7	75.0	100.0	86.7	68.2	78.9	85.9
SERL	–	–	–	–	–	–	–	–	84.3	75.8	97.0	83.3	93.5	100.0	68.2	81.8	<u>87.5</u>
HASD	47.2	64.9	47.7	44.4	45.4	20.4	71.8	<u>48.8</u>	88.2	<u>78.9</u>	97.0	100.0	96.2	91.7	88.5	91.3	**93.8**

5 Experiments

5.1 Experimental Setup

Benchmarks and metrics. We evaluate HASD on ALFWorld (Shridhar et al., 2021), WebShop (Yao et al., 2022), and Search-QA following Search-R1 (Jin et al., 2025), covering household interaction, web shopping, and search-based question answering. We report success rates on ALFWorld and WebShop and mean accuracy over seven Search-QA datasets. See Appendix B.1 for details.

Models and baselines. We use Qwen3-1.7B (Yang et al., 2025) and Qwen2.5-Instruct at 3B and 7B scales (Yang et al., 2024a). Baselines include **training-free and outcome-based RL**: Vanilla, Skill-Prompt, and GRPO (Shao et al., 2024); **trajectory-level OPSD**: Skill-SD (Wang et al., 2026a), RLSD (Yang et al., 2026a), and SDAR (Lu et al., 2026); and **step-level OPSD**: SERL (Li et al., 2026b) and StepOPSD (Zhang et al., 2026). See Appendices B.2–B.3 for baseline and implementation details. We use $\epsilon_H = 0.7$ and $\epsilon_C = 0$ by default; see Appendix E.3 for sensitivity analysis.

5.2 Main Results

HASD achieves the strongest overall performance. Table 1 shows that HASD ranks first in seven of nine settings based on aggregate performance. With Qwen3-1.7B, HASD reaches 43.6 on Search-QA, 71.1 on WebShop, and 71.9 on ALFWorld, exceeding the second-best results by 1.7, 3.9, and 14.1 points. It also leads in all three aggregates at 3B. At 7B, HASD achieves the best ALFWorld result while remaining competitive on Search-QA and WebShop.

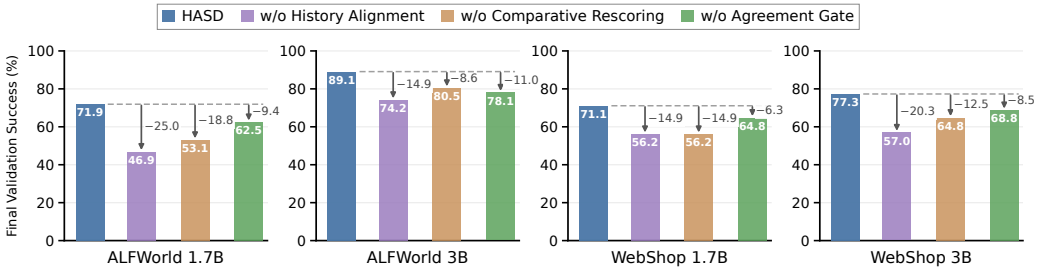


Figure 3: Component ablations with 1.7B and 3B backbones on ALFWorld and WebShop. Bars show validation success rates (%) at the final checkpoint.

HASD improves over both trajectory- and step-level OPSD. At 1.7B and 3B, HASD outperforms trajectory-level OPSD baselines Skill-SD, RLSD, and SDAR across all aggregate metrics, exceeding the strongest baseline by 18.0 points on ALFWorld and 12.5 points on WebShop at 1.7B. It also outperforms step-level baselines SERL and StepOPSD on all available aggregates, exceeding the stronger baseline by 14.1 points on ALFWorld at 1.7B and 14.0 points on WebShop at 3B. These results support history-aligned local supervision for addressing granularity-dependent limitations.

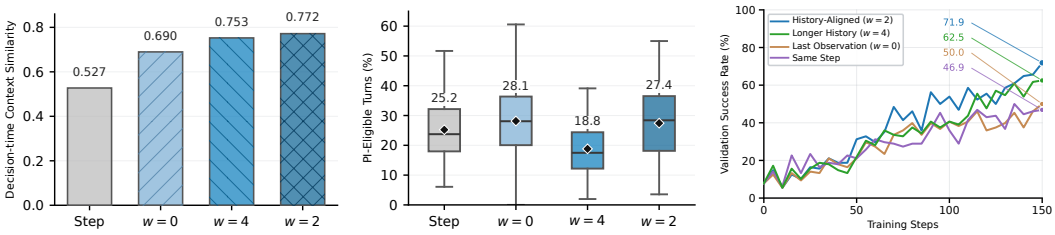
HASD yields larger gains on smaller backbones. Table 1 shows that HASD achieves larger gains over GRPO with Qwen3-1.7B than with larger backbones: 10.7 points on average across three aggregate metrics, compared with 9.8 points at 3B and 6.4 points at 7B. These results suggest that smaller backbones benefit more from the history-aligned and selectively filtered PI provided by HASD. Appendix E.4 further reports a multi-seed evaluation showing that HASD's gains persist across seeds in representative settings, supporting the stability of the main results.

5.3 Component Ablations

Figure 3 ablates the three core components of HASD on ALFWorld and WebShop with 1.7B and 3B backbones. The w/o History Alignment variant replaces history-aligned grouping with same-step grouping; the w/o Comparative Rescoring variant replaces the comparative signal with the one-sided signal from the original-label PI view; and the w/o Agreement Gate variant removes the gate and applies the comparative signal to all eligible decision groups. Across all four settings, HASD achieves the highest final validation success rate. Averaged over the two model scales, removing history alignment causes the largest performance drops, reaching 20.0 points on ALFWorld and 17.6 on WebShop, highlighting the importance of comparing turns under matched decision contexts. Removing comparative rescoring also causes an average drop of 13.7 points in both environments, supporting its advantage over one-sided PI rescoring. Removing the agreement gate produces a smaller but consistent decline, reducing the average final validation success to 70.3 on ALFWorld and 66.8 on WebShop, supporting the benefit of outcome-grounded gating for auxiliary self-distillation. Together, these components make complementary contributions to the full performance of HASD.

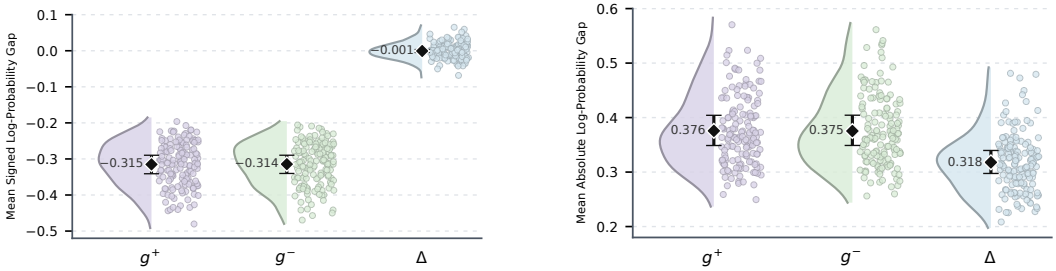
5.4 Further Analyses

Does history alignment balance comparability and PI availability? Figure 4 compares four grouping strategies: Same Step, Last Observation ($w = 0$), Longer History ($w = 4$), and History-Aligned ($w = 2$) used in HASD. Same Step groups same-task turns by step index without history matching, whereas the other three use the similarity-based grouping rule in Eq. (4) with the same threshold ϵ_H, differing only in w, the number of prior observations retained in the Observation Trace. The $w = 2$ setting matches the recent-history window used in the decision-time context. Figure 4(a) evaluates decision comparability using pairwise similarity between full decision-time contexts, as detailed in Appendix B.3. On average, $w = 2$ achieves the highest similarity at 77.2%, suggesting better decision comparability. Figure 4(b) shows it retains 27.4% PI-eligible turn coverage, close to 28.1% for $w = 0$ and above 18.8% for $w = 4$. Figure 4(c) shows that $w = 2$ attains the highest final validation success rate of 71.9, compared with 62.5 for $w = 4$, 50.0 for $w = 0$, and 46.9 for Same Step. These results indicate that aligning the Observation Trace with the decision-time context best balances decision comparability and PI availability, yielding the strongest validation performance.



