How Far Can Language Agents Learn from the Internet? A Critical Survey of Web Agents

Anonymous ACL submission

Abstract

As large language models (LLMs) increasingly power autonomous web agents, a central question is: **how can LLM-based agents be made robust, scalable, and sustainable on the real internet?** This critical survey shifts attention from algorithmic performance in simulated benchmarks to the practical barriers of real-world application. We identify recurring bottlenecks from three perspectives: **environment and benchmarking, cognitive agent frameworks, and data synthesis for agent tuning**. First, we argue that current benchmarks, often built on static datasets or simplified web simulators, obscure key deployment challenges such as website variability, open-ended task ambiguity, and evaluation instability. Second, we deconstruct existing agent frameworks, showing how perception, decision-making, and execution modules remain fragile under dynamic layouts, long interaction contexts, iterative LLM latency, cost-error trade-offs, and unexpected UI states. Third, we examine data synthesis pipelines, where both top-down and bottom-up approaches struggle to produce diverse, executable, and generalizable training data. Beyond these, we propose a verifiable checklist to guarantee the high quality of real-world web agent for reference in future works. Only through such a systems-first approach can we progress toward practical web agents capable of long-term, adaptive operation on the real internet.

1 Introduction

Large language models (LLMs) (OpenAI, 2024; Singh et al., 2025; DeepSeek-AI, 2025; Team, 2025) have expanded autonomous agents from rigid, rule-based systems to open-world actors capable of advanced reasoning and planning (Liu et al., 2025d; Tang et al., 2023, 2024). This evolution has driven their application across diverse domains, notably finance (Ding et al., 2024), software engineering (Liu et al., 2024; Jimenez et al., 2024), and, crucially, autonomous web task (Wu et al., 2025; Li et al., 2025a; Drouin et al., 2024) execution (as shown in Figure 1).

Recent benchmarks (Zhou et al., 2024; Drouin et al., 2024) and frameworks (Yang et al., 2025c; Xi et al., 2025; de Chezelles et al., 2025) suggest rapid progress in web agents. However, their impressive performance in controlled settings may create a misleading optimism. Agents that perform well in curated settings still often fail catastrophically in the dynamic, noisy reality of the live web, exposing critical deficits in robustness, generalization, and long-term reasoning. We therefore advocate for a shift **from imitation in closed worlds to adaptive agents evaluated under real deployment pressures**.

While prior surveys often isolate specific aspects like planning (Huang et al., 2024b), memory (Hu et al., 2025d; Jia et al., 2026), benchmarking (Zhu et al., 2025b), or general progress in web agent (Ning et al., 2025), we provide a **panoramic** diagnosis of the **bottlenecks hindering web agents**. Rather than merely cataloging techniques, we structure our analysis around three critical challenges: (1) **Benchmarking** (Section 3), where we call for realistic environments and rigorous evaluation to expose the failures of static setups; (2) **Agent Frameworks** (Section 4), diagnosing what happens when LLM agents meet cognitive bottlenecks such as perceptual overload; and (3) **Data Synthesis** (Section 5), exploring how to synthesize data widely, deeply, and efficiently. Finally, in Section 6, we provide a checklist (Table 1) and discuss how these challenges might be addressed to move towards more adaptable web agents.

2 Preliminary

Following prior work (Xi et al., 2024, 2025), we formalize web tasks as a partially observable Markov decision process (POMDP) $\mathcal{M} =$

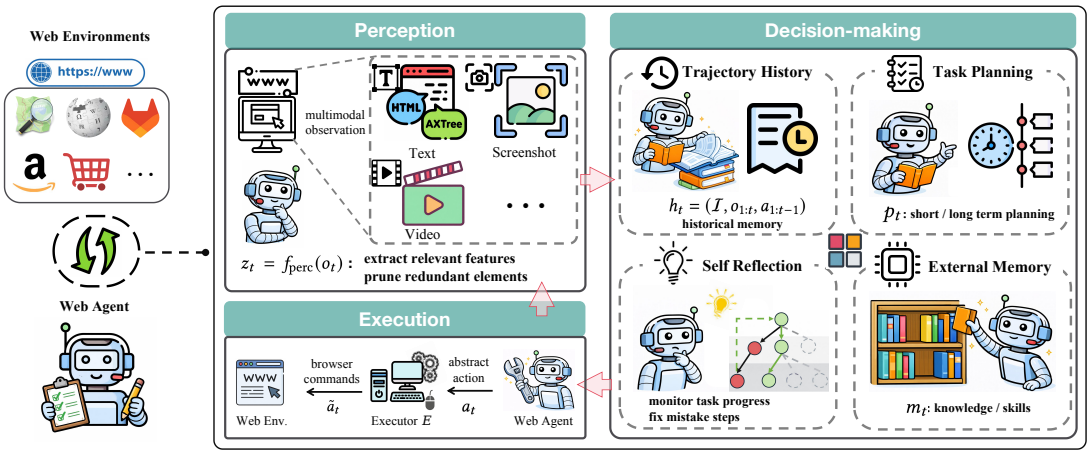

Figure 1: Overview of web agent.

$(\mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R})$. Here, $\mathcal{I}$ denotes the instruction space, and the state $\mathcal{S}$ consists of persistent components, such as URLs and backend records, and volatile components, such as DOM structures and text-box inputs (Dihan et al., 2025). The agent acts in action space $\mathcal{A}$, induces transitions $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, receives observations through $\phi : \mathcal{S} \to \mathcal{O}$, and is evaluated by a task reward $\mathcal{R} : \mathcal{I} \times \mathcal{S} \to \mathbb{R}$.

For a given task $I \in \mathcal{I}$, web agent generates a trajectory τ defined as a sequence of observations and actions: $\tau = (o_1, a_1, o_2 \ldots, a_{n-1}, o_n)$. At each step t, the agent receives an observation o_t and maintains a history $h_t = (\mathcal{I}, o_{1:t}, a_{1:t-1})$. We decompose the policy π_θ into three stages:

- Perception: $z_t = f_{\text{perc}}(o_t)$ extracts task-relevant features from the observation.
- Decision-making: $d_t = \psi(z_t, h_t, m_t, p_t)$ integrates perception, interaction history, external memory, and planning.
- Execution: the policy selects an action a_t, which an executor E grounds into low-level browser commands $\tilde{a}_t$.

The interaction ends upon a terminating action or exceeding the budget N. The objective is to optimize the policy parameters θ to maximize the expected return R over the task distribution $\mathcal{D}$: $\theta^* = \arg\max_\theta \mathbb{E}_{\mathcal{I}\sim\mathcal{D}, \tau\sim\pi_\theta}[R(\tau, \mathcal{I})]$

3 Benchmarking: Call for Realistic Environment and Rigorous Evaluation

In real world deployment, web agents operate in dynamic and failure-prone web environments, where success depends not only on final outcomes but also on how tasks are executed. Current benchmarks often simplify both the environment and the evaluation, failing to capture the complexity of the live web. In this section, we analyze the limitations of current benchmarking setups by proposing the following research questions:

- ***RQ1:*** *How to build realistic yet controllable web environments for agent interaction?*
- ***RQ2:*** *How to guarantee robust metrics and reproducible process for evaluation reliability?*

As illustrated in Figure 2 (a) and Table 3, we will analyze the problem of current web agent environments (§3.1) and evaluation methodologies (§3.2) in the following section.

3.1 Environment

3.1.1 Sim-to-Real Gap

Web agent benchmarks face an inherent trade-off between scalability and realism (Table section B.1). Early synthetic environments (Yao et al., 2022; Shi et al., 2017; Liu et al., 2018) are limited to toy tasks, whereas containerized sandboxes (Zhou et al., 2024; He et al., 2024; Drouin et al., 2024) improve realism but still do not address real-world challenges like CAPTCHAs and dynamic content. Conversely, live-web benchmarks (Xue et al., 2025; Anupam et al., 2025) suffer from irreproducibility due to frequent UI updates and transient noise.

To bridge this gap, recent frameworks explicitly inject realistic complexities—including adversarial noise, dynamic changes, and infrastructural failures—into standardized environments (Kara et al., 2025; Boisvert et al., 2025). The severe performance degradation of state-of-the-art agents under these conditions exposes a critical vulnerability to real-world operational risks.

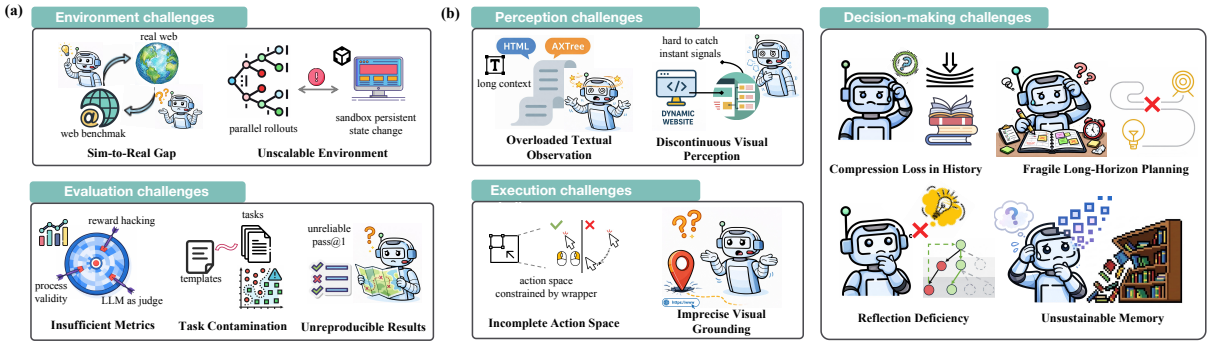

Figure 2: Current Challenges in benchmark (a) and agent framework (b).

