Skill0.5: Joint Skill Internalization and Utilization for Out-of-Distribution Generalization in Agentic Reinforcement Learning

Anonymous ACL submission

Abstract

Equipping large language models with explicit skills has emerged as a promising paradigm for enabling autonomous agents to solve complex tasks. Agent skills can be inherently divided into general skills for broad cognitive transfer and task-specific skills for dynamic execution. However, existing skill-based reinforcement learning (RL) methods typically force a rigid choice between full externalization, which incurs prohibitive context overhead, and full internalization, which risks overfitting and knowledge conflicts. To address this dilemma, we propose Skill0.5 a novel agentic RL framework that explicitly differentiates skill treatments by combining general skill internalization with task-specific skill utilization. Driven by a dynamic, difficulty-aware router, Skill0.5 streams tasks into distinct mastery tiers to apply tailored optimization strategies: it internalizes general skills via privileged distillation to build a cognitive foundation for hard tasks, while using diagnostic probing on easy tasks to penalize shortcuts and enforce specific skill utilization. Experiments on ALFWorld and WebShop demonstrate that Skill0.5 outperforms both memory-based and skill-based RL baselines, yielding performance improvements across both in-distribution and out-of-distribution scenarios. The code is available at: `https://anonymous.4open.science/r/Skill0_5-1234/`.

1 Introduction

As Large Language Models (LLMs) (Zeng et al., 2026; Singh et al., 2025; Team et al., 2026) evolve into autonomous problem solvers, they are increasingly entrusted with challenging agentic tasks (Guan et al., 2026; Du et al., 2025; Ding et al., 2026). To enable agents to master the complex operational logic of real-world tasks, *agent skills* have emerged as a promising solution to break through performance bottlenecks (Xu and Yan, 2026; Ling et al., 2026). A skill encapsulates procedural knowledge into modular, reusable textual directives that codify standard operating procedures and heuristics (Li et al., 2026a). In practice, these skills are dynamically retrieved and injected into the agent's prompt to explicitly guide it through intricate workflows (Zhou et al., 2026b).

While introducing skills via zero-shot prompting offers immediate utility, to further empower agents to robustly navigate complex environments (Xia et al., 2026b), recent research has expanded into skill-based training methods. These methods primarily diverge into two extreme paradigms. One paradigm advocates for *full externalization*, where all skills are maintained as external contextual guidance throughout training and inference (Xia et al., 2026a; Shi et al., 2026). Conversely, another line of research explores *full internalization*, aiming to completely assimilate the skills into model parameters (Lu et al., 2026b; Wang et al., 2026b). In authentic deployment scenarios, however, skill libraries expand dynamically through user contributions, frequently confronting agents with unfamiliar tasks alongside unseen Out-of-Distribution (OOD) task-specific skills (Ma et al., 2026b).

Consequently, both paradigms exhibit notable limitations: *full externalization* imposes severe challenges on LLMs' In-Context Learning (ICL) capabilities (Liu et al., 2024; Zhou et al., 2026a); as the prompt expands with numerous skills, the excessive length can degrade reasoning and instruction-following performance, especially in long-horizon tasks (Si et al., 2026; Hsieh et al., 2024). On the other hand, *full internalization* is fundamentally constrained by model capacity (Allen-Zhu and Li, 2025) and potentially introduces knowledge conflict risks (Xu et al., 2024; Wang et al., 2025a). Agents may fail to absorb and utilize new instructions when these unfamiliar external skills contradict their internalized skill patterns, leading to execution hallucinations (Liu et al., 2026b;

Zhang et al., 2026a). Therefore, the efficacy of existing skill-based training approaches in dynamic, real-world environments remains underexplored.

Fundamentally, agentic skills fall into two complementary categories: general and task-specific (Xia et al., 2026a; Li et al., 2026b,c). General skills (e.g., meta-reasoning and error recovery) are domain-agnostic but contextually lengthy (Zhou et al., 2026a). Conversely, task-specific skills encode granular execution rules that are dynamically updated and susceptible to retrieval noise (Lu et al., 2026b). However, existing methods treat these categories uniformly, creating a critical dilemma: *fully externalizing* lengthy general heuristics incurs prohibitive context overhead (Wu and Zhang, 2026), while *fully internalizing* volatile specific skills risks severe overfitting and knowledge conflicts (Ma et al., 2026a; Alzubi et al., 2026). To resolve this, we advocate differentiated treatments: **internalizing general skills to establish a context-efficient cognitive foundation, while dynamically utilizing plug-and-play task-specific skills to enhance adaptability**, especially in skill-augmented OOD scenarios.

Intuitively, an agent must grasp foundational strategies before exploiting fine-grained rules. Motivated by this, we propose **Skill0.5**, a unified agentic RL framework that jointly optimizes decoupled general and specific skills based on real-time task mastery. Specifically, a *difficulty-aware router* streams tasks into three tiers for tailored optimization: hard tasks internalize general skills via privileged distillation; medium tasks undergo standard RL to maximize success; and easy tasks employ diagnostic probing to enforce faithful specific skill utilization. Evaluations show Skill0.5 outperforms the strongest skill-augmented baseline by **+2.2%** (ID) and **+8.5%** (OOD) across ALFWorld and WebShop. Our contributions are three-fold:

- We identify the necessity for differentiated skill treatment in agentic RL, advocating that general skills should be internalized while task-specific skills are dynamically utilized, especially for authentic OOD deployment scenarios.
- We propose Skill0.5, a novel RL framework featuring an adaptive difficulty-aware router that applies tailored optimization objectives across distinct mastery tiers to jointly internalize and utilize skills.
- We conduct extensive evaluations on ALFWorld and WebShop under both ID and OOD settings, experimentally demonstrating the effectiveness of joint optimization based on functional skill decoupling.

2 Preliminary

2.1 Task Formulation

We consider an agent interacting with a text-based environment $\mathcal{E}$ modeled as a Partially Observable Markov Decision Process (POMDP), designated by the tuple $\mathcal{M} = (\mathcal{S}_{\text{env}}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \Omega, \mathcal{R}_{\text{step}}, \gamma)$. At each observation turn t, the agent receives a partial textual observation $o_t \in \mathcal{O}$ that exposes a localized view of the environment state $s_t \in \mathcal{S}_{\text{env}}$. The agent then selects a free-form natural language action $a_t \in \mathcal{A}$, which triggers an environment state transition via $\mathcal{T}(s_{t+1} \mid s_t, a_t)$ and emits the next observation via $\Omega(o_{t+1} \mid s_{t+1}, a_t)$. A complete interactive sequence is captured by an episodic trajectory $\tau = (o_1, a_1, o_2, a_2, \ldots, o_T)$.

Each task is specified by a textual instruction x sampled from a task dataset $\mathcal{D}$. The LLM-based agent, parameterized by θ, generates each action $a_t \sim \pi_\theta(\cdot \mid o_{\leq t}, x, c_t)$ conditioned on the interaction history, the task instruction, and an additional context c_t (e.g., skills) injected into the prompt at turn t. For simplicity, we abbreviate this execution history as $h_t = (o_{\leq t}, x_i)$. Our goal is to optimize the policy parameters θ to maximize the expected cumulative return across the task distribution:

$$\max_\theta \mathbb{E}_{x\sim\mathcal{D},\tau\sim\pi_\theta}\left[\sum_{t=1}^{T}\gamma^{t-1}\mathcal{R}_{\text{step}}(s_t, a_t)\right]$$

For outcome-based agentic tasks, this formulation simplifies to a sparse, binary terminal reward $R(\tau) \in \{0, 1\}$, reducing the optimization goal to maximizing the expected success rate $\max_\theta \mathbb{E}[R(\tau)]$. To improve the task success rate, procedural skills are incorporated into the prompt as the *runtime context* c_t, which we formalize in the next subsection.

