Learning Agentic Policy from Action Guidance

Yuxiang Ji[1,2]* **Zengbin Wang**[2]* **Yong Wang**[2]†‡ **Shidong Yang**[2] **Ziyu Ma**[2]
Guanhua Chen[3] **Zonghua Sun**[1] **Liaoni Wu**[1]‡ **Xiangxiang Chu**[2]
[1]Xiamen University [2]AMAP, Alibaba Group [3]Southern University of Science and Technology

GitHub: https://github.com/Yux1angJi/ActGuide-RL

Abstract

Agentic reinforcement learning (RL) for Large Language Models (LLMs) critically depends on the exploration capability of the base policy, as training signals emerge only within its in-capability region. For tasks where the base policy cannot reach reward states, additional training or external guidance is needed to recover effective learning signals. Rather than relying on costly iterative supervised fine tuning (SFT), we exploit the abundant action data generated in everyday human interactions. We propose ACTGUIDE-RL, which injects action data as plan-style reference guidance, enabling the agentic policy to overcome reachability barriers to reward states. Guided and unguided rollouts are then jointly optimized via mixed-policy training, internalizing the exploration gains back into the unguided policy. Motivated by a theoretical and empirical analysis of the benefit-risk trade-off, we adopt a minimal intervention principle that invokes guidance only as an adaptive fallback, matching task difficulty while minimizing off-policy risk. On search-agent benchmarks, ACTGUIDE-RL substantially improves over zero RL (+10.7 pp on GAIA and +19 pp on XBench with Qwen3-4B), and performs on par with the SFT+RL pipeline without any cold start. This suggests a new paradigm for agentic RL that reduces the reliance on heavy SFT data by using scalable action guidance instead.

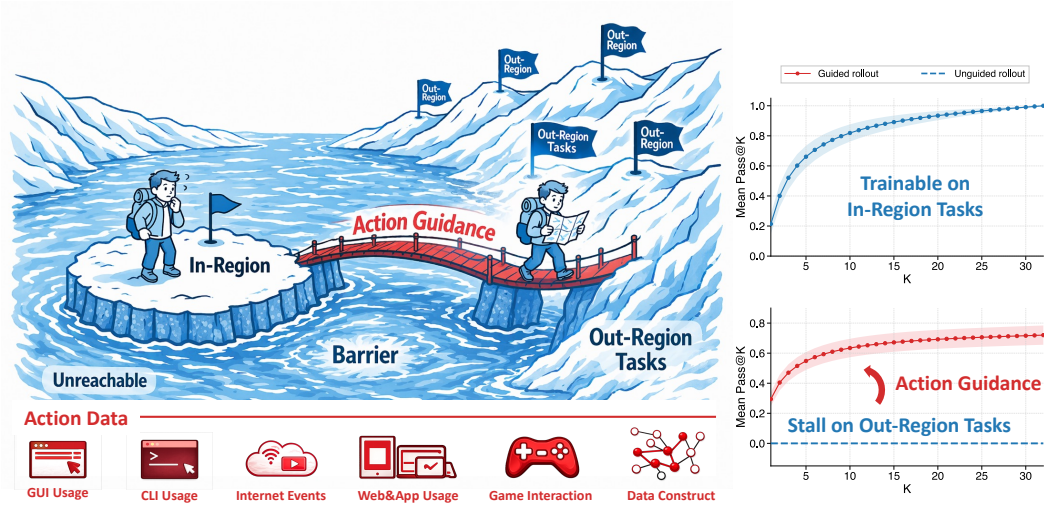

Figure 1: Agentic RL is typically confined to the *in-capability region*[1] of the base policy, and stalls on *out-region* tasks beyond this exploration frontier. ACTGUIDE-RL leverages diverse and scalable **action data** as plan-style reference to guide effective state visitation in *out-region* tasks.

*Equal contribution. Work done during internship at AMAP, Alibaba Group. †Project lead. ‡Corresponding author.
[1]*In-region* is where reward signals are reachable during rollout (pass@K > 0).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

The role of Large Language Models (LLMs) has shifted from simple chatbots to agents capable of independently solving complex tasks [36, 39, 63, 72, 73]. With targeted agentic training, recent frontier models can autonomously plan and accomplish a wide range of complex tasks [1, 44, 53]. This ability has been demonstrated in general tool-use [2, 13, 25], GUI [46, 66, 84], and CLI [27] settings, including in-the-wild real-world scenarios [12, 61]. A key factor behind such targeted training is agentic reinforcement learning (RL), in which LLM-based policies are optimized through repeated interaction with specific or diverse environments toward verifiable or heuristic rewards [24, 62, 80].

Unlike static supervised training, online RL is highly sensitive to task difficulty because the training signal comes only from exploration by the model itself. As Figure 1, we refer to tasks within the reachable capability of the base policy as *in-region*, and those beyond this boundary as *out-region*. When reward states fall into the out-region, group-based advantage estimates collapse to zero gradient, causing training to stall. As a result, a common view is that current RL-based methods are fundamentally limited by the capabilities of the base model [9, 22, 65, 71, 78, 87]. To address the cold-start problem of RL on difficult or unseen tasks, a typical practice is to perform corresponding Supervised Fine-Tuning (SFT) followed by dynamic difficulty adjustment or curriculum learning. However, such pipelines shift the burden to warm-start data and careful curriculum design. This dependence makes agentic RL complex and difficult to scale to new environments.

Stepping back to the original motivation for developing agentic capabilities, the goal is to move beyond reasoning and enable models to act, interact, and make decisions in a human-like manner to accomplish long-horizon tasks. From this perspective, a direct and currently underutilized training source is the abundant *action data* generated in open-world settings or during task construction. As shown in Figure 1, examples include step-by-step GUI/CLI interactions with computers or phones, API-mediated task execution, and long-horizon gameplay. In addition, some agentic RL tasks are constructed through a reverse process [15, 27, 29], where a valid action trajectory is first constructed and then used to instantiate the task, making the correct actions naturally available. These action data are inherently diverse and large in scale, yet their direct use for model training is often limited by the absence of explicit reasoning traces. Existing approaches either augment such data with synthesized chain-of-thought [16, 70] or directly leverage it through behavior imitation [3, 10]. However, synthesized reasoning can suffer from post-hoc rationalization [57], while behavior imitation tends to fit surface action patterns rather than induce the reasoning abilities of agentic policy.

In this work, we investigate how to leverage action data to enhance agentic RL. Through empirical analysis, we first characterize the capability barrier of agentic policies, where reward states fall outside the current reachable region and training signals become unavailable. To address this issue, we propose ACTGUIDE-RL, which injects action data as plan-style reference guidance to help the policy cross such barriers and perform effective out-region state visitation. We further analyze the benefit-risk trade-off introduced by guidance, where stronger guidance improves exploration but also increases off-policy distribution shift. Based on this, we draw two main conclusions from our experiments: **C1**: Action guidance works best when it serves as a zero-reward fallback and is minimized adaptively, following a principle of minimal intervention. **C2**: Under such minimal intervention, guided rollouts can be directly internalized into the unguided model through a mixed-policy optimization paradigm.

We evaluate ACTGUIDE-RL on search-agent benchmarks across different base models, task difficulty levels, and both in-domain and out-of-domain settings. Compared with zero RL, ACTGUIDE-RL consistently improves all tested base models, with especially large gains on harder benchmarks where unguided RL struggles to obtain effective training signals. Specifically, based on Qwen3-4B-Instruct, ACTGUIDE-RL improves zero RL by +10.68 pp on GAIA, +27.79 pp on WebWalkerQA, +19.00 pp on XBench, and +5.15 pp on BC-ZH. Notably, it also performs on par with the SFT+RL pipeline even without any cold-start initialization. This substantially alleviates the dependence on SFT and offers a new perspective for agentic post-training.

2 Method

2.1 Preliminaries: Agentic RL

We follow existing works to formulate Agentic RL as a Partially Observable Markov Decision Process (POMDP), where a language model acts as a policy π_θ. Given a task instance $x \sim \mathcal{D}$, the policy

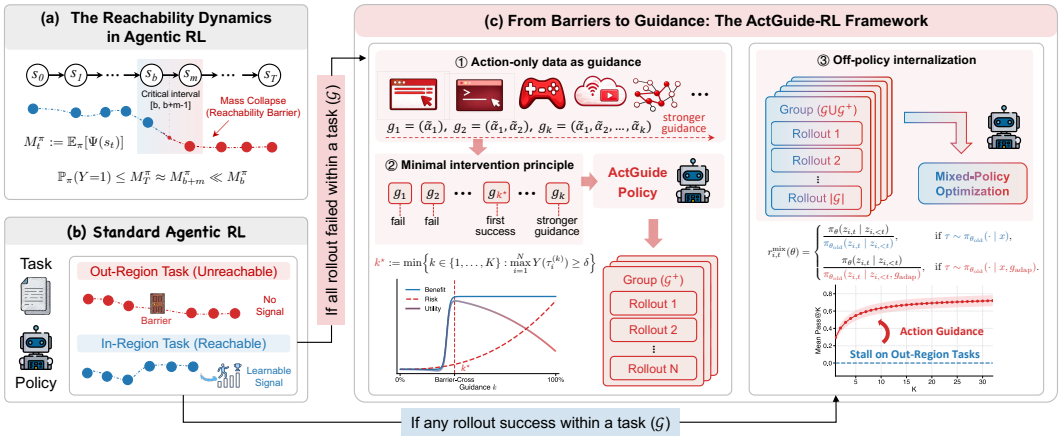

Figure 2: **Overview of ACTGUIDE-RL framework.** Conventional agentic RL can only obtain training signals within the base model in-capability region. ACTGUIDE-RL follows the principle of minimal intervention, dynamically introducing action data to guide the model toward out-region exploration. Such mixed rollouts are trained through mixed-policy optimization.

receives the interaction history as its state s_t at each step t, and predicts the next step $\alpha_t \sim \pi_\theta(\cdot \mid s_t)$. A full rollout yields a trajectory τ with a binary outcome reward $Y(\tau) \in \{0, 1\}$ indicating whether the task is successfully solved. The overall training objective is to maximize the expected reward:

$$\max_\theta \quad \mathcal{J}(\theta) \;:=\; \mathbb{E}_{x\sim\mathcal{D}}\,\mathbb{E}_{\tau\sim\pi_\theta(\cdot|x)}[\,Y(\tau)\,]\,. \tag{1}$$

Since $Y(\tau)$ is binary, this naturally amounts to maximizing the expected success rate over a task distribution that may contain tasks of *varying difficulty*.

