

Avenir-Web¹: Human-Experience-Imitating Multimodal Web Agents with Mixture of Grounding Experts

Aiden Yiliu Li^{1,2} Xinyue Hao³ Shilong Liu² Mengdi Wang²

Abstract

Despite advances in Multimodal Large Language Models (MLLMs), autonomous web agents still struggle to reliably execute long-horizon tasks on complex and dynamic web interfaces. Existing agents often suffer from inaccurate element grounding, the absence of site-specific procedural knowledge, and unstable long-term task tracking and memory, particularly when operating in complex Document Object Model (DOM) structures. To address these limitations, we introduce AVENIR-WEB, a web agent that achieves a new open-source state-of-the-art (SOTA) on the ONLINE-MIND2WEB benchmark in real-world deployment. AVENIR-WEB leverages Mixture of Grounding Experts (MoGE), Experience-Imitation Planning (EIP) for incorporating procedural priors, and a Task-Tracking Checklist combined with Adaptive Memory to enable robust and seamless interaction across diverse UI paradigms. We evaluate AVENIR-WEB on ONLINE-MIND2WEB (Xue et al., 2025), a rigorous benchmark of live and user-centered web tasks. Our results demonstrate that AVENIR-WEB significantly surpasses prior open-source agents and attains performance parity with top-tier proprietary models, thereby establishing a new open-source state-of-the-art for reliable web agents on live websites.

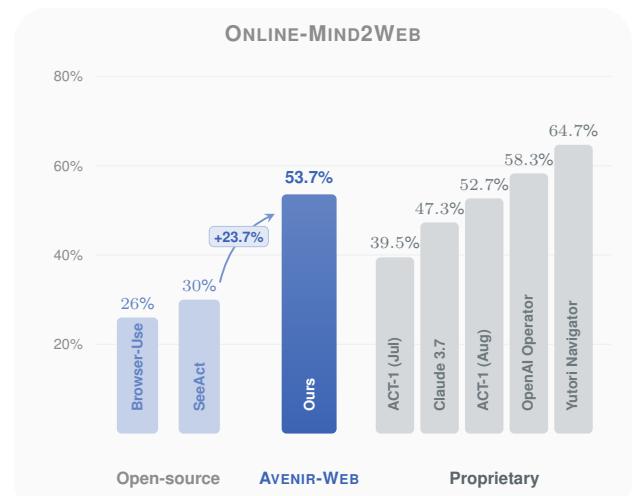


Figure 1. Performance of AVENIR-WEB on the ONLINE-MIND2WEB (Xue et al., 2025) benchmark (300 live tasks). The figure compares the success rate of our agent with existing open-source baselines and proprietary models. AVENIR-WEB achieves a **53.7%** success rate, which is shown alongside specialized agents such as ACT-1 (Enhans, 2025), OPERATOR (OpenAI, 2025a), and NAVIGATOR (Yutori, 2025).

1. Introduction

See it. Say it. Sorted! In modern computing environments characterized by complex and dynamic graphical user interfaces, a web agent is defined as an autonomous system (Deng et al., 2023; Zheng et al., 2024) capable of perceiving web user interfaces, interpreting natural language instructions, and executing multi-step action sequences within a standard browser environment. These agents facilitate end-to-end assistance for intricate workflows, ranging from flight booking and form completion to enterprise data extraction and application configuration. Recent advancements in general-purpose multimodal large language models (MLLMs), such as GPT-4o (OpenAI, 2025b), Gemini (Google DeepMind, 2025), and Claude (Anthropic, 2025), have established formidable reasoning and agentic planning capabilities (Yao et al., 2023; Shinn et al., 2023), enabling sophisticated interpretation of web content via structured HTML or accessibility tree analysis. Concurrently, the emergence of specialized GUI models (Hong et al., 2024; Lu et al., 2024; Lin et al., 2024) (foundational

¹, ‘Avenir’ is the French word for ‘future’, and also refers to the classic geometric sans-serif typeface known for its forward-looking aesthetic.