(a) Decision-time context similarity (b) PI-eligible turn coverage (c) Validation success rate

Figure 4: History-alignment analysis on ALFWorld with Qwen3-1.7B over 150 training steps. (a) Mean pairwise similarity of full decision-time contexts within each decision group. (b) Distribution of PI-eligible turn coverage across training steps, defined as the fraction of turns for which comparative PI can be constructed. (c) Validation success rate over training.



(a) Signed gap (b) Absolute gap

Figure 5: Empirical analysis of one-sided and comparative PI rescoring on Search-QA with Qwen3-1.7B over 150 training steps. (a) Distributions of per-step mean signed log-probability gaps. (b) Distributions of per-step mean absolute log-probability gaps. Each point denotes one training step.

Does comparative PI rescoring mitigate the one-sided negative tendency? Figure 5 compares the one-sided signals g^+ and g^- from the original-label and label-reversed views with the comparative signal $\Delta = g^+ - g^-$. Figure 5(a) shows that, averaged over the 150 training steps, g^+ and g^- have signed means of -0.315 and -0.314, respectively, with their per-step means remaining below -0.2. In contrast, Δ has a mean of -0.001, whose magnitude is 99.68% smaller than the average magnitude of the two one-sided means. This empirical pattern is consistent with the theoretical analysis: each one-sided signal has a non-positive conditional expectation, whereas the comparative signal has no fixed-sign expectation. Although Δ has a near-zero signed mean, the absolute-value analysis shows that Δ is not uniformly small. Figure 5(b) shows that Δ has a mean absolute value of 0.318, compared with 0.376 for g^+ and 0.375 for g^-, meaning that Δ retains approximately 84.69% of the average one-sided absolute magnitude. These results show that comparative PI rescoring effectively mitigates the one-sided negative tendency while retaining substantial signal magnitude.

6 Conclusion

We studied self-distillation supervision for long-horizon agents from the perspective of supervision quality at individual decision turns. Our findings show that effective local supervision depends not only on its granularity, but also on whether its supporting experience matches the target decision context and provides sufficiently informative evidence about action quality. To address these challenges, we introduced HASD, a history-aligned self-distillation framework that groups comparable turns using bounded Observation Traces, converts the resulting decision contrasts into token-level supervision through comparative PI rescoring, and retains this supervision only when its turn-level ordering agrees with detached GRPO advantages. Across environments and model scales, HASD demonstrates strong performance with modest training overhead, while ablations confirm the effectiveness of its core components and further analyses substantiate the roles of history alignment and comparative rescoring. Future work may explore adaptive history representations and continuous aggregation from graded or partial rewards to improve reliability and coverage.

AI USE STATEMENT

In this work, generative AI tools were used to assist with the writing and presentation of mathematical proofs, as well as with translation, language polishing, and grammar checking to improve clarity and readability. All AI-assisted text was carefully reviewed and verified by the authors. The mathematical arguments and derivations themselves were developed by the authors. Generative AI tools were not used to develop the theoretical or conceptual framework of the work, formulate mathematical claims, provide critical ingredients for proofs, propose or refine research hypotheses, design or provide feedback on the methodology or experiments, implement the proposed method, or interpret experimental results. Synthetic dataset generation, dataset cleaning or reformatting, and qualitative or thematic data analysis were not applicable to this work. The authors take full responsibility for the final content of the paper and all AI-assisted text.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the main method and experimental details in the paper: Section 4 defines HASD, Section 5.1 summarizes the experimental setup, and Appendices A–C provide the training algorithm, implementation details, and exact PI prompt. Core implementation code and configurations are included in the supplementary materials and the full code will be publicly released. All benchmarks and pretrained backbone checkpoints are publicly available.

REFERENCES

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

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2605.26684.

Lucio M. Dery, Yann Dauphin, and David Grangier. Auxiliary task update decomposition: The good, the bad and the neutral. In *International Conference on Learning Representations*, 2021.

Yifeng Ding, Hung Le, Songyang Han, Kangrui Ruan, Zhenghui Jin, Varun Kumar, Zijian Wang, and Anoop Deoras. Empowering multi-turn tool-integrated agentic reasoning with group turn policy optimization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42409–42423. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1962.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity, 2018. URL https://arxiv.org/abs/1812.02224.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for LLM agent training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 51797–51830, 2025. doi: 10.52202/085713-1544.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Gheshlaghi Azar, Bilal Piot, Nicolas Heess, Hado P. van Hasselt, Gregory Wayne, Satinder Singh, Doina Precup, and Remi Munos. Hindsight credit assignment. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12467–12476, 2019.

Qiangqiang He, Zhongheng Wu, and ZiJian Wang. Not all tokens deserve equal credit: Counterfactual sensitivity credit reallocation for Long-CoT reasoning, 2026. URL https://arxiv.org/abs/2607.27888.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580.

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, 2025. URL https://arxiv.org/abs/2503.09516.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

Jiazheng Li, Yawei Wang, Qiaojing Yan, Yijun Tian, Zhichao Xu, Huan Song, Panpan Xu, and Lin Lee Cheong. SALT: Step-level advantage assignment for long-horizon agents via trajectory graph. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4709–4725. Association for Computational Linguistics, 2026a. doi: 10.18653/v1/2026.findings-eacl.247.

Xiaozhe Li, Tianyi Lyu, Yang Li, Yichuan Ma, Peiji Li, Linyang Li, Qipeng Guo, Dahua Lin, and Kai Chen. What and when to distill: Selective hindsight distillation for multi-turn agents, 2026b. URL https://arxiv.org/abs/2605.19447.

Yugu Li, Zehong Cao, Jianglin Qiao, and Siyi Hu. SSVPO: Effective step-level credit assignment for RL training of language models. In *International Conference on Learning Representations*, 2026c.

Dongyi Liu, Yifan Niu, Qinwen Wang, Han Xiao, and Jia Li. PiCA: Pivot-based credit assignment for search agentic reinforcement learning, 2026a. URL https://arxiv.org/abs/2605.09287.

Haoran Liu, Yuwei Zhang, Xiyao Li, Bohan Lyu, and Jingbo Shang. HERO: Hindsight-enhanced reflection from environment observations for agentic self-distillation, 2026b. URL https://arxiv.org/abs/2606.11559.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. Agentic reinforcement learning with implicit step rewards. In *International Conference on Learning Representations*, 2026c.

Yuze Liu, Chaochao Lu, and Chao Yang. SHADOW: Dynamic-aware credit assignment against long-horizon tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23935–23944, 2026d. doi: 10.1609/aaai.v40i28.39570.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Self-distilled agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2605.15155.

Kangyang Luo, Shuzheng Si, Yuzhuo Bai, Cheng Gao, Zhitong Wang, Cheng Huang, Yingli Shen, Yufeng Han, Wenhao Li, Cunliang Kong, and Maosong Sun. MEIC-DT: Memory-efficient incremental clustering for long-text coreference resolution with dual-threshold constraints. In *Findings*

of the Association for Computational Linguistics: ACL 2026, pp. 17787–17804. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.881.

Yufei Ma, Zihan Liang, Ben Chen, Zhipeng Qian, Huangyu Dai, Lingtao Mao, Xuxin Zhang, Chenyi Lei, and Wenwu Ou. SD-Search: On-policy hindsight self-distillation for search-augmented reasoning, 2026. URL https://arxiv.org/abs/2605.18299.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546.

Zibin Meng and Kani Chen. CRAFT: Counterfactual credit assignment from free sibling rollouts for self-distilled agentic reinforcement learning, 2026. URL https://arxiv.org/abs/2606.29476.

