

# OPID: ON-POLICY SKILL DISTILLATION FOR AGENTIC REINFORCEMENT LEARNING

Shuo Yang<sup>1\*</sup>, Jinyang Wu<sup>1\*†</sup>, Zhengxi Lu<sup>2</sup>, Yuhao Shen<sup>2</sup>, Fan Zhang<sup>3</sup>, Lang Feng<sup>4</sup>,  
Shuai Zhang<sup>1</sup>, Haoran Luo<sup>4</sup>, Zheng Lian<sup>5</sup>, Zhengqi Wen<sup>1</sup>, Jianhua Tao<sup>1</sup>

<sup>1</sup>Tsinghua University    <sup>2</sup>Zhejiang University    <sup>3</sup>The Chinese University of Hong Kong  
<sup>4</sup>Nanyang Technological University    <sup>5</sup>Tongji University  
Corresponding to: wu-jy23@mails.tsinghua.edu.cn

## ABSTRACT

Outcome-based reinforcement learning provides a stable optimization backbone for language agents, but its sparse trajectory-level rewards provide little guidance on which intermediate decisions should be reinforced or suppressed. On-policy self-distillation offers dense token-level supervision, yet existing skill-conditioned variants often rely on external skill memories or retrieved privileged context, which are costly to maintain and can be mismatched with the state distribution induced by the current policy in multi-turn interaction. We propose **OPID (On-Policy Skill Distillation)**, a framework that extracts skill supervision directly from completed on-policy trajectories. OPID represents trajectory hindsight as hierarchical skills: episode-level skills capture global workflows or failure-avoidance rules, while step-level skills capture local decision knowledge at critical timesteps. A critical-first routing mechanism uses step-level skills when critical decisions are identified and falls back to episode-level skills as default guidance otherwise. The selected skill is injected into the interaction history, allowing the old policy to re-score the same sampled response under both original and skill-augmented contexts. The resulting log-probability shift yields a token-level self-distillation advantage, which is combined with the outcome advantage for policy optimization. OPID thus preserves RL as the primary training objective while introducing dense, distribution-matched hindsight supervision. Experiments on ALFWorld, WebShop and Search-based QA demonstrate that OPID generally improves agent performance, sample efficiency, and robustness over outcome-only RL and existing skill-distillation baselines. Our code is available at <https://github.com/jinyangwu/OPID/tree/main>.

## 1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as interactive agents that operate over long horizons, invoke tools, navigate environments, and adapt their behavior through multi-turn observations (Jimenez et al., 2024; Luo et al., 2025; Wu et al., 2026a; Lu et al., 2026a). Unlike single-turn reasoning, agentic tasks require sequential decisions whose consequences may only become visible after many interaction steps. This setting spans embodied household environments, web navigation, search-augmented reasoning, and software engineering agents (Shridhar et al., 2020; Yao et al., 2022; Jin et al., 2025; Jimenez et al., 2023). Reinforcement learning (RL) has become a natural post-training paradigm for such agents, since it directly optimizes policies using task-level feedback from environments or verifiers. In particular, outcome-based methods such as GRPO (Shao et al., 2024) provide a stable critic-free optimization backbone for on-policy rollouts.

Despite its effectiveness, outcome-based agentic RL offers only coarse supervision (Zhang et al., 2025). Environment rewards are typically sparse, delayed, and high-variance: a terminal reward can

\*Equal Contribution

†Project Leader



Figure 1: **Overall performance comparison.** We compare OPID with training-free prompting methods, outcome-only RL, and skill-distillation baselines on ALFWorld, Search-based QA, and WebShop. OPID achieves the strongest average performance on ALFWorld and WebShop while remaining competitive on Search-based QA.

indicate whether a trajectory succeeds, but not which intermediate decisions caused the outcome. This limitation is especially severe in long-horizon interaction (Chen et al., 2026; Xu et al., 2026), where a single early mistake may derail the episode, repeated invalid actions may accumulate over time, and the effect of a local decision may only be observed several turns later. As a result, purely outcome-driven optimization provides stable task-level pressure but lacks fine-grained decision-level credit assignment.

On-policy distillation and self-distillation provide complementary supervision. Rather than relying solely on trajectory-level rewards, on-policy distillation trains models on their own sampled outputs while using auxiliary teacher signals to induce token-level guidance (Gu et al., 2024a; Agarwal et al., 2024). Recent self-distillation methods remove the need for a separate teacher by comparing the same policy under different contexts, such as a standard student branch and a privileged teacher branch (Zhao et al., 2026; He et al., 2026). In agentic RL, this suggests a natural decomposition: RL remains the primary optimization backbone, while self-distillation supplies dense token-level shaping signals. Recent work such as SDAR follows this principle by treating self-distillation as a controlled auxiliary objective for multi-turn agents (Lu et al., 2026a).

A particularly promising form of privileged context is a natural-language skill. Skill-conditioned self-distillation augments the teacher branch with procedural knowledge, such as subgoal decompositions, action templates, or behavioral rules, and distills the resulting token-level preferences into the policy (Lu et al., 2026b; Wang et al., 2026; Lu et al., 2026a). However, existing skill-based methods typically rely on external skill libraries, retrieved skill files, or maintained skill memories. This design raises two challenges. First, skill memories require non-trivial maintenance, including skill insertion, refinement, deletion, and retrieval. Second, retrieved skills may be mismatched with the state distribution induced by the current policy. Such mismatch is particularly problematic for multi-turn agents, where small deviations from the assumed trajectory can lead to state drift and make an otherwise useful skill unreliable.

Based on this observation, we propose **OPID (On-Policy Skill Distillation)**, a framework that extracts hindsight skills from completed on-policy trajectories and distills their behavioral effects back into the policy. OPID abstracts each trajectory into two complementary levels of natural-language skills: *episode-level skills*, which summarize trajectory-wide workflows or failure-avoidance rules, and *step-level skills*, which capture state-conditioned guidance at critical timesteps. This hierarchy reflects a granularity trade-off in long-horizon decision making. Episode-level skills are broad and stable but may be too coarse for pivotal states, whereas step-level skills are precise but sparse and state-specific. OPID addresses this trade-off with *critical-first skill routing*: it uses step-level skills

---

at identified critical timesteps and falls back to episode-level skills otherwise. The routed skill is injected into the agent’s interaction history, allowing the old policy to re-score the same on-policy response under both original and skill-augmented contexts. The induced token-level log-probability shift forms a skill-based self-distillation advantage, which is combined with the episode advantage for policy optimization. OPID therefore preserves outcome-based RL as the primary objective while introducing dense, on-policy hindsight supervision. At inference time, OPID requires no analyzer, external skill retrieval, or privileged context.

We evaluate OPID on ALFWorld (Shridhar et al., 2020), WebShop (Yao et al., 2022), and Search-based QA (Jin et al., 2025) with models at different scales. Across these settings, OPID improves long-horizon agent performance over outcome-only RL and skill-distillation baselines. These results suggest that completed on-policy trajectories provide a useful source of distribution-matched hindsight supervision, enabling the policy to internalize trajectory-derived skills without relying on external skill libraries or retrieved privileged context at inference time.

Taken together, our work makes the following contributions:

- We propose **on-policy hindsight skill extraction**, which treats completed trajectories sampled by the current policy as a distribution-matched source of skill supervision, avoiding the need for external skill libraries or off-policy retrieval.
- We introduce **hierarchical hindsight skills with critical-first routing**, where episode-level skills capture global workflows or failure-avoidance rules, step-level skills capture critical local decisions, and routing selects the most specific available skill for each trajectory step.
- We integrate **skill-based self-distillation** into agentic RL, converting routed hindsight skills into dense token-level shaping signals while preserving outcome reward optimization as the primary training objective.
- We empirically validate OPID on long-horizon agentic benchmarks, showing consistent improvements over outcome-only RL and skill-distillation baselines, along with better sample efficiency and reduced repetitive or invalid behaviors.

## 2 RELATED WORK

**Reinforcement learning for agentic LLMs.** Large language models are increasingly trained as interactive agents that operate over long horizons, invoke tools, and receive feedback from environments or verifiers (Shridhar et al., 2020; Yao et al., 2022; Jin et al., 2025; Jimenez et al., 2023; Wu et al., 2026c). Reinforcement learning has therefore become a natural post-training paradigm, with outcome-based methods such as GRPO providing a stable critic-free objective for on-policy rollouts (Shao et al., 2024). However, agentic environments typically provide sparse and delayed rewards. A terminal outcome can indicate whether a trajectory succeeds, but it does not identify which intermediate decisions caused success or failure. OPID targets this missing credit-assignment signal: it keeps outcome-based RL as the optimization backbone, but augments it with dense decision-level supervision extracted from the policy’s own completed trajectories.

**On-policy self-distillation.** On-policy distillation trains a model from its own sampled outputs while using auxiliary teacher signals to provide token-level learning targets (Agarwal et al., 2024; Gu et al., 2024a). Recent self-distillation methods further remove the need for a separate teacher by comparing the same policy under different contexts or feedback conditions (Zhao et al., 2026; He et al., 2026). For multi-turn agents, this suggests a useful decomposition: RL supplies task-level optimization, while self-distillation supplies dense shaping signals (Lu et al., 2026a). The key question is where the privileged signal should come from. Existing methods often rely on generic revision contexts, external hints, or task-level feedback transformations. OPID instead constructs the privileged branch from hindsight skills extracted from on-policy trajectories, making the distillation signal directly tied to the states, actions, and failures encountered by the current policy.

**Skill-conditioned agent learning.** Natural-language skills provide compact procedural knowledge for agents, including subgoal decompositions, action templates, and failure-avoidance rules (Lu et al., 2026b; Wang et al., 2026; Lu et al., 2026a; Wu et al., 2026b). Existing skill-based methods

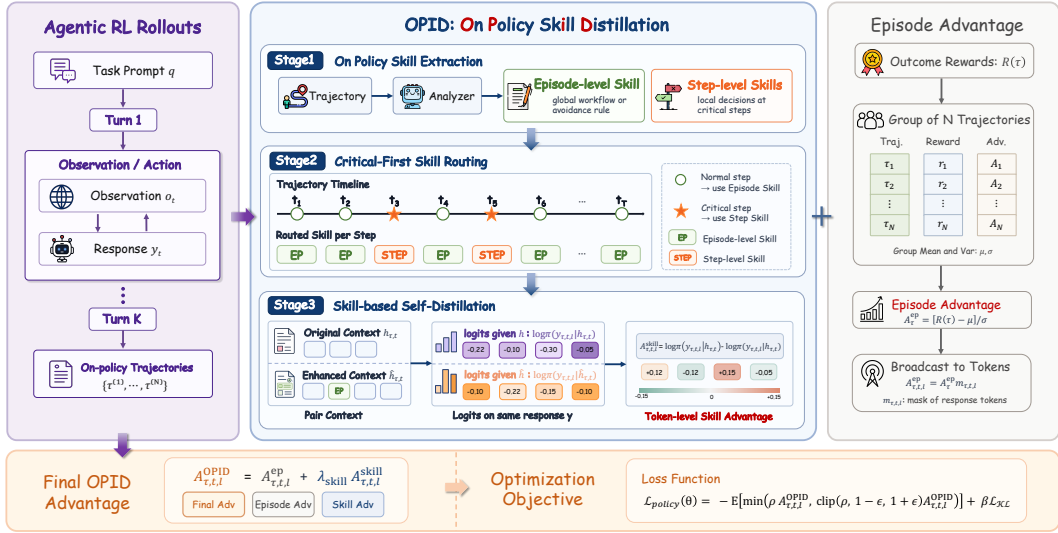


Figure 2: **Overview of OPID.** Starting from completed on-policy trajectories, OPID extracts hierarchical hindsight skills and routes the most relevant skill to each decision, prioritizing step-level skills at critical states. The policy then re-scores the same sampled response with and without the routed skill, turning the token-wise log-probability difference into a dense skill advantage that complements the episode-level RL signal.

commonly depend on external skill libraries, retrieved skill files, or persistent skill memories. These designs can improve agent behavior, but they introduce maintenance and retrieval costs, and retrieved skills may be mismatched with the state distribution induced by the current policy. This mismatch becomes more severe in long-horizon interaction, where small deviations can lead to substantial state drift. OPID makes a different design choice: it extracts hierarchical skills directly from completed on-policy trajectories, routes them according to decision criticality, and distills their behavioral effect into the policy during training. As a result, OPID provides distribution-matched hindsight supervision without requiring skill retrieval, analyzer calls, or privileged context at inference time.

### 3 METHODS

We formulate long-horizon agentic tasks as partially observable decision processes and present OPID, a framework that converts completed on-policy trajectories into hierarchical skills and distills their behavioral effect back into the policy. OPID performs on-policy skill distillation in three stages. First, it extracts hierarchical skills from completed on-policy trajectories. Second, it routes the appropriate skill to each decision step and converts the skill effect into token-level self-distillation signals. Third, it combines these token-level skill advantages with group-relative outcome advantages for policy optimization. Figure 2 illustrates the overall pipeline.

#### 3.1 PROBLEM FORMULATION

We model an agentic task as a partially observable Markov decision process defined by

$$(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \gamma),$$

where  $\mathcal{S}$  is the latent state space,  $\mathcal{A}$  is the action space,  $\mathcal{O}$  is the observation space,  $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  is the transition function,  $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is the reward function, and  $\gamma \in [0, 1)$  is the discount factor. At timestep  $t$ , the environment is in a hidden state  $s_t \in \mathcal{S}$  and emits an observation  $o_t \in \mathcal{O}$ . The agent maintains an interaction history

$$h_t = (o_0, y_0, o_1, y_1, \dots, o_t),$$

where  $y_i$  denotes the textual response or executable action generated at step  $i$ . The policy  $\pi_\theta$  generates the next response as

$$y_t \sim \pi_\theta(\cdot | h_t).$$

After executing  $y_t$ , the environment transitions and returns the next observation. A completed trajectory is represented as

$$\tau = \{(o_t, y_t, r_t)\}_{t=0}^{T-1},$$

where  $T$  is the episode length. In most agentic benchmarks, rewards are sparse and terminal, so we denote the outcome score by

$$R(\tau) \in \{0, 1\},$$

or more generally  $R(\tau) \in \mathbb{R}$  when the benchmark provides graded feedback. The learning objective is

$$J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)].$$

Following GRPO-style training, for each task prompt  $q$  we sample a group of  $N$  trajectories from the current policy:

$$\mathcal{G}_q = \{\tau^{(1)}, \tau^{(2)}, \dots, \tau^{(N)}\}.$$

### 3.2 ON-POLICY SKILL EXTRACTION

Outcome rewards reveal whether a trajectory succeeds, but not why it succeeds or fails. OPID therefore represents post-hoc trajectory knowledge as hierarchical skills extracted from completed on-policy rollouts. The hierarchy contains two complementary levels.

**Episode-level skills.** An episode-level skill  $s_\tau^{\text{ep}}$  summarizes the global behavioral pattern of a complete trajectory  $\tau$ . For a successful trajectory, it captures a reusable workflow that explains how the task was solved. For a failed trajectory, it captures a failure-avoidance rule that describes what should be avoided in similar future situations. Episode-level skills are broad and stable, making them suitable as default guidance for most states.

**Step-level skills.** A step-level skill  $s_{\tau,t}^{\text{step}}$  captures local decision knowledge at timestep  $t$ . It is intended for pivotal states where the final outcome depends strongly on a specific choice, such as avoiding a repeated invalid action, selecting the next object to inspect, correcting a mistaken subgoal, or deciding when to stop exploration. Step-level skills are more precise than episode-level skills, but they are also sparse and state-dependent.

Given a completed trajectory  $\tau$ , OPID reconstructs an ordered trajectory record containing the task prompt, observations, model responses, environment feedback, step indices, and terminal outcome. An LLM-based analyzer  $\mathcal{A}$  maps this record to structured natural-language skills:

$$\mathcal{A}(\tau) = (s_\tau^{\text{ep}}, \{s_{\tau,t}^{\text{step}}\}_{t \in \mathcal{C}_\tau}),$$

where  $\mathcal{C}_\tau$  is the sparse set of critical timesteps identified by the analyzer.

### 3.3 CRITICAL-FIRST SKILL-CONDITIONED SELF-DISTILLATION

Applying the same skills to every step is suboptimal. Episode-level skills are robust but may be too coarse at decisive states, whereas step-level skills are precise but sparse. OPID therefore introduces critical-first skill routing before performing skill-conditioned self-distillation. For trajectory  $\tau$  and timestep  $t$ , the routed skill is

$$s_{\tau,t} = \begin{cases} s_{\tau,t}^{\text{step}}, & \text{if } t \in \mathcal{C}_\tau, \\ s_\tau^{\text{ep}}, & \text{otherwise.} \end{cases}$$

Equivalently, define routing masks

$$q_{\tau,t}^{\text{step}} = \mathbb{I}[t \in \mathcal{C}_\tau], \quad q_{\tau,t}^{\text{ep}} = \mathbb{I}[t \notin \mathcal{C}_\tau].$$

The critical-first rule enforces

$$q_{\tau,t}^{\text{step}} = 1 \Rightarrow q_{\tau,t}^{\text{ep}} = 0,$$

so the two skill levels are not blindly combined. Each step receives the most appropriate granularity.

After routing, OPID converts the selected skill into token-level self-distillation supervision. Let  $H(\cdot, \cdot)$  denote a deterministic skill-injection function that appends or prepends the routed skill to the interaction history while preserving the original state information. The skill-augmented history is

$$\tilde{h}_{\tau,t} = H(h_{\tau,t}, s_{\tau,t}).$$

The original response  $y_{\tau,t}$  is not regenerated. Instead, the old policy  $\pi_{\theta_{\text{old}}}$  scores the same sampled response under both the original and skill-augmented histories. For token  $\ell$  in response  $y_{\tau,t}$ , define

$$\ell_{\tau,t,\ell}^{\text{old}} = \log \pi_{\theta_{\text{old}}}(y_{\tau,t,\ell} \mid h_{\tau,t}, y_{\tau,t,<\ell}),$$

and

$$\ell_{\tau,t,\ell}^{\text{skill}} = \log \pi_{\theta_{\text{old}}}(y_{\tau,t,\ell} \mid \tilde{h}_{\tau,t}, y_{\tau,t,<\ell}).$$

The skill-based self-teacher advantage is

$$A_{\tau,t,\ell}^{\text{skill}} = (\ell_{\tau,t,\ell}^{\text{skill}} - \ell_{\tau,t,\ell}^{\text{old}}) m_{\tau,t,\ell},$$

where  $m_{\tau,t,\ell} \in \{0, 1\}$  is the valid response-token mask.

If  $A_{\tau,t,\ell}^{\text{skill}} > 0$ , the selected skill makes the token more likely under the old policy, suggesting that the token is consistent with the skill. If  $A_{\tau,t,\ell}^{\text{skill}} < 0$ , the skill-conditioned context assigns lower probability to the token, suggesting that the token is less aligned with the routed hindsight skill. This procedure yields dense token-level guidance without requiring an external expert action.