¹University College London, London, UK ²Princeton University, Princeton, NJ, USA ³The University of Edinburgh, Edinburgh, UK. Correspondence to: Shilong Liu <slongliu86@gmail.com>, Mengdi Wang <mengdiw@princeton.edu>.

backbones explicitly trained for native OS-level interaction) has significantly enhanced visual grounding accuracy and native computer-interaction reliability. This trend has been further extended by recent open-source scaling efforts and vision-language-action paradigms (Liu et al., 2025; Lin et al., 2024). Together, these developments have catalyzed the emergence of autonomous systems (He et al., 2024; Gur et al., 2024) capable of grounding complex natural language commands into executable browser actions on live websites.

Despite this rapid progress, current web agents still suffer from critical reliability bottlenecks in real-world deployment. Overall, these limitations can be summarized into three major bottlenecks: inaccurate *element grounding* (Yang et al., 2023; Wu et al., 2025), the lack of site-specific procedural knowledge (Gur et al., 2024; Furuta et al., 2024), and unstable *long-term task tracking and memory* (Xue et al., 2025). (1) First, regarding *element grounding*, relying on a single modality or failing to deeply fuse multiple modalities is often insufficient for modern web applications (Deng et al., 2023; Gur et al., 2024). Agents can break on iframes, canvas elements, shadow DOMs, and other non-standard structures, or miss fine-grained semantic attributes and scrolling context. (2) Second, regarding site-specific procedural knowledge, existing systems generally lack the ability to learn from human experience encoded in external resources such as human-authored online guides. Without such guidance, agents are forced into trial-and-error exploration (Xue et al., 2025), which significantly increases token consumption and the risk of task incompleteness due to reaching step limits or encountering irreversible navigation errors. (3) Third, regarding *long-term task tracking and memory*, many agents remain largely reactive and short-horizon (Shinn et al., 2023; Yao et al., 2023). When tasks span multiple pages and state transitions, their internal state representation degrades, subgoal progress becomes ambiguous, and the agent exhibits navigational drift and compounding execution errors.

To address these limitations, we introduce AVENIR-WEB, a web agent that achieves a new open-source state-of-the-art (SOTA) on ONLINE-MIND2WEB, targeting the three core bottlenecks in real-world deployment. Specifically, AVENIR-WEB integrates *Mixture of Grounding Experts* (*MoGE*) for robust element grounding, *Experience-Imitation Planning* (*EIP*) for incorporating external procedural priors, and a combination of *Task-Tracking Checklist* and *Adaptive Memory* for long-term task tracking and memory (see Table 1 for a comparison with existing agents).

(1) To improve element grounding, we develop *Mixture of Grounding Experts* (*MoGE*) for web tasks (Figure 6), which prioritizes a visual-first grounding path using an MLLM. This approach emulates human behavior by interacting with the interface as a unified visual canvas, effectively resolv-

ing elements within complex structures like nested iframes that typically paralyze DOM-centric agents (Zheng et al., 2024; He et al., 2024). For specific edge cases requiring high-precision manipulation, MoGE utilizes a semantic structural reasoner as a robust fallback. Compared with DOM-centric pipelines that often require separate inferences for Action Generation and Action Grounding (Zheng et al., 2024), MoGE typically emits an executable action in a single inference step, while triggering extra rounds only for delicate operations (e.g., selection) or when the initially grounded point is non-responsive. (2) To incorporate site-specific knowledge, we propose *Experience-Imitation Planning* (*EIP*) (Figure 3), which retrieves and comprehends human-authored online guides (Gur et al., 2024) to produce high-level, site-specific plans, avoiding expensive open-ended exploration and reducing reliance on transient parametric memory. (3) To stabilize long-horizon execution, we further introduce a *Task-Tracking Checklist* that explicitly records subgoal completion (Deng et al., 2023) and preserves a structured task state across page transitions, mitigating navigational drift and memory decay. Finally, we implement *Adaptive Memory* (Figure 7) to manage interaction history through *Chunked Recursive Summarization* and *Failure Reflection* (Shinn et al., 2023). This module distills long-horizon execution traces and immediate failures into a persistent summary buffer, ensuring the agent maintains strategic situational awareness and can reason over past errors without exceeding context constraints.