2.2 Skill Bank and Runtime Context

Following Xia et al. (2026a), we assume a hierarchical skill bank $\mathcal{S}$ comprising general skills $\mathcal{S}_G$ and specific skills $\mathcal{S}_S$. While $\mathcal{S}_G$ provides universally applicable strategic heuristics, $\mathcal{S}_S$ stores fine-grained execution rules explicitly tied to distinct task domains. At each interaction turn t, general skills can be fully provided to the agent due to their broad applicability. For specific skills, which are numerous and semantically fine-grained, an embedding model is used to retrieve a subset most

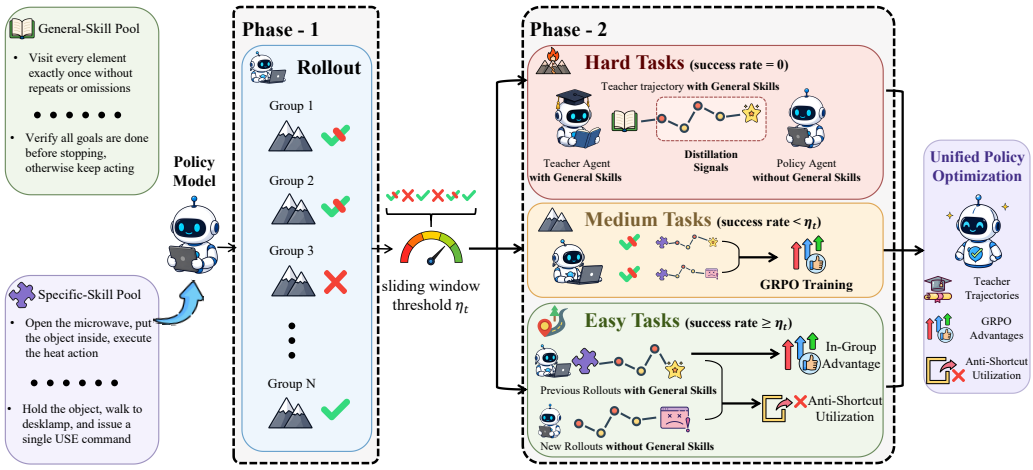

Figure 1: Overall workflow of the Skill0.5 framework. Skills are explicitly decoupled into general and specific pools. The difficulty-aware router dynamically streams tasks into three tiers for tailored optimization: hard tasks distill general skills, medium tasks apply standard GRPO to improve success rates, and easy tasks employ anti-shortcut probing to provide a utilization advantage.

relevant to the task. Let e_x and e_s denote the embeddings of the task instruction x and a candidate skill s. The selected specific skill subset $\mathcal{K}_t(x)$ is retrieved via Top-K semantic matching, measured by cosine similarity, across the available specific skill pool:

$$\mathcal{K}_t(x) = \mathrm{TopK}_{s \in \mathcal{S}_S}\big(\cos(e_x, e_s),\ K\big) \qquad (1)$$

Together with the general skills $\mathcal{S}_G$, this retrieved subset $\mathcal{K}_t(x)$ serves as the candidate guidance for constructing the auxiliary context c_t. Different skill-augmented approaches diverge in how they formulate this runtime c_t during training and inference phases:

- **Full Externalization** (e.g., SkillRL (Xia et al., 2026a)): Involves both general and selected specific skills into the context $c_t = \mathcal{S}_G \cup \mathcal{K}_t(x)$ throughout both phases.
- **Full Internalization** (e.g., SKILL0 (Lu et al., 2026b)): Progressively assimilates the full context $c_t = \mathcal{S}_G \cup \mathcal{K}_t(x)$ into model parameters during training to achieve context vacancy $c_t = \emptyset$ at deployment.
- **Hybrid Paradigms:** SLIM (Shen et al., 2026) dynamically maintains $c_t \subset \mathcal{S}$ as an updating active subset during training, and utilizes the final active skill set $c_T \subset \mathcal{S}$ at inference. For our Skill0.5, we tailor c_t for tasks of varying difficulties during training (elaborated in 3), while solely relying on specific skills $c_t = \mathcal{K}_t(x)$ during inference.

2.3 ID and OOD Settings

We simulate an authentic skill deployment scenario. The complete task domain space is partitioned into ID domains $\mathcal{D}_{\text{id}}$ and OOD domains $\mathcal{D}_{\text{ood}}$, partitioning the entire specific skill pool accordingly into $\mathcal{S}_S^{\text{id}}$ and $\mathcal{S}_S^{\text{ood}}$. The ID tasks are further divided into training splits $\mathcal{X}_{\text{train}}^{\text{id}}$ and validation splits $\mathcal{X}_{\text{val}}^{\text{id}}$. Note that all the general skills $\mathcal{S}_G$ remain globally accessible across all phases, due to their cross-domain applicability.

During training, the agent encounters ID training tasks $x \sim \mathcal{X}_{\text{train}}^{\text{id}}$ alongside their corresponding ID specific skills $\mathcal{S}_S^{\text{id}}$, whereas OOD tasks and the paired specific skills $\mathcal{S}_S^{\text{ood}}$ remain strictly unobserved. During evaluation, we assess the agent under two settings: ID evaluation samples tasks from $x \sim \mathcal{X}_{\text{val}}^{\text{id}}$ with retrieval $\mathcal{K}_t(x)$ performed exclusively over $\mathcal{S}_S^{\text{id}}$, while OOD evaluation samples from the unseen $x \sim \mathcal{X}^{\text{ood}}$ with retrieval conducted over the previously unobserved $\mathcal{S}_S^{\text{ood}}$.

Different methods reflect their design principles by how they expose accessible skills at inference. Our philosophy is to fully internalize the strategic essence of $\mathcal{S}_G$ during ID training, and to generalize to unseen tasks by exclusively utilizing plug-and-play specific skills $\mathcal{K}_t(x)$ at evaluation.

3 Method

Achieving joint skill internalization and utilization requires strategic training design. In cognitive science (Sweller, 1988), expertise acquisition follows a sequential progression: learners must first construct foundational cognitive schemas before effi-

ciently processing domain-specific rules to prevent cognitive overload. Analogously, an agent cannot effectively utilize task-specific guidance until it has internalized the general logical foundation to interact with the environment.

Motivated by this cognitive progression, we propose **Skill0.5**, an agentic RL framework that dynamically decouples the optimization towards general and specific skills based on the agent's real-time task mastery. To achieve this, our framework operates in a streamlined two-phase sampling and optimization paradigm, as depicted in Figure 1. Specifically, **Phase-1** (§3.1) executes a difficulty-aware router based on empirical pass rates to stream tasks into three mastery tiers. Subsequently, **Phase-2** (§3.2) applies tier-tailored optimization: *hard* tasks necessitate the internalization of general heuristics; *medium* tasks prioritize maximizing pass rates; and *easy* tasks ensure that specific skills are genuinely utilized. By providing tailored optimization objectives for each tier, Skill0.5 promotes the joint internalization and utilization of hierarchical skills.

3.1 Phase-1: Difficulty-Aware Routing

We measure task difficulties using the empirical task pass rate. At step t, for each task x_i in batch $\mathcal{B}_t$, we sample G independent trajectories $\tau^{(1)} \sim \pi_\theta(\cdot \mid h_t, c_t^{\text{std}})$ on the STANDARD PROMPT, where $c_t^{\text{std}} = \mathcal{K}_t(x_i)$ ensures only retrieved specific skills are used. This rollout configuration shares the exact same prompt construction as the inference phase. The difficulty of x_i is then evaluated by $p_i = \frac{1}{G}\sum_{g=1}^{G} R(\tau_i^{(1,g)})$, where $R \in \{0, 1\}$ is the binary environmental outcome. We strictly adhere to the ID training setting formulated in §2.3, thus omitting the "id" superscripts for brevity.

Crucially, these Phase-1 trajectories serve a dual purpose: they act as probing signals to dynamically route the tasks, and are opportunistically reused to support tier-tailored optimization in Phase-2.

Based on the evaluated pass rates, tasks with complete failure, i.e., $p_i = 0$, are directly routed to the *Hard* tier. To further delineate *Medium* from *Easy* tasks, we use a cross-step sliding window average as a dynamic threshold η_t. This running average is more robust against the limited task amount within a single batch. Given window size W and the batch-level mean $\bar{p}_t = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} p_i$, the threshold η_t averages these means over the past window $[t - \min(W, t) + 1, t]$:

$$\eta_t = \frac{1}{\min(W, t)} \sum_{j=t-\min(W,t)+1}^{t} \bar{p}_j \tag{2}$$

Task x_i is treated as *Easy* if $p_i > \eta_t$, and *Medium* otherwise. We formalize this difficulty-aware router $\mathcal{M}(x_i)$ as:

$$\mathcal{M}(x_i) = \begin{cases} \text{Hard}, & \text{if } p_i = 0 \\ \text{Medium}, & \text{if } 0 < p_i \le \eta_t \\ \text{Easy}, & \text{if } p_i > \eta_t \end{cases} \tag{3}$$

3.2 Phase-2: Tier-Tailored Optimization

Driven by the real-time mastery reflected from Phase-1, the agent now applies targeted optimization objectives for tasks at each tier.