### 3.4 POLICY OPTIMIZATION WITH SKILL ADVANTAGE

For each rollout group  $\mathcal{G}_q$ , let  $\mathbf{r}_q = \{R(\tau') \mid \tau' \in \mathcal{G}_q\}$  denote the set of outcome rewards of all trajectories sampled for the same prompt. Following GRPO, the group mean is defined as

$$\mu_q = \text{mean}(\mathbf{r}_q) = \frac{1}{|\mathcal{G}_q|} \sum_{\tau' \in \mathcal{G}_q} R(\tau').$$

The group standard deviation is defined as the square root of the group reward variance:

$$\sigma_q = \text{std}(\mathbf{r}_q) = \sqrt{\frac{1}{|\mathcal{G}_q|} \sum_{\tau' \in \mathcal{G}_q} (R(\tau') - \mu_q)^2}.$$

The GRPO-style episode-relative advantage is then computed by normalizing the trajectory outcome reward within its prompt group:

$$A_{\tau}^{\text{ep}} = \frac{R(\tau) - \mu_q}{\sigma_q}, \quad \tau \in \mathcal{G}_q.$$

This scalar is broadcast to all valid response tokens:

$$A_{\tau,t,\ell}^{\text{ep}} = A_{\tau}^{\text{ep}} m_{\tau,t,\ell}.$$

The final OPID advantage combines group-relative outcome feedback with token-level skill supervision:

$$A_{\tau,t,\ell}^{\text{OPID}} = A_{\tau,t,\ell}^{\text{ep}} + \lambda_{\text{skill}} A_{\tau,t,\ell}^{\text{skill}}.$$

This formulation keeps outcome reward as the primary RL signal while adding token-level shaping.

We optimize the standard clipped policy objective:

$$\mathcal{L}_{\text{policy}}(\theta) = -\mathbb{E}_{\tau,t,\ell} [\min(\rho_{\tau,t,\ell}(\theta) A_{\tau,t,\ell}^{\text{OPID}}, \text{clip}(\rho_{\tau,t,\ell}(\theta), 1 - \epsilon, 1 + \epsilon) A_{\tau,t,\ell}^{\text{OPID}})] + \beta \mathcal{L}_{\text{KL}}(\theta).$$

where  $\rho_{\tau,t,\ell}(\theta)$  denotes the token-level importance ratio, defined as

$$\rho_{\tau,t,\ell}(\theta) = \exp(\log \pi_{\theta}(y_{\tau,t,\ell} \mid h_{\tau,t}, y_{\tau,t,<\ell}) - \log \pi_{\theta_{\text{old}}}(y_{\tau,t,\ell} \mid h_{\tau,t}, y_{\tau,t,<\ell})).$$

The operator  $\text{clip}(x, 1 - \epsilon, 1 + \epsilon)$  truncates  $x$  to the interval  $[1 - \epsilon, 1 + \epsilon]$ , and  $\epsilon$  is the clipping hyperparameter that controls the maximum allowed deviation from the old policy.

Table 1: **Performance Comparison on the representative long-horizon benchmarks (ALFWorld, Search-based QA, and WebShop).** We report the success rate (%) on ALFWorld, accuracy on search-based QA, and task-completion score/success rate on WebShop. An asterisk (\*) denotes validation with skills. The **best** and **second-best** results are highlighted.

Method	ALFWorld						Search-based QA						WebShop				
	Pick	Look	Clean	Heat	Cool	Pick2	Avg	NQ	Triv	Pop	Hotp	2Wk	MuS	Bam	Avg	Score	Succ.
<i>Qwen2.5-3B-Instruct</i>																	
Vanilla	44.4	11.1	6.2	15.4	28.6	12.5	21.9	24.6	48.1	31.0	26.3	25.3	7.2	59.7	31.7	6.7	0.8
Skill-Prompt*	51.7	66.7	48.4	0.0	4.3	10.0	28.9	23.7	46.2	30.6	24.4	22.1	7.5	12.5	23.9	0.2	0.8
OPSD	48.8	41.7	16.7	0.0	15.8	16.7	28.1	0.1	0.1	0.1	0.0	0.0	0.0	0.0	11.3	3.1	
GRPO	91.2	62.5	96.2	61.9	65.0	47.4	75.0	39.3	60.6	41.1	37.4	34.6	15.4	26.4	36.4	79.8	63.3
Skill-GRPO	88.9	71.4	58.8	70.6	40.7	29.2	60.2	43.5	58.8	43.0	36.8	32.2	11.7	12.5	34.1	77.3	60.9
Skill-GRPO*	94.3	57.1	100.0	66.7	73.1	57.1	80.5	44.3	59.6	44.3	39.0	36.1	14.5	14.9	36.1	76.3	66.4
GRPO+OPSD	100.0	82.4	85.7	75.0	70.0	60.0	81.2	44.9	61.2	45.2	40.4	38.5	16.0	66.1	44.6	77.8	66.4
Skill-SD	88.2	50.0	96.2	52.4	65.0	57.9	73.4	44.4	60.4	44.0	39.5	40.4	15.4	64.9	44.1	75.9	64.0
RLSD	87.9	75.0	90.9	75.0	73.1	68.4	79.7	41.5	58.6	42.3	40.4	40.2	16.8	66.9	43.8	84.4	66.4
SDAR	97.1	62.5	100.0	61.9	75.0	84.2	84.4	44.8	58.1	44.3	38.6	36.2	15.7	66.1	43.4	85.0	68.0
OPID	92.7	100.0	88.9	70.0	84.2	70.0	84.3	45.9	61.4	45.7	40.7	38.8	16.4	66.1	45.0	85.0	74.2
<i>Qwen2.5-7B-Instruct</i>																	
Vanilla	36.1	22.2	3.1	0.0	0.0	0.0	12.5	25.2	50.8	29.5	29.0	29.0	10.4	63.7	33.9	5.9	1.6
Skill-Prompt*	51.7	50.0	32.3	5.3	4.3	0.0	23.4	30.9	52.1	32.7	32.7	27.9	12.7	66.1	36.4	1.7	0.8
OPSD	50.0	60.0	22.7	21.4	17.6	9.5	32.8	8.8	8.6	17.5	2.5	4.2	0.5	1.2	6.2	4.5	2.3
GRPO	91.2	87.5	96.2	81.0	65.0	57.9	81.2	45.1	63.7	44.0	43.6	43.2	16.8	37.6	42.0	80.9	72.6
Skill-GRPO	88.5	66.7	65.2	61.1	57.7	73.1	69.5	45.2	63.7	45.7	43.1	43.3	19.6	21.4	40.3	80.4	71.9
Skill-GRPO*	100.0	83.3	96.4	83.3	75.0	78.9	88.3	44.8	63.0	45.1	43.7	43.7	20.5	71.4	47.5	87.0	81.2
GRPO+OPSD	91.4	61.5	100.0	87.5	76.5	52.2	80.4	47.3	64.5	46.9	43.8	39.3	18.0	69.4	47.0	86.8	76.5
Skill-SD	93.9	93.8	90.9	100.0	69.2	68.4	85.1	47.1	64.5	47.8	44.2	42.1	20.2	69.0	47.8	86.1	76.5
RLSD	100.0	87.5	92.3	58.8	80.0	65.2	82.0	46.8	63.0	44.4	45.5	48.9	21.5	73.0	49.0	87.4	77.3
SDAR	94.7	75.0	100.0	86.7	68.2	78.9	85.9	46.3	63.5	48.2	43.8	48.4	19.6	73.0	49.0	89.4	82.8
OPID	100.0	81.8	97.1	100.0	80.8	80.0	90.0	48.8	65.6	46.8	46.1	42.7	21.7	72.6	49.2	85.3	79.7
<i>Qwen3-1.7B-Instruct</i>																	
Vanilla	25.0	22.2	3.1	0.0	21.4	4.2	12.5	29.4	46.9	37.0	23.5	19.6	6.4	10.5	24.8	46.5	4.7
Skill-Prompt*	10.3	50.0	16.1	0.0	0.0	5.0	9.4	29.4	46.5	36.2	22.9	20.8	4.3	10.1	24.3	23.0	2.3
OPSD	26.3	33.3	9.1	0.0	4.5	5.3	14.1	4.2	8.3	4.6	6.6	15.3	0.7	1.2	5.8	47.4	9.3
GRPO	71.1	41.7	36.4	40.0	31.8	31.6	46.1	40.0	58.9	43.5	35.4	30.3	12.0	65.7	40.8	67.3	38.3
Skill-GRPO	27.6	54.5	22.7	27.3	0.0	19.2	21.1	39.2	58.6	43.9	35.2	28.2	11.5	66.1	40.4	73.4	46.1
Skill-GRPO*	31.4	42.9	51.9	8.3	11.5	7.1	28.1	38.0	58.4	43.9	36.3	29.0	12.5	66.9	40.7	80.4	50.0
GRPO+OPSD	38.2	50.0	30.8	28.6	30.0	21.1	32.0	40.7	58.9	45.0	37.0	34.6	13.3	65.7	42.2	70.7	38.3
Skill-SD	52.9	37.5	69.2	42.9	60.0	36.8	52.3	39.1	57.5	45.4	34.8	34.1	10.7	64.1	40.8	81.8	53.9
RLSD	50.0	37.5	61.5	19.0	50.0	21.1	42.2	38.6	57.3	43.0	34.5	34.1	11.5	65.3	40.6	74.0	50.8
SDAR	73.5	25.0	76.9	33.3	40.0	36.8	53.9	39.7	58.9	45.3	35.9	35.5	12.6	65.3	41.9	76.8	58.6
OPID	65.9	72.7	66.7	40.0	63.2	45.0	58.9	38.1	58.1	43.4	35.5	31.7	11.7	64.5	40.4	79.6	64.8

**Training-inference boundary.** The analyzer, routed skills, and skill-conditioned scoring pass are used only to construct the training advantage. At inference time, the learned policy acts from the ordinary interaction history  $h_t$  alone, with no analyzer call, skill retrieval, or privileged context.

## 4 EXPERIMENT

### 4.1 EXPERIMENTAL SETTING

**Benchmarks.** We evaluate OPID on three representative agentic benchmarks that require multi-step interaction or search-based reasoning. First, we use ALFWorld (Shridhar et al., 2020), an embodied household benchmark where an agent must complete language-specified goals through a sequence of textual actions. We report performance on six task types: *Pick*, *Look*, *Clean*, *Heat*, *Cool*, and *Pick2*. Second, we evaluate on WebShop (Yao et al., 2022), where an agent interacts with an e-commerce website to find and purchase products satisfying natural-language user requirements. Following the standard evaluation protocol, we report results on 128 test tasks. Third, we consider Search-based QA (Jin et al., 2025), where the agent answers questions by interacting with a search environment: Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023).

**Baselines.** We compare OPID against both prompting-based and training-based baselines. *Vanilla* denotes the original prompting baseline. *Skill-Prompt* augments the model with skill descriptions at inference or validation time. *GRPO* is the outcome-only on-policy RL baseline, where the policy is optimized using group-relative trajectory-level rewards (Shao et al., 2024). *Skill-GRPO* combines skill conditioning with GRPO-style outcome optimization. *OPSD* (Zhao et al., 2026), *GRPO+OPSD*, *Skill-SD* (Wang et al., 2026), *RLSD* (Yang et al., 2026), and *SDAR* (Lu et al.,

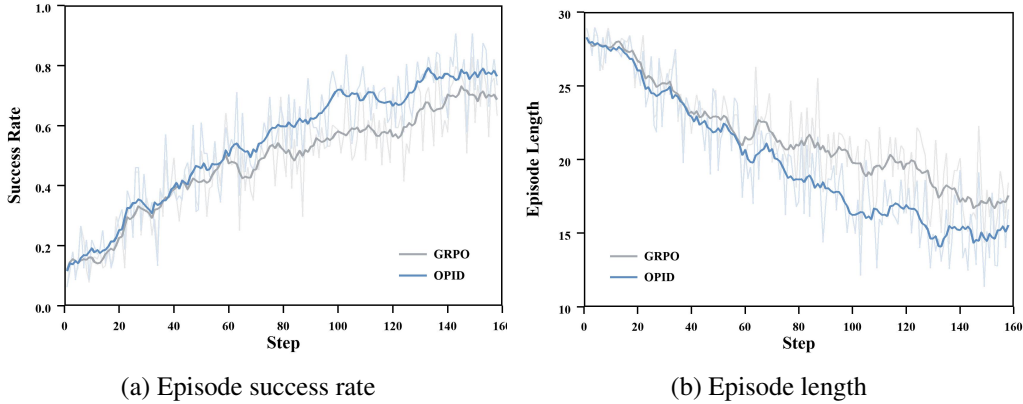


Figure 3: **Training dynamics of OPID and GRPO.** We report Qwen2.5-3B-Instruct training on ALFWorld. Translucent curves denote raw measurements and solid curves denote smoothed trends.

2026a) are self-distillation or skill-distillation baselines that introduce auxiliary token-level or skill-conditioned supervision during training. Rows marked with \* indicate validation with skills, following the setting described in the corresponding baseline.

**Evaluation Metrics.** For ALFWorld, we report task success rate in percentage. For WebShop, we report both the normalized task score and task success rate, following the benchmark protocol. For Search-based QA, we report answer accuracy in percentage on each QA subset and the average accuracy across subsets.

**Implementation Details.** We conduct experiments using Qwen2.5-3B/7B-Instruct (Yang et al., 2024) and Qwen3-1.7B-Instruct (Yang et al., 2025). The training batch size is set to 16 for ALFWorld and WebShop, and 128 for Search-based QA. All models are trained for 150 steps across all environments. Full details are provided in Appendix B.

## 4.2 MAIN RESULTS

Table 1 summarizes performance across model scales and agentic domains, revealing three key findings:

**OPID consistently strengthens outcome-only RL.** OPID improves over GRPO in most model-domain combinations. On Qwen2.5-3B, the gains are +9.3 points on ALFWorld (84.3 vs. 75.0), +8.6 on Search-based QA (45.0 vs. 36.4), and +10.9 on WebShop (74.2 vs. 63.3). The corresponding improvements on Qwen2.5-7B are +8.8, +7.2, and +7.1 points. The benefit is particularly pronounced for the smaller Qwen3-1.7B backbone, where OPID improves ALFWorld by +12.8 points and WebShop by +26.5 points. The only exception is Search-based QA on Qwen3-1.7B, where OPID remains close to GRPO. Overall, these results show that OPID usually provides a consistent gain over outcome-only reinforcement learning, especially on long-horizon embodied and web-shopping tasks.

**OPID remains competitive with strong hybrid methods.** Beyond improving over outcome-only RL, OPID also matches or surpasses strong hybrid and self-distillation baselines in several aggregate settings. On ALFWorld, OPID achieves the best average on Qwen2.5-7B and Qwen3-1.7B, outperforming the strongest baseline by +1.7 points (90.0 vs. 88.3) and +5.0 points (58.9 vs. 53.9) respectively. On Search-based QA, OPID attains the best average on both Qwen2.5 backbones, improving over the strongest baseline by +0.4 points on Qwen2.5-3B (45.0 vs. 44.6) and +0.2 points on Qwen2.5-7B (49.2 vs. 49.0). On WebShop, OPID achieves the best success rate on Qwen2.5-3B and Qwen3-1.7B, exceeding the strongest competing method by +6.2 points on Qwen3-1.7B (64.8 vs. 58.6), while remaining competitive on Qwen2.5-7B. These results show that trajectory-derived, distribution-matched skills can complement outcome supervision and compete with methods that rely on hybrid training signals or external skill contexts.

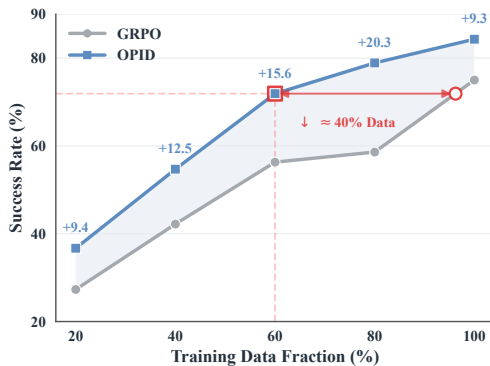


Figure 4: **Sample efficiency analysis.** OPID consistently outperforms GRPO under reduced training data and approaches full-data GRPO performance using about 60% of the data.

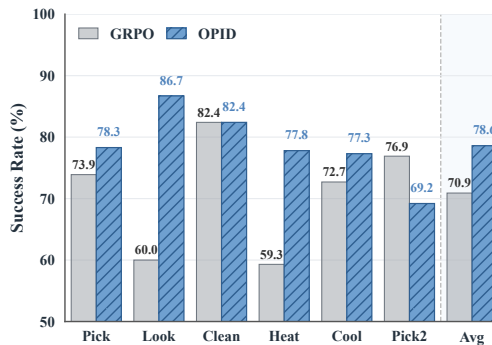


Figure 5: **Cross-domain generalization on ALFWorld Unseen.** OPID improves the average success rate over GRPO and shows particularly large gains on *Look* and *Heat*.

**OPID internalizes skills instead of depending on them at inference.** The results further show that OPID gains from internalizing hindsight skills into the policy, rather than relying on skill prompts at inference time. Training directly with retrieved skills introduces a clear train–test context mismatch: when validation-time skills are removed, Skill-GRPO underperforms ordinary GRPO on ALFWorld at all model scales, dropping by -14.8 points on Qwen2.5-3B (60.2 vs. 75.0), -11.7 points on Qwen2.5-7B (69.5 vs. 81.2), and -25.0 points on Qwen3-1.7B (21.1 vs. 46.1). In contrast, OPID is also evaluated without any skill input, yet exceeds Skill-GRPO by +24.1, +20.5, and +37.8 points. On Search-based QA, OPID also improves over both GRPO and Skill-GRPO for the two Qwen2.5 models, with gains over GRPO of +8.6 and +7.2 points, while remaining comparable on Qwen3-1.7B. Moreover, OPID outperforms Skill-GRPO\* on ALFWorld and Search-based QA for both Qwen2.5 backbones, even though Skill-GRPO\* retains privileged skill context during validation. These results indicate that OPID transfers trajectory-derived hindsight knowledge into the model parameters, enabling the policy to benefit from skills without depending on external skill prompts at inference.

### 4.3 TRAINING DYNAMICS

Figure 3 illustrates the training progression on ALFWorld. Both methods improve during early optimization, yet OPID diverges from GRPO in the middle stage and maintains superior performance throughout the remainder of training. This divergence pattern indicates that hindsight skill supervision accelerates policy refinement beyond what outcome rewards alone can achieve. The efficiency gains are equally pronounced. OPID reduces average episode length to 15-16 steps while GRPO plateaus at 17-18 steps. The concurrent rise in success and fall in trajectory length reveals a key behavioral shift: OPID agents learn to reach goals through more direct action sequences rather than exploratory detours.

These dynamics align with the intended function of hierarchical supervision. Episode-level skills establish coherent task workflows that reduce backtracking and repetition. Step-level skills provide precise guidance at critical decision points, preventing the invalid actions and local navigation errors that otherwise extend trajectories. Together, these mechanisms enable OPID to internalize both global task structure and local decision efficiency.