We rigorously evaluate AVENIR-WEB on ONLINE-MIND2WEB (Xue et al., 2025), a challenging benchmark reflecting diverse, live web domains. Our experiments demonstrate that AVENIR-WEB achieves a **23.7%** absolute improvement in task success rate over existing open-source baselines (Figure 1), establishing a new open-source state-of-the-art for reliable web agents on live websites. To further promote accessibility, we also introduce a *fully open-source* configuration utilizing the lightweight *Qwen-3-VL-8B*. This version achieves a **25.7%** success rate, rivaling the performance of earlier baselines that rely on significantly larger proprietary backbones such as GPT-4o. This result underscores the framework’s capacity to empower compact open-source models with robust, industry-standard agentic capabilities. These results validate that unifying Experience-Imitation Planning, Mixture of Grounding Experts (*MoGE*), the Task-Tracking Checklist, and Adaptive Memory significantly enhances an agent’s ability to navigate the complexity of the modern web.

In summary, our primary contributions are:

- We introduce AVENIR-WEB, a new ONLINE-MIND2WEB’s open-source state-of-the-art web agent designed for resilient long-horizon execution on live websites.

- We proposed *Mixture of Grounding Experts (MoGE)* for robust element grounding, *Experience-Imitation Planning (EIP)* for incorporating external procedural priors, and a combination of *Task-Tracking Checklist* and *Adaptive Memory* for resilient long-term task tracking and memory.
- We demonstrate ONLINE-MIND2WEB’s open-source state-of-the-art performance on the ONLINE-MIND2WEB (Xue et al., 2025) benchmark, achieving a **53.7%** success rate and substantially outperforming existing open-source baselines (Figure 1). Furthermore, our framework allows a *fully open-source* setup with a lightweight 8B model to reach the performance of earlier state-of-the-art baselines. We release the codebase to facilitate reproducible research.¹

2. Related Works

Autonomous Web Agents. The development of autonomous web agents has evolved from heuristic-based scripts to sophisticated multimodal systems. Early efforts primarily focused on leveraging HTML document structures or visual information for web interaction (Deng et al., 2023; Gur et al., 2024; Shaw et al., 2023; Furuta et al., 2024; Lee et al., 2023; Yan et al., 2023). Building on these foundations, several prominent agent systems have emerged, including NAVIGATOR (Yutori, 2025), ACT-1 (Enhans, 2025), SEEACT (Zheng et al., 2024), WEB VOYAGER (He et al., 2024), BROWSER-USE (use Team, 2024), OPERATOR (OpenAI, 2025a), COLORBROWSERAGENT (Zhou et al., 2026), SCALECUA (Liu et al., 2025), IBM CUGA (Marreed et al., 2025), WEBOPERATOR (Dihan et al., 2025), AGENTSYMBIOTIC (Zhang et al., 2025), LEARN-BY-INTERACT (Su et al., 2025), and WEBPILOT (Zhang et al., 2024). These systems explore different trade-offs between open-source availability, grounding modalities, and planning strategies (see Table 1). While these works show promise, generalizing to diverse web environments remains challenging. Recent agent frameworks have explored the potential of powerful LMMs (e.g., GPT-4V, Gemini) as generalist agents, demonstrating zero-shot capabilities on live websites. AVENIR-WEB builds on these foundations by integrating an *Experience-Imitation Planner* and *Adaptive Memory* to handle the reliability bottlenecks of long-horizon tasks.

MLLMs and GUI Grounding. Advanced Multimodal Large Language Models (MLLMs) such as QWEN3 (Bai et al., 2025; Xu et al., 2025), CLAUDE 4.5 (Anthropic, 2025), SEED-1.8 (ByteDance Seed Team, 2025), and GEMINI 3 PRO (Google DeepMind, 2025) have transformed web agents into strategic planners capable of sophisticated task decomposition and self-correction via reasoning traces (Yao