3.2.1 Hard Tasks: Internalization via Privileged Distillation

When encountering hard tasks, the agent exposes a lack of foundational reasoning logic. To teach the agent how to think, we introduce the PRIVILEGED PROMPT, expanding the runtime context to include general heuristics: $c_t^{\text{priv}} = \mathcal{S}_G \cup \mathcal{K}_t(x_i)$. Specifically, we leverage $\mathcal{S}_G$ as privileged information to elicit correct reasoning traces. The agent re-attempts each hard task for G times under this enriched context, performing Phase-2 rollouts as a *teacher*: $\tau^{(2)} \sim \pi_\theta(\cdot \mid h_t, c_t^{\text{priv}})$. These rollouts are filtered for successful trajectories to construct a golden set $\mathcal{T} = \{\tau^{(2)} \mid R(\tau^{(2)}) = 1\}$. Discarding the zero-reward Phase-1 trajectories, we employ teacher forcing to distill this oracle behavior into the student. Specifically, by computing the student's probability distribution along the teacher's successful rollouts $\tau^{(2)} \in \mathcal{T}$, we force the student policy (given only c_t^{std}) to mimic the exact reasoning steps of the teacher (guided by c_t^{priv}). This alignment is optimized via token-level Jensen-Shannon Divergence (JSD), inspired by Ding (2026):

$$\mathcal{L}_{\text{hard}} = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} \frac{1}{|\tau|}\sum_{k=1}^{|\tau|} \text{JSD}\big(\text{sg}[\pi_\theta^{\text{t}}(k)] \,\big\|\, \pi_\theta^{\text{s}}(k)\big) \tag{4}$$

where $\pi_\theta^{\text{t}}(k) := \pi_\theta(\cdot \mid h_k, c_t^{\text{priv}})$ and $\pi_\theta^{\text{s}}(k) := \pi_\theta(\cdot \mid h_k, c_t^{\text{std}})$. Here, $\text{sg}[\cdot]$ represents the stop-gradient operator, guaranteeing that the student policy actively aligns with the teacher. This enables the agent to handle basic heuristics as if it were guided by $\mathcal{S}_G$ without explicitly conditioning on it, presenting a natural internalization process compatible for inference.

3.2.2 Medium Tasks: Capability Reinforcement

For medium tasks whose pass rates fall below the router threshold η_t, the agent has bypassed the complete cold-start stage but still exhibits substantial room for capability optimization. We directly reuse the Phase-1 trajectories collected during the routing phase, comprising G rollouts for each task. Standard GRPO (Shao et al., 2024) is applied to maximize the agent's success rate.

Let the policy ratio for a trajectory τ at step k be $\rho_k^{(g)} = \frac{\pi_\theta(a_k|h_k, c_t^{\text{std}})}{\pi_{\theta_{\text{old}}}(a_k|h_k, c_t^{\text{std}})}$. The RL objective $\mathcal{L}_{\text{med}}$ for these medium tasks is formulated as:

$$\mathcal{L}_{\text{medium}} = \frac{1}{G} \sum_{g=1}^{G} \sum_{k=1}^{|\tau^{(g)}|} \min \Big(\rho_k^{(g)} A_i^{(g)}, \text{clip}\big(\rho_k^{(g)}, 1-\epsilon, 1+\epsilon\big) A_i^{(g)} \Big) \quad (5)$$

where ϵ is the clipping hyperparameter. The advantage $A_i^{(g)}$ is computed via intra-group normalization: $A_i^{(g)} = \frac{R_i^{(g)} - \text{mean}(\mathbf{R}_i)}{\text{std}(\mathbf{R}_i)}$, where $\mathbf{R}_i = \{R_i^{(1)}, \ldots, R_i^{(G)}\}$ denotes the rewards for the G trajectories sampled for task x_i.

This medium tier functions as the optimization sweet spot (Yu et al., 2026). Through trial and error driven by reward signals, $\mathcal{L}_{\text{med}}$ reinforces the agent's active utilization of specific skills, elevating the sampling efficiency of correct reasoning paths and ultimately maximizing the task success rates.

3.2.3 Easy Tasks: Anti-Shortcut Utilization

With the success rate continuously escalating in the easy tier, the policy risks falling into shortcut learning (Sun et al., 2025). Rather than genuinely utilizing the retrieved specific skills $\mathcal{K}_t(x_i)$, the agent tends to memorize spurious mappings from task instructions directly to actions. This superficial overfitting severely hurts genuine skill utilization and degrades OOD generalization, where dynamically adapting to unseen specific skills is mandatory.

To penalize such shortcut behaviors, we introduce a counterfactual diagnostic probe: the NO-SKILL PROMPT, where specific skills are deliberately ablated, i.e., $c_t^{none} = \emptyset$. For each easy task x_i, we force the agent to perform Phase-2 rollouts under c_t^{none} to sample G trajectories, and measure the intra-group empirical pass rate p_i^{none}. Crucially, these diagnostic trajectories serve strictly as a counterfactual anchor to isolate the utilization gain, without participating in the policy gradient computation.

We quantify the agent's reliance on specific skills via the utilization gain $u_i = p_i - p_i^{\text{none}}$ where p_i is the original Phase-1 pass rate of the same task x_i conditioned on c_t^{std}. Intuitively, this gain captures the causal impact of the specific skills on task success. A robust agent equipped with necessary skills should strictly outperform its unguided counterpart. When u_i shrinks or becomes negative, it exposes the agent's behavior of bypassing the external guidance.

To optimize for this reliance, we apply a sliding window to track the mean utilization gain over the recent W steps, denoted as $\bar{u}_t$. By treating $\bar{u}_t$ as a dynamic anchor, we naturally construct an auxiliary task-level utilization advantage A_i^u for the tasks in the current batch:

$$A_i^u = \frac{u_i - \overline{u}_t}{\sigma_u} \quad (6)$$

where σ_u is the batch-level standard deviation of $(u_i - \overline{u}_t)$. Unlike the standard intra-group advantage $A_i^{(g)}$ which performs zero-mean normalization to evaluate the *relative* quality among trajectories, A_i^u serves as a *global task-level modulator*. It shifts the entire advantage landscape of the task. The composite advantage for the g-th rollout (sampled from Phase-1 under c_t^{std}) is thus formulated as:

$$\hat{A}_i^{(g)} = \underbrace{A_i^{(g)}}_{\text{Trajectory-level quality}} + \underbrace{A_i^u}_{\text{Task-level utilization}} \quad (7)$$

If a task exposes shortcut learning ($u_i < \overline{u}_t$), the negative offset A_i^u globally suppresses the optimization gradients for this task, penalizing the distribution of actions that bypass specific skills. Finally, the objective $\mathcal{L}_{easy}$ is optimized by substituting the standard $A_i^{(g)}$ with the composite advantage $\hat{A}_i^{(g)}$ into the identical GRPO framework.

Overall Objective. Ultimately, the global optimization objective of Skill0.5 is formulated as the joint aggregation of the tier-specific losses:

$$\mathcal{L} = \mathcal{L}_{hard} + \mathcal{L}_{medium} + \mathcal{L}_{easy} \quad (8)$$

For any single task x_i within a training batch, these optimization signals are mutually exclusive due to the routing boundaries. This dynamic routing mechanism establishes a structured curriculum synchronized with the agent's real-time mastery dynamics. By allocating tailored learning objectives