3.1.2 Unscalable Environment

Scaling web agent evaluation is severely bottlenecked by heavy container overhead (e.g., Docker) (Zhou et al., 2024; Xie et al., 2024) and frequent environment resets. Because agents often execute destructive actions (Dihan et al., 2025) that permanently alter the system state $\mathcal{S}_{per}$ (e.g., modifying database records), benchmarks must enforce strict state resets before each rollout. This requirement cripples data collection throughput and parallelizable reinforcement learning algorithms (e.g., GRPO (Shao et al., 2024)) that depend on massive rollouts, as concurrent execution on a single instance causes state pollution and invalidates the reward signal R.

To mitigate these bottlenecks, recent infrastructure-level optimizations introduce heuristic backtracking to revert system changes (Dihan et al., 2025) or leverage lightweight containerization to alleviate virtualization overhead (Lu et al., 2025). While these initial efforts remain limited to specific benchmark suites (Zhou et al., 2024), they underscore the urgent need for scalable, high-throughput environment management to support large-scale reinforcement learning.

3.2 Evaluation

3.2.1 Insufficient Metrics.

Current evaluation often relies on **rigid, outcome-based metrics** (e.g., exact string or URL matching) (Zhou et al., 2024; Drouin et al., 2024), which notoriously penalize valid execution trajectories for minor formatting mismatches (see case study in Appendix B.3). More critically, they ignore process validity, omitting crucial factors like task progress, operational cost and unintended side effects. Conversely, transitioning to process-based rewards (Chae et al., 2025b; Fan et al., 2026) or LLM-as-a-judge evaluators (Gandhi and Neubig, 2025; Xue et al., 2025) introduces vulnerabilities to reward hacking. For instance, evaluators often mistakenly reward low-utility intermediate actions, incentivizing agents to produce redundant tool calls (He et al., 2025a). Designing robust process metrics that accurately reflect progress without incentivizing such suboptimal behavior remains an open challenge.

3.2.2 Task Contamination Risks.

Template-based task construction is widely adopted in web agent benchmarks (Zhou et al., 2024; Drouin et al., 2024) to ensure scalable generation and automatic validation. However, this paradigm introduces structural generalization risks. Because evaluations frequently rely on splits where training and test tasks share identical templates, a substantial template-level overlap often persists beneath instance-disjoint tasks (Lewis et al., 2021; Roig, 2025). As detailed in Appendix B, such hidden overlaps can partially inflate reported performance gains through procedural memorization rather than generalized reasoning. This pattern suggests that high in-distribution success rates can mask underlying brittleness when agents encounter genuinely novel task structures or cross-domain workflows.

3.2.3 Unreproducible Results.

Single-run evaluations are fundamentally unreliable due to compounding variance from external factors (API instability, UI updates, rate limits) and the inherent stochasticity of LLM policies. As demonstrated in complex agentic benchmarks (e.g., GAIA Level 3 (Ko et al., 2026)), trial-to-trial stochasticity—rather than actual task difficulty—accounts for a significant portion of performance variance (Mustahsan et al., 2025). Consequently, reliable evaluation protocols must mandate multiple rollouts and report stability metrics alongside raw accuracy to distinguish genuinely robust policies from random

outliers.

4 Agent Frameworks: When LLM Agent Meets Cognitive Bottlenecks

Web agents are commonly organized as perception, decision-making, and execution during a robust cognitive loop (see Section 2). However, it encounters significant bottlenecks when exposed to dynamic and noisy web environments. In this section, we will discuss the following questions:

- ***RQ1:*** *How can agents extract robust and task-relevant perceptions from web pages?*
- ***RQ2:*** *How can agents maintain a precise and sustainable decision state in long-horizon web tasks ?*
- ***RQ3:*** *How can agents' actions be effectively grounded during interactions with web elements?*

We provide an analysis from three perspectives of an agent framework including perception (§4.1), decision-making (§4.2), and execution (§4.3).

4.1 Perception

The perception function f_{perc} maps noisy web inputs (HTML (Gur et al., 2023), accessibility trees (axtree) (Kim et al., 2023; Lo et al., 2023), and screenshots (Gao et al., 2023; Zhang and Zhang, 2024)) into structured representations for decision-making d_t. Existing proxy-based or trainable implementations (Lee et al., 2025; Deng et al., 2023; Kerboua et al., 2025a) reduce raw page complexity, yet two gaps remain.

4.1.1 Overloaded Textual Observation

Most web agents still rely on textualized DOM or accessibility-tree inputs for grounding. These inputs are long, noisy, and only partially aligned with tasks. Heuristic pruning (Yang et al., 2025c; Zhang et al., 2025c) is efficient but can remove necessary context; retrieval-based reranking (Lù et al., 2024; Deng et al., 2023) can prioritize semantically similar distractors (e.g., *Best Sellers* widget at homepage) while missing fine-grained constraints (e.g., the 2023 time restriction); and plan-aware rewriting (Lee et al., 2025; Kerboua et al., 2025b) improves selectivity at the cost of extra LLM calls and latency. Thus, perception remains caught between context budget, recall, and responsiveness.

4.1.2 Discontinuous Visual Perception

Multimodal observations help recover spatial information that text-only inputs miss (Koh et al., 2024; Qin et al., 2025), but current frameworks usually perceive the web through static snapshots (Liu et al., 2025c; de Chezelles et al., 2025; Xi et al., 2024). This creates a temporal blind spot: loading states, hover menus, animations, and transient pop-ups can appear and disappear between observations. Progressive-Disclosure (Springer and Whittaker, 2019) interfaces further force agents to act before they have enough visual evidence (Qiu et al., 2026), producing unnecessary retries, premature termination, or mistaken assumptions about task completion.

4.2 Decision-making

Constructing a comprehensive decision state $d_t = \psi(z_t, h_t, m_t, p_t)$ is fundamental for effective web operation. In existing methods, d_t typically includes several key components: (1) **Interaction History** h_t, which tracks past observations and actions to ensure task continuity; (2) **Task Planning** p_t , which breaks down high-level goals into executable sub-tasks; (3) **Self-Reflection**, which monitors task progress and corrects mistakes; (4) **External Memory** m_t, which provides persistent knowledge and skills(e.g., API tools or expert "tips") that enhance agent's reasoning.

4.2.1 Compression Loss in History

ReAct-style agents (Yao et al., 2023b; Li et al., 2025c) append raw $(Obs, Action)$ triplets into history, quickly exhausting the context window in long-horizon tasks. Sliding history window or action-only histories (Gandhi and Neubig, 2025; Qin et al., 2025; Liu et al., 2025c) reduce length but discard early constraints, leading to information gaps. While active compression and folding methods (Yu et al., 2025a; Ye et al., 2025; Liu et al., 2025b; Chen et al., 2025a) add cost and can overwrite evidence that later becomes decisive. The core issue is not simply context length, but irreversible loss of task-critical state.

4.2.2 Misalignment between High-level Planning and Low-level Action

Plan-then-act agents (Yao et al., 2023b; Shinn et al., 2023) decompose a user goal into substeps, but static plans often diverge from live page states and introduce hallucination. Appendix C gives representative failure cases. Linear or tree-structured decomposition (Erdogan et al., 2025; Lo et al., 2023; Wei et al., 2022; Yao et al., 2023a) can mismatch the granularity of executable actions, while adap-

tive refinement (Shinn et al., 2023; Hu et al., 2025b; Yang et al., 2025b) depends on reliable feedback that may be missing or misleading.

Search-based methods such as Best-of-N (Gonzalez-Pumariega et al., 2025), self-consistency (Wang et al., 2023; Sodhi et al., 2023), state-space machines (Ma et al., 2023; Zhang et al., 2025b), and MCTS (Yu et al., 2025d; Zhang et al., 2025e; Dihan et al., 2025) improve exploration, but are difficult to deploy safely because web actions can be irreversible and page rendering is slow.

World models (Gu et al., 2025; Gao et al., 2025; Chae et al., 2025a; Fang et al., 2025) reduce this cost by simulating transitions, yet long-horizon hallucination remains common on unseen interfaces.

4.2.3 Fragile Reflection

Reflection mechanisms aim to turn execution feedback into self-correction, but they often rely on internal judgments rather than grounded process verification. Without external verifiers, agents frequently misdiagnose failures (Wang et al., 2024; He et al., 2025a). Furthermore, because these internal reflections lack granular grounding in specific DOM observations, agents tend to encode generic advice rather than actionable insights. This shallow feedback ultimately yields rigid, template-overfitted rules that fail to generalize to novel layouts (Azam et al., 2025; He et al., 2025a; Li et al., 2025b, 2024b).

This process also introduces severe inefficiency. Search-based frameworks like Reflective-MCTS (Yu et al., 2025d; Dihan et al., 2025) scale token consumption exponentially with reasoning depth, rendering them economically impractical for real-world deployment (Wang et al., 2024).

4.2.4 Unsustainable Memory Mechanisms

Long-term memory introduces parallel deployment bottlenecks in generalization, robustness, and scalability. First, abstracting past experiences into structured workflows (Wang et al., 2025e,d), API skills (Zheng et al., 2025; Yu et al., 2025b), or trajectory records (Zheng et al., 2024b; Chen et al., 2025c) often sacrifices essential environmental grounding. These site-specific heuristics and human-provided tips (Zhou et al., 2026) struggle to transfer across changing interfaces. Consequently, while boosting in-benchmark scores, agents risk overfitting by memorizing specific templates (Zhou et al., 2024; Koh et al., 2024) rather than learning generalizable skills, failing significantly under distribution shifts (Appendix D).

