

HarnessForge: Joint Harness and Policy Evolution for Adaptive Agent Systems

Mingju Chen¹, Can Lv¹, Guibin Zhang³, Heng Chang⁴, Shiji Zhou^{1,2}

¹Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing,
School of Artificial Intelligence, Beihang University,

²Beijing Academy of Artificial Intelligence (BAAI),

³National University of Singapore, ⁴Tsinghua University

Project Lead: Heng Chang, Corresponding to: Shiji Zhou <zhoushiji25@buaa.edu.cn>

Abstract

LLM agents are increasingly expected to operate across heterogeneous task regimes that require distinct execution paradigms. This challenges fixed agent systems and motivates system-level meta-adaptation beyond isolated component updates. While existing works have adapted external harness or trained underlying reasoning policies, full-system adaptation remains insufficiently characterized. The adaptation space between structure and execution is rarely made explicit, and the compatibility between the external harness and the internal reasoner is not optimized jointly. We propose HarnessForge, a meta-adaptive framework for evolving LLM agent systems. HarnessForge formulates an agent system as a harness-policy pair, defining a stable adaptation space that separates harness-level execution structure from policy-level reasoning behavior. It then performs harness-policy co-evolution through fault-guided harness tailoring and harness-conditioned policy alignment. Experiments across five benchmarks from diverse domains show that HarnessForge consistently improves both Qwen3-4B and Qwen3-8B backbones, outperforming harness-only and policy-only baselines with gains of up to 12.0% over the strongest baseline and achieving favorable rollout-efficiency tradeoffs, demonstrating that harness-policy co-evolution is effective, and that executable compatibility between the harness and reasoning policy is essential for agent-system adaptation. The code is available at <https://github.com/mingju-c/HarnessForge>.

1 Introduction

LLM agents are increasingly deployed in complex and heterogeneous task regimes, including multi-step tool use (Schick et al., 2023; Qian et al., 2026), retrieval-heavy reasoning (Jin et al., 2025; Shen et al., 2025), web interaction (Yao et al., 2022), and

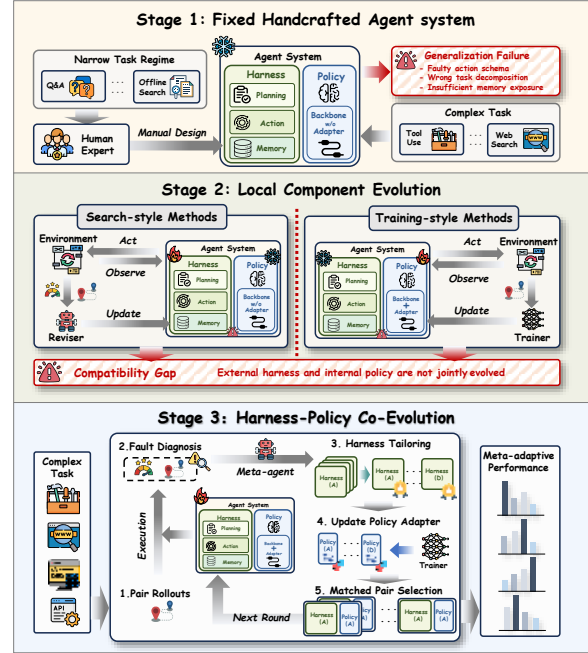


Figure 1: A three-stage view of LLM agent adaptation: fixed handcrafted systems, local component adaptation, and harness-policy co-evolution (HarnessForge).

stateful multi-turn tasks (Yao et al., 2024; Shinn et al., 2023). These regimes differ not only in task difficulty, but also in the structural requirements they impose on agent execution. Some require explicit task decomposition and verification, some rely on strict action schemas and tool-use protocols, while others demand persistent memory exposure, or state tracking. Such diversity suggests that there is unlikely to be a single fixed agent system that performs optimally across regimes. Instead, agent systems should be able to meta-adapt their execution paradigms to the target task regime.

As shown in Fig. 1, recent work has begun to move toward this goal by treating agent components as adaptive objects rather than fixed hand-written artifacts (Hu et al., 2025; Zhang et al., 2025c; Shang et al., 2025; Zhang et al., 2025a). Concretely, some search-style methods optimize external execution structures, including workflows,

tool-use procedures, role assignments, memory management, or execution graphs (Wu et al., 2024; Hong et al., 2024; Zhong et al., 2023; Packer et al., 2024; Zhang et al., 2025b), showing that external execution structures can be searched, revised, or evolved. Other training-style methods adapt the model policy through supervised learning, preference learning, or reinforcement learning on agentic trajectories (Shinn et al., 2023; Madaan et al., 2023; Zhang et al., 2026; Li et al., 2025; Shao et al., 2024; Lv et al., 2026b), improving the model’s internal execution behavior. These works demonstrate that different components of an agent system can be adapted.

However, this component-level view remains insufficient for system-level meta-adaptation. The core issue lies in the **adaptation target**: existing methods typically optimize external harness or internal policies as separate objects, whereas a full LLM agent system operates as a coupled harness–policy pair. This coupling is especially salient in resource-constrained settings, where the harness provides structural support for limited model capabilities and the policy must learn to execute the harness-induced execution paradigm. Moreover, existing adaptation still faces a **compatibility gap**. A more expressive harness may expose useful planning, action, or memory structures, yet fail if the reasoner cannot reliably execute them; conversely, a stronger policy may still be constrained by a harness that exposes unsuitable states, actions, or control signals. Effective system-level adaptation should therefore go beyond optimizing either side alone and instead co-evolve the external harness and internal policy to improve their compatibility.

Motivated by this view, we formulate an LLM agent system as a harness–policy pair, making the coupled harness and policy the explicit unit of adaptation. The harness specifies the external execution interface, including planning, action, and memory structures that shape agent behavior; the policy captures how the reasoner executes within this interface. Based on this formulation, we propose HarnessForge, a framework for harness–policy co-adaptation. On the harness side, HarnessForge uses rollout diagnostics with a meta-agent to perform fault-guided harness tailoring over planning, action, and memory components. On the policy side, it trains harness-conditioned adapters from curated trajectories to align the reasoner with the selected harness. Finally, HarnessForge selects and evolves harness–policy pairs, optimizing task-

regime-adapted agent systems rather than isolated workflows or policies across evolutionary rounds.

Our contributions are summarized as follows:

- We reformulate system-level LLM agent adaptation from local component optimization to harness–policy pair evolution, treating the coupled external harness and internal policy as the basic unit of optimization.
- We propose HarnessForge, a meta-adaptive co-evolution framework that coordinates fault-guided harness tailoring with harness-conditioned policy alignment to improve pair-level executable compatibility.
- We evaluate HarnessForge across five benchmarks and two backbones, improving over the strongest harness-only and policy-only baselines by 3.56% on average and up to 12.0%, while preserving favorable rollout–performance trade offs and harness–policy compatibility.

2 Related Work

Optimization for Agent System Design Prior work shows that LLM-agent capabilities are strongly shaped by the external harness used to organize reasoning and execution. Early prompting and interaction paradigms, such as Chain-of-Thought (Wei et al., 2022), Plan-and-Solve (Wang et al., 2023), and ReAct (Yao et al., 2023), introduce explicit reasoning traces, plans, actions, and observations. Agent frameworks further expose more components, including roles, protocols, and memory modules in AutoGen (Wu et al., 2024), MetaGPT (Hong et al., 2024), and memory-augmented agent systems (Zhong et al., 2023; Packer et al., 2024; Zhang et al., 2025b; Lv et al., 2026a; Chen et al., 2026). More recent search-style methods automate harness design through workflow or architecture search, including ADAS (Hu et al., 2025), AFlow (Zhang et al., 2025c), AgentSquare (Shang et al., 2025), MaAS (Zhang et al., 2025a), AutoHarness (Lou et al., 2026), Meta-Harness (Lee et al., 2026) and MermaidFlow (Zheng et al., 2025a). These works reduce manual engineering and establish harnesses as important optimization targets, but they mainly optimize external structures while leaving compatibility with the model-side executor implicit.

Agentic RL for Policy Evolution Another line of work optimizes the model-side policy. Recent agentic RL methods train models from in-

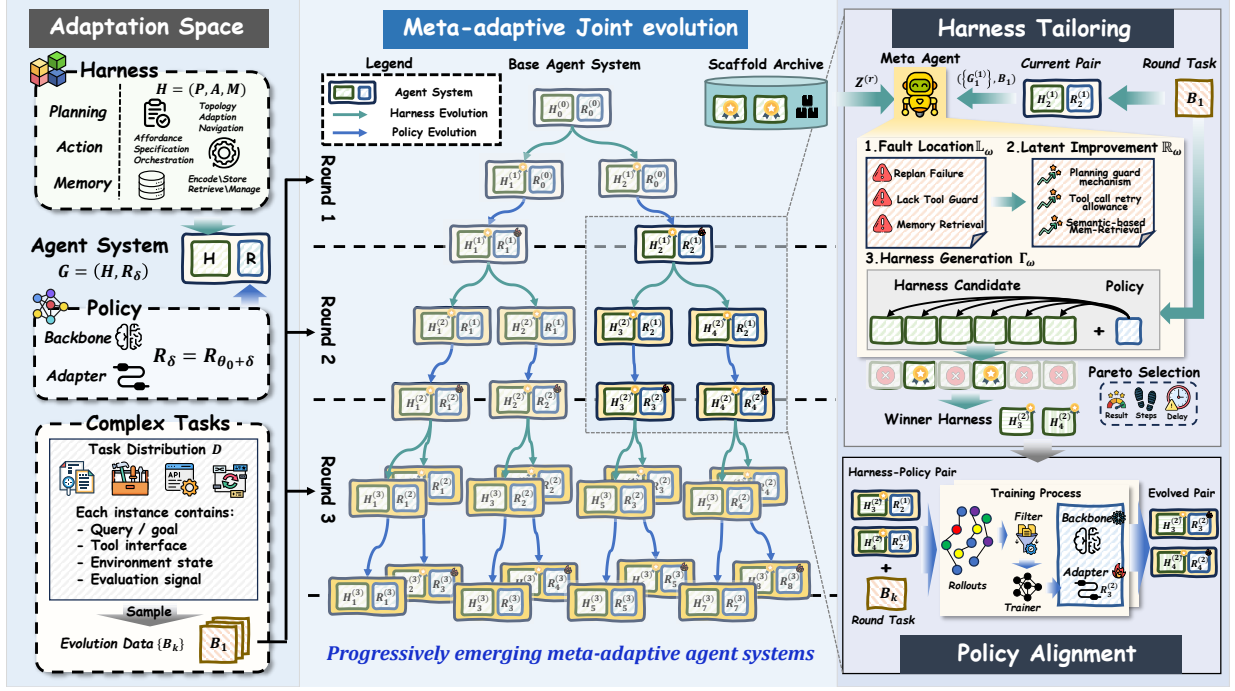


Figure 2: Overview of the HarnessForge co-evolution workflow. Starting from a harness–policy pair, each round diagnoses execution failures, tailors the harness over planning, action, and memory modules, trains a harness–conditioned adapter from curated trajectories, and selects improved matched pairs for the next round.

teractive trajectories with task rewards, tool feedback, or environment signals. For search and tool use, Search-R1 (Jin et al., 2025), ToolRL (Qian et al., 2026), and ToRL (Li et al., 2025) optimize when and how models issue external actions. For long-horizon agent execution, GiGPO (Feng et al., 2025), TreeRL (Hou et al., 2025), and ARPO (Dong et al., 2026) address credit assignment and trajectory-level optimization, while Planner-R1 (Zhu et al., 2025) and Memory-R1 (Yan et al., 2026) extend RL to planning and memory management. Broader reasoning-RL systems, including DeepseekMATH (GRPO) (Shao et al., 2024), DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025b), Satori (Shen et al., 2025), and Absolute Zero (Zhao et al., 2025), further show that RL can improve reasoning, exploration, and self-generated curricula. These methods strengthen the internal executor, but typically assume a fixed, externally specified interaction loop or task interface.

3 Methodology

Fig. 2 shows the overall workflow of HarnessForge. Sec. 3.1 formalizes the agent system, and evaluation criteria. Sec. 3.2 presents our meta-adaptive joint evolution mechanism. Secs. 3.3 and 3.4 detail harness and policy evolution, respectively.

3.1 Preliminary

Agent-System Formulation. We formulate an LLM agent system $\mathcal{G}$ as the coupling of an external harness $\mathcal{H}$ and an adapted reasoner $\mathcal{R}_\delta$:

$$\mathcal{G} = (\mathcal{H}, \mathcal{R}_\delta), \mathcal{H} = (\mathcal{P}, \mathcal{A}, \mathcal{M}), \mathcal{R}_\delta = \mathcal{R}_{\theta_0 + \delta}. \quad (1)$$

Here $\mathcal{H}$ is the editable execution harness, which decomposes into three execution-layer components: $\mathcal{P}$ denotes the planning component, including task decomposition, replanning, and termination. $\mathcal{A}$ denotes the action component, including tool interfaces, role assignment, and orchestration rules. $\mathcal{M}$ denotes the memory component, including what is written, retrieved, summarized, and exposed to future decisions. The adapted reasoner $\mathcal{R}_\delta$ denotes the reasoning component, which parameterizes the policy that executes under this harness, and δ is a lightweight adapter on base reasoner $\mathcal{R}_{\theta_0}$.

Evaluation Criteria. Given a task x and trajectory τ_x , we let $\phi(\tau, x)$ collect final response quality within environment, negative token cost, negative latency to evaluate the agent system, with larger values preferred in every dimension. For a batch $B \subset \mathcal{D}$, we define fitness indicator $\mathbf{J}(\mathcal{G}; B)$:

$$\mathbf{J}(\mathcal{G}; B) = \frac{1}{|B|} \sum_{x \in B} \phi(\tau_x(\mathcal{G}), x). \quad (2)$$

These criteria are used for Pareto-based validation and selection of candidate systems; the detailed evaluation procedure is provided in App. C.1.