Method	ID					OOD				
	Pick	Cool	Clean	Avg.	Rank ↓	Look	Heat	Pick2	Avg.	Rank ↓
Prompt-based Methods										
Zero-shot	28.6	12.0	18.5	20.7	17.2	38.5	12.5	12.5	18.9	15.5
Few-shot	62.9	44.0	63.0	57.5	12.0	46.2	31.2	8.3	24.5	12.8
ReAct	71.4	28.0	33.3	47.1	13.3	46.2	18.8	12.5	22.6	12.8
Reflexion	85.7	44.0	44.4	60.9	10.3	46.2	31.3	29.2	34.0	8.7
Mem0	54.3	4.0	18.5	28.7	17.5	38.5	18.8	4.2	17.0	15.7
ExpeL	80.0	44.0	66.7	65.5	9.5	46.2	18.8	20.8	18.9	11.2
MemP	65.7	12.0	33.3	40.2	15.3	46.2	37.5	12.5	28.3	11.0
SimpleMem	71.4	16.0	44.4	47.1	13.3	53.8	18.8	20.8	28.3	9.5
RL-based Methods										
RLOO	91.4	80.0	81.5	85.1	4.3	61.5	56.3	20.8	41.5	5.8
GRPO	80.0	72.0	88.9	80.5	6.2	**76.9**	56.3	16.7	43.4	5.5
Memory-Augmented RL Methods										
MemRL	74.3	12.0	55.6	50.6	12.7	46.2	12.5	45.8	35.8	9.7
EvolveR	88.6	52.0	81.5	75.9	6.2	46.2	6.2	**50.0**	35.8	10.0
Mem0+GRPO	65.7	20.0	51.9	48.3	13.2	23.1	6.2	20.8	17.0	15.0
SimpleMem+GRPO	85.7	52.0	70.3	71.3	7.7	61.5	43.8	41.7	47.2	4.5
Skill-Augmented RL Methods										
SkillRL	91.4	84.0	**96.3**	90.8	2.7	69.2	75.0	12.5	45.3	6.3
SKILL0	**94.3**	76.0	81.5	85.1	3.8	46.2	50.0	29.2	39.6	7.5
SLIM	91.4	84.0	70.4	82.8	4.5	53.8	31.3	29.2	35.8	7.0
Skill0.5	**94.3**	**88.0**	**96.3**	**93.1**	**1.3**	69.2	**87.5**	33.3	**58.5**	**2.5**

Table 1: Performance comparison on ALFWorld under ID and OOD task settings. **Best** and second-best results in each column are highlighted, respectively. Average Rank is computed across all evaluated settings.

based on real-time task mastery, Skill0.5 achieves joint optimization of foundational reasoning internalization and task-specific guidance utilization within a unified RL framework. The full procedure is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Environments and ID/OOD Partition. We evaluate our framework on two multi-turn interactive benchmarks that offer clear domain segmentation, enabling a rigorous study of OOD generalization.

- **ALFWorld** (Shridhar et al., 2020) is a text-based embodied environment where agents complete household tasks through natural language actions. We evaluate on its six canonical task types. We designate {Pick, Cool, Clean} as ID and {Look, Heat, Pick2} as OOD domains.
- **WebShop** (Yao et al., 2022a) is a web-based shopping environment where agents search for products and make purchases matching user instructions. We split product categories into ID = {Apparel, Electronics, Footwear, Other} and OOD = {Accessories, Beauty & Health, Home Decor} domains following a balanced protocol detailed in Appendix B. The OOD categories exhibit distinct attribute vocabularies and product matching heuristics entirely absent from training.

For **agent skills**, we adopt the hierarchical Skill Bank proposed by Xia et al. (2026a) as our foundational skill set. The library comprises 12 and 15 general skills for ALFWorld and WebShop, respectively, while each task domain maintains around 5 task-specific skills.

Baselines. We compare Skill0.5 against diverse spectrum of methods: (1) **Prompt-based Methods**: Zero-shot and Few-shot prompting. (2) **Prompt-based Agentic or Memory-based Methods**: ReAct (Yao et al., 2022b) and Reflexion (Shinn et al., 2023), which rely on in-context prompting for multi-step reasoning, alongside Mem0 (Chhikara et al., 2025), ExpeL (Zhao et al., 2024), MemP (Fang et al., 2025), and SimpleMem (Liu et al., 2026a), which utilize external experience pools to guide behavior without parameter updates. (3) **RL-based Methods**: Group-based RL algorithms such as RLOO (Ahmadian et al., 2024) and GRPO (Shao et al., 2024). (4) **Memory-Augmented RL**: MemRL (Zhang et al., 2026c), EvolveR (Wu et al., 2025), Mem0+GRPO, and SimpleMem+GRPO, which integrate persistent memory directly into RL optimization. (5) **Skill-Augmented RL**: SkillRL (Xia et al., 2026a), SKILL0 (Lu et al., 2026b), and SLIM (Shen et al., 2026), which represent the current frontier of skill-based agent training.

The implementation details and hyperparameter configurations are provided in Appendix D and F.

4.2 Main Results

Table 1 and Table 3 report the comprehensive success rates across distinct domains of ALFWorld and WebShop, detailing both the task-specific performance and the aggregated ID and OOD averages. We draw the following key observations:

Skill0.5 achieves the highest overall performance. Across all settings, Skill0.5 establishes a decisive performance advantage over the full spectrum of baselines. Compared to the strongest skill-augmented baseline, SkillRL, Skill0.5 achieves absolute improvements of **+2.3%** (ID) and **+13.2%** (OOD) on ALFWorld, with consistent gains of **+2.1%** and **+3.9%** on the respective splits of WebShop. These results confirm that our joint internalization and utilization framework ensures steady ID progress while unlocking significant generalization leaps in OOD scenarios.

Prompt-based methods establish a performance floor. Although prompting with skills or memory (e.g., Few-shot, ReAct) improves upon zero-shot baselines, these methods significantly lag behind Skill0.5, trailing by an average of over 45% on ALFWorld and 28% on WebShop. This profound gap highlights that relying solely on in-context learning is insufficient to effectively synergize with external skills (Liu et al., 2026b).

Memory-based methods are bottlenecked by context noise. Methods integrating trajectory storage into the prompt are highly sensitive to memory quality. For instance, SimpleMem+GRPO achieves 47.2% on ALFWorld OOD thanks to its sophisticated memory management, whereas Mem0+GRPO collapses to 17.0%. However, even the strongest memory-augmented methods fail to surpass skill-augmented approaches. This discrepancy arises because raw memory retrieval tends to inject overly detailed context noise, whereas effective knowledge transfer demands high-level procedural abstractions and reusable heuristics (Xia et al., 2026a).

4.3 Training Dynamics

Here, we investigate the training dynamics to elucidate why our approach consistently surpasses existing skill-augmented RL baselines. Figure 2 illustrates the success rate curves on the training and validation (ID and OOD) sets of ALFWorld across training steps. Concurrently, Figure 3 tracks the dynamic proportions of hard, medium, and easy tasks allocated by our difficulty-aware router.

Early-stage training: Overcoming the zero-gradient dilemma for rapid internalization. Initially, hard tasks dominate the distribution, causing zero reward variance and vanishing advantages that entirely eliminate gradient signals (Yu et al., 2026). Our difficulty-aware router resolves this by triggering privileged distillation as a surrogate gradient source, which breaks the exploration deadlock and swiftly steers optimization, enabling Skill0.5 to achieve markedly faster initial ascent compared to baselines.

Mid-to-late training: Anti-shortcut utilization driving robust OOD generalization. As training progresses, easy tasks gradually dominate, creating a phase prone to shortcut learning. Owing to our anti-shortcut diagnostic probing, Skill0.5 maintains steady growth on training and ID validation sets while achieving consistent, un-decaying improvements on OOD tasks. This demonstrates that our method genuinely learns to utilize novel skills, rather than overfitting to bypass skill guidance.

Limitations of baseline methods. SkillRL exhibits overfitting: despite a surging training success rate, its performance on validation and OOD tasks declines in later stages. This indicates a collapse into shortcut learning, where the agent memorizes domain-specific actions at the expense of adaptability. Conversely, SKILL0 relies on pure internalization but is consistently suppressed on OOD tasks, confirming that a fully internalized policy is too rigid to integrate novel task-skill pairs. Finally, SLIM suffers from severe oscillations due to its dynamic retirement mechanism prematurely discarding general skills. Losing this foundational reasoning halts progress on hard tasks, and the resulting disjointed active skill set inevitably mismatches with OOD tasks.

4.4 Ablation Study

To isolate the contributions of our tier-tailored optimization, we evaluate two ablated variants on ALFWorld: **Internalize-Only** (retains distillation for hard tasks; standard GRPO elsewhere) and **Utilize-Only** (retains contrastive utilization for easy tasks; standard GRPO elsewhere).