Furthermore, these systems are highly vulnerable to memory poisoning. Relying on imperfect self-judgments (Ouyang et al., 2025; Huang et al., 2024a) allows hallucinated successes to be permanently encoded. Without external verification, this internal pollution accumulates over time (Huang et al., 2025; Zhang et al., 2025a), causing the system to degenerate.

Finally, handling an endless stream of real-world experiences exposes an unscalable architecture. Most current memory banks rely on an append-only design (Yuan et al., 2025; Ouyang et al., 2025; Ye et al., 2025; Jia et al., 2026; Yan et al., 2025; Hu et al., 2025c). The lack of maintenance protocols such as active forgetting and consolidation inevitably leads to memory overload, where outdated or context-mismatched retrievals actively inject noise into the agent's reasoning.

4.3 Execution

Execution bridges a web agent's digital reasoning with its physical browser interaction. However, current frameworks face fundamental grounding bottlenecks:

4.3.1 Incomplete Action Space

Most web agents rely on environmental wrappers (Xi et al., 2024; de Chezelles et al., 2025) that restrict interactions to a limited set of discrete, element-bound operations (Yang et al., 2025c). This abstraction creates an incomplete action space, rendering tasks that require continuous control (e.g., dragging) fundamentally unsolvable. While recent works attempt to patch this limitation by heuristically incorporating continuous gestures (Wang et al., 2026a) or automatically mining composite skills (Zheng et al., 2025), these high-level abstractions demand rigorous verification and consistently fail to generalize across varied web layouts.

4.3.2 Imprecise Visual Grounding

Visual grounding approaches predict actions directly from pixels to bypass the long-context bottleneck of raw HTML (Li et al., 2024a; Mao et al., 2025; He et al., 2025c). However, this introduces severe spatial hallucinations, as vision-language models natively struggle with precise pixel-level localization (Zheng et al., 2024a; Cheng et al., 2024). Although visual overlays like Set-of-Mark and OmniParser (Yang et al., 2023; Wan et al., 2024; Yu et al., 2025c) attempt to mitigate this by tagging

interactable elements with bounding boxes, these methods remain highly brittle to slight UI scaling and frequently fail to distinguish visually similar but functionally distinct components.

5 Data Synthesis: How to Explore Widely, Deeply and Efficiently?

Training reliable web agents requires diverse, executable, and scalable data (Shen et al., 2024; Pandit et al., 2025). Manually collecting such data is expensive and hard to cover the width and depth of real web tasks.

This data scarcity has driven researchers toward **automated data synthesis**, where agents explore environments and generate `<task, trajectory>` pairs $(\mathcal{I}, \tau)$ as training samples (Mai et al., 2025). Current paradigms bifurcate into two main routes: **top-down** synthesis (instruction-guided exploration) and **bottom-up** synthesis (post-hoc task extraction from unguided interactions).

In this section, we analyze the key bottlenecks in data synthesis (Figure 3) regarding exploration (§5.1) and efficiency (§5.2), mainly focusing on the questions below:

- ***RQ1:*** *How can agents efficiently balance exploration width and depth to generate diverse and executable trajectories?*
- ***RQ2:*** *How to improve exploration efficiency and data utilization during synthesis?*

5.1 Exploration Width and Depth

5.1.1 Top-down: Instruction First

Tasks generated by a top-down policy $\pi_{\text{task}}(\cdot|\mathcal{P}_{\text{ref}})$ act as pre-defined hypotheses without environmental verification, where $\mathcal{P}_{\text{ref}}$ represents human-annotated templates or partial observations.

Template-based methods (Lù et al., 2024; Zhou et al., 2025; Lai et al., 2024; Fan et al., 2026) typically mimic benchmark patterns, restricting synthesis to a narrow set of low-variety intents. While subsequent approaches (Gandhi and Neubig, 2025; Wang et al., 2025c; Fan et al., 2026) leverage LLMs to propose tasks from snapshots, this soft guidance still restricts task type heavily toward simple page navigation or information seeking (Gandhi and Neubig, 2025).

Crucially, generating long-horizon tasks from static observations introduces severe hallucination, as models blindly predict unvisited states (e.g., proposing to "fork a repository" without confirming its existence) and fail to enforce precise difficulty control for curriculum learning.

Fundamentally, this paradigm lacks environmental grounding; unlike a site administrator with absolute system knowledge, the model operates blindly. Although recent work (Fan et al., 2026) mitigates hallucination by positioning the LLM as an environment architect that constructs the web state alongside the task, synthesis diversity remains heavily bottlenecked by the model's parametric knowledge.

5.1.2 Bottom-up: Interaction First

Bottom-up synthesis inverts this paradigm by exploring the web environment before abstracting tasks. This interaction-driven approach ensures feasible trajectories, while it often fails to reach the **depth** of complex workflows without explicit goals.

Without predefined tasks, exploration easily becomes a "random walk" (Sun et al., 2025a; Gandhi and Neubig, 2025), leading to logically **fragmented** or **purposeless** trajectories. Even after post filtering, many task-trajectory pairs can remain incoherent. Recent works reintroduce soft prior knowledge like concept pools (Mai et al., 2025) or guidance documents (Su et al., 2025), but the challenge remains in guiding exploration effectively while maintaining task diversity.

Consequently, synthesized tasks lack depth. Real-world tasks often require specific actions, such as typing a precise search query or applying multiple filters. A wandering agent is statistically unlikely to "stumble upon" these high-intent inputs, favoring simple clicks instead, leaving many functional elements within the website untouched (Fan et al., 2026). This results in overly simple, shallow tasks that fail to capture the long-tail of complex, multi-step real-world operations.

5.2 Efficiency Bottleneck

Beyond width-depth trade-off, the synthesis process suffers from severe inefficiency. Whether starting from task (Top-down) or trajectory (Bottom-up), the misalignment between agent's policy and web environment leads to massive computational waste and low data utilization.

5.2.1 Top-down: Task-Trajectory Misalignment

Directly executing ungrounded tasks often results in severe misalignment and corrupts entire trajectory. Some methods (Wang et al., 2025c; Yang et al., 2025b; Wang et al., 2025b) allow the task

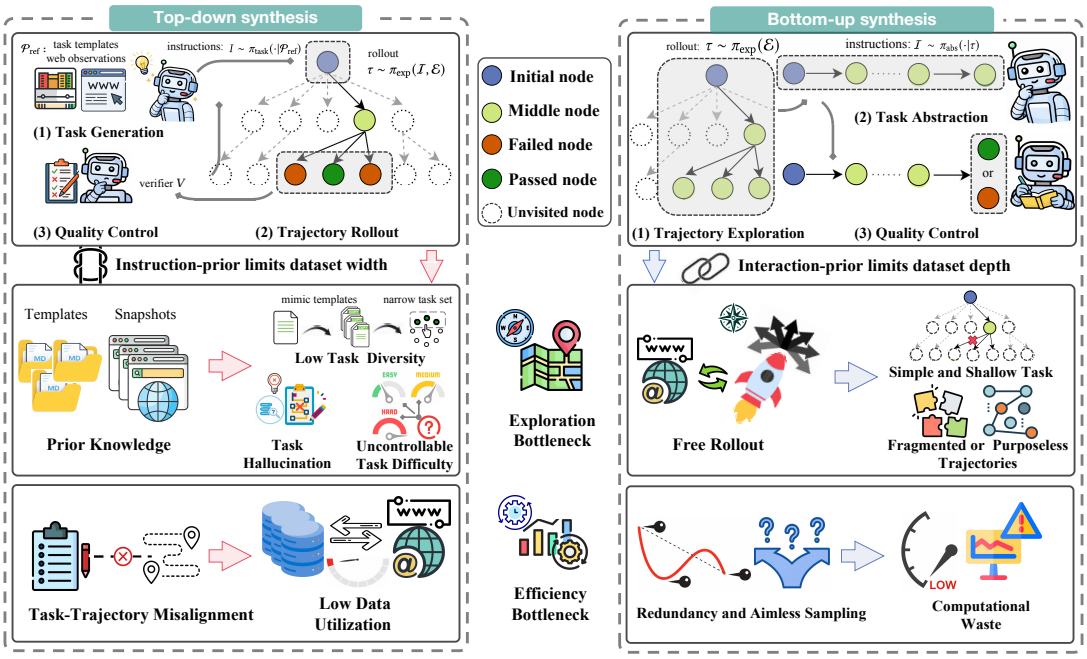

Figure 3: Challenges in Data Synthesis approaches and their current challenges. (a) Top-down. (b)Bottom-up.

description to be dynamically refined during the execution process, but this can lead to overly complex tasks or create a logic disconnect, where the final task no longer matches the agent's early actions. Furthermore, executing tasks inevitably introduces trial-and-error noise (He et al., 2025b), such as redundant clicks or loops, which, without mechanisms to transform these errors into reflection signals, teach the model inefficient planning instead of efficient decision-making.

5.2.2 Bottom-up: Redundant Sampling

In bottom-up approaches, because agents typically start each exploration episode from the same homepage without a memory of previous runs, the system suffers from extreme redundancy. Agents frequently select the same prominent UI elements (e.g., clicking every item on a shop page) and produce nearly duplicate samples. Existing works attempt to increase diversity by deduplicating `<state, action>` pairs (Mai et al., 2025) or categorizing UI elements (Wang et al., 2025c) for action selection, but they fail to prevent agents from retracing the same paths across different episodes due to a lack of shared memory.

6 Checklist: Towards Realistic Web Agents

Our analysis shows that many studies claim improvements in robustness, reasoning, or scalability without specifying the tested bottlenecks. To make such claims more testable, we provide an executable checklist for web-agent research in Table 1.