2.2 The Reachability Barrier in Agentic RL

To optimize the above objective, recent RL algorithms [7, 50, 77] often sample a group of N rollout trajectories $\{\tau_i\}_{i=1}^N$ per task and compute advantages from the contrast between successful and failed ones. This mechanism works well when reward states lie within the in-capability region. However, when reward states fall into the out-region and become unreachable, no learning signal is obtained. We formalize this phenomenon through the concept of ***reachability dynamics***.

Definition 2.1 (Reachability Dynamics). *Let $\Psi(s) := \sup_\pi \mathbb{P}_\pi(Y{=}1 \mid s)$ denote the least upper bound on the success probability achievable by any continuation policy from state s. We define the effective state-visiting mass*

$$M_t^\pi := \mathbb{E}_\pi[\Psi(s_t)], \tag{2}$$

which measures the average remaining success potential along rollouts induced by policy π. The ratio $\bar\kappa_t^\pi := M_{t+1}^\pi / M_t^\pi$ quantifies the one-step reachability retention. By telescoping, the mass over any interval $[u, v)$ satisfies the multiplicative recursion

$$M_v^\pi = M_u^\pi \prod_{t=u}^{v-1} \bar\kappa_t^\pi. \tag{3}$$

Remark 2.1 (Mass Collapse as Reachability Barrier). *Since $\Psi(s)$ upper-bounds the success probability achievable from s under any policy, the terminal success probability satisfies*

$$\mathbb{P}_\pi(Y{=}1) = \mathbb{E}_\pi[\mathbb{P}_\pi(Y{=}1 \mid s_T)] \le \mathbb{E}_\pi[\Psi(s_T)] = M_T^\pi. \tag{4}$$

*Suppose a short critical interval $[b, b{+}m{-}1]$ exhibits low cumulative retention, i.e., $\prod_{t=b}^{b+m-1} \bar\kappa_t^\pi \ll 1$, so that $M_{b+m}^\pi \ll M_b^\pi$. Once such a sharp drop in reachability mass occurs, the remaining rollout tends to stay in low-reachability regions, so the terminal mass remains close to the post-collapse level, i.e., $M_T^\pi \approx M_{b+m}^\pi$. We call such a regime a **reachability barrier**.*

A reachability barrier makes rollouts beyond step $b{+}m$ receive $Y(\tau){=}0$, collapsing the group-based advantage to zero gradient. *This confines the model to in-region training and prevents learning on out-region tasks.* Unlike insufficient sampling, this failure is structural, so increasing N cannot help. The policy itself must first be steered across the critical interval, which motivates our method below.

2.3 From Barriers to Guidance: The ACTGUIDE-RL Framework

To address the fundamental barrier in agentic RL, we propose ACTGUIDE-RL to use action as guidance, illustrated in Figure 2. ACTGUIDE-RL is driven by three core questions along with two empirical findings: whether action data can repair reachability barriers (§2.3.1, Finding 1), how much guidance to inject (§2.3.2, Finding 2), and how to optimize from guided samples (§2.3.3).

2.3.1 How to Guide: Action Data Repairs Barriers

To explore whether action-only data can repair reachability barriers, we treat the action trajectory as a reference plan $g = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_L)$ and condition the policy as $\pi_\theta(\cdot \mid s, g)$. We then compare the guided and unguided behavior along the guided rollout. Specifically, we measure:

$$\underbrace{|\Delta\text{Logit}| = \left|\text{logit}_{\pi_\theta}(\cdot \mid s_t, g) - \text{logit}_{\pi_\theta}(\cdot \mid s_t)\right|}_{\text{token-level guidance influence}}, \quad \underbrace{\text{Pass@K} = \mathbb{P}_{\tau_{1:K} \sim \pi_\theta(\cdot|s_t)}\left[\max_{i \le K} Y(\tau_i) = 1\right]}_{\text{prefix-level reachability}} \tag{5}$$

where $|\Delta\text{Logit}|$ computes the token logits difference between the guided policy $\pi_\theta(\cdot \mid s, g)$ and the unguided policy $\pi_\theta(\cdot \mid s)$, capturing how much guidance changes the policy locally. Prefix-level Pass@K instead samples unguided continuations from the current guided state s_t and measures whether they can recover reward, reflecting the remaining reachability after that state.

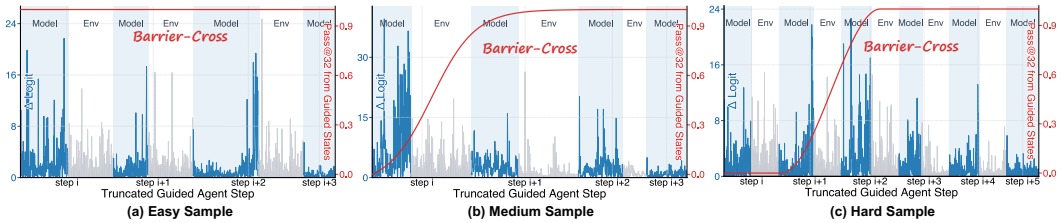

Figure 3: Action guidance repairs reachability barriers along guided rollouts. Blue bars show $|\Delta\text{Logit}|$, and red curves show prefix-level Pass@K (K=32). Barriers emerge where unguided Pass@K collapses and the guidance-induced logit shift spikes.

Finding 1: Action guidance repairs reachability barriers. As shown in Figure 3, easy tasks[2] already show non-zero Pass@K from early guided states, while harder tasks[3] keep zero unguided Pass@K until the guided trajectory crosses the barrier. Within these barrier intervals, $|\Delta\text{Logit}|$ spikes sharply, showing that action trajectories diverge from the current policy exactly where it fails. After the barrier is crossed, unguided Pass@K recovers to non-trivial levels, showing that action guidance brings the policy to reachable reward states rather than simply replacing its decisions.

Motivated by Finding 1, we formally leverage action data (g) as the effective guidance signal and simply append it to the task prompt as a list of future reference actions (Appendix B). This provides a non-intrusive reference plan, rather than forcing the model to generate the actions as a fixed prefix. Moreover, recognizing that different barriers may require varying amounts of guidance to cross, we organize guidance into an ordered family

$$g_0 = \varnothing \prec g_1 \prec \cdots \prec g_K, \tag{6}$$

where $g_k = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_k)$ provides the first k reference actions. This gives guidance a monotone strength parameter, which later allows us to search for the minimal sufficient intervention. For a barrier interval $[b, b+m-1]$ of the base policy $\pi_\theta(\alpha_t \mid s_t)$, we measure the *barrier-repair benefit* of guidance level g_k by the increase of effective state-visiting mass after the barrier:

$$B_k := \log \frac{M_{b+m}^{\pi_\theta(\cdot|s,g_k)}}{M_{b+m}^{\pi_\theta(\cdot|s)}}, \tag{7}$$

where a larger B_k implies that the guidance better preserves reachable success potential.

[2]Easy samples: the model can discover reward from early guided states.

[3]Harder samples: rewarding states become reachable only at much later guided states.

2.3.2 How Much to Guide: Minimal Intervention Principle

While stronger guidance raises the barrier-repair benefit B_k (Eq. 7), it also induces a larger distribution shift from the base policy, increasing the risk of off-policy optimization error [58, 83]. Let $\tau = (z_1, \ldots, z_{|\tau|})$ be the generated token sequence. To quantify the distribution shift under guidance level g_k, we measure the cumulative token-level log-ratio shift of a rollout τ:

$$\mathcal{L}_k(\tau) := \sum_{j=1}^{|\tau|} \log \frac{\pi_\theta(z_j \mid z_{<j})}{\pi_\theta(z_j \mid z_{<j}, g_k)}. \tag{8}$$

The corresponding *off-policy risk* is the variance of this shift:

$$R_k := \mathrm{Var}_{\tau \sim \pi_\theta(\cdot|s,g_k)}(\mathcal{L}_k(\tau)). \tag{9}$$

Finding 2: Over-guidance inflates off-policy risk. As shown in Figure 4, the mean log-ratio shift (blue) and its variance (red) describe the guidance-induced distribution shift from complementary perspectives. As the guidance level k increases, the off-policy risk R_k keeps rising, indicating that stronger guidance makes guided rollouts increasingly unstable for off-policy optimization.

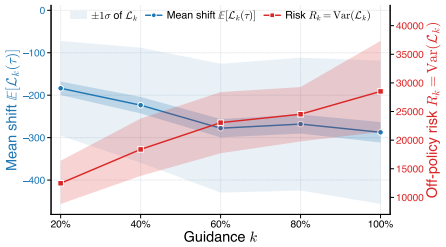

Figure 4: Guidance-induced distribution shift under different guidance proportions. The blue curve shows the mean log-ratio shift, while the red curve shows its variance, *i.e.*, the off-policy risk R_k.

Motivated by Finding 2, we adopt a *minimal intervention principle*: for each task, use the least guidance level that recovers reward. This principle can be viewed as approximately maximizing a guidance utility $J_k = B_k - \lambda R_k$, where the barrier-repair benefit B_k exhibits a sharp increase after reward recovery while the off-policy risk R_k grows with the guidance level. In practice, we first collect an unguided rollout group per task, invoking guidance only as a fallback when the entire group fails. Under a mild monotonicity assumption (stronger levels do not decrease recovery probability), we can efficiently identify the smallest sufficient level $k^\star$ via binary search:

$$k^\star := \min\Big\{k \in \{1, \ldots, K\} : \max_{i=1}^{N} Y(\tau_i^{(k)}) \geq \delta\Big\}, \tag{10}$$

where $\{\tau_i^{(k)}\}_{i=1}^N$ are N rollouts under guidance g_k and $\delta > 0$ is the success threshold. We denote the resulting adaptive guidance as $g_{\text{adap}} := g_{k^\star}$, which keeps guided rollouts close to the unguided distribution and enables the off-policy optimization studied next. Note that under binary rewards, B_k exhibits threshold behavior (near zero until the barrier is crossed, then jumping sharply), while R_k grows monotonically. The guidance utility J_k therefore peaks near the minimal successful level, making the binary search in Eq. 10 a practical proxy for approximately maximizing J_k.