Table 2 details the results, revealing two critical insights: (1) **Internalization is a strict prerequisite for overall capability.** The *Utilize-Only* variant suffers a catastrophic decline across both ID and OOD splits. Although the adaptive router

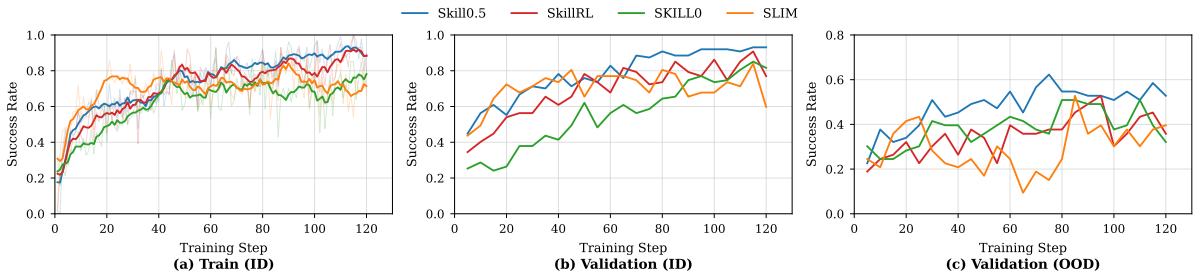

Figure 2: Success rates across the training and validation sets on ALFWorld, compared to skill-based RL baselines.

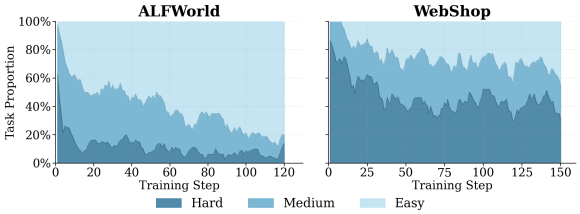

Figure 3: Dynamic distribution of task difficulties allocated by our difficulty-aware router.

Variant	ID Avg.	OOD Avg.
Internalize-Only	89.6	52.8
Utilize-Only	85.1	50.9
Skill0.5	**93.1**	**58.5**

Table 2: Ablation study on ALFWorld. We report the average success rates across ID and OOD task splits.

still allocates tasks to the easy tier based on relative performance, the contrastive utilization objective becomes futile: lacking basic competence, the agent fails regardless of whether specific skills are provided, yielding a negligible contrastive advantage that severely stalls the entire training process. (2) **Utilization unlocks peak OOD adaptability.** The *Internalize-Only* variant experiences a moderate performance drop. While the internalized general skills successfully provide a robust reasoning baseline (preventing a complete collapse), the absence of the contrastive utilization loss prevents the model from faithfully grounding its actions in novel, task-specific OOD guidance, limiting its peak generalization.

Together, these results validate the necessity of our decoupled design: internalization builds the indispensable reasoning foundation, upon which utilization further maximizes OOD adaptability.

4.5 Case Study

We conduct qualitative trajectory analysis on ALFWorld OOD tasks to diagnose why each baseline fails (full details in Appendix E). We identify three distinct failure mechanisms: SkillRL suffers from *Contextual Interference*, where OOD-specific skills are overridden by ID-trained habitual associations from the dominant general context; SKILL0 exhibits *Parametric Knowledge Conflict*, where internalized ID templates produce actions incompatible with OOD procedures despite the agent referencing the correct skill in its reasoning; and SLIM demonstrates *Behavioral Collapse*, progressively degenerating from success to task hallucination after prematurely retiring general skills. In all cases, Skill0.5 succeeds by internalizing general reasoning into parameters while faithfully executing novel OOD-specific skills, validating our thesis that the two skill types require differentiated treatments.

5 Conclusion

In this paper, we tackled the skill-treatment dilemma for LLM agents, where existing RL methods typically suffered from a rigid choice between full context externalization and indiscriminate internalization. To resolve this, we introduced Skill0.5, a unified framework that achieves joint optimization by explicitly differentiating the treatment of decoupled general and task-specific skills. Driven by a dynamic, difficulty-aware router, Skill0.5 orchestrated optimization strategies tailored to different task tiers: it compressed general skills via privileged distillation to build a cognitive foundation for hard tasks, while enforcing specific skill utilization via contrastive advantage probing on easy tasks. Extensive experiments on the ALFWorld and WebShop benchmarks demonstrated that Skill0.5 significantly outperformed prompt-, memory-, and skill-based baselines, validating that our framework yields moderate improvements on ID tasks while substantially enhancing the agent's generalization capability across practical deployment settings involving unseen OOD tasks.

Limitations

While our framework is validated on text-based interactive environments, the principle of differentiated skill treatment is broadly applicable. In future work, we plan to extend Skill0.5 to more complex domains such as code generation, multi-modal environments, and open-ended web navigation, as well as to settings with longer horizons and larger action spaces.

Ethical Considerations

This work focuses on improving the reasoning and generalization capabilities of LLM-based agents within simulated environments (ALFWorld and WebShop). We identify no direct ethical risks arising from our research. All experiments are conducted in controlled, sandboxed settings with no real-world deployment or interaction with human users. The benchmarks used are publicly available and do not involve personally identifiable information or sensitive content. Our method does not introduce new data collection procedures, and all training data is derived from synthetic environment interactions. We acknowledge that advances in autonomous agent capabilities carry broader societal implications; however, the household and e-commerce simulation domains studied here pose minimal risk of misuse.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267.

Zeyuan Allen-Zhu and Yuanzhi Li. 2025. Physics of Language Models: Part 3.3, Knowledge Capacity Scaling Laws. In *Proceedings of the 13th International Conference on Learning Representations*, ICLR '25. Full version available at https://ssrn.com/abstract=5250617.

Salaheddin Alzubi, Noah Provenzano, Jaydon Bingham, Weiyuan Chen, and Tu Vu. 2026. Evoskill: Automated skill discovery for multi-agent systems. *arXiv preprint arXiv:2603.02766*.

Shuzhen Bi, Mengsong Wu, Hao Hao, Keqian Li, Wentao Liu, Siyu Song, Hongbo Zhao, and Aimin Zhou. 2026. Automating skill acquisition through large-scale mining of open-source agentic repositories: A framework for multi-agent procedural knowledge extraction. *arXiv preprint arXiv:2603.11808*.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*.

Ken Ding. 2026. Hdpo: Hybrid distillation policy optimization via privileged self-distillation. *arXiv preprint arXiv:2603.23871*.

Shuangrui Ding, Xuanlang Dai, Long Xing, Shengyuan Ding, Ziyu Liu, Yang JingYi, Penghui Yang, Zhixiong Zhang, Xilin Wei, Xinyu Fang, and 1 others. 2026. Wildclawbench: A benchmark for real-world, long-horizon agent evaluation. *arXiv preprint arXiv:2605.10912*.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2025. Memp: Exploring agent procedural memory. *arXiv preprint arXiv:2508.06433*.

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

Lang Feng, Longtao Zheng, Shuo He, Fuxiang Zhang, and Bo An. 2026b. Dr. mas: Stable reinforcement learning for multi-agent llm systems. *arXiv preprint arXiv:2602.08847*.

Hao Guan, Lingyue Fu, Shao Zhang, Yaoming Zhu, Kangning Zhang, Lin Qiu, Xunliang Cai, Xuezhi Cao, Weiwen Liu, Weinan Zhang, and 1 others. 2026. Swe-cycle: Benchmarking code agents across the complete issue resolution cycle. *arXiv preprint arXiv:2605.13139*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*.

Chenliang Li, Adel Elmahdy, Alex Boyd, Zhongruo Wang, Alfredo Garcia, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, and Mingyi Hong. 2025. Stppo: Stabilized off-policy proximal policy optimization for multi-turn agents training. *arXiv preprint arXiv:2511.20718*.

Hao Li, Chunjiang Mu, Jianhao Chen, Siyue Ren, Zhiyao Cui, Yiqun Zhang, Lei Bai, and Shuyue Hu. 2026a. Organizing, orchestrating, and benchmarking agent skills at ecosystem scale. *arXiv preprint arXiv:2603.02176*.

Xiangyi Li, Wenbo Chen, Yimin Liu, Shenghan Zheng, Xiaokun Chen, Yifeng He, Yubo Li, Bingran You, Haotian Shen, Jiankai Sun, and 1 others. 2026b. Skillsbench: Benchmarking how well agent skills work across diverse tasks. *arXiv preprint arXiv:2602.12670*.

Xiaoyuan Li, Moxin Li, Keqin Bao, Yubo Ma, Wenjie Wang, Dayiheng Liu, and Fuli Feng. 2026c. Skillgraph: Skill-augmented reinforcement learning for agents via evolving skill graphs. *arXiv preprint arXiv:2605.12039*.

Yu Li, Rui Miao, Zhengling Qi, and Tian Lan. 2026d. Arise: Agent reasoning with intrinsic skill evolution in hierarchical reinforcement learning. *arXiv preprint arXiv:2603.16060*.

Qiliang Liang, Hansi Wang, Zhong Liang, and Yang Liu. 2026a. From skill text to skill structure: The scheduling-structural-logical representation for agent skills. *arXiv preprint arXiv:2604.24026*.