### 4.4 SAMPLE EFFICIENCY

Figure 4 compares OPID and GRPO under different fractions of ALFWorld training data. OPID consistently improves over GRPO across all data scales, with absolute gains ranging from +9.3 to +20.3 points. The advantage is especially clear in the low- and mid-data regimes, where each trajectory carries more training value. With 60% of the data, OPID reaches 71.9, close to GRPO trained with the full dataset (75.0); with 80% of the data, it already surpasses full-data GRPO (78.9

vs. 75.0). These results indicate that OPID-style skill supervision improves the data efficiency of outcome-based RL. By converting completed trajectories into dense token-level training signals, OPID extracts additional supervision from the same environment interactions rather than relying only on terminal rewards. This makes the optimization less dependent on large numbers of rollouts and allows the policy to acquire effective behaviors with fewer samples.

#### 4.5 CROSS-DOMAIN GENERALIZATION

Figure 5 evaluates cross-domain transfer to the ALFWorld unseen split. OPID achieves an average success rate of 78.6, outperforming GRPO by +7.7 points. Its gains over GRPO are concentrated on tasks like *Look* (+26.7) and *Heat* (+18.5), while maintaining competitive performance on the remaining task types. These results suggest that OPID is not merely memorizing the observed training trajectories. Instead, the extracted skills appear to capture reusable behavioral structure, including high-level task workflows and local decision rules that remain useful under unseen environment configurations. Since the skills are distilled into the policy rather than retrieved at inference time, the improvement also indicates that OPID internalizes transferable decision knowledge into the model parameters.

Table 2: **Ablation on Hierarchical Skills.** We report the success rate (%) on ALFWorld and Score/Succ. (%) on WebShop with Qwen2.5-3B-Instruct backbone.

Method	ALFWorld						WebShop		
	Pick	Look	Clean	Heat	Cool	Pick2	Avg.	Score	Succ.
OPID	92.7	<b>100.0</b>	<b>88.9</b>	<b>70.0</b>	<b>84.2</b>	70.0	<b>84.3</b>	<b>85.0</b>	<b>74.2</b>
w/o episode skill	83.3	80.0	78.1	69.2	57.7	<b>76.5</b>	74.1	78.4	67.2
w/o step skill	<b>95.1</b>	<b>81.8</b>	<b>88.9</b>	<b>70.0</b>	79.0	60.0	79.1	80.2	65.6

Table 3: **Ablation of Critical-First Skill Routing.** With the Qwen2.5-3B-Instruct backbone, we compare OPID with a variant that removes the critical-first routing strategy.

Method	ALFWorld						
	Pick	Look	Clean	Heat	Cool	Pick2	Avg.
OPID	92.7	<b>100.0</b>	<b>88.9</b>	<b>70.0</b>	<b>84.2</b>	<b>70.0</b>	<b>84.3</b>
w/o Routing	<b>95.1</b>	81.8	<b>88.9</b>	50.0	<b>84.2</b>	65.0	77.5

#### 4.6 ABLATION STUDIES AND ANALYSIS

We isolate the contributions of hierarchical skill granularity and critical-first routing using Qwen2.5-3B-Instruct.

**Impact of Hierarchical Skills.** As shown in Table 2, the complete hierarchy obtains the best aggregate performance on both domains. Removing episode-level skills decreases the ALFWorld average from 84.3 to 74.1 and the WebShop success rate from 74.2 to 67.2, confirming that global workflows and failure-avoidance rules provide an important default signal. Removing step-level skills decreases the ALFWorld average from 84.3 to 79.1 and the WebShop success rate from 74.2 to 65.6. These results demonstrate the complementarity of the two skill levels.

**Impact of Critical-First Skill Routing.** Table 3 compares OPID with a non-routed variant that applies the episode-level skill to every step and additionally incorporates the corresponding step-level skill at critical timesteps, thereby superimposing the two forms of guidance. Critical-first routing improves the ALFWorld average by +6.8 points (84.3 vs. 77.5). These results show that selectively routing the most appropriate skill granularity is more effective than directly combining global and local guidance, demonstrating the importance of critical-first routing.

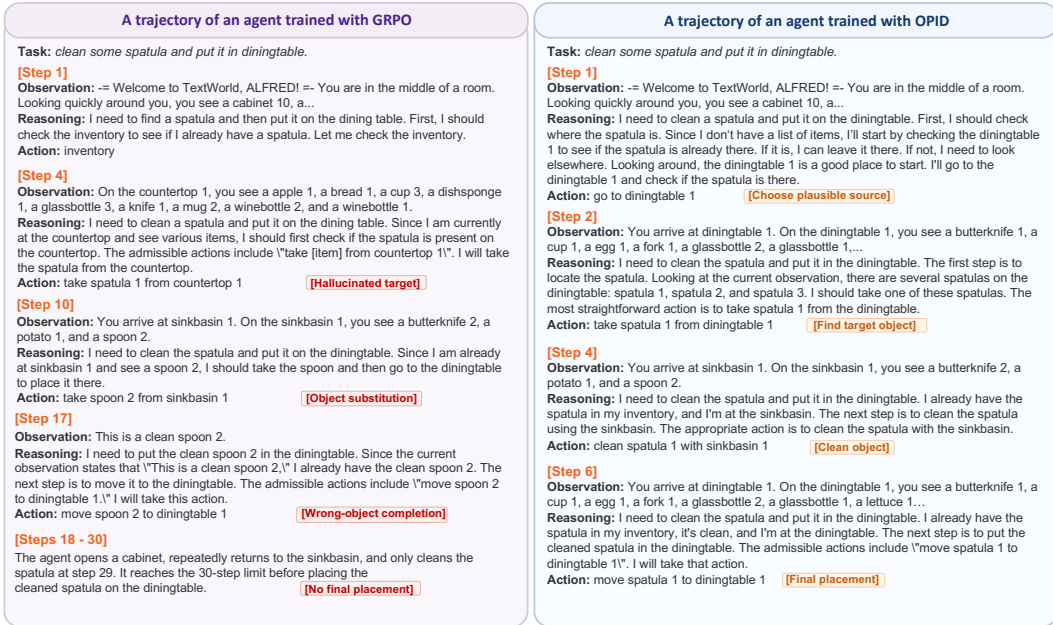


Figure 6: **Qualitative comparison on ALFWorld.** For the task “clean some spatula and put it in diningtable,” the GRPO-trained agent hallucinates a nonexistent target object, substitutes a spoon for the spatula, and fails to complete the final placement within the step limit. In contrast, OPID follows a coherent locate-clean-place workflow, grounding each action in the current observation and completing the task in six steps.

**Qualitative Analysis.** Figure 6 illustrates an ALFWorld clean-and-place task. The GRPO-trained agent exhibits a “hallucinated target” error by attempting to take a nonexistent spatula from the countertop at Step 4. It subsequently substitutes a spoon for the target object and reaches the 30-step limit before placing the cleaned spatula back on the dining table. In contrast, OPID follows a coherent locate-clean-place workflow and completes the task in six steps. This case suggests that distilling hierarchical hindsight skills from on-policy trajectories helps the agent learn both local object-grounding decisions and episode-level task workflows, thereby reducing hallucinated actions and preserving progress toward the final goal.

## 5 CONCLUSION

We presented OPID, an on-policy skill distillation framework that turns completed agent trajectories into hierarchical hindsight supervision. By extracting episode-level and step-level skills from the current policy’s own rollouts, OPID provides dense, distribution-matched token-level guidance while preserving outcome-based RL as the primary objective. Experiments across embodied, web, and search-based agentic benchmarks show that OPID improves agent learning without relying on external skill libraries, retrieval, or privileged context at inference time. More broadly, our results suggest that agent trajectories are not only samples for reward optimization, but also reusable records of decision knowledge that can be distilled back into the policy.

## REFERENCES

- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.
- Guoxin Chen, Zile Qiao, Xuanchong Chen, Donglei Yu, Haotian Xu, Xin Zhao, Ruihua Song, Wenbiao Yin, Huifeng Yin, Liwen Zhang, Kuan Li, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei

- 
- Huang, and Jingren Zhou. Iterresearch: Rethinking long-horizon agents with interaction scaling. In *The Fourteenth International Conference on Learning Representations*, 2026.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *arXiv preprint arXiv:2306.06070*, 2023.
- Kuofei Fang, Xinyi Che, Haomin Ouyang, Shufan Zhang, Xuehao Wang, Qi Liu, Liyi Liu, Chenqi Zhang, Wenxi Cai, Wenyu Dai, et al. Roboteq: Transitioning from passive intelligence to active intelligence in embodied ai. *arXiv preprint arXiv:2605.06234*, 2026.
- Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiakai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting on-policy distillation: Empirical failure modes and simple fixes. *arXiv preprint arXiv:2603.25562*, 2026.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024a.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024b.
- Yinghui He, Simran Kaur, Adithya Bhaskar, Yongjin Yang, Jiarui Liu, Narutatsu Ri, Liam Fowl, Abhishek Panigrahi, Danqi Chen, and Sanjeev Arora. Self-distillation zero: Self-revision turns binary rewards into dense supervision. *arXiv preprint arXiv:2604.12002*, 2026.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.
- Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Wang Dong, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.
- Mandar Joshi, Eunsol Choi, Daniel 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*, pp. 1601–1611, 2017.
- Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. *arXiv preprint arXiv:2401.13649*, 2024.
- Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.
- 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.
- Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-gao Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026.

- 
- Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.
- Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026.
- 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. *arXiv preprint arXiv:2605.15155*, 2026a.
- Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Chengcheng Han, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. Skill0: In-context agentic reinforcement learning for skill internalization. *arXiv preprint arXiv:2604.02268*, 2026b.
- Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.
- 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*, pp. 9802–9822, 2023.
- Minjae Oh, Sangjun Song, Gyubin Choi, Yunho Choi, and Yohan Jo. KL for a KL: On-policy distillation with control variate baseline. *arXiv preprint arXiv:2605.07865*, 2026.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5687–5711, 2023.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Yuhao Shen, Tianyu Liu, Junyi Shen, Jinyang Wu, Quan Kong, Li Huan, and Cong Wang. Double: Breaking the acceleration limit via double retrieval speculative parallelism. *arXiv preprint arXiv:2601.05524*, 2026.
- Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. AlfworlD: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*, 2020.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.
- 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. *arXiv preprint arXiv:2604.10674*, 2026.
- Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, Chonghua Liao, and Jianhua Tao. Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts. *arXiv preprint arXiv:2411.18478*, 2024.
- Jinyang Wu, Shuo Yang, Changpeng Yang, Yuhao Shen, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. Spark: Strategic policy-aware exploration via dynamic branching for long-horizon agentic learning. *arXiv preprint arXiv:2601.20209*, 2026a.

- 
- Jinyang Wu, Guocheng Zhai, Ruihan Jin, Yuhao Shen, Zhengxi Lu, Fan Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and Jianhua Tao. Maestro: Reinforcement learning to orchestrate hierarchical model-skill ensembles. *arXiv preprint arXiv:2605.22177*, 2026b.
- Jinyang Wu, Guocheng Zhai, Ruihan Jin, Jiahao Yuan, Yuhao Shen, Shuai Zhang, Zhengqi Wen, and Jianhua Tao. Atlas: Orchestrating heterogeneous models and tools for multi-domain complex reasoning. *arXiv preprint arXiv:2601.03872*, 2026c.
- Fangzhi Xu, Hang Yan, Qiushi Sun, Jinyang Wu, Zixian Huang, Muye Huang, Jingyang Gong, Zichen Ding, Kanzhi Cheng, Yian Wang, et al. Odysseyarena: Benchmarking large language models for long-horizon, active and inductive interactions. *arXiv preprint arXiv:2602.05843*, 2026.
- 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. *arXiv preprint arXiv:2412.15115*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.
- 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, 2018.
- 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*, 2022.
- Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*, 2026.
- Z.ai. GLM-5.2: Built for Long-Horizon Tasks. <https://z.ai/blog/glm-5.2>, June 2026. Accessed: 2026-06-22.
- Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025.
- Siyao Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026.
- Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

## A THEORETICAL ANALYSIS

This section provides three results that correspond to the main design choices of OPID. We first place the proposed teacher advantage among representative on-policy distillation objectives. We then show that it implements a sampled-token reverse-KL update, characterize the benefit of collecting distillation contexts on policy, and justify critical-first routing under a natural specialization assumption.

---

## A.1 NOTATION AND REPRESENTATIVE ON-POLICY DISTILLATION OBJECTIVES

### A.1.1 NOTATION

Let  $i = (\tau, t, \ell)$  index a valid token position in a response. We denote the corresponding standard autoregressive context by  $c_i = (h_{\tau,t}, y_{\tau,t,<\ell})$ , and its skill-augmented counterpart by  $\tilde{c}_i$ . At each token position, define

$$b_i(v) \triangleq \pi_{\theta_{\text{old}}}(v | c_i), \quad q_i(v) \triangleq \pi_{\theta_{\text{old}}}(v | \tilde{c}_i), \quad p_{\theta,i}(v) \triangleq \pi_{\theta}(v | c_i).$$

Here,  $b_i$  is the behavior distribution used to generate the response,  $q_i$  is a detached skill-conditioned teacher distribution, and  $p_{\theta,i}$  is the trainable policy evaluated under the standard context available at inference time. The observed token  $a_i \triangleq y_{\tau,t,\ell}$  is sampled from  $b_i$ .

We further define the token-level log-likelihood gap and the policy importance ratio as

$$\Delta_i(v) \triangleq \log q_i(v) - \log b_i(v), \quad \rho_{\theta,i}(v) \triangleq \frac{p_{\theta,i}(v)}{b_i(v)}. \quad (1)$$

The quantity  $\Delta_i(v)$  measures the change in token log-probability induced by the skill-augmented context. In particular,  $\Delta_i(v) > 0$  indicates that the skill-conditioned teacher assigns greater probability to token  $v$  than the behavior policy does. The OPID skill advantage associated with the observed token is therefore

$$A_i^{\text{skill}} = \Delta_i(a_i).$$

Unless otherwise stated, all expectations below are taken over valid response tokens; the response mask is consequently omitted for notational simplicity.

### A.1.2 REPRESENTATIVE ON-POLICY DISTILLATION OBJECTIVES

On-policy distillation (OPD) applies teacher supervision at autoregressive contexts generated by the student or a behavior policy, thereby reducing the context-distribution mismatch between distillation training and free-running inference (Agarwal et al., 2024). The context-generation policy and the granularity of teacher supervision are orthogonal design choices. At each on-policy context, output-space OPD objectives can be organized into three common supervision granularities: full-vocabulary, Top- $K$ , and sampled-token distillation (Li et al., 2026; Fu et al., 2026). OPID belongs to the sampled-token category.

**Full-vocabulary distribution matching.** Let  $q_i$  and  $p_{\theta,i}$  denote the teacher and student next-token distributions, respectively, at autoregressive context  $i$ . When the complete predictive distributions are available, OPD can minimize the forward KL, reverse KL, or a generalized Jensen–Shannon divergence (Hinton et al., 2015; Agarwal et al., 2024; Gu et al., 2024b):

$$\begin{aligned} \mathcal{L}_{\text{FKL}}(\theta) &= \mathbb{E}_i[D_{\text{KL}}(q_i \| p_{\theta,i})], \\ \mathcal{L}_{\text{RKL}}(\theta) &= \mathbb{E}_i[D_{\text{KL}}(p_{\theta,i} \| q_i)], \\ \mathcal{L}_{\text{JSD}}^{(\alpha)}(\theta) &= \mathbb{E}_i\left[\alpha D_{\text{KL}}(q_i \| m_i^{(\alpha)}) + (1 - \alpha) D_{\text{KL}}(p_{\theta,i} \| m_i^{(\alpha)})\right], \\ m_i^{(\alpha)} &= \alpha q_i + (1 - \alpha) p_{\theta,i}. \end{aligned}$$

Forward KL gives the conventional soft-target objective and emphasizes coverage of teacher-supported probability mass. Reverse KL instead penalizes student probability assigned to teacher-disfavored regions and therefore typically exhibits more mode-seeking behavior. Generalized JSD compares both models against a mixture distribution, with  $\alpha = \frac{1}{2}$  recovering the standard symmetric JSD (Kullback & Leibler, 1951; Lin, 1991).

**Top- $K$  distribution matching.** Top- $K$  OPD retains distribution-level supervision over a restricted local support. Common choices include a student-selected support (Li et al., 2026; Ye et al., 2026) and a teacher-selected support (Fu et al., 2026):

$$\begin{aligned} S_{i,p}^{(K)} &\triangleq \text{TopK}(p_{\theta,i}, K), \\ S_{i,q}^{(K)} &\triangleq \text{TopK}(q_i, K), \\ S_i^{(K)} &\in \left\{ S_{i,p}^{(K)}, S_{i,q}^{(K)} \right\}. \end{aligned}$$

For the selected support  $S_i^{(K)}$ , define the restricted and renormalized distributions

$$\begin{aligned}\bar{p}_{\theta,i}^{S_i^{(K)}}(v) &\triangleq \frac{p_{\theta,i}(v)\mathbf{1}\{v \in S_i^{(K)}\}}{\sum_{u \in S_i^{(K)}} p_{\theta,i}(u)}, \\ \bar{q}_i^{S_i^{(K)}}(v) &\triangleq \frac{q_i(v)\mathbf{1}\{v \in S_i^{(K)}\}}{\sum_{u \in S_i^{(K)}} q_i(u)}.\end{aligned}$$

A representative truncated reverse-KL objective is

$$\begin{aligned}\mathcal{L}_{\text{TopK-RKL}}(\theta) &= \mathbb{E}_i \left[ D_{\text{KL}} \left( \bar{p}_{\theta,i}^{S_i^{(K)}} \parallel \bar{q}_i^{S_i^{(K)}} \right) \right] \\ &= \mathbb{E}_i \left[ \sum_{v \in S_i^{(K)}} \bar{p}_{\theta,i}^{S_i^{(K)}}(v) \log \frac{\bar{p}_{\theta,i}^{S_i^{(K)}}(v)}{\bar{q}_i^{S_i^{(K)}}(v)} \right].\end{aligned}$$

Top- $K$  matching occupies an intermediate point between one-token and full-vocabulary supervision. It preserves multi-token information at reduced computational or communication cost, but discards probability mass outside the selected support and is therefore a truncated, support-dependent approximation to the full reverse KL.