2.3.3 How to Learn: Off-Policy Internalization

Action guidance is available only at training time. At inference, the agent must act under the unguided policy $\pi_\theta(\cdot \mid x)$, so any learning signal extracted from guided rollouts has to be internalized. Since the guided policy $\pi_\theta(\cdot \mid x, g)$ shares parameters with the unguided one, we treat guided samples as off-policy data w.r.t. $\pi_\theta(\cdot \mid x)$ and optimize the mixed objective

$$\begin{aligned}\mathcal{J}_{\text{mix}}(\theta) = \mathbb{E}_{(x,\bar{g})\sim\mathcal{D},\ \mathcal{G}\sim q^{\text{mix}}_{\theta_{\text{old}}}(\cdot|x,\bar{g})}\Bigg[\frac{1}{\sum_i T_i}\sum_{i=1}^{|\mathcal{G}|}\sum_{t=1}^{T_i}\min\Big(r^{\text{mix}}_{i,t}(\theta)\,\hat{A}(\tau_i),\\ \text{clip}\Big(r^{\text{mix}}_{i,t}(\theta), 1-\epsilon, 1+\epsilon\Big)\,\hat{A}(\tau_i)\Big) - \beta\frac{1}{|\mathcal{G}|}\sum_{i=1}^{|\mathcal{G}|}\mathbb{D}_{\text{KL}}(\pi_\theta(\tau_i \mid x)\,\|\,\pi_{\text{ref}}(\tau_i \mid x))\Bigg],\end{aligned} \tag{11}$$

where $q^{\text{mix}}_{\theta_{\text{old}}}$ denotes the mixed rollout collection process in Algorithm 1, $\hat{A}(\tau_i)$ is the group-based advantage, and the token-level importance ratio adapts to the rollout source:

$$r^{\text{mix}}_{i,t}(\theta) = \begin{cases} \dfrac{\pi_\theta(z_{i,t} \mid z_{i,<t})}{\pi_{\theta_{\text{old}}}(z_{i,t} \mid z_{i,<t})}, & \text{if } \tau_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x), \\ \dfrac{\pi_\theta(z_{i,t} \mid z_{i,<t})}{\pi_{\theta_{\text{old}}}(z_{i,t} \mid z_{i,<t}, g_{\text{adap}})}, & \text{if } \tau_i \sim \pi_{\theta_{\text{old}}}(\cdot \mid x, g_{\text{adap}}). \end{cases} \tag{12}$$

Algorithm 1: Adaptive Minimal-Intervention Training with Action Guidance

Input: policy π_θ, dataset $\mathcal{D} = \{(x, \bar{g})\}$, minibatch size M, group size N, threshold δ, search budget B, steps S

for $s = 1$ **to** S **do**

  Sample $\mathcal{B} = \{(x_b, \bar{g}_b)\}_{b=1}^{M} \sim \mathcal{D}$; $\mathcal{G} \leftarrow \emptyset$;

  foreach $(x_b, \bar{g}_b) \in \mathcal{B}$ **do**

    Define $g_{b,k} = (\tilde{\alpha}_{b,1}, \ldots, \tilde{\alpha}_{b,k})$;

    $\mathcal{G}_b \leftarrow \{(\tau_{b,i}, r_{b,i})\}_{i=1}^{N},\ \tau_{b,i} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_b),\ r_{b,i} = Y(\tau_{b,i})$;

    if $\max_i r_{b,i} < \delta$ **then**

      $k_b^\star \leftarrow \min\left\{k : \max_j r_{b,j}^{(k)} \geq \delta\right\}$ via binary search under budget B;

      if $k_b^\star$ *exists* **then**

        $\mathcal{G}_b^+ \leftarrow \{(\tau_{b,i}^+, r_{b,i}^+)\}_{i=1}^{N},\ \tau_{b,i}^+ \sim \pi_{\theta_{\text{old}}}(\cdot \mid x_b, g_{b,k_b^\star}),\ r_{b,i}^+ = Y(\tau_{b,i}^+)$;

        $\mathcal{G}_b \leftarrow \mathcal{G}_b \cup \mathcal{G}_b^+$;

    $\mathcal{G} \leftarrow \mathcal{G} \cup \mathcal{G}_b$;

  Compute advantages on $\mathcal{G}$; Update $\pi_\theta(\cdot \mid x)$ by $\mathcal{J}_{\text{mix}}$;

For unguided rollouts this is the standard importance ratio; for guided rollouts the denominator uses the guided distribution, transferring credit back to the unguided target $\pi_\theta(\cdot \mid x)$. Unlike prior off-policy RL methods that include ratio shaping [43, 69], we keep the optimization objective unchanged because minimal intervention limits the shift between guided rollouts and the base policy.

3 Experiment

3.1 Experimental Setup

Benchmarks. To evaluate the effectiveness of our proposed ACTGUIDE-RL in LLM agentic RL, we conduct experiments in the search-agent setting, which is stateless and facilitates the collection of action data. Our evaluation covers two categories of benchmarks. The first category is in-domain search-agent benchmarks, including four representative datasets, *GAIA* [40], *WebWalkerQA* [64], *XBench* [5], and *BrowseComp-ZH (BC-ZH)* [86], which span diverse difficulty levels, multiple languages, and real-world multi-step reasoning scenarios. The second category is out-of-domain benchmarks, including *GPQA* [48], *TruthfulQA* [34], and *IFEval* [85], which are used to evaluate the out-of-domain generalization ability of models beyond the search-agent setting. The detailed RL and SFT training data source are provided in Appendix A.

Baselines. Under the same evaluation protocol, we compare ACTGUIDE-RL against several baselines, including foundation models [35, 41, 52], specified search-agent-trained models [14, 29, 31], and vanilla RL trained from the same backbones without action guidance. For the RL baseline, we adopt the standard GRPO objective with token-level policy optimization, using the same training data but without action guidance.

Implementation Details. Following Tongyi-DeepResearch [55], we equip the agent with two tools, *web-search* and *web-visit*, whose schemas are included in the system prompt. Given the limited interaction budget and context length in our setup, we use raw tool outputs directly without a separate summary model. For both training reward and test-time evaluation, we adopt the few-shot, reference-based binary LLM-judge template from Tongyi-DeepResearch. Full implementation details are provided in Appendix B.

3.2 Main Results

Overall Comparison. Table 1 reports overall accuracy on four in-domain benchmarks, from which three observations stand out.

- **ACTGUIDE-RL mitigates in-region RL capability regression.** When the exploration difficulty of the RL training data does not match the base model, vanilla RL restricted to in-region exploration can lead to partial performance regression on some benchmarks. For example, RL

Table 1: Main results of ACTGUIDE-RL on search-agent benchmarks comparing foundation models, search-agent trained models, and RL baseline. The best results are indicated in **bold**.

Method	General AI Assistant				WebWalkerQA				XBench	BC-ZH
	Lv.1	Lv.2	Lv.3	Avg.	Easy	Med.	Hard	Avg.	Avg.	Avg.
Foundation Model										
MiniMax-M2.1	-	-	-	64.3	-	-	-	-	68.0	66.6
DeepSeek-V3.2	-	-	-	75.1	-	-	-	-	78.0	65.0
GPT-5 High	-	-	-	76.4	-	-	-	-	77.0	65.0
Search-Agent-Trained Models										
WebSailor-7B	-	-	-	37.9	-	-	-	-	34.0	14.2
ARPO-8B	53.9	32.7	16.7	38.8	26.7	33.3	29.6	30.5	25.0	-
WebThinker-32B-RL	56.4	50.0	16.7	48.5	58.8	44.6	40.4	46.5	24.0	7.3
***Baseline and* ACTGUIDE-RL**										
Qwen2.5-3B-Instruct	15.38	7.69	0.00	9.71	5.00	7.14	4.58	5.73	8.00	2.08
+ RL	15.38	7.69	16.66	11.65	15.00	15.00	15.83	15.29	10.00	2.42
+ **ACTGUIDE-RL**	**28.21**	**11.54**	**16.66**	**18.45**	**18.75**	**16.07**	**22.08**	**18.82**	**16.00**	**4.50**
ΔDelta	+12.83	+3.85	+0.00	+6.80	+3.75	+1.07	+6.25	+3.53	+6.00	+2.08
Qwen2.5-7B-Instruct	35.89	15.38	8.33	22.32	18.75	19.28	16.25	18.09	19.00	4.50
+ RL	20.51	7.69	0.00	11.65	14.37	20.35	19.58	18.67	22.00	4.84
+ **ACTGUIDE-RL**	**41.02**	**17.30**	**8.33**	**25.24**	**24.37**	**21.07**	**21.66**	**22.05**	**24.00**	**8.31**
ΔDelta	+20.51	+9.61	+8.33	+13.59	+10.00	+0.72	+2.08	+3.38	+2.00	+3.47
Qwen3-4B-Instruct	17.94	17.30	0.00	15.53	8.75	3.57	0.83	3.82	14.00	7.96
+ RL	33.33	25.00	0.00	25.24	13.12	13.92	9.17	12.06	18.00	15.26
+ **ACTGUIDE-RL**	**46.15**	**32.69**	**16.66**	**35.92**	**43.75**	**41.78**	**35.00**	**39.85**	**37.00**	**20.41**
ΔDelta	+12.82	+7.69	+16.66	+10.68	+22.50	+30.63	+25.83	+27.79	+19.00	+5.15
Qwen3-8B	43.58	26.92	16.66	32.03	41.87	31.78	26.25	32.20	32.00	23.52
+ RL	46.15	32.69	25.00	36.89	43.75	44.64	39.16	42.50	33.00	21.79
+ **ACTGUIDE-RL**	**51.28**	**36.53**	**33.33**	**41.74**	**50.00**	**46.79**	**44.58**	**46.77**	**44.00**	**26.64**
ΔDelta	+5.13	+3.84	+8.33	+4.85	+6.25	+2.15	+5.42	+4.27	+11.00	+4.85

degrades Qwen2.5-7B-Instruct on GAIA and Qwen3-8B on BC-ZH, whereas ACTGUIDE-RL alleviates these regressions through adaptive guidance and more effective state visitation.

- **ACTGUIDE-RL improves exploration beyond the current reachable region.** When vanilla RL fails to access sufficiently effective states on harder tasks, action guidance helps the policy move beyond its current reachable region and enables more effective state visitation. This is most evident on Qwen3-4B-Instruct, where ACTGUIDE-RL brings broad gains across all four benchmarks, with especially large improvements on WebWalker ($12.06\% \rightarrow 39.85\%$) and XBench ($18.00\% \rightarrow 37.00\%$).
- **ACTGUIDE-RL delivers stable gains across base models.** For base models with different capability levels, action guidance can adaptively help the policy access more effective states on each training sample according to its difficulty. As a result, compared with vanilla RL, ACTGUIDE-RL consistently improves all four base models, underscoring the strong adaptability of action guidance across different capability levels.