Zhenwen Liang, Yujun Zhou, Sidi Lu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2026b. Too correct to learn: Reinforcement learning on saturated reasoning data. *arXiv preprint arXiv:2604.18493*.

George Ling, Shanshan Zhong, and Richard Huang. 2026. Agent skills: A data-driven analysis of claude skills for extending large language model functionality. *arXiv preprint arXiv:2602.08004*.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. 2026a. Simplemem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Yujian Liu, Jiabao Ji, Li An, Tommi Jaakkola, Yang Zhang, and Shiyu Chang. 2026b. How well do agentic skills work in the wild: Benchmarking llm skill usage in realistic settings. *arXiv preprint arXiv:2604.04323*.

Zhengxi Lu, Zhiyuan Yao, Zhuowen Han, Zi-Han Wang, Jinyang Wu, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and 1 others. 2026a. Self-distilled agentic reinforcement learning. *arXiv preprint arXiv:2605.15155*.

Zhengxi Lu, Zhiyuan Yao, Jinyang Wu, Chengcheng Han, Qi Gu, Xunliang Cai, Weiming Lu, Jun Xiao, Yueting Zhuang, and Yongliang Shen. 2026b. Skill0: In-context agentic reinforcement learning for skill internalization. *arXiv preprint arXiv:2604.02268*.

Yingwei Ma, Yue Liu, Xinlong Yang, Yanhao Li, Kelin Fu, Yibo Miao, Yuchong Xie, Zhexu Wang, and Shing-Chi Cheung. 2026a. Scaling coding agents via atomic skills. *arXiv preprint arXiv:2604.05013*.

Ziyu Ma, Shidong Yang, Yuxiang Ji, Xucong Wang, Yong Wang, Yiming Hu, Tongwen Huang, and Xiangxiang Chu. 2026b. Skillclaw: Let skills evolve collectively with agentic evolver. *arXiv preprint arXiv:2604.08377*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Junhao Shen, Teng Zhang, Xiaoyan Zhao, and Hong Cheng. 2026. Dynamic skill lifecycle management for agentic reinforcement learning. *arXiv preprint arXiv:2605.10923*.

Yaorui Shi, Yuxin Chen, Zhengxi Lu, Yuchun Miao, Shugui Liu, Qi Gu, Xunliang Cai, Xiang Wang, and An Zhang. 2026. Skill1: Unified evolution of skill-augmented agents via reinforcement learning. *arXiv preprint arXiv:2605.06130*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.

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

Shuzheng Si, Haozhe Zhao, Yu Lei, Qingyi Wang, Dingwei Chen, Zhitong Wang, Zhenhailong Wang, Kangyang Luo, Zheng Wang, Gang Chen, and 1 others. 2026. From context to skills: Can language models learn from context skillfully? *arXiv preprint arXiv:2604.27660*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Xinyi Sun, Hongye Tan, Yaxin Guo, Pengpeng Qiang, Ru Li, and Hu Zhang. 2025. Mitigating shortcut learning via smart data augmentation based on large language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8160–8172.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Meituan LongCat Team, Anchun Gui, Bei Li, Bingyang Tao, Bole Zhou, Borun Chen, Chao Zhang, Chen Gao, Chen Zhang, Chengcheng Han, and 1 others. 2026. Longcat-flash-thinking-2601 technical report. *arXiv preprint arXiv:2601.16725*.

Chenxi Wang, Zhuoyun Yu, Xin Xie, Wuguannan Yao, Runnan Fang, Shuofei Qiao, Kexin Cao, Guozhou Zheng, Xiang Qi, Peng Zhang, and 1 others. 2026a. Skillx: Automatically constructing skill knowledge bases for agents. *arXiv preprint arXiv:2604.04804*.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan O Arik. 2025a. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30553–30571.

Hanlin Wang, Chak Tou Leong, Jiashuo Wang, Jian Wang, and Wenjie Li. 2025b. Spa-rl: Reinforcing llm agents via stepwise progress attribution. *arXiv preprint arXiv:2505.20732*.

Hao Wang, Guozhi Wang, Han Xiao, Yufeng Zhou, Yue Pan, Jichao Wang, Ke Xu, Yafei Wen, Xiaohu Ruan, Xiaoxin Chen, and 1 others. 2026b. Skill-sd: Skill-conditioned self-distillation for multi-turn llm agents. *arXiv preprint arXiv:2604.10674*.

Jiawei Wang, Jiacai Liu, Yuqian Fu, Yingru Li, Xintao Wang, Yuan Lin, Yu Yue, Lin Zhang, Yang Wang, and Ke Wang. 2025c. Harnessing uncertainty: Entropy-modulated policy gradients for long-horizon llm agents. *arXiv preprint arXiv:2509.09265*.

Zihan Wang, Chi Gui, Xing Jin, Qineng Wang, Licheng Liu, Kangrui Wang, Shiqi Chen, Linjie Li, Zhengyuan Yang, Pingyue Zhang, and 1 others. 2026c. Understanding reasoning collapse in multi-turn agent reinforcement learning. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*.

Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, and 1 others. 2025d. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning. *arXiv preprint arXiv:2504.20073*.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, and 1 others. 2025. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *arXiv preprint arXiv:2505.11821*.

Rong Wu, Xiaoman Wang, Jianbiao Mei, Pinlong Cai, Daocheng Fu, Cheng Yang, Licheng Wen, Xuemeng Yang, Yufan Shen, Yuxin Wang, and 1 others. 2025. Evolver: Self-evolving llm agents through an experience-driven lifecycle. *arXiv preprint arXiv:2510.16079*.

Yaxiong Wu and Yongyue Zhang. 2026. Agent skills from the perspective of procedural memory: A survey. *Authorea Preprints*.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, and 1 others. 2026a. Skillrl: Evolving agents via recursive skill-augmented reinforcement learning, 2026. *URL https://arxiv. org/abs/2602.08234*.

Peng Xia, Jianwen Chen, Xinyu Yang, Haoqin Tu, Jiaqi Liu, Kaiwen Xiong, Siwei Han, Shi Qiu, Haonian Ji, Yuyin Zhou, and 1 others. 2026b. Metaclaw: Just talk–an agent that meta-learns and evolves in the wild. *arXiv preprint arXiv:2603.17187*.

Renjun Xu and Yang Yan. 2026. Agent skills for large language models: Architecture, acquisition, security, and the path forward. *arXiv preprint arXiv:2602.12430*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565.

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

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

Woongyeng Yeo, Yumin Choi, Taekyung Ki, and Sung Ju Hwang. 2026. Hint-sd: Targeted hindsight self-distillation for long-horizon agents. *arXiv preprint arXiv:2605.17873*.

Zhiqi Yu, Zhangquan Chen, Mengting Liu, Heye Zhang, and Liangqiong Qu. 2026. Unveiling implicit advantage symmetry: Why grpo struggles with exploration and difficulty adaptation. *arXiv preprint arXiv:2602.05548*.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Genrui Zhang, Erle Zhu, Jinfeng Zhou, Caiyan Jia, and Hongning Wang. 2026a. Skillevolver: Skill learning as a meta-skill. *arXiv preprint arXiv:2605.10500*.

Hanrong Zhang, Shicheng Fan, Henry Peng Zou, Yankai Chen, Zhenting Wang, Jiayu Zhou, Chengze Li, Wei-Chieh Huang, Yifei Yao, Kening Zheng, and 1 others. 2026b. Coevoskills: Self-evolving agent skills via co-evolutionary verification. *arXiv preprint arXiv:2604.01687*.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, and 1 others. 2026c. Memrl: Self-evolving agents via runtime reinforcement learning on episodic memory. *arXiv preprint arXiv:2601.03192*.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. 2024. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642.

Chenyu Zhou, Huacan Chai, Wenteng Chen, Zihan Guo, Rong Shan, Yuanyi Song, Tianyi Xu, Yingxuan Yang, Aofan Yu, Weiming Zhang, and 1 others. 2026a. Externalization in llm agents: A unified review of memory, skills, protocols and harness engineering. *arXiv preprint arXiv:2604.08224*.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl, 2024. *URL https://arxiv. org/abs/2402.19446*.

Yingli Zhou, Wang Shu, Yaodong Su, Wenchuan Du, Yixiang Fang, and Xuemin Lin. 2026b. A comprehensive survey on agent skills: Taxonomy, techniques, and applications. *arXiv preprint arXiv:2605.07358*.