Thomas Mesnard, Theophane Weber, Fabio Viola, Shantanu Thakoor, Alaa Saade, Anna Harutyunyan, Will Dabney, Thomas S. Stepleton, Nicolas Heess, Arthur Guez, Eric Moulines, Marcus Hutter, Lars Buesing, and Remi Munos. Counterfactual credit assignment in model-free reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7654–7664. PMLR, 2021.

Xuan-Phi Nguyen, Zeyu Leo Liu, Yang Li, Shrey Pandit, Yiran Zhao, Anurag Koul, and Shafiq Joty. Privileged likelihood is not automatically value: Three checks for token credit in on-policy self-distillation, 2026. URL https://arxiv.org/abs/2608.09263.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged information distillation for language models, 2026. URL https://arxiv.org/abs/2602.04942.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.378.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551, 2023. doi: 10.52202/075280-2997.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297. ACM, 2025. doi: 10.1145/3689031.3696075.

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

Hui-Ze Tan, Xiao-Wen Yang, Hao Chen, Jie-Jing Shao, Yi Wen, Yuteng Shen, Weihong Luo, Xiku Du, Lan-Zhe Guo, and Yu-Feng Li. Hindsight credit assignment for long-horizon LLM agents, 2026a. URL https://arxiv.org/abs/2603.08754.

Qitai Tan, Zefang Zong, Mo Li, Yipeng Shi, Yang Li, and Peng Chen. ATOD: Annealed turn-aware on-policy distillation for multi-turn agentic tasks, 2026b. URL https://arxiv.org/abs/2606.27814.

Qiaoyu Tang, Hao Xiang, Le Yu, Bowen Yu, Yaojie Lu, Xianpei Han, Le Sun, WenJuan Zhang, Pengbo Wang, Shixuan Liu, Zhenru Zhang, Jianhong Tu, Hongyu Lin, and Junyang Lin. Beyond turn limits: Training deep search agents with dynamic context window, 2025. URL https://arxiv.org/abs/2510.08276.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and Honggang Qi. Skill-SD: Skill-conditioned self-distillation for multi-turn LLM agents, 2026a. URL https://arxiv.org/abs/2604.10674.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. doi: 10.1007/s11704-024-40231-1.

Tao Wang, Suhang Zheng, and Xiaoxiao Xu. RTMC: Step-level credit assignment via rollout trees, 2026b. URL https://arxiv.org/abs/2604.11037.

Quan Wei, Siliang Zeng, Chenliang Li, Zhongruo Wang, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and Mingyi Hong. Reinforcing multi-turn reasoning in LLM agents via fine-grained reward structure and credit assignment, 2025a. URL https://arxiv.org/abs/2505.11821.

Yumou Wei, Paulo Carvalho, and John Stamper. KCluster: An LLM-based clustering approach to knowledge component discovery. In *Proceedings of the 18th International Conference on Educational Data Mining*, pp. 228–240. International Educational Data Mining Society, 2025b. doi: 10.5281/zenodo.15870197.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, Qi Zhang, and Tao Gui. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68:121101, 2025. doi: 10.1007/s11432-024-4222-0.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024a. URL https://arxiv.org/abs/2412.15115.

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, 2025. URL https://arxiv.org/abs/2505.09388.

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

Haojin Yang, Ai Jian, Yiwei Wang, Xinyue Huang, Weipeng Zhang, Ke Zeng, Xunliang Cai, and Jingqing Ruan. Harmonizing dense and sparse signals in multi-turn RL: Dual-horizon credit assignment for industrial sales agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 1067–1076. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-industry.74.

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. In *Advances in Neural Information Processing Systems*, volume 37, pp. 50528–50652, 2024b. doi: 10.52202/079017-1601.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and Jianhua Tao. OPID: On-policy skill distillation for agentic reinforcement learning, 2026c. URL https://arxiv.org/abs/2606.26790.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757, 2022. doi: 10.52202/068431-1508.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836. Curran Associates, Inc., 2020.

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

Zijing Zhang and Xiajie Yang. CRPS: Curriculum replay via progressive suffixes from successful trajectories for long-horizon LLM agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13891–13904. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.680.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026. URL https://arxiv.org/abs/2601.18734.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. PyTorch FSDP: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12): 3848–3860, 2023. doi: 10.14778/3611540.3611569.

Xuhui Zheng, Kang An, Ziliang Wang, Yuhang Wang, and Yichao Wu. StepSearch: Igniting LLMs search ability via step-wise proximal policy optimization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21805–21830. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1106.

Zefang Zong, Dingwei Chen, Yang Li, Qi Yi, Bo Zhou, Chengming Li, BO Qian, Peng Chen, and Jie Jiang. AT^{2}PO: Agentic turn-based policy optimization via tree search. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24120–24143. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1106.

A HASD ALGORITHM

Algorithm 1 The pseudocode of History-Aligned Self-Distillation

Require: actor π_θ; tasks; $w, \alpha_H, \epsilon_H, \epsilon_C$; λ_{HASD}
$\theta_{\text{old}} \leftarrow \theta$; sample rollout groups $\mathcal{B}_x$ with $\pi_{\theta_{\text{old}}}$
Record R_i and S_i; compute detached A_i and assign $A_u \leftarrow A_i$ to each turn $u = (i, k)$
$\mathcal{G} \leftarrow \emptyset, \quad \mathcal{G}' \leftarrow \emptyset$
for all tasks x **do**
    $\mathcal{G}_x \leftarrow \emptyset$
    for all parser-valid decision turns $u = (i, k)$ of x in stable order **do**
        Construct $T_u \equiv T_{i,k}$ by Eq. (3); normalize its observations for matching
        $G^\star \leftarrow$ first decision group $G \in \mathcal{G}_x$ whose representative turn v satisfies $\text{Sim}(T_u, T_v) \geq \epsilon_H$
        if $G^\star$ exists **then**
            Add u to $G^\star$
        else
            Create a new decision group represented by u
        end if
    end for
    $\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{G}_x$
end for
for all decision groups $G \in \mathcal{G}$ **do**
    Normalize parsed actions; compute $q_G(a)$ from binary S_i
    Construct the complete argmax/argmin sets $\mathcal{A}_G^{\text{hi}}$ and $\mathcal{A}_G^{\text{lo}}$
    if $|\mathcal{A}_G| < 2$ or $q_G(a)$ is constant over $\mathcal{A}_G$ **then**
        continue
    end if
    Instantiate P_G^+ and P_G^- for comparative PI rescoring
    if either P_G^+ or P_G^- exceeds the PI token budget **then**
        continue
    end if
    $\mathcal{G}' \leftarrow \mathcal{G}' \cup \{G\}$
    for all $u \in G$ **do**
        Teacher-force the sampled response with $\pi_{\theta_{\text{old}}}$ under P_G^+ and P_G^-
        Compute token contrasts $\Delta_{u,t}$ and turn-level comparative score d_u
    end for
    $W_G \leftarrow \sum_{u<v} |A_u - A_v|$
    if $W_G = 0$ **then**
        $m_G \leftarrow 0$
    else
        Compute C_G by Eq. (9); $m_G \leftarrow \mathbb{I}[C_G > \epsilon_C]$
    end if
end for
Update θ using Eq. (11); use zero auxiliary loss if no gated token remains

B EXPERIMENTAL AND IMPLEMENTATION DETAILS

This section provides the detailed experimental and implementation settings used in the primary experiments.

B.1 BENCHMARKS AND METRICS

Following SDAR (Lu et al., 2026), we evaluate HASD on three representative agentic benchmarks.

ALFWorld. ALFWorld (Shridhar et al., 2021) evaluates text-based household interaction across six task types: Pick, Look, Clean, Heat, Cool, and Pick2. We report the success rate for each task type and the overall success rate across all evaluation episodes, denoted by Avg in Table 1.

WebShop. WebShop (Yao et al., 2022) requires an agent to search for and purchase products that satisfy natural-language user constraints. We report both the task score, which measures the degree to which the selected product satisfies the requested attributes, and the binary success rate.