Comparison with SFT + RL. Another commonly used strategy to address training stalls caused by limited policy exploration is a targeted SFT cold start. To further analyze the role of ACTGUIDE-RL relative to the SFT + RL paradigm, we also initialize the policy with an SFT cold start constructed by partially distilling Tongyi-DeepResearch-30B-A3b. This setting aims to explore a new possibility beyond the standard SFT + RL pipeline through action-level guidance, rather than merely pursuing performance improvements over a comprehensive SFT baseline. As shown in Table 2, even without any cold start, ACTGUIDE-RL achieves performance comparable to the two-stage SFT+RL pipeline. Moreover, when built on the same cold-start initialized model, ACTGUIDE-RL still can obtain additional gains from action guidance. Meanwhile, due to the mode-covering nature of SFT, cold-start initialization often degrades out-of-domain performance as the consistent performance drop on GPQA-CoT (Zero Shot), TruthfulQA and IFEVAL, whereas such degradation does not occur in ACTGUIDE-RL with zero RL setting.

Overall, ACTGUIDE-RL offers a new alternative paradigm for agentic RL, alleviating the dependence on heavy SFT data throught the use of lighter-weight action data instead.

Table 2: Comparison of ACTGUIDE-RL and SFT + RL on in-domain and out-of-domain benchmarks.

Method	In-Domain				Out-of-Domain		
	GAIA	WebWalker	XBench	BC-ZH	GPQA-CoT (ZS)	TruthQA	IFEval
ZeroRL	25.24	12.06	18.00	15.26	35.45	62.17	81.33
+ **ACTGUIDE**	**35.92**	**39.85**	**37.00**	**20.41**	**36.93**	**62.30**	**82.99**
SFT	34.95	31.18	25.00	25.61	29.15	56.95	**77.82**
+ RL	36.89	32.20	17.00	26.30	**29.85**	57.02	76.34
+ **ACTGUIDE**	**40.77**	**37.06**	**25.00**	**28.02**	29.57	**57.11**	77.43

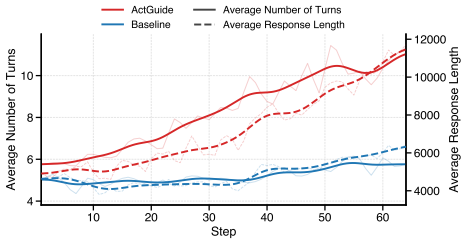

Figure 6: Training dynamics on number of interaction turns and response length.

Table 3: Agent performance under different interaction turn budgets.

Turn Budget	GAIA	WebWalker	XBench	BC-ZH
2	0.97	9.26	5.00	1.04
4	18.44	33.97	33.00	4.84
8	19.41	35.00	33.00	16.96
16	27.18	37.55	35.00	17.99
32	35.92	39.85	37.00	20.41

3.3 Further Analysis and Ablation

Training Dynamics. To further analyze the effect of action guidance on training dynamics, we track the proportion of rollout groups that provide effective learning signals during training, as shown in Figure 5. Specifically, we find action data helps the policy discover effective training signals in a higher proportion of samples, while the unguided baseline is frequently hindered by exploration barriers and therefore wastes many rollouts on ineffective state visitation. This suggests that ACTGUIDE-RL improves exploration beyond the current reachable region, allowing the policy to learn from out-region tasks.

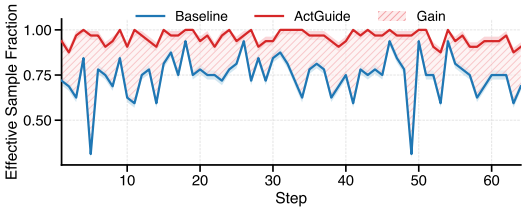

Figure 5: Trainable groups dynamic.

Towards Complex Interaction. A central challenge of agentic RL without cold-start is that the policy struggles to acquire complex interaction skills within its in-region tasks. Fortunately, we find that ACTGUIDE-RL enables even a small model such as Qwen3-4B-Instruct without any cold-start initialization, to gradually acquire complex interaction capability, as reflected by the steady increase in the number of interaction turns and generated tokens over training in Figure 6. To further verify whether these increased interactions are indeed effective, we vary the interaction budget at evaluation time and observe that performance consistently improves as the budget increases in Table 3.

Ablation Study on ACTGUIDE-RL. We conduct ablation studies on several key design choices in ACTGUIDE-RL, including the adaptive guidance mechanism, the fallback guidance, and mixed-policy optimization. As shown in Table 4, removing either the adaptive or fallback guidance mechanism causes performance degradation to different extents. We further compare fixed guidance ratios in Figure 7, and again find that dynamic guidance performs best. These results indicate that action guidance is not effective simply because more guidance is provided, nor is less always better. Rather, the best performance comes from minimally introducing guidance in an adaptive manner according to the policy capability.

Table 4: Ablation study of ACTGUIDE-RL.

Method	GAIA	WebWalker	XBench
ACTGUIDE-RL	35.92	39.85	37.00
– Minimal-Intervention (Adaptive)	27.18	35.00	34.00
– Minimal-Intervention (Fallback)	24.27	23.82	19.00
– Mixed-Policy Optimization	22.32	21.76	21.00

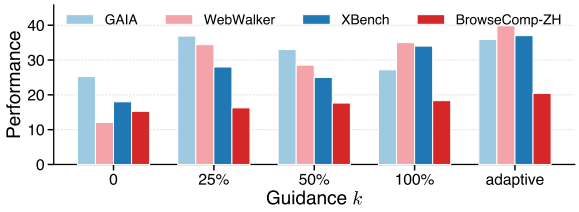

Figure 7: Performance of different guidance strength.

Removing mixed-policy optimization also causes a substantial performance drop, since it breaks the pathway that transfers behaviors acquired under guidance into the test-time unguided capability.

Sensitivity to Action Noise. When considering scaling up the collection of action data, an important factor is data noise, as human demonstrations may contain a substantial amount of meaningless or irrelevant actions while completing certain tasks. Here we simulate such noise by randomly inserting task-irrelevant actions into the original per-sample action trajectories, and then perform the same ACTGUIDE-RL training. As shown in Table 5, ACTGUIDE-RL is not overly sensitive to action noise. It maintains stable performance under a 10% noise ratio, while a further increase to 20% leads to a performance drop.

Table 5: Results of different action noise ratio.

Noise Ratio	GAIA	WebWalker	XBench	BC-ZH
0%	35.92	**39.85**	37.00	**20.41**
10%	**39.81**	39.26	**38.00**	19.03
20%	29.12	37.94	35.00	17.64

4 Related Work

4.1 Agentic RL

Recent advancements in RL [7, 17, 38, 49, 50, 67, 77] enable end-to-end training of agents that can interact with environments, make sequential decisions, and optimize toward long-horizon objectives. This makes agentic RL a pivotal paradigm for both foundation-model capability building [1, 44, 53, 54] and domain-specific agent post-training [8, 25, 28, 55, 76]. Since effective agentic RL strongly depends on the base model to explore valid training signals, existing methods often rely on a cold-start before RL or on alternating SFT and RL to dynamically align the model capabilities with the target tasks [4, 6, 11, 13, 45]. Some works instead adopt dynamic task scheduling [20, 74, 79] or curriculum learning [26, 30] to ensure that the difficulty of training tasks is well matched to the evolving capabilities of the model. A line of work most closely related to ours constructs curriculum learning examples from existing SFT data [60, 75], or directly uses this data as hints to guide the model toward obtaining meaningful learning signals on hard tasks [22, 43, 65]. Unlike these approaches, ACTGUIDE-RL seeks to leverage more readily available action data, offering greater practical value and stronger scaling potential.

4.2 RL from Demonstration

Our work is also related to reinforcement learning from demonstrations (RLfD) [33, 42], where demonstrations usually take the form of expert trajectories, typically as full reasoning-and-action traces in agent settings. Classical RLfD methods often use demonstration trajectories to bootstrap exploration in sparse-reward settings, for example by retaining them in the replay buffer and combining RL updates with auxiliary imitation losses [21, 47, 59]. Following a similar intuition, several recent LLM studies incorporate off-policy expert trajectories into online RL to mitigate sparse-reward and hard-exploration challenges [18, 32, 37, 81]. Specifically, LUFFY [68] incorporates off-policy expert trajectories into online RL through mixed-policy optimization, using regularized importance shaping to avoid rigid imitation. Guide [43] utilizes adaptive hint-guided off-policy trajectories into online RL, reweighting them to improve exploration while training a policy that no longer relies on hints at inference time. Unlike these demonstration-based approaches, this work focuses on learning agentic policy from action guidance, with minimal intervention that adopts to tasks of different difficulty.

5 Conclusion

We present ACTGUIDE-RL, a framework that leverages readily available action data as plan-style guidance to help agentic RL overcome exploration barriers beyond the base policy's reachable region. By introducing guidance only as an adaptive fallback and optimizing guided and unguided rollouts jointly, ACTGUIDE-RL internalizes exploration gains while reducing the off-policy risks of excessive intervention. Across search-agent benchmarks, these design choices yield consistent gains over vanilla RL and performance comparable to SFT+RL, without requiring a supervised cold start. Further analyses show that these gains are accompanied by more effective multi-step interaction and arise from adaptive, minimally intrusive guidance rather than simply stronger intervention. These findings suggest that scalable action-only traces can serve as a practical post-training signal for complex agentic interaction, complementing or partially replacing costly supervised demonstrations.

References

[1] Anthropic. Claude Opus 4.6 model card, 2026. URL https://www-cdn.anthropic.com/bf10f64990cfda0ba858290be7b8cc6317685f47.pdf.

[2] Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. tau2 Bench: Evaluating Conversational Agents in a Dual-Control Environment. *arXiv preprint arXiv:2506.07982*, 2025.

[3] Massimo Caccia, Megh Thakkar, Léo Boisvert, Thibault Le Sellier De Chezelles, Alexandre Piché, Nicolas Chapados, Alexandre Drouin, Maxime Gasse, and Alexandre Lacoste. Fine-tuning web agents: It works, but it's trickier than you think. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.

[4] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*, 2025.

[5] Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, Chen Sun, Han Hou, Hui Yang, James Pan, Jianan Lou, Jiayi Mao, Jizheng Liu, Jinpeng Li, Kangyi Liu, Kenkun Liu, Rui Wang, Run Li, Tong Niu, Wenlong Zhang, Wenqi Yan, Xuanzheng Wang, Yuchen Zhang, Yi-Hsin Hung, Yuan Jiang, Zexuan Liu, Zihan Yin, Zijian Ma, and Zhiwen Mo. xbench: Tracking Agents Productivity Scaling with Profession-Aligned Real-World Evaluations, June 2025. URL http://arxiv.org/abs/2506.13651. arXiv:2506.13651 [cs].

[6] Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. Beyond two-stage training: Cooperative sft and rl for llm reasoning. *arXiv preprint arXiv:2509.06948*, 2025.

[7] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=inccdtfx8x.

[8] Zheng Chu, Xiao Wang, Jack Hong, Huiming Fan, Yuqi Huang, Yue Yang, Guohai Xu, Chenxiao Zhao, Cheng Xiang, Shengchao Hu, et al. Redsearcher: A scalable and cost-efficient framework for long-horizon search agents. *arXiv preprint arXiv:2602.14234*, 2026.