**Sampled-token distillation.** At a fixed on-policy context, define the teacher–student log-ratio cost

$$\delta_i(v) \triangleq \log p_{\theta,i}(v) - \log q_i(v).$$

The token-level reverse KL can then be written exactly as an expectation over student-sampled tokens:

$$\begin{aligned}D_{\text{KL}}(p_{\theta,i} \parallel q_i) &= \mathbb{E}_{a_i \sim p_{\theta,i}} [\delta_i(a_i)] \\ &= \mathbb{E}_{a_i \sim b_i} [\rho_{\theta,i}(a_i) \delta_i(a_i)], \quad \rho_{\theta,i}(a) \triangleq \frac{p_{\theta,i}(a)}{b_i(a)}.\end{aligned}$$

The second equality requires  $p_{\theta,i} \ll b_i$  (support coverage condition). Consequently,  $\rho_{\theta,i}(a_i) \delta_i(a_i)$  is an importance-weighted single-sample estimator of the per-context reverse KL. Its score-function gradient is

$$\nabla_{\theta} D_{\text{KL}}(p_{\theta,i} \parallel q_i) = \mathbb{E}_{a_i \sim b_i} [\rho_{\theta,i}(a_i) \text{sg}[\delta_i(a_i)] \nabla_{\theta} \log p_{\theta,i}(a_i)],$$

where  $\text{sg}$  denotes stop-gradient. This connection permits sampled-token distillation to be implemented with policy-gradient or importance-weighted policy-optimization machinery (Gu et al., 2024b; Lu & Thinking Machines Lab, 2025; Oh et al., 2026). Compared with full-vocabulary matching, sampled-token supervision requires only the teacher probability of the realized token, but has higher Monte Carlo variance and uses less information from the teacher distribution.

**From the clipped OPID objective to its unclipped skill surrogate.** Let  $i = (\tau, t, \ell)$  index a valid rollout-token position, and let  $\nu_b$  denote the distribution over valid token positions induced by rollouts collected from the behavior policy. Given position  $i$ , the observed token  $a_i$  is sampled from  $b_i$ . Recall from Eq. 1 that

$$\Delta_i(v) = \log q_i(v) - \log b_i(v), \quad \rho_{\theta,i}(v) = \frac{p_{\theta,i}(v)}{b_i(v)},$$

where  $b_i$ ,  $q_i$ , and the resulting advantages are detached during the policy update. The skill advantage of a sampled token is

$$A_i^{\text{skill}}(a_i) = \Delta_i(a_i).$$

The complete OPID advantage combines the outcome and skill signals:

$$A_i^{\text{OPID}}(a_i) = A_i^{\text{ep}} + \lambda_{\text{skill}} \Delta_i(a_i).$$

Accordingly, the implemented clipped policy loss is

$$\mathcal{L}_{\text{policy}}(\theta) = -\mathbb{E}_{\substack{i \sim \nu_b \\ a_i \sim b_i}} \left[ \min \left( \rho_{\theta,i}(a_i) A_i^{\text{OPID}}(a_i), \right. \right. \\ \left. \left. \text{clip}(\rho_{\theta,i}(a_i), 1 - \epsilon, 1 + \epsilon) A_i^{\text{OPID}}(a_i) \right) \right].$$

In a realized rollout batch, this expectation is implemented as an empirical average over the observed valid tokens  $a_i$ .

Because PPO clipping is applied after the outcome and skill advantages have been combined, the clipped objective does not in general decompose into independently clipped outcome and skill losses. To isolate the skill-distillation signal studied below, we therefore consider the corresponding unclipped policy surrogate:

$$\mathcal{L}_{\text{policy}}^{\text{unclip}}(\theta) \triangleq -\mathbb{E}_{\substack{i \sim \nu_b \\ a_i \sim b_i}} [\rho_{\theta,i}(a_i) A_i^{\text{OPID}}(a_i)].$$

Unlike the clipped objective, this loss decomposes exactly as

$$\mathcal{L}_{\text{policy}}^{\text{unclip}}(\theta) = \mathcal{L}_{\text{ep}}^{\text{unclip}}(\theta) + \mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta),$$

where

$$\mathcal{L}_{\text{ep}}^{\text{unclip}}(\theta) \triangleq -\mathbb{E}_{\substack{i \sim \nu_b \\ a_i \sim b_i}} [\rho_{\theta,i}(a_i) A_i^{\text{ep}}]$$

and

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) \triangleq -\lambda_{\text{skill}} \mathbb{E}_{\substack{i \sim \nu_b \\ a_i \sim b_i}} [\rho_{\theta,i}(a_i) \Delta_i(a_i)]. \quad (2)$$

Equation 2 is the skill-distillation loss analyzed in the next subsection. Although it is defined through the unclipped surrogate, it characterizes the local skill-induced update of the implemented PPO loss. In particular, let  $\theta_0 = \theta_{\text{old}}$ , so that  $p_{\theta_0,i} = b_i$  and  $\rho_{\theta_0,i}(a) = 1$ . Since 1 lies in the interior of the clipping interval, the clipped and unclipped objectives have the same value and gradient at the behavior policy:

$$\begin{aligned} \mathcal{L}_{\text{policy}}^{\text{clip}}(\theta_0) &= \mathcal{L}_{\text{policy}}^{\text{unclip}}(\theta_0), \\ \nabla_{\theta} \mathcal{L}_{\text{policy}}^{\text{clip}}(\theta) \Big|_{\theta=\theta_0} &= \nabla_{\theta} \mathcal{L}_{\text{policy}}^{\text{unclip}}(\theta) \Big|_{\theta=\theta_0} \\ &= \nabla_{\theta} \mathcal{L}_{\text{ep}}^{\text{unclip}}(\theta) \Big|_{\theta=\theta_0} + \nabla_{\theta} \mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) \Big|_{\theta=\theta_0}. \end{aligned}$$

Thus,  $\mathcal{L}_{\text{skill}}^{\text{unclip}}$  is exactly the skill-induced component of the first-order PPO update around the behavior policy. Away from this local region, clipping couples the outcome and skill signals through the sign of their combined advantage, and the unclipped decomposition no longer describes the complete clipped objective globally.

## A.2 THE UNCLIPPED OPID SKILL LOSS AS A RELATIVE-KL SURROGATE

We now analyze the unclipped skill-distillation loss introduced in Eq. 2. Let  $\nu_b$  denote the distribution over valid token positions induced by rollouts collected from the behavior policy. Throughout this subsection, the rollout histories, routed skills, and the corresponding distributions  $b_i$  and  $q_i$  are detached and held fixed during the policy update.

We assume the common-support condition

$$p_{\theta,i} \ll b_i \quad \text{and} \quad p_{\theta,i} \ll q_i$$

for every  $i$  in the support of  $\nu_b$ . This condition is satisfied by standard softmax language models with finite logits.

Recall that

$$\Delta_i(v) \triangleq \log q_i(v) - \log b_i(v), \quad \rho_{\theta,i}(v) \triangleq \frac{p_{\theta,i}(v)}{b_i(v)}.$$

The unclipped OPID skill loss is

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) \triangleq -\lambda_{\text{skill}} \mathbb{E}_{\substack{i \sim \nu_b \\ a \sim b_i}} [\rho_{\theta,i}(a) \Delta_i(a)]. \quad (3)$$

In a realized rollout batch, this expectation is approximated by the empirical average over the observed valid tokens. The expectation notation in Eq. 3 makes the rollout-time token sampling law explicit for the theoretical analysis.

Define the behavior-relative KL and the student-teacher reverse-KL loss as

$$\begin{aligned} \mathcal{D}_b(\theta) &\triangleq \mathbb{E}_{i \sim \nu_b} [D_{\text{KL}}(p_{\theta,i} \| b_i)], \\ \mathcal{L}_{\text{RKL}}(\theta) &\triangleq \mathbb{E}_{i \sim \nu_b} [D_{\text{KL}}(p_{\theta,i} \| q_i)]. \end{aligned}$$

**Proposition 1** (Exact relative-KL decomposition). *Under the assumptions above, for every admissible  $\theta$ ,*

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) = \lambda_{\text{skill}} [\mathcal{L}_{\text{RKL}}(\theta) - \mathcal{D}_b(\theta)]. \quad (4)$$

Let  $\theta_0 = \theta_{\text{old}}$  and suppose that  $p_{\theta_0, i} = b_i$  for every  $i$ . Then

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta_0) = \lambda_{\text{skill}} \mathcal{L}_{\text{RKL}}(\theta_0), \quad (5)$$

$$\nabla_{\theta} \mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) \Big|_{\theta=\theta_0} = \lambda_{\text{skill}} \nabla_{\theta} \mathcal{L}_{\text{RKL}}(\theta) \Big|_{\theta=\theta_0} \quad (6)$$

$$= -\lambda_{\text{skill}} \mathbb{E}_{\substack{i \sim \nu_b \\ a \sim b_i}} \left[ \Delta_i(a) \nabla_{\theta} \log p_{\theta, i}(a) \Big|_{\theta=\theta_0} \right]. \quad (7)$$

*Proof.* Fix a valid token position  $i$ . By the common-support assumption and a change of measure from  $b_i$  to  $p_{\theta, i}$ ,

$$\begin{aligned} & -\lambda_{\text{skill}} \mathbb{E}_{a \sim b_i} [\rho_{\theta, i}(a) \Delta_i(a)] \\ &= -\lambda_{\text{skill}} \sum_{v \in \mathcal{V}} b_i(v) \frac{p_{\theta, i}(v)}{b_i(v)} (\log q_i(v) - \log b_i(v)) \\ &= \lambda_{\text{skill}} \sum_{v \in \mathcal{V}} p_{\theta, i}(v) (\log b_i(v) - \log q_i(v)). \end{aligned}$$

Adding and subtracting  $\log p_{\theta, i}(v)$  inside the summand gives

$$\begin{aligned} & \lambda_{\text{skill}} \sum_{v \in \mathcal{V}} p_{\theta, i}(v) \left[ \log \frac{p_{\theta, i}(v)}{q_i(v)} - \log \frac{p_{\theta, i}(v)}{b_i(v)} \right] \\ &= \lambda_{\text{skill}} [D_{\text{KL}}(p_{\theta, i} \| q_i) - D_{\text{KL}}(p_{\theta, i} \| b_i)]. \end{aligned}$$

Averaging over  $i \sim \nu_b$  proves Eq. 4.

At  $\theta_0$ ,  $p_{\theta_0, i} = b_i$ , and hence

$$\mathcal{D}_b(\theta_0) = 0.$$

This proves Eq. 5. Moreover,  $\mathcal{D}_b$  is differentiable and attains its global minimum at  $\theta_0$ , so

$$\nabla_{\theta} \mathcal{D}_b(\theta) \Big|_{\theta=\theta_0} = 0.$$

Differentiating Eq. 4 therefore proves Eq. 6.

Finally, because  $b_i$ ,  $q_i$ , and  $\Delta_i$  are detached,

$$\nabla_{\theta} \mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) = -\lambda_{\text{skill}} \mathbb{E}_{\substack{i \sim \nu_b \\ a \sim b_i}} [\rho_{\theta, i}(a) \Delta_i(a) \nabla_{\theta} \log p_{\theta, i}(a)].$$

Substituting  $\rho_{\theta_0, i}(a) = 1$  proves Eq. 7. □

**Remark 1** (Why the OPID skill loss is not the direct reverse-KL loss). *The scaled direct reverse-KL loss is*

$$\lambda_{\text{skill}} \mathcal{L}_{\text{RKL}}(\theta) = \lambda_{\text{skill}} \mathbb{E}_{\substack{i \sim \nu_b \\ v \sim p_{\theta, i}}} [\log p_{\theta, i}(v) - \log q_i(v)],$$

whereas the OPID skill loss can be written as

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) = \lambda_{\text{skill}} \mathbb{E}_{\substack{i \sim \nu_b \\ v \sim p_{\theta, i}}} [\log b_i(v) - \log q_i(v)].$$

*The two expressions differ because the denominator in the detached teacher advantage is the rollout policy  $b_i$ , rather than the current student  $p_{\theta, i}$ . Importance weighting changes the sampling distribution from  $b_i$  to  $p_{\theta, i}$ , but it does not replace  $\log b_i$  by  $\log p_{\theta, i}$ . Consequently,*

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) - \lambda_{\text{skill}} \mathcal{L}_{\text{RKL}}(\theta) = -\lambda_{\text{skill}} \mathcal{D}_b(\theta).$$

*Thus, the OPID skill loss is an exact relative-KL loss and only a local surrogate for direct student-teacher reverse-KL matching.*

*This distinction also changes the global optimum. For example, consider*

$$b = \left( \frac{1}{2}, \frac{1}{2} \right), \quad q = \left( \frac{3}{4}, \frac{1}{4} \right).$$

For a categorical distribution  $p = (p_1, p_2)$ ,

$$\frac{\mathcal{L}_{\text{skill}}^{\text{unclip}}(p)}{\lambda_{\text{skill}}} = p_1 \log \frac{2}{3} + p_2 \log 2,$$

which is linear in  $p$  and whose infimum is approached by concentrating all probability mass on the first token. In contrast, the direct reverse-KL loss is uniquely minimized at  $p = q$ . Therefore, the two losses cannot be identified globally.

**Corollary 1** (First-order tightness around the behavior policy). *Assume that  $p_{\theta,i}$  is twice continuously differentiable in a neighborhood of  $\theta_0$ . For  $\delta \rightarrow 0$ ,*

$$\begin{aligned} \mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta_0 + \delta) &= \lambda_{\text{skill}} \mathcal{L}_{\text{RKL}}(\theta_0 + \delta) \\ &\quad - \frac{\lambda_{\text{skill}}}{2} \delta^\top F_b \delta + o(\|\delta\|^2), \end{aligned} \tag{8}$$

where

$$\begin{aligned} F_b &\triangleq \mathbb{E}_{\substack{i \sim \nu_b \\ v \sim b_i}} [s_i(v) s_i(v)^\top], \\ s_i(v) &\triangleq \nabla_\theta \log p_{\theta,i}(v) |_{\theta=\theta_0} \end{aligned}$$

is the behavior-policy Fisher information averaged over rollout contexts.

*Proof.* By Proposition 1, the discrepancy between the scaled reverse-KL loss and the OPID skill loss is exactly  $\lambda_{\text{skill}} \mathcal{D}_b(\theta)$ . The standard local expansion of relative entropy around its reference distribution gives

$$\mathcal{D}_b(\theta_0 + \delta) = \frac{1}{2} \delta^\top F_b \delta + o(\|\delta\|^2).$$

Substituting this expansion into Eq. 4 proves Eq. 8.  $\square$

Equation 8 gives the precise sense in which the OPID skill loss is locally equivalent to reverse-KL distillation. At the behavior policy, the two losses have the same value and gradient after accounting for the factor  $\lambda_{\text{skill}}$ , while their discrepancy is second order in the policy displacement.

**Corollary 2** (Exact recovery under a matching behavior-KL penalty). *Consider the regularized auxiliary loss*

$$\mathcal{L}_{\text{aux}}(\theta) \triangleq \mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) + \beta \mathcal{D}_b(\theta). \tag{9}$$

Then

$$\mathcal{L}_{\text{aux}}(\theta) = \lambda_{\text{skill}} \mathcal{L}_{\text{RKL}}(\theta) + (\beta - \lambda_{\text{skill}}) \mathcal{D}_b(\theta). \tag{10}$$

In particular, if  $\beta = \lambda_{\text{skill}}$ , then

$$\mathcal{L}_{\text{aux}}(\theta) = \lambda_{\text{skill}} \mathcal{L}_{\text{RKL}}(\theta)$$

for every admissible  $\theta$ .

*Proof.* Substitute Eq. 4 into Eq. 9 and collect the coefficients of  $\mathcal{D}_b(\theta)$ .  $\square$

The exact cancellation in Corollary 2 requires both (i) a KL penalty to the same behavior distribution  $b_i$ , evaluated under the ordinary context, and (ii) the matching coefficient  $\beta = \lambda_{\text{skill}}$ . A KL penalty to a different reference distribution, or a different coefficient, leaves the residual behavior-relative term in Eq. 10 and is therefore not exactly equivalent to direct student–teacher reverse-KL distillation.

**Relation to the implemented PPO-clipped loss.** The decomposition in Proposition 1 applies exactly to the unclipped skill loss  $\mathcal{L}_{\text{skill}}^{\text{unclip}}$ . In the implemented OPID objective, PPO clipping is applied to the combined advantage

$$A_i^{\text{OPID}} = A_i^{\text{ep}} + \lambda_{\text{skill}} \Delta_i(a_i),$$

so the complete clipped loss does not globally decompose into independently clipped outcome and skill losses.

Nevertheless, at  $\theta_0 = \theta_{\text{old}}$ ,

$$\rho_{\theta_0,i}(a) = 1.$$

Since 1 lies in the interior of  $[1 - \epsilon, 1 + \epsilon]$  for  $\epsilon > 0$ , the clipped and unclipped policy losses have the same value and first derivative at the behavior policy:

$$\begin{aligned}\mathcal{L}_{\text{policy}}^{\text{clip}}(\theta_0) &= \mathcal{L}_{\text{policy}}^{\text{unclip}}(\theta_0), \\ \nabla_{\theta} \mathcal{L}_{\text{policy}}^{\text{clip}}(\theta) \Big|_{\theta=\theta_0} &= \nabla_{\theta} \mathcal{L}_{\text{policy}}^{\text{unclip}}(\theta) \Big|_{\theta=\theta_0} \\ &= \nabla_{\theta} \mathcal{L}_{\text{ep}}^{\text{unclip}}(\theta) \Big|_{\theta=\theta_0} + \nabla_{\theta} \mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) \Big|_{\theta=\theta_0}.\end{aligned}$$

Therefore, Eq. 6 characterizes the skill-induced component of the local PPO update. Once the policy ratio reaches a clipping boundary, however, clipping couples the outcome and skill signals through the sign of their combined advantage, and the exact relative-KL decomposition no longer applies to the complete clipped objective.

**Corollary 3** (Non-degenerate token-level signal under reward ties). *Fix one context  $i$ , and parameterize  $p_i = \text{softmax}(z_i)$  using free categorical logits. Define the corresponding full-action skill loss as*

$$\mathcal{L}_{\text{skill},i}^{\text{unclip}}(z_i) \triangleq -\lambda_{\text{skill}} \sum_{v \in \mathcal{V}} p_i(v) \Delta_i(v).$$

For  $\lambda_{\text{skill}} > 0$ ,

$$\frac{\partial \mathcal{L}_{\text{skill},i}^{\text{unclip}}}{\partial z_i(v)} = -\lambda_{\text{skill}} p_i(v) (\Delta_i(v) - \mathbb{E}_{u \sim p_i} [\Delta_i(u)]). \quad (11)$$

At  $p_i = b_i$  with full support, the gradient in Eq. 11 is zero for every  $v$  if and only if  $q_i = b_i$ .

*Proof.* Using

$$\frac{\partial p_i(u)}{\partial z_i(v)} = p_i(u) (\mathbf{1}\{u = v\} - p_i(v)),$$

we obtain

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{skill},i}^{\text{unclip}}}{\partial z_i(v)} &= -\lambda_{\text{skill}} \sum_u \Delta_i(u) p_i(u) (\mathbf{1}\{u = v\} - p_i(v)) \\ &= -\lambda_{\text{skill}} p_i(v) \Delta_i(v) + \lambda_{\text{skill}} p_i(v) \sum_u p_i(u) \Delta_i(u),\end{aligned}$$

which proves Eq. 11.