et al., 2023; Shinn et al., 2023). Modern architectures move beyond linear prompting to externalize latent reasoning through structured state monitors and reflection loops, mitigating error propagation in long-horizon navigation. Simultaneously, specialized GUI models, including UI-TARS (Qin et al., 2025; Wang et al., 2025a), CO-GAGENT (Hong et al., 2024), OMNIPARSER (Lu et al., 2024), SHOWUI (Lin et al., 2024), OPENCUA (Wang et al., 2025b), MAI-UI (Zhou et al., 2025), EVOCUA (Xue et al., 2026), GTA1 (Yang et al., 2025), and others (OpenAI, 2025b; H Company, 2025; Lei et al., 2025; Cheng et al., 2026; Gou et al., 2025), serve as powerful backbones for precise coordinate-based grounding. Despite these advances, GUI agents still face challenges with context loss and detection errors, prompting iterative strategies like *Chain-of-Ground* (Li et al., 2025; Wu et al., 2025; Hsieh et al., 2025) to enhance interaction robustness. Further research continues to address model hallucinations (Huang et al., 2025) and refine reasoning-eliciting techniques such as Set-of-Mark (SoM) and Chain-of-Thought (Yang et al., 2023; Wei et al., 2023).

3. Methods

The AVENIR-WEB framework introduces a modular architecture designed to address the reliability bottlenecks of autonomous web navigation through four primary components: *Experience-Imitation Planning (EIP)*, the *Task-Tracking Checklist*, *Adaptive Memory*, and the *Mixture of Grounding Experts (MoGE)*. Our system distributes reasoning and execution responsibilities across specialized modules to ensure consistency, interpretability, and resilience in dynamic web environments. The overall design is structured around two functional phases that directly target the core bottlenecks: Initialization for strategic synthesis, and the Execution Loop for resilient task tracking and robust element interaction (Figure 2).

The system’s execution pipeline follows a structured transition from strategic synthesis to iterative interaction. During the Initialization phase, the *Experience-Imitation Planner* establishes a strategic roadmap by breaking down the user goal into a high-level plan and an initial *Task-Tracking Checklist* of verifiable success criteria. Once initialized, the Action Agent enters a persistent Execution Loop to navigate the target interface. In each step, the agent integrates its current perception with the global strategy and historical context from *Adaptive Memory* to determine the optimal next operation. This intent is then operationalized through the MoGE module, which prioritizes direct visual grounding using general-purpose multimodal models while leveraging semantic structural reasoning to resolve complex interface elements. Finally, the system updates the *Task-Tracking Checklist* to monitor progress and detect execution anomalies.

¹Code is available at <https://github.com/Princeton-AI2-Lab/Avenir-Web>.

Table 1. Comparison of AVENIR-WEB with existing autonomous web agents across key functional dimensions. The table lists features such as Experience-Imitation Planning (EIP), Task-Tracking Checklist, and Adaptive Memory. We leave some entries blank when implementation details are unavailable.

Agent	Open Source	Grounding	External Knowledge	Checklist	Adapt. Mem.
Navigator (Yutori, 2025)	✗	Vision Only	—	—	—
ACT-1 (Enhans, 2025)	✗	—	—	—	—
SeeAct (Zheng et al., 2024)	✓	DOM Only	✗	✗	✗
Agent-E (Abuelsaad et al., 2024)	✓	DOM Only	✗	✗	✗
Browser-Use (use Team, 2024)	✓	DOM Only	✗	✗	✗
Avenir-Web (Ours)	✓	Mixture	✓	✓	✓

lies, ensuring robust goal-oriented navigation.

3.1. Experience-Imitation Planning

The Experience-Imitation Planner is a reasoning module that provides the system with procedural guidance by emulating human experience derived from online knowledge (Figure 3). Many web tasks require knowledge of site-specific workflows or familiarity with navigation patterns that are unique to the target website (Gur et al., 2024; Deng et al., 2023). To bridge this gap, the planner uses Claude 4.5 Sonnet (Anthropic, 2025) with its online search capability to retrieve site-specific resources such as forums, help centers, or user guides, identifying the most effective interaction sequences for a given goal. This strategic exploration phase is similar to paradigms employed by advanced multi-agent systems (Zhang et al., 2024) and tree-search-based operators (Dihan et al., 2025) that prioritize high-level roadmap synthesis over immediate reaction.

Without such human-derived experience, agents are often forced into open-ended exploration, which carries significant disadvantages: it leads to excessive token consumption, risks reaching maximum step limits before task completion, and increases the likelihood of encountering irreversible navigation errors that can terminate the entire session. This procedural imitation allows the agent to anticipate common obstacles and prioritize high-success navigation paths, significantly reducing the exploration time required to solve unfamiliar tasks.