[9] Yanqi Dai, Yuxiang Ji, Xiao Zhang, Yong Wang, Xiangxiang Chu, and Zhiwu Lu. Harder is better: Boosting mathematical reasoning via difficulty-aware GRPO and multi-aspect question reformulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=nfURupkdRJ.

[10] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

[11] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. *arXiv preprint arXiv:2503.17352*, 2025.

[12] Shuangrui Ding, Xuanlang Dai, Long Xing, Shengyuan Ding, Ziyu Liu, Jingyi Yang, Penghui Yang, Zhixiong Zhang, Xilin Wei, Yubo Ma, Haodong Duan, Jing Shao, Jiaqi Wang, Dahua Lin, Kai Chen, and Yuhang Zang. Wildclawbench. https://github.com/InternLM/WildClawBench, 2026. GitHub repository.

[13] Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. Tool-star: Empowering llm-brained multi-tool reasoner via reinforcement learning. *arXiv preprint arXiv:2505.16410*, 2025.

[14] Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. Agentic Reinforced Policy Optimization, July 2025. URL http://arxiv.org/abs/2507.19849. arXiv:2507.19849 [cs].

[15] Guanting Dong, Junting Lu, Junjie Huang, Wanjun Zhong, Longxiang Liu, Shijue Huang, Zhenyu Li, Yang Zhao, Xiaoshuai Song, Xiaoxi Li, et al. Agent-world: Scaling real-world environment synthesis for evolving general agent intelligence. *arXiv preprint arXiv:2604.18292*, 2026.

[16] Lutfi Eren Erdogan, Nicholas Lee, Sehoon Kim, Suhong Moon, Hiroki Furuta, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. Plan-and-act: Improving planning of agents for long-horizon tasks. *arXiv preprint arXiv:2503.09572*, 2025.

[17] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-Group Policy Optimization for LLM Agent Training, May 2025. URL http://arxiv.org/abs/2505.10978. arXiv:2505.10978 [cs].

[18] Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*, 2025.

[19] Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl, 2025. URL https://arxiv.org/abs/2508.07976.

[20] Zhengyao Gu, Jonathan Light, Raul Astudillo, Ziyu Ye, Langzhou He, Henry Peng Zou, Wei Cheng, Santiago Paternain, Philip S Yu, and Yisong Yue. Actor-curator: Co-adaptive curriculum learning via policy-improvement bandits for rl post-training. *arXiv preprint arXiv:2602.20532*, 2026.

[21] Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, et al. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[22] Qihan Huang, Weilong Dai, Jinlong Liu, Wanggui He, Hao Jiang, Mingli Song, Jingyuan Chen, Chang Yao, and Jie Song. Boosting mllm reasoning with text-debiased hint-grpo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4848–4857, 2025.

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

[24] Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoni Wu. Tree search for LLM agent reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ZpQwAFhU13.

[25] Yuxiang Ji, Yong Wang, Ziyu Ma, Yiming Hu, Hailang Huang, Xuecai Hu, Guanhua Chen, Liaoni Wu, and Xiangxiang Chu. Thinking with map: Reinforced parallel map-augmented agent for geolocalization. *ACL*, 2026.

[26] Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuewei Zhang, Guohua Liu, and Hao Wang. Vcrl: Variance-based curriculum reinforcement learning for large language models. *arXiv preprint arXiv:2509.19803*, 2025.

[27] 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.

[28] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning, April 2025. URL http://arxiv.org/abs/2503.09516. arXiv:2503.09516 [cs].

[29] Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. WebSailor: Navigating Super-human Reasoning for Web Agent, July 2025. URL http://arxiv.org/abs/2507.02592. arXiv:2507.02592 [cs].

[30] Renda Li, Hailang Huang, Fei Wei, Feng Xiong, Yong Wang, and Xiangxiang Chu. Adacurl: Adaptive curriculum reinforcement learning with invalid sample mitigation and historical revisiting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23123–23131, 2026.

[31] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. WebThinker: Empowering Large Reasoning Models with Deep Research Capability, April 2025. URL http://arxiv.org/abs/2504.21776. arXiv:2504.21776 [cs].

[32] Jing Liang, Hongyao Tang, Yi Ma, Jinyi Liu, Yan Zheng, Shuyue Hu, Lei Bai, and Jianye Hao. Squeeze the soaked sponge: Efficient off-policy reinforcement finetuning for large language model. *arXiv preprint arXiv:2507.06892*, 2025.

[33] Gabriele Libardi, Gianni De Fabritiis, and Sebastian Dittert. Guided exploration with proximal policy optimization using a single demonstration. In *International Conference on Machine Learning*, pages 6611–6620. PMLR, 2021.

[34] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022.

[35] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[36] 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.

[37] Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, et al. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*, 2025.

[38] Ziyu Ma, Hailang Huang, Shun Zou, Yong Wang, Shidong Yang, Yiming Hu, Fei Wei, and XiangXiang Chu. Longhorizon-harness: Advancing long-horizon agents for real-world tasks. *arXiv preprint arXiv:2608.01964*, 2026.

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

[40] Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: a benchmark for General AI Assistants, November 2023. URL http://arxiv.org/abs/2311.12983. arXiv:2311.12983 [cs].

[41] MiniMax. Minimax m2.1 system card, 2025. URL https://www.minimax.io/news/minimax-m21.

[42] Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6292–6299. IEEE, 2018.

[43] Vaskar Nath, Elaine Lau, Anisha Gunjal, Manasi Sharma, Nikhil Baharte, and Sean Hendryx. Adaptive guidance accelerates reinforcement learning of reasoning models. *arXiv preprint arXiv:2506.13923*, 2025.

[44] OpenAI. Gpt-5.4 thinking system card, 2025. URL https://openai.com/index/gpt-5-4-thinking-system-card/.

[45] Richard Y Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637, 2024.

[46] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.

[47] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

[48] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

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

[50] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models, April 2024. URL http://arxiv.org/abs/2402.03300. arXiv:2402.03300 [cs].

[51] Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-Distillation Enables Continual Learning, January 2026. URL http://arxiv.org/abs/2601.19897. arXiv:2601.19897 [cs].

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

[53] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[54] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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

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

[57] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

[58] Hado Van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

[59] Mel Vecerik, Todd Hester, Jonathan Scholz, Fumin Wang, Olivier Pietquin, Bilal Piot, Nicolas Heess, Thomas Rothörl, Thomas Lampe, and Martin Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

[60] Weixun Wang, XiaoXiao Xu, Wanhe An, Fangwen Dai, Wei Gao, Yancheng He, Ju Huang, Qiang Ji, Hanqi Jin, Xiaoyang Li, et al. Let it flow: Agentic crafting on rock and roll, building the rome model within an open agentic learning ecosystem. *arXiv preprint arXiv:2512.24873*, 2025.

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

[62] Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang, Yejin Choi, and Manling Li. RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning, May 2025. URL http://arxiv.org/abs/2504.20073. arXiv:2504.20073 [cs].

[63] Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, et al. Agentic reasoning for large language models. *arXiv preprint arXiv:2601.12538*, 2026.

[64] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking LLMs in Web Traversal, August 2025. URL http://arxiv.org/abs/2501.07572. arXiv:2501.07572 [cs].

[65] Yangzhen Wu, Shanda Li, Zixin Wen, Xin Zhou, Ameet Talwalkar, Yiming Yang, Wenhao Huang, and Tianle Cai. Learn hard problems during rl with reference guided fine-tuning. *arXiv preprint arXiv:2603.01223*, 2026.

[66] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024.

[67] Feng Xiong, Zengbin Wang, Yong Wang, Xuecai Hu, Jinghan He, Liang Lin, Yuan Liu, and Xiangxiang Chu. Ace-skill: Bootstrapping multimodal agents with prioritized and clustered evolution. *arXiv preprint arXiv:2605.08887*, 2026.

[68] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

[69] Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to Reason under Off-Policy Guidance, June 2025. URL http://arxiv.org/abs/2504.14945. arXiv:2504.14945 [cs].

[70] Rui Yang, Qianhui Wu, Zhaoyang Wang, Hanyang Chen, Ke Yang, Hao Cheng, Huaxiu Yao, Baoling Peng, Huan Zhang, Jianfeng Gao, et al. Gui-libra: Training native gui agents to reason and act with action-aware supervision and partially verifiable rl. *arXiv preprint arXiv:2602.22190*, 2026.

[71] Shidong Yang, Ziyu Ma, Tongwen Huang, Xucong Wang, Renda Li, Yiming Hu, Yong Wang, and Xiangxiang Chu. Skillforge: Evolving verifiable skills for reinforcement learning agents. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 10448–10469, 2026.

[72] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

[73] Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. tau bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. *arXiv preprint arXiv:2406.12045*, 2024.

[74] Zhiyuan Yao, Yi-Kai Zhang, Yuxin Chen, Yueqing Sun, Zishan Xu, Yu Yang, Tianhao Hu, Qi Gu, Hui Su, and Xunliang Cai. Coba-rl: Capability-oriented budget allocation for reinforcement learning in llms. *arXiv preprint arXiv:2602.03048*, 2026.

[75] Junkeun Yi, Damon Mosk-Aoyama, Baihe Huang, Ritu Gala, Charles Wang, Sugam Dipak Devare, Khushi Bhardwaj, Abhibha Gupta, Oleksii Kuchaiev, Jiantao Jiao, et al. Pivotrl: High accuracy agentic post-training at low compute cost. *arXiv preprint arXiv:2603.21383*, 2026.

[76] Ailing Yu, Lan Yao, Jingnan Liu, Zhe Chen, Jiajun Yin, Yuan Wang, Xinhao Liao, Zhiling Ye, Ji Li, Yun Yue, et al. Medresearcher-r1: Expert-level medical deep researcher via a knowledge-informed trajectory synthesis framework. *arXiv preprint arXiv:2508.14880*, 2025.

[77] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale, May 2025. URL http://arxiv.org/abs/2503.14476. arXiv:2503.14476 [cs].

[78] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?, May 2025. URL http://arxiv.org/abs/2504.13837. arXiv:2504.13837 [cs].

[79] Yunpeng Zhai, Shuchang Tao, Cheng Chen, Anni Zou, Ziqian Chen, Qingxu Fu, Shinji Mai, Li Yu, Jiaji Deng, Zouying Cao, et al. Agentevolver: Towards efficient self-evolving agent system. *arXiv preprint arXiv:2511.10395*, 2025.

[80] 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.

[81] Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*, 2026.

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

[83] Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy rl reach with stale data on llms? *arXiv preprint arXiv:2510.01161*, 2025.