Search-QA. Search-QA follows the search-agent setting of Search-R1 (Jin et al., 2025). It reports accuracy on Natural Questions, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle (Kwiatkowski et al., 2019; Joshi et al., 2017; Mallen et al., 2023; Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2023). The Avg column in Table 1 is the unweighted mean accuracy over these seven datasets.

B.2 Models and Baselines

Model backbones. We evaluate HASD with Qwen3-1.7B (Yang et al., 2025), Qwen2.5-3B-Instruct, and Qwen2.5-7B-Instruct (Yang et al., 2024a). Thinking is disabled for Qwen3-1.7B.

Training-free and outcome-based RL baselines. Vanilla evaluates the corresponding backbone without any additional training. Skill-Prompt supplies skill guidance without parameter updates. GRPO (Shao et al., 2024) learns directly from environment outcomes and serves as the outcome-based RL baseline.

Trajectory-level OPSD baselines. Skill-SD (Wang et al., 2026a) summarizes completed trajectories into reusable skills used as trajectory-level PI for self-distillation. RLSD (Yang et al., 2026a) reweights token-level advantages without changing their signs, while its agentic instantiation in Lu et al. (2026) uses retrieved skills as PI. SDAR (Lu et al., 2026) retrieves task-relevant skills as trajectory-level PI and applies a sigmoid gate to teacher–student log-probability gaps.

Step-level OPSD baselines. SERL (Li et al., 2026b), in its primary step-level configuration, attaches environment hindsight, such as feedback after an action or later observations, to individual steps and uses it to adjust the updates of action tokens. StepOPSD (Zhang et al., 2026) independently rescores each step by injecting hindsight from a successful sibling trajectory at the corresponding step boundary, and uses the resulting log-probability gaps to shape step-level advantages.

B.3 Implementation Settings

Optimization and evaluation. We use verl (Sheng et al., 2025) with FSDP (Zhao et al., 2023) on a single node with eight NVIDIA A100 GPUs. vLLM (Kwon et al., 2023) rollouts use tensor parallelism of size 1 for the 1.7B and 3B models and size 2 for the 7B model. We evaluate Qwen3-1.7B, Qwen2.5-3B-Instruct, and Qwen2.5-7B-Instruct, with thinking disabled for Qwen3-1.7B. We use a rollout group size of 8, with temperature 1.0, top-$p = 1.0$, and a maximum response length of 512 tokens. Following common practice in recent implementations of multi-turn agentic RL (Li et al., 2026b; Lu et al., 2026), we compute GRPO reward normalization over the expanded turn rows within each task group. The primary experiments use seed 0, while Appendix E.4 repeats representative settings with three training seeds under a fixed evaluation configuration and task sequence. We use AdamW with a learning rate of 1×10^{-6}, one PPO epoch per update, $\epsilon_{\text{clip}} = 0.2$, $\beta_{\text{KL}} = 0.001$, and an invalid-action penalty coefficient of 0.1. We use 15 optimizer warm-up steps for Search-QA and none for ALFWorld or WebShop.

For ALFWorld, WebShop, and Search-QA, respectively, the training batch sizes are $16/16/128$ task prompts, maximum prompt lengths are $2048/4096/4096$, PPO mini-batch sizes are $256/64/256$ expanded turn rows, and interaction horizons are $50/15/4$. Search-QA follows the Search-R1 retrieval configuration with E5 retrieval. ALFWorld and WebShop are evaluated every five steps at temperature 0.4 and top-$p = 1.0$, using 128 sampled environment episodes per validation checkpoint. Search-QA is evaluated only at the final step with a validation batch size of 512. All main runs use 150 optimization steps. Curves show raw measurements unless otherwise specified, with MA10 denoting a ten-step moving average.

HASD-specific settings. The HASD implementation follows the three-stage procedure described in Section 4. Here we report the additional choices needed for reproduction. For history matching, each parser-valid turn is assigned to the first matching representative from the same task using its normalized pre-action Observation Trace, with turns scanned in the current batch row order and each group's representative fixed as the turn that instantiated it. Observations are normalized by removing the outer information wrapper when present, lowercasing, trimming, and normalizing whitespace.

A turn is parser-valid when the task-specific action parser successfully extracts the expected action block, namely `<search>` or `<answer>` for Search-QA and `<action>` for ALFWorld/WebShop, and is marked valid during rollout processing. For outcome aggregation, actions are canonicalized using their action type and the extracted content, with the latter lowercased, trimmed, and whitespace-normalized. Invalid-format penalties affect GRPO advantages and the agreement gate, but not S_i or $q_G(a)$. Following prior work (Feng et al., 2025; Zhang & Yang, 2026), we use $w = 2$ for ALFWorld and WebShop and the full available observation history for Search-QA. Observation similarity is computed with Python's `difflib.SequenceMatcher` after lexicographically ordering the two normalized observations in each pair, yielding a sequence-similarity ratio in $[0, 1]$. We set $\epsilon_H = 0.7$ and $\alpha_H = 0.95$. For the diagnostic analyses in Figures 1(b) and 4(a), similarity is computed as the `SequenceMatcher` ratio between full decision-time contexts. These contexts contain the current observation and bounded recent interaction history available to the policy, with the task instruction shared within each same-task comparison. For self-teacher rescoring, the teacher-only P_G^+ and P_G^- are separately appended to each turn $u \in G$ after its causal history and before the assistant generation marker, and are not propagated to turns outside G. The PI budget is 1024 tokens, the agreement threshold is $\epsilon_C = 0$, and the auxiliary-loss coefficient is $\lambda_{\text{HASD}} = 0.01$.

C Comparative PI Prompt Template and Case Study

C.1 Privileged-Information Prompt Template

Table 2 gives the exact field assignment. The two rows preserve the same action evidence and the same ordering and content within each action set.

Table 2: Instantiation of the paired comparative PI prompts.

Prompt	`HIGHER-OUTCOME` field	`LOWER-OUTCOME` field
P_G^+	$\mathcal{A}_G^{\text{hi}}$	$\mathcal{A}_G^{\text{lo}}$
P_G^-	$\mathcal{A}_G^{\text{lo}}$	$\mathcal{A}_G^{\text{hi}}$

The following is the shared prompt template for both PI views used by HASD:

Shared Privileged-Information Prompt Template

You are given privileged hindsight from sampled rollout turns under the same or highly similar decision contexts.

The following action alternative(s) have the highest empirical outcome evidence:

[HIGHER-OUTCOME ACTION ALTERNATIVES]

{higher_actions}

[/HIGHER-OUTCOME ACTION ALTERNATIVES]

The following action alternative(s) have the lowest empirical outcome evidence:

[LOWER-OUTCOME ACTION ALTERNATIVES]

{lower_actions}

[/LOWER-OUTCOME ACTION ALTERNATIVES]

When multiple actions appear in the same set, treat them as tied alternatives; the available evidence does not establish a preference among them.

Use the contrast between the two sets as soft guidance for the current decision. Favor applicable patterns from the higher-outcome alternatives and avoid those associated with the lower-outcome alternatives. Ground the response in the current context and do not copy unsupported task-specific details. Treat these labels as relative empirical evidence, not guarantees of local optimality or unique correctness. Do not explain the hindsight itself.

For P_G^+, {higher_actions} is instantiated with $\mathcal{A}_G^{\text{hi}}$ and {lower_actions} with $\mathcal{A}_G^{\text{lo}}$. For P_G^-, {higher_actions} is instantiated with $\mathcal{A}_G^{\text{lo}}$ and {lower_actions} with $\mathcal{A}_G^{\text{hi}}$. Thus, the prompts share identical instruction text and action evidence, differing only in the assignment of the high- and low-outcome action sets to the two fields.

