ANSWER2PATH: OUTCOME-GUIDED ACQUISITION OF REUSABLE WEB-AGENT EXPERIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Successful web-agent trajectories are costly to acquire, and terminal success alone does not show whether an agent gathered evidence or learned a reusable procedure. Yet many tasks already have a known answer or desired final state, typically used only for evaluation. Can a small pool of such outcomes guide trajectory acquisition and seed further supervision? We propose ANSWER2PATH (A2P), a closed-loop framework that uses outcomes during exploration and expands supervision from verified webpage evidence. For queries, the agent seeks a path supporting the known answer; for modifications, the desired state guides action and post-state verification. We compare when the same outcome is revealed and whether it explicitly guides evidence seeking. A condition-blind companion records evidence, and independently validated page facts become provenance-preserving task–outcome pairs. Fixed workflow induction then tests whether acquired trajectories help on held-out tasks without outcome access. On 61 WebArena Shopping queries over three seeds, A2P raises pooled official success from 40.4% to 84.7%; at fixed K, induced workflows raise held-out success from 88.3% to 97.2%. Across six modification families, eligible source success rises from 22.9% to 54.2%, while source tokens per success fall by 75.0%. Late guidance has the highest aggregate yield in a two-family placement study, and a two-round single-family pilot validates backend-free, evidence-grounded task expansion. Together, these results show how limited outcome supervision can improve procedural experience acquisition and seed further verified tasks.

1 INTRODUCTION

Web agents now operate realistic websites, desktop environments, and mobile applications using language, accessibility trees, screenshots, and executable actions (Zhou et al., 2024; Koh et al., 2024; Deng et al., 2023; Drouin et al., 2024; Xie et al., 2024; Rawles et al., 2025). Their progress has been accelerated by stronger multimodal policies (Hong et al., 2024; Cheng et al., 2024; Wu et al., 2025; Lu et al., 2024; Wang et al., 2025a; Liu et al., 2025) and interactive training environments (de Chezelles et al., 2025; Lu et al., 2025b; Pan et al., 2024; Peeters et al., 2025). Yet the experience used to improve these agents remains expensive. Successful trajectories are sparse, annotation-heavy, and brittle under website changes; terminal success alone says little about whether the trajectory contains evidence or reusable procedural structure.

This paper studies a simple but underused resource: many web tasks already come with a known answer, desired final state, or executable evaluator. Benchmarks need these outcomes to score agents, operators often know what state a workflow should reach, and deployed systems may receive outcome feedback after execution. Standard evaluation leaves this information at the end of the episode. We ask whether moving it earlier can improve the acquisition of procedural experience without requiring action demonstrations.

Our target is procedural experience rather than immediate answer accuracy. A useful trajectory records the evidence or state transition behind its success, and its value should persist when the answer is removed. We therefore evaluate outcome supervision at three levels: (i) eligible yield and evidence-supported yield per rollout or token; (ii) transfer from a fixed number K of successful trajectories to held-out tasks without outcomes; and (iii) generation of new, independently verifiable task–outcome pairs from webpage evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

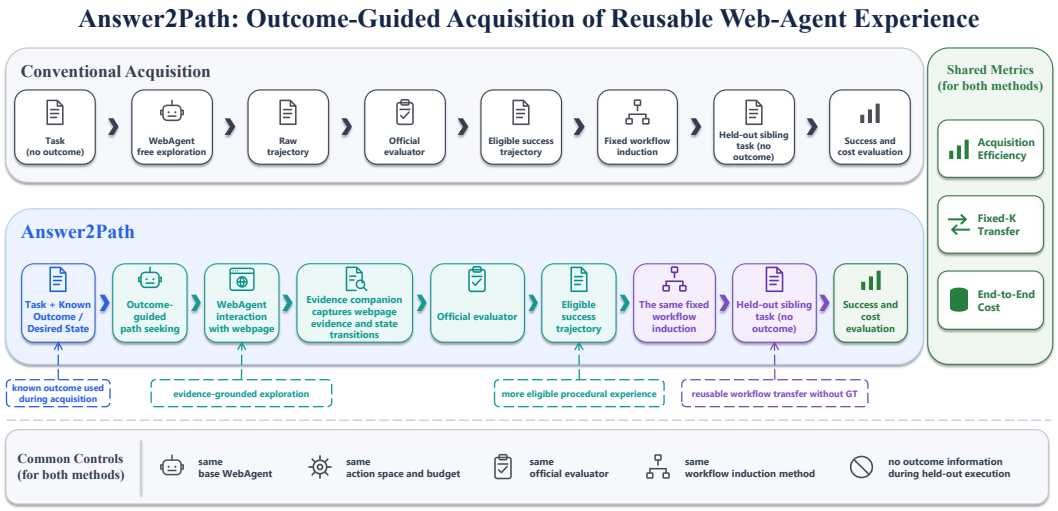


Figure 1: ANSWER2PATH overview. Both pipelines use the same base agent, budget, evaluator, workflow induction method, and no-outcome held-out evaluation. A2P uses a known outcome to guide acquisition; official success determines eligibility, while webpage or state evidence is audited separately.

ANSWER2PATH connects these levels in one acquisition loop. For query tasks, a textual answer becomes a target for evidence seeking. For modification tasks, an evaluator-defined desired state anchors entity localization, action execution, and post-state verification. A condition-blind companion records evidence and scores steps; complete, infrastructure-valid, officially successful rollouts form the eligible pool. A fixed induction method abstracts eligible trajectories into workflows, which are evaluated on sibling tasks without answers. Verified webpage observations also yield provenance-preserving tasks for the next acquisition round.

Our contributions are:

- We formulate *outcome-conditioned experience acquisition*, separating acquisition efficiency from the conditional transfer value of successful trajectories.
- We introduce an auditable closed-loop pipeline that combines supervision placement, evidence auditing, workflow induction, GT-free transfer, and GUI-only task expansion.
- We provide controlled evidence across query and modification tasks. On 61 query tasks over three seeds, A2P improves success by 44.3 percentage points while using fewer total tokens. A four-family fixed-K study improves held-out transfer by 8.9 points. On six modification families, eligible source success rises from 22.9% to 54.2% while tokens per success fall by 75.0%.
- We study where outcome information should enter the interaction. Across two clean functional families, late guided placement attains 78.1% source yield versus 40.6% without an answer and improves $K{=}1$ transfer from 65.0% to 80.0%.

2 RELATED WORK

2.1 WEB AGENTS AND INTERACTIVE BENCHMARKS

WebArena, VisualWebArena, WebShop, and WorkArena provide interactive browser tasks and evaluation (Zhou et al., 2024; Koh et al., 2024; Yao et al., 2022; Drouin et al., 2024). Mind2Web and WebLINX provide recorded web-interaction data, while BrowserGym/AgentLab standardizes agent interfaces and evaluation (Deng et al., 2023; Lù et al., 2024; de Chezelles et al., 2025). Subsequent suites broadened long-horizon, multimodal, compositional, safety, and live-web coverage (Boisvert et al., 2024; Jang et al., 2025; Yang et al., 2025; Yoran et al., 2024; Song et al., 2025; Zheng et al.,

2025b; Gou et al., 2025a). Representative agents combine language-model planning, multimodal grounding, memory, and browser-specific actions (Gur et al., 2024; He et al., 2024; 2025; Lai et al., 2024; Abuelsaad et al., 2024; Zheng et al., 2024b;a; Yu et al., 2025). ANSWER2PATH operates as an acquisition protocol that can wrap an existing agent and evaluator.

2.2 TRAJECTORY ACQUISITION AND SELF-IMPROVEMENT

Recent systems acquire interaction experience through curricula, exploration, and task synthesis. WebRL uses a self-evolving curriculum (Qi et al., 2025); OS-Genesis explores GUI states before retrospectively synthesizing tasks (Sun et al., 2025a). Explorer and HATS expand trajectory pools through exploration and hardness-aware synthesis (Pahuja et al., 2025; Shao et al., 2026). Other systems scale synthetic tasks and environments (Ramrakhya et al., 2026; Cao et al., 2026; Zhou, 2026; Lv et al., 2026), develop synthetic supervision or reusable skills (Patel et al., 2024; Zhou et al., 2025; Wang et al., 2026d; Zhang et al., 2025b; Sun et al., 2025b; Xiao et al., 2025; Xue et al., 2026; He et al., 2026b), mine tutorial trajectories (Zhang et al., 2026a), or build cumulative data flywheels (Ye et al., 2026; Wang et al., 2026b; Zhang et al., 2026b). We study how an already available outcome enters interaction and affects evidence-supported yield and fixed-K transfer.