3.2 Meta-Adaptive Joint Evolution

Previous work typically improves only individual components of an agent system and overlooks the compatibility between the external harness and the internal reasoning policy. To remedy this gap, HarnessForge adopts a meta-adaptive joint evolution mechanism over iterative rounds. The two evolution processes are mutually reinforcing: *better harnesses induce more structured and informative trajectories, while stronger reasoning policies execute harness protocols more faithfully.*

Round Structure. At round r , HarnessForge maintains a population of the agent systems $\mathbb{G}^{(r)}$:

$$\mathbb{G}^{(r)} = \{\mathcal{G}_i^{(r)} = (\mathcal{H}_i^{(r)}, \mathcal{R}_{\delta_i}^{(r)})\}_{i \in I^{(r)}}, \quad (3)$$

where each element is an executable harness-policy pair. The initial round $r = 0$ starts from a singleton population containing a manually designed base harness and the frozen base reasoner.

Given the evolution batch data B_r , HarnessForge first performs harness tailoring. For each agent, $\mathcal{G}_i^{(r)}$ executes tasks on B_r and collects trajectories, execution statistics, and environment feedback. A meta-agent tailoring operator T_ψ then updates the harness population through controlled executable harness editing, including fault attribution, archive-guided improvement, candidate generation, and budgeted Pareto selection:

$$\mathcal{C}^{(r+1)} = T_\psi(\{\mathcal{G}_i^{(r)}\}_{i \in I^{(r)}}, \mathcal{Z}^{(r)}, B_r), \quad (4)$$

where $\mathcal{C}^{(r+1)}$ is the survivor harness set retained for the next policy-evolution stage.

Conditioned on the survivor harness set $\mathcal{C}^{(r+1)}$, HarnessForge performs policy alignment for each harness through a policy-evolution operator E_η :

$$\mathcal{R}_{\delta_k}^{(r+1)} = E_\eta((\mathcal{H}_k^{(r+1)}, \mathcal{R}_{\delta_k}^{(r)}), B_r) \quad (5)$$

Unlike traditional post-training, the goal of policy evolution is to improve compatibility between the reasoning policy and the execution harness instead of optimizing a universally stronger reasoner.

Overview. At a higher level, each round alternates between (i) evolving harness structures from

trajectory-level execution evidence, and (ii) evolving harness-conditioned reasoning policies from the resulting survivor population:

$$\begin{aligned} \mathbb{G}^{(r)} &= \{(\mathcal{H}_k^{(r)}, \mathcal{R}_{\delta_k}^{(r)})\} \rightarrow \{(\mathcal{H}_k^{(r+1)}, \mathcal{R}_{\delta_k}^{(r)})\} \\ &\rightarrow \{(\mathcal{H}_k^{(r+1)}, \mathcal{R}_{\delta_k}^{(r+1)})\} = \mathbb{G}^{(r+1)}. \end{aligned} \quad (6)$$

By iterating this co-evolution process, both the execution harness and the governing reasoning policy evolve jointly, yielding increasingly adaptive agent systems over time. The implementation details of harness evolution and policy evolution are presented in Sec. 3.3 and Sec. 3.4, respectively.

3.3 Fault-Guided Harness Tailoring

The main challenge in agent system evolution is to locate which part of the harness causes failed execution. We therefore introduce a locate-and-refine mechanism to tailor better harnesses. We use $T_\psi = (\mathbb{L}_\omega, \mathbb{R}_\omega, \Gamma_\omega)$ to denote the overall evolution pipeline, where $\mathbb{L}_\omega$ performs fault attribution, $\mathbb{R}_\omega$ produces archive-guided improvement reports, Γ_ω generates revised harness candidates.

Fault Attribution. For each active system $\mathcal{G}_i^{(r)} = (\mathcal{H}_i^{(r)}, \mathcal{R}_{\delta_i}^{(r)})$, HarnessForge first evaluates $\mathcal{G}_i^{(r)}$ on batch B_r , producing rollout traces $\mathcal{T}_i^{(r)}$ and the evaluation vector $\mathbf{J}(\mathcal{G}_i^{(r)}; B_r)$ defined in Eq. 2. A meta-agent then performs fault-attribution operation $\mathbb{L}_\omega$ by jointly inspecting the current harness design and its representative failure trajectories:

$$\mathbf{F}_{\mathcal{H}_i}^{(r)} = \mathbb{L}_\omega(\mathcal{H}_i^{(r)}, \mathcal{T}_i^{(r)}, \mathbf{J}(\mathcal{G}_i^{(r)}; B_r)), \quad (7)$$

where $\mathbf{F}_{\mathcal{H}_i}^{(r)}$ is the fault report, attributing failures to planning, action and memory components.

Archive-Guided Improvement. HarnessForge maintains an archive $\mathcal{Z}^{(r)}$ of historical harnesses, storing compact summaries of harness designs and their corresponding evaluation vectors $\mathbf{J}$. Given the current harness $\mathcal{H}_i^{(r)}$ and its fault report $\mathbf{F}_{\mathcal{H}_i}^{(r)}$, the meta-agent samples exemplar cases $\mathcal{S}_{\mathcal{H}_i}^{(r)} \subset \mathcal{Z}^{(r)}$ based on fault relevance and Pareto-front quality, and produces an improvement report:

$$\mathbf{I}_{\mathcal{H}_i}^{(r)} = \mathbb{R}_\omega(\mathcal{H}_i^{(r)}, \mathbf{F}_{\mathcal{H}_i}^{(r)}, \mathcal{S}_{\mathcal{H}_i}^{(r)}). \quad (8)$$

The improvement report summarizes the likely improvement directions for the current harness, such as which component should be edited and which historical designs should be referenced. It serves as the input to the subsequent generation operator.

Refine-and-Filter. Based on the improvement report, HarnessForge generates revised harness candidates. For each active harness $\mathcal{H}_i^{(r)}$, the generation operator Γ_ω proposes K_{gen} revised harnesses:

$$\mathcal{C}_{\mathcal{H}_i}^{(r)} = \Gamma_\omega \left(\mathcal{H}_i^{(r)}, \mathbf{I}_{\mathcal{H}_i}^{(r)} \right), \quad |\mathcal{C}_{\mathcal{H}_i}^{(r)}| = K_{\text{gen}}. \quad (9)$$

The operator only edits the execution-layer components $\mathcal{P}$, $\mathcal{A}$, and $\mathcal{M}$. Let $\mathcal{C}_0^{(r)} = \bigcup_i \mathcal{C}_{\mathcal{H}_i}^{(r)}$ be the pooled candidate set. Since fully evaluating every candidate is expensive, HarnessForge applies *half-selection* over progressively larger task subsets. At filtering stage t , each candidate harness $\mathcal{H} \in \mathcal{C}_{t-1}^{(r)}$ is paired with the corresponding policy to form an executable system $\mathcal{G}_{\mathcal{H}}$ and evaluated on $B_{r,t}$ using the batch fitness $\mathbf{J}(\mathcal{G}_{\mathcal{H}}; B_{r,t})$. Candidates are selected according to Pareto optimality over the dimensions of this evaluation vector, and the retained subset becomes $\mathcal{C}_t^{(r)}$. After filtering, HarnessForge updates the archive $\mathcal{Z}^{(r+1)}$ with each evaluated harness and its evaluation vector $\mathbf{J}(\mathcal{G}_{\mathcal{H}}; B_{r,t})$, enabling later Pareto-aware retrieval. The survivor harnesses $\mathcal{C}^{(r+1)} = \mathcal{C}_T^{(r)}$ are then passed to policy evolution. Implementation details of the whole tailoring process are provided in App. C.4.

3.4 Harness-Conditioned Policy Alignment

After harness evolution, each survivor harness must be paired with an executor that can reliably operate under its evolved planning, action, and memory interface. We relabel the survivor set $\mathcal{C}^{(r+1)} = \{\mathcal{H}_k^{(r+1)}\}_k$ by lineage, so each survivor inherits the corresponding round- r parent policy $\mathcal{R}_{\delta_k}^{(r)}$. Its goal is not to train a universally stronger reasoner, but to align the inherited policy with the execution conventions induced by a particular harness.

Parent Initialization. For each survivor harness $\mathcal{H}_k^{(r+1)}$, HarnessForge initializes the child policy from its parent lineage and trains a new harness-specific LoRA (Hu et al., 2021) update, denoted as $\delta_k^{(r+1)} = \text{Merge}(\delta_k^{(r)}) \oplus \Delta\delta_k^{(r+1)}$. Here, $\text{Merge}(\cdot)$ folds the parent adapter into the child initialization, while $\Delta\delta_k^{(r+1)}$ adapts the policy to the evolved harness. Details are provided in App. D.1.

Trajectory Curation. Under the same rollout budget, policy evolution should not introduce a separate data-collection stage. HarnessForge therefore reuses the rollout pool already produced when $\mathcal{H}_k^{(r+1)}$ is evaluated during budgeted harness selection. Denote this pool by $\mathcal{T}_k^{(r+1)}$. We keep the

successful trajectories:

$$\mathcal{T}_k^+ = \{\tau_x \in \mathcal{T}_k^{(r+1)} \mid S(\tau_x) = 1\}. \quad (10)$$

Here $S(\tau_x)$ is the task-success indicator used in evaluation. This keeps trajectory curation tied to the same success signal that determines the harness success rate, while avoiding extra rollout cost.

Harness-Conditioned Evolution. HarnessForge converts the retained successful rollouts into step-level supervision for the selected harness. Each trajectory $\tau_x \in \mathcal{T}_k^+$ is decomposed into decision pairs (z_t, y_t) over its time steps, yielding the harness-conditioned dataset $\mathcal{D}_{\mathcal{H}_k} = \{(z_t, y_t)\}$ used by the alignment loss. The input $z_t = (x, \mathcal{H}_k^{(r+1)}, o_{\leq t}, m_t, a_t)$ packages the task instruction, active harness interface, accumulated observations, current memory state, and available actions; the target y_t is the corresponding next behavior, such as a reasoning step, tool action, memory operation, or final response. More general policy-update objectives can instantiate this stage, as discussed in App. D. In the main experiments, HarnessForge uses supervised trace alignment because it reuses the successful trajectories above and offers a favorable rollout–performance tradeoff:

$$\Delta\delta_k^{(r+1)} = \arg \min_{\Delta\delta} \mathcal{L}(\mathcal{R}_{\delta_k}^{(r)} \oplus \Delta\delta; \mathcal{D}_{\mathcal{H}_k}). \quad (11)$$

This objective learns only an incremental adapter on top of the inherited policy, aligning the executor to $\mathcal{H}_k^{(r+1)}$ without spending additional rollouts. The result is a matched next-round harness–policy pair $\mathcal{G}_k^{(r+1)} = (\mathcal{H}_k^{(r+1)}, \mathcal{R}_{\delta_k}^{(r+1)})$, which enters the next-round population rather than being treated as a general-purpose model upgrade.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate HarnessForge on diverse benchmarks: ToolHop (Ye et al., 2025), RestBench-TMDB (Song et al., 2023), and API-Bank (Li et al., 2023) measure tool selection, and API-grounded execution capabilities. SearchQA, built from HotpotQA (Yang et al., 2018) and 2Wiki-MultiHopQA (Ho et al., 2020), evaluates retrieval-heavy multi-hop question answering. Detailed descriptions and dataset statistics are in App. B.

Implementation Details. We instantiate HarnessForge with Qwen3-4B and Qwen3-8B (Yang et al., 2025) as the default backbones to evaluate its

Method	ToolHop		SearchQA			TMDB		API-Bank		
	Ans.	Path	Hotpot	2Wiki	Overall	Succ.	Path	Succ.	Path	API.
Qwen3-4B										
ADAS	40.00	44.99	28.67	28.00	28.33	45.00	57.43	51.75	57.31	60.99
AgentSquare	29.23	39.46	29.33	30.00	29.67	35.00	49.83	39.47	47.82	46.10
AFlow	31.28	43.81	30.67	29.33	30.00	32.00	47.64	37.72	45.72	41.13
MaAS	42.05	51.80	30.67	30.00	30.33	37.00	46.57	52.63	62.74	63.83
MermaidFlow	44.62	58.74	31.33	28.00	29.67	39.00	51.95	54.39	64.33	60.28
SFT	45.13	63.90	36.67	38.67	37.67	61.00	75.70	69.30	75.10	73.76
RLOO	46.15	64.87	35.33	40.00	37.67	61.00	77.42	72.81	78.82	76.60
GRPO	49.74	66.41	36.00	43.33	39.67	64.00	76.67	71.93	76.46	75.18
HarnessForge	52.82	68.10	42.00	42.00	42.00	76.00	85.10	77.19	80.07	82.27
Qwen3-8B										
ADAS	42.05	61.32	31.33	32.00	31.67	51.00	60.41	54.39	60.14	59.57
AgentSquare	30.77	41.22	29.33	30.00	29.50	38.00	53.83	45.61	54.37	53.90
AFlow	32.82	54.12	31.33	32.67	32.00	39.00	51.28	43.86	49.74	47.52
MaAS	44.62	51.80	32.00	32.00	32.00	43.00	54.78	57.89	66.72	60.28
MermaidFlow	47.39	59.31	32.67	30.67	31.67	47.00	58.67	57.02	67.46	62.41
SFT	48.72	72.20	40.67	41.33	41.00	69.00	82.92	68.42	73.80	73.05
RLOO	50.77	72.75	40.00	39.33	39.67	74.00	84.25	66.67	73.76	71.63
GRPO	51.28	75.62	41.33	42.00	41.67	70.00	83.20	69.30	74.71	73.76
HarnessForge	54.87	74.05	41.33	44.33	42.83	80.00	88.75	74.56	78.22	78.01

Table 1: Main results across benchmark groups for the Qwen3-4B and Qwen3-8B backbones. pale blue rows indicate search-style methods, soft blue rows indicate training-style methods, and blue-purple rows indicate our HarnessForge framework. Detailed benchmark and baseline configurations are provided in App. B and App. E.

effectiveness. The meta-agent used for harness evolution is GPT-5.5. We run $r = 3$ evolution rounds and report the main setting that retains $|C| = 2$ survivor harnesses per round. Additional implementation details are provided in App. C.