C.2 Case Study: Decision Contrasts and Comparative PI

Figure 6 illustrates the construction of comparative privileged information with an ALFWorld example using $w = 2$, i.e., the current observation and two preceding observations. The history-aligned decision group contains four normalized actions with different empirical outcomes. Action 1 succeeds in both sampled turns and therefore has $q_G(a) = 1.0$, whereas Action 2 succeeds once and fails once, giving $q_G(a) = 0.5$. Actions 3 and 4 are both associated with failed episodes, so both have $q_G(a) = 0$. Accordingly, Action 1 forms $\mathcal{A}_G^{\text{hi}}$, Actions 3 and 4 form $\mathcal{A}_G^{\text{lo}}$, and the intermediate Action 2 is excluded. The resulting prompts contain the same action evidence: P_G^+ assigns the high- and low-outcome sets to their corresponding labels, while P_G^- reverses these assignments.

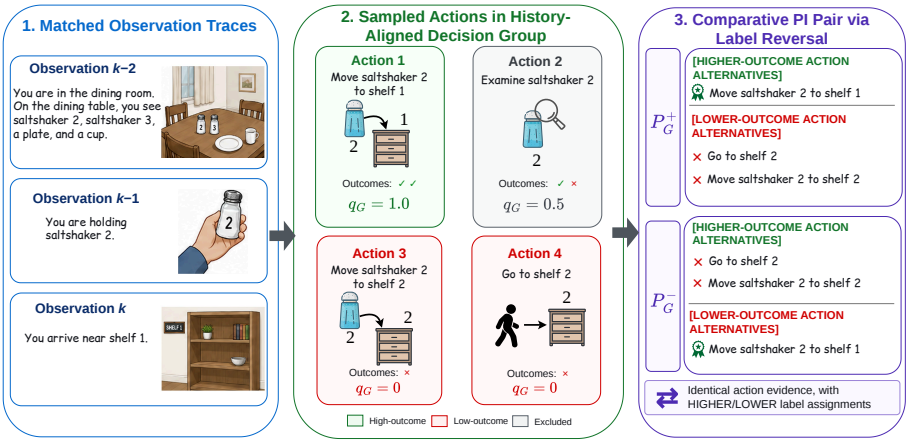


Figure 6: Case study of comparative PI construction in HASD on ALFWorld. History-aligned turns within the same decision group provide empirical high- and low-outcome action sets; the two PI prompts contain the same evidence but reverse the assignments of these sets to the HIGHER−OUTCOME and LOWER−OUTCOME fields. Images and icons are illustrative only. HASD operates on textual observations and actions.

D Theoretical Analyses

D.1 One-Sided and Comparative Signal Analysis

For a fixed token context (u, t), let $p_{u,t}^0(\cdot) = \pi_{\theta_{\text{old}}}(\cdot \mid H_u, y_{u,<t})$ denote the PI-free next-token distribution, and let $p_{u,t}^+(\cdot)$ and $p_{u,t}^-(\cdot)$ denote the corresponding distributions conditioned additionally on P_G^+ and P_G^-, respectively. As shown in Eq. (8), the shared PI-free reference term cancels exactly, leaving $\Delta_{u,t}(y) = \log p_{u,t}^+(y) - \log p_{u,t}^-(y)$. For either fixed PI view $\sigma \in \{+, -\}$, the one-sided sampled-token self-distillation signal satisfies

$$\mathbb{E}_{y \sim p_{u,t}^0}\left[\log p_{u,t}^\sigma(y) - \log p_{u,t}^0(y)\right] = -D_{\text{KL}}\left(p_{u,t}^0 \| p_{u,t}^\sigma\right) \le 0. \tag{16}$$

Thus, Eq. (16) formalizes the negative tendency introduced in Section 4.2: the one-sided OPSD signal has a non-positive conditional mean for every fixed token context, although individual sampled tokens can receive either positive or negative scores. In finite samples, this tendency can manifest as a negative empirical mean over sampled training tokens. Although turns within the same decision group share the same P_G^σ, their decision contexts H_u and response prefixes $y_{u,<t}$ can differ, so these conditional means need not be identical.

In contrast, with both PI views fixed, the comparative signal satisfies

$$\mathbb{E}_{y\sim p^0_{u,t}}[\Delta_{u,t}(y)] = D_{\mathrm{KL}}\left(p^0_{u,t}\|p^-_{u,t}\right) - D_{\mathrm{KL}}\left(p^0_{u,t}\|p^+_{u,t}\right), \tag{17}$$

which has no fixed sign and compares the two PI-conditioned views under the same token context. Importantly, comparative rescoring is not equivalent to subtracting a constant mean: $g^-_{u,t}$ is itself token-dependent, so $\Delta_{u,t} = g^+_{u,t} - g^-_{u,t}$ performs a token-wise contrast between the two PI views.

We next examine how this context-conditioned negative tendency enters the sampled-token optimization. Consider any fixed set of turns used for an auxiliary update, and let $Z = \sum_u L_u$ denote its number of sampled tokens. For view σ, define the empirical mean

$$\bar{g}^\sigma = \frac{1}{Z}\sum_u \sum_{t=1}^{L_u} g^\sigma_{u,t}. \tag{18}$$

Here, $\bar{g}^\sigma$ is computed from the same sampled tokens in the update set and is therefore sample-dependent. It is a shared scalar within the realized set, whereas the conditional means in Eq. (16) generally vary across token contexts. With the self-teacher scores detached during optimization, the corresponding finite-sample one-sided objective decomposes exactly as

$$\begin{aligned}\mathcal{L}^\sigma_{\mathrm{one}} &= -\frac{1}{Z}\sum_u \sum_{t=1}^{L_u} \mathrm{sg}(g^\sigma_{u,t}) \log \pi_\theta(y_{u,t} \mid H_u, y_{u,<t}) \\ &= \mathcal{L}^\sigma_{\mathrm{ctr}} + \bar{g}^\sigma \mathcal{L}_{\mathrm{NLL}},\end{aligned} \tag{19}$$

where

$$\mathcal{L}^\sigma_{\mathrm{ctr}} = -\frac{1}{Z}\sum_u \sum_{t=1}^{L_u} \mathrm{sg}(g^\sigma_{u,t} - \bar{g}^\sigma) \log \pi_\theta(y_{u,t} \mid H_u, y_{u,<t}), \tag{20}$$

$$\mathcal{L}_{\mathrm{NLL}} = -\frac{1}{Z}\sum_u \sum_{t=1}^{L_u} \log \pi_\theta(y_{u,t} \mid H_u, y_{u,<t}). \tag{21}$$

Let

$$c^\sigma_{u,t} = g^\sigma_{u,t} - \bar{g}^\sigma$$

denote the centered token signal, so that $g^\sigma_{u,t} = c^\sigma_{u,t} + \bar{g}^\sigma$. When $\bar{g}^\sigma < 0$,

$$c^\sigma_{u,t} > 0 \implies g^\sigma_{u,t} < c^\sigma_{u,t}, \qquad c^\sigma_{u,t} < 0 \implies g^\sigma_{u,t} < c^\sigma_{u,t} < 0,$$

and, in particular,

$$0 < c^\sigma_{u,t} < |\bar{g}^\sigma| \implies g^\sigma_{u,t} < 0.$$

Thus, a negative empirical mean lowers the one-sided signals relative to their centered values, regardless of whether the centered signal is positive or negative. It attenuates positive relative signals, strengthens negative ones, and can even reverse sufficiently small positive signals. Eq. (19) isolates the contribution of the empirical mean as $\bar{g}^\sigma \mathcal{L}_{\mathrm{NLL}}$ in the finite-sample auxiliary objective.

For the comparative signal, using the same sampled tokens gives

$$\bar{\Delta} = \bar{g}^+ - \bar{g}^-. \tag{22}$$

Hence, when the two one-sided views contain similar negative mean components, their subtraction substantially reduces the negative tendency. Unlike constant centering, however, comparative rescoring subtracts the token-dependent $g^-_{u,t}$, so the resulting signal can encode token-dependent relative information between the two PI views rather than merely subtracting a constant mean term. Section 5.4 empirically examines these one-sided and comparative signal statistics during training.

D.2 Agreement-Score Characterization and Gating Robustness

This section proves Lemma 1 and Proposition 1. The first result gives both a weighted average of pairwise margins and a scaled covariance. The second uses this pairwise representation to prove invariance to group-shared additive shifts and show that token-level perturbations induce bounded changes in the agreement score after being aggregated into turn-level scores.