A Related Work

A.1 Skill-Augmented Agentic Training

Early work on agent skills predominantly employed skills as training-free in-context augmentations (Wang et al., 2026a; Bi et al., 2026; Liang et al., 2026a; Zhang et al., 2026b). However, constrained by model capacities and the increasing complexity of agentic tasks, recent research has expanded into training agents to effectively harness skills, which primarily diverge into two paradigms. One paradigm advocates for full externalization (Shi et al., 2026; Xia et al., 2026b; Li et al., 2026d,c). SkillRL (Xia et al., 2026a), for example, constructs a hierarchical skill bank where general skills are appended with task-specific skills, persistently maintaining all guidance within the contextual window throughout training and inference. Another line of research explores *full internalization* to eliminate runtime context overhead (Wang et al., 2026b; Lu et al., 2026a; Yeo et al., 2026). SKILL0 (Lu et al., 2026b), for instance, leverages a dynamic curriculum to progressively withdraw skills from the context until the agent operates without any external input, completely assimilating all guidance utility into model parameters.

Yet, joint skill internalization and utilization remains underexplored. A closely related work, SLIM (Shen et al., 2026), dynamically decides whether to retain skills for active utilization or retire them upon internalization. Still, SLIM treats all skills uniformly, and its final active skill set risks incompatibility with OOD tasks paired with unseen specific skills. In contrast, we explicitly decouple general and task-specific skills, jointly optimizing them for foundational internalization and adaptive utilization, respectively. This parameterizes foundational reasoning logic to actively exploit tailored external guidance in authentic OOD settings.

A.2 Agentic Reinforcement Learning

Reinforcement learning, particularly Group Relative Policy Optimization (GRPO) (Shao et al., 2024), has become a core backbone for training LLMs as interactive agents. Recent algorithmic advances build upon this framework to address central challenges in multi-turn environments, including temporal credit assignment (Feng et al., 2026a; Wei et al., 2025; Wang et al., 2025b), long-horizon optimization (Zhou et al., 2024; Li et al., 2025; Wang et al., 2025c), and training stability against degenerate action cycles (Wang et al., 2025d; Feng et al., 2026b; Wang et al., 2026c). While these advances provide a solid optimization foundation, the objective of GRPO remains difficulty-agnostic—the intra-group reward variance collapses to zero on either impossibly hard tasks or overly simple tasks, which leads to exploration stagnation (Yu et al., 2026) and saturation-induced mode collapse (Liang et al., 2026b), respectively. Therefore, we dynamically perceive task difficulty and assign tailored auxiliary optimization objectives specifically for excessively hard and near-saturated tasks, ensuring the effectiveness of agentic RL training.

B WebShop Domain Split Statistics.

We use the 12,087 human-annotated goals from WebShop and partition them into seven domains via keyword-based classification of goal instructions. Four domains serve as ID categories for training and ID evaluation, while three are held out as out-of-distribution OOD categories for OOD evaluation only. To address the severe imbalance in the miscellaneous *Other* category (which accounts for over 60% of all goals before processing), we apply farthest-point sampling (FPS) based on sentence embeddings to downsample it to a scale comparable with the other domains.

The resulting per-domain statistics are as follows. **Training set (ID, 3,320 goals):** Apparel 776, Electronics 938, Footwear 606, Other 1,000 (downsampled via FPS from ∼6,600). **ID validation set (454 goals):** Apparel 113, Electronics 152,

Footwear 89, Other 100 (downsampled via FPS from ∼940). **OOD validation set (207 goals):** Accessories 80, Beauty & Health 54, Home Decor 73.

The training range corresponds to goal indices 1,500+ in the original WebShop ordering, while the validation pool merges the original test (0–499) and development (500–1,499) splits to ensure sufficient per-domain coverage.

C Pseudo Code

Algorithm 1 Skill0.5: Joint Skill Internalization and Utilization

Require: Policy π_θ, General Skills $\mathcal{S}_G$, Specific Skills $\mathcal{S}_S$
for each training step t **do**
  Sample batch $\mathcal{B}_t \sim \mathcal{X}^{id}_{train}$. Initialize $\mathcal{L} \leftarrow 0$.
  % Phase-1: Difficulty-Aware Routing
  for each $x_i \in \mathcal{B}_t$ **do**
    $c_t^{std} \leftarrow \mathcal{K}_t(x_i)$. Sample G rollouts $\tau^{(1)} \sim \pi_\theta(\cdot|h_t, c_t^{std})$.
    Evaluate empirical pass rate p_i.
  end for
  Compute dynamic routing threshold η_t (Eq. 2).
  % Phase-2: Tier-Tailored Optimization
  for each $x_i \in \mathcal{B}_t$ **do**
    if $p_i == 0$ **then**
      % Hard Tier
      Sample $\tau^{(2)}$ guided by $c_t^{priv} \leftarrow \mathcal{S}_G \cup \mathcal{K}_t(x_i)$.
      $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{hard}$ via token-level JSD (Eq. 4).
    else if $p_i \leq \eta_t$ **then**
      % Medium Tier
      $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{medium}$ via standard GRPO on $\tau^{(1)}$ (Eq. 5).
    else
      % Easy Tier
      Sample diagnostic τ_{diag} using $c_t^{none} \leftarrow \emptyset$ to get p_i^{none}.
      $A_i^u \leftarrow$ offset derived from utilization gain $p_i - p_i^{none}$ (Eq. 6).
      $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}_{easy}$ via GRPO using $\hat{A}_i^{(g)} = A_i^{(g)} + A_i^u$.
    end if
  end for
  Update policy parameters θ using $\nabla_\theta \mathcal{L}$.
end for

D Implementation Details

Inference Protocol. To ensure a fair comparison, we strictly control the prompt conditions at test time to match each skill-bases method's design principle and OOD setting constraints. Specifically, **SkillRL** receives both general skills and ID or OOD-specific skills retrieved from the skill set. **SKILL0** receives no skill context on ID tasks due to full internalization, and receives only OOD-specific skills on OOD tasks. **SLIM** receives its final trained active skill set (comprising a subset of general and ID-specific skills) for ID testing, whereas its ID-specific skills are replaced with OOD ones for OOD evaluation. Our **Skill0.5** receives only ID or OOD-specific skills during inference , as general skills have been internalized.

Training and Implementation Details. We use Qwen2.5-7B-Instruct as the base model. For policy optimization, we employ GRPO as the backbone with a group size of $G = 8$, a learning rate of 1×10^{-6}. Training is conducted on 4 H800 GPUs with a batch size of 16 tasks per iteration, spanning 120 steps for ALFWorld and 150 steps for WebShop. The maximum interaction horizon is set to 30 steps for ALFWorld and 15 steps for WebShop. Task-specific skills are retrieved via Qwen3-Embedding-0.6B with a retrieval capacity of $K = 3$. A sliding window of size $W = 5$ is maintained for routing threshold and utilization gain tracking. For privileged distillation, the token-level JSD optimization is performed over the top-64 tokens following Ding (2026).

E Case Study

To qualitatively demonstrate why differentiated skill treatment is essential, we present representative OOD failure cases from each skill-augmented baseline on ALFWorld, contrasting them with Skill0.5. Table 4 summarizes the failure mechanisms, and Figure 4 provides detailed trajectory comparisons.

Case 1: Contextual Interference (SkillRL). In the "Heat & Place" task, SkillRL's context window contains ∼1,617 tokens of general principles, common mistakes, and OOD-specific skills, with the latter occupying only ∼12% of the total context. After successfully heating the potato and navigating to the fridge, the agent executes `cool potato with fridge`—directly contradicting both the task goal and the injected skill "*Direct Post-Heat Placement: place the object once*". The agent's reasoning reveals that the ID-trained association "fridge → cool" (from Cool & Place tasks) is activated by the general heuristic "Use State-Changing Tools Early", overpowering the novel OOD instruction. In contrast, Skill0.5 internalizes general skills into parameters, leaving only ∼214 tokens of OOD-specific guidance in the context. This ensures the novel "place" instruction receives undiluted attention, enabling correct execution in 7 steps.