[84] Yuhao Zheng, Li'an Zhong, Yi Wang, Rui Dai, Kaikui Liu, Xiangxiang Chu, Linyuan Lv, Philip Torr, and Kevin Qinghong Lin. Code2world: A gui world model via renderable code generation. *arXiv preprint arXiv:2602.09856*, 2026.

[85] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[86] Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, et al. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*, 2025.

[87] Shun Zou, Ziyu Ma, Yi Zou, Yong Wang, Lin Chen, Zehui Chen, Guanhua Chen, Xiangxiang Chu, and Feng Zhao. Scaling agentic data for long-horizon terminal intelligence. *arXiv preprint arXiv:2610.13166*, 2026.

Appendix

A Datasets

A.1 Train

We adopted the search agent RL training data from ASearcher [19]. Specifically, we sampled 2k instances to be used for RL training across all our experimental settings.

Additionally, to acquire the action data, we utilized Tongyi-DeepResearch-30B-A3B [56] as the expert model to conduct rejection sampling. Consistent with our experimental settings, we restricted the toolset to only two types of tools: *web-search* and *web-visit*. The correct trajectories generated by

the expert model were collected, from which we exclusively extracted the atomic per-step operations (i.e., the tool call names and corresponding arguments) to serve as the candidate complete action guidance trajectories for each sample instance. The action turns statistics of the RL training data are shwon in Figure 8.

For the SFT data, we sampled another disjoint subset from the ASearcher dataset in a similar manner. We also employed Tongyi-DeepResearch-30B-A3B to conduct rejection sampling, yielding 4k complete search agent trajectories. Unlike the action data, the SFT data inherently preserves the comprehensive elements of the trajectory, emphasizing the retention of the full Chain-of-Thought (CoT) reasoning, explicit tool calls, and corresponding tool responses.

A.2 Evaluation

To comprehensively evaluate our proposed search agent's capabilities in complex reasoning and deep search, we adopt several standard and challenging deep search benchmarks. The details of the utilized datasets are outlined below:

- **GAIA** [40] is a challenging general AI assistant benchmark comprising real-world questions that require deep reasoning and web browsing. Following previous works, we utilize a subset of 103 text-only questions to test the fundamental capabilities of our system.
- **WebWalkerQA** [64] evaluates LLMs in complex web traversal and information gathering. It contains 680 QA tasks requiring agents to systematically traverse multiple dynamic web pages to discover multi-layered information via multi-hop reasoning.
- **XBench** [5] specifically assesses the deep search capabilities of AI agents. It comprises 100 questions and dynamically evaluates high-order information retrieval and tool usage abilities across real-world scenarios, considering both search breadth and reasoning depth.
- **BrowseComp-ZH** [86] is a complex benchmark measuring web browsing and reasoning within the Chinese internet ecosystem. It comprises 289 native, multi-hop retrieval questions strictly cross-validated across major search engines to test sophisticated multi-step reasoning.

To further assess out-of-domain generalization beyond the search-agent setting, we also evaluate on three general-purpose benchmarks:

- **GPQA** [48] is a graduate-level, Google-proof question-answering benchmark covering difficult scientific domains. We use it as an out-of-domain reasoning benchmark beyond the search-agent setting.
- **TruthfulQA** [34] evaluates whether language models generate truthful answers rather than imitating common misconceptions, providing an out-of-domain test of factual robustness.
- **IFEval** [85] measures instruction-following ability with verifiable constraints, serving as an out-of-domain benchmark for general alignment and controllability.

B Experiment Details

Implementation Details. Our implementation is built upon VeRL. All the experimental hyperparameter settings are listed in Table 6. During guided rollout, we inject the action data into the query prompt as plan-style reference guidance, so that the policy can follow the partial action trajectory while still completing any missing steps by itself. The exact prompt format is shown in Template B.

ACTGUIDE Prompt Template

```
Answer the given question using the given tools.
For each step, you must conduct a thought section to reason before calling any tools.

Question: {question}

Follow the partial action trajectory hint to take actions. Note that the trajectory may not be
complete, and you may still need to make extra tool calls to finish the task.

Reference action trajectory hint: {action trajectory}
```

Table 6: Hyperparameters for search-agent RL training.

Config	Setting
optimizer	AdamW
learning rate	1e-6
KL coefficient	0.001
training data	2,000
total training steps	64
training batch size	32
PPO mini batch size	16
group size	8
max response length	40,960
max observation length	8,000
max turns	30
$\epsilon_{\text{clip}_{\text{low}}}$	0.2
$\epsilon_{\text{clip}_{\text{high}}}$	0.2

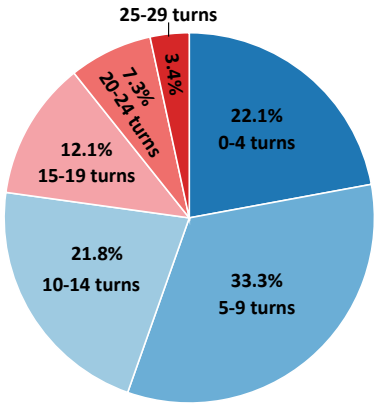

Figure 8: Action turns statistics of RL training data.

LLM-Judge Prompt Template

```
Based on the given question, standard answer, and model-predicted answer,
evaluate whether the model's response is correct. Your task is to classify
the result as: [CORRECT] or [INCORRECT].

First, we'll list examples for each category, then you'll evaluate a new
question's predicted answer.

Here are examples of [CORRECT] responses:
```
Question: What are the names of Barack Obama's children?
Standard Answer: Malia Obama and Sasha Obama
Model Prediction 1: Malia Obama and Sasha Obama
Model Prediction 2: Malia and Sasha
Model Prediction 3: Most would say Malia and Sasha, but I'm not sure, I should verify
Model Prediction 4: Barack Obama has two daughters, Malia Ann and Natasha Marian,
                    commonly known as Malia Obama and Sasha Obama.
```

These responses are all [CORRECT] because they:
  - Fully include the important information from the standard answer.
  - Don't contain any information that contradicts the standard answer.
  - Focus only on semantic content; language, capitalization, punctuation,
    grammar, and order aren't important.
  - Vague statements or guesses are acceptable as long as they include the
    standard answer and don't contain incorrect information or contradictions.

Here are examples of [INCORRECT] responses:
```
Question: What are the names of Barack Obama's children?
Standard Answer: Malia Obama and Sasha Obama
Model Prediction 1: Malia
Model Prediction 2: Malia, Sasha and Susan or Sasha Obama or Malia Obama,
                    or Natasha Marian, or Einstein
Model Prediction 3: While I don't know their exact names, I can tell you
                    Barack Obama has two children.
Model Prediction 4: You might be thinking of Betsy and Olivia. But you should
                    verify the details with the latest references. Is that
                    the correct answer?
Model Prediction 5: Barack Obama's children
```

These responses are all [INCORRECT] because they:
  - Contain factual statements that contradict the standard answer.
  - Are empty or merely repeat the question.
  - Enumerate multiple answers or repeat the answer.

Pay special attention to the following:
  - The standard answer may contain responses to multiple aspects of the
    question. Within the same aspect, there may be different descriptions,
```

```
    all of which are correct and are given in the same bracket, connected by
    commas. For example:
      Question: What is the name of ByteDance's AI model?
      Standard Answer: [[Doubao, Skylark]]
      Correct Predictions: Doubao; Doubao, Skylark; Skylark

  - For standard answers containing responses to different aspects, the model
    needs to provide answers to all aspects to be considered correct. There is
    no [PARTIALLY CORRECT] output option. These answers will be given in
    different brackets. For example:
      Question: Who are the members of TFBOYS?
      Standard Answer: [[Wang Junkai][Wang Yuan][Yi Yangqianxi]]
      Correct Prediction: Wang Junkai, Wang Yuan, Yi Yangqianxi
      Incorrect Prediction: Wang Junkai, Yi Yangqianxi

Also note the following points:
  - For questions with numerical standard answers, the predicted answer should
    match the standard answer. For example:
      Question: What is the total length in meters of the Huangpu River Bridge
                on the Jinshan Railway?
      Standard Answer: 3518.17
      Correct Predictions: 3518; 3518.1; 3518.17
      Incorrect Predictions: 3520; 3600

  - If the model prediction doesn't directly answer the question, attempts to
    circumvent, or fails to directly provide the standard answer, it's
    considered an [INCORRECT] answer. For example:
      Question: Who is JJ Lin's wife?
      Standard Answer: Ding Wenqi
      Incorrect Predictions: JJ Lin's wife; JJ Lin's wife should be excellent;
                             JJ Lin's wife might be a public figure

  - If the standard answer contains more information than the question asks
    for, the predicted answer only needs to include the information mentioned
    in the question. For example:
      Question: What is the main chemical component of magnesite?
      Standard Answer: Magnesium carbonate (MgCO3)
      Correct Predictions: Magnesium carbonate; MgCO3

  - If information omitted in the predicted answer can be clearly inferred
    from the question, it's considered correct. For example:
      Question: The Nuragic ruins of Barumini were listed as a World Cultural
                Heritage by UNESCO in 1997, so where is this site located?
      Standard Answer: Sardinia, Italy
      Correct Prediction: Sardinia

  - If it's clear that different translations of a name refer to the same
    person, it's considered correct. For example, if the standard answer is
    Robinson, answers like Lubinson or Lubinsun are both correct.

  - Focus more on the match between the standard answer and the model
    prediction than whether the standard answer itself is correct.

Below is a new question example. Please reply with only [CORRECT] or
[INCORRECT], without apologies or corrections to your own errors. Just
evaluate the answer.
```
Question: {question}
Standard Answer: {correct_answer}
Predicted Answer: {response}
```

Evaluate this new question's predicted answer as one of the following:
A. [CORRECT]
B. [INCORRECT]

Return only the option representing [CORRECT] or [INCORRECT], i.e., just
return A or B, without adding any other text.
```

Compute Resources. All training and rollout experiments were conducted on nodes equipped with 8 NVIDIA H20 GPUs. The LLM judge used for reward assignment and test-time evaluation required additional serving resources, for which we used a separate node with 8 NVIDIA H20 GPUs.

Different Guidance Methods. We also compare different ways of injecting action guidance for LLM-based agent. Besides the unguided setting, we consider an assistant-prefix format following prior hint-based methods, where the action reference is prepended as a generated prefix and the model

continues generation from it. We also consider a user-assistant message format, where the action data are converted into the corresponding tool calls and tool responses and then assembled as multi-turn messages before the model continues generation. As shown in Table 7, inserting the action trajectory as a reference plan in the query prompt achieves the best Reward@1, suggesting that lightweight plan-style guidance is more effective than directly prefixing or replaying actions for LLM agent.