Baselines. We compare HarnessForge against two main baseline groups: Search-style harness baselines include AFlow (Zhang et al., 2025c), ADAS (Hu et al., 2025), AgentSquare (Shang et al., 2025), MaAS (Zhang et al., 2025a), and MermaidFlow (Zheng et al., 2025a). Training-style baselines include SFT, GRPO (Shao et al., 2024), and RLOO (Ahmadian et al., 2024). Detailed configurations of baselines are provided in App. E.

4.2 Main Results

Tab. 1 compares HarnessForge with harness-search and policy-training baselines across five agentic benchmark and two backbone sizes, averaging +3.56% over per-metric strongest baselines. HarnessForge delivers strong performance gains and reaches SOTA results on most benchmarks, spanning both tool-use and retrieval settings. Notably, the policy-training baselines RLOO and GRPO require larger rollout budgets than HarnessForge (App. E.3), yet still fall behind on most metrics. The largest gains appear on TMDB: HarnessForge

Round	Variant	ToolHop	SearchQA
Round0	Vanilla	41.03	34.33
Round1	HarnessForge	46.15	38.67
	-w/o Harness Evo.	43.08 ($\downarrow 3.07$)	35.67 ($\downarrow 3.00$)
	-w/o Policy Evo.	44.62 ($\downarrow 1.53$)	36.33 ($\downarrow 2.34$)
Round2	HarnessForge	50.77	40.33
	-w/o Harness Evo.	44.62 ($\downarrow 6.15$)	36.33 ($\downarrow 4.00$)
	-w/o Policy Evo.	48.72 ($\downarrow 2.05$)	38.33 ($\downarrow 2.00$)
Round3	HarnessForge	52.82	42.00
	-w/o Harness Evo.	46.67 ($\downarrow 6.15$)	37.00 ($\downarrow 5.00$)
	-w/o Policy Evo.	50.26 ($\downarrow 2.56$)	39.00 ($\downarrow 3.00$)

Table 2: Module ablation of HarnessForge using Qwen3-4B. ToolHop reports final-answer correctness, and SearchQA reports the macro-average answer F1.

improves success by 12.00% with Qwen3-4B and 6.00% with Qwen3-8B over the strongest baseline. On API-Bank, it improves API accuracy by average 4.96% across backbones. It also remains strong on reasoning-heavy benchmarks, with average 3.34% ToolHop answer gains across backbones and best SearchQA overall scores 42.83%.

4.3 Ablation Study

Tab. 2 shows that removing either harness tailoring or policy alignment consistently degrades performance on both ToolHop and SearchQA, indicating that both modules are necessary for Harness-

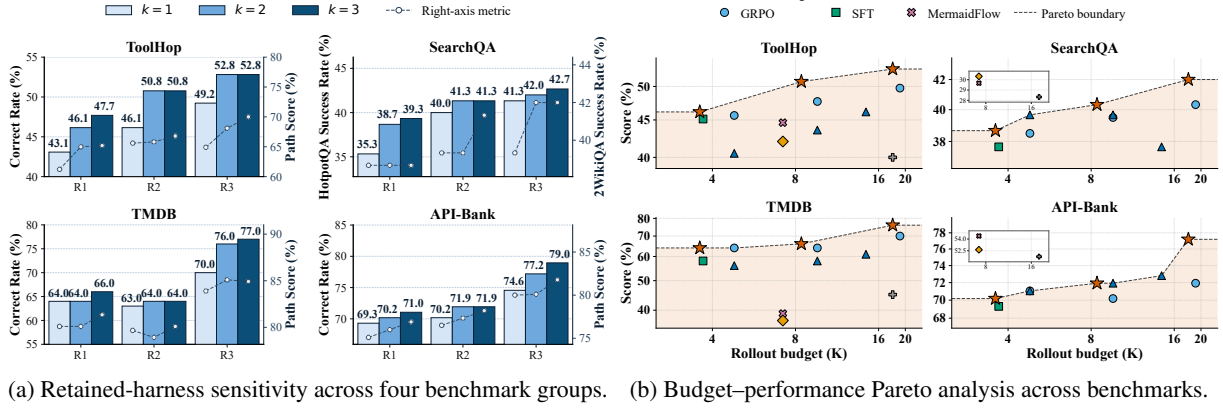


Figure 3: Framework analysis of HarnessForge. **Left:** performance sensitivity to the number of retained harnesses per evolution round. **Right:** rollout-budget efficiency compared with alternative adaptation settings.

Forge. Harness tailoring is the dominant factor: disabling it causes the largest drop in every round, and the gap becomes larger as evolution proceeds, increasing from -3.07%/-3.00% in Round 1 to -6.15%/-5.00% in Round 3 on ToolHop/SearchQA. Removing policy alignment also hurts performance, with final-round drops of -2.56% on ToolHop and -3.00% on SearchQA, suggesting that the reasoner must adapt to the evolved execution interface. Overall, the widening gaps show that HarnessForge’s gains come from harness-policy co-evolution.

4.4 Framework Analysis

Retained-harness sensitivity. Fig. 3a studies how the survivor pool size affects harness evolution. Retaining a single harness is often too restrictive: at the final round, increasing from $k = 1$ to $k = 2$ improves the main metric by 3.6% on ToolHop, 0.7% on SearchQA, 6.0% on TMDB, and 2.6% on API-Bank, averaging 3.2% points. Further increasing to $k = 3$ brings only marginal additional gains on most benchmarks, suggesting that excessive retention weakens selection pressure. Across rounds, $k = 2$ also yields consistent improvements from R1 to R3 on the four benchmarks groups. These results indicate that a small survivor population preserves useful harness diversity while still maintaining effective selection, enabling agent systems to emerge through population-level harness-policy evolution rather than one-shot scaffold optimization.

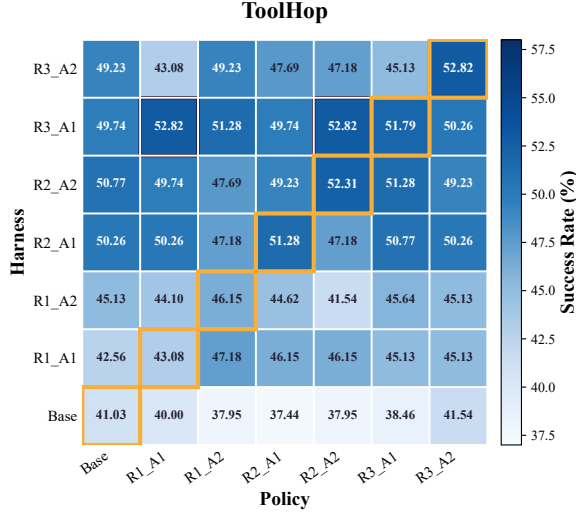
Rollout-budget efficiency. Fig. 3b compares different methods under varying rollout budgets. HarnessForge consistently lies on or near the Pareto frontier across all four benchmark groups, indicating that the proposed co-evolution process is both performance-effective and rollout-efficient. Unlike policy-only training methods that must

Round	Training	ToolHop		API-Bank		Budget (K)
		Ans.	Path	Succ.	Path	
1	SFT	45.13	60.66	69.30	74.85	73.05
	GRPO	45.64	59.49	70.18	76.02	74.47
	RLOO	46.67	61.17	70.18	75.73	73.76
2	SFT	48.72	63.70	71.05	75.15	75.89
	GRPO	49.23	64.28	69.30	76.90	73.76
	RLOO	49.74	63.80	71.05	77.78	75.18
3	SFT	50.77	65.87	71.05	76.90	75.18
	GRPO	52.31	67.32	71.93	77.78	75.18
	RLOO	51.28	66.45	72.80	79.09	76.60

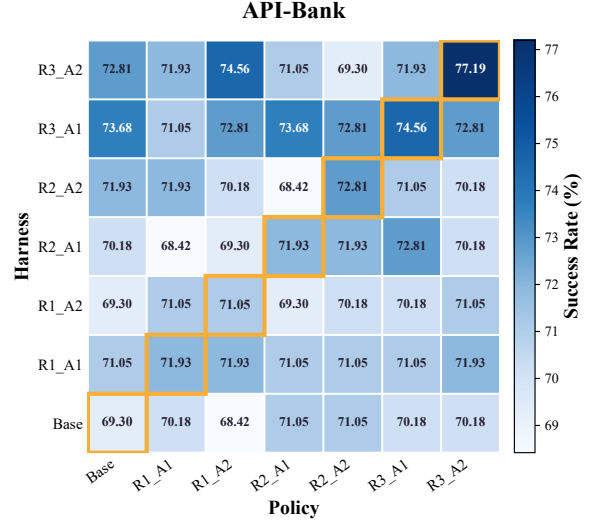
Table 3: Training agnostic analysis. We evaluate HarnessForge+SFT, HarnessForge+GRPO, and HarnessForge+RLOO on ToolHop, and report API-Bank as an auxiliary transfer/evaluation setting.

spend additional rollouts to explore better behaviors under a fixed interface, HarnessForge reuses rollout evidence to improve the external harness and then aligns the policy to the selected interface. This allows useful structural revisions and harness-conditioned execution patterns to accumulate across rounds, reducing the need for large-scale exploration. Consequently, HarnessForge provides a stronger performance-budget tradeoff than other on-policy training-style baselines. The rollout budget above counts executable environment interactions, while meta-agent overhead is reported separately. Detailed statistics and sensitivity results are provided in Apps. F.1 and G.1.

Training-method agnosticism. HarnessForge is not tied to a specific policy-training method. In the main experiments, we use supervised fine-tuning (SFT) as the default instantiation because it directly internalizes harness-induced execution patterns from curated trajectories. Tab. 3 compares SFT, GRPO, and RLOO within the same harness-



(a) Compatibility matrix on ToolHop.



(b) Compatibility matrix on API-Bank.

Figure 4: Harness–policy cross-pairing analysis on (a) ToolHop and (b) API-Bank. R_iA_j denotes the j -th adaptation checkpoint within the i -th co-evolution round. Rows and columns correspond to harness and policy checkpoints, respectively. Orange boxes mark matched pairs from the same checkpoint, while off-diagonal cells evaluate mismatched cross-checkpoint combinations.

policy co-evolution framework. Across rounds, replacing SFT with GRPO or RLOO further improves several metrics, e.g., in Round 3 GRPO improves ToolHop answer accuracy from 50.77% to 52.31%, while RLOO improves API-Bank success from 71.05% to 72.80%. These gains, however, require substantially larger rollout budgets: Round-3 RL-style instantiations use 45.6K rollouts compared with 12.0K for SFT. This suggests that HarnessForge’s framework-level gains are not specific to SFT, while SFT remains a rollout-efficient default and RL-style objectives provide additional improvement potential at higher cost.

Adaptation necessity analysis. Fig. 4 cross-pairs every evolved harness checkpoint (rows) with every policy checkpoint (columns) on ToolHop and API-Bank. Here, BASE denotes the initial harness–policy pair, while R_iA_j denotes the checkpoint after the j -th adaptation step within the i -th co-evolution round. The orange diagonal therefore marks *matched* harness–policy pairs obtained at the same checkpoint, whereas off-diagonal cells evaluate *mismatched* pairs assembled from different checkpoints. From the base checkpoint to the final matched pair, the success rate increases from 41.03% to 52.82% on ToolHop and from 69.30% to 77.19% on API-Bank. More importantly, the final matched pair achieves the highest performance in both its row and column on both benchmarks. When the final harness is combined with earlier policies, performance averages only 46.92% on

ToolHop and 71.93% on API-Bank; when the final policy is combined with earlier harnesses, the corresponding averages are 46.93% and 71.06%. Thus, adapting only one side leaves a gap of approximately 5.9 percentage points on ToolHop and 5.3–6.1 points on API-Bank relative to the final matched pair. These cross-pairing results indicate that the gains cannot be attributed solely to a universally stronger harness or policy; instead, a substantial portion arises from checkpoint-specific harness–policy compatibility induced by co-evolution.

4.5 Case Study

Fig. 5 traces a representative ToolHop lineage of HarnessForge across three co-evolution rounds, each consisting of harness tailoring followed by policy alignment. Round 1 addresses tool-interface and grounding failures: the harness introduces tool-contract notes, schema preflight checks, and observation-grounded memory, contributing a 2.14% gain, while the subsequent adapter update adds 1.57%. Round 2 strengthens long-horizon evidence management through explicit evidence slots, mutable-field tracking, and support-record verification, with the harness and adapter updates yielding 2.51% and 2.10%, respectively. Round 3 introduces failure-aware memory retrieval that selects compact recovery experience according to task topology and execution phase, followed by gains of 0.94% and 1.11%. Overall, the lineage shows a progression from tool-use ground-

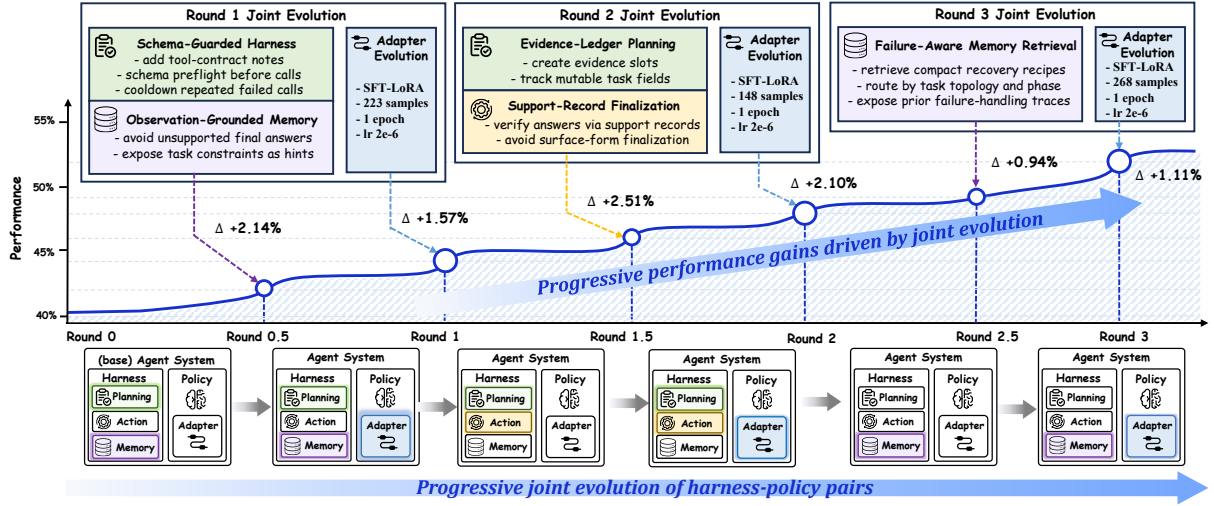


Figure 5: Representative ToolHop lineage of HarnessForge across three harness-policy co-evolution rounds. Each round alternates failure-driven harness tailoring with adapter-based policy alignment. The three rounds progressively improve tool-use grounding, evidence-consistent execution, and failure-aware memory retrieval. Dashed connectors associate each update with its incremental gain, while the lower panels indicate the component updated at each checkpoint.

ing, to evidence-consistent execution, and finally to experience-based recovery, with policy adaptation consistently helping the model exploit each evolved harness. Additional lineages are provided in App. G.3.