Method	ID					OOD			
	Apparel	Elec.	Footwear	Other	Avg.	Access.	Beauty	Home	Avg.
Prompt-based Methods									
Zero-shot	4.4	4.6	3.4	1.0	3.5	2.5	3.7	5.5	3.9
Few-shot	14.2	15.1	13.5	24.0	16.5	18.8	29.6	5.5	16.9
Prompt-based Agentic or Memory-based Methods									
ReAct	12.4	11.1	4.5	12.0	10.4	11.2	24.1	2.7	11.6
Reflexion	3.5	8.6	1.1	7.0	5.5	6.3	5.6	1.4	4.4
Mem0	8.9	9.2	2.3	11.0	8.2	10.0	16.7	4.1	9.7
ExpeL	6.2	12.5	9.0	21.0	12.1	12.5	25.9	8.2	14.5
MemP	15.9	13.2	9.0	19.0	14.3	16.2	14.8	9.6	13.5
SimpleMem	11.5	13.8	6.7	13.0	11.7	10.0	20.4	5.5	11.1
RL-based Methods									
RLOO	34.9	23.9	**41.5**	32.1	31.1	31.4	46.5	22.5	32.9
GRPO	35.1	22.6	39.0	49.5	33.6	27.7	47.3	25.9	32.3
Memory-Augmented RL Methods									
MemRL	22.2	15.2	25.0	48.3	26.2	13.2	17.9	27.8	19.6
EvolveR	32.5	31.1	25.0	20.9	28.0	20.8	28.6	18.2	21.9
Mem0+GRPO	36.5	22.0	40.6	23.1	29.5	10.2	32.1	30.2	23.0
SimpleMem+GRPO	25.4	28.0	26.6	25.1	26.4	16.2	32.1	29.4	25.0
Skill-Augmented RL Methods									
SkillRL	36.0	34.2	41.4	49.3	38.3	36.3	48.5	27.6	36.7
SKILL0	**39.2**	33.0	38.1	37.9	35.2	**42.1**	38.6	26.5	35.4
SLIM	31.9	36.8	31.5	33.0	33.7	35.0	29.6	**35.6**	33.8
Skill0.5	39.1	**37.3**	41.1	**50.9**	**40.4**	36.6	**54.2**	31.4	**40.6**

Table 3: Performance comparison on WebShop under ID and OOD task settings. **Best** and second-best results in each column are highlighted.

Method	OOD Task	Steps	Result	Failure Diagnosis
SkillRL	Heat & Place	30	✗	Contextual Interference
Skill0.5	Heat & Place	7	✓	—
SKILL0	Examine in Light	20	✗	Parametric Conflict
Skill0.5	Examine in Light	4	✓	—
SLIM	Examine in Light	30	✗	Behavioral Collapse
Skill0.5	Examine in Light	3	✓	—

Table 4: Summary of representative OOD failure cases. Each baseline exhibits a distinct failure mechanism, while Skill0.5 consistently succeeds with significantly fewer steps.

Case 2: Parametric Knowledge Conflict (SKILL0). For the "Examine in Light" task, SKILL0 receives the OOD-specific skill "*Switch Lamp On: Issue 'use desklamp' as soon as you reach it*". Despite the explicit instruction, the agent activates internalized ID procedural templates: it executes `move bowl to sidetable` (Pick & Place terminal action) and repeatedly attempts `take desklamp` (treating it as a portable tool). Notably, the agent's reasoning even references the skill ("According to the Single Toggle Rule...") yet immediately violates it—demonstrating that context-free training has atrophied the model's instruction-following capability, allowing parametric priors to dominate over novel textual guidance. Skill0.5 avoids this conflict entirely: general skills (domain-agnostic, non-procedural) are internalized, while task-specific skills remain external. The contrastive utilization training (§3.2.3) explicitly builds the "read instruction → execute" capability, enabling faithful compliance with novel OOD skills.

Case 3: Behavioral Collapse (SLIM). SLIM's lifecycle management retires "Systematic Exploration" at step 5 (utility ≈ 0.001) alongside 7 other general skills by step 50 (66.7% retired). At step 85, the agent still succeeds on this task. However, by step 120—after 70 additional training steps without these general constraints—the policy exhibits catastrophic degradation: task hallucination (reasoning about "cd" when the task specifies "pillow"), think-action decoupling (reasoning says "fridge" but action goes to "bed"), and cross-task pattern collapse (identical degenerate template across 5 unrelated tasks). This irreversible cognitive erosion stems from treating general skills as retirable commodities rather than permanent cognitive foundations. Skill0.5 permanently embeds general skills via JSD distillation, making them immune to retirement decisions and ensuring stable reasoning throughout extended training.

Unified Diagnosis. These cases demonstrate that each uniform skill treatment strategy produces a characteristic failure mode in OOD settings: full externalization causes *context-level* interference (fixable in principle by reducing context), full internalization causes *parameter-level* conflict (unfixable at inference time), and indiscriminate lifecycle management causes *temporal* degradation (irreversible once retired). Skill0.5 eliminates all three failure modes by design through its type-differentiated treatment: permanently internalizing domain-agnostic general skills while maintaining faithful utilization of dynamic task-specific skills.

F Detailed Hyperparameters

Hyperparameter	Value
Policy Optimization Backbone	GRPO
Learning Rate	1e-6
KL Regularization Coeff.	0.01
Entropy Bonus Coeff.	0.001
Invalid Action Penalty Coeff.	0.1
GRPO Group Size	8
Batch size	16
Mini-Batch size	128
Token-level Top-k for JSD Loss	64
Maximum Prompt Token Length	6000
Maximum Response Token Length	768
Evaluation Sampling Temperature	0.4

Table 5: Detailed configuration of training hyperparameters.

G The Use of LLMs

LLMs were employed exclusively for grammatical correction, linguistic polishing, and readability enhancement during the drafting of this paper. All substantive intellectual content, scientific interpretations, and conclusions were generated solely by the human authors.

Case 1: SkillRL (Full Externalization) — "Put a hot potato in fridge" [FAIL, 30 steps]

Injected OOD Skill: *"Direct Post-Heat Placement: After heating, navigate straight to the destination and place the object once."*

Key Trajectory Segment:

`Step 5: heat potato 2 with microwave 1`	✓ Heated
`Step 6: go to fridge 1`	✓ Correct destination
`Step 7: cool potato 2 with fridge 1`	✗ Contradicts goal!
`think: "need to cool it before placing..."`	← ID habit: fridge → cool

Skill0.5 — Same Task [SUCCESS, 7 steps]

`Step 0: go to countertop` → `Step 1: take potato` → `Step 2: go to microwave` → `Step 3: heat potato` → `Step 4: go to fridge` → `Step 5: open fridge` → `Step 6: move potato to fridge` ✓
`think: "heated the potato... next step is to place the hot potato in the fridge."` No "cool" interference

Case 2: SKILL0 (Full Internalization) — "Examine the bowl with the desklamp" [FAIL, 20 steps]

Injected OOD Skill: *"Switch Lamp On: Issue 'use desklamp' as soon as you reach it."*

Key Trajectory Segment:

`Step 1: take bowl 1 from desk 1`	✓ Bowl in hand, desklamp visible
`Step 2: go to sidetable 1`	✗ Leaves desklamp location!
`Step 3: move bowl 1 to sidetable 1`	✗ Pick&Place template activated
`Step 5: take desklamp 1 from desk 1`	✗ Tries to pick up fixture
`Step 6-8: take desklamp` ×3 (repeated)	← Policy collapse

Skill0.5 — Same Task [SUCCESS, 4 steps]

`Step 0: go to desk` → `Step 1: take bowl` → `Step 2: go to desk` → `Step 3: use desklamp 1` ✓
`think: "take the bowl and use the desklamp to examine it."` Directly follows "Switch Lamp On"

Case 3: SLIM (Lifecycle Management) — "Look at pillow under desklamp" [Step 85: PASS → Step 120: FAIL]

Retired General Skill: *gen_001 "Systematic Exploration" (retired at step 5, utility ≈ 0.001)*

Step 85 (Success): `think: "go to desk 1 to check if there is a desklamp"` ✓

Step 120 (Collapse):

`think: "go to fridge or sink to find a cd"`	✗ Task hallucination (task is "pillow")
`action: go to bed 1`	✗ Think-action decoupling

Same degenerate pattern across 5 different tasks at step 120.

Skill0.5 — Examine Task at Step 120 [SUCCESS, 3 steps]

`Step 0: go to dresser` → `Step 1: take keychain` → `Step 2: use desklamp 1` ✓
`think: "find the desklamp and use it to examine the keychain under its light."` Stable reasoning at step 120

Figure 4: Trajectory comparisons on ALFWorld OOD tasks. Skill0.5 succeeds in all cases by internalizing general reasoning while faithfully utilizing novel OOD-specific skills.