6.1 Realistic Benchmarking: Unified Protocols and Robust Assessment.

Benchmarks must specify their target environments (e.g., static sandboxes, live websites, or simulators) to balance auditable control with realistic complexity. LLMs can serve as environment architects (Fan et al., 2026; Zhang et al., 2026b) for constructing such sandboxes, covering safety (Levy et al., 2024; Tur et al., 2025), bias (Wang et al., 2026b), dynamic layouts (Wang et al., 2025a), resource competition (Zhu et al., 2025a), and adversarial attacks (Liao et al., 2025; Kara et al., 2025; Boisvert et al., 2025). Beyond static metrics (Zhou et al., 2024), evaluations require process-based assessments (Chae et al., 2025b; Fan et al., 2026; Zhang et al., 2026a) and network monitoring (Pan et al., 2025) to verify intended behaviors and reduce reward hacking, alongside reporting statistically valid multi-run results or significance tests.

Furthermore, securing action space completeness (Qiu et al., 2026; Wang et al., 2026a) raises a fundamental choice: forcing agents to master existing human GUIs via end-to-end training (Qin et al., 2025; Wang et al., 2026a), or rearchitecting the web to expose machine-centric semantic primitives (Qiu et al., 2026; Lù et al., 2025).

Target RQ			Methodological Requirements and Guidelines	Ref.
Environment and Benchmarking (§3)				
RQ1	**RQ2**			
✓			Characterize whether the environment is static or live, and report contamination controls.	§3.1
✓			Report how environments are reset between runs, including state isolation and reset cost.	§3.1
	✓		Report evaluation metrics with confidence intervals or equivalent uncertainty estimates.	§3.2
	✓		Validate evaluation logic against semantic equivalents, redundant outputs, edge cases.	§3.2
	✓		Report human agreement, self-consistency, and adversarial robustness for LLM-as-judge.	§3.2
	✓		Report evidence against reward hacking, such as redundant-action or invalid-call rates.	§3.2
	✓		Report train-test distribution differences, including task templates, websites, and domains.	§3.2
Agent Frameworks (§4)				
RQ1	**RQ2**	**RQ3**		
✓		✓	Clearly define the observation and action spaces, and report unsupported operations.	§4.1 & §4.3
✓			Evaluate perception robustness under dynamic UI states, pop-ups, and layout shifts.	§4.1
	✓		Report whether memory modules include explicit pruning or consolidation mechanisms.	§4.2
	✓		Evaluate memory and reflection mechanisms on unseen websites or cross-domain interfaces.	§4.2
	✓		Exclude oracle states, ground-truth actions, and validator outputs from memory or reflection.	§4.2
	✓		Report whether reflection is based on internal judgments or external tools.	§4.2
	✓		Compare search or reflection methods with budget-matched strong baselines.	§4.2
Data Synthesis and Training (§5)				
RQ1	**RQ2**			
✓			Report structural quality of synthesized data, including diversity, redundancy, and executability.	§5.1
	✓		Report the prior knowledge, task templates, or validators used for data generation.	§5.2
	✓		When training, report detailed settings, including data scale, template sources, annotation cost, and computational cost required to achieve the reported performance.	§5.2

Table 1: Checklist for rigorous and deployable web-agent research.

6.2 Architectural Reform: Grounded Planning and Sustainable Memory.

Architectural claims must isolate specific improvements to prevent confounds among perception, planning, memory, and execution. Open-ended tasks require hierarchical cognitive architectures (Ouyang et al., 2025; Liu et al., 2025e; Hu et al., 2025d) that decompose raw traces into reusable components like planning logic and user preferences (Zhang et al., 2025d). This fine-grained memory (Ouyang et al., 2025) enables skill recombination for unseen tasks.

Memory lifecycles must be explicit, utilizing reasoning-driven maintenance (adaptive decay, state pruning, and consolidating transient observations (Hu et al., 2025a) into long-term knowledge (Liu et al., 2025a; Hu et al., 2025d)) to reduce stale contexts. Finally, reflection and planning must be evaluated under matched budgets without oracle states or validators to prevent leakage and compute confounds, clarifying true architectural robustness.

6.3 Closing the Loop: From Static Imitation to Self-Evolution.

Rather than static behavioral imitation, the ultimate vision for web agents is autonomous self-evolution (Shao et al., 2025). Training recipes should always report which prior knowledge, templates, validators, and websites are used to construct data, and whether the generated trajectories are executable and diverse, enabling contamination and data-quality audits. To achieve true generalization, agent must decouple site-specific layout knowledge from the transferable "plan–act–reflect" reasoning loop (Tian et al., 2025), retrieving specific site attributes dynamically at test time. A preliminary investigation (Appendix A) indicates that injecting abstract planning capabilities from unseen distributions enhances target performance, paving a possible way for self-evolving agents (Zhao et al., 2025; Shao et al., 2025).

7 Conclusion

This paper critically examines the gap between benchmark progress and deployable web agents, identifying bottlenecks in evaluation, agent architecture, and data synthesis. While notable progress has been made, web agents remain fragile in the dynamic reality of the open web. Meaningful advancement demands a fundamental shift toward true open-world robustness, empowering web agents to self-evolve in complex, real-world environments.

Limitations

This paper provides a critical survey and diagnostic analysis of recent web-agent research, but it does not aim to exhaustively evaluate all existing agents, benchmarks, or training pipelines. Some observations may also change as web environments, benchmark protocols, and agent architectures continue to evolve.

References

Sagnik Anupam, Davis Brown, Shuo Li, Eric Wong, Hamed Hassani, and Osbert Bastani. 2025. Browserarena: Evaluating LLM agents on real-world web navigation tasks. *CoRR*, abs/2510.02418.

Ruhana Azam, Aditya Vempaty, and Ashish Jagmohan. 2025. Reflection-based memory for web navigation agents. *CoRR*, abs/2506.02158.

Léo Boisvert, Mihir Bansal, Chandra Kiran Reddy Evuru, Gabriel Huang, Abhay Puri, Avinandan Bose, Maryam Fazel, Quentin Cappart, Jason Stanley, Alexandre Lacoste, Alexandre Drouin, and Krishnamurthy Dvijotham. 2025. Doomarena: A framework for testing AI agents against evolving security threats. *CoRR*, abs/2504.14064.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier de Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. 2024. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. *CoRR*, abs/2407.05291.

Hyungjoo Chae, Namyoung Kim, Kai Tzu-iunn Ong, Minju Gwak, Gwanwoo Song, Jihoon Kim, Sunghwan Kim, Dongha Lee, and Jinyoung Yeo. 2025a. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Hyungjoo Chae, Sunghwan Kim, Junhee Cho, Seungone Kim, Seungjun Moon, Gyeom Hwangbo, Dongha Lim, Minjin Kim, Yeonjun Hwang, Minju Gwak, Dongwook Choi, Minseok Kang, Gwanhoon Im, ByeongUng Cho, Hyojun Kim, Jun Hee Han, Taeyoon Kwon, Minju Kim, Beong-woo Kwak, and 2 others. 2025b. Web-shepherd: Advancing prms for reinforcing web agents. volume abs/2505.15277.

Gongwei Chen, Xurui Zhou, Rui Shao, Yibo Lyu, Kaiwen Zhou, Shuai Wang, Wentao Li, Yinchuan Li, Zhongang Qi, and Liqiang Nie. 2025a. Less is more: Empowering GUI agent with context-aware simplification. volume abs/2507.03730.

Sen Chen, Tong Zhao, Yi Bin, Fei Ma, Wenqi Shao, and Zheng Wang. 2025b. D-GARA: A dynamic benchmarking framework for GUI agent robustness in real-world anomalies. *CoRR*, abs/2511.16590.

Zhaorun Chen, Zhuokai Zhao, Kai Zhang, Bo Liu, Qi Qi, Yifan Wu, Tarun Kalluri, Sara Cao, Yuanhao Xiong, Haibo Tong, Huaxiu Yao, Hengduo Li, Jiacheng Zhu, Xian Li, Dawn Song, Bo Li, Jason Weston, and Dat Huynh. 2025c. Scaling agent learning via experience synthesis. *CoRR*, abs/2511.03773.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. 2024. Seeclick: Harnessing GUI grounding for advanced visual GUI agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9313–9332. Association for Computational Linguistics.

Thibault Le Sellier de Chezelles, Maxime Gasse, Alexandre Lacoste, Massimo Caccia, Alexandre Drouin, 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, Graham Neubig, Quentin Cappart, Russ Salakhutdinov, and Nicolas Chapados. 2025. The browsergym ecosystem for web agent research. *Trans. Mach. Learn. Res.*, 2025.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samual Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Mahir Labib Dihan, Tanzima Hashem, Mohammed Eunus Ali, and Md. Rizwan Parvez. 2025. Weboperator: Action-aware tree search for autonomous agents in web environment. *CoRR*, abs/2512.12692.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *CoRR*, abs/2408.06361.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vázquez, Nicolas Chapados, and Alexandre Lacoste. 2024. Workarena: How capable are web agents at solving common knowledge work tasks?

Lutfi Eren Erdogan, Nicholas Lee, Sehoon Kim, Suhong Moon, Hiroki Furuta, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. 2025. Plan-and-act: Improving planning of agents for long-horizon tasks. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.

Sicheng Fan, Qingyun Shi, Shengze Xu, Shengbo Cai, Tieyong Zeng, LI LING, Yanyi Shang, and Dehan Kong. 2026. Webfactory: Automated compression of

foundational language intelligence into grounded web agents. In *The Fourteenth International Conference on Learning Representations*.

Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. 2025. Webevolver: Enhancing web agent self-improvement with coevolving world model. *CoRR*, abs/2504.21024.

Apurva Gandhi and Graham Neubig. 2025. Go-browse: Training web agents with structured exploration. *CoRR*, abs/2506.03533.