Table 7: Injection method comparison.

Inject Method	Unguidance	Assistant Prefix	User-Assistant Messages	Reference Plan in Query Prompt
Reward@1	57.90	74.50	80.10	**85.70**

Action Data for On-policy Self Distillation. Beyond using action data to guide the policy toward better state visitation, we also explore whether it can be used for on-policy self distillation (OPSD) [23, 51, 82]. Specifically, OPSD still samples trajectories from the unguided policy, but uses action-conditioned guided logits as the distillation target on these on-policy rollouts. Formally, for an unguided rollout $\tau \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$, we re-evaluate each visited prefix $z_{<t}$ with the same model additionally conditioned on the action guidance g, and optimize

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\,\tau\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{D}_{\text{KL}}(\text{sg}[\pi_{\theta_{\text{old}}}(\cdot \mid z_{<t}, g)] \,\|\, \pi_\theta(\cdot \mid z_{<t}))\right], \tag{13}$$

where $\text{sg}[\cdot]$ denotes stop-gradient, so the guided distribution serves only as a token-level teacher while the learned policy remains unguided at inference time. As shown in Table 8, OPSD can improve model performance, but the gains remain limited because the visited states are still determined by the base unguided policy. Therefore, it does not fundamentally resolve the ineffective state-visitation problem when the agent cannot reach useful states by itself.

Table 8: Comparison between ACTGUIDE-RL and OPSD.

Guidance Use	GAIA	WebWalker	XBench
ACTGUIDE-RL	35.92	**39.85**	**37.00**
OPSD	**36.89**	30.29	26.00

C Theoretical Analysis

C.1 Covariance Form of the Token-Level Off-Policy Risk

In Section 2.3.2, let $\tau = (z_1, \ldots, z_{|\tau|})$ be the generated token sequence. We define the token-level importance ratio under guidance level g_k as

$$r_j^{(k)} := \frac{\pi_\theta(z_j \mid z_{<j})}{\pi_\theta(z_j \mid z_{<j}, g_k)}, \tag{14}$$

and the corresponding cumulative log-ratio shift as

$$\mathcal{L}_k(\tau) := \sum_{j=1}^{|\tau|} \log r_j^{(k)}. \tag{15}$$

The off-policy risk is then defined as

$$R_k := \text{Var}_{\tau\sim\pi_\theta(\cdot|s,g_k)}(\mathcal{L}_k(\tau)). \tag{16}$$

By variance expansion, we have

$$R_k = \text{Var}_{\tau\sim\pi_\theta(\cdot|s,g_k)}\left(\sum_{j=1}^{|\tau|}\log r_j^{(k)}\right). \tag{17}$$

Therefore,

$$R_k = \sum_{j=1}^{|\tau|} \mathrm{Var}_{\tau\sim\pi_\theta(\cdot|s,g_k)}\left(\log r_j^{(k)}\right) + 2\sum_{j<j'} \mathrm{Cov}_{\tau\sim\pi_\theta(\cdot|s,g_k)}\left(\log r_j^{(k)}, \log r_{j'}^{(k)}\right), \tag{18}$$

where the second summation ranges over all distinct token pairs in the rollout.

This decomposition shows that the off-policy risk consists of two components: (1) token-wise variance terms, which capture local distribution mismatch at each generation step, and (2) cross-token covariance terms, which capture the dependence structure of these mismatches along the autoregressive trajectory. Hence, stronger guidance may increase not only the magnitude of token-level deviations, but also their correlation across the rollout, both of which contribute to larger internalization risk.

In particular, if the token-level log-ratio shifts were independent, then all covariance terms would vanish and R_k would reduce to the sum of token-wise variances. However, in autoregressive agent generation, token dependencies are intrinsic, and the covariance terms generally cannot be ignored. This motivates using the variance of the cumulative log-ratio shift as a compact measure of off-policy risk.

C.2 Risk-Constrained View of Minimal Intervention

We formalize the minimal-intervention rule in Section 2.3.2 as a risk-constrained selection problem. For a guidance level g_k, define the group recovery probability

$$Q_k := \mathbb{P}_{\{\tau_i^{(k)}\}_{i=1}^N\sim\pi_\theta(\cdot|x,g_k)}\left(\max_{i\le N} Y(\tau_i^{(k)}) \ge \delta\right), \tag{19}$$

where N is the rollout group size and δ is the success threshold. Given a target recovery level $\rho \in (0,1)$, the least-risk sufficient guidance level is the solution of

$$\min_{k\in\{0,\dots,K\}} R_k \quad \text{s.t.} \quad Q_k \ge \rho. \tag{20}$$

Assumption C.1 (Monotone recovery and risk)**.** *The ordered guidance family $g_0 \prec g_1 \prec \cdots \prec g_K$ satisfies:*

$$Q_0 \le Q_1 \le \cdots \le Q_K, \qquad R_0 \le R_1 \le \cdots \le R_K. \tag{21}$$

Proposition C.1 (Minimal sufficient guidance is risk-optimal)**.** *Under Assumption C.1, if the feasible set of Eq. 20 is non-empty, then*

$$k_\rho^\star := \min\{k \in \{0,\dots,K\} : Q_k \ge \rho\} \tag{22}$$

is an optimal solution of Eq. 20.

Proof. By definition, $k_\rho^\star$ is feasible. For any other feasible level k, minimality of $k_\rho^\star$ implies $k \ge k_\rho^\star$. Since R_k is non-decreasing in k, we have $R_k \ge R_{k_\rho^\star}$. Hence no feasible guidance level has smaller off-policy risk than $k_\rho^\star$. □

This proposition gives a constrained interpretation of Eq. 10: minimal intervention does not maximize guidance strength, but selects the lowest-risk level that satisfies a recovery requirement. When Q_k is not known exactly, it can be estimated by repeated rollout groups. Let $\widehat{Q}_k$ be the empirical mean of m independent group-recovery indicators at level k.

Corollary C.1 (Empirical identification under a margin)**.** *Suppose Assumption C.1 holds and there exists a margin $\Delta > 0$ such that*

$$Q_k \le \rho - \Delta \quad \forall k < k_\rho^\star, \qquad Q_k \ge \rho + \Delta \quad \forall k \ge k_\rho^\star. \tag{23}$$

If

$$m \ge \frac{1}{2\Delta^2}\log\frac{2(K+1)}{\xi}, \tag{24}$$

then with probability at least $1-\xi$, the empirical rule

$$\widehat{k}_\rho := \min\{k : \widehat{Q}_k \ge \rho\} \tag{25}$$

recovers $k_\rho^\star$.

Proof. By Hoeffding's inequality and a union bound over $K+1$ levels,

$$\mathbb{P}\left(\max_k |\widehat{Q}_k - Q_k| \geq \Delta\right) \leq 2(K+1)\exp(-2m\Delta^2) \leq \xi. \tag{26}$$

On the complementary event, every $k < k^\star_\rho$ has $\widehat{Q}_k < \rho$, while every $k \geq k^\star_\rho$ has $\widehat{Q}_k \geq \rho$. Therefore the empirical minimal feasible level equals $k^\star_\rho$. □

D Training Cases of ACTGUIDE-RL

We present representative training cases to illustrate how ACTGUIDE-RL injects action data as plan-style guidance. Each case uses the full guidance level g_K, where the prompt contains the complete available reference action trajectory while still requiring the policy to reason and complete the task with tools.

Training Case 1 of ACTGUIDE-RL with Full Guidance g_K

Example from RL Training Data

Original Question: Which political party last held overall control of the West Berkshire Council before the election that first introduced the ward named after the Berkshire village where the Kennet & Avon Canal route was restored to Hungerford Wharf in July 1974, resulting in a tied outcome necessitating by-elections and enabling the Conservatives to secure a majority?

Ground Truth: The Liberal Democrats

Prompt for ActGuide-RL

Answer the given question using the given tools.

For each step, you must conduct a thought section to reason before calling any tools.

Question: Which political party last held overall control of the West Berkshire Council before the election that first introduced the ward named after the Berkshire village where the Kennet & Avon Canal route was restored to Hungerford Wharf in July 1974, resulting in a tied outcome necessitating by-elections and enabling the Conservatives to secure a majority?

Follow the partial action trajectory hint to take actions, note that the trajectory may not complete and you still need do some extra tool calls to finish the task.

Reference action trajectory hint:

<1>: search [''Kennet & Avon Canal restored to Hungerford Wharf July 1974'', ''Hungerford Wharf July 1974 canal restoration'', ''Kennet and Avon canal restoration July 1974'', ''Berkshire village Hungerford Wharf restored July 1974'']

<2>: search [''West Berkshire Council election tied result by-elections Conservatives secured majority new ward named after Hungerford'', ''West Berkshire Council election tied outcome by-elections Conservative majority ward introduction'', ''West Berkshire Council ward first introduced election tied by-elections'']

<3>: visit [https://en.wikipedia.org/wiki/West_Berkshire_Council_elections]

<4>: search [''Hungerford ward West Berkshire Council'', ''Hungerford electoral ward West Berkshire'', ''Hungerford West Berkshire Council ward'']

<5>: search [''tied West Berkshire council election by-election Conservative majority'', ''tied West Berkshire by-election'', ''by-elections West Berkshire Conservative majority'']

<6>: search [''West Berkshire Council tied election'', ''West Berkshire tied election'', ''West Berkshire tied council'']

<7>: search [''tied result West Berkshire council by-election'', ''tied by-elections West Berkshire'', ''tie West Berkshire council election'']

<8>: search [''West Berkshire 2003 new ward'', ''West Berkshire 2003 ward boundary'', ''West Berkshire 2003 electoral changes'']

<9>: visit [https://grokipedia.com/page/2003_west_berkshire_council_election]

<10>: search [''Hungerford ward West Berkshire 2003'', ''Hungerford West Berkshire new ward 2003'']

<11>: visit [https://www.andrewteale.me.uk/leap/ward/2226/]

<12>: visit [https://en.wikipedia.org/wiki/West_Berkshire_Council_elections]

Training Case 2 of ACTGUIDE-RL with Full Guidance g_K

Example from RL Training Data

Original Question: In a chess tournament, all players faced each other twice and scored 1 point for a win, half a point for a draw, and zero points for a loss. The winner was the one who scored the most points. Luíza, a curious mathematician, found a paper stating that the sum of the scores of all participants was 210 points. a) How many players participated in the tournament? b) Luíza noticed that the tournament logo depicted a knight in the center, with black squares illustrating the knight's possible moves. She wondered: If the board were infinite and the knight continued its pattern of movement, onto how many different squares could the knight move in two moves? Initially, she thought ''64'' but quickly realized she was wrong. Now, it's your turn to correctly answer the question Luíza thought of. c) Luíza received a message from a friend who participated in the tournament, informing

her that he scored 12 points. Luíza replied: ‘‘- Don’t be sad, you will win the next championship!’’ How did Luíza know that her friend did not win the tournament?