2.3 WORKFLOW AND PROCEDURAL MEMORY

Agent Workflow Memory (AWM) abstracts common subroutines from past trajectories, while later work induces programmatic skills, discovers reusable procedures, or maintains evolving experience memories (Wang et al., 2025c;b; Zheng et al., 2025a; He et al., 2026a; Deng et al., 2026; Xiao et al., 2026). Related systems use critic-guided exploration, hierarchical memory, executable skills, and continual experience learning (Li et al., 2025; Zhang et al., 2026c; Liu et al., 2026; Hu et al., 2026). We fix the AWM induction method and source-selection order; the modification transfer study separately accounts for deterministic compilation of A2P workflows.

2.4 TRAJECTORY VERIFICATION AND REWARD MODELING

AgentRewardBench and CUARewardBench benchmark trajectory and reward judgments (Lù et al., 2025; Lin et al., 2025). STEVE verifies individual actions, WebGraphEval models cross-trajectory structure, and reward or verifier systems assess progress (Lu et al., 2025a; Qian et al., 2025; Xiong et al., 2025; Zhang et al., 2026d). Closely related work develops interactive critics, milestone rewards, action verification, and proactive evidence seeking (Cui et al., 2026; Zhang et al., 2025a; Dai et al., 2026b;a; Lee et al., 2025; Zhang et al., 2026e; Cai et al., 2025; Yang et al., 2026; Rosset et al., 2026). ANSWER2PATH uses a condition-blind companion for evidence capture and analysis; the official evaluator alone determines task success.

Appendix Table 10 contrasts the closest methodological families by what they acquire, verify, or reuse.

3 PROBLEM FORMULATION

Let a web task be x, an acquisition-time outcome be y, an interactive environment be $\mathcal{E}$, and an agent policy be π. The outcome may be a textual answer for a query or an evaluator-defined target state for a modification. A trajectory is

$$\tau = (o_0, a_0, o_1, a_1, \ldots, o_T), \tag{1}$$

where o_t is the browser observation and a_t is an executable action. The official evaluator $V(x, \tau)$ determines task success. We write $A(x, \tau) = 1$ when a rollout is officially successful, complete, and free of infrastructure errors. Evidence support is recorded separately from webpage observations or a verified state transition; evidence-qualified trajectories satisfy both criteria.

An acquisition strategy m maps a task and, when available, its outcome to a trajectory distribution. Baseline has no access to y; A2P variants differ in when and how y enters. Given a fixed rollout or token budget B, acquisition efficiency is

$$\mathrm{Yield}(m; B) = \frac{\sum_{\tau \in \mathcal{D}_m(B)} A(x, \tau)}{|\mathcal{D}_m(B)|}, \qquad \mathrm{Cost/Eligible} = \frac{\mathrm{Tokens}(\mathcal{D}_m)}{\sum A(x, \tau)}. \tag{2}$$

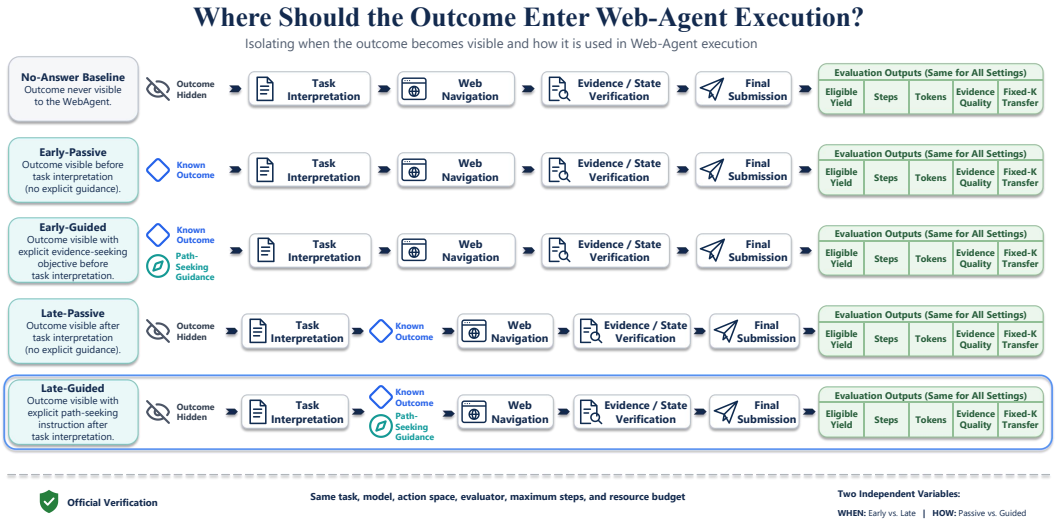


Figure 2: Controlled outcome-placement interventions. The four A2P variants isolate *when* the same outcome is revealed and *how* it is used, under matched tasks, models, budgets, and evaluators.

To measure reuse independently of source success count, a fixed induction operator G maps exactly K eligible trajectories to a workflow $w = G(\tau_1, \ldots, \tau_K)$. Held-out sibling tasks x' receive only (x', w), never y or source entities. A strict conditional transfer comparison holds G, executor, and budget identical; the query fixed-K study follows this design. The modification study keeps AWM fixed but adds A2P-side deterministic workflow compilation, so its held-out comparison measures the resulting pipeline rather than trajectory-only value.

Finally, a webpage evidence extractor proposes new pairs $(\tilde{x}, \tilde{y}, p)$, where provenance p identifies the page, observation, and generating trajectory. An independent GUI validation step must accept the pair before it can enter the next acquisition round. This creates a bounded flywheel rather than an unconstrained self-labeling loop.

4 ANSWER2PATH

4.1 OUTCOME PLACEMENT

We compare a no-answer Baseline with four A2P variants:

Early-Passive. The outcome is visible as context from the beginning, without an explicit evidence objective.

Early-Guided. The outcome is visible from the beginning and the agent is instructed to seek supporting evidence or a verified state transition.

Late-Passive. The outcome is introduced after initial task interpretation, reducing early anchoring while retaining outcome context.

Late-Guided. The outcome is introduced after initial interpretation together with a targeted path-seeking objective.

The 61-task query study and clean query fixed-K study reveal the full answer at rollout start. Their frozen prompts ask the agent to verify it through a webpage path. The four variants above form a separate placement study; its results are reported separately. All answer-visible variants receive the same outcome and use the same action space, observation, maximum steps, evaluator, model, and token scope. Thus the experiment asks not merely whether answers help, but how their placement changes the distribution of acquired experience.

4.2 Query and modification tasks

For a query, the agent must navigate until webpage observations support y; repeating the known answer is not sufficient. For a modification, y is a target state. The agent may first perform a bounded *shadow query* to locate the target entity, read dynamic fields, and bind the correct control before modifying state. Easy tasks can skip this phase. Success is always determined by the official evaluator, and state-modification runs reset the environment between conditions.

4.3 Evidence companion

The companion stores evidence snapshots, action effects, provenance, and a condition-blind step score. It rejects incomplete finishes and checks task, entity, and field binding. For queries it validates that the required webpage fields were actually observed; for modifications it requires the official post-state transition. Companion calls are reported separately from agent execution so that trajectory quality does not become hidden inference cost.

4.4 Workflow induction and leakage control

Eligible trajectories are passed to a fixed AWM implementation (Wang et al., 2025c). Baseline and A2P use the same induction commit and sampling order. For each K, the first K eligible trajectories in a frozen order are selected without replacement. A condition with fewer than K successes is marked not estimable. Workflow preparation retains action-bearing steps while removing source-specific entities, values, and ground-truth literals. The standard compiler replaces brittle element identifiers with semantic bindings where possible. In the modification fixed-K experiment, A2P workflows additionally undergo deterministic standard-compact compilation; Baseline workflows remain uncompressed. This comparison therefore evaluates the acquired-trajectory-to-workflow pipeline, including its condition-specific compilation.

4.5 Evidence-grounded data flywheel

Verified observations can expose additional products, orders, attributes, or page facts. The companion converts them into candidate task–outcome pairs, freezes the candidate before evaluation, and validates it through a separate no-GT GUI rollout. Accepted pairs carry source-page and observation provenance and enter only the next round. The agent never receives database generation logic. Backends may still be used by the benchmark's official evaluator, but not to create query tasks.

5 Experimental Setup

Environment and tasks. We use the Shopping site in WebArena (Zhou et al., 2024) through BrowserGym/AgentLab-compatible browser interaction (de Chezelles et al., 2025). Query experiments include 61 benchmark tasks and clean functional families for order details, monthly spend, purchased options, and review ratings. Modification experiments include cart, Wishlist, contact/draft, purchase, and review state changes. A clean family holds navigation procedure, required fields, and computation semantics fixed while varying entities and values.

Agents and models. The main experiments use an accessibility/DOM-based OpenAI-compatible web agent with temperature zero and at most 20 steps. Confirmatory query experiments use GPT-5.5; additional experiments use DeepSeek-V3.2 and an independent AgentLab GenericAgent scaffold. The modification fixed-K study reuses historical GPT-5.5 source trajectories for both conditions and evaluates no-GT held-out execution with DeepSeek-V3.2. This complements work on end-to-end multimodal and grounding-oriented agents (He et al., 2024; Luo et al., 2025; Wei et al., 2025; Gou et al., 2025b) while isolating the acquisition intervention from model training.