This module operates during the initialization phase to establish a strategic roadmap before the interaction loop begins. By accessing site-specific online resources, the planner synthesizes the user instruction and target URL into a high-level plan consisting of 2-4 imperative directives (see Figure 4 for an example). For instance, while a reactive agent might attempt to find a Careers link by searching the main navigation bar, the Experience-Imitation Planner identifies from the site’s specific documentation that such links are located exclusively in the page footer, directing the agent to scroll to the bottom of the page immediately. This plan is restricted to abstract action descriptions rather than precise selectors to ensure it remains robust across different interface layouts.

This approach allows the Action Agent to follow a proven procedural path, decoupling high-level strategy from low-level execution and preventing the system from becoming stuck in local navigation errors.

3.2. Task-Tracking Checklist

To maintain long-term goal focus and prevent repetitive failure loops, AVENIR-WEB utilizes a Task-Tracking Checklist that decomposes complex instructions into verifiable milestones (Deng et al., 2023) (Figure 5). This structured decomposition is critical for overcoming the compounding error problem observed in long-horizon web navigation benchmarks (Xue et al., 2025; He et al., 2024). Figure 4 illustrates the parallel initialization of the Experience-Imitation Planning (EIP) and the Task-Tracking Checklist for a navigation task on [petfinder.com](https://www.petfinder.com). The checklist defines *atomic outcome states* that remain valid regardless of the specific navigation path taken.

Checklist Generation. When a new task starts, the system establishes the initial success criteria. The Experience-Imitation Planner constructs a prompt that instructs the model to create 2–6 atomic items, each limited to 10 words and output in a structured JSON format. To ensure reliability, the raw output is passed through a refinement layer that splits complex sentences, deduplicates similar goals, and initializes all statuses to a pending state. If the model fails to provide a valid structure after retries, the system utilizes a hardcoded fallback consisting of general execution and completion checkpoints.

Iterative Updating. After every interaction step t , the checklist is synchronized with the environment. Let $C_t = \{(m_i, s_{i,t})\}_{i=1}^k$ be the checklist at step t , where m_i is the i -th milestone and $s_{i,t} \in \{P, IP, C, F\}$ is its status (Pending, In Progress, Completed, Failed). The updated checklist C_{t+1} is computed as:

$$C_{t+1} = \mathcal{F}_\theta(C_t, a_t, o_t, S_t) \quad (1)$$

where a_t is the executed action, o_t is the observed outcome, and S_t is the current page state. This update is performed by a lightweight *Qwen-3-VL-8B* (Bai et al., 2025) model tasked with updating exactly one item most directly affected by the