Difei Gao, Lei Ji, Zechen Bai, Mingyu Ouyang, Peiran Li, Dongxing Mao, Qinchen Wu, Weichen Zhang, Peiyi Wang, Xiangwu Guo, Hengxu Wang, Luowei Zhou, and Mike Zheng Shou. 2023. ASSISTGUI: task-oriented desktop graphical user interface automation. *CoRR*, abs/2312.13108.

Yifei Gao, Junhong Ye, Jiaqi Wang, and Jitao Sang. 2025. Websynthesis: World-model-guided MCTS for efficient webui-trajectory synthesis. *CoRR*, abs/2507.04370.

Divyansh Garg, Shaun VanWeelden, Diego Caples, Andis Draguns, Nikil Ravi, Pranav Putta, Naman Garg, Tomas Abraham, Michael Lara, Federico Lopez, James Liu, Atharva Gundawar, Prannay Hebbar, Youngchul Joo, Jindong Gu, Charles London, Christian A. Schröder de Witt, and Sumeet Motwani. 2025. REAL: benchmarking autonomous agents on deterministic simulations of real websites. *CoRR*, abs/2504.11543.

Gonzalo Gonzalez-Pumariega, Vincent Tu, Chih-Lun Lee, Jiachen Yang, Ang Li, and Xin Eric Wang. 2025. The unreasonable effectiveness of scaling agents for computer use. *CoRR*, abs/2510.02250.

Yu Gu, Kai Zhang, Yuting Ning, Boyuan Zheng, Boyu Gou, Tianci Xue, Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi, Huan Sun, and Yu Su. 2025. Is your LLM secretly a world model of the internet? model-based planning for web agents. *Trans. Mach. Learn. Res.*, 2025.

Izzeddin Gur, Ofir Nachum, Yingjie Miao, Mustafa Safdari, Austin V. Huang, Aakanksha Chowdhery, Sharan Narang, Noah Fiedel, and Aleksandra Faust. 2023. Understanding HTML with large language models. EMNLP 2023:2803–2821.

Guanzhong He, Zhen Yang, Jinxin Liu, Bin Xu, Lei Hou, and Juanzi Li. 2025a. Webseer: Training deeper search agents through reinforcement learning with self-reflection. *CoRR*, abs/2510.18798.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024. Webvoyager: Building an end-to-end web agent with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 6864–6890. Association for Computational Linguistics.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Hongming Zhang, Tianqing Fang, Zhenzhong Lan, and Dong Yu. 2025b. Openwebvoyager: Building multimodal web agents via iterative real-world exploration, feedback and optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 27545–27564. Association for Computational Linguistics.

Ziyuan He, Yuxuan Wang, Jiaqi Li, Kexin Liang, and Muhan Zhang. 2025c. Loogle v2: Are llms ready for real world long dependency challenges? *arXiv preprint arXiv:2510.22548*.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025a. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 32779–32798. Association for Computational Linguistics.

Minda Hu, Tianqing Fang, Jianshu Zhang, Junyu Ma, Zhisong Zhang, Jingyan Zhou, Hongming Zhang, Haitao Mi, Dong Yu, and Irwin King. 2025b. Webcot: Enhancing web agent reasoning by reconstructing chain-of-thought in reflection, branching, and rollback. *CoRR*, abs/2505.20013.

Yuanzhe Hu, Yu Wang, and Julian J. McAuley. 2025c. Evaluating memory in LLM agents via incremental multi-turn interactions. *CoRR*, abs/2507.05257.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2025d. Memory in the age of AI agents. *CoRR*, abs/2512.13564.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024a. Large language models cannot self-correct reasoning yet.

Tenghao Huang, Kinjal Basu, Ibrahim Abdelaziz, Pavan Kapanipathi, Jonathan May, and Muhao Chen. 2025. R2D2: remembering, replaying and dynamic decision making with a reflective agentic memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30318–30330. Association for Computational Linguistics.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024b. Understanding

the planning of LLM agents: A survey. *CoRR*, abs/2402.02716.

Zixia Jia, Jiaqi Li, Yipeng Kang, Yuxuan Wang, Tong Wu, Quansen Wang, Xiaobo Wang, Shuyi Zhang, Junzhe Shen, Qing Li, and 1 others. 2026. The ai hippocampus: How far are we from human memory? *arXiv preprint arXiv:2601.09113*.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R. Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Su Kara, Fazle E. Faisal, and Suman Nath. 2025. WAREX: web agent reliability evaluation on existing benchmarks. *CoRR*, abs/2510.03285.

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

Imene Kerboua, Sahar Omidi Shayegan, Megh Thakkar, Xing Han Lù, Massimo Caccia, Véronique Eglin, Alexandre Aussem, Jérémy Espinas, and Alexandre Lacoste. 2025b. Lineretriever: Planning-aware observation reduction for web agents. *CoRR*, abs/2507.00210.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Dogyu Ko, Chanyoung Yeo, Daeho Kim, Jaeho Kim, and Hyoseok Hwang. 2026. GAIA: generating task instruction aware simulation grounded in real contexts using vision-language models. *IEEE Robotics Autom. Lett.*, 11(1):626–633.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. 2024. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 881–905. Association for Computational Linguistics.

Hanyu Lai, Xiao Liu, Iat Long Iong, Shuntian Yao, Yuxuan Chen, Pengbo Shen, Hao Yu, Hanchen Zhang, Xiaohan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autowebglm: A large language model-based web navigating agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 5295–5306. ACM.

Dongjun Lee, Juyong Lee, Kyuyoung Kim, Jihoon Tack, Jinwoo Shin, Yee Whye Teh, and Kimin Lee. 2025. Learning to contextualize web pages for enhanced decision making by LLM agents.

Ido Levy, Ben Wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. 2024. St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. volume abs/2410.06703.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1000–1008. Association for Computational Linguistics.

Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, Xixi Wu, and 36 others. 2025a. Tongyi deepresearch technical report. *CoRR*, abs/2510.24701.

Jiaqi Li, Xinyi Dong, Yang Liu, Zhizhuo Yang, Quansen Wang, Xiaobo Wang, Song-Chun Zhu, Zixia Jia, and Zilong Zheng. 2025b. Reflectevo: Improving meta introspection of small llms by learning self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16948–16966.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Jiaqi Li, Xiaobo Wang, Wentao Ding, Zihao Wang, Yipeng Kang, Zixia Jia, and Zilong Zheng. 2024b. Ram: Towards an ever-improving memory system by learning from communications. *arXiv preprint arXiv: 2404.12045*.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025c. Websailor: Navigating super-human reasoning for web agent. *CoRR*, abs/2507.02592.

Zeyi Liao, Lingbo Mo, Chejian Xu, Mintong Kang, Jiawei Zhang, Chaowei Xiao, Yuan Tian, Bo Li, and Huan Sun. 2025. Eia: Environmental injection attack on generalist web agents for privacy leakage. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Evan Zheran Liu, Kelvin Guu, Panupong Pasupat, Tianlin Shi, and Percy Liang. 2018. Reinforcement learning on web interfaces using workflow-guided exploration. In *6th International Conference on Learning*

Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Genglin Liu, Shijie Geng, Sha Li, Hejie Cui, Sarah Zhang, Xin Liu, and Tianyi Liu. 2025a. Webcoach: Self-evolving web agents with cross-session memory guidance. *CoRR*, abs/2511.12997.

Junwei Liu, Kaixin Wang, Yixuan Chen, Xin Peng, Zhenpeng Chen, Lingming Zhang, and Yiling Lou. 2024. Large language model-based agents for software engineering: A survey. *CoRR*, abs/2409.02977.

Tao Liu, Chongyu Wang, Rongjie Li, Yingchen Yu, Xuming He, and Song Bai. 2025b. Gui-rise: Structured reasoning and history summarization for GUI navigation. *CoRR*, abs/2510.27210.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Xixuan Song, Yifan Xu, Shudan Zhang, Hanyu Lai, Jiadai Sun, Xinyue Yang, Yu Yang, Zehan Qi, Shuntian Yao, Xueqiao Sun, Siyi Cheng, Qinkai Zheng, Hao Yu, Hanchen Zhang, Wenyi Hong, and 9 others. 2025c. Visualagentbench: Towards large multimodal models as visual foundation agents.

Yang Liu, Jiaqi Li, and Zilong Zheng. 2025d. Rulereasoner: Reinforced rule-based reasoning via domain-aware dynamic sampling. *arXiv preprint arXiv:2506.08672*.

Zhipeng Liu, Xuefeng Bai, Kehai Chen, Xinyang Chen, Xiucheng Li, Yang Xiang, Jin Liu, Hong-Dong Li, Yaowei Wang, Liqiang Nie, and Min Zhang. 2025e. A survey on the feedback mechanism of llm-based AI agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 10582–10592. ijcai.org.

Robert Lo, Abishek Sridhar, Frank F. Xu, Hao Zhu, and Shuyan Zhou. 2023. Hierarchical prompting assists large language model on web navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pages 10217–10244. Association for Computational Linguistics.

Xing Han Lù, Gaurav Kamath, Marius Mosbach, and Siva Reddy. 2025. Build the web for agents, not agents for the web. *CoRR*, abs/2506.10953.

Xing Han Lù, Zdenek Kasner, and Siva Reddy. 2024. Weblinx: Real-world website navigation with multi-turn dialogue.

Yuxuan Lu, Jing Huang, Hui Liu, Jiri Gesi, Yan Han, Shihan Fu, Tianqi Zheng, and Dakuo Wang. 2025. WEBSERV: A browser-server environment for efficient training of reinforcement learning-based web agents at scale. *CoRR*, abs/2510.16252.