Eligibility and evaluation. The primary eligibility rule is official success, trajectory completeness, and no infrastructure error. A condition-blind runtime flag, the literal evidence check, separately records webpage support for query conditions; a retrospective semantic review examines A2P traces that fail this check. Neither audit changes primary eligibility. Held-out tasks receive no GT. We report success, eligible yield, steps, known tokens, wall time where available, and cost per eligible trajectory.

Table 1: Query acquisition on 61 Shopping tasks per seed. Δ is the success-rate change; token Δ is relative total rollout cost. The pooled gain has exact McNemar $p = 1.89 \times 10^{-22}$ and paired-bootstrap 95% CI $[+36.6, +51.9]$ pp.

Seed	Baseline	A2P	Δ pp	A2P-only / B-only	Token Δ
0	21/61 (34.4%)	**47/61 (77.0%)**	**+42.6**	26 / 0	-3.3%
1	24/61 (39.3%)	**52/61 (85.2%)**	**+45.9**	30 / 2	-18.0%
2	29/61 (47.5%)	**56/61 (91.8%)**	**+44.3**	27 / 0	-16.5%
Pooled	74/183 (40.4%)	**155/183 (84.7%)**	**+44.3**	**83 / 2**	-13.1%

Query fixed-K end-to-end tokens include source acquisition, AWM induction, and held-out execution. Modification fixed-K token totals cover held-out execution; the placement study reports main-agent tokens and excludes companion calls. These choices follow GUI-agent efficiency surveys (Wang et al., 2024; Nguyen et al., 2025; Shi et al., 2025; Bai et al., 2026) and context/compression work (Song et al., 2026; Kerboua et al., 2025; Wang et al., 2026a;c).

Statistical protocol. The 61-task confirmatory study uses three seeds and paired McNemar tests. The query fixed-K experiment uses four clean families, $K \in \{1, 2, 4\}$, and three seeds. Modification and placement experiments use frozen task splits, matched protocols, and exact denominators.

6 Results

6.1 Outcome conditioning improves query acquisition

Table 1 reports the confirmatory 61-task study. A2P succeeds on 155/183 rollouts versus 74/183 for Baseline, a 44.3-point gain. The paired result is highly significant ($p = 1.89 \times 10^{-22}$), with 83 A2P-only successes and two Baseline-only successes. Despite more than doubling successful trajectories, A2P uses 13.1% fewer total tokens, reducing tokens per success from approximately 125,925 to 52,217.

Task-family experiments corroborate the result (Appendix Table 11). Across 40 tasks and three seeds, same-family path transfer raises official success from 95/120 to 107/120; adding outcome guidance reaches 112/120. A one-seed cross-family control attains 31/40 strict successes, versus 106/120 for same-family transfer across three seeds, consistent with family-specific procedural structure.

6.2 Acquired query experience transfers without the answer

The fixed-K experiment holds successful source count constant and uses the same induction pipeline. Across 180 paired held-out observations, Baseline-AWM succeeds on 159 (88.3%) and A2P-AWM on 175 (97.2%), with exact McNemar $p = .000855$. Gains are consistent across $K = 1, 2, 4$ (8.3, 8.3, and 10.0 points). End-to-end tokens per held-out success fall from 92,157 to 83,174, even though held-out execution itself is near ceiling in several families.

The frozen manifest separates source and held-out IDs and entity fingerprints while preserving each family's three-part functional signature. All 72 induced-workflow records pass the stored leakage audit with zero source-entity literals. Of 360 held-out prompts, 359 are observable and contain no GT marker; one is unobservable because of a retained infrastructure error. Configuration symmetry, AWM version, and adapter checks also pass.

6.3 Modification tasks benefit from grounded outcome use

Modification tasks require action, not answer repetition. Across six families and three seeds, grounded A2P obtains 78/144 successful source trajectories versus 33/144 for Baseline. Tokens per success fall from 263,206 to 65,857. A separate four-family, 20-task run with GPT-5.6-sol yields 16/20 successes versus 6/20 while reducing total tokens by 28.7%. These experiments support the use of desired states as acquisition anchors.

Table 2: Clean query fixed-K transfer: four families, three seeds, identical induction, and no outcome at held-out time. The pooled paired gain is significant ($p = .000855$; 95% CI $[+1.1, +18.3]$ pp). ↓ denotes lower cost.

Metric	Baseline-AWM	A2P-AWM	Relative change
$K = 1$ held-out success	54/60 (90.0%)	**59/60 (98.3%)**	**+8.3 pp**
$K = 2$ held-out success	54/60 (90.0%)	**59/60 (98.3%)**	**+8.3 pp**
$K = 4$ held-out success	51/60 (85.0%)	**57/60 (95.0%)**	**+10.0 pp**
Pooled held-out success	159/180 (88.3%)	**175/180 (97.2%)**	**+8.9 pp**
Eligible source trajectories	82/96 (85.4%)	**93/96 (96.9%)**	**+11.5 pp**
Source tokens	6,150,109	**5,386,862**	**↓12.4%**
Source tokens / eligible	75,001	**57,923**	**↓22.8%**
E2E tokens / held-out success	92,157	**83,174**	**↓9.7%**

Table 3: Modification-task acquisition and fixed-K transfer. Source rows report the frozen six-family study; per-family rows detail the four families contributing eligible trajectories. Transfer uses Wishlist and cart, three seeds, and no outcome at held-out time. The transfer A2P workflows receive deterministic standard-compact compilation after the same AWM induction; Baseline workflows are uncompressed. Transfer token counts cover held-out execution.

Stage / scope	Metric	Baseline	A2P	Change
Source, all six	Success	33/144 (22.9%)	**78/144 (54.2%)**	**+31.3 pp**
Wishlist	Success	21/24	**21/24**	0.0 pp
Cart	Success	11/24	**15/24**	**+16.7 pp**
Draft amount	Success	1/24	**24/24**	**+95.8 pp**
Draft SKU	Success	0/24	**18/24**	**+75.0 pp**
Source, all six	Tokens / success	263,206	**65,857**	**↓75.0%**
Fixed-K, pooled	Held-out success	57/90 (63.3%)	**72/90 (80.0%)**	**+16.7 pp**
Cart, $K = 1$	Held-out success	9/15	**12/15**	**+20.0 pp**
Cart, $K = 2$	Held-out success	8/15	**12/15**	**+26.7 pp**
Cart, $K = 4$	Held-out success	9/15	**12/15**	**+20.0 pp**
Fixed-K, pooled	Total tokens	2,686,767	**1,711,673**	**↓36.3%**
Fixed-K, pooled	Tokens / success	47,136	**23,773**	**↓49.6%**
Fixed-K, pooled	Avg. steps / time	10.01 / 159.0s	**5.22 / 86.7s**	**↓47.9 / 45.5%**

State-transition verification agrees with official success at 33/144 for Baseline and 78/144 for A2P; the stricter query-style page flag yields 32/144 and 36/144. On held-out modification tasks, state support matches official success at 57/90 and 72/90 (Appendix Table 8).

The modification fixed-K comparison covers Wishlist and cart over three seeds. Baseline-AWM succeeds on 57/90 held-out runs, while A2P-AWM with standard-compact compilation succeeds on 72/90 using 36.3% fewer held-out tokens. Cart improves from 9/15, 8/15, and 9/15 to 12/15 at $K = 1, 2, 4$, respectively. Wishlist improves at $K = 1, 2$ but falls from 10/15 to 9/15 at $K = 4$. Thus the pooled pipeline gain is strongest in Cart and does not hold for every family–K cell.

The frozen manifest, artifact hashes, source/held-out isolation, and workflow leakage audit support this no-GT comparison; Appendix C.1 gives the checks and remaining element-ID dependencies.

6.4 Where should the answer be placed?

Table 4 combines the frozen order-detail and review-rating protocols, each with eight source tasks, five held-out tasks, and two seeds. Late-Guided is the strongest aggregate acquisition strategy: 25/32 eligible trajectories versus 13/32 for no-answer Baseline, with 35.4% fewer main-agent tokens. At fixed $K = 1$, it reaches 16/20 held-out successes versus 13/20 and reduces main tokens per success by 35.3%.

The ordering is informative. Early-Passive improves source yield but transfers poorly, while Late-Passive is competitive on transfer. Thus the result is not explained by simply adding answer tokens. The benefit is concentrated on the harder review-rating family (7/10 versus 4/10); order-detail is near ceiling (9/10 for both Baseline and Late-Guided). Outcome placement therefore interacts with task difficulty and the structure preserved by workflow induction.