5 Conclusion

We introduced HarnessForge, a meta-adaptive framework that reformulates LLM agent adaptation as harness-policy pair evolution. Instead of optimizing external workflows or internal policies in isolation, HarnessForge co-evolves the execution harness and the reasoning policy through fault-guided harness tailoring and harness-conditioned policy alignment. Experiments across diverse agent benchmarks show consistent gains over diverse baselines, favorable rollout-performance tradeoffs, and strong matched-pair compatibility. These results demonstrate that effective agent-system adaptation depends on optimizing the executable compatibility between the harness and the policy.

Limitations

HarnessForge is primarily evaluated with Qwen3-4B and Qwen3-8B backbones. This setting is important for resource-constrained agent deployment, where the coupling between the external harness and the internal policy is especially salient: the harness can provide structural support for limited model capabilities, while the policy must learn to execute the harness-induced paradigm. Whether

the same magnitude of harness-policy compatibility gains holds for substantially larger frontier-scale models remains an open direction.

HarnessForge also requires repeated rollouts for harness profiling, selection, and policy alignment. Although our design reuses rollout trajectories across harness evolution and policy alignment to improve rollout efficiency, long-horizon environments can still make the evolution process costly. Future work could reduce this cost through proxy evaluation, adaptive rollout allocation, or learned early-stopping criteria.

Finally, the HarnessForge currently uses a structured meta-evolution protocol with constrained edit operators over planning, action, and memory components. This design improves executability and makes the evolution process auditable, but it does not exhaustively explore the full space of possible agent-system implementations, such as arbitrary code-level harness rewrites, new tool abstractions, or learned verifier modules. Future work could extend the operator space and study cost-aware meta-evolution with alternative closed- or open-source meta-agents.

Acknowledgments

This work is supported by the National Science Foundation of China (62401327, 62672026), and the Beijing Major Science and Technology Project under Contract no. Z25110008125031. This work was supported by Beijing Academy of Artificial Intelligence (BAAI).

References

- Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. [Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms](#). *Preprint*, arXiv:2402.14740.
- Mingju Chen, Guibin Zhang, Heng Chang, Yuchen Guo, and Shiji Zhou. 2026. [A-mapreduce: Executing wide search via agentic mapreduce](#). *Preprint*, arXiv:2602.01331.
- 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. 2026. [Agentic reinforced policy optimization](#). In *The Fourteenth International Conference on Learning Representations*.
- Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. [Group-in-group policy optimization for llm agent training](#). In *Advances in Neural Information Processing Systems*, volume 38, pages 46375–46408. Curran Associates, Inc.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. [MetaGPT: Meta programming for a multi-agent collaborative framework](#). In *The Twelfth International Conference on Learning Representations*.
- Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. 2025. [TreeRL: LLM reinforcement learning with on-policy tree search](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12355–12369, Vienna, Austria. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Shengran Hu, Cong Lu, and Jeff Clune. 2025. [Automated design of agentic systems](#). In *International Conference on Learning Representations*, volume 2025, pages 21344–21377.
- Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. [Search-rl: Training llms to reason and leverage search engines with reinforcement learning](#). *Preprint*, arXiv:2503.09516.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Yoonho Lee, Roshen Nair, Qizheng Zhang, Kangwook Lee, Omar Khattab, and Chelsea Finn. 2026. [Meta-harness: End-to-end optimization of model harnesses](#). *Preprint*, arXiv:2603.28052.
- Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. [API-bank: A comprehensive benchmark for tool-augmented LLMs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3102–3116, Singapore. Association for Computational Linguistics.
- Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. [Torl: Scaling tool-integrated rl](#). *Preprint*, arXiv:2503.23383.
- Xinghua Lou, Miguel Lázaro-Gredilla, Antoine Dedieu, Carter Wendelken, Wolfgang Lehrach, and Kevin P. Murphy. 2026. [Autoharness: improving llm agents by automatically synthesizing a code harness](#). *Preprint*, arXiv:2603.03329.
- Can Lv, Heng Chang, Shengyu Tao, Mingju Chen, Zhaoxin Fan, Ziwei Zhang, Yuchen Guo, and Shiji Zhou. 2026a. [All-mem: Agentic lifelong memory via dynamic topology evolution](#). *Preprint*, arXiv:2603.19595.
- Can Lv, Mingju Chen, Heng Chang, and Shiji Zhou. 2026b. [Mitigating false credit propagation: Probabilistic graphical reward aggregation for rubric-based reinforcement learning](#). *Preprint*, arXiv:2606.03361.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. [Memgpt: Towards llms as operating systems](#). *Preprint*, arXiv:2310.08560.
- Cheng Qian, Emre Can Acikgoz, Qi He, Hongru WANG, Xiushi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2026. [ToolRL: Reward is all tool](#)

- learning needs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Yu Shang, Yu Li, Keyu Zhao, Likai Ma, Jiahe Liu, Fengli Xu, and Yong Li. 2025. [Agentsquare: Automatic LLM agent search in modular design space](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory Wornell, Subhro Das, David Cox, and Chuang Gan. 2025. [Satori: Reinforcement learning with chain-of-action-thought enhances llm reasoning via autoregressive search](#). *Preprint*, arXiv:2502.02508.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. [Reflexion: language agents with verbal reinforcement learning](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Xiaoshuai Song, Haofei Chang, Guanting Dong, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2026. [Envscaler: Scaling tool-interactive environments for llm agent via programmatic synthesis](#). *Preprint*, arXiv:2601.05808.
- Yifan Song, Weimin Xiong, Dawei Zhu, Wenhao Wu, Han Qian, Mingbo Song, Hailiang Huang, Cheng Li, Ke Wang, Rong Yao, Ye Tian, and Sujian Li. 2023. [Restgpt: Connecting large language models with real-world restful apis](#). *Preprint*, arXiv:2306.06624.
- Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. [Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. [Autogen: Enabling next-gen LLM applications via multi-agent conversations](#). In *First Conference on Language Modeling*.
- Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2026. [Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning](#). *Preprint*, arXiv:2508.19828.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. [Webshop: Towards scalable real-world web interaction with grounded language agents](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc.
- Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. [\$\tau\$ -bench: A benchmark for tool-agent-user interaction in real-world domains](#). *Preprint*, arXiv:2406.12045.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. [React: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xu-anjing Huang, and Jiecao Chen. 2025. [ToolHop: A query-driven benchmark for evaluating large language models in multi-hop tool use](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2995–3021, Vienna, Austria. Association for Computational Linguistics.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, juncai liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2025. [Dapo: An open-source llm reinforcement learning](#)

system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pages 113222–113244. Curran Associates, Inc.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, LEI BAI, and Xiang Wang. 2025a. [Multi-agent architecture search via agentic supernet](#). In *Forty-second International Conference on Machine Learning*.

Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, Junhao Wang, He Zhu, Wangchunshu Zhou, and Shuicheng Yan. 2025b. [Memevolve: Meta-evolution of agent memory systems](#). *Preprint*, arXiv:2512.18746.

Huan Zhang, Mingju Chen, Dongxu Zhou, Can Lv, Heng Chang, Sen Cui, Faguo Wu, and Shiji Zhou. 2026. [Trca: Transition-wise rubric credit assignment for long-horizon llm agents](#). *Preprint*, arXiv:2608.16156.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. 2025c. [AFlow: Automating agentic workflow generation](#). In *The Thirteenth International Conference on Learning Representations*.

Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. 2025. [Absolute zero: Reinforced self-play reasoning with zero data](#). *Preprint*, arXiv:2505.03335.

Chengqi Zheng, Jianda Chen, Yueming Lyu, Wen Zheng Terence Ng, Haopeng Zhang, Yew-Soon Ong, Ivor Tsang, and Haiyan Yin. 2025a. [Mermaidflow: Redefining agentic workflow generation via safety-constrained evolutionary programming](#). *Preprint*, arXiv:2505.22967.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025b. [Group sequence policy optimization](#). *Preprint*, arXiv:2507.18071.

WanJun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. [Memorybank: Enhancing large language models with long-term memory](#). *Preprint*, arXiv:2305.10250.

Siyu Zhu, Yanbin Jiang, Hejian Sang, Shao Tang, Qingquan Song, Biao He, Rohit Jain, Zhipeng Wang, and Alborz Geramifard. 2025. [Planner-r1: Reward shaping enables efficient agentic rl with smaller llms](#). *Preprint*, arXiv:2509.25779.

A Algorithm and Notation

Algorithm 1 HarnessForge Co-evolution Round

Require: parent population $\mathcal{G}^{(r)}$, evolution batch B_r , archive $\mathcal{Z}^{(r)}$

- 1: **for** each parent pair $G_i^{(r)} = (H_i^{(r)}, R_{\delta_i}^{(r)})$ **do**
- 2: Roll out $G_i^{(r)}$ on B_r and collect traces $\mathcal{T}_i^{(r)}$
- 3: Generate fault report $F_{H_i}^{(r)}$
- 4: Retrieve archive cases and produce improvement brief $\mathbf{I}_{H_i}^{(r)}$
- 5: Generate K_{gen} child harnesses
- 6: Discard invalid children using interface and smoke tests
- 7: **end for**
- 8: Select survivor harnesses by staged Pareto filtering over $B_{r,1}, \dots, B_{r,T}$
- 9: **for** each survivor harness $H_k^{(r+1)}$ **do**
- 10: Reuse successful rollout traces from filtering to construct D_{H_k}
- 11: Materialize an independent copy of the parent lineage policy and train a new harness-specific adapter
- 12: Return matched pair $(H_k^{(r+1)}, R_{\delta_k}^{(r+1)})$
- 13: **end for**
- 14: Update archive with evaluated harnesses, metrics, reports, and selection logs

A.1 Notation

Tab. 4 summarizes the notation used in the main text and appendix.

B Datasets Details

B.1 Evaluation Benchmark and Metrics

The main experiments cover five datasets organized into four benchmark families. ToolHop, RestBench-TMDB, and API-Bank each form one benchmark family. SearchQA is the retrieval-heavy benchmark family and contains two datasets, HotpotQA and 2WikiMultiHopQA. We therefore use “five datasets” when describing dataset coverage and “four benchmark families” or “four benchmark groups” when referring to the columns in Tab. 1.

Let $\mathcal{D}$ denote the evaluated split, $N = |\mathcal{D}|$, $\hat{a}_i$ and a_i denote the predicted and gold final answers for instance i , and $\hat{\pi}_i$ and π_i denote the predicted and gold tool/API-call paths. We use $\mathbf{1}[\cdot]$ for the indicator function. For answer matching, $\text{norm}(\cdot)$ lowercases text and removes articles, punctuation,

Table 4: Notation used by HarnessForge.

Symbol	Meaning
$\mathcal{G} = (\mathcal{H}, \mathcal{R}_\delta)$	Agent system consisting of an executable harness and an adapted reasoner.
$\mathcal{H}$	Harness with planning, action, and memory components.
$(\mathcal{P}, \mathcal{A}, \mathcal{M})$	Frozen base reasoner before harness-conditioned adaptation.
$\mathcal{R}_{\theta_0}$	Lightweight adapter parameters associated with a policy lineage.
δ	Population of harness-policy pairs at evolution round r .
$\mathbb{G}^{(r)}$	Evolution batch at round r , and the subset used at filtering stage t .
$B_r, B_{r,t}$	Rollout traces collected by pair i at round r .
$\mathcal{T}_i^{(r)}$	Multi-objective evaluation vector over task performance and efficiency.
$\mathbf{J}(\mathcal{G}; B)$	Archive of evaluated harnesses, rollout summaries, diagnostics, and selection records.
$\mathcal{Z}^{(r)}$	Fault report and archive-guided improvement brief for harness $\mathcal{H}$.
$\mathbf{F}_{\mathcal{H}}^{(r)}, \mathbf{I}_{\mathcal{H}}^{(r)}$	Survivor harness set passed to policy alignment for the next round.
$\mathcal{C}^{(r+1)}$	

and extra whitespace, following standard QA evaluation.

ToolHop. ToolHop evaluates multi-hop tool use, where each instance requires the agent to decompose a question, call tools for intermediate evidence, and return a final answer (Ye et al., 2025). Our current held-out split contains 195 evaluated instances. We report **Correct**, the final-answer correctness judged by the benchmark evaluator. Let $m_i^{\text{ans}} = 1$ iff the normalized prediction matches the gold answer. Then

$$\text{Correct}_{\text{ToolHop}} = \frac{\sum_{i=1}^N m_i^{\text{ans}}}{N}, \quad (12)$$

and **Path**, the average fraction of required intermediate subgoals solved during the trajectory. If G_i is the set of required intermediate subgoals and $\hat{G}_i$ is the set credited by the evaluator, then

$$\text{Path}_{\text{ToolHop}} = \frac{1}{N} \sum_{i=1}^N \frac{|\hat{G}_i \cap G_i|}{|\hat{G}_i|}. \quad (13)$$

SearchQA. SearchQA evaluates retrieval-heavy multi-hop question answering over local evidence corpora constructed from HotpotQA and 2WikiMultiHopQA (Yang et al., 2018; Ho et al., 2020). We use the normalized token-level answer F1 score.