Proof of Lemma 1. Let $G \in \mathcal{G}'$ contain n_G turns, and define

$$\bar{A}_G = \frac{1}{n_G}\sum_{u\in G} A_u, \qquad \bar{d}_G = \frac{1}{n_G}\sum_{u\in G} d_u,$$

and

$$\mathrm{Cov}(A,d) = \frac{1}{n_G}\sum_{u\in G}(A_u - \bar{A}_G)(d_u - \bar{d}_G).$$

For $W_G > 0$, using $x = |x|\,\mathrm{sgn}(x)$, the agreement score in Eq. (9) becomes

$$\begin{aligned} C_G &= \frac{1}{W_G}\sum_{u<v}(A_u - A_v)(d_u - d_v) \\ &= \sum_{u<v}\frac{|A_u - A_v|}{W_G}\,\mathrm{sgn}(A_u - A_v)(d_u - d_v). \end{aligned} \tag{23}$$

Because

$$\frac{|A_u - A_v|}{W_G} \geq 0, \qquad \sum_{u<v}\frac{|A_u - A_v|}{W_G} = 1,$$

Eq. (23) is a weighted average of signed pairwise comparative margins. For a pair with $A_u > A_v$, the margin is $d_u - d_v$ and is positive when the turn with the higher inherited GRPO advantage also receives the larger comparative score.

To derive the covariance form, expand the unordered pair sum:

$$\begin{aligned} \sum_{u<v}(A_u - A_v)(d_u - d_v) &= \frac{1}{2}\sum_{u,v}(A_u - A_v)(d_u - d_v) \\ &= n_G\sum_{u\in G} A_u d_u - \left(\sum_{u\in G} A_u\right)\left(\sum_{u\in G} d_u\right) \\ &= n_G\sum_{u\in G}(A_u - \bar{A}_G)(d_u - \bar{d}_G) \\ &= n_G^2\,\mathrm{Cov}(A,d). \end{aligned} \tag{24}$$

Dividing Eq. (24) by W_G yields

$$C_G = \frac{n_G^2}{W_G}\,\mathrm{Cov}(A,d),$$

which completes the two representations in Eq. (12). Since $n_G^2/W_G > 0$, C_G and $\mathrm{Cov}(A,d)$ have the same sign. This is an aggregate property and does not require every turn pair to have matching advantage and comparative-score orderings. □

Proof of Proposition 1. We first consider an additive shift b shared across the decision group G. Since the agreement score uses d only through pairwise differences,

$$\begin{aligned} C_G(A, d + b\mathbf{1}) &= \frac{1}{W_G}\sum_{u<v}(A_u - A_v)\big[(d_u + b) - (d_v + b)\big] \\ &= \frac{1}{W_G}\sum_{u<v}(A_u - A_v)(d_u - d_v) \\ &= C_G(A,d). \end{aligned} \tag{25}$$

Thus, a component shared by all turns cancels separately within every pairwise difference, proving Eq. (13).

We next connect token-level variation to the turn-level perturbation used in Eq. (15). Let $\varepsilon_{u,t} \in \mathbb{R}$ be an arbitrary additive perturbation and set $\widetilde{\Delta}_{u,t} = \Delta_{u,t} + \varepsilon_{u,t}$. Averaging within turn u gives the compact relation

$$\widetilde{d}_u = \frac{1}{L_u}\sum_{t=1}^{L_u}(\Delta_{u,t} + \varepsilon_{u,t}) = d_u + \underbrace{\frac{1}{L_u}\sum_{t=1}^{L_u}\varepsilon_{u,t}}_{\xi_u}. \tag{26}$$

Hence,

$$\xi_u = \frac{1}{L_u} \sum_{t=1}^{L_u} \varepsilon_{u,t}$$

is precisely the turn-level perturbation induced by token averaging. No probabilistic assumption is required: the perturbations may be arbitrary deterministic changes. Moreover, if $|\varepsilon_{u,t}| \le \delta$ for every token, then

$$|\xi_u| \le \frac{1}{L_u} \sum_{t=1}^{L_u} |\varepsilon_{u,t}| \le \delta, \qquad \|\xi\|_\infty \le \delta.$$

Thus, turn averaging does not enlarge a uniform token-level perturbation bound.

For $\xi = (\xi_u)_{u \in G}$, subtracting the original agreement score yields

$$C_G(A, d + \xi) - C_G(A, d) = \frac{1}{W_G} \sum_{u<v} (A_u - A_v)(\xi_u - \xi_v). \tag{27}$$

Using the normalized pairwise weights from Eq. (23),

$$\begin{aligned} |C_G(A, d + \xi) - C_G(A, d)| &\le \sum_{u<v} \frac{|A_u - A_v|}{W_G} |\xi_u - \xi_v| \\ &\le \max_{u,v \in G} |\xi_u - \xi_v| \\ &\le 2\|\xi\|_\infty. \end{aligned} \tag{28}$$

The intermediate bound shows that the sensitivity of C_G is controlled by variation in the perturbations across turns, rather than by their absolute magnitude. The final inequality follows from

$$|\xi_u - \xi_v| \le |\xi_u| + |\xi_v| \le 2\|\xi\|_\infty.$$

This proves Eq. (15) without distributional or independence assumptions.

Combining the token-level bound with Eq. (28) gives

$$|\varepsilon_{u,t}| \le \delta \quad \forall u, t \qquad \Longrightarrow \qquad |C_G(A, d + \xi) - C_G(A, d)| \le 2\delta.$$

Thus, a uniform token-level perturbation bound translates directly into a deterministic bound on the agreement score.

Finally, let $\widetilde{C}_G = C_G(A, d + \xi)$. If

$$|C_G - \epsilon_C| > 2\|\xi\|_\infty, \tag{29}$$

then C_G and $\widetilde{C}_G$ remain on the same side of the agreement threshold. Consequently,

$$\mathbb{I}[\widetilde{C}_G > \epsilon_C] = \mathbb{I}[C_G > \epsilon_C]. \tag{30}$$

Therefore, decision groups farther from the threshold retain the same accept/reject decision under all perturbations satisfying the bound. For the primary setting $\epsilon_C = 0$, the condition reduces to $|C_G| > 2\|\xi\|_\infty$.

The factor 2 is tight in general: in a two-turn decision group, $\xi_u = \delta$ and $\xi_v = -\delta$ change the pairwise comparative margin by exactly 2δ. □

Degenerate case. If $W_G = 0$, then

$$\sum_{u<v} |A_u - A_v| = 0,$$

so all A_u are identical within decision group G. No GRPO-advantage ordering is then available for evaluating the comparative scores, and the gate is set to $m_G = 0$ as specified in Section 4.3.

The derivations above concern the agreement gate itself: they show that the group-shared components cancel in pairwise differences and that the effect of token-level variation remains bounded after aggregation to the turn and decision-group levels. No stochastic model of the perturbations is required.

E Extended Analyses

E.1 Empirical Analysis of Decision-Group Structure and Representative Matching

To characterize the empirical structure of decision groups in $\mathcal{G}'$, Table 3 summarizes their action–outcome evidence from Eqs. (5)–(6) across environments and backbones. The results show that 83.36–95.11% of groups contain turns from more than two unique trajectories, with an average of 4.89–7.00 unique trajectories per group. Given the rollout group size of 8, these values indicate that a large fraction of the sampled trajectories typically contribute to each decision group in $\mathcal{G}'$. Meanwhile, 46.15–77.70% of groups contain more than two retained high/low actions, with 2.96–4.77 retained per group on average. The number of unique trajectories consistently exceeds the number of retained actions, indicating that normalized actions recur across trajectories, which allows $q_G(a)$ to aggregate empirical outcomes from multiple trajectories. These results show that the PI is commonly constructed from multiple action alternatives rather than a single high/low action pair.