Table 4: Two-family controlled answer-placement study. Main-agent tokens exclude the condition-blind companion. Bold marks the best aggregate condition; arrows are relative to no-answer Baseline.

Condition	Source eligible	Avg. steps	Source tokens	Tokens / eligible	$K=1$ held-out	Transfer tokens / success
No-answer Baseline	13/32 (40.6%)	13.22	1,534,909	118,070	13/20 (65.0%)	60,626
Early-Passive	19/32 (59.4%)	11.81	1,303,833	68,623	6/20 (30.0%)	193,528
Early-Guided	22/32 (68.8%)	10.97	1,139,259	51,785	11/20 (55.0%)	81,847
Late-Passive	20/32 (62.5%)	10.50	1,141,247	57,062	14/20 (70.0%)	52,805
Late-Guided	**25/32 (78.1%)**	**9.63**	**992,008**	**39,680**	**16/20 (80.0%)**	**39,211**
LG vs. Baseline	**+37.5 pp**	**↓27.2%**	**↓35.4%**	**↓66.4%**	**+15.0 pp**	**↓35.3%**

6.5 Evidence and the supervision flywheel

A two-stage audit provisionally classified 124 of 155 officially successful A2P query trajectories as webpage-supported (80.0%). The original literal check passed 83; semantic review of the other 72 recovered 41 supported trajectories. The 83 literal passes were not re-audited with the same semantic rubric. Among the 72 reviewed traces, no conclusive deliberate answer copying or shortcut was identified, while 31 lacked sufficient or consistent recorded support. Trajectory collection applies task/entity/field binding and a common finish guard.

The separate, condition-symmetric mechanical evidence flag gives 53/183 (29.0%) evidence-supported successful trajectories for Baseline and 83/183 (45.4%) for A2P, a 16.4-point gain in evidence-supported yield. Tokens per mechanically supported success fall by 44.5%. The mechanically supported fraction among official successes is a separate composition metric (71.6% versus 53.5%); the acquisition quantity is supported success per rollout.

The matched supervision ladder applies the same mechanical flag across seeds 1–2. Full GT attains the highest immediate success (110/122), while A2P attains 59/122 evidence-supported successes versus 52/122 for Full GT. Tokens per mechanically supported success are also slightly lower for A2P (92,508 versus 94,323). Full outcome exposure favors immediate completion; answer-conditioned path seeking yields more trajectories with mechanically recorded webpage support in this comparison.

A one-family terminal-verification control admits 5/8 source trajectories, versus 8/8 for full GT and 7/8 for evidence seeking. At fixed $K = 1$, held-out success is 5/5, 4/5, and 5/5, respectively (Appendix Table 7). Companion step scores provide a separate behavioral audit (Appendix C.2).

In the closed-loop pilot, three placement conditions each completed two rounds. Starting from eight official order-detail task–outcome pairs per condition, the system generated ten condition-local provenance-preserving records, corresponding to seven unique semantic task–outcome pairs after deduplicating repeats across isolated conditions. All ten records passed independent GUI validation and evidence checks; five subsequent acquisitions produced evidence-supported trajectories, and workflow memory was updated in four of six rounds (Appendix Table 5). This demonstrates an operational, backend-free supervision-expansion loop that converts webpage evidence into new tasks, new acquisitions, and updated workflow memory.

6.6 Cross-scaffold and cross-backbone evidence

An independent AgentLab GenericAgent replication on the full 61-task query set improves from 22/61 to 40/61, reduces tokens by 5.7%, and yields exact McNemar $p = 4.005 \times 10^{-5}$. With DeepSeek-V3.2 on the main scaffold, success improves from 24/61 to 49/61 and tokens per success fall by 70.3%. These replications extend the acquisition effect across model backbones and agent scaffolds (Appendix Table 12).

The condition-symmetric evidence flag also favors A2P on both replications (Appendix C.3). DeepSeek counts use the frozen provenance correction for finish attempts with environment-level zero rewards; no rollouts were replaced.

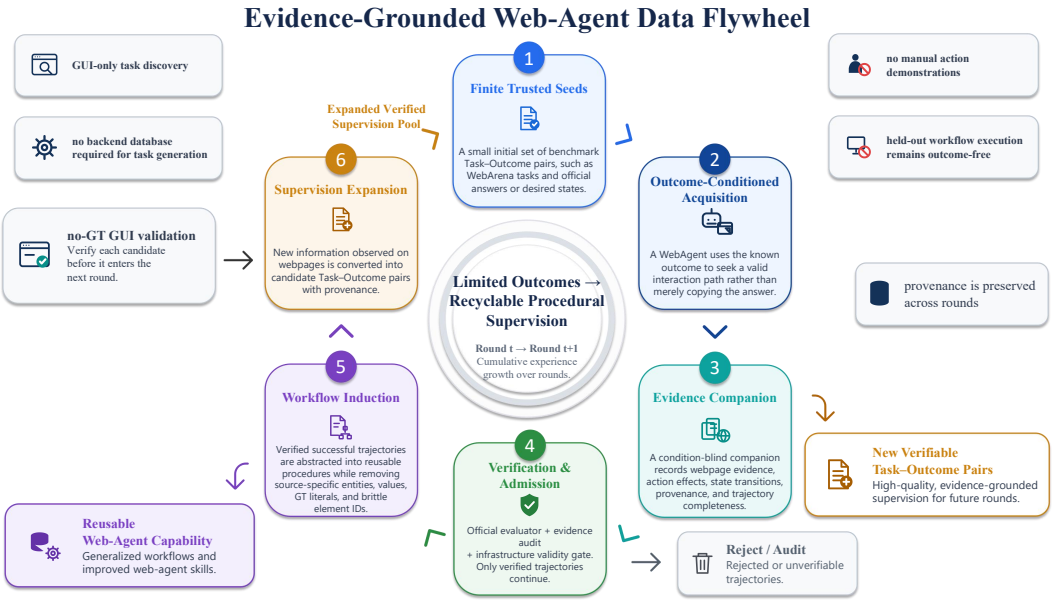


Figure 3: Evidence-grounded supervision flywheel. Finite trusted outcomes guide acquisition; independently admitted trajectories induce reusable workflows and expose provenance-preserving candidate tasks for the next validated round.

7 Analysis and Discussion

Acquisition and transfer are distinct. Fixed-budget acquisition measures the yield of successful experience, while fixed-K transfer holds the number of successful source trajectories constant. Query gains on both measures extend beyond source-task success. In modification transfer, A2P-side compact compilation contributes to the tested pipeline, so its pooled gain cannot be attributed to source trajectories alone. Placement results show that high source yield need not imply strong transfer.

Why late guidance can help. Early exposure may anchor the agent before it has parsed the page and task structure. Late guidance permits an initial task model, then narrows search toward evidence. The larger gain on review-rating than on near-ceiling order-detail tasks is consistent with this interpretation and suggests that outcome timing interacts with task difficulty.

Operational trajectory value. We assess experience through official success, recorded evidence, and no-GT workflow transfer. Companion step scores characterize behavior without determining task success, complementing automatic trajectory evaluation (Lù et al., 2025; Lin et al., 2025; Gu et al., 2026).

Scope and assumptions. The empirical scope is WebArena Shopping: 61 query tasks, four clean query-transfer families, six modification-acquisition families, and Wishlist/cart modification transfer. The GUI-only flywheel completes two rounds in one family. Acquisition starts from trusted outcome seeds; held-out tasks receive no outcome supervision.

8 Conclusion

Answer2Path uses known outcomes to acquire procedural experience rather than only score completed tasks. It raises eligible yield and lowers source tokens per success on query and modification tasks. Fixed-K query workflows transfer better to held-out tasks without outcomes. Outcome placement changes acquisition, while a two-round GUI-only pilot turns verified page evidence into new tasks. Together, these results connect targeted exploration, reusable workflows, and supervision expansion.

REFERENCES

Tamer Abuelsaad, Deepak Akkil, Prasenjit Dey, Ashish Jagmohan, Aditya Vempaty, and Ravi Kokku. Agent-e: From autonomous web navigation to foundational design principles in agentic systems, 2024. URL https://arxiv.org/abs/2407.13032.

Bizhe Bai, Jiakang Yuan, Hongming Wu, Xinyue Wang, Jie Ren, Siyao Chen, Yuchen Ya, Fan Bai, Pai Peng, Huafeng Qin, and Tao Chen. Efficient gui agents: A systems survey of observation, memory, action, and runtime optimization, 2026. URL https://arxiv.org/abs/2609.02309.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier De Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0b82662b6c32e887bb252a74d8cb2d5e-Abstract-Datasets_and_Benchmarks_Track.html.