For token multisets $T(\hat{a}_i)$ and $T(a_i)$, let

$$\begin{aligned} P_i &= \frac{|T(\hat{a}_i) \cap T(a_i)|}{|T(\hat{a}_i)|}, \\ R_i &= \frac{|T(\hat{a}_i) \cap T(a_i)|}{|T(a_i)|}, \\ F1_i &= \frac{2P_i R_i}{P_i + R_i}. \end{aligned} \quad (14)$$

The reported HotpotQA and 2WikiMultiHopQA scores are

$$\begin{aligned} \text{Score}_d &= \frac{\sum_{i \in \mathcal{D}_d} F1_i}{|\mathcal{D}_d|}, \\ d &\in \{\text{Hotpot}, 2\text{Wiki}\}, \end{aligned} \quad (15)$$

and **Overall** is the macro-average over the two subsets:

$$\text{Overall}_{\text{SearchQA}} = \frac{\text{Score}_{\text{Hotpot}} + \text{Score}_{2\text{Wiki}}}{2}. \quad (16)$$

RestBench-TMDB. RestBench-TMDB evaluates REST-style API use over a movie-database environment (Song et al., 2023). Each task specifies an information need that must be satisfied through one or more API calls. Our current evaluation split contains 100 instances. We report **Success**, the fraction of tasks whose final answer satisfies the verifier,

$$\text{Success}_{\text{TMDB}} = \frac{\sum_{i=1}^N v_i}{N}, \quad v_i \in \{0, 1\}, \quad (17)$$

and **Path**, the evaluator’s path-rate metric for matching the required API execution path. Following the RestBench notion of a correct API path, we count a path as matched if the gold API-call sequence is preserved as an ordered subsequence of the predicted path:

$$\text{Path}_{\text{TMDB}} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\pi_i \preceq \hat{\pi}_i], \quad (18)$$

where $\preceq$ denotes ordered-subsequence matching.

API-Bank. API-Bank evaluates structured API calling across diverse user requests (Li et al., 2023). Each instance provides a user instruction and a gold API-call specification, including the function name and arguments. Our current evaluation split contains 114 instances. We report **Success**, the fraction of instances whose full call trajectory and

final response satisfy the evaluator,

$$\text{Success}_{\text{APIBank}} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\text{Eval}(\hat{\pi}_i, \hat{a}_i) = 1]. \quad (19)$$

We report **Path** as an ordered API-name overlap score. Let L_i be the length of the longest common subsequence between the predicted and gold API-name sequences, $P_i^\pi = L_i/|\hat{\pi}_i|$, and $R_i^\pi = L_i/|\pi_i|$. Then

$$\text{Path}_{\text{APIBank}} = \frac{1}{N} \sum_{i=1}^N \frac{2P_i^\pi R_i^\pi}{P_i^\pi + R_i^\pi}. \quad (20)$$

Finally, **API Accuracy** measures call-level exactness after aligning predicted calls to gold calls. Let $m_{i,t}^{\text{api}} = 1$ iff the aligned predicted call has the correct API name and schema-normalized arguments, i.e., $\text{name}(\hat{c}_{i,t}) = \text{name}(c_{i,t})$ and $\text{args}(\hat{c}_{i,t}) \simeq \text{args}(c_{i,t})$. Then

$$\text{APIAcc}_{\text{APIBank}} = \frac{1}{\sum_i |\pi_i|} \sum_{i=1}^N \sum_{t=1}^{|\pi_i|} m_{i,t}^{\text{api}}, \quad (21)$$

where $\simeq$ denotes schema-normalized argument equivalence.

B.2 Training & Evolution Data

We construct a 3.8K(3,800)-task training pool covering three complementary agent capabilities: general tool use, complex environment interaction, and offline retrieval. The training pool is strictly disjoint from all held-out evaluation splits used in the main experiments. Training tasks are used only for HarnessForge harness evolution, trajectory curation, and policy-adapter training, as well as for training the training-style baselines (SFT, GRPO, and RLOO); the test splits are reserved for final evaluation.

- **Complex Environment Interaction:** We sample 2.0k tasks from EnvScaler-RL (Song et al., 2026). Each task provides an executable environment, an initial configuration, a task instruction, available tools, and verifier-based feedback. This subset targets multi-step tool use in stateful environments.
- **General Tool-Use:** We sample 0.8k instances from ToolHop (Ye et al., 2025). These tasks emphasize tool selection, tool chaining, and precise function-call execution.

- **Offline QA & Retrieval:** We sample 1.0k tasks from Wikipedia-based QA datasets, including Natural Questions (NQ) (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), and 2WikiMultiHopQA (Ho et al., 2020). In our setup, we convert these datasets into an offline/local-corpus retrieval setting: each question is paired with a fixed local document collection, and the model must retrieve supporting evidence from that corpus before answering. This subset trains reproducible evidence retrieval and answer generation without relying on live web access.

Overall, the resulting training pool contains **3.8k samples** and provides a compact mixture of tool-use, environment interaction, and offline retrieval tasks.

B.3 Split and Deduplication Protocol

All reported test scores are computed on held-out evaluation splits. The 3.8K training and evolution pool is used for harness evolution, trajectory curation, and policy-adapter training, but is disjoint from the held-out test splits used for final reporting. When a source dataset contributes to both training/evolution and evaluation, we use non-overlapping split identifiers and remove potential duplicates by dataset identifier and normalized task instruction. For converted offline retrieval tasks, the local evidence corpus is fixed before adaptation and the held-out question identifiers are not exposed during harness search, policy training, or model selection.

C Harness Tailoring Details

C.1 Evaluator Setting

HarnessForge uses a multi-objective evaluator to select harness candidates and retrieve useful historical cases from the archive. We reuse the benchmark metrics defined in App. B as the task-performance dimension, and add efficiency dimensions for candidate selection. For each candidate agent system $\mathcal{G}$ and evaluation batch B , we compute

$$\mathbf{J}(\mathcal{G}; B) = (\text{Perf}(\mathcal{G}; B), -\text{Tok}(\mathcal{G}; B), -\text{Delay}(\mathcal{G}; B)), \quad (22)$$

where larger values are preferred in every dimension. Perf denotes benchmark-specific task performance, Tok denotes token usage, and Delay

denotes wall-clock latency. The negative signs convert efficiency objectives into maximization dimensions.

The performance term is instantiated according to the domain of each task. Our evolution data mainly covers ToolHop, EnvScaler-RL, and SearchQA. For ToolHop, we combine the final-answer correctness and path-completion metrics defined in App. B:

$$\text{Perf}_{\text{ToolHop}}(\tau, x) = \lambda_{\text{ans}} \mathbb{1}[\text{AnsCorrect}(\tau, x)] + \lambda_{\text{path}} \text{PathScore}(\tau, x), \quad (23)$$

where $\lambda_{\text{ans}} + \lambda_{\text{path}} = 1$, and we use $\lambda_{\text{ans}} = \lambda_{\text{path}} = 0.5$ by default.

For SearchQA, we use the normalized token-level answer F1 score defined in App. B:

$$\text{Perf}_{\text{SearchQA}}(\tau, x) = \text{F1}(\hat{a}(\tau), a_x), \quad (24)$$

where $\hat{a}(\tau)$ is the final answer produced by trajectory τ and a_x is the gold answer.

For EnvScaler-RL, we follow the official environment-state verifier. Each task scenario provides a checklist of terminal-state validation functions $\mathcal{C}_x = \{c_j^x\}_{j=1}^{m_x}$, where each function checks whether one required condition is satisfied in the final environment state. Let $s_T(\tau, x)$ denote the terminal state reached after executing trajectory τ . We define

$$\text{Perf}_{\text{Env}}(\tau, x) = \text{Done}(\tau, x) = \frac{1}{m_x} \sum_{j=1}^{m_x} \mathbb{1}[c_j^x(s_T(\tau, x)) = \text{True}]. \quad (25)$$

This score measures task completion by verifying the final environment state, rather than matching a single reference action sequence.

For a batch B , each metric is averaged over all tasks in the batch:

$$\text{Perf}(\mathcal{G}; B) = \frac{1}{|B|} \sum_{x \in B} \text{Perf}_{d(x)}(\tau_x(\mathcal{G}), x), \quad (26)$$

where $d(x)$ denotes the benchmark domain of task x .

Given two candidate harnesses $\mathcal{H}_a$ and $\mathcal{H}_b$, paired with their corresponding policies to form executable systems $\mathcal{G}_a$ and $\mathcal{G}_b$, we say $\mathcal{H}_a$ Pareto-dominates $\mathcal{H}_b$ on batch B if

$$\mathcal{H}_a \succ_B \mathcal{H}_b \iff \mathbf{J}(\mathcal{G}_a; B) \succeq \mathbf{J}(\mathcal{G}_b; B) \wedge \mathbf{J}(\mathcal{G}_a; B) \neq \mathbf{J}(\mathcal{G}_b; B). \quad (27)$$

That is, $\mathcal{G}_a$ is no worse in every objective and strictly better in at least one objective. During budgeted filtering, HarnessForge ranks candidates by Pareto fronts and uses the primary task performance as the tie-breaker when candidates are otherwise comparable. The retained Pareto-competitive harnesses form the survivor set passed to policy alignment.

The same evaluator is also used to maintain the archive. For each evaluated harness, HarnessForge stores its design summary, fault report, rollout summary, and evaluation vector $\mathbf{J}$. During archive-guided improvement, reference cases are retrieved based on both fault relevance and Pareto quality: the meta-agent prioritizes historical harnesses that address similar planning, action, or memory failures and lie on or near the Pareto frontier. Thus, the evaluator provides a unified criterion for both harness selection and archive retrieval.

C.2 Hyperparameter Configuration

Tab. 5 summarizes the main implementation configuration used by HarnessForge. Unless otherwise stated, the same configuration is used across benchmarks and backbones.

Table 5: Implementation configuration of HarnessForge.

Parameter	Value
Base reasoner	Qwen3-4B and Qwen3-8B
Meta-agent	GPT-5.5
Evolution rounds	3
Harness proposals per round	8 candidates
Filtering stages	Two-stage half-selection
Retained harnesses	$ C = 2$ survivor harnesses per round
Harness components	Planning, Action, Memory
Tailoring target	Harness edits under task-interface constraints
Adapter type	Harness-specific LoRA adapter
Trainable parameters	LoRA adapters only; base model frozen
Adapter objective	SFT on filtered successful trajectories
LoRA target modules	Attention and MLP projections
LoRA rank / alpha / dropout	8 / 16 / 0.05
Adapter learning rate	2×10^{-6}
Adapter epochs	1

C.3 Harness Representation and Edit Space

Each harness is stored as an executable code/configuration bundle. The editable scope is restricted to the harness controller, so all candidate children can be run under the same evaluator. Tab. 6 lists the main edit categories used by the meta-agent.

Table 6: Executable harness representation and edit boundaries.

Component	Editable harness-side fields
Planning $\mathcal{P}$	Decomposition templates, replanning triggers, verification steps, termination conditions, and controller order.
Action $\mathcal{A}$	Tool descriptions, argument schemas, routing rules, validity guards, retry policies, and role/tool orchestration.
Memory $\mathcal{M}$	State keys, write conditions, retrieval rules, summarization format, and which memory fields are exposed to the reasoner.

A proposed child harness must instantiate the same controller interface as its parent and pass a lightweight validity check before rollout evaluation: required fields must be present, tool schemas must be parseable, action names must map to available tools, and memory keys referenced by planning or action code must be defined. Invalid children are discarded before budgeted selection. This representation makes the parent harness $\rightarrow$ fault report $\rightarrow$ child harness transition reproducible as a constrained code-editing problem rather than an unconstrained natural-language redesign.

C.4 Meta Tailoring Operator and Prompt Protocol

The meta tailoring operator is implemented as a staged prompt-driven meta-agent pipeline. Instead of relying on a single monolithic generation call, the pipeline decomposes tailoring into diagnosis, improvement planning, executable harness generation, and smoke-test repair.

This staged design improves controllability and noise isolation: each stage consumes a compact, task-specific intermediate artifact rather than the full long-context evidence bundle, which reduces context drift and prevents later generation steps from being dominated by irrelevant trajectory details. Concretely, one call diagnoses failures, one converts the diagnosis into transferable improvement directions, one writes executable harness code, and a bounded retry loop repairs implementation errors discovered by smoke tests. The full prompt for this operation is available in our open-source codebase, and the appendix includes a representative excerpt for reference.

C.4.1 Fault-Attribution Operation

The first stage localizes observed execution failures to specific harness modules. Instead of treating every wrong final answer as a generic model-reasoning failure, the meta-agent distinguishes

planning, action, and memory failures.

As shown in Eq. (7), the fault-attribution operation produces a structured report $\mathbf{F}_{\mathcal{H}_i}^{(r)}$ that localizes observed failures to harness modules. In brief, it instructs the meta-agent to analyze the current harness, rollout metrics, and representative trajectories, and to produce an evidence-grounded diagnosis rather than task-specific patches. The prompt requires the meta-agent to (i) reverse-engineer the implemented planning, action, memory, and wiring behavior, (ii) identify stable failure modes and their first meaningful error points, (iii) assign module ownership with concrete trajectory evidence, (iv) distinguish transferable harness weaknesses from benchmark-specific artifacts, and (v) output repair priorities, preservation constraints, and generation-ready instructions for the next tailoring stage.