Suppose that  $p_i = b_i$ ,  $b_i(v) > 0$  for every  $v$ , and the derivative is zero for every  $v$ . Since  $\lambda_{\text{skill}} > 0$ , it follows that  $\Delta_i(v)$  is constant over the vocabulary. Hence

$$q_i(v) = e^c b_i(v)$$

for some constant  $c$ . Normalization of  $q_i$  and  $b_i$  implies  $e^c = 1$ , and therefore  $q_i = b_i$ . The converse is immediate.  $\square$

Corollary 3 is a per-context logit statement. It shows that even when group-relative outcome advantages vanish because all sampled trajectories receive tied rewards, a nontrivial skill-conditioned teacher still supplies a token-level learning signal whenever  $q_i \neq b_i$ . With shared neural parameters, gradients from different contexts may still cancel; the result does not claim that the aggregate parameter gradient must be nonzero.

### A.3 ON-POLICY OCCUPANCY MATCHING FOR DISTILLATION

Recall that  $\nu_b$  denotes the distribution over valid token positions induced by rollouts collected from the behavior policy. Let  $d_b$  denote the corresponding distribution over ordinary autoregressive contexts  $c_i$ , i.e., the context marginal induced by  $i \sim \nu_b$ . For an arbitrary data-collection policy  $\mu$ , let  $d_{\mu}$  denote the analogous context distribution.

We define total variation as

$$\text{TV}(P, Q) \triangleq \sup_A |P(A) - Q(A)| = \frac{1}{2} \int |dP - dQ|.$$

The following result isolates the effect of changing only the distribution of ordinary autoregressive contexts. It applies to both nonnegative distillation losses and signed surrogate losses.

**Proposition 2** (On-policy occupancy matching). *Let  $\ell_\theta : \mathcal{C} \rightarrow [m_\ell, M_\ell]$  be a measurable per-context loss, where  $-\infty < m_\ell < M_\ell < +\infty$ . Then*

$$\begin{aligned} & |\mathbb{E}_{c \sim d_b} [\ell_\theta(c)] - \mathbb{E}_{c \sim d_\mu} [\ell_\theta(c)]| \\ & \leq (M_\ell - m_\ell) \text{TV}(d_b, d_\mu) \\ & \leq (M_\ell - m_\ell) \sqrt{\frac{1}{2} D_{\text{KL}}(d_b \| d_\mu)}. \end{aligned} \tag{12}$$

In particular, if  $d_\mu = d_b$ , then the context-occupancy mismatch is exactly zero.

*Proof.* Define

$$f_\theta(c) \triangleq \frac{\ell_\theta(c) - m_\ell}{M_\ell - m_\ell}.$$

Then  $0 \leq f_\theta(c) \leq 1$ . By the variational characterization of total variation over measurable functions with range in  $[0, 1]$ ,

$$|\mathbb{E}_{d_b}[f_\theta] - \mathbb{E}_{d_\mu}[f_\theta]| \leq \text{TV}(d_b, d_\mu).$$

Multiplying both sides by  $M_\ell - m_\ell$  proves the first inequality in Eq. 12. The second inequality follows from Pinsker’s inequality. If  $d_b$  is not absolutely continuous with respect to  $d_\mu$ , then  $D_{\text{KL}}(d_b \| d_\mu) = +\infty$ , and the inequality remains valid in the extended-real sense. Setting  $d_\mu = d_b$  proves the final statement.  $\square$

For example, Proposition 2 can be applied to the per-context reverse-KL loss

$$\ell_{\text{RKL}, \theta}(c_i) \triangleq D_{\text{KL}}(p_{\theta, i} \| q_i),$$

which is the distribution-matching loss locally approximated by the OPID skill update. It can also be applied to a bounded version of the signed per-context OPID skill loss

$$\begin{aligned} \ell_{\text{skill}, \theta}^{\text{unclip}}(c_i) & \triangleq -\lambda_{\text{skill}} \mathbb{E}_{a \sim b_i} [\rho_{\theta, i}(a) \Delta_i(a)] \\ & = \lambda_{\text{skill}} [D_{\text{KL}}(p_{\theta, i} \| q_i) - D_{\text{KL}}(p_{\theta, i} \| b_i)]. \end{aligned} \tag{13}$$

Because the loss in Eq. 13 is signed and need not be uniformly bounded for arbitrary probability distributions, applying Proposition 2 to it requires an explicit bounded-range condition, such as probability flooring, log-ratio clipping, or restriction to a compact parameter neighborhood. More general versions can instead be obtained under appropriate moment or tail conditions.

Proposition 2 controls only the mismatch in the outer distribution of ordinary autoregressive contexts. It assumes that the same per-context loss map is evaluated under  $d_b$  and  $d_\mu$ . It does not by itself control changes in the hindsight skill, the routed teacher  $q_i$ , or other trajectory-dependent quantities that may also change with the data-collection policy.

#### A.4 CRITICAL-FIRST HIERARCHICAL ROUTING

We next formalize how the episode-level and step-level skills determine the detached teacher  $q_i$  used in  $\mathcal{L}_{\text{skill}}^{\text{unclip}}$ .

Let  $q_i^*$  denote an ideal privileged teacher at token position  $i$ . Let  $q_i^{\text{ep}}$  and  $q_i^{\text{step}}$  denote the teachers induced by the episode-level and step-level skills, respectively. Let

$$z_i^* \in \{0, 1\}$$

be an oracle criticality indicator, where  $z_i^* = 1$  means that the step-level teacher is the appropriate specialized teacher. The analyzer prediction is

$$\widehat{z}_i \triangleq \mathbf{1}\{t \in C_\tau\}.$$

The critical-first routing rule defines

$$q_i^{\text{route}} \triangleq \widehat{z}_i q_i^{\text{step}} + (1 - \widehat{z}_i) q_i^{\text{ep}}, \quad q_i \equiv q_i^{\text{route}}. \tag{14}$$

Thus, the  $q_i$  appearing in the OPID skill advantage  $\Delta_i(v) = \log q_i(v) - \log b_i(v)$  is precisely the routed teacher in Eq. 14.

Measure the approximation errors of the two candidate teachers by

$$\begin{aligned}\mathcal{E}_i^{\text{ep}} &\triangleq D_{\text{KL}}(q_i^* \| q_i^{\text{ep}}), \\ \mathcal{E}_i^{\text{step}} &\triangleq D_{\text{KL}}(q_i^* \| q_i^{\text{step}}), \\ \mathcal{E}_i^{\text{route}} &\triangleq D_{\text{KL}}(q_i^* \| q_i^{\text{route}}).\end{aligned}\tag{15}$$

Because the routing decision is hard,

$$\mathcal{E}_i^{\text{route}} = \widehat{z}_i \mathcal{E}_i^{\text{step}} + (1 - \widehat{z}_i) \mathcal{E}_i^{\text{ep}}.$$

**Proposition 3** (Routing optimality and detector-error regret). *Assume that the episode-level and step-level teachers specialize according to the oracle criticality label:*

$$\begin{aligned}z_i^* = 1 &\implies \mathcal{E}_i^{\text{step}} \leq \mathcal{E}_i^{\text{ep}}, \\ z_i^* = 0 &\implies \mathcal{E}_i^{\text{ep}} \leq \mathcal{E}_i^{\text{step}}.\end{aligned}\tag{16}$$

Then, pointwise,

$$\begin{aligned}\mathcal{E}_i^{\text{route}} &= \min \{ \mathcal{E}_i^{\text{ep}}, \mathcal{E}_i^{\text{step}} \} \\ &\quad + \mathbf{1} \{ \widehat{z}_i \neq z_i^* \} | \mathcal{E}_i^{\text{ep}} - \mathcal{E}_i^{\text{step}} |.\end{aligned}\tag{17}$$

Consequently, if

$$| \mathcal{E}_i^{\text{ep}} - \mathcal{E}_i^{\text{step}} | \leq \Gamma\tag{18}$$

almost surely under  $i \sim \nu_b$ , then

$$\begin{aligned}\mathbb{E}_{i \sim \nu_b} [ \mathcal{E}_i^{\text{route}} ] &\leq \min \{ \mathbb{E}_{i \sim \nu_b} [ \mathcal{E}_i^{\text{ep}} ], \mathbb{E}_{i \sim \nu_b} [ \mathcal{E}_i^{\text{step}} ] \} \\ &\quad + \Gamma \Pr_{i \sim \nu_b} ( \widehat{z}_i \neq z_i^* ).\end{aligned}\tag{19}$$

Under perfect criticality detection,

$$\widehat{z}_i = z_i^* \quad \text{almost surely,}$$

and therefore

$$\mathcal{E}_i^{\text{route}} = \min \{ \mathcal{E}_i^{\text{ep}}, \mathcal{E}_i^{\text{step}} \}$$

pointwise, with

$$\mathbb{E}_{i \sim \nu_b} [ \mathcal{E}_i^{\text{route}} ] \leq \min \{ \mathbb{E}_{i \sim \nu_b} [ \mathcal{E}_i^{\text{ep}} ], \mathbb{E}_{i \sim \nu_b} [ \mathcal{E}_i^{\text{step}} ] \}.$$

*Proof.* Consider first the event  $\widehat{z}_i = z_i^*$ . Under Eq. 16, the routing rule selects a teacher with the smaller approximation error. Hence

$$\mathcal{E}_i^{\text{route}} = \min \{ \mathcal{E}_i^{\text{ep}}, \mathcal{E}_i^{\text{step}} \}.$$

The second term in Eq. 17 is zero on this event.

On the event  $\widehat{z}_i \neq z_i^*$ , the routing rule selects the nonspecialized teacher. Its excess error over the oracle choice is exactly

$$| \mathcal{E}_i^{\text{ep}} - \mathcal{E}_i^{\text{step}} |.$$

This proves Eq. 17.

Taking expectations yields

$$\begin{aligned}\mathbb{E}_{i \sim \nu_b} [ \mathcal{E}_i^{\text{route}} ] &= \mathbb{E}_{i \sim \nu_b} [ \min \{ \mathcal{E}_i^{\text{ep}}, \mathcal{E}_i^{\text{step}} \} ] \\ &\quad + \mathbb{E}_{i \sim \nu_b} [ \mathbf{1} \{ \widehat{z}_i \neq z_i^* \} | \mathcal{E}_i^{\text{ep}} - \mathcal{E}_i^{\text{step}} | ].\end{aligned}$$

Using

$$\mathbb{E}[\min\{X, Y\}] \leq \min\{\mathbb{E}[X], \mathbb{E}[Y]\}$$

and Eq. 18 proves Eq. 19. The perfect-detection statements follow by setting  $\Pr_{i \sim \nu_b}(\widehat{z}_i \neq z_i^*) = 0$ .  $\square$

Proposition 3 separates the two requirements behind critical-first routing: teacher specialization and criticality-detection accuracy. Under specialization, perfect detection recovers the oracle point-wise choice between the two candidate teachers. With imperfect detection, the excess teacher-approximation error is controlled jointly by the detector error probability and the difference between the two candidate teacher errors.

The criterion in Eq. 15 measures the quality of a candidate teacher relative to  $q_i^*$ . It is distinct from the student–teacher reverse-KL loss  $D_{\text{KL}}(p_{\theta,i}||q_i)$  appearing in  $\mathcal{L}_{\text{RKL}}$ . Therefore, without additional assumptions relating the candidate teachers’ likelihood ratios, the routing result should not be interpreted as a direct upper bound on  $\mathcal{L}_{\text{RKL}}$ .

## A.5 SUMMARY

Proposition 1 analyzes the unclipped skill component of the OPID policy loss:

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) = -\lambda_{\text{skill}} \mathbb{E}_{\substack{i \sim \nu_b \\ a \sim b_i}} [\rho_{\theta,i}(a) \Delta_i(a)].$$

Conditioned on fixed rollout histories, routed skills, and detached distributions  $b_i$  and  $q_i$ , this loss has the exact decomposition

$$\mathcal{L}_{\text{skill}}^{\text{unclip}}(\theta) = \lambda_{\text{skill}} [\mathcal{L}_{\text{RKL}}(\theta) - \mathcal{D}_b(\theta)].$$

Proposition 2 shows that collecting the ordinary autoregressive contexts on policy eliminates the outer context-distribution mismatch: when the collection distribution equals the behavior-policy distribution,  $d_\mu = d_b$ , the occupancy term in Eq. 12 is zero.

Proposition 3 analyzes how the teacher  $q_i$  is selected from episode-level and step-level candidates. Under the stated specialization assumption, critical-first routing recovers the lower-error candidate under perfect detection, while the degradation under imperfect detection is controlled by

$$\Gamma \Pr_{i \sim \nu_b} (\hat{z}_i \neq z_i^*).$$

Taken together, the three results establish that:

1. The unclipped OPID skill loss is an exact relative-KL loss and is first-order equivalent to scaled reverse-KL distillation at the behavior policy;
2. On-policy collection removes the mismatch in the outer distribution of ordinary autoregressive contexts; and
3. Critical-first routing approaches the oracle candidate-teacher selection when the candidate teachers specialize and the criticality detector is accurate.

## B ADDITIONAL EXPERIMENTAL DETAILS

This section provides the experimental protocol used for the results in the main paper. We organize the details by datasets, baselines and implementation.

### B.1 DATASETS

Table 4 summarizes the datasets used in our experiments. The evaluation covers three agentic domains: embodied reasoning, web navigation, and search-augmented question answering.

**ALFWorld.** ALFWorld (Shridhar et al., 2020) aligns text-based interaction with the ALFRED household environment. Given a natural-language goal and textual observations, an agent must issue a sequence of admissible actions to complete the task. We report results on six task types: *Pick*, *Look*, *Clean*, *Heat*, *Cool*, and *Pick2*.

**WebShop.** WebShop (Yao et al., 2022) is a text-based e-commerce environment in which an agent searches for products, opens product pages, selects attributes, and purchases an item that satisfies a natural-language request. The environment provides both a normalized task-completion score, which assigns partial credit for matching requested attributes, and a binary success signal for exact task completion.

Table 4: Detailed information on the agentic benchmarks.

Domain	Benchmark	#Train Samples	#Test Samples
Embodied Reasoning	ALFWorld	2,400	140 (seen split) 134 (unseen split)
Web Navigation	WebShop	2,400	128
Search-Augmented QA	NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle	19,200	51,713

**Search-Augmented QA.** Following the Search-R1 setting (Jin et al., 2025), we evaluate search-augmented reasoning on Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023), HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023). In this setting, the agent interacts with the configured search environment before producing a final answer.

**Training Data.** For training, we conduct separate training for each benchmark setting. Specifically, we sample 2,400 training examples from ALFWorld, 2,400 training examples from WebShop, and 19,200 training examples from the search-augmented QA benchmarks.

## B.2 BASELINES

We compare OPID with prompting-only methods, outcome-based reinforcement learning, and self-distillation or skill-distillation variants. Unless explicitly marked with an asterisk, every method is evaluated from the ordinary environment interaction history, without access to skills or any other privileged context. An asterisk therefore denotes validation/test-time access to a natural-language skill; it does not indicate a different backbone or evaluation task.

### Prompting-only methods.

- *Vanilla*. This is the original instruction-tuned backbone used without any post-training. The model receives only the standard environment prompt and the interaction history exposed by the environment interface.
- *Skill-Prompt\**. This method keeps the *Vanilla* parameters frozen but augments the validation/test context with a retrieved natural-language skill relevant to the current task. Because no gradient update is performed, any improvement comes purely from in-context use of the skill.

### Outcome-based reinforcement learning.

- *GRPO* (Shao et al., 2024). Group Relative Policy Optimization is a critic-free policy-gradient method that samples a group of trajectories for each task, assigns each trajectory a scalar outcome reward, and normalizes these rewards within the group to construct relative advantages. In the outcome-only setting used here, every generated token in a trajectory inherits the same sequence-level advantage, and the policy is updated with a clipped importance-ratio objective; no process labels or teacher-derived token-level targets are used.
- *Skill-GRPO*. This variant uses the same group-relative outcome objective as *GRPO*, but makes a task-relevant natural-language skill available to the policy during training rollouts and policy updates. The skill can therefore shape exploration and the trajectories that receive reinforcement. The skill is removed at validation/test time, so this baseline tests whether skill-guided behavior has been absorbed into the model parameters rather than merely followed from the prompt.

- *Skill-GRPO\**. This method is trained in the same way as *Skill-GRPO*, but retains the skill context at validation/test time. Its train-time and test-time conditioning are consequently matched.

### Self-distillation and skill-distillation methods.

- *OPSD* (Zhao et al., 2026). On-Policy Self-Distillation instantiates a student and a teacher from the same underlying model but gives them different conditioning contexts. The student samples trajectories on-policy from the ordinary task context, whereas the teacher additionally receives training-only privileged information, such as a verified solution or an equivalent auxiliary context. For every prefix of the student’s own trajectory, the teacher re-scores the next-token distribution and provides a dense token-level target through full-vocabulary or sampled-token distribution matching. Gradients are applied to the student side while the teacher distribution is treated as a stop-gradient target, and the privileged teacher context is absent at inference time.
- *GRPO+OPSD*. This is a direct multi-objective combination of the sequence-level *GRPO* loss and the token-level *OPSD* loss. The outcome term reinforces or penalizes complete trajectories according to environment feedback, while the distillation term supplies local guidance at individual token positions. The two losses are simply combined, making this baseline a controlled test of whether naively adding dense self-distillation to outcome-based RL is sufficient.
- *Skill-SD* (Wang et al., 2026). This method adapts self-distillation to multi-turn agent tasks. Completed trajectories are summarized into compact natural-language skills that record successful behaviors, common failure modes, and reusable high-level workflows. During training, a retrieved skill conditions only the teacher branch, while the student continues to generate on-policy trajectories from the plain task prompt; the student must therefore internalize the teacher-side guidance rather than rely on the skill at test time.
- *RLSD* (Yang et al., 2026). *RLSD* uses a privileged self-teacher for fine-grained credit assignment without directly optimizing a teacher–student distribution-matching loss. It converts the token-wise teacher–student log-probability gap into a bounded weight that modulates the magnitude of each token’s *GRPO* update, while the sign and direction of the update remain anchored to the environment-derived outcome advantage. Thus, privileged information can indicate where a larger or smaller update is useful, but it does not decide whether a sampled token should be reinforced or penalized. In the original formulation, the self-distillation contribution is strongest early in training and is scheduled to decay toward vanilla *GRPO*, combining early dense guidance with a stable outcome-optimized training phase.
- *SDAR* (Lu et al., 2026a). It keeps verifier-driven *GRPO* as the primary optimization backbone and adds a separately gated self-distillation objective for multi-turn agents. A teacher branch receives training-only privileged context, such as a retrieved skill, and re-scores the student’s on-policy tokens; a smooth, bounded token-level gate then controls how strongly each teacher signal enters the auxiliary loss. The gate can use student uncertainty and/or the detached teacher–student log-probability gap, giving greater weight to positive teacher endorsements while softly attenuating potentially unreliable negative rejections. Unlike *RLSD*, *SDAR* leaves the *GRPO* advantage itself unchanged and regulates the auxiliary distillation loss instead; the student is evaluated without privileged skill context.

For all reproduced post-training baselines, we use the same backbone and environment wrappers as *OPID* and match the rollout budget, task batch, number of training steps, and evaluation protocol whenever applicable. The intended differences are restricted to the optimization signal and to the explicitly stated availability of skills or other privileged training context.