Shaofei Cai, Yulei Qin, Haojia Lin, Zihan Xu, Gang Li, Yuchen Shi, Zongyi Li, Yong Mao, Siqi Cai, Xiaoyu Tan, Yitao Liang, Ke Li, and Xing Sun. Smartsnap: Proactive evidence seeking for self-verifying agents, 2025. URL https://arxiv.org/abs/2512.22322.

Yuan Cao, Dezhi Ran, Mengzhou Wu, Yuzhe Guo, Xin Chen, Ang Li, Gang Cao, Gong Zhi, Hao Yu, Linyi Li, Wei Yang, and Tao Xie. Gui-genesis: Automated synthesis of efficient environments with verifiable rewards for gui agent post-training, 2026. URL https://arxiv.org/abs/2602.14093.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. In *ACL 2024*, 2024. URL https://aclanthology.org/2024.acl-long.505/.

Chaoqun Cui, Jing Huang, Shijing Wang, Liming Zheng, Qingchao Kong, and Zhixiong Zeng. Agentic reward modeling: Verifying gui agent via progressive trajectory-grounded interaction, 2026. URL https://arxiv.org/abs/2602.00575.

Gaole Dai, Shiqi Jiang, Ting Cao, Yuanchun Li, Yuqing Yang, Rui Tan, Mo Li, and Lili Qiu. Advancing mobile gui agents: A verifier-driven approach to practical deployment. In *MobiCom 2026*, 2026a. URL https://arxiv.org/abs/2503.15937.

Gaole Dai, Shiqi Jiang, Ting Cao, Yuqing Yang, Yuanchun Li, Rui Tan, Mo Li, and Lili Qiu. Prore: A proactive reward system for gui agents via reasoner-actor collaboration. In *ICLR 2026*, 2026b. URL https://arxiv.org/abs/2509.21823.

Thibault Le Sellier de Chezelles, Maxime Gasse, Alexandre Lacoste, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, and Nicolas Chapados. The browsergym ecosystem for web agent research. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=5298fKGmv3.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *NeurIPS 2023 Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html.

Zheyuan Deng, Binghang Lu, Hanqi Feng, Shirley Huang, Dianzhuo Wang, Yuanda Xu, Zhiwei Zhang, Yige Sun, Changhong Mou, Runyu Zhang, Yuexing Hao, Barnabas Poczos, and Xiaomin Li. Contramem: Learning self-evolving procedural memory from contrasting multi-model trajectories, 2026. URL https://arxiv.org/abs/2608.22533.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? In *ICML 2024*, 2024. URL https://proceedings.mlr.press/v235/drouin24a.html.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, Tianshu Zhang, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, Kai Zhang, Boyuan Zheng, Zhaowei Cai, Viktor Rozgic, Morteza Ziyadi, Huan Sun, and Yu Su. Mind2web 2: Evaluating agentic search with agent-as-a-judge. In *NeurIPS 2025 Datasets & Benchmarks Track*, 2025a. URL https://openreview.net/forum?id=AUaW6DS9si.

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. Navigating the digital world as humans do: Universal visual grounding for gui agents. In *ICLR 2025*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4ca0e369689dadb25a5345ba9755ad6f-Abstract-Conference.html.

Chengyang Gu, Le Zhang, Jingbo Zhou, Yize Chen, Yu Shi, Siqi Bao, Zheng-Fan Wu, Hua Wu, and Hui Xiong. Beyond success and failure: Length-aware contrastive learning for gui agents, 2026. URL https://arxiv.org/abs/2608.21830.

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. A real-world webagent with planning, long context understanding, and program synthesis. In *ICLR 2024*, 2024. URL https://openreview.net/forum?id=9JQtrumvg8.

Bowei He, Xiaokun Zhang, Meng Ding, and Xue Liu. Scaffold: Self-improving web agents via recursive parametric skill abstraction, 2026a. URL https://arxiv.org/abs/2609.05511.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. In *ACL 2024*, 2024. URL https://aclanthology.org/2024.acl-long.371/.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Hongming Zhang, Tianqing Fang, Zhenzhong Lan, and Dong Yu. Openwebvoyager: Building multimodal web agents via iterative real-world exploration, feedback and optimization. In *ACL 2025*, 2025. URL https://aclanthology.org/2025.acl-long.1336/.

Tiantian He, Yihang Chen, Keyue Jiang, Ka Yiu Lee, Kaiwen Zhou, Kun Shao, and Shuai Wang. Ee-mcp: Self-evolving mcp-gui agents via automated environment generation and experience learning, 2026b. URL https://arxiv.org/abs/2604.09815.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. Cogagent: A visual language model for gui agents. In *CVPR 2024*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Hong_CogAgent_A_Visual_Language_Model_for_GUI_Agents_CVPR_2024_paper.html.

Longtao Hu, Xiao Liang, and Linchao Zhu. From interaction traces to persistent skills: Online evolution for computer-use agents, 2026. URL https://arxiv.org/abs/2609.04869.

Lawrence Jang, Yinheng Li, Dan Zhao, Charles Ding, Justin Lin, Paul Pu Liang, Rogerio Bonatti, and Kazuhito Koishida. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. In *ICLR 2025*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5b555804d495321df2e3208cc27f4fbc-Abstract-Conference.html.

Imene Kerboua, Sahar Omidi Shayegan, Megh Thakkar, Xing Han Lù, Léo Boisvert, Massimo Caccia, Jérémy Espinas, Alexandre Aussem, Véronique Eglin, and Alexandre Lacoste. Focusagent: Simple yet effective ways of trimming the large context of web agents, 2025. URL https://arxiv.org/abs/2510.03204.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *ACL 2024*, 2024. URL https://aclanthology.org/2024.acl-long.50/.

Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. Autowebglm: A large language model-based web navigating agent. In *KDD 2024*, 2024. URL https://dl.acm.org/doi/10.1145/3637528.3671620.

Jungjae Lee, Dongjae Lee, Chihun Choi, Youngmin Im, Jaeyoung Wi, Kihong Heo, Sangeun Oh, Sunjae Lee, and Insik Shin. Verisafe agent: Safeguarding mobile gui agent via logic-based action verification. In *MobiCom 2025*, 2025. URL https://arxiv.org/abs/2503.18492.

Runze Li, Yuwen Zhai, Bo Xu, LiWu Xu, Nian Shi, Wei Zhang, Ran Lin, and Liang Wang. Echotrail-gui: Building actionable memory for gui agents via critic-guided self-exploration, 2025. URL https://arxiv.org/abs/2512.19396.

Haojia Lin, Xiaoyu Tan, Yulei Qin, Zihan Xu, Yuchen Shi, Zongyi Li, Gang Li, Shaofei Cai, Siqi Cai, Chaoyou Fu, Ke Li, and Xing Sun. Cuarewardbench: A benchmark for evaluating reward models on computer-using agent, 2025. URL https://arxiv.org/abs/2510.18596.

Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners, 2025. URL https://arxiv.org/abs/2504.14239.

Ziwei Liu, Borui Kang, Hangjie Yuan, Zixiang Zhao, Wei Li, Yifan Zhu, and Tao Feng. Continual gui agents, 2026. URL https://arxiv.org/abs/2601.20732.

Fanbin Lu, Zhisheng Zhong, Ziqin Wei, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Steve: A step verification pipeline for computer-use agent training, 2025a. URL https://arxiv.org/abs/2503.12532.

Yadong Lu, Jianwei Yang, Yelong Shen, and Ahmed Awadallah. Omniparser for pure vision based gui agent, 2024. URL https://arxiv.org/abs/2408.00203.

Yuxuan Lu, Ziyi Wang, Jing Huang, Hui Liu, Jiri Gesi, Yan Han, Shihan Fu, Tianqi Zheng, Xianfeng Tang, Chen Luo, Yisi Sang, Jin Lai, and Dakuo Wang. Webserv: A full-stack and rl-ready web environment for training web agents at scale, 2025b. URL https://arxiv.org/abs/2510.16252.

Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents, 2025. URL https://arxiv.org/abs/2504.10458.

Bowen Lv, Xiao Liu, Yanyu Ren, Hanyu Lai, Bohao Jing, Hanchen Zhang, Yanxiao Zhao, Shuntian Yao, Jie Tang, and Yuxiao Dong. Scalecua: Scaling computer use agents with verifiable task synthesis and efficient online rl, 2026. URL https://arxiv.org/abs/2607.11185.

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with multi-turn dialogue. In *ICML 2024*, 2024. URL https://proceedings.mlr.press/v235/lu24e.html.

Xing Han Lù, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stańczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. Agentrewardbench: Evaluating automatic evaluations of web agent trajectories. In *COLM 2025*, 2025. URL https://arxiv.org/abs/2504.08942.