Phase I Fault-Location Prompt (Partial)

```
# You are a Harness Failure Localization
Agent
Your goal is to analyze the execution
failures of the current winner harness
'{winner_harness_name}' and produce a
**module-localized, evidence-grounded,
generation-ready diagnosis of
transferable harness weaknesses**.
This is not a leaderboard summary and not a
request for task-specific patches.
This is a structural postmortem whose
main purpose is to answer:
**Which module-level capability gap is
responsible for each dominant failure
mode, and why is that diagnosis likely
to
transfer beyond the observed training tasks
?**
---
## Harness Under Analysis

### Winner / Current Best Harness
'{winner_harness_name}'

### Harness Snapshot
{winner_harness_snapshot}

### Module Files
'{module_files_info}'
You may inspect the harness implementation
files if tools are available. Pay
special attention to:
- 'builder.py'
- 'planning_module/provider.py'
- 'action_module/provider.py'
- 'memory_module/provider.py'
- prompt files under 'planning_module/', '
  action_module/', or 'memory_module/'
  if present
- 'Description.md' and prior reports if
  present
---
### PART 7: GENERATION CONSTRAINTS
Flat bullets only.
Each bullet must start with one of:
- '[Planning]'
- '[Action]'
- '[Memory]'
- '[Builder]'
- '[Interface]'
- '[Avoid]'
- '[Preserve]'
These should be operational constraints for
```

```

        the next harness generation step.
Include at least one '[Avoid]' constraint
that prevents benchmark-specific or
trajectory-specific patching.
---
### PART 8: READY SIGNAL
End with exactly:
```text
READY_FOR_IMPROVEMENT_DIRECTION
```
---
## Critical Guidelines
- Be evidence-first.
- Treat trajectories as evidence for
  transferable harness weaknesses, not
  as templates for special-case fixes.
- Do not collapse all errors into planning.
- Treat final-answer failure as different
  from path-search failure.
- Treat tool schema failure as different
  from reasoning failure.
- Separate module ownership from general
  blame.
- Prefer module-level capability diagnoses
  over small defensive patches.
- Avoid overfitting to the current training
  task distribution.
- Preserve working behaviors.
- Make the report directly usable by the
  next prompt.
"""

```

C.4.2 Archive-Guided Improvement

The second stage converts the fault report into a generation-ready improvement brief. Rather than generating the next harness from scratch, HarnessForge retrieves reference cases from the archive $\mathcal{Z}^{(r)}$, which stores previously explored harnesses, rollout summaries, validation scores, and diagnosis reports. Retrieval is conditioned on both fault relevance and Pareto quality: the meta-agent prioritizes historical harnesses that address similar planning, action, or memory failures and lie on or near the Pareto frontier of validation performance and execution cost. This allows the improvement stage to reuse transferable design patterns without overfitting to narrow task-specific patches.

We retrieve the reference cases as

$$\mathcal{S}_{\mathcal{H}_i}^{(r)} = \text{Retrieve} \left(\mathcal{Z}^{(r)}, \mathbf{F}_i^{(r)} \right), \quad (28)$$

where $\mathcal{S}_{\mathcal{H}_i}^{(r)}$ denotes the selected Pareto-competitive reference harnesses. The improvement agent then produces an implementation brief:

$$\mathbf{I}_i^{(r)} = \mathbb{R}_\omega \left(\mathcal{H}_i^{(r)}, \mathbf{F}_i^{(r)}, \mathcal{E}_i^{(r)} \right). \quad (29)$$

In brief, the prompt instructs the meta-agent to read the current harness, the module-localized fault report $\mathbf{F}_{\mathcal{H}_i}^{(r)}$, and the retrieved archive examples, then produce a transferable improvement brief rather than executable code. The brief specifies which failure modes to prioritize, which

planning/action/memory or cross-module interface should own each fix, which historical design patterns should be reused or avoided, and what preservation constraints should guide the subsequent harness-generation stage. This separates diagnosis, improvement abstraction, and code generation, reducing trajectory-level overfitting and making harness edits more transferable. Concretely, we instantiate this operation with the following prompt:

Phase II Improvement Prompt (Partial)

```

# You are a Harness Improvement Direction
Agent
Your goal is to read the Stage 1 module-
localization report for '{
  winner_harness_name}', inspect
relevant harness
examples from the harness archive / harness
pool, and produce a **generation-
ready improvement direction brief for
transferable harness improvements**.
This phase does **not** generate code.
It decides:
- which failure modes should be fixed first
- which module should own each fix
- which existing harness examples are
  useful as few-shot references
- what to borrow
- what to avoid
- how to keep directions transferable to
  unseen tasks
- what exact design direction the Stage 3
  generator should follow
---
## Inputs

### Winner / Current Best Harness
'{winner_harness_name}'

### Winner Harness Snapshot
{winner_harness_snapshot}

### Stage 1 Module Localization Report
{module_localization_report}

### Harness Pool / Archive Overview
{harness_pool_overview}

### Candidate Harness Examples
{harness_examples}

### Existing Harness Names / Identities
{existing_harness_names}
---
## System Boundaries
The generator can create or modify a
harness candidate inside the harness
factory, but it must respect the local
harness structure:
- 'builder.py'
- '__init__.py'
- 'Description.md'
- 'planning_module/provider.py'
- 'action_module/provider.py'
- 'memory_module/provider.py'
The generator may change internal module
behavior, prompts, helper functions,
and module wiring as long as it
remains compatible with the harness
factory.

### PART 8: READY SIGNAL
End with exactly:
```text
READY_FOR_HARNESS_GENERATION
```
---
## Critical Guidelines

```

- The Stage 1 localization report is authoritative.
- Preserve Stage 1's distinction between surface trajectory evidence and transferable harness weakness.
- Do not invent new failure modes unless the examples reveal a direct contradiction.
- Use peer harnesses selectively rather than copying a whole architecture.
- Prefer targeted module transfer over whole-architecture replacement.
- Prefer reusable module capabilities over narrow fallback patches.
- Preserve the winner's useful behaviors.
- Make the final blueprint specific enough for code generation.

C.4.3 Refinement and Generation

The third stage converts the improvement brief $I_i^{(r)}$ into executable harness candidates and filters them before policy alignment. The generator may revise the internal planning, action, memory, and wiring logic, but must preserve the dataset, evaluator, backend model, benchmark runner, and task labels. This constraint ensures that improvements come from harness tailoring rather than task-specific shortcuts.

In brief, it instructs the meta-agent to generate a production-ready harness candidate from the current harness, the module-localized fault report $F_i^{(r)}$, the archive-guided improvement brief $I_i^{(r)}$, and selected reference examples. The prompt requires the generated harness to be self-contained, importable, compatible with the expected builder interface, and organized into the required planning, action, memory, and wiring files. It also asks the meta-agent to implement evidence-grounded module-level repairs, preserve useful parent-harness behaviors, avoid cosmetic or unrelated changes, and prevent hard-coded benchmark shortcuts. After generation, candidates are checked by smoke tests and evaluated under a bounded rollout budget; only valid and Pareto-competitive survivors are passed to the policy-alignment stage.

Concretely, we instantiate this operation with the following prompt:

Phase III Generation Prompt (Partial)

```
## Stage 3 - Harness Generation

### Objective
Build one creative but evolutionary,
production-grade harness candidate
that improves the current winner by
following
the module-localized diagnosis and the
improvement direction brief.
The generated harness must be a complete
local harness bundle that can be
placed under:
'harness_factory/rounds/{target_round}/{
```

```
candidate_name}/{
Each generated harness candidate must be
independent and self-contained. It may
learn from the winner and selected
examples, but it must not depend on files
from sibling candidates, previous
generated candidates, or mutable
external
harness directories at runtime.
The generator should implement concrete
module-level repairs, not produce
another analysis report or a from-
scratch
unrelated harness.
---
### Context
**Harness Integration**:
A harness is composed of:
- 'builder.py' - wires the harness and
exposes 'build_agent_from_context(
context)'
- '__init__.py' - exports harness constants
and build entry point
- 'Description.md' - explains the harness
design
- 'planning_module/provider.py' - planning
implementation
- 'action_module/provider.py' - action /
tool-use implementation, including
concrete agent collaboration /
orchestration
- 'memory_module/provider.py' - memory
provider or memory wrapper
**Core Agent Constraint**:
The benchmark runner, backend model,
dataset, evaluator, and outer agent
loop are fixed unless the local
harness
examples clearly show an approved extension
point. The harness may evolve
internally through planning behavior,
action behavior, memory guidance, and
simple interfaces between them.
**Module Ownership**:
- Planning owns decomposition, plan
topology, progress state, verification
questions, evidence-chain intent,
replan triggers/timing, milestone gates,
and commit criteria.
- Action owns concrete agent collaboration
/ orchestration, agent topology, tool
selection, tool-call arguments,
schema preflight/repair, observation
handling, retry/stop policy,
aggregation/arbitration, and final-
answer submission.
- Memory owns compact reusable guidance,
prior-case reminders, memory encoding,
storage scope/format,
read/retrieval timing, routing, update/
write policy, and ingestion.
- Builder owns wiring, constants, metadata,
provider connection, and harness
import compatibility.
- Interfaces should be simple and explicit.
Use them only when one module must
pass state or guidance to another.
---
## Final Checklist
Use flat bullets only.
Include:
- unique candidate name
- candidate is independent and self-
contained
- required files present
- builder exports correct entry point
- module constants consistent
- harness is intended to be executable
without manual repair
- imports and provider exports are complete
- Stage 1 top failures addressed
- Stage 2 borrow / avoid constraints
followed
- no hard-coded answers
- no whole-harness copy
```

```

- no unnecessary architecture expansion
---
### Final Reminder
The generated harness should be ambitious
in completeness and creative in local
mechanisms, but conservative in
overall
architecture.
It should feel like a precise module-level
evolution of the current winner, not a
new unrelated system.

```

C.4.4 Retry Mechanism

Generated harnesses may contain implementation errors even when their high-level design is useful. Before rollout validation, we apply a lightweight smoke-test for importability, builder-interface compatibility, provider signatures, tool-call format validity, and execution on diagnostic examples. Failed candidates are returned to the meta-agent with the error trace and intended improvement brief for repair. We allow at most $N_{\text{retry}} = 3$ repair attempts; candidates that still fail are discarded before validation. This prevents invalid harnesses from consuming rollout budget, and the surviving executable harnesses are passed to harness-conditioned policy alignment.

C.4.5 Filter and Archive Update

HarnessForge uses a budgeted half-selection protocol to filter generated harness candidates before policy alignment. Each evolution round uses an evolution batch of about 1.2K tasks. Instead of evaluating every candidate on the full batch, we apply staged filtering on progressively consumed subsets. In our default setting, each filtering stage evaluates candidates on 200 tasks.

Let $\mathcal{C}_0^{(r)}$ denote the pooled candidate set generated at round r . At filtering stage t , each candidate harness $\mathcal{H} \in \mathcal{C}_{t-1}^{(r)}$ is paired with its corresponding policy to form an executable system $\mathcal{G}_{\mathcal{H}}$, and is evaluated on a fresh filtering subset $B_{r,t}$ with $|B_{r,t}| = 200$. We compute the multi-objective evaluation vector $\mathbf{J}(\mathcal{G}_{\mathcal{H}}; B_{r,t})$, rank candidates by Pareto dominance with task performance as the primary tie-breaker, and retain the top half:

$$\mathcal{C}_t^{(r)} = \text{Half}_{\text{Pareto}} \left(\mathcal{C}_{t-1}^{(r)}, \{ \mathbf{J}(\mathcal{G}_{\mathcal{H}}; B_{r,t}) \}_{\mathcal{H} \in \mathcal{C}_{t-1}^{(r)}} \right). \quad (30)$$

We use two half-selection stages by default. Thus, starting from eight generated candidates, the first 200-task filtering stage retains four harnesses, and the second 200-task filtering stage retains two survivor harnesses. The remaining tasks in the

round batch are then used to finish rollout evaluation for the survivor harnesses. These survivor rollouts are reused for harness-conditioned policy alignment, so the policy-alignment stage does not require a separate data-collection phase.

After each filtering stage, HarnessForge updates the archive $\mathcal{Z}^{(r)}$. For every evaluated harness, the archive stores its design summary, parent lineage, fault report, improvement brief, rollout summary, selection status, and evaluation vector $\mathbf{J}$. The updated archive is used in later rounds for Pareto-aware retrieval: archive cases are selected not only by fault relevance, but also by whether they lie on or near the Pareto frontier of performance and efficiency. This makes the filtering process both budget-efficient and reproducible, while also turning past evaluations into reusable evidence for subsequent harness tailoring.

D Policy Alignment Details

The policy side of HarnessForge is always conditioned on the selected harness and its lineage. This appendix separates the alignment protocol from the baseline-training configuration in App. E.

D.1 Policy Lineage and Adapter Operation

HarnessForge keeps a lineage for each survivor pair. When multiple survivor harnesses descend from the same parent pair, each child receives an independent copy of the parent policy state. The copied policy is then adapted separately using trajectories collected under the corresponding child harness, so sibling survivor pairs do not share trainable adapter parameters after branching.

Operationally, we first materialize the parent lineage policy as the initialization for the child pair. The parent adapter state is folded into the child initialization, and a new harness-specific LoRA adapter is attached and trained on successful trajectories collected under the child harness. We write this update as

$$\delta_k^{(r+1)} = \text{Mat} \left(\delta_{\text{parent}(k)}^{(r)} \right) \oplus \Delta \delta_k^{(r+1)}, \quad (31)$$

where $\text{Mat}(\cdot)$ denotes materializing the parent lineage policy for the child branch, $\oplus$ denotes attaching the new round-specific adapter to that materialized initialization, and $\Delta \delta_k^{(r+1)}$ is trained independently for harness $\mathcal{H}_k^{(r+1)}$. This convention makes pair lineage explicit while preserving the matched harness-policy interpretation used in the compatibility analysis.

D.2 Trajectory Curation and Success Filtering

For each survivor harness, HarnessForge reuses rollout traces already collected during staged harness filtering. A trajectory is retained for policy alignment only when it satisfies the task-success signal used by the corresponding benchmark evaluator. The retained trajectories are then decomposed into step-level decision pairs over the active harness interface, observation history, memory state, available actions, and next behavior. This reuse keeps the default SFT instantiation rollout-neutral with respect to the harness-selection budget.