Ground Truth: 33

Prompt for ActGuide-RL

Answer the given question using the given tools.

For each step, you must conduct a thought section to reason before calling any tools.

Question: In a chess tournament, all players faced each other twice and scored 1 point for a win, half a point for a draw, and zero points for a loss. The winner was the one who scored the most points. Luíza, a curious mathematician, found a paper stating that the sum of the scores of all participants was 210 points. a) How many players participated in the tournament? b) Luíza noticed that the tournament logo depicted a knight in the center, with black squares illustrating the knight’s possible moves. She wondered: If the board were infinite and the knight continued its pattern of movement, onto how many different squares could the knight move in two moves? Initially, she thought ‘‘64’’ but quickly realized she was wrong. Now, it’s your turn to correctly answer the question Luíza thought of. c) Luíza received a message from a friend who participated in the tournament, informing her that he scored 12 points. Luíza replied: ‘‘- Don’t be sad, you will win the next championship!’’ How did Luíza know that her friend did not win the tournament?

Follow the partial action trajectory hint to take actions, note that the trajectory may not complete and you still need do some extra tool calls to finish the task.

Reference action trajectory hint:

<1>: search [‘‘knight moves two moves reachable squares infinite board 33’’]

<2>: search [‘‘knight two moves reachable squares infinite board’’]

<3>: visit [https://www.reddit.com/r/chess/comments/bem4zw/given_an_infinite_chess_board_how_many_spaces/]

<4>: search [‘‘how many squares can a knight reach in two moves’’]

<5>: search [‘‘knight reach two moves how many squares’’]

<6>: visit [https://file.scirp.org/Html/8-1200157_34513.htm]

<7>: search [‘‘33 squares knight two moves’’]

<8>: visit [https://www.mathworks.com/matlabcentral/cody/problems/60541-compute-the-number-of-squares-a-knight-can-reach-after-n-moves]

<9>: visit [https://www.mathworks.com/content/dam/mathworks/images/cody/graphics/ChessKnight.svg]

<10>: search [‘‘33 squares knight two moves infinite board’’]

<11>: search [‘‘33 squares after two moves knight’’]

<12>: search [‘‘knight can reach after two moves 33’’]

<13>: visit [https://www.mathworks.com/matlabcentral/cody/problems/60541-compute-the-number-of-squares-a-knight-can-reach-after-n-moves]

<14>: search [‘‘knight two moves 33 squares’’]

<15>: visit [https://www.reddit.com/r/chess/comments/1j3roeo/the_amount_of_space_a_single_knight_can_control/]

Training Case 3 of ActGuide-RL with Full Guidance g_K

Example from RL Training Data

Original Question: In which year, during the early 1970s, did a player, known for his time with the New York Yankees, join the team, and who is the former MLB pitcher, now a pitching coach for a minor league team affiliated with the Miami Marlins, whose first name starts with ‘M’ and who started the opening game of the 2009 World Baseball Classic against Venezuela, pitching 4 shutout innings?

Ground Truth: 1972, Mark DiFelice

Prompt for ActGuide-RL

Answer the given question using the given tools.

For each step, you must conduct a thought section to reason before calling any tools.

Question: In which year, during the early 1970s, did a player, known for his time with the New York Yankees, join the team, and who is the former MLB pitcher, now a pitching coach for a minor league team affiliated with the Miami Marlins, whose first name starts with ‘M’ and who started the opening game of the 2009 World Baseball Classic against Venezuela, pitching 4 shutout innings?

Follow the partial action trajectory hint to take actions, note that the trajectory may not complete and you still need do some extra tool calls to finish the task.

Reference action trajectory hint:

<1>: search [‘‘2009 World Baseball Classic opening game Venezuela 4 shutout innings pitcher’’, ‘‘Venezuela 2009 WBC opening game pitcher 4 shutout innings’’, ‘‘World Baseball Classic 2009 Venezuela opener pitcher’’, ‘‘2009 WBC Venezuela vs Opening Game pitcher’’, ‘‘2009 WBC Venezuela 4 shutout innings start’’]

<2>: search [‘‘2009 World Baseball Classic opening game March 5 2009 opening games schedule’’, ‘‘first game of 2009 World Baseball Classic opening day’’]

<3>: visit [https://www.aol.com/articles/miami-opening-game-world-baseball-183425977.html]

<4>: search [‘‘2009 World Baseball Classic Venezuela vs Italy opening game 4 shutout innings pitcher’’, ‘‘Venezuela vs Italy March 7 2009 WBC pitcher started 4 shutout innings’’]

<5>: visit [https://www.arkansasonline.com/news/2009/mar/11/world-baseball-classic-venezuelas-power-too-much-for-italy/]

<6>: visit [https://www.espn.com/world-baseball-classic/boxscore/_/gameId/290307116]
<7>: search [‘‘Mark DiFelice pitching coach Miami Marlins affiliate’’, ‘‘Mark DiFelice minor league pitching coach Marlins’’]
<8>: search [‘‘joined the New York Yankees in 1970’’, ‘‘joined the New York Yankees in 1971’’, ‘‘joined the New York Yankees in 1972’’, ‘‘joined the New York Yankees in 1973’’, ‘‘joined the New York Yankees in 1974’’, ‘‘joined the New York Yankees in 1975’’]
<9>: search [‘‘joined the New York Yankees 1970’’, ‘‘joined the New York Yankees 1971’’, ‘‘joined the New York Yankees 1972’’, ‘‘joined the New York Yankees 1973’’, ‘‘joined the New York Yankees 1974’’, ‘‘joined the New York Yankees 1975’’]
<10>: search [‘‘Sparky Lyle joined the New York Yankees 1972’’]
<11>: search [‘‘Graig Nettles joined the New York Yankees 1972’’]
<12>: search [‘‘known for his time with the New York Yankees’’]
<13>: search [‘‘Thurman Munson joined the New York Yankees’’]
<14>: visit [https://pinstripesnation.com/thurman-munson-the-captain-who-never-left-the-yankees-2023-08-02/]
<15>: search [‘‘Graig Nettles known for his time with the New York Yankees’’]
<16>: search [‘‘Mark DiFelice 2009 World Baseball Classic opening game Venezuela 4 shutout innings’’]
<17>: visit [https://en.wikipedia.org/wiki/Mark_DiFelice]
<18>: visit [https://www.ebay.com/itm/389131297484]
<19>: visit [https://en.wikipedia.org/wiki/Graig_Nettles]
<20>: visit [https://en.wikipedia.org/wiki/Mark_DiFelice]
<21>: search [‘‘Mark DiFelice Marlins’’, ‘‘Mark DiFelice Miami Marlins’’]
<22>: visit [https://en.wikipedia.org/wiki/Graig_Nettles]
<23>: search [‘‘Sparky Lyle New York Yankees known for his time with’’]
<24>: visit [https://en.wikipedia.org/wiki/Sparky_Lyle]
<25>: search [‘‘best known for his time with the New York Yankees Sparky Lyle’’]
<26>: search [‘‘best known for his time with the New York Yankees Graig Nettles’’]

E Limitations

Due to the relatively simple experimental setup, the ease of obtaining task queries with different difficulty levels, and the natural availability of action data, our main experiments are conducted in the search-agent setting. This setting provides a controlled testbed for studying reachability barriers and guidance-induced off-policy risk. Nevertheless, ACTGUIDE-RL is designed for general agentic training rather than being specific to search agents, and its effectiveness in other agent tasks, such as CLI, GUI, API-based, and embodied environments, remains to be further explored.

This work utilizes action data through plan-style guidance, where reference actions are injected as a high-level action plan to help the policy cross exploration barriers. This simple formulation keeps the method lightweight, broadly applicable, and independent of costly reasoning traces. More fine-grained ways of using action data, such as step-level guidance injection, also remain to be further explored.

This work focuses on how to leverage action data for agentic RL, but does not discuss how such data should be systematically collected and processed. In practice, structured collection, cleaning, and filtering of existing interaction records, such as backend logs from different agent applications, are also important for action-data-based training and remain worth exploring.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

   Answer: [Yes]

   Justification: The main claims are stated in the abstract and introduction, including the motivation, method design, and empirical improvements. The reported claims are supported by the theoretical analysis, main results, and ablation studies in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section in Appendix E. It discusses the search-agent experimental scope, the plan-style use of action data, and the structured collection and processing of action data.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The theoretical assumptions, propositions, corollaries, and proofs are provided in Appendix C. The method section also introduces the main definitions and analysis used to motivate the proposed algorithm.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the training data, evaluation benchmarks, tool setting, reward and evaluation protocol, prompt format, and training hyperparameters in the experiment section and Appendix A and Appendix B. These details are intended to make the main experimental results reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides the data sources, action data construction procedure, prompt format, implementation details, and hyperparameters needed to reproduce the main results.

The accompanying code and data are made available with instructions for reproducing the reported experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup specifies the benchmarks, baselines, tool environment, reward and evaluation protocol, and implementation details. Appendix A and Appendix B provide the data sources, prompt template, and hyperparameter table.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports single-run benchmark results and ablation results without error bars, confidence intervals, or statistical significance tests. We do not include repeated-run statistics due to the computational cost of agentic RL training and LLM-judge evaluation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are described in Appendix B. The reported experiments were conducted on nodes with 8 NVIDIA H20 GPUs, with a separate 8-H20 node used for serving the LLM judge.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses standard public benchmarks, existing open-source infrastructure, and LLM-based agents in controlled experimental settings. We have reviewed the NeurIPS Code of Ethics and believe the work conforms to its requirements.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work focuses on a general method for improving agentic RL with action guidance and does not introduce a deployed system or application-specific technology with direct societal impact. Therefore, a broader societal impact discussion is not applicable.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk pretrained foundation models, image generators, or scraped datasets. The released assets are limited to code and documented action-guidance resources for reproducing the experiments.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original sources of the training data, evaluation benchmarks, base models, expert model, and implementation framework used in the experiments. These assets are used according to their public release terms, and the relevant sources are documented in the experiment section and Appendix A.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces action guidance trajectories constructed from expert successful rollouts, and documents their construction process in Appendix A. Documentation and instructions are provided alongside the released assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted using existing benchmarks and model-generated trajectories.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or an equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs were only used for writing, editing, and formatting assistance, and did not affect the core methodology, scientific rigor, or originality of the research. The paper separately describes the use of an LLM judge for reward assignment and evaluation in the experimental setup.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.