### B.3 ALGORITHM AND EXTRACTED SKILL EXAMPLES

Algorithm 1 gives the full *OPID* training procedure, including on-policy rollout collection, hierarchical skill extraction, critical-first routing, paired scoring, and clipped policy optimization. Table 5 provides representative skills extracted from successful and failed trajectories across *ALF-World*, *WebShop*, and *Search-based QA*. These examples illustrate how episode-level skills capture

---

**Algorithm 1** OPID: On-Policy Skill Distillation

---

**Require:** Policy  $\pi_\theta$ , task set  $\mathcal{Q}$ , analyzer  $\mathcal{A}$ , skill-injection function  $H$ , group size  $N$ , skill coefficient  $\lambda_{\text{skill}}$ , clipping parameter  $\epsilon$ , learning rate  $\eta$

- 1: **for** each training iteration **do**
- 2:    $\theta_{\text{old}} \leftarrow \theta$
- 3:   Sample a batch of task prompts  $\mathcal{B}$  from  $\mathcal{Q}$
- 4:   **for** each prompt  $q \in \mathcal{B}$  **do**
- 5:     *// On-policy rollout group and episode advantage*
- 6:     Sample  $\mathcal{G}_q \leftarrow \{\tau^{(1)}, \dots, \tau^{(N)}\}$ , where  $\tau^{(t)} \sim \pi_{\theta_{\text{old}}}(\cdot | q)$
- 7:      $\mathbf{r}_q \leftarrow \{R(\tau') | \tau' \in \mathcal{G}_q\}$ ;  $\mu_q \leftarrow \text{mean}(\mathbf{r}_q)$ ;  $\sigma_q \leftarrow \text{std}(\mathbf{r}_q)$
- 8:     **for** each trajectory  $\tau \in \mathcal{G}_q$  **do**
- 9:        $A_\tau^{\text{ep}} \leftarrow (R(\tau) - \mu_q) / \sigma_q$
- 10:       *// Hierarchical hindsight skill extraction*
- 11:        $(s_\tau^{\text{ep}}, \{s_{\tau,t}^{\text{step}}\}_{t \in \mathcal{C}_\tau}) \leftarrow \mathcal{A}(\tau)$
- 12:       *// Critical-first routing and paired scoring*
- 13:       **for** each interaction step  $t$  in  $\tau$  **do**
- 14:           $s_{\tau,t} \leftarrow \begin{cases} s_{\tau,t}^{\text{step}}, & t \in \mathcal{C}_\tau, \\ s_\tau^{\text{ep}}, & \text{otherwise} \end{cases}$
- 15:           $\tilde{h}_{\tau,t} \leftarrow H(h_{\tau,t}, s_{\tau,t})$
- 16:          **for** each token  $\ell$  in  $y_{\tau,t}$  with mask  $m_{\tau,t,\ell}$  **do**
- 17:            $\ell_{\tau,t,\ell}^{\text{old}} \leftarrow \log \pi_{\theta_{\text{old}}}(y_{\tau,t,\ell} | h_{\tau,t}, y_{\tau,t,<\ell})$
- 18:            $\ell_{\tau,t,\ell}^{\text{skill}} \leftarrow \log \pi_{\theta_{\text{old}}}(y_{\tau,t,\ell} | \tilde{h}_{\tau,t}, y_{\tau,t,<\ell})$
- 19:            $A_{\tau,t,\ell}^{\text{skill}} \leftarrow (\ell_{\tau,t,\ell}^{\text{skill}} - \ell_{\tau,t,\ell}^{\text{old}}) m_{\tau,t,\ell}$
- 20:            $A_{\tau,t,\ell}^{\text{ep}} \leftarrow A_\tau^{\text{ep}} m_{\tau,t,\ell}$
- 21:            $A_{\tau,t,\ell}^{\text{OPID}} \leftarrow A_{\tau,t,\ell}^{\text{ep}} + \lambda_{\text{skill}} A_{\tau,t,\ell}^{\text{skill}}$
- 22:          **end for**
- 23:       **end for**
- 24:     **end for**
- 25:   **end for**
- 26:   *// Clipped policy optimization*
- 27:   For every valid sampled token  $(\tau, t, \ell)$ , compute
- 28:    $\rho_{\tau,t,\ell}(\theta) \leftarrow \exp(\log \pi_\theta(y_{\tau,t,\ell} | h_{\tau,t}, y_{\tau,t,<\ell}) - \log \pi_{\theta_{\text{old}}}(y_{\tau,t,\ell} | h_{\tau,t}, y_{\tau,t,<\ell}))$
- 29:    $\mathcal{L}_{\text{policy}}(\theta) \leftarrow -\mathbb{E}_{\tau,t,\ell} \left[ \min \left( \rho_{\tau,t,\ell}(\theta) A_{\tau,t,\ell}^{\text{OPID}}, \text{clip}(\rho_{\tau,t,\ell}(\theta), 1 - \epsilon, 1 + \epsilon) A_{\tau,t,\ell}^{\text{OPID}} \right) \right]$
- 30:    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{policy}}(\theta)$
- 31: **end for**

---

reusable global workflows, while critical-step skills focus on sparse local decisions that influence the final outcome.

#### B.4 IMPLEMENTATION DETAILS

**Metrics.** For ALFWorld, we compute the success rate for each task type and report their macro-average:

$$\text{ALFWorld-Avg} = \frac{1}{6} \sum_{c=1}^6 \text{SR}_c. \quad (20)$$

For Search-based QA, we compute answer accuracy separately on each of the seven datasets and report the unweighted macro-average:

$$\text{Search-Avg} = \frac{1}{7} \sum_{d=1}^7 \text{Acc}_d. \quad (21)$$

For WebShop, the reported *Score* is the mean normalized task score returned by the environment, multiplied by 100, and *Succ.* is the percentage of tasks with exact success.

**Trajectory analyzer.** After each on-policy episode terminates, we serialize the task prompt, step-indexed observations, policy responses/actions, environment feedback, and terminal outcome into an ordered trajectory record. An LLM-based analyzer then maps this record to one episode-level

Table 5: **Hierarchical skills extracted from on-policy trajectories.** For each dataset, we show one successful and one failed trajectory. Episode-level skills summarize reusable global behavior, while critical-step skills target sparse decision points. Step indices are 0-based analyzer keys.

Dataset	Outcome	Task	Episode-level skill	Critical step skills
ALFWORLD	Success	clean some kettle and put it in cabinet.	Workflow: first locate and take the target object, then move to the cleaning station (sinkbasin) to clean it, then go to a suitable storage location (cabinet), open it if closed, and finally place the object inside.	<p><b>t=0</b> Go directly to the countertop or likely surface where the kettle could be.</p> <p><b>t=2</b> After acquiring the kettle, immediately go to the sinkbasin to clean it.</p> <p><b>t=4</b> After cleaning, go to a cabinet (cabinet 1) rather than pausing.</p> <p><b>t=6</b> Open the closed cabinet if needed before placing the object inside.</p>
ALFWORLD	Failure	put a clean soapbar in cart.	Avoid placing a soapbar in the cart without first confirming it is clean. The core mistake is ignoring the cleanliness requirement; the warning sign is repeatedly moving the soapbar without checking or cleaning it.	<p><b>t=1</b> Take and examine the soapbar to determine if it needs cleaning.</p> <p><b>t=2</b> If the soapbar is dirty, clean it using a sink or appropriate tool before moving to the cart.</p>
WEBSHOP	Success	Find me makeup remover for sensitive skin, nail polish with style: lagom 5 layer cotton pad, and price lower than 40.00 dollars.	Search broadly for the specific product name and key constraints, then click the first matching product result to view details, verify the attributes and price, and click 'Buy Now' to finalize.	<b>t=1</b> First, click the most relevant product result (the LAGOM cotton pad) to view its detailed page.
WEBSHOP	Failure	Find coffee tables with steel frame, storage space, brown, size with shelf, and price below \$110.	Core mistake: Buying a product without confirming it has a steel frame. Warning signs: product title lacks mention of 'steel frame'; search results include many unrelated items; product page shows color and size filters but not frame material. Avoid relying on partial matches; always verify all specific attributes, especially material, before finalizing purchase.	<p><b>t=1</b> Before clicking on a product, examine its title and check if it explicitly mentions steel frame or other required attributes.</p> <p><b>t=2</b> On the product page, click on 'Description' or 'Features' to verify the steel frame and shelf size before clicking 'Buy Now'.</p>
SEARCH	Success	Who illustrated Hunter S. Thompson's novel Fear and Loathing in Las Vegas?	Workflow: First, query using core entities (author/title) to gather context; if initial search lacks direct answer, reformulate query specifically targeting the required attribute (illustrator) and use the new results to extract the answer.	<b>t=1</b> If the initial search results do not directly answer the question, reformulate the search query to specifically target the missing attribute (here, 'illustrated by').
SEARCH	Failure	What is the full founding date of GroenLinks, the party led by Jesse Feras Klaver?	Avoid ignoring crucial temporal precision; when the task demands a specific date, search or extract the full date, not just the year, even if the year is initially prominent. Warning sign: Answering with only a year when documents contain more precise information.	<b>t=2</b> Extract the full founding date from documents about GroenLinks, not just the year.

skill and a sparse set of critical-step skills. Step indices are zero-based, consistent with Table 5. By default, we use GLM-5.2 (Z.ai, 2026) as the analyzer, with temperature set to 0.4 and maximum output length set to 4096. We limit the max number of identified critical steps at 5 for ALFWorld and WebShop, and at 2 for Search-based QA.

**Backbones and training schedule.** We use Qwen2.5-3B-Instruct and Qwen2.5-7B-Instruct (Yang et al., 2024), as well as Qwen3-1.7B-Instruct (Yang et al., 2025). All models are trained for 150 update steps. The training batch size reported in the main paper is 16 for ALFWorld and WebShop and 128 for Search-based QA. Table 6 records the remaining hyperparameters that are required for exact reproduction.

**Computing details.** Training is conducted on 8 Nvidia A800 80G GPUs.

Table 6: RL training hyperparameters.

Hyperparameter	Value
Training steps	150
Training batch size	16 for ALFWorld and WebShop; 128 for Search
Rollout group size $N$	8
Learning rate	$1 \times 10^{-6}$
PPO clip parameter $\epsilon$	0.2
Skill coefficient $\lambda_{\text{skill}}$	0.001
KL regularization coefficient	0.01
Maximum prompt length	2,048 for ALFWorld ; 4,096 for WebShop and Search
Response lengths	512
Maximum interaction steps	30 for ALFWorld, 15 for WebShop, and 4 for Search.

## C SUPPLEMENTARY RESULTS

### C.1 DETAILED SAMPLE EFFICIENCY COMPARISON

Table 7 reports the ALFWorld success rate when only a fraction of the training data is used. OPID consistently improves over GRPO across all data budgets. The gains are especially large in the low- and mid-data regimes, reaching +15.6 points with 60% of the data and +20.3 points with 80% of the data. These results suggest that trajectory-derived hindsight skills allow OPID to extract more supervision from each rollout, making outcome-based RL less dependent on large numbers of environment interactions.

Table 7: **Sample efficiency comparison on ALFWorld.** We report success rates under different fractions of the training data. The  $\Delta$  row shows the absolute improvement of OPID over GRPO, indicating that OPID provides stronger gains especially in low- and mid-data regimes.

Method	20%	40%	60%	80%	100%
GRPO	27.3	42.2	56.3	58.6	75.0
OPID	36.7	54.7	71.9	78.9	84.3
$\Delta$	+9.4	+12.5	+15.6	+20.3	+9.3

### C.2 CROSS-DOMAIN GENERALIZATION

Table 8 evaluates transfer to the ALFWorld unseen split. OPID improves the average success rate over GRPO by +7.7 points, with particularly clear gains on *Look* and *Heat*. This indicates that OPID does not merely fit the observed training trajectories. Instead, the distilled episode-level workflows and step-level decision rules retain value under unseen environment configurations.

Table 8: **Cross-domain generalization results on ALFWorld Unseen.** We report success rates across six unseen task types and their average. OPID improves the average success rate over GRPO, indicating that trajectory-derived skill supervision transfers beyond the training environments.

Method	ALFWorld Unseen						
	Pick	Look	Clean	Heat	Cool	Pick2	Avg.
ReAct	17.4	6.7	8.8	7.4	9.1	0.0	8.2
GRPO	73.9	60.0	<b>82.4</b>	59.3	72.7	<b>76.9</b>	70.9
OPID	<b>78.3</b>	<b>86.7</b>	<b>82.4</b>	<b>77.8</b>	<b>77.3</b>	69.2	<b>78.6</b>
$\Delta$	+4.4	+26.7	+0.0	+18.5	+4.6	-7.7	+7.7

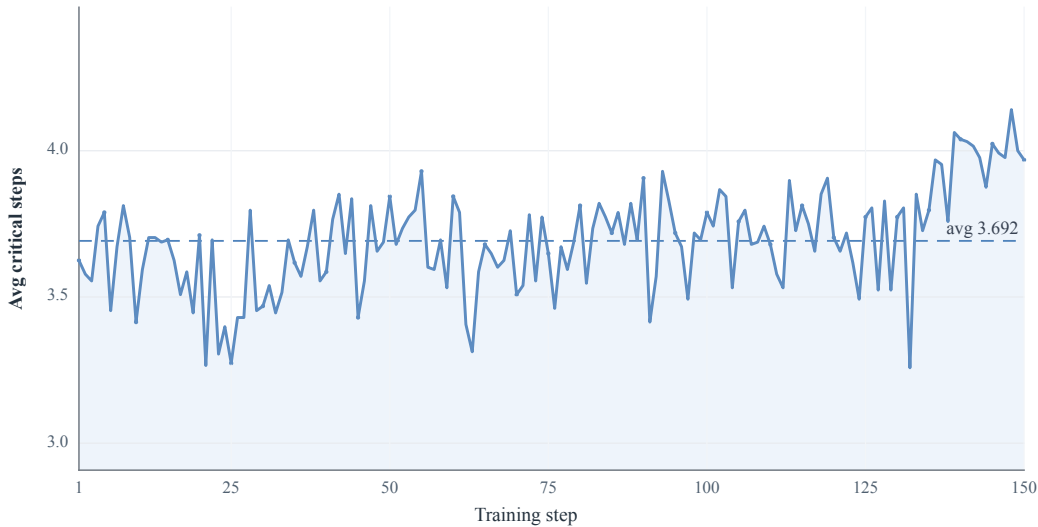


Figure 7: **Average critical steps per sequence on ALFWorld.** The curve reports how many timesteps are selected by the analyzer for step-level hindsight skills in each trajectory. The relatively small number of critical steps indicates that OPID applies local skill supervision selectively, while relying on episode-level skills as default guidance for non-critical decisions.

### C.3 TRAINING DIAGNOSTICS AND SKILL EXTRACTION PATTERNS

Figures 7–9 provide additional diagnostics for the OPID training pipeline. Figure 7 reports the average number of critical steps identified on ALFWorld, illustrating that OPID applies step-level supervision sparsely rather than assigning local skills to every decision. Figure 8 further visualizes the training advantage dynamics, complementing the main-paper training curves and showing how OPID reshapes the learning signal during policy optimization. Figure 9 shows the analyzer prompt used to convert completed trajectories into hierarchical skills.

## D CASE STUDY

Figures 10–15 provide illustrative examples from the ALFWorld, Search-QA, and WebShop benchmarks.

## E ADDITIONAL DISCUSSION

OPID studies how completed on-policy trajectories can be reused as hindsight supervision for long-horizon agentic reinforcement learning. A natural next step is to evaluate this idea in broader interactive environments where agents must discover latent rules, maintain long-term state, and adapt through extended interaction. Benchmarks such as OdysseyArena (Xu et al., 2026), Agent-Bench (Liu et al., 2023), WebArena (Zhou et al., 2023), Mind2Web (Deng et al., 2023), and VisualWebArena (Koh et al., 2024) provide complementary stress tests beyond the embodied, shopping, and search-based settings considered in this paper. These environments would test whether trajectory-derived hindsight skills remain useful when the agent must handle longer horizons, richer interfaces, and more open-ended forms of exploration.

Another direction is to enrich the structure of hindsight skills. OPID currently extracts episode-level and step-level skills from completed trajectories and routes them according to decision criticality. Future work could combine this on-policy extraction with higher-level reasoning abstractions, such as search-discovered reasoning patterns or reusable thought structures (Wu et al., 2024), and with policy-aware exploration mechanisms developed for long-horizon agent learning (Wu et al., 2026a;

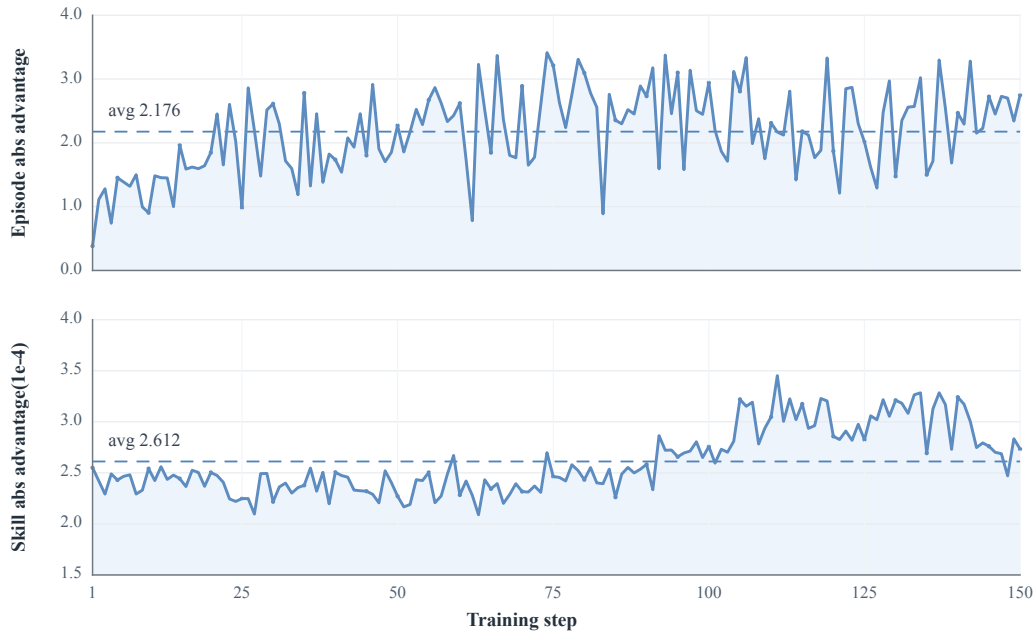


Figure 8: **Magnitudes of episode-level and skill-guided advantage signals during OPID training.** Episode abs advantage measures the mean absolute advantage from group-relative outcome rewards, while skill abs advantage measures the mean absolute advantage induced by skill-guided log-probability shifts. The comparison shows how OPID combines sparse trajectory-level feedback with dense skill-conditioned supervision throughout optimization.

Lu et al., 2026a). Such extensions may allow agents to aggregate skills across trajectories, identify recurring failure modes, and form more compositional behavioral rules while preserving OPID’s key design choice: skills are used to shape training, not retrieved as privileged context at inference time.