Beyond group composition, Table 3 further characterizes the retained high/low action sets and their supporting evidence in terms of coverage, outcome separation, and other-turn contribution. Across all settings, 87.42–97.42% of distinct normalized actions are retained in $\mathcal{A}_G^{hi} \cup \mathcal{A}_G^{lo}$, leaving only a small fraction as intermediate actions. The mean high–low outcome gap $\max_{a \in \mathcal{A}_G} q_G(a) - \min_{a \in \mathcal{A}_G} q_G(a)$ ranges from 0.697 to 0.917, indicating clear empirical outcome separation between the retained high- and low-outcome actions. Moreover, 84.34–91.40% of the raw high/low action occurrences used as PI evidence come from turns other than the target turn, showing that most PI evidence comes from other history-aligned decisions on average. Together, these statistics show that eligible decision groups provide the empirical diversity needed for decision contrasts, with multiple action alternatives across trajectories, clear outcome separation, and substantial evidence from other history-aligned turns, rather than relying on a single high/low action pair.

Table 3: Empirical structure of action–outcome evidence in $\mathcal{G}'$ across environments and backbones. Here, high/low action coverage denotes the fraction of distinct normalized actions retained in $\mathcal{A}_G^{hi} \cup \mathcal{A}_G^{lo}$; other-turn PI evidence denotes the average fraction of raw high/low action occurrences originating from turns other than the target turn.

	Search-QA			**WebShop**			**ALFWorld**		
Property	1.7B	3B	7B	1.7B	3B	7B	1.7B	3B	7B
>2 unique traj. (%)	95.11	91.17	90.52	92.54	92.92	90.72	87.25	85.86	83.36
Mean unique traj.	7.00	6.51	6.39	6.36	6.39	6.31	5.21	5.06	4.89
>2 high/low actions (%)	53.58	69.31	65.31	46.15	56.40	58.03	75.11	77.70	75.51
Mean high/low actions	3.09	4.00	3.72	2.96	3.82	3.59	4.60	4.77	4.49
High/low action coverage (%)	97.30	97.42	96.60	89.33	90.53	91.15	87.42	87.81	88.83
Mean high–low outcome gap	0.863	0.917	0.892	0.697	0.761	0.762	0.817	0.857	0.852
Other-turn PI evidence (%)	85.71	84.64	84.34	90.00	91.09	89.22	91.40	89.86	89.28

Table 4 further evaluates the consistency and efficiency of representative matching based on Eq. (4) for decision groups in $\mathcal{G}'$. The pairwise match rate denotes the fraction of within-group turn pairs satisfying $\mathrm{Sim}(T_u, T_v) \geq \epsilon_H$, while mean pairwise similarity is the average Observation-Trace similarity over all within-group pairs. For comparison, strict pairwise matching requires a candidate turn to satisfy the history-matching threshold with every existing turn in the group. The final row reports the ratio of similarity-comparison counts under strict pairwise and representative matching. Although each turn is matched only to the group representative, 74.34–92.28% of within-group pairs also satisfy the same history-matching threshold. Their mean pairwise similarity ranges from 0.739 to 0.881, remaining above the default $\epsilon_H = 0.7$ across all environments and backbones. Enforcing strict pairwise matching instead requires 2.25–$7.01\times$ as many similarity comparisons. This difference is particularly pronounced on ALFWorld, reaching 6.45–$7.01\times$, where larger within-group turn counts make exhaustive pairwise comparisons more costly. These results show that representative matching preserves substantial within-group consistency with far fewer comparisons, aligning with recent text-clustering work that uses representative items for cluster representation or efficient incremental maintenance (Wei et al., 2025b; Luo et al., 2026).

Table 4: Consistency and efficiency of representative matching in $\mathcal{G}'$ across environments and backbones.

	Search-QA			WebShop			ALFWorld		
Property	1.7B	3B	7B	1.7B	3B	7B	1.7B	3B	7B
Pairwise match rate (%)	89.78	91.29	92.28	82.49	88.94	83.96	79.15	74.34	77.73
Mean pairwise similarity	0.842	0.872	0.881	0.838	0.815	0.814	0.766	0.739	0.751
Strict/Rep. comparison ratio	2.25×	2.28×	2.35×	3.06×	2.69×	3.25×	6.93×	7.01×	6.45×

E.2 Leave-One-Trajectory-Out Analysis

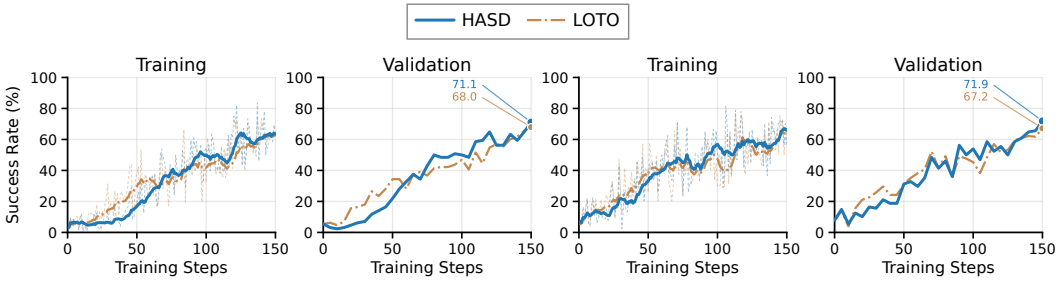
Although Table 3 shows that most raw PI evidence comes from turns other than the target turn, this does not determine the respective contributions of cross-trajectory and target-trajectory self-derived evidence to training performance. We therefore evaluate a leave-one-trajectory-out (LOTO) training variant. For each target turn, LOTO first removes all turns from the same trajectory and then recomputes the empirical action outcomes and high/low action sets used for PI construction. The resulting LOTO PI is then used throughout training in place of the original HASD PI, while all other components and hyperparameters are kept unchanged.

Table 5: Structural stability under LOTO with Qwen3-1.7B on WebShop and ALFWorld. All entries are percentages.

Env.	Group Retained	Turn Retained	Hi/Lo Unchanged	Target Supported
WebShop	94.57	89.03	71.76	80.28
ALFWorld	88.80	88.49	70.32	72.34

Figure 7 shows that LOTO retains strong performance despite this intervention. It reaches final validation success rates of 68.0 on WebShop versus 71.1 for HASD, and 67.2 on ALFWorld versus 71.9, with broadly similar training trends. Table 5 reports four structural stability metrics under LOTO. Group Retained and Turn Retained measure the ratios of the numbers of PI-eligible groups and target turns under LOTO to those under the original construction, reaching 88.80–94.57% and 88.49–89.03%, respectively. Hi/Lo Unchanged denotes the average of the exact-set unchanged rates for the high- and low-outcome action sets, reaching 70.32–71.76%. Target Supported denotes the fraction of original PI-eligible turns whose canonical target action remains supported by at least one other trajectory in the same decision group, reaching 72.34–80.28%. All four statistics are averaged over 150 training steps.

Thus, substantial cross-trajectory evidence remains available even after removing the target trajectory, helping explain why LOTO preserves much of the performance and showing that HASD's gains are not solely driven by target-trajectory evidence. At the same time, the remaining gap between LOTO and full HASD indicates that target-trajectory evidence provides complementary supervision, suggesting that combining self-derived evidence with evidence from other history-aligned trajectories forms a richer basis for PI construction. This is consistent with prior self-distillation results showing that retaining self-derived evidence together with other-decision evidence performs better than relying on other-decision evidence alone (Ma et al., 2026).



(a) WebShop (b) ALFWorld

Figure 7: Leave-one-trajectory-out analysis with Qwen3-1.7B on WebShop and ALFWorld. In the training panels, the thin curves show raw per-step measurements, while the thick curves show MA10. The validation panels show raw checkpoint measurements.