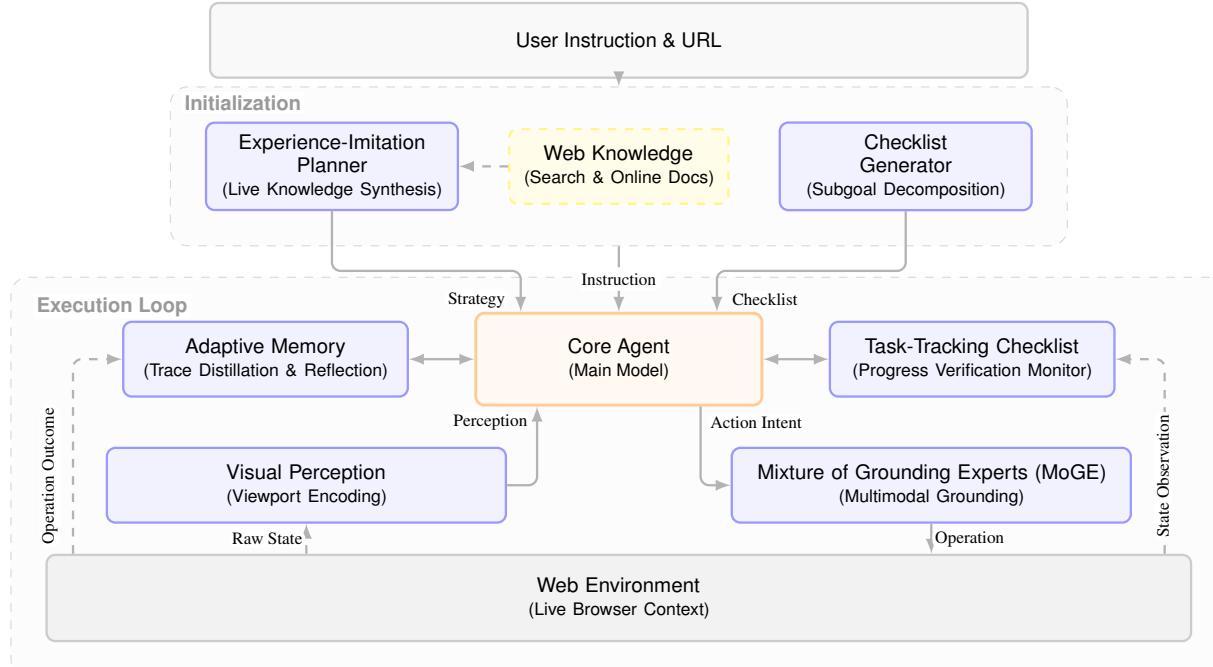


Figure 2. System architecture of AVENIR-WEB, featuring a decoupled strategic planning and execution framework. The *Initialization* phase utilizes the **Experience-Imitation Planner (EIP)** to transform external procedural knowledge into a verifiable **Task-Tracking Checklist**. During the iterative *Execution Loop*, the agent maintains strategic consistency through **Adaptive Memory** while the **Mixture of Grounding Experts (MoGE)** ensures robust element interaction via hierarchical visual-semantic grounding. A closed-loop feedback mechanism propagates environmental state observations back to the checklist and memory modules to prevent navigational drift in long-horizon tasks.

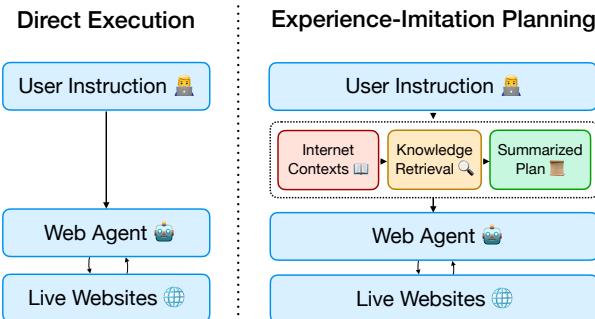


Figure 3. Comparison of Experience-Imitation Planning (EIP). Without EIP, the agent executes instructions directly against the live website. With EIP, external how-to knowledge is searched and summarized into a site-specific plan that guides the agent’s interaction.

action. This selective update strategy minimizes latency and ensures that the checklist remains a reliable source of truth for the Action Agent throughout the execution loop.

3.3. Mixture of Grounding Experts (MoGE)

The core operative module of AVENIR-WEB is the Mixture of Grounding Experts (MoGE), which prioritizes direct visual grounding using general-purpose multimodal models such as QWEN3-VL(Bai et al., 2025) and GEMINI 3

PRO(Google DeepMind, 2025). This design emulates human interaction by treating the entire viewport, including nested iframes, as a unified visual canvas, thereby bypassing the structural complexity of underlying HTML code. By focusing on coordinate-based interaction, an approach increasingly adopted by native GUI agents (Lin et al., 2024; Qin et al., 2025; Hong et al., 2024), MoGE naturally resolves elements within iframes that frequently hinder DOM-centric agents. In edge cases where visual cues are insufficient or fine-grained structural control is required, the system falls back to semantic structural reasoning to ensure reliable grounding across diverse web paradigms (Zheng et al., 2024; Gou et al., 2025; Yan et al., 2023) (see Figure 8). Crucially, MoGE functions as a one-step grounding system for standard interactions, producing an executable action with a single model inference. To improve precision, it incorporates a visual annotation layer inspired by Set-of-Mark (SoM) prompting (Yang et al., 2023), overlaying interactive elements with unique identifiers to disambiguate closely spaced controls. This design enables less than one model call per step on average, streamlining the interaction loop while avoiding spatial context loss during HTML parsing (Lu et al., 2024; Hsieh et al., 2025). For more complex grounding scenarios, MoGE can be extended with iterative reasoning steps, following the *Chain-of-Ground* paradigm (Li et al., 2025; Lei et al., 2025).

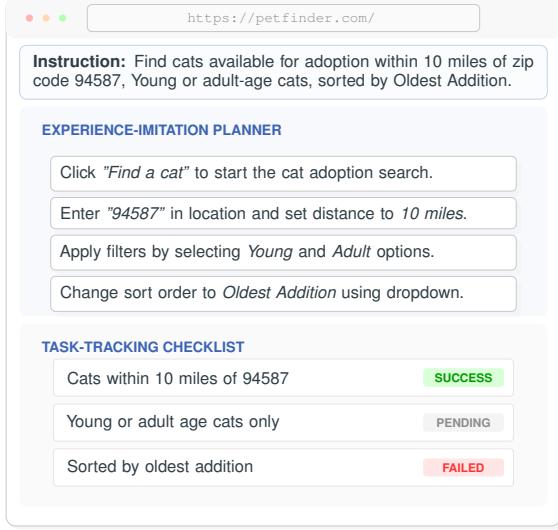


Figure 4. AVENIR-WEB execution context for [petfinder.co](https://petfinder.com)m. The figure illustrates the integration of Experience-Imitation Planning (EIP) and Task-Tracking Checklist with status feedback.



Figure 5. The Task-Tracking Checklist lifecycle. (1) Initialization generates atomic outcome states from the user instruction. (2) During the execution loop, a lightweight model iteratively updates the checklist based on action feedback and environment observations.

For specialized interactions, MoGE adopts a hierarchical decision-making architecture that prioritizes execution efficiency while maintaining robustness through progressively richer grounding strategies. The framework supports heterogeneous modalities via a tiered design in which point-based actions emphasize high-precision spatial localization in a normalized coordinate space, while preserving parallel semantic context to enable structural recovery when visual grounding fails to induce a valid state transition. Text entry follows a staged protocol that first attempts direct coordinate-based input via a virtual keyboard; upon failure or verification inconsistency, the system falls back to structural element targeting, and finally to a global search over candidate input fields ranked by spatial proximity or semantic similarity, with an LLM resolving remaining ambiguities. Dropdown interactions use a complementary fallback strategy, beginning with direct value assignment through script-level manipulation to handle system-native components that are often inaccessible to coordinate-based clicking; if unsuccessful, the system performs a semantic search over selectable elements and applies LLM-based reasoning

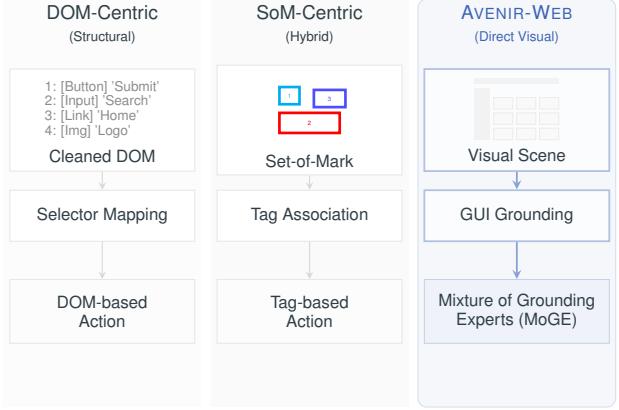


Figure 6. Architectural comparison of web agent grounding paradigms. DOM-Centric models rely on structural parsing of the page hierarchy. SoM-Centric systems utilize visual tagging to bridge the pixel-element gap. AVENIR-WEB uses MoGE for semantic-first grounding with hierarchical fallbacks, leveraging structured metadata to resolve complex interactions while maintaining direct visual grounding capabilities.

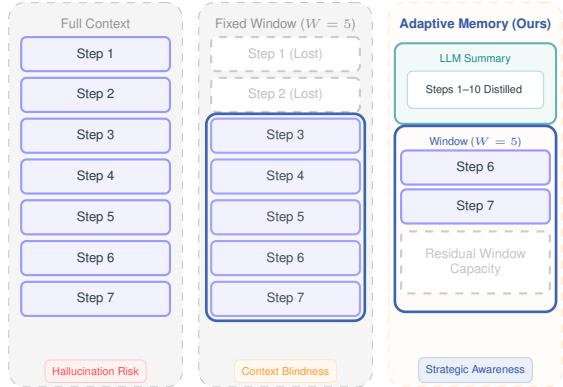


Figure 7. Comparison of memory architectures for long-horizon web navigation. While *Full Context* risks hallucination due to excessive tokens and *Fixed Window* suffers from context blindness as early steps are lost, our **Adaptive Memory** maintains strategic awareness by combining a distilled recursive summary of past actions with a precise sliding window of recent interactions.

to identify and select the most appropriate option.

3.4. Adaptive Memory

Sequential decision-making in long-horizon web tasks requires a memory mechanism that can balance detailed execution history with long-term strategic awareness. Without such a mechanism, agents often operate in a purely reactive manner, taking one step at a time without analyzing why previous actions may have failed to reach a goal (Yao et al., 2023; Shinn et al., 2023). This often results in repetitive failure loops where the system repeats the same unsuccessful action until reaching a step limit (see Figure 7).

To address the propensity for navigational drift and repetitive errors, we introduce Adaptive Memory, which balances tactical interaction history with strategic awareness

via Chunked Recursive Summarization. This mechanism is inspired by the need for long-term reasoning and self-correction in complex agentic workflows (Yao et al., 2023; Shinn et al., 2023). The system operates over a sliding window of size W (default $W = 5$), periodically distilling its raw interaction buffer \mathcal{B}_k into a more abstract, persistent memory state \mathcal{M}_k . The buffer for the k -th chunk is defined as:

$$\mathcal{B}_k = \{(a_i, o_i, S_i)\}_{i=(k-1)W+1}^{kW} \quad (2)$$

This recursive distillation preserves high-level strategic awareness while preventing the context-induced hallucination (Huang et al., 2025) that often occurs when an agent is overwhelmed by a long and repetitive history of low-level actions. Upon the completion of each interaction chunk, the memory state is updated as:

$$\mathcal{M}_k = \mathcal{G}_\phi(\mathcal{M}_{k-1}, \mathcal{B}_k, \mathcal{E}_k) \quad (3)$$

where \mathcal{G}_ϕ is the distillation function parameterized by ϕ , and \mathcal{E}_k represents the failure reflection buffer containing distilled traces of execution errors within the chunk. This ensures that the agent retains long-term situational awareness without context overflow. Furthermore, to ensure that critical failures are never lost during summarization, any execution error or unexpected feedback is immediately summarized by the LLM and added to the summary buffer. This *Failure Reflection* allows the agent to reason over past errors with distilled clarity, even after the corresponding step has been summarized or removed from the active window.

To facilitate this reflection, the system implements a robust outcome detection mechanism that verifies the impact of every action. It compares the page state before and after execution to determine whether an interaction has resulted in a meaningful transition. The mechanism evaluates state changes through a prioritized hierarchy, looking for changes in visible text, interactive elements, focus, the URL, scroll position, or modal popups. For key interactions such as point-based clicking or textual input, the absence of detected changes triggers a failure flag. When a failure is confirmed, the system logs a warning and captures a coordinate-annotated screenshot to assist in future reasoning, while propagating the failure status to the Task-Tracking Checklist to prompt a strategic retry.

4. Experiments

4.1. Experimental Setup

Benchmark. We evaluate AVENIR-WEB on ONLINE-MIND2WEB (Xue et al., 2025), an online evaluation benchmark consisting of 300 diverse and realistic tasks spanning 136 websites. This environment enables the evaluation of web agents under a setting that approximates real-world user interaction, encompassing dynamic content, complex Docu-

ment Object Model (DOM) structures, and state-dependent workflows.

Metrics. We report the *Task Success Rate (TSR)*, which measures the percentage of tasks where the agent successfully reaches the target goal state. To facilitate scalable evaluation, we utilize a novel *LLM-as-a-Judge* automatic evaluation method. This approach, powered by *o4-mini*, achieves an average agreement rate of 85.7% with human judgment and maintains a