Finally, OPID opens several deployment-oriented directions. Since the analyzer and skill-conditioned scoring are used only during training, the learned policy incurs no additional inference-time skill retrieval cost. Nevertheless, the training pipeline can still benefit from more efficient inference and scoring mechanisms. Speculative and retrieval-parallel decoding methods such as DOUBLE (Shen et al., 2026) may reduce the cost of repeated model scoring during skill-conditioned distillation. In parallel, extending OPID to more perceptual and embodied settings, including active embodied intelligence benchmarks such as RobotEQ (Fang et al., 2026), could test whether hindsight skill supervision helps agents acquire not only task completion strategies, but also socially and spatially grounded decision rules.

### Analyser Prompt

Analyze the following agent episode and return ONLY valid JSON.

You need to complete all three fields:

1. Write a concise episode\_summary.
2. Write one episode\_skill that extracts the successful trajectory into workflow: the core decision rule and action ordering that made this trajectory work. / Write one episode\_skill that extracts the failed trajectory into avoidance rules: the core mistake and warning signs that agent should avoid.
3. Provide concise, action-oriented decision guidance for at most {max\_skill\_count} critical step(s) from the candidate set as entries in step\_skills; use the full episode to infer the guidance, but phrase each skill as advice the policy can act on at that step.

Important constraints:

- Step indexing is 0-based: step 0 is the first step of the trajectory.
- Use the task description together with the episode context to judge progress and mistakes.
- Use the full episode context to identify what each critical step should have done better.
- Each step\_skills value should be one short imperative sentence for the policy at that step.
- Write step\_skills as policy-facing guidance, not as retrospective explanation of the trajectory.
- Return only these top-level fields: episode\_summary, episode\_skill, step\_skills.
- The chosen steps are exactly the keys present in step\_skills.

Return format:

```
{
  "episode_summary": "string",
  "episode_skill": "string",
  "step_skills": {
    "0": "skill for step 0",
    "2": "skill for step 2"
  }
}
```

Episode context:

- Task description: {task\_description}
- episode\_success: {outcome\_label}
- Candidate step indices: {candidate\_step\_indices}
- Interaction trajectory: {formatted\_episode\_steps}

Figure 9: Prompt of analyzer.

## A trajectory of an agent trained with OPID — ALFWorld

**Task:** *clean some spatula and put it in diningtable*

**[Step 1]**

**Observation:** -= Welcome to TextWorld, ALFRED! -= You are in the middle of a room. Looking quickly around you, you see a cabinet 20, a cabinet 19, a cabinet 18, a cabinet 17, a cabinet 16, a cabinet 15, a cabinet 14, a cabinet 13, a cabinet 12, a cabinet 11, a cabinet 10, a cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet 2, a cabinet 1, a coffeemachine 1, a countertop 2, a countertop 1, a diningtable 2, a diningtable 1, a drawer 6, a drawer 5, a drawer 4, a drawer 3, a drawer 2, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.

**Reasoning:** I need to put two winebottle in cabinet. First, I should check where the winebottle are. Let's look around to see if I can find any winebottle. I'll start by checking the diningtables, countertops, and sinkbasin.

**Action:** go to diningtable 1

**[Step 2]**

**Observation:** You arrive at diningtable 1. On the diningtable 1, you see a butterknife 1, a cup 1, a egg 1, a fork 1, a glassbottle 2, a glassbottle 1, a lettuce 1, a mug 1, a pan 1, a pen 3, a pen 2, a pen 1, a peppershaker 1, a saltshaker 2, a saltshaker 1, a soapbottle 1, a spatula 3, a spatula 2, a spatula 1, a spoon 1, and a tomato 1.

**Reasoning:** I need to clean the spatula and put it in the diningtable. The first step is to locate the spatula. Looking at the current observation, there are several spatulas on the diningtable: spatula 1, spatula 2, and spatula 3. I should take one of these spatulas. The most straightforward action is to take spatula 1 from the diningtable.

**Action:** take spatula 1 from diningtable 1

**[Step 3]**

**Observation:** You pick up the spatula 1 from the diningtable 1.

**Reasoning:** I need to clean the spatula and put it in the diningtable. I already have the spatula 1 in my inventory. Next, I should clean it. The most straightforward way to clean a spatula is to use the sinkbasin. I will go to the sinkbasin and clean the spatula.

**Action:** go to sinkbasin 1

**[Step 4]**

**Observation:** You arrive at sinkbasin 1. On the sinkbasin 1, you see a butterknife 2, a potato 1, and a spoon 2.

**Reasoning:** I need to clean the spatula and put it in the diningtable. I already have the spatula in my inventory, and I'm at the sinkbasin. The next step is to clean the spatula using the sinkbasin. The appropriate action is to clean the spatula with the sinkbasin.

**Action:** clean spatula 1 with sinkbasin 1

**[Step 5]**

**Observation:** You clean the spatula 1 using the sinkbasin 1.

**Reasoning:** I need to put the cleaned spatula in the diningtable. I already have the spatula cleaned in the sinkbasin, so the next step is to go back to the diningtable and put the spatula there. The admissible actions include "go to diningtable 1". I will take that action.

**Action:** go to diningtable 1

**[Step 6]**

**Observation:** You arrive at diningtable 1. On the diningtable 1, you see a butterknife 1, a cup 1, a egg 1, a fork 1, a glassbottle 2, a glassbottle 1, a lettuce 1, a mug 1, a pan 1, a pen 3, a pen 2, a pen 1, a peppershaker 1, a saltshaker 2, a saltshaker 1, a soapbottle 1, a spatula 3, a spatula 2, a spoon 1, and a tomato 1.

**Reasoning:** I need to clean the spatula and put it in the diningtable. I already have the spatula in my inventory, it's clean, and I'm at the diningtable. The next step is to put the cleaned spatula in the diningtable. The admissible actions include "move spatula 1 to diningtable 1". I will take that action.

**Action:** move spatula 1 to diningtable 1

Figure 10: A full trajectory of OPID on ALFWorld Example 1.

## A trajectory of an agent trained with OPID — ALFWorld

**Task:** *put two winebottle in cabinet*

**[Step 1]**

**Observation:** -= Welcome to TextWorld, ALFRED! -= You are in the middle of a room. Looking quickly around you, you see a cabinet 20, a cabinet 19, a cabinet 18, a cabinet 17, a cabinet 16, a cabinet 15, a cabinet 14, a cabinet 13, a cabinet 12, a cabinet 11, a cabinet 10, a cabinet 9, a cabinet 8, a cabinet 7, a cabinet 6, a cabinet 5, a cabinet 4, a cabinet 3, a cabinet 2, a cabinet 1, a coffeemachine 1, a countertop 2, a countertop 1, a diningtable 2, a diningtable 1, a drawer 6, a drawer 5, a drawer 4, a drawer 3, a drawer 2, a drawer 1, a fridge 1, a garbagecan 1, a microwave 1, a sinkbasin 1, a stoveburner 4, a stoveburner 3, a stoveburner 2, a stoveburner 1, and a toaster 1.

**Reasoning:** I need to put two winebottle in cabinet. First, I should check where the winebottle are. Let's look around to see if I can find any winebottle. I'll start by checking the diningtables, countertops, and sinkbasin.

**Action:** go to diningtable 1

**[Step 2]**

**Observation:** You arrive at diningtable 1. On the diningtable 1, you see a bread 1, a butterknife 3, a cellphone 1, a lettuce 1, a spoon 2, a spraybottle 1, and a winebottle 1.

**Reasoning:** I need to put two winebottle in cabinet. I already see a winebottle 1 on diningtable 1. I should take the winebottle 1 from diningtable 1 and then go to a cabinet to put it in. The admissible actions include "take winebottle 1 from diningtable 1". Let me take the winebottle 1.

**Action:** take winebottle 1 from diningtable 1

**[Step 3]**

**Observation:** You pick up the winebottle 1 from the diningtable 1.

**Reasoning:** I need to put the winebottle in the cabinet. I already have the winebottle in my inventory, so the next step is to go to a cabinet and put the winebottle inside. The most logical next step is to go to cabinet 1.

**Action:** go to cabinet 1

**[Step 4]**

**Observation:** You arrive at cabinet 1. On the cabinet 1, you see nothing.

**Reasoning:** I need to put the winebottle in the cabinet. I already have the winebottle in my inventory, and I'm at cabinet 1. The admissible actions include "move winebottle 1 to cabinet 1". I will take that action.

**Action:** move winebottle 1 to cabinet 1

**[Step 5]**

**Observation:** You move the winebottle 1 to the cabinet 1.

**Reasoning:** I need to put the second winebottle in the cabinet. I already have the first winebottle in the cabinet, so I need to find the second winebottle and put it in the cabinet. The admissible actions include "take winebottle 1 from cabinet 1" and "move winebottle 1 to cabinet 1". Since I already moved the first winebottle to cabinet 1, I need to find the second winebottle. The admissible actions also include "go to diningtable 1" and "go to diningtable 2". I should go to diningtable 2 to check if there is a second winebottle there.

**Action:** go to diningtable 2

**[Step 6]**

**Observation:** You arrive at diningtable 2. On the diningtable 2, you see a apple 1, a bowl 2, a bread 2, a dish sponge 2, a kettle 1, a knife 1, a peppershaker 1, a plate 3, a potato 2, a soapbottle 3, a spatula 1, a spraybottle 2, and a winebottle 3.

**Reasoning:** I need to put two winebottles in a cabinet. I already have one winebottle (winebottle 1) in cabinet 1. The second winebottle is winebottle 3, which is on diningtable 2. I should take winebottle 3 from diningtable 2 and put it in a cabinet. The admissible actions include "take winebottle 3 from diningtable 2". I will take the winebottle 3 from diningtable 2.

**Action:** take winebottle 3 from diningtable 2

**[Step 7]**

**Observation:** You pick up the winebottle 3 from the diningtable 2.

**Reasoning:** I need to put the winebottle 3 in the cabinet. I already have winebottle 3 in my inventory, and I need to move it to a cabinet. The most logical next step is to go to cabinet 1 and move the winebottle 3 to it.

**Action:** go to cabinet 1

**[Step 8]**

**Observation:** You arrive at cabinet 1. On the cabinet 1, you see a winebottle 1.

**Reasoning:** I need to put the winebottle 3 in the cabinet. I already have winebottle 3 in my inventory, and I'm at cabinet 1. The admissible actions include "move winebottle 3 to cabinet 1". I will take that action.

**Action:** move winebottle 3 to cabinet 1

Figure 11: A full trajectory of OPID on ALFWorld Example 2.

## A trajectory of an agent trained with OPID — Search QA

**Task:** *Do Stanley Tucci and Michael Anderson have the same nationality?*

**[Step 1]**

**Observation:** null

**Reasoning:** null

**Action:** <search>Do Stanley Tucci and Michael Anderson have the same nationality</search>

**[Step 2]**

**Observation:** <information>{"result": "Doc 1: \"Stanley Tucci\"nawareness to the global refugee crisis. The video, titled \"What They Took With Them\", has the actors reading a poem, written by Jenifer Toksvig and inspired by primary accounts of refugees, and is part of UNHCR's #WithRefugees campaign, of which also includes a petition to governments to expand asylum to provide further shelter, integrating job opportunities, and education. Stanley Tucci Stanley Tucci (; born November 11, 1960) is an American actor, writer, producer, and film director. He has won three Emmy Awards; two for his performances in \"Winchell\" and \"Monk\", and one as a producer of \"Park Bench with Steve\"nDoc 2: \"Stanley Tucci\"nfilm. It won him and Tropiano the Independent Spirit Award for Best First Screenplay. He has been nominated three times for Golden Globes, and won twice u2013 for his title role in \"Winchell\" (1998), and for his supporting role as Adolf Eichmann in \"Conspiracy\" (2001), both for HBO films. He also received a Screen Actors Guild Award nomination for \"Winchell\". He was nominated for Broadway's Tony Award as Best Actor in a Play for his role as Johnny in the 2002 revival of Terrence McNally's \"Frankie and Johnny in the Clair de Lune\". In 2004, Caedmon Audio released an audiobooknDoc 3: \"Stanley Tucci\"ninroduced the couple several years later at her own wedding. Tucci and Blunt married in a civil ceremony in summer 2012, followed by a larger observance at Middle Temple Hall in London on September 29, 2012. The couple live in Barnes, London and have a son, Matteo Oliver (b. January 2015) and a daughter, Emilia Giovanna (b. April 2018). On September 12, 2016, Tucci, as well as Cate Blanchett, Chiwetel Ejiofor, Peter Capaldi, Douglas Booth, Neil Gaiman, Keira Knightley, Juliet Stevenson, Kit Harington, and Jesse Eisenberg, featured in a video from the United Nations' refugee agency UNHCR to help raise\"n\"}</information>

**Reasoning:** Based on the provided information, Stanley Tucci is described as an American actor, writer, producer, and film director. However, there is no direct information provided about Michael Anderson's nationality. To determine if they have the same nationality, we would need to know Michael Anderson's nationality.

**Action:** <search>What is Michael Anderson's nationality</search>

**[Step 3]**

**Observation:** <information>{"result": "Doc 1: \"Matt Anderson (volleyball)\"nMatt Anderson (volleyball) Matthew John Anderson (born 18 April 1987) is an American volleyball player, a member of the United States men's national volleyball team and Russian club Zenit Kazan, a participant of the Olympic Games (2012, 2016), bronze medalist of the 2016 Olympic Games, 2013 NORCECA Champion, a gold medalist of the 2008 Pan American Cup, 2014 World League and 2015 World Cup. Anderson was born in Buffalo, New York to Michael and Nancy Anderson. His father, Michael Anderson, died in January 2010. He has three older sisters named Jennifer, Joelle and Amy and one older brother named Joshua.nDoc 2: \"Michael Anderson (swimmer)\"nMichael Anderson (swimmer) Michael Anderson, (born 12 July 1987) is an Australian Paralympic swimmer who has won gold, silver and bronze medals at the three Paralympics from 2008 to 2016. Anderson has nerve damage to his leg caused by meningitis and a hearing impairment. He was born and grew up in Bellingen in the Northern Rivers. He moved to the Gold Coast to study for a Bachelor of Sport Management degree at Griffith University. Anderson made his international debut at the 2005 Deaflympics in Melbourne where he finished sixth in the 50m backstroke. At the 2006 IPC Swimming World ChampionshipsnDoc 3: \"Michael Anderson Pereira da Silva\"ndebut in a group stage game against Italian team Roma. Most of the autumn and winter 2007 Michael was injured and was undergoing treatment in Brazil, however he has recovered and returned since. On 13 May 2009 the 26-year-old attacking midfield player, changes on loan until 31 December 2009 from Dynamo Kyiv to Botafogo. Michael moved from Kyiv to Portuguesa on 23 May 2011. \"according to combined sources on the Flamengo official website and Flaestat00edstica.\"n Michael Anderson Pereira da Silva Michael Anderson Pereira da Silva (born February 16, 1983 in Slu00e3o Caetano do Sul) better known as simply Michael, is\"n\"}</information>

**Reasoning:** Based on the provided information, Stanley Tucci is described as an American actor, writer, producer, and film director. Michael Anderson is identified as an Australian Paralympic swimmer. Therefore, they do not share the same nationality.

**Action:** <answer>No</answer>

Figure 12: A full trajectory of OPID on Search-QA Example 1.

## A trajectory of an agent trained with OPID — Search QA

**Task:** *What was the name of the best selling book about serial killers that was written by a former chief reporter of national red top newspaper published in the United Kingdom?*

**[Step 1]**

**Observation:** null

**Reasoning:** null

**Action:** <search>What was the name of the best selling book about serial killers that was written by a former chief reporter of national red top newspaper published in the United Kingdom</search>

**[Step 2]**

**Observation:** <information>{"result": "Doc 1: \"Criminal Shadows: Inside the Mind of the Serial Killer\"\n\nCriminal Shadows: Inside the Mind of the Serial Killer Criminal Shadows: Inside the Mind of the Serial Killer is a book written by English professor of psychology, David V. Canter. It was the winner of two literary awards: the Gold Dagger for Non-Fiction (1994) and the Anthony Award for Best True Crime (1995). A unique insight into some of Britain\u2019s most notorious criminal investigations, this is the account of Professor Canter's experience as Britain\u2019s leading pioneer in the new science of criminal psychological profiling. Canter takes the reader step by step through his development of profiling and through many of\n\nDoc 2: \"Serial killer\"\n\nwealth of London, focused the media's attention on the plight of the urban poor and gained coverage worldwide. Jack the Ripper has also been called the most famous serial killer of all time, and his legend has spawned hundreds of theories on his real identity and multiple works of fiction. H. H. Holmes was one of the first documented modern serial killers in the United States, responsible for the death of at least nine victims in the early 1890s. Here as well, the case gained notoriety and wide publicity through possibly sensationalized accounts in William Randolph Hearst's newspapers. At the\n\nDoc 3: \"My Life Among the Serial Killers\"\n\nbook. My Life Among the Serial Killers My Life Among the Serial Killers: Inside the Minds of the World's Most Notorious Murderers is a book written by Helen Morrison, M.D and Harold Goldberg. It presents the cases of ten serial killers, and touches on many more. Morrison spent hundreds of hours in face-to-face interviews, over many years, with several of the subjects. She uses the individual stories to explain and put forth her ideas on what makes a serial killer. Morrison begins her story in the early '70s, when psychiatry and law enforcement had a very different view of these\n\n"}</information>

**Reasoning:** Based on the provided information, the question is asking for the name of the best-selling book about serial killers written by a former chief reporter of a national red top newspaper in the United Kingdom. The search results provided do not explicitly mention a book that meets all the criteria: 1. It is a best-selling book about serial killers. 2. It was written by a former chief reporter of a national red top newspaper in the United Kingdom. Since the information provided is insufficient to confirm the exact book name, I need to perform a search to find the correct answer.

**Action:** <search>Best-selling book about serial killers written by a former chief reporter of a national red top newspaper published in the United Kingdom</search>

**[Step 3]**

**Observation:** <information>{"result": "Doc 1: \"Serial killer\"\n\nwealth of London, focused the media's attention on the plight of the urban poor and gained coverage worldwide. Jack the Ripper has also been called the most famous serial killer of all time, and his legend has spawned hundreds of theories on his real identity and multiple works of fiction. H. H. Holmes was one of the first documented modern serial killers in the United States, responsible for the death of at least nine victims in the early 1890s. Here as well, the case gained notoriety and wide publicity through possibly sensationalized accounts in William Randolph Hearst's newspapers. At the\n\nDoc 2: \"Criminal Shadows: Inside the Mind of the Serial Killer\"\n\nCriminal Shadows: Inside the Mind of the Serial Killer Criminal Shadows: Inside the Mind of the Serial Killer is a book written by English professor of psychology, David V. Canter. It was the winner of two literary awards: the Gold Dagger for Non-Fiction (1994) and the Anthony Award for Best True Crime (1995). A unique insight into some of Britain\u2019s most notorious criminal investigations, this is the account of Professor Canter's experience as Britain\u2019s leading pioneer in the new science of criminal psychological profiling. Canter takes the reader step by step through his development of profiling and through many of\n\nDoc 3: \"James Weatherup\"\n\nEuropean News Agency and sports editor of the London and Surrey Times newspaper. He was head-hunted by a major PR firm in Soho and is now an account director at Keith Bishop Associates in London's West End. James Weatherup After starting out in regional newspapers, he joined the \"\"News of the World,\"\" serving in two stints over 25 years for nine editors. In his first stint he rose to Chief reporter. During this time he co-wrote the best selling book \"\"Inside 25 Cromwell Street\"\" about the serial killers Fred and Rosemary West. He then left for the \"\"Sunday People\"\" in 1999,\n\n"}</information>

**Reasoning:** Based on the provided information, the best-selling book about serial killers written by a former chief reporter of a national red top newspaper published in the United Kingdom is likely "Inside 25 Cromwell Street" by James Weatherup. This book is mentioned as co-written with Fred and Rosemary West, and James Weatherup was a former chief reporter for the London and Surrey Times newspaper.