E.3 Parameter Sensitivity

Table 6 studies the effects of the history-matching threshold ϵ_H in Eq. (4) and the agreement threshold ϵ_C in Eq. (9) on Search-QA with Qwen3-1.7B. A lower ϵ_H admits looser history matches, producing larger decision groups and exposing more action–outcome contrasts. Increasing ϵ_H improves matching specificity, but also reduces usable support by decreasing the decision-group size. The aggregate performance is highest at $\epsilon_H = 0.7$, reaching 43.6, compared with 41.2 at $\epsilon_H = 0.5$ and 42.1 at $\epsilon_H = 0.9$. This pattern suggests that overly permissive matching can admit less comparable decisions, whereas overly strict matching limits the comparative evidence available for PI construction. We therefore use $\epsilon_H = 0.7$ as the default, which provides the best balance among the tested values.

Table 6: Parameter sensitivity to ϵ_H and ϵ_C on Search-QA using Qwen3-1.7B.

Setting	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg
Primary	41.4	59.5	45.9	38.5	38.8	14.1	67.0	**43.6**
$\epsilon_H = 0.5$	42.9	58.1	43.8	36.7	29.9	12.3	64.5	41.2
$\epsilon_H = 0.9$	42.3	59.3	45.9	37.6	32.0	12.3	65.6	42.1
$\epsilon_C = -0.1$	42.7	59.0	47.1	36.9	32.9	13.4	67.3	42.8
$\epsilon_C = 0.1$	40.8	57.9	43.3	35.3	28.9	12.4	68.2	41.0

The agreement threshold ϵ_C sets the required ordering agreement with GRPO advantages for decision-group acceptance. A negative threshold permits weakly conflicting directions, while a positive threshold requires an additional positive agreement margin. In the Search-QA sweep, $\epsilon_C = -0.1$ obtains an average of 42.8, the default $\epsilon_C = 0$ reaches 43.6, and $\epsilon_C = 0.1$ decreases the result to 41.0. These results indicate that the neutral boundary at $\epsilon_C = 0$ provides a better balance than either admitting weak conflict or imposing a stricter positive margin. Accordingly, we use $\epsilon_C = 0$ as the default. Across all PI-eligible groups before gating, the group-level C_G scores have a mean of 0.734 and a standard deviation of 1.043. Applying the default threshold $\epsilon_C = 0$ retains 91.81% of these groups and 92.55% of PI-eligible turns.

E.4 Multi-Seed Evaluation

To assess whether the main performance gains are stable across independent training runs, we repeat GRPO, SERL, and HASD with three training seeds on ALFWorld and WebShop across all three backbones. We focus on these two environments because they use smaller evaluation sample sizes than Search-QA, making it particularly important to verify that the observed performance differences persist across independent training runs. ALFWorld additionally shows the largest HASD gains in Table 1 and serves as a primary setting for our detailed analysis, while WebShop provides a complementary setting. We include GRPO as the direct outcome-based RL baseline, allowing us to measure the additional contribution of HASD beyond its base objective. We also include SERL because it is a methodologically relevant step-level OPSD baseline and the strongest baseline in both environments at 1.7B, the model scale where HASD shows its largest gains on average. All runs use the same hyperparameters and 150-step training budget as the primary experiments. We report final-checkpoint mean and standard deviation over three training seeds.

Table 7 shows that the performance gains over GRPO and SERL remain stable across independent training seeds. HASD achieves the highest mean success rate in all six backbone–environment settings, outperforming the stronger of GRPO and SERL by 10.20, 9.07, and 4.13 points on WebShop and by 12.54, 6.00, and 6.53 points on ALFWorld at 1.7B, 3B, and 7B, respectively. The largest margins occur at 1.7B in both environments, consistent with the main-results trend that smaller backbones benefit more from HASD. Overall, the improvements persist after averaging over three training seeds, supporting the stability of the main performance gains.

Table 7: Multi-seed evaluation on WebShop and ALFWorld across three backbones. Final-checkpoint success rates (%) are reported as mean ± standard deviation over three training seeds. Bold values indicate the best mean performance in each setting.

	WebShop			**ALFWorld**		
Method	1.7B	3B	7B	1.7B	3B	7B
GRPO	58.33 ± 4.45	60.70 ± 2.39	74.73 ± 0.46	55.70 ± 2.97	81.77 ± 0.40	83.87 ± 2.01
SERL	60.87 ± 7.16	64.60 ± 1.61	75.27 ± 1.67	56.23 ± 2.71	76.03 ± 1.96	85.67 ± 2.51
HASD	$\mathbf{71.07 \pm 1.55}$	$\mathbf{73.67 \pm 3.93}$	$\mathbf{79.40 \pm 2.39}$	$\mathbf{68.77 \pm 3.15}$	$\mathbf{87.77 \pm 1.22}$	$\mathbf{92.20 \pm 1.39}$

F Computational Efficiency

On-policy self-distillation enriches sparse outcome supervision but introduces additional training-time computation relative to GRPO. The key efficiency question is how often teacher rescoring is required. SDAR applies a single PI-conditioned rescoring pass to each sampled turn, whereas HASD performs paired $P_G^{\pm}$ rescoring only for turns in eligible history-aligned decision groups.

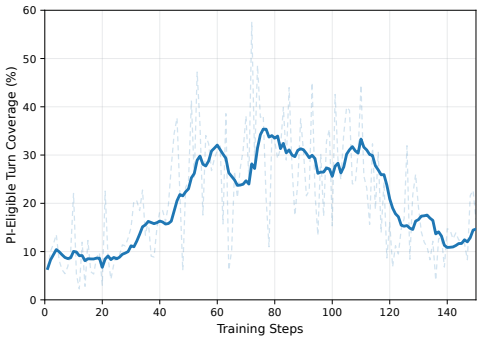


Figure 8: PI-eligible turn coverage for HASD on WebShop with Qwen3-1.7B.

Figure 8 reports PI-eligible turn coverage, i.e., the fraction of all rollout turns that belong to decision groups $G \in \mathcal{G}'$ and enter paired rescoring at each training step. The dashed curve shows the per-step coverage, while the solid curve shows its 10-step moving average. The coverage remains well below 100%, showing that HASD activates OPSD selectively within eligible history-aligned decision groups that provide comparative action–outcome evidence. This selective activation further limits the additional PI preparation and teacher computation.

Table 8 further quantifies the resulting training operations and wall-clock cost per training step on WebShop with Qwen3-1.7B. For this efficiency comparison, we run SDAR under the same hardware and runtime configuration as GRPO and HASD. Rollout generation and policy-update costs remain comparable across the three methods, so the additional overhead is concentrated in PI preparation and teacher rescoring. Together, these two stages require 21.74 seconds per step for SDAR but only 9.01 seconds for HASD, corresponding to a 58.6% reduction in auxiliary overhead. Consequently, the total training cost is $1.12\times$ GRPO for SDAR and only $1.06\times$ for HASD. Although HASD performs two teacher evaluations for each selected turn, selective activation limits the number of turns requiring rescoring and keeps the overall auxiliary cost low. All additional computation is training-only and incurs no inference-time overhead.

Table 8: Training operations and wall-clock cost per training step on WebShop with Qwen3-1.7B. Times are mean ± standard deviation across training steps 1–150 on eight GPUs; validation and checkpoint I/O are excluded.

Stage	**GRPO**	**SDAR**	**HASD**
Operations			
Rollout generation	Student rollouts	Student rollouts	Student rollouts
PI preparation	—	Skill retrieval	History grouping + $P_G^{\pm}$
Teacher rescoring	—	$1\times$, all turns	$2\times$, eligible turns
Policy update	GRPO	GRPO + SDAR	GRPO + HASD
Wall-clock time (s/step)			
Rollout generation	125.17 ± 6.93	123.54 ± 6.09	125.33 ± 6.93
PI preparation	0.00	9.54 ± 1.31	4.80 ± 2.08
Teacher rescoring	0.00	12.20 ± 1.79	4.21 ± 2.22
Policy update	55.15 ± 9.44	55.81 ± 7.10	56.16 ± 8.51
Total	180.32 ± 11.35	201.09 ± 12.32	**190.50 ± 11.35**
Relative cost	1.00×	1.12×	**1.06×**