Dang Nguyen, Jian Chen, Yu Wang, Gang Wu, Namyong Park, Zhengmian Hu, Hanjia Lyu, Junda Wu, Ryan Aponte, Yu Xia, Xintong Li, Jing Shi, Hongjie Chen, Viet Dac Lai, Zhouhang Xie, Sungchul Kim, Ruiyi Zhang, Tong Yu, Mehrab Tanjim, Nesreen K. Ahmed, Puneet Mathur, Seunghyun Yoon, Lina Yao, Branislav Kveton, Jihyung Kil, Thien Huu Nguyen, Trung Bui, Tianyi Zhou, Ryan A. Rossi, and Franck Dernoncourt. Gui agents: A survey. In *Findings of ACL 2025*, 2025. URL https://arxiv.org/abs/2412.13501.

Vardaan Pahuja, Yadong Lu, Corby Rosset, Boyu Gou, Arindam Mitra, Spencer Whitehead, Yu Su, and Ahmed Hassan Awadallah. Explorer: Scaling exploration-driven web trajectory synthesis for multimodal web agents. In *Findings of ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.326/.

Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, and Zhengyang Wu. Webcanvas: Benchmarking web agents in online environments, 2024. URL https://arxiv.org/abs/2406.12373.

Ajay Patel, Markus Hofmarcher, Claudiu Leoveanu-Condrei, Marius-Constantin Dinu, Chris Callison-Burch, and Sepp Hochreiter. Large language models can self-improve at web agent tasks, 2024. URL https://arxiv.org/abs/2405.20309.

Ralph Peeters, Aaron Steiner, Luca Schwarz, Julian Yuya Caspary, and Christian Bizer. Webmall – a multi-shop benchmark for evaluating web agents, 2025. URL https://arxiv.org/abs/2508.13024.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, Jie Tang, and Yuxiao Dong. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *ICLR 2025*, 2025. URL https://openreview.net/forum?id=oVKEAFjEqv.

Yaoyao Qian, Yuanli Wang, Jinda Zhang, Yun Zong, Meixu Chen, Hanhan Zhou, Jindan Huang, Yifan Zeng, Xinyu Hu, Chan Hee Song, and Danqing Zhang. Webgrapheval: Multi-turn trajectory evaluation for web agents using graph representation. In *NeurIPS 2025 Workshop on Multi-Turn Interactions in Large Language Models*, 2025. URL https://arxiv.org/abs/2510.19205.

Ram Ramrakhya, Andrew Szot, Omar Attia, Yuhao Yang, Anh Nguyen, Bogdan Mazoure, Zhe Gan, Harsh Agrawal, and Alexander Toshev. Scaling synthetic task generation for agents via exploration. In *ICLR 2026*, 2026. URL https://arxiv.org/abs/2509.25047.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William E Bishop, Wei Li, Folawiyo Campbell-Ajala, Daniel Kenji Toyama, Robert James Berry, Divya Tyamagundlu, Timothy P Lillicrap, and Oriana Riva. Androidworld: A dynamic benchmarking environment for autonomous agents. In *ICLR 2025*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/01a83bc2f2732a58e6aa731e659e7101-Abstract-Conference.html.

Corby Rosset, Pratyusha Sharma, Andrew Zhao, Miguel Gonzalez-Fernandez, and Ahmed Awadallah. The art of building verifiers for computer use agents, 2026. URL https://arxiv.org/abs/2604.06240.

Rui Shao, Ruize Gao, Bin Xie, Yixing Li, Kaiwen Zhou, Shuai Wang, Weili Guan, and Gongwei Chen. Hats: Hardness-aware trajectory synthesis for gui agents. In *CVPR 2026*, 2026. URL https://arxiv.org/abs/2603.12138.

Yucheng Shi, Wenhao Yu, Jingyuan Huang, Wenlin Yao, Wenhu Chen, and Ninghao Liu. Towards trustworthy gui agents: A survey, 2025. URL https://arxiv.org/abs/2503.23434.

Yixiao Song, Katherine Thai, Chau Minh Pham, Yapei Chang, Mazin Nadaf, and Mohit Iyyer. Bearcubs: A benchmark for computer-using web agents. In *COLM 2025*, 2025. URL https://arxiv.org/abs/2503.07919.

Yurun Song, Jiong Yin, Rongjunchen Zhang, and Ian G. Harris. Compress to focus: Efficient coordinate compression for policy optimization in multi-turn gui agents, 2026. URL https://arxiv.org/abs/2601.11631.

Qiushi Sun, Kanzhi Cheng, Zichen Ding, Chuanyang Jin, Yian Wang, Fangzhi Xu, Zhenyu Wu, Chengyou Jia, Liheng Chen, Zhoumianze Liu, Ben Kao, Guohao Li, Junxian He, Yu Qiao, and Zhiyong Wu. Os-genesis: Automating gui agent trajectory construction via reverse task synthesis. In *ACL 2025*, 2025a. URL https://aclanthology.org/2025.acl-long.277/.

Zeyi Sun, Ziyu Liu, Yuhang Zang, Yuhang Cao, Xiaoyi Dong, Tong Wu, Dahua Lin, and Jiaqi Wang. Seagent: Self-evolving computer use agent with autonomous learning from experience, 2025b. URL https://arxiv.org/abs/2508.04700.

Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, Wanjun Zhong, Yining Ye, Yujia Qin, Yuwen Xiong, Yuxin Song, Zhiyong Wu, Aoyan Li, Bo Li, Chen Dun, Chong Liu, Daoguang Zan, Fuxing Leng, Hanbin Wang, Hao Yu, Haobin Chen, Hongyi Guo, Jing Su, Jingjia Huang, Kai Shen, Kaiyu Shi, Lin Yan, Peiyao Zhao, Pengfei Liu, Qinghao Ye, Renjie Zheng, Shulin Xin, Wayne Xin Zhao, Wen Heng, Wenhao Huang, Wenqian Wang, Xiaobo Qin, Yi Lin, Youbin Wu, Zehui Chen, Zihao Wang, Baoquan Zhong, Xinchun Zhang, Xujing Li, Yuanfan Li, Zhongkai Zhao, Chengquan Jiang, Faming Wu, Haotian Zhou, Jinlin Pang, Li Han, Qi Liu, Qianli Ma, Siyao Liu, Songhua Cai, Wenqi Fu, Xin Liu, Yaohui Wang, Zhi Zhang, Bo Zhou, Guoliang Li, Jiajun Shi, Jiale Yang, Jie Tang, Li Li, Qihua Han, Taoran Lu, Woyu Lin, Xiaokang Tong, Xinyao Li, Yichi Zhang, Yu Miao, Zhengxuan Jiang, Zili Li, Ziyuan Zhao, Chenxin Li, Dehua Ma, Feng Lin, Ge Zhang, Haihua Yang, Hangyu Guo, Hongda Zhu, Jiaheng Liu, Junda Du, Kai Cai, Kuanye Li, Lichen Yuan, Meilan Han, Minchao Wang, Shuyue Guo, Tianhao Cheng, Xiaobo Ma, Xiaojun Xiao, Xiaolong Huang, Xinjie Chen, Yidi Du, Yilin Chen, Yiwen Wang, Zhaojian Li, Zhenzhu Yang, Zhiyuan Zeng, Chaolin Jin, Chen Li, Hao Chen, Haoli Chen, Jian Chen, Qinghao Zhao, and Guang Shi. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning, 2025a. URL https://arxiv.org/abs/2509.02544.

Junjie Wang, Zequn Xie, Dan Yang, Jie Feng, Yue Shen, Duolin Sun, Meixiu Long, Yihan Jiao, Zhehao Tan, Jian Wang, Peng Wei, and Jinjie Gu. Webclipper: Efficient evolution of web agents with graph-based trajectory pruning, 2026a. URL https://arxiv.org/abs/2602.12852.

Shaokang Wang, Pei Fu, Ruoceng Zhang, Shaojie Zhang, Xiuwen Xi, Jiahui Yang, Bin Qin, Ying Huang, Zhenbo Luo, and Jian Luan. Gaia: A data flywheel system for training gui test-time scaling critic models, 2026b. URL https://arxiv.org/abs/2601.18197.

Shuai Wang, Weiwen Liu, Jingxuan Chen, Yuqi Zhou, Weinan Gan, Xingshan Zeng, Yuhan Che, Shuai Yu, Xinlong Hao, Kun Shao, Bin Wang, Chuhan Wu, Yasheng Wang, Ruiming Tang, and Jianye Hao. Gui agents with foundation models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2411.04890.

Yuhao Wang, Mu Qiao, Xindong Zhang, Yunzhi Zhuge, Lei Zhang, and Huchuan Lu. Trace: Trajectory-robust admission with evidence ordering for efficient gui agents, 2026c. URL https://arxiv.org/abs/2609.10297.

Zhaoyang Wang, Yiming Liang, Xuchao Zhang, Qianhui Wu, Siwei Han, Anson Bastos, Rujia Wang, Chetan Bansal, Baolin Peng, Jianfeng Gao, Saravan Rajmohan, and Huaxiu Yao. Synthagent: Adapting web agents with synthetic supervision. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15730–15752, 2026d. doi: 10.18653/v1/2026.acl-long.716. URL https://aclanthology.org/2026.acl-long.716/.