Yaxin Luo, Zhaoyi Li, Jiacheng Liu, Jiacheng Cui, Xiaohan Zhao, and Zhiqiang Shen. 2025. Open captcha-world: A comprehensive web-based platform for testing and benchmarking multimodal LLM agents. *CoRR*, abs/2505.24878.

Yougang Lyu, Xiaoyu Zhang, Lingyong Yan, Maarten de Rijke, Zhaochun Ren, and Xiuying Chen. 2025. Deepshop: A benchmark for deep research shopping agents. *CoRR*, abs/2506.02839.

Kaixin Ma, Hongming Zhang, Hongwei Wang, Xiaoman Pan, and Dong Yu. 2023. LASER: LLM agent with state-space exploration for web navigation. *CoRR*, abs/2309.08172.

Shinji Mai, Yunpeng Zhai, Ziqian Chen, Cheng Chen, Anni Zou, Shuchang Tao, Zhaoyang Liu, and Bolin Ding. 2025. Cues: A curiosity-driven and environment-grounded synthesis framework for agentic RL. *CoRR*, abs/2512.01311.

Yansheng Mao, Yufei Xu, Jiaqi Li, Fanxu Meng, Haotong Yang, Zilong Zheng, Xiyuan Wang, and Muhan Zhang. 2025. Lift: Improving long context understanding of large language models through long input fine-tuning. *arXiv preprint arXiv:2502.14644*.

Zairah Mustahsan, Abel Lim, Megna Anand, Saahil Jain, and Bryan McCann. 2025. Stochasticity in agentic evaluations: Quantifying inconsistency with intra-class correlation. *CoRR*, abs/2512.06710.

Liangbo Ning, Ziran Liang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Wenqi Fan, Xiaoyong Wei, Shanru Lin, Hui Liu, Philip S. Yu, and Qing Li. 2025. A survey of webagents: Towards next-generation AI agents for web automation with large foundation models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 6140–6150. ACM.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. 2025. Reasoningbank: Scaling agent self-evolving with reasoning memory. *CoRR*, abs/2509.25140.

Yichao Pan, Yujia Qin, Yutong Zhu, Weizhi Wang, Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu, and Yankai Lin. 2025. Webarena-verified: Improving the reliability of web agent evaluation. In *International Conference on Learning Representations (ICLR)*.

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

Shrey Pandit, Xuan-Phi Nguyen, Yifei Ming, Austin Xu, Jiayu Wang, Caiming Xiong, and Shafiq Joty. 2025. Synthesizing agentic data for web agents with progressive difficulty enhancement mechanisms. *CoRR*, abs/2510.13913.

Ralph Peeters, Aaron Steiner, Luca Schwarz, Julian Yuya Caspary, and Christian Bizer. 2025. Webmall - A multi-shop benchmark for evaluating web agents. *CoRR*, abs/2508.13024.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, Wanjun Zhong, Kuanye Li, Jiale Yang, Yu Miao, Woyu Lin, Longxiang Liu, Xu Jiang, Qianli Ma, Jingyu Li, and 16 others. 2025. UI-TARS: pioneering automated GUI interaction with native agents. *CoRR*, abs/2501.12326.

Zhi Qiu, Jiazheng Sun, Chenxiao Xia, Jun Zheng, and Xin Peng. 2026. Ci4a: Semantic component interfaces for agents empowering web automation. *arXiv preprint arXiv:2601.14790*.

JV Roig. 2025. Towards a standard, enterprise-relevant agentic AI benchmark: Lessons from 5.5 billion tokens' worth of agentic AI evaluations. *CoRR*, abs/2511.08042.

Zhihong Shao, Yuxiang Luo, Chengda Lu, Z. Z. Ren, Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, and Xiaokang Zhang. 2025. Deepseekmath-v2: Towards self-verifiable mathematical reasoning. *CoRR*, abs/2511.22570.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Junhong Shen, Atishay Jain, Zedian Xiao, Ishan Amlekar, Mouad Hadji, Aaron Podolny, and Ameet Talwalkar. 2024. Scribeagent: Towards specialized web agents using production-scale workflow data. *CoRR*, abs/2411.15004.

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. 2017. World of bits: An open-domain platform for web-based agents. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3135–3144. PMLR.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Paloma Sodhi, SRK Branavan, Yoav Artzi, and Ryan McDonald. 2023. Step: Stacked llm policies for web actions. *arXiv preprint arXiv:2310.03720*.

Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*, pages 107–120. ACM.

Hongjin Su, Ruoxi Sun, Jinsung Yoon, Pengcheng Yin, Tao Yu, and Sercan Ö. Arik. 2025. Learn-by-interact: A data-centric framework for self-adaptive agents in realistic environments.

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. 2025a. Os-genesis: Automating GUI agent trajectory construction via reverse task synthesis. pages 5555–5579.

Zihao Sun, Meng Fang, and Ling Chen. 2025b. Webarxiv: Evaluating multimodal agents on time-invariant arxiv tasks. *CoRR*, abs/2507.00938.

Xiaojuan Tang, Jiaqi Li, Yitao Liang, Muhan Zhang, and Zilong Zheng. 2024. Mars: Situated inductive reasoning in an open-world environment. In *38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks*.

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. 2023. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv: 2305.14825*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Shizuo Tian, Hao Wen, Yuxuan Chen, Jiacheng Liu, Shanhui Zhao, Guohong Liu, Ju Ren, Yunxin Liu, and Yuanchun Li. 2025. Agentprog: Empowering long-horizon GUI agents with program-guided context management. *CoRR*, abs/2512.10371.

Ada Defne Tur, Nicholas Meade, Xing Han Lù, Alejandra Zambrano, Arkil Patel, Esin Durmus, Spandana Gella, Karolina Stanczak, and Siva Reddy. 2025. Safearena: Evaluating the safety of autonomous web agents. 267.

Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. 2024. OMNIPARSER: A unified framework for text spotting, key information extraction and table recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*

2024, Seattle, WA, USA, June 16-22, 2024, pages 15641–15653. IEEE.

Dakuo Wang, Ting-Yao Hsu, Yuxuan Lu, Hansu Gu, Limeng Cui, Yaochen Xie, William Headean, Bingsheng Yao, Akash Veeragouni, Jiapeng Liu, Sreyashi Nag, and Jessie Wang. 2025a. Agenta/b: Automated and scalable web a/btesting with interactive LLM agents. *CoRR*, abs/2504.09723.

Haoyu Wang, Tao Li, Zhiwei Deng, Dan Roth, and Yang Li. 2024. Devil's advocate: Anticipatory reflection for LLM agents. EMNLP 2024:966–978.

Xuan Wang, Siyuan Su, Quantong Fu, Yongxiang Hu, and Yangfan Zhou. 2026a. Swipegen: Bridging the execution gap in gui agents via human-like swipe synthesis. In *unknown*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Yaoxiang Wang, Zhiyong Wu, Junfeng Yao, and Jinsong Su. 2025b. TDAG: A multi-agent framework based on dynamic task decomposition and agent generation. *Neural Networks*, 185:107200.

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. 2025c. Adapting web agents with synthetic supervision. *CoRR*, abs/2511.06101.

Zongwei Wang, Bincheng Gu, Hongyu Yu, Junliang Yu, Tao He, Jiayin Feng, Chenghua Lin, and Min Gao. 2026b. When agents see humans as the outgroup: Belief-dependent bias in llm-powered agents. *arXiv e-prints*, pages arXiv–2601.

Zora Zhiruo Wang, Apurva Gandhi, Graham Neubig, and Daniel Fried. 2025d. Inducing programmatic skills for agentic tasks. *CoRR*, abs/2504.06821.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. 2025e. Agent workflow memory. 267.

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.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. Webdancer: Towards autonomous information seeking agency. *CoRR*, abs/2505.22648.

Zhiheng Xi, Yiwen Ding, Wenxiang Chen, Boyang Hong, Honglin Guo, Junzhe Wang, Dingwen Yang, Chenyang Liao, Xin Guo, Wei He, Songyang Gao, Lu Chen, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024. Agentgym: Evolving large language model-based agents across diverse environments. *CoRR*, abs/2406.04151.

Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, Wei He, Yiwen Ding, Guanyu Li, Zehui Chen, Zhengyin Du, Xuesong Yao, Yufei Xu, Jiecao Chen, Tao Gui, and 4 others. 2025. Agentgym-rl: Training LLM agents for long-horizon decision making through multi-turn reinforcement learning. *CoRR*, abs/2509.08755.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. 2024. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Tianci Xue, Weijian Qi, Tianneng Shi, Chan Hee Song, Boyu Gou, Dawn Song, Huan Sun, and Yu Su. 2025. An illusion of progress? assessing the current state of web agents. *CoRR*, abs/2504.01382.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *CoRR*, abs/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 40 others. 2025a. Qwen3 technical report. *CoRR*, abs/2505.09388.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *CoRR*, abs/2310.11441.

Jingbo Yang, Bairu Hou, Wei Wei, Shiyu Chang, and Yujia Bao. 2025b. Webdart: Dynamic decomposition and re-planning for complex web tasks. *CoRR*, abs/2510.06587.

Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, and Huzefa Rangwala. 2025c. Agentoccam: A simple yet strong baseline for llm-based web agents.

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 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a.

Tree of thoughts: Deliberate problem solving with large language models.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models.

Rui Ye, Zhongwang Zhang, Kuan Li, Huifeng Yin, Zhengwei Tao, Yida Zhao, Liangcai Su, Liwen Zhang, Zile Qiao, Xinyu Wang, Pengjun Xie, Fei Huang, Siheng Chen, Jingren Zhou, and Yong Jiang. 2025. Agentfold: Long-horizon web agents with proactive context management. *CoRR*, abs/2510.24699.

Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2025a. Memagent: Reshaping long-context LLM with multi-conv rl-based memory agent. *CoRR*, abs/2507.02259.

Simon Yu, Gang Li, Weiyan Shi, and Peng Qi. 2025b. Polyskill: Learning generalizable skills through polymorphic abstraction. *CoRR*, abs/2510.15863.

Wenwen Yu, Zhibo Yang, Jianqiang Wan, Sibo Song, Jun Tang, Wenqing Cheng, Yuliang Liu, and Xiang Bai. 2025c. Omniparser V2: structured-points-of-thought for unified visual text parsing and its generality to multimodal large language models. *CoRR*, abs/2502.16161.

Xiao Yu, Baolin Peng, Vineeth Vajipey, Hao Cheng, Michel Galley, Jianfeng Gao, and Zhou Yu. 2025d. Exact: Teaching AI agents to explore with reflective-mcts and exploratory learning.

Qianhao Yuan, Jie Lou, Zichao Li, Jiawei Chen, Yaojie Lu, Hongyu Lin, Le Sun, Debing Zhang, and Xianpei Han. 2025. Memsearcher: Training llms to reason, search and manage memory via end-to-end reinforcement learning. *CoRR*, abs/2511.02805.

Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, Junhao Wang, He Zhu, Wangchunshu Zhou, and Shuicheng Yan. 2025a. Memevolve: Meta-evolution of agent memory systems. *CoRR*, abs/2512.18746.

Jenny Zhang, Shengran Hu, Cong Lu, Robert T. Lange, and Jeff Clune. 2025b. Darwin godel machine: Open-ended evolution of self-improving agents. *CoRR*, abs/2505.22954.

Jiayuan Zhang, Kaiquan Chen, Zhihao Lu, Enshen Zhou, Qian Yu, and Jing Zhang. 2025c. Prune4web: DOM tree pruning programming for web agent. *CoRR*, abs/2511.21398.

Wentao Zhang, Ce Cui, Yilei Zhao, Rui Hu, Yang Liu, Yahui Zhou, and Bo An. 2025d. Agentorchestra: A hierarchical multi-agent framework for general-purpose task solving. *CoRR*, abs/2506.12508.

Yao Zhang, Zijian Ma, Yunpu Ma, Zhen Han, Yu Wu, and Volker Tresp. 2025e. Webpilot: A versatile and autonomous multi-agent system for web task execution with strategic exploration. pages 23378–23386.

Yao Zhang, Shijie Tang, Zeyu Li, Zhen Han, and Volker Tresp. 2026a. Webarbiter: A principle-guided reasoning process reward model for web agents. *arXiv preprint arXiv:2601.21872*.

Zhuosheng Zhang and Aston Zhang. 2024. You only look at screens: Multimodal chain-of-action agents. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 3132–3149. Association for Computational Linguistics.

Ziyun Zhang, Zezhou Wang, Xiaoyi Zhang, Zongyu Guo, Jiahao Li, Bin Li, and Yan Lu. 2026b. Infiniteweb: Scalable web environment synthesis for gui agent training. *arXiv preprint arXiv:2601.04126*.

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. *CoRR*, abs/2505.03335.

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. 2025. Skillweaver: Web agents can self-improve by discovering and honing skills. *CoRR*, abs/2504.07079.

Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. 2024a. Gpt-4v(ision) is a generalist web agent, if grounded. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Longtao Zheng, Rundong Wang, Xinrun Wang, and Bo An. 2024b. Synapse: Trajectory-as-exemplar prompting with memory for computer control.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024c. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372.

Jiamu Zhou, Jihong Wang, Weiming Zhang, Weiwen Liu, Zhuosheng Zhang, Xingyu Lou, Weinan Zhang, Huarong Deng, and Jun Wang. 2026. Colorbrowseragent: An intelligent gui agent for complex long-horizon web automation. *arXiv preprint arXiv:2601.07262*.

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. 2024. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yifei Zhou, Qianlan Yang, Kaixiang Lin, Min Bai, Xiong Zhou, Yu-Xiong Wang, Sergey Levine, and Li Erran Li. 2025. Proposer-agent-evaluator (PAE): autonomous skill discovery for foundation model internet agents. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and Jiaxuan You. 2025a. Multiagentbench : Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 8580–8622. Association for Computational Linguistics.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Jasjeet S. Sekhon, Jacob Steinhardt, and 6 others. 2025b. Establishing best practices for building rigorous agentic benchmarks. *CoRR*, abs/2507.02825.

A Experimental Setup

To explore the feasibility of isolating the sources of fine-tuning gains, we conducted a tentative experiment to determine whether improvements stem primarily from target-site knowledge, explicit planning supervision, or their combination. Under controlled training settings and budgets, we compare two factors: (1) **site-specific knowledge**, such as DOM structures and website workflows; and (2) **general planning capability**, which decomposes instructions into intermediate subgoals before acting.

Data Sources We use three public data sources. *Go-Browse* (Gandhi and Neubig, 2025) and *OS-Genesis* (Sun et al., 2025a) are treated as **unseen task distributions**; *AgentOccam* (Yang et al., 2025c), sampled from WebArena while excluding the *lite* evaluation subset, is treated as the **target task distribution**.

For each source family, the **plan** variant is derived from the same successful-trajectory pool as its non-plan counterpart. We use `GPT-4o-mini` (OpenAI, 2024) to infer a high-level plan from each execution log and reformat the sample into `<Plan, Reasoning, Action>`, keeping filtering, formatting, and training pipeline matched within each source family.

Training Budget and Implementation All models are fine-tuned from `Qwen3-8B-Base` (Yang et al., 2025a) with `LLaMA-Factory` (Zheng et al., 2024c) on 8× NVIDIA A800 (80GB) GPUs, using a learning rate of 1.0×10^{-5} and cosine decay. Epoch counts and preprocessing recipes follow the corresponding reference setups. The released sources differ in scale: *Go-Browse* has 9.5K successful trajectories, *OS-Genesis* has 7.4K web trajectories, and *AgentOccam* has 0.6K non-*lite* WebArena target-site trajectories.

We also evaluate a two-stage training setting: the model is first fine-tuned on *AgentOccam* for target-site knowledge, and then continued on *OS-Genesis(Plan)* for explicit planning supervision from an unseen trajectory source. The two-stage setting has a **comparable overall budget** to the 7K–10K-scale baselines because it only uses the same 0.6K *AgentOccam* target-site budget in the first stage.

Evaluation Protocol We align action spaces between training and evaluation. *Go-Browse* variants use **BrowserGym (BG)** (de Chezelles et al., 2025); *OS-Genesis* and *AgentOccam* variants use **Visual-Agent-Bench (VAB)** (Liu et al., 2025c). Models are served with `vLLM`; we use a 32k-token context for VAB and the original 80,000-character setting for BG. We report **pass@3** and reset WebArena before each instance.

Method	Env.	pass@1	pass@3
Unseen Task Distribution			
Qwen3-8B Baseline	BG	14.55	22.42
+ Go-Browse	BG	14.55	16.97
+Go-Browse(Plan)	BG	16.97	17.58
Qwen3-8B Baseline	VAB	17.58	24.85
+ OS-Genesis	VAB	18.18	26.67
+ OS-Genesis(Plan)	VAB	11.52	18.79
Target Task Distribution			
+ AgentOccam (Target Knowledge)	VAB	20.60	30.91
+ OS-Genesis(Plan) (Capability Injection)	VAB	24.24	38.18

Note: **Env.** refers to evaluation framework where **BG**: BrowserGym (de Chezelles et al., 2025); **VAB**: Visual-Agent-Bench (Liu et al., 2025c).

Table 2: **Study on decoupling knowledge and capability.** Success rates (%) on WebArena-lite (Liu et al., 2025c) across 165 tasks over 5 sites.

Experimental Results First, blind scaling is insufficient. Direct fine-tuning on unseen tasks improves pass@1 but may degrades intrinsic generalization (significant drop in pass@3 (Go-Browse)), indicating that the model overfits to source-domain shortcuts rather than learning transferable reasoning. Furthermore, naively enforcing a Plan output without domain grounding may leads to hallucinated reasoning, causing performance drops(OS-Genesis(Plan)).

This pattern is also consistent with the template-overlap analysis in Appendix B: only 0.6K target-site trajectories from *AgentOccam* outperform much larger 7K–10K-scale unseen-source data, suggesting that same-template target-domain exposure can dominate nominal data scale.

Second, decoupling knowledge and capability are promising. The *Capability Injection* setting achieves superior results by combining target domain knowledge with general planning skills. This confirms our hypothesis: the agent requires domain knowledge to "ground" the abstract planning logic. Without knowledge, planning is hallucination; with knowledge, planning becomes a capability amplifier.