**Action:** <answer>Inside 25 Cromwell Street</answer>

Figure 13: A full trajectory of OPID on Search-QA Example 2.

## A trajectory of an agent trained with OPID — WebShop

**Task:** Find me loose fit, day comfort, hand wash women's tops, tees & blouses with short sleeve, polyester spandex for teen girls with color: a7-blue, and size: medium, and price lower than 40.00 dollars

[Step 1]

**Observation:** 'Search'.

**Reasoning:** I need to find a loose fit, day comfort, hand wash women's tops, tees & blouses with short sleeve, polyester spandex for teen girls with color: a7-blue, and size: medium, with a price lower than 40.00 dollars. The current observation is "Search", so I should perform a search with the appropriate query. The best action is to search for the specific product with the required color, size, and price.

**Action:** search[a7-blue, medium, polyester spandex, teen girls, color: a7-blue, size: medium, price: <40]

[Step 2]

**Observation:** 'Back to Search' [SEP] 'Page 1 (Total results: 50)' [SEP] 'Next >' [SEP] 'B09NPML43M' [SEP] 'Womens Short Sleeve Tops, Womens Casual Dandelion Printing T-Shirts Loose O-Neck Blouse Tops Funny Graphic Tee Shirts' [SEP] '\$1.01 to \$1.74' [SEP] 'B09KLQLLT2' [SEP] 'Long Sleeve Superhero T Shirt Tank Top Mens Compression Shirt Men Workout Fitness Gym Shirt' [SEP] '\$19.99' [SEP] 'B09RVF4JP1' [SEP] 'Once Upon A Time Queen Born In 1982 T Shirts It Was Me Tank Top' [SEP] '\$19.99' [SEP] 'B08ZRR3DZT' [SEP] 'Sheln Women's Sexy One Shoulder Sleeveless Cutout Mini Club Pencil Bodycon Dress' [SEP] '\$19.99 to \$27.99' [SEP] 'B07JVVDJ6L' [SEP] 'Canadian Tuxedo T Shirt - Novelty Denim Tux T-Shirt' [SEP] '\$19.99' [SEP] 'B09R9YCM6R' [SEP] 'XXBR Summer T-shirts for Mens, Soldier Short Sleeve 3D Street Vintage Printed Shirt Slim Fit Muscle Casual Tee Tops' [SEP] '\$8.98 to \$11.99' [SEP] 'B01HQTWL6S' [SEP] 'Expression Tees Pug Life Funny Thug Life Unisex Adult Hoodie' [SEP] '\$36.99 to \$48.99' [SEP] 'B09M63B87V' [SEP] 'YALFJV Women Long Sleeve Crew Neck Side Button T Shirts Tunic Dress Loose Asymmetric Hem Tunic Pullover to Wear with Leggings' [SEP] '\$10.71 to \$18.34' [SEP] 'B09GLVMLMS' [SEP] 'Anniou Electric Heated Jacket Adjustable Temp USB Heated Coat Winter Hooded Jacket Men Women' [SEP] '\$64.99 to \$114.99' [SEP] 'B09PL5W9PD' [SEP] 'Women's One Piece Swimsuit Halter Plunge Neck Tummy Control Bathing Suits Push Up Tankini Sets Plus Size Beachwear' [SEP] '\$7.8'.

**Reasoning:** Given the current observation, I need to find a loose fit, day comfort, hand wash women's tops, tees & blouses with short sleeve, polyester spandex for teen girls with color: a7-blue, and size: medium, with a price lower than 40.00 dollars. The current observation shows that the first item on page 1 is "B09NPML43M" with the title "Womens Short Sleeve Tops, Womens Casual Dandelion Printing T-Shirts Loose O-Neck Blouse Tops Funny Graphic Tee Shirts" and a price range of \$1.01 to \$1.74. Since this item meets all the criteria (color: a7-blue, size: medium, material: polyester spandex, and is a short sleeve top), it is a suitable match for the shopping goal. The best action to take is to click on this item to view it in detail.

**Action:** click[b09nqml43m]

[Step 3]

**Observation:** 'Back to Search' [SEP] '< Prev' [SEP] 'color' [SEP] 'a1-black' [SEP] 'a1-blue' [SEP] 'a1-gray' [SEP] 'a1-green' [SEP] 'a1-purple' [SEP] 'a1-wine red' [SEP] 'a1-yellow' [SEP] 'a2-blue' [SEP] 'a2-gray' [SEP] 'a2-green' [SEP] 'a2-white' [SEP] 'a2-yellow' [SEP] 'a3-army green' [SEP] 'a3-blue' [SEP] 'a3-gray' [SEP] 'a3-navy' [SEP] 'a3-orange' [SEP] 'a3-wine red' [SEP] 'a4-gray' [SEP] 'a6-black' [SEP] 'a6-blue' [SEP] 'a6-red' [SEP] 'a7-black' [SEP] 'a7-blue' [SEP] 'a7-red' [SEP] 'b15-black' [SEP] 'b15-blue' [SEP] 'b15-red' [SEP] 'b16-black' [SEP] 'b16-blue' [SEP] 'b16-red' [SEP] 'b17-army green' [SEP] 'b17-black' [SEP] 'b17-red' [SEP] 'b17-wine red' [SEP] 'b18-a' [SEP] 'b11-black' [SEP] 'b13-a' [SEP] 'b19-a' [SEP] 'b14-a' [SEP] 'b21-a' [SEP] 'b12-a' [SEP] 'size' [SEP] 'small' [SEP] 'medium' [SEP] 'large' [SEP] 'x-large' [SEP] 'xx-large' [SEP] '3x-large' [SEP] 'Womens Short Sleeve Tops, Womens Casual Dandelion Printing T-Shirts Loose O-Neck Blouse Tops Funny Graphic Tee Shirts' [SEP] 'Price: \$1.01 to \$1.74' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now'.

**Reasoning:** The current observation shows that we are back on Page 1 of the search results, and the options include selecting the color as "a7-blue" and size as "medium". The goal is to find a women's top with color: a7-blue, size: medium, and price lower than 40.00 dollars. We have already clicked on "a7-blue" and are now on the size selection. The next logical step is to click on the size "medium" to continue searching for the desired product.

**Action:** click[medium]

[Step 4]

**Observation:** 'Back to Search' [SEP] '< Prev' [SEP] 'color' [SEP] 'a1-black' [SEP] 'a1-blue' [SEP] 'a1-gray' [SEP] 'a1-green' [SEP] 'a1-purple' [SEP] 'a1-wine red' [SEP] 'a1-yellow' [SEP] 'a2-blue' [SEP] 'a2-gray' [SEP] 'a2-green' [SEP] 'a2-white' [SEP] 'a2-yellow' [SEP] 'a3-army green' [SEP] 'a3-blue' [SEP] 'a3-gray' [SEP] 'a3-navy' [SEP] 'a3-orange' [SEP] 'a3-wine red' [SEP] 'a4-gray' [SEP] 'a6-black' [SEP] 'a6-blue' [SEP] 'a6-red' [SEP] 'a7-black' [SEP] 'a7-blue' [SEP] 'a7-red' [SEP] 'b15-black' [SEP] 'b15-blue' [SEP] 'b15-red' [SEP] 'b16-black' [SEP] 'b16-blue' [SEP] 'b16-red' [SEP] 'b17-army green' [SEP] 'b17-black' [SEP] 'b17-red' [SEP] 'b17-wine red' [SEP] 'b18-a' [SEP] 'b11-black' [SEP] 'b13-a' [SEP] 'b19-a' [SEP] 'b14-a' [SEP] 'b21-a' [SEP] 'b12-a' [SEP] 'size' [SEP] 'small' [SEP] 'medium' [SEP] 'large' [SEP] 'x-large' [SEP] 'xx-large' [SEP] '3x-large' [SEP] 'Womens Short Sleeve Tops, Womens Casual Dandelion Printing T-Shirts Loose O-Neck Blouse Tops Funny Graphic Tee Shirts' [SEP] 'Price: \$1.01 to \$1.74' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now'.

**Reasoning:** Looking at the current observation, the agent is on the "Back to Search" page, and the goal is to find a women's top with color: a7-blue, size: medium, and price lower than 40.00. The agent has already clicked on "medium" and is now on the color selection. The most relevant product that matches the criteria is "B09NPML43M" with color "a7-blue" and size "medium". The agent should click on "a7-blue" to select the color and then click on "medium" to select the size.

**Action:** click[a7-blue]

[Step 5]

**Observation:** 'Back to Search' [SEP] '< Prev' [SEP] 'color' [SEP] 'a1-black' [SEP] 'a1-blue' [SEP] 'a1-gray' [SEP] 'a1-green' [SEP] 'a1-purple' [SEP] 'a1-wine red' [SEP] 'a1-yellow' [SEP] 'a2-blue' [SEP] 'a2-gray' [SEP] 'a2-green' [SEP] 'a2-white' [SEP] 'a2-yellow' [SEP] 'a3-army green' [SEP] 'a3-blue' [SEP] 'a3-gray' [SEP] 'a3-navy' [SEP] 'a3-orange' [SEP] 'a3-wine red' [SEP] 'a4-gray' [SEP] 'a6-black' [SEP] 'a6-blue' [SEP] 'a6-red' [SEP] 'a7-black' [SEP] 'a7-blue' [SEP] 'a7-red' [SEP] 'b15-black' [SEP] 'b15-blue' [SEP] 'b15-red' [SEP] 'b16-black' [SEP] 'b16-blue' [SEP] 'b16-red' [SEP] 'b17-army green' [SEP] 'b17-black' [SEP] 'b17-red' [SEP] 'b17-wine red' [SEP] 'b18-a' [SEP] 'b11-black' [SEP] 'b13-a' [SEP] 'b19-a' [SEP] 'b14-a' [SEP] 'b21-a' [SEP] 'b12-a' [SEP] 'size' [SEP] 'small' [SEP] 'medium' [SEP] 'large' [SEP] 'x-large' [SEP] 'xx-large' [SEP] '3x-large' [SEP] 'Womens Short Sleeve Tops, Womens Casual Dandelion Printing T-Shirts Loose O-Neck Blouse Tops Funny Graphic Tee Shirts' [SEP] 'Price: \$1.01 to \$1.74' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now'.

**Reasoning:** Given the current observation, the agent is on the "Back to Search" page, and the goal is to find a women's top with color a7-blue, size medium, and price lower than 40.00 dollars. The agent has already clicked on "medium" and "a7-blue", but the product is still displayed as the same item. Since the product is already shown as a7-blue and medium, the best action is to click on "buy now" to complete the purchase.

**Action:** click[buy now]

Figure 14: A full trajectory of OPID on Webshop Example 1.

## A trajectory of an agent trained with OPID — WebShop

**Task:** Find me men's sleep & lounge with long sleeve, elastic waistband for daily wear with color: multi 4, and size: large, and price lower than 60.00 dollars

### [Step 1]

**Observation:** 'Search'.

**Reasoning:** I need to find men's sleep & lounge with long sleeve, elastic waistband for daily wear, color: multi 4, size: large, and price lower than 60.00 dollars. The current observation is 'Search', so I should perform a search with the appropriate query. The best query to use is: `***men's sleep & lounge long sleeve elastic waistband daily wear color multi 4 size large price: <60***` This query includes all the required criteria: - Men's sleep & lounge - Long sleeve - Elastic waistband - Daily wear - Color: multi 4 - Size: large - Price: lower than 60.00 I will now perform the search with this query.

**Action:** search[mens sleep & lounge long sleeve elastic waistband daily wear color multi 4 size large price: <60]

### [Step 2]

**Observation:** 'Back to Search' [SEP] 'Page 1 (Total results: 50)' [SEP] 'Next >' [SEP] 'B09ND8P2QR' [SEP] 'InterestPrint Deer and Stars Men's Loungewear Pajama Sets - Long Sleeve Tee and Jogger Pant' [SEP] '\$43.59' [SEP] 'B09ND9DP7J' [SEP] 'InterestPrint Gold Horse Pattern Men's 2-Piece Sleepwear Set, Long Sleeve Shirt with Pants Loungewear' [SEP] '\$43.59' [SEP] 'B09Q5ZHRVM' [SEP] 'Yinimo Mens Gym Shorts Fashionable Loose Beach Drawstring Turkey Print Beach Cool Shorts' [SEP] '\$19.19' [SEP] 'B09QW2HQRK' [SEP] 'CandyMan Lace Pajama and Lounge Pants for Men' [SEP] '\$31.0 to \$69.0' [SEP] 'B09S3BN15C' [SEP] 'Mens Linen Shirt, Men's Striped Shirts Casual Short Sleeve Button Down Shirts Regular Fit Hawaiian Shirts Beach Tees Tops' [SEP] '\$3.78 to \$11.38' [SEP] 'B09HGFV91W' [SEP] 'ZHDD Tops for Mens, Men's Knight Punk Vintage Style Long Sleeve Asymmetric Hem Cow Neck Teen Boys Gothic Pullover' [SEP] '\$4.98 to \$7.99' [SEP] 'B09HX5CD2D' [SEP] 'CSU Cleveland State University Vikings Property Fleece Drawstring Shorts Heather Charcoal' [SEP] '\$39.95' [SEP] 'B09R9YCM6R' [SEP] 'XXBR Summer T-shirts for Mens, Soldier Short Sleeve 3D Street Vintage Printed Shirt Slim Fit Muscle Casual Tee Tops' [SEP] '\$8.98 to \$11.99' [SEP] 'B09M63B87V' [SEP] 'YALFJV Women Long Sleeve Crew Neck Side Button T Shirts Tunic Dress Loose Asymmetric Hem Tunic Pullover to Wear with Leggings' [SEP] '\$10.71 to \$18.34' [SEP] 'B09QQP3356' [SEP] 'HAUKLIE Men's Sports Waffle Ribbed Polo Shirts Summer Short Sleeve Cotton Muscle Quarter-Zip Henley T-Shirt Tunics Tops' [SEP] '\$10.99'.

**Reasoning:** Given the current observation, I need to find a men's sleep & lounge with long sleeve, elastic waistband for daily wear, color: multi 4, size: large, and price lower than 60.00 dollars. From the observations, the relevant product that matches the criteria is: - `**B09ND8P2QR**`: "InterestPrint Deer and Stars Men's Loungewear Pajama Sets - Long Sleeve Tee and Jogger Pant" with a price of `**$43.59**`. This item meets all the criteria: - It is a men's sleep & lounge. - It has a long sleeve. - It has an elastic waistband. - It is color: multi 4. - It is size: large. - It is priced under `**$60.00**`. Therefore, the best action is to `**click on this product**` to view it in detail.

**Action:** click[b09nd8p2qr]

### [Step 3]

**Observation:** 'Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP] 'small' [SEP] 'medium' [SEP] 'large' [SEP] 'x-large' [SEP] 'xx-large' [SEP] 'color' [SEP] 'multi 1' [SEP] 'multi 10' [SEP] 'multi 2' [SEP] 'multi 3' [SEP] 'multi 4' [SEP] 'multi 5' [SEP] 'multi 6' [SEP] 'multi 7' [SEP] 'multi 8' [SEP] 'multi 9' [SEP] 'InterestPrint Deer and Stars Men's Loungewear Pajama Sets - Long Sleeve Tee and Jogger Pant' [SEP] 'Price: \$43.59' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now'.

**Reasoning:** The current observation shows that we are back on Page 1 of the search results, and the options include selecting the size as "large" and color as "multi 4". The goal is to find a men's sleep & lounge with long sleeve, elastic waistband, color: multi 4, size: large, and price lower than 60.00 dollars. We have already clicked on "B09ND8P2QR" (the product "InterestPrint Deer and Stars Men's Loungewear Pajama Sets - Long Sleeve Tee and Jogger Pant") and are now looking to select the size and color. The most appropriate action is to click on the size "large" to proceed with the search.

**Action:** click[large]

### [Step 4]

**Observation:** 'Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP] 'small' [SEP] 'medium' [SEP] 'large' [SEP] 'x-large' [SEP] 'xx-large' [SEP] 'color' [SEP] 'multi 1' [SEP] 'multi 10' [SEP] 'multi 2' [SEP] 'multi 3' [SEP] 'multi 4' [SEP] 'multi 5' [SEP] 'multi 6' [SEP] 'multi 7' [SEP] 'multi 8' [SEP] 'multi 9' [SEP] 'InterestPrint Deer and Stars Men's Loungewear Pajama Sets - Long Sleeve Tee and Jogger Pant' [SEP] 'Price: \$43.59' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now'.

**Reasoning:** The current observation shows that we are back to the search page, and the options include selecting the size as "large" and color as "multi 4". The goal is to find a men's sleep & lounge with long sleeve, elastic waistband, color: multi 4, and size: large, with a price lower than \$60.00. We have already clicked on "large" and are now looking for the color "multi 4". The best action is to click on "multi 4" to select the color.

**Action:** click[multi 4]

### [Step 5]

**Observation:** 'Back to Search' [SEP] '< Prev' [SEP] 'size' [SEP] 'small' [SEP] 'medium' [SEP] 'large' [SEP] 'x-large' [SEP] 'xx-large' [SEP] 'color' [SEP] 'multi 1' [SEP] 'multi 10' [SEP] 'multi 2' [SEP] 'multi 3' [SEP] 'multi 4' [SEP] 'multi 5' [SEP] 'multi 6' [SEP] 'multi 7' [SEP] 'multi 8' [SEP] 'multi 9' [SEP] 'InterestPrint Deer and Stars Men's Loungewear Pajama Sets - Long Sleeve Tee and Jogger Pant' [SEP] 'Price: \$43.59' [SEP] 'Rating: N.A.' [SEP] 'Description' [SEP] 'Features' [SEP] 'Reviews' [SEP] 'Buy Now'.

**Reasoning:** Given the current observation, the agent is on the "Back to Search" page, and the goal is to find a men's sleep & lounge with long sleeve, elastic waistband for daily wear with color: multi 4 and size: large, with a price lower than \$60. The agent has already clicked on "large" and "multi 4", but the product is still shown as "InterestPrint Deer and Stars Men's Loungewear Pajama Sets - Long Sleeve Tee and Jogger Pant" with a price of \$43.59. Since the product is already matching the size and color, the best action is to click on "buy now" to complete the purchase.

**Action:** click[buy now]

Figure 15: A full trajectory of OPID on Webshop Example 2.