Zora Zhiruo Wang, Apurva Gandhi, Graham Neubig, and Daniel Fried. Inducing programmatic skills for agentic tasks. In *COLM 2025*, 2025b. URL https://openreview.net/forum?id=lsAY6fWsog.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *ICML 2025*, 2025c. URL https://proceedings.mlr.press/v267/wang25bx.html.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. In *EMNLP 2025*, 2025. URL https://arxiv.org/abs/2505.16421.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. Os-atlas: A foundation action model for generalist gui agents. In *ICLR 2025*, 2025. URL https://arxiv.org/abs/2410.23218.

Han Xiao, Guozhi Wang, Yuxiang Chai, Zimu Lu, Weifeng Lin, Hao He, Lue Fan, Liuyang Bian, Rui Hu, Liang Liu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Aojun Zhou, and Hongsheng Li. Ui-genie: A self-improving approach for iteratively boosting mllm-based mobile gui agents. In *NeurIPS 2025*, 2025. URL https://arxiv.org/abs/2505.21496.

Han Xiao, Guozhi Wang, Hao Wang, Shilong Liu, Yuxiang Chai, Yue Pan, Yufeng Zhou, Xiaoxin Chen, Yafei Wen, and Hongsheng Li. Ui-mem: Self-evolving experience memory for online reinforcement learning in mobile gui agents, 2026. URL https://arxiv.org/abs/2602.05832.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Jing Hua Toh, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Tao Xiong, Xavier Hu, Yurun Chen, Yuhang Liu, Changqiao Wu, Pengzhi Gao, Wei Liu, Jian Luan, and Shengyu Zhang. Gui-pra: Process reward agent for gui tasks, 2025. URL https://arxiv.org/abs/2509.23263.

Taofeng Xue, Chong Peng, Mianqiu Huang, Linsen Guo, Tiancheng Han, Haozhe Wang, Jianing Wang, Xiaocheng Zhang, Xin Yang, Dengchang Zhao, Jinrui Ding, Xiandi Ma, Yuchen Xie, Peng Pei, Xunliang Cai, and Xipeng Qiu. Evocua: Evolving computer use agents via learning from scalable synthetic experience, 2026. URL https://arxiv.org/abs/2601.15876.

Pei Yang, Hai Ci, and Mike Zheng Shou. macosworld: A multilingual interactive benchmark for gui agents. In *NeurIPS 2025*, 2025. URL https://arxiv.org/abs/2506.04135.

Pengshuai Yang, Zijing Gao, Xue Yu, Benhui Zhuang, Bo Yuan, and Junlan Feng. Automated trajectory evaluation for mobile agents via step-level consequence reasoning and aggregation, 2026. URL https://arxiv.org/abs/2608.20797.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS 2022*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html.

Tong Ye, Kunyang Han, Guozhi Wang, Longqiang Luo, Zhifeng Ding, Yongxiang Zhang, Xiaolei Shen, Yuxuan Zhang, Zhuping Zhang, Tao Xu, Yue Pan, Yucheng Zhao, Yupei Hu, Yuanjiang Ouyang, Danfeng Shen, Runqi Lin, Hongda Cai, Zhaoxiong Wang, Mengjia Yan, Yingjie Zhong, Chen Zhou, Zeyu Zhang, Xuwen Zhu, Penggang Shi, Mingcheng Luo, Ziyang Wu, Min Jin, Mingfu Shen, Zairong Xu, Fan Zhang, Hao Wang, Liang Liu, Zhulin Xie, Lijun Yao, Xiao Liang, Liangmin Wen, Liqiang Feng, Feilong Wu, Min Hu, Min Chen, Guanjing Xiong, Xiaohu Ruan, and Xiaoxin Chen. Bluelm-gui technical report: A real-device-centric flywheel for self-improving mobile gui agents, 2026. URL https://arxiv.org/abs/2609.12394.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? In *EMNLP 2024*, 2024. URL https://aclanthology.org/2024.emnlp-main.505/.

Tao Yu, Zhengbo Zhang, Zhiheng Lyu, Junhao Gong, Hongzhu Yi, Xinming Wang, Yuxuan Zhou, Jiabing Yang, Ping Nie, Yan Huang, and Wenhu Chen. Browseragent: Building web agents with human-inspired web browsing actions. *TMLR*, 2025. URL https://arxiv.org/abs/2510.10666.

Bofei Zhang, Zirui Shang, Zhi Gao, Wang Zhang, Rui Xie, Xiaojian Ma, Tao Yuan, Xinxiao Wu, Song-Chun Zhu, and Qing Li. Tongui: Internet-scale trajectories from multimodal web tutorials for generalized gui agents. In *AAAI 2026*, 2026a. URL https://arxiv.org/abs/2504.12679.

Danyang Zhang, Situo Zhang, Ziyue Yang, Zichen Zhu, Zihan Zhao, Ruisheng Cao, Lu Chen, and Kai Yu. Progrm: Build better gui agents with progress rewards, 2025a. URL https://arxiv.org/abs/2505.18121.

Junlei Zhang, Zichen Ding, Chang Ma, Zijie Chen, Qiushi Sun, Zhenzhong Lan, and Junxian He. Breaking the data barrier – building gui agents through task generalization. In *COLM 2025*, 2025b. URL https://openreview.net/forum?id=QDtORaZt8K.

Long Zhang, Yuhan Chen, Chaoran Zhang, Wanxia Cao, Kun Huang, Pengzhi Gao, Wei Liu, Jian Luan, Chenliang Li, and Lixin Zou. Gsar: Goal-state-anchor rewards for mobile gui agents with self-evolving data synthesis, 2026b. URL https://arxiv.org/abs/2608.22847.

Wenyuan Zhang, Xinghua Zhang, Haiyang Yu, Shuaiyi Nie, Bingli Wu, Juwei Yue, Tingwen Liu, and Yongbin Li. Expseek: Self-triggered experience seeking for web agents, 2026c. URL https://arxiv.org/abs/2601.08605.

Yao Zhang, Shijie Tang, Zeyu Li, Zhen Han, and Volker Tresp. Webarbiter: A principle-guided reasoning process reward model for web agents. In *ICLR 2026*, 2026d. URL https://arxiv.org/abs/2601.21872.

Yuzhe Zhang, Xianwei Xue, Xingyong Wu, Mengke Chen, Chen Liu, Xinran He, Run Shao, Feiran Liu, Huanmin Xu, Qiutong Pan, and Haiwei Wang. Don't act blindly: Robust gui automation via action-effect verification and self-correction. In *ACL 2026*, 2026e. URL https://arxiv.org/abs/2604.05477.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. Gpt-4v(ision) is a generalist web agent, if grounded. In *ICML 2024*, 2024a. URL https://proceedings.mlr.press/v235/zheng24e.html.

Boyuan Zheng, Michael Y. Fatemi, Xiaolong Jin, Zora Zhiruo Wang, Apurva Gandhi, Yueqi Song, Yu Gu, Jayanth Srinivasa, Gaowen Liu, Graham Neubig, and Yu Su. Skillweaver: Web agents can self-improve by discovering and honing skills, 2025a. URL https://arxiv.org/abs/2504.07079.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. Synapse: Trajectory-as-exemplar prompting with memory for computer control. In *ICLR 2024*, 2024b. URL https://openreview.net/forum?id=Pc8AU1aF5e.

Zihan Zheng, Tianle Cui, Taoran Wang, Fengtao Wang, Jiahui Pan, Lewei He, and Qianglong Chen. Naturalgaia: A verifiable benchmark and hierarchical framework for long-horizon gui tasks, 2025b. URL https://arxiv.org/abs/2508.01330.

Shuyan Zhou. Webarena-infinity: Generating browser environments with verifiable tasks at scale. Project website and code, 2026. URL https://github.com/web-arena-x/webarena-infinity.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *ICLR 2024*, 2024. URL https://arxiv.org/abs/2307.13854.

Yifei Zhou, Qianlan Yang, Kaixiang Lin, Min Bai, Xiong Zhou, Yu-Xiong Wang, Sergey Levine, and Li Erran Li. Proposer-agent-evaluator (pae): Autonomous skill discovery for foundation model internet agents. In *ICML 2025*, 2025. URL https://proceedings.mlr.press/v267/zhou25ah.html.

AI ASSISTANCE

Yes, to aid or polish writing. Details are described in the paper. Yes, for retrieval and discovery (e.g., finding related work). Details are described in the paper. Yes, for research ideation or execution. Details are described in the paper. Yes, to draft sections of the paper. Details are described in the paper.

Specifically, AI tools assisted with English phrasing, organization, and copyediting; identifying candidate papers for related-work review; discussing research questions and organizing experimental comparisons; and preparing draft text for selected manuscript sections.