B Environment and Evaluation

B.1 Benchmark Analysis

Table 3 summarizes representative web-agent benchmarks along dimensions relevant to our analysis, including observation modality, scalability, evaluation protocol, and whether process rewards

Benchmark	Task Domain	Observation	Scalability	Evaluation	Process Reward
Simple Synthetic Environments					
WebShop (Yao et al., 2022)	Shopping	TXT	SL	EM	✗
MiniWoB++ (Liu et al., 2018)	Comprehensive	MM	SL	EM	✗
Static Sandbox Environments					
WebArXiv (Sun et al., 2025b)	Research	MM	SL	EM	✗
Open CaptchaWorld (Luo et al., 2025)	Safety	VIS	SL	EM	✓
REAL (Garg et al., 2025)	Comprehensive	MM	RR	EM, FM	✗
WebMall (Peeters et al., 2025)	Shopping	MM	RR	EM	✗
SafeArena (Tur et al., 2025)	Safety	MM	RR	EM, FM	✗
WorkArena++ (Boisvert et al., 2024)	Enterprise	MM	RR	EM, FM	✗
VisualWebArena (Koh et al., 2024)	Comprehensive	MM	RR	EM, FM	✗
WebArena (Zhou et al., 2024)	Comprehensive	TXT	RR	EM, FM	✗
Real-World Environments					
BrowserArena (Anupam et al., 2025)	Comprehensive	MM	LW	FM	✓
DeepShop (Lyu et al., 2025)	Shopping	MM	LW	EM, FM	✗
D-GARA (Chen et al., 2025b)	Safety	MM	LW	EM, FM	✗
WebCanvas (Pan et al., 2024)	Comprehensive	MM	LW	EM, FM	✓
WebVoyager (He et al., 2024)	Comprehensive	MM	LW	FM	✗

Table 3: Comparison of representative web agent benchmarks. Abbreviations are defined as follows: **Observation** (*TXT*: Text Only, *VIS*: Vision Only, *MM*: Multimodal); **Scalability** (*SL*: Stateless, *RR*: Reset Required, *LW*: Live Web); **Evaluation** (*EM*: Exact Match (Rule-based), *FM*: Fuzzy Match (LLM-as-Judge)).

are provided. These dimensions contextualize the risks discussed in Section 3.

B.2 Case Studies on Template Overlap

We further analyze WebArena-Lite (Liu et al., 2025c) as a concrete case for template overlap and evaluation reliability. Let $\mathcal{D}^{\text{test}}$ and $\mathcal{D}^{\text{train}}$ denote the test and training sets. Each task x is mapped to a canonical template $t(x)$ that removes instance-specific slot values such as entity names, dates, or prices while preserving the underlying procedure. Let $\mathcal{T}^{\text{test}}$ and $\mathcal{T}^{\text{train}}$ be the corresponding unique template sets.

Train-test Overlap		Reproducibility
Overlap Rate	Jaccard Similarity	Intra-template Var.
78.79%	0.5394	0.0629

Table 4: Quantitative analysis of evaluation reliability on WebArena-Lite.

The overlap rate reports the fraction of test tasks whose templates also appear in training, $\textsc{OverlapRate} = |\mathcal{C}|/|\mathcal{D}^{\text{test}}|$, where $\mathcal{C} = \{x \in \mathcal{D}^{\text{test}} : t(x) \in \mathcal{T}^{\text{train}}\}$. Jaccard similarity reports the overlap between unique training and test templates, $\textsc{Jaccard} = |\mathcal{T}^{\text{test}} \cap \mathcal{T}^{\text{train}}|/|\mathcal{T}^{\text{test}} \cup \mathcal{T}^{\text{train}}|$. These two views show that an instance-disjoint split can still preserve substantial procedural overlap. This connects to the training study in Appendix A, where 0.6K target-site trajectories outperform larger unseen-source datasets, suggesting that template-level exposure can substantially affect reported performance.

We also report intra-template variance, $\textsc{IntraTemplateVar} = |\mathcal{T}^{\text{test}}|^{-1} \sum_{t \in \mathcal{T}^{\text{test}}} \textsc{Var}_t$, where $\textsc{Var}_t$ is the variance of success scores among test tasks sharing template t. Success scores are derived from AgentOccam (Yang et al., 2025c) with GPT-4.1 trajectories. This metric captures whether nominally similar tasks produce stable outcomes, complementing prior evidence that agent scores can vary across repeated runs (Mustahsan et al., 2025). Together, these results motivate reporting both train-test template structure and statistical stability, rather than only aggregate success rates.

B.3 Case Studies on Environment and Evaluation

The following cases show benchmark and environment flaws that are difficult to see from aggregate scores alone, including strict answer matching, unreachable target states, and unrealistic environment behavior.

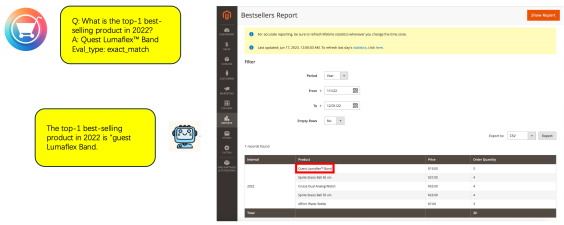

Figure 4: An unreasonable evaluation case in WebArena using strict string matching. The model correctly identified the answer but was penalized for missing the "TM" superscript, highlighting the fragility of exact match metrics.

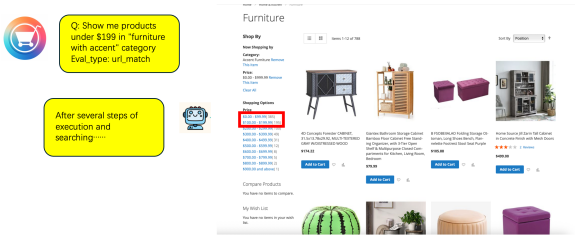

Figure 5: An impossible task case in WebArena. The instruction requires filtering for a $0–199 range, but the UI only permits disjoint ranges (0–99 or 100–199), making the target state unreachable via standard interaction.

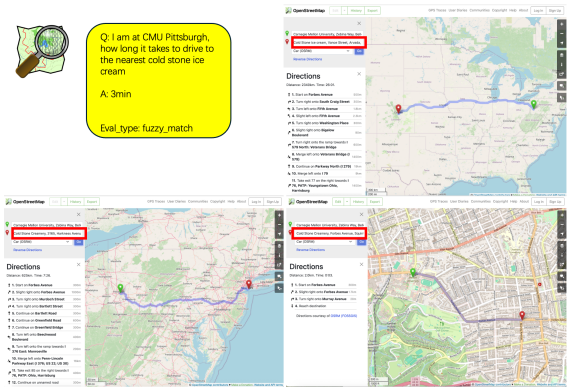

Figure 6: A robustness failure in WebArena's map environment. The search function fails to resolve common queries (e.g., 'Cold Stone Ice Cream') without explicit location qualifiers, unlike real-world map applications.

C Case Studies on Planning Hallucination

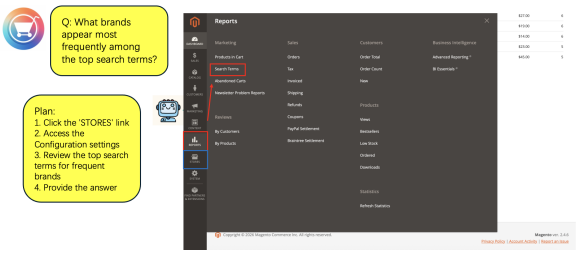

Figure 7: Planning hallucination example I. The agent incorrectly navigates to 'Stores' instead of 'Reports', fails to backtrack or re-plan, and ultimately hallucinates an irrelevant answer.

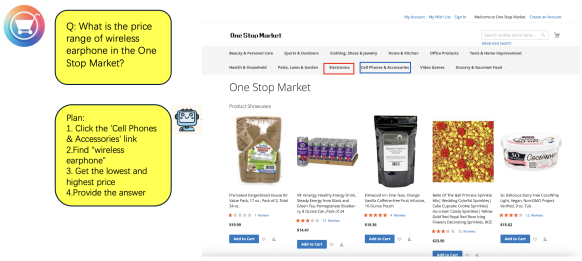

Figure 8: Planning hallucination example II. The agent selects an incorrect category ('Cell Phones') for a 'wireless headphones' query and forces a search in the wrong domain rather than correcting its navigation path.

D Case Studies on Memory Generalization

4. When the task asks you to draft an email. DO NOT send the email. Just draft it and provide the content in the last message
5. For questions ONLY about **Nintendo Switch game cards** please navigate back to the url of the product you end up choosing. This is because the benchmark evalautes using the page you are on for the Nintendo task. If you are own the wrong product page it will assume you selected that product.
6. When asked to "show me products under a price" the benchmark evaluates based on the url. so just ensure the price is under the price asked for in the url.
7. When asked about product type don't return the acutal product name but rather the type of product it is. Ex: If the most common products are "harry potter" and "twilight" the product type is "books".
8. **For SHOPPING/BROWSING tasks**: When asked to browse products in a particular category, navigate using the dropdown menus (not search) when possible. This may require hovering over nested dropdowns (e.g., hover over "Electronics" → hover over "Computers" → click "Laptops"). Use the hover tool to reveal these nested menus before clicking.
9. Task wording can contain typos or singular/plural mismatches. Treat singular terms like "order" or "transaction" as potentially plural. If multiple entries match the described criteria, summarize all of them, list each matching item with its key details, make the quantity explicit, and when the question asks for an amount/total, include the combined sum alongside the itemized numbers. Never assume "the most recent one" unless the instructions explicitly say so.

Figure 9: Specific knowledge injection in ColorBrowserAgent (Zhou et al., 2026). The memory includes site-specific advice (e.g., regarding Nintendo Switch game cards) that is unlikely to generalize to other domains, potentially consuming context window inefficiently.

E Ethics Considerations

This work is a critical survey and methodological analysis of web-agent research. We do not collect new user data, conduct human-subject experiments,

or deploy agents on real user accounts. Our analyses use publicly released papers, benchmark descriptions, task metadata, and released evaluation artifacts when available. We use these artifacts only for scholarly analysis, follow their original licenses and terms of use, and do not redistribute raw benchmark data or user-generated web content.

Although our work does not introduce a new deployed system, web agents can raise practical risks when interacting with real websites, including privacy leakage, unintended actions, unsafe tool use, and exploitation of weak evaluators. We therefore emphasize controlled environments, state isolation, transparent evaluation protocols, contamination auditing, and explicit reporting of safety and robustness limitations. AI assistants were used for language polishing; all claims, analyses, and final text were reviewed and revised by the authors.