D.3 Objective Instantiations

Training-objective agnostic alignment. The policy-alignment step in HarnessForge is not tied to a specific training objective. Given a harness $\mathcal{H}$ and a set of harness-conditioned trajectories collected under $\mathcal{H}$, the goal is to update the adapter so that the reasoner can execute reliably within the interface defined by that harness. We therefore view this step as a harness-conditioned policy-alignment operator rather than an SFT-specific operator. This operator can be instantiated with supervised fine-tuning, preference optimization, or reinforcement learning objectives. In this sense, HarnessForge defines the harness-conditioned adaptation problem, while the specific optimizer determines how the policy side is updated.

Default policy-evolution objective. In the main experiments, HarnessForge instantiates policy evolution with supervised fine-tuning. We use SFT for two practical reasons. First, it preserves the rollout budget: the successful trajectories used for policy learning are already collected during budgeted harness selection, so the policy step does not require a new environment-interaction phase,

$$\mathcal{T}_k^+ \subseteq \mathcal{T}_k^{(r+1)}. \quad (32)$$

Second, SFT gives a direct step-level supervision signal. Each retained successful rollout is decomposed into decision-level pairs,

$$\mathcal{D}_{\mathcal{H}_k} = \{(z_t, y_t) \mid \tau \in \mathcal{T}_k^+, t = 1, \dots, |\tau|\}. \quad (33)$$

Here z_t contains the task, active harness specification, observation history, memory state, and available actions; y_t is the demonstrated next behavior under that harness, such as a reasoning step, tool

call, memory write, or final answer. The corresponding objective is

$$\mathcal{L}_{\text{sft}} = - \sum_{(z_t, y_t) \in \mathcal{D}_{\mathcal{H}_k}} \log p_{\mathcal{R}_{\delta_k}^{(r)} \oplus \Delta \delta}(y_t \mid z_t). \quad (34)$$

Thus, SFT does not optimize a separate scalar reward; it teacher-forces the executor to imitate successful harness-conditioned decisions. This makes it a natural and rollout-efficient objective for learning harness-specific execution behaviors, including action formatting, tool-use discipline, memory utilization, verification habits, and termination control.

Policy-alignment objective agnosticism. The policy-alignment step in HarnessForge is defined at the framework level rather than being tied to one particular optimization objective. Its role is to adapt the executor to the selected harness, so the same operator can be instantiated with SFT, preference-style objectives, or RL-style objectives. In this view, different objectives mainly provide different ways of converting harness-conditioned execution evidence into an updated policy: SFT imitates curated successful trajectories, preference-style training contrasts stronger and weaker executions, and RL-style training directly optimizes reward signals under the current harness.

This design makes the choice of policy optimizer a practical tradeoff rather than a methodological constraint. As shown in Tab. 3, GRPO and RLOO can bring additional improvements in several settings, suggesting that HarnessForge still has headroom when stronger policy optimization and larger interaction budgets are available. However, these methods typically require extra rollout sampling for candidate generation, preference construction, or reward estimation. By contrast, SFT can reuse trajectories already collected during harness evolution, making it a more budget-efficient default. We therefore use SFT in the main experiments as a conservative tradeoff between performance and rollout cost, while treating preference- and RL-style objectives as higher-budget instantiations of the same alignment step.

E Baselines and Fairness Protocol

E.1 Shared Fairness Protocol

For fair comparison, training-style baselines such as SFT, GRPO, and RLOO are paired with the same

fixed base scaffold and trained on the same 3.8K trajectory pool used by HarnessForge for trajectory curation and policy alignment. These baselines therefore isolate the effect of policy optimization under a fixed execution scaffold, without performing harness evolution. Meta-agent-based harness-search baselines use the same GPT-5.5 meta-agent backend as HarnessForge. Following prior agent-search work, rollout budget counts executable environment/model rollouts, while meta-agent generation cost is treated as implementation overhead and is not included in rollout counts. The same convention is applied to HarnessForge and all meta-agent-based harness-search baselines.

E.2 Search-Style Baselines

Search-style harness baselines use the same frozen base reasoners as HarnessForge and optimize only the external execution structure. The executor is served through an OpenAI-compatible local vLLM endpoint with deterministic decoding (temperature = 0, top-p = 1), Qwen3 thinking disabled, and a 300-second request timeout. The held-out test split is never used during search. TMDB and API-Bank are used only for transfer evaluation rather than task-specific evolution. For TMDB and API-Bank, all methods start from the ToolHop-evolved harness or workflow and are evaluated under the target API interface without using TMDB or API-Bank held-out test examples. Tab. 7 summarizes the executable search budgets used in our local baseline ports.

Baseline-specific notes. **ADAS** is used as a free-form executable agent-code search baseline; we port generated agents to the ToolHop and SearchQA runtimes and discard invalid or schema-breaking generations before held-out evaluation. **AFlow** searches Python workflow programs with task-specific tool-call operators, so it tests whether workflow-level code evolution alone can match paired harness-policy adaptation. **AgentSquare** provides a structured modular-search counterpart, combining planning, reasoning, tool-use, and memory modules under the same task adapters. **MaAS** searches an operator controller for multi-agent execution rather than a single explicit workflow; deterministic path-repair shortcuts are disabled in our port. **MermaidFlow** follows the AFlow-style workflow search protocol but constrains candidates to a Mermaid graph plus executable code representation. Across all search-style baselines, the executor

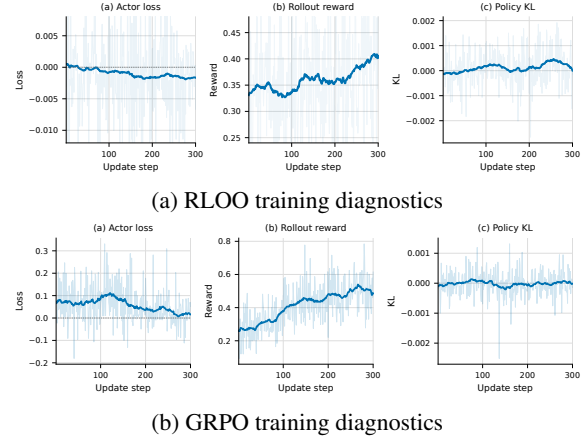


Figure 6: Online optimization diagnostics for the RLOO and GRPO training-style baselines over the first 300 update steps. We report actor loss, rollout reward, and policy KL from the training learner. Curves are smoothed for readability; these traces are used only as training diagnostics, while final comparisons are based on blind-test evaluation.

policy is not trained; only the external harness or workflow structure is selected.

Transfer setting for TMDB and API-Bank. TMDB and API-Bank are used only for transfer evaluation rather than task-specific harness search, because they do not provide dedicated evolution or training splits in our setting. For these two benchmarks, we start from the workflow or harness evolved on ToolHop and adapt it to the new task interface by replacing the tool APIs, environment wrappers, action schemas, and benchmark-specific execution constraints. No TMDB or API-Bank held-out test examples are used during this adaptation. This setting evaluates whether a harness discovered on ToolHop encodes reusable agent-system execution patterns that can be re-instantiated under new tools and environments.

E.3 Training-Style Baselines

Training-style baselines keep the harness fixed and update only the model-side executor. Unless otherwise stated, the base reasoner is frozen and only LoRA adapters are updated. Tab. 8 summarizes the shared adapter-training configuration, while Fig. 6 reports online optimization traces for the RL-style baselines. Final comparisons are based on held-out test evaluation in Tab. 1.

SFT. SFT trains LoRA adapters with multi-turn trajectories collected under the selected fixed-harness interface. We use rejection sampling to

Table 7: Executable configuration of search-style harness baselines. Validation statistics are used only for model or workflow selection; held-out test scores are reported in Tab. 1.

| Method | Evolution Data | Search Budget / Max Rounds | Validation Selection Criterion | Task Interface / Executor |
|-------------|---|---|---|---|
| ADAS | ToolHop: 800 online-dev examples; SearchQA: 1000 offline validation examples. | Target 30 valid agent generations; up to 3 debug attempts and 3 meta retries; invalid generations skipped. | Select by ToolHop hybrid validation score or SearchQA validation exact-match score. | ToolHop closed-mode tool calls; SearchQA E5 retrieval, top-3, max 4 searches, 6000-char observations; frozen executor. |
| AFlow | ToolHop: 800 online-dev examples; SearchQA: 1000 offline validation examples. | Top-4 workflow sampling per round; one validation pass per workflow; SearchQA search capped at 20 rounds. | Select the workflow with the best validation score on the corresponding task split. | Code-represented workflows with ToolHop/SearchQA tool-call operators and no executor training. |
| AgentSquare | ToolHop: 800 online-dev examples; SearchQA: 1000 offline validation examples. | Enumerates 1050 modular combinations from planning, reasoning, tool-use, and memory modules; target 30 valid generations. | Select the best valid modular agent by validation performance. | Same adapters as ADAS; invalid generated modules skipped; retrieval settings matched to ADAS. |
| MaAS | ToolHop: 800 online-dev examples; SearchQA: 1000 offline validation examples. | Adam controller search with lr 0.01, batch size 4, sample size 4; up to 9 recorded training rounds. | Select by controller validation objective on the development split. | Operator-controller search over 5–6 operators; no deterministic path repairs. |
| MermaidFlow | ToolHop: 800 online-dev examples; SearchQA: 1000 offline validation examples. | Elite workflow sampling with Mermaid-code validation; recombination after round 4 with probability 0.1; search capped by recorded evolution rounds. | Select the best executable Mermaid workflow by validation performance. | Mermaid graph plus executable code; SearchQA E5 retrieval, top-3, max 4 searches; deterministic executor, no policy tuning. |

construct the training set: only trajectories with valid final answers and positive task feedback are retained, while failed executions, malformed action traces, and trajectories without valid answers are discarded. Given a curated trajectory dataset $\mathcal{D}_{\text{sft}} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_{\text{sft}}|}$, where x_i denotes the task prompt together with the harness-conditioned interaction history and y_i denotes the target assistant action/response sequence, we optimize the token-level negative log-likelihood

$$\mathcal{L}_{\text{SFT}}(\delta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{sft}}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_{\theta_0 + \delta}(y_t \mid x, y_{<t}) \right], \quad (35)$$

where the backbone parameters θ_0 are frozen and only the LoRA adapter δ is updated. This instantiation is rollout-budget efficient because it reuses successful trajectories already produced during harness evolution rather than requiring an additional exploration stage.

Reward Design. All RL-style baselines use the same function-based verifier reward rather than a learned reward model. The rollout worker executes the complete scaffold-conditioned tool trajectory and assigns a scalar reward from the final trace. For ToolHop, the reward combines final-answer correctness and a path score that measures whether intermediate tool observations cover the annotated subtask answers, with equal weights 0.5/0.5. Missing or malformed final answers incur a -0.1 penalty, and non-final tool calls beyond four are penalized by 0.01 per call; the resulting

score is clipped to $[-0.1, 1]$. For SearchQA, exact-match final answers receive full credit, substring matches receive 0.5 partial credit, and responses without an extractable answer receive zero. For EnvScaler-style stateful tasks, the terminal reward is the fraction of task postcondition checks satisfied after the agent calls the completion action. GRPO and RLOO share this reward function and the same terminal environment feedback, so their comparison isolates the advantage estimator rather than reward design.

GRPO. GRPO normalizes group-relative advantages within the $N = 4$ completions for each prompt. KL is not added to the reward, instead, both methods use an actor-side low-variance KL loss with coefficient 10^{-3} .

RLOO. RLOO uses the same on-policy scaffold agent loop as GRPO. For each prompt, the rollout worker samples $N = 4$ trajectories. The advantage of a trajectory is computed against the mean reward of the other trajectories in the same prompt group, giving a leave-one-out baseline without training a separate value critic. We keep the clipped actor objective, token-mean loss aggregation, entropy coefficient, KL loss, and rollout budget identical to GRPO so that the comparison isolates the advantage estimator.

F Reproducibility Artifacts

Reproducibility of Meta-Agent Harness Evolution. HarnessForge uses a strong meta-agent for harness evolution, but implements it as a staged operator with fixed input-output contracts rather

Table 8: Training configuration for training-style baselines.

| Parameter | Value |
|----------------------|--|
| Trainable parameters | LoRA adapters only |
| LoRA target modules | Attention and MLP projections |
| LoRA rank | 8 |
| LoRA alpha | 16 |
| LoRA dropout | 0.05 |
| Epoch | 1.0 |
| Optimizer | AdamW with bf16 training |
| Learning rate | 2×10^{-6} |
| LR sweep | $\{5 \times 10^{-6}, 1 \times 10^{-5}, 2 \times 10^{-5}\}$ |
| Scheduler | Cosine decay, 3% warmup |
| KL coefficient | 0.01 |
| Clip range | 0.2 |
| Reward | Success, validity, tool call counts |
| GRPO $rollout_n$ | 4 rollouts per prompt |
| RLOO $rollout_n$ | 4 rollouts per prompt |

than an unconstrained generation call. Each stage takes structured inputs, including the parent harness, rollout evidence, evaluation summaries, and archive records, and produces schema-controlled artifacts such as fault reports, evolution manifests, harness-candidate manifests, and smoke-test logs. These artifacts record the diagnosed failure modes, responsible modules, supporting evidence, edited components, repair priorities, and expected behavioral changes, making the evolution process auditable rather than opaque.

We support reproducibility at two levels. For result-level replay, we release the final selected harness-policy pairs, including evolved harnesses, adapter checkpoints and configurations, benchmark wrappers, evaluation scripts, and split identifiers. This allows the reported systems to be evaluated without re-running meta-agent evolution. For process-level auditability, we release the prompts, operator schemas, generated reports, candidate manifests, smoke-test outcomes, and survivor-selection logs used during evolution. These records expose how each harness was generated, what failure it aimed to repair, and why it was retained.

Because meta-agent outputs may vary across model versions and providers, we do not assume exact regeneration of every intermediate candidate. Instead, our artifact release separates replaying the final executable systems from re-instantiating the evolution protocol. The released schemas and stage-wise contracts allow future work to replace the proprietary meta-agent with alternative closed- or open-source models under the same evolution interface and selection procedure.

Table 9: Meta-agent process statistics and cost averaged across evolution rounds. Stage-wise costs are rounded, while the total is computed from the original billing records.

| Metric | Average |
|-----------------------------------|-----------|
| <i>Generation and validation</i> | |
| Candidate-generation success rate | 96.43% |
| Candidates repaired by smoke test | 37.50% |
| Meta-agent calls per round | 3.43 |
| <i>Resource usage per round</i> | |
| Input tokens | 10.35M |
| Output tokens | 0.298M |
| Wall-clock time | 33.43 min |
| <i>Stage-wise monetary cost</i> | |
| Fault attribution | \$50.80 |
| Improvement planning | \$0.28 |
| Proposal generation | \$6.21 |
| Refinement and repair | \$2.80 |
| Total meta-agent cost per round | \$60.14 |
| Cost per refined candidate | ~\$7.5 |

F.1 Meta-Agent Process Statistics and Cost

HarnessForge invokes the meta-agent only during harness tailoring; policy alignment and final evaluation do not require additional meta-agent queries. Tab. 9 reports process-level statistics averaged across the three evolution rounds. The statistics are collected from the generation manifests, smoke-test logs, API usage records, and wall-clock execution logs.

The high candidate-generation success rate shows that staged input-output contracts make open-ended harness generation reliable, while the 37.50% smoke-test repair rate highlights the importance of execution validation. Fault attribution dominates meta-agent cost because it processes full rollout and implementation context, whereas later stages operate on compact reports. With eight candidates per round, the overhead is amortized to approximately \$7.5 per candidate. Final results can be reproduced without rerunning the meta-agent using the released harnesses, adapters, rollout traces, and evolution logs.

G Additional Results and Analysis

G.1 Sensitivity to Meta-Agent Capability

To examine whether the gains of HarnessForge depend on a particular strong meta-agent, we repeat one evolution round on ToolHop with Qwen3-4B while replacing GPT-5.5 with GPT-5.4-mini. We keep the harness factory, candidate budget, eval-

Table 10: One-round sensitivity to meta-agent capability on ToolHop with Qwen3-4B. Harness-only disables policy alignment, while HarnessForge uses the same evolved harness followed by harness-conditioned policy alignment. Meta-agent tokens and costs are shared by the two variants under each backend because policy alignment does not invoke the meta-agent.

| Meta-Agent | Variant | ToolHop Ans. | ToolHop Path | Tokens | Cost |
|--------------|---------------------|--------------|--------------|--------|---------|
| – | Vanilla | 41.03 | 56.07 | – | – |
| GPT-5.5 | Harness-only | 45.13 | 63.90 | 10.6M | \$60.14 |
| | HarnessForge | 46.15 | 64.99 | | |
| GPT-5.4-mini | Harness-only | 42.05 | 61.22 | 18.1M | \$15.30 |
| | HarnessForge | 43.59 | 62.50 | | |

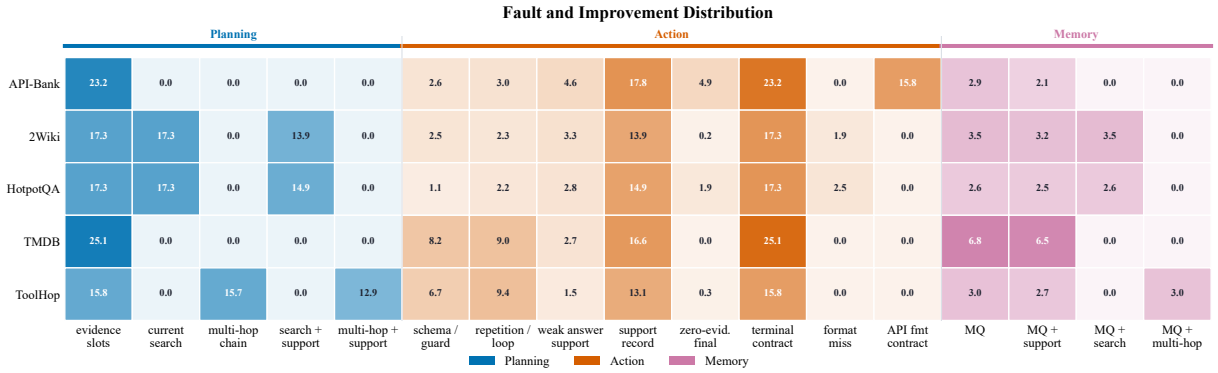


Figure 7: Fault and improvement-signal distribution across benchmarks. Each row is normalized within a benchmark, and columns group repair signals into planning-, action-, and memory-related categories.

uation protocol, and policy-alignment procedure unchanged. For each meta-agent, we compare the resulting harness-only system against the full HarnessForge pair after harness-conditioned policy alignment.

As shown in Tab. 10, GPT-5.5 produces stronger absolute results than GPT-5.4-mini. Nevertheless, policy alignment consistently improves over the corresponding harness-only control under both backends: 1.02/1.09 answer/path points with GPT-5.5 and 1.54/1.28 points with GPT-5.4-mini. Thus, meta-agent capability affects harness quality, but the additional benefit of harness-policy co-evolution is not specific to GPT-5.5. GPT-5.4-mini further provides a lower-cost, albeit weaker, alternative.

G.2 Module-Level Repair Statistics

Fig. 7 shows that harness failures are module-specific and benchmark-dependent. Action-side signals dominate on API-heavy benchmarks such as API-Bank and TMDb, where common failures include invalid API formats, schema mismatches, missing guards, repeated actions, and incorrect endpoint or tool selection. These tasks therefore benefit most from action-layer repairs such as stricter format contracts, schema preflight, and loop pre-

vention.

Planning-side signals are more visible on retrieval-heavy and multi-hop benchmarks. The SearchQA variants mainly require current-query repair and stronger coupling between search actions and supporting evidence, while ToolHop more often requires preserving multi-hop tool chains and grounding final answers in the accumulated evidence. These patterns indicate that planning repairs are most useful when success depends on maintaining the right intermediate intent across several reasoning or retrieval steps.

Memory-related signals are smaller but still meaningful. They typically appear as co-repairs with planning or action modules, suggesting that memory quarantine acts as a stabilizing layer rather than an isolated repair target. It helps prevent stale or irrelevant traces from disrupting otherwise valid plans and tool actions. Overall, the heatmap suggests a clear division of repair pressure: API tasks stress action reliability, search tasks stress query planning and evidence coupling, multi-hop tool-use tasks stress path preservation, and memory repairs provide cross-cutting execution stability.

G.3 Case Study

To complement the aggregate repair statistics in Fig. 7, we present five representative parent-child trajectory comparisons in Figs. 8–12. Each case shows how HarnessForge turns a recurring trajectory-level failure into a reusable repair over planning, action, or memory.

Case 1: Multi-hop comparison. Fig. 8 shows that the child harness replaces repeated unsupported finalization with entity-level evidence slots and support-record verification.

Case 2: Structured API execution. Fig. 9 illustrates API-route repair: the child harness follows the required authentication–deletion sequence and enforces the expected API-request output format.

Case 3: Retrieval-heavy multi-hop QA. Fig. 10 shows current-query repair, where the child harness avoids stale repeated searches and grounds the final answer in question-specific evidence.

Case 4: Multi-hop tool-chain execution. Fig. 11 demonstrates schema-guarded tool use: the child harness preserves the source–intermediate–transform chain while normalizing invalid date-tool arguments.

Case 5: REST-style endpoint routing. Fig. 12 shows endpoint-routing repair, where the child harness avoids unnecessary season-level calls and retrieves the answer from the correct TV-detail endpoint.

Overall, these cases show that HarnessForge improves execution behavior by modifying reusable interfaces, such as evidence slots, schema guards, support records, API contracts, and memory quarantine, rather than applying task-specific answer patches.

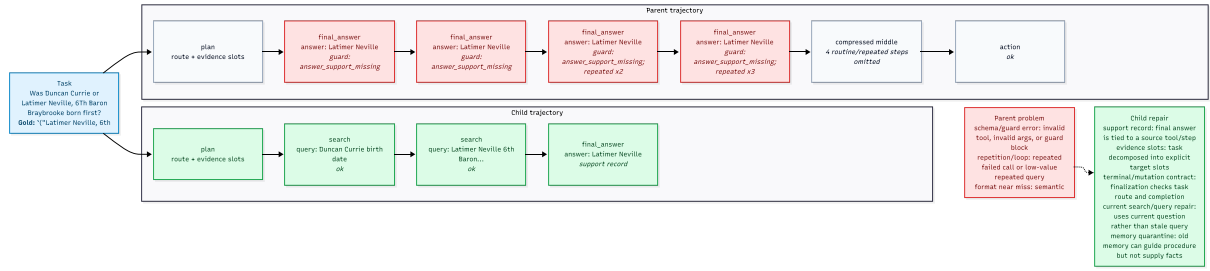


Figure 8: Case Study 1: evidence-supported finalization for multi-hop comparison. The child harness replaces repeated unsupported finalization with entity-level evidence slots and support-record verification.

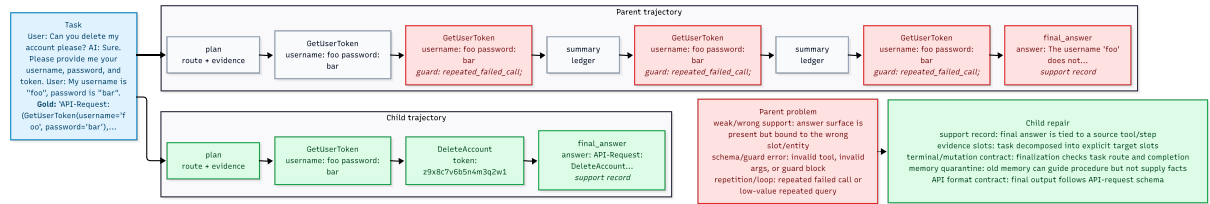


Figure 9: Case Study 2: API-contract repair for account deletion. The child harness follows the required API route and emits the final answer in the expected API-request format.

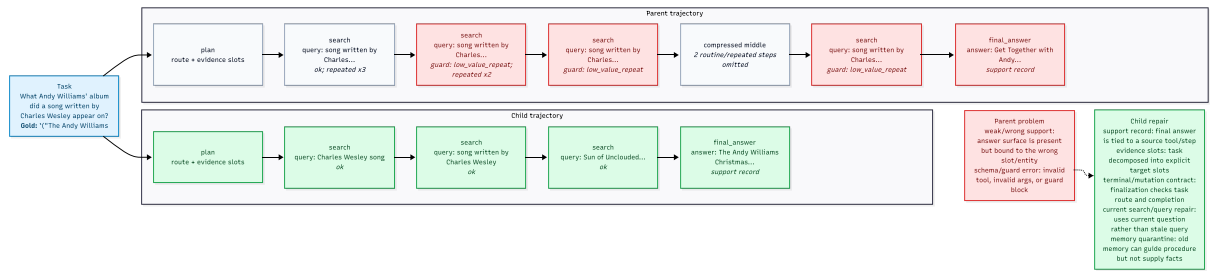


Figure 10: Case Study 3: current-query repair for retrieval-heavy multi-hop QA. The child harness replaces stale repeated searches with question-grounded queries and support-record finalization.

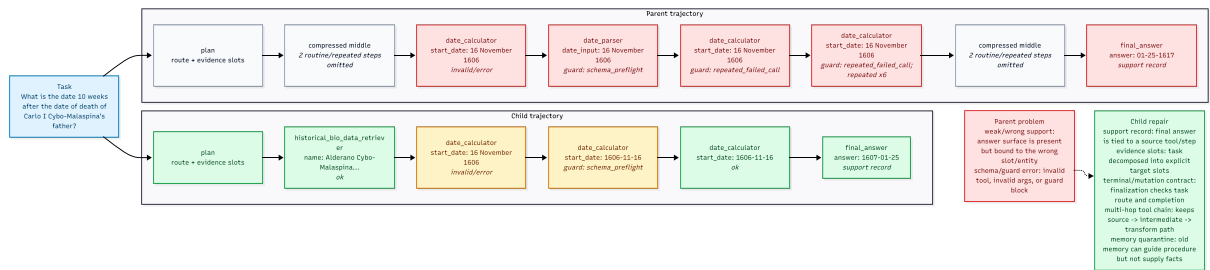


Figure 11: Case Study 4: schema-guarded multi-hop tool-chain repair. The child harness preserves the source–intermediate–transform path while repairing invalid date-tool arguments.

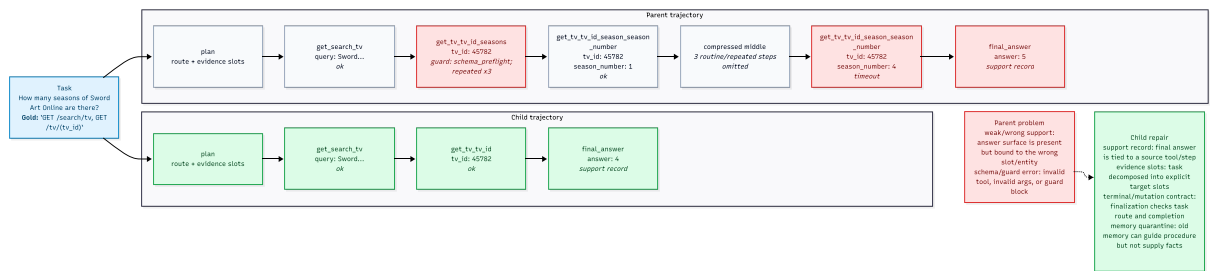


Figure 12: Case Study 5: endpoint-routing repair in TMDB-style API use. The child harness routes to the correct TV-detail endpoint and avoids unnecessary repeated season-level calls.