A PROTOCOL SUMMARY

Query confirmatory. Sixty-one tasks, three seeds, GPT-5.5, temperature zero, 20-step maximum, paired Baseline/A2P, no repair or retry in the formal result.

Query fixed-K. Four clean functional families, frozen source/held-out split, $K = 1, 2, 4$, three seeds, identical AWM commit and executor, 180 paired held-out observations per condition.

Modification acquisition. Six functional families, eight source tasks and five held-out tasks per family, three seeds, environment reset between state-changing rollouts.

Answer placement. Two clean query families, eight source and five held-out tasks per family, seeds 0 and 1, five conditions, DeepSeek-V3.2, maximum 20 steps, matched task order and budgets.

B ADDITIONAL CONTROLLED RESULTS

Table 5: Evidence and closed-loop accounting. The 124/155 query figure combines 83 literal-check passes with semantic review of 72 literal-check failures; the 83 passes were not semantically re-audited. Flywheel rows report the single-family, two-round pilot.

Audited quantity	Result
Provisionally supported A2P query successes, two-stage audit	**124/155 (80.0%)**
Source trajectories passing official + evidence checks	**20/24 (83.3%)**
Independent flywheel rounds completed	**6/6**
Condition-local Task–Outcome records validated	**10**
Unique semantic Task–Outcome pairs	**7**
Downstream evidence-supported acquisitions from new pairs	**5**
Rounds updating reusable workflow memory	**4/6**
Unknown-use calls / child errors	**0 / 0**

Table 6: Matched query supervision ladder on the same 61 tasks and seeds 1–2. These conditions isolate how much outcome information is exposed during source acquisition; fixed-K transfer is evaluated separately.

Condition	Success	Avg. steps	Total tokens
No outcome	53/122 (43.4%)	6.81	6,593,513
Binary verifier	48/122 (39.3%)	12.78	12,765,863
Partial outcome	75/122 (61.5%)	5.02	4,836,972
Full GT	110/122 (90.2%)	5.09	4,904,795
A2P evidence-seeking	108/122 (88.5%)	5.61	5,457,969

Table 7: Terminal-verification control on one clean order-detail family. Main-agent source tokens exclude the condition-blind companion; all three conditions completed the attached two-round flywheel.

Condition	Source success	Avg. steps	Source tokens	K=1 held-out
Terminal verification	5/8	8.75	273,478	5/5
Full GT only	8/8	4.625	112,348	4/5
Evidence-seeking guidance	7/8	5.00	124,797	5/5

Table 8: Modification source evidence audit on the frozen v8 protocol. State support is the primary criterion; page support is auxiliary for state-ending tasks.

Metric	Baseline	A2P
Official + state support	33/144	78/144
Page-supported success	32/144	36/144
Total source tokens	8,685,791	5,136,852
Tokens / state-supported success	263,206	65,857
Tokens / page-supported success	271,431	142,690

Table 9: Fixed-K breakdowns. Query transfer improves for every source seed, while modification transfer yields positive pooled gains across Wishlist and cart (Tables 2 and 3).

(a) Query transfer by source seed

Seed	Baseline-AWM	A2P-AWM	Δ pp	A2P-only / B-only
6	54/60 (90.0%)	**60/60 (100.0%)**	**+10.0**	6 / 0
7	53/60 (88.3%)	**57/60 (95.0%)**	**+6.7**	6 / 2
8	52/60 (86.7%)	**58/60 (96.7%)**	**+10.0**	7 / 1

(b) Modification transfer by family and K

Family	K	Baseline-AWM	A2P-AWM	Δ pp
Wishlist	1	12/15 (80.0%)	**14/15 (93.3%)**	**+13.3**
Wishlist	2	9/15 (60.0%)	**13/15 (86.7%)**	**+26.7**
Wishlist	4	**10/15 (66.7%)**	9/15 (60.0%)	−6.7
Cart	1	9/15 (60.0%)	**12/15 (80.0%)**	**+20.0**
Cart	2	8/15 (53.3%)	**12/15 (80.0%)**	**+26.7**
Cart	4	9/15 (60.0%)	**12/15 (80.0%)**	**+20.0**

C Additional Comparisons and Audits

Table 10: Closest methodological families and the distinction of ANSWER2PATH. "Outcome placement" means the same known outcome is deliberately inserted at different stages.

Line of work	Representative methods	Main object	Distinction of ANSWER2PATH
Workflow memory	AWM, SkillWeaver, SCAFFOLD	Reuse past experience	Studies how outcome guidance changes acquired experience
Trajectory synthesis	OS-Genesis, Explorer, HATS	Generate or explore tasks	Starts from finite known outcomes and audits webpage evidence
Self-improving agents	WebRL, UI-Genie, SEAgent	Optimize policy/data loop	Separates acquisition yield from fixed-K transfer value
Trajectory evaluation	AgentRewardBench, STEVE, WebGraphEval	Judge success, steps, or trajectory structure	Audits evidence separately from official task success
Environment/task generation	GUI-GENESIS, SCALECUA, WebArena-Infinity	Scale tasks/worlds	Extracts GUI-grounded Task–GT pairs without backend task generation

Table 11: Query family generalization over 40 tasks and three seeds. Strict success additionally requires trajectory verification. Cross-family has one seed and is not pooled with the 120-run conditions. $^{\dagger}p < .001$ and $^{*}p < .05$ versus Baseline by exact paired test.

Condition	n	Official	Strict	Avg. steps	Δ off.
Baseline	120	95/120 (79.2%)	93/120 (77.5%)	9.09	–
A2P	120	**110/120 (91.7%)**†	**109/120 (90.8%)**	8.03	**+12.5**
Same-family path	120	107/120 (89.2%)*	106/120 (88.3%)	8.41	**+10.0**
A2P + same-family path	120	**112/120 (93.3%)**	**111/120 (92.5%)**	**7.82**	**+14.1**
Cross-family control	40	34/40 (85.0%)	31/40 (77.5%)	8.68	n/a

Table 12: Acquisition replications across agent scaffolds and model backbones. Protocols differ and are reported separately. Token Δ is total rollout cost.

Task	Scaffold	Model	Seeds	Baseline	A2P	Δ / token Δ
Query	Main DOM	GPT-5.5	3	74/183	**155/183**	**+44.3 pp / −13.1%**
Query	AgentLab GenericAgent	GPT-5.5	1	22/61	**40/61**	**+29.5 pp / −5.7%**
Query	Main DOM	DeepSeek-V3.2	1	24/61	**49/61**	**+41.0 pp / −39.3%**
Modification	Main DOM	GPT-5.6-sol	1	6/20	**16/20**	**+50.0 pp / −28.7%**

C.1 Modification transfer provenance

The six-family source audit distinguishes state-transition verification from query-style browser/page evidence. A successful state change need not leave the specific page fact captured by the auxiliary

query-style flag. Applying the same audit to the 90 fixed-K held-out rollouts gives 57/90 state-supported successes for Baseline and 72/90 for A2P, with page support matching both counts. Transfer tokens per state-supported success are 47,136 and 23,773, respectively.

The frozen v8 standard-compact manifest hash matches its recorded file. Source and held-out IDs and instance payloads are disjoint within Wishlist and cart, and each family retains one functional signature. All 12 Baseline/compact workflow artifacts match their manifest hashes; all six A2P artifacts record standard compression. Held-out input contains only the task and induced workflow, with no GT, source value, source trajectory, or backend access.

A content-level audit of the same 12 artifacts finds zero source-instance or GT literal hits and zero position-dependent phrases. Compact A2P artifacts are shorter than their paired Baseline artifacts in four of six cells; two retain longer structural grounding and final-state checks. A small number of artifacts retain element-ID references, recorded separately from source-entity leakage.

C.2 Companion step audit

The condition-blind companion rubric was applied to saved browser traces without new model calls or score-based trajectory admission. Among 23 scored source traces, terminal-only runs have mean step score 1.060, useful-step fraction 18.0%, and harmful-step fraction 26.0%; the corresponding figures are 1.346/37.8%/0.0% for GT-only and 1.230/30.0%/5.0% for evidence seeking. Official evaluation remains the task-success measure.

C.3 Cross-system evidence and provenance

In the independent AgentLab replication, the mechanical evidence flag gives 15/61 evidence-supported successes for Baseline and 25/61 for A2P, a 16.4-point yield gain. Tokens per evidence-supported success fall from 454,082 to 256,951.

The frozen DeepSeek trajectory audit reconciles executed finish attempts with official scores under environment-level zero rewards. The pre-registered correction changes Baseline/A2P success counts from 19/61 and 48/61 to 24/61 and 49/61, respectively, without replacing rollouts. Evidence-supported success is 15/61 versus 24/61; tokens per evidence-supported success are 177,824 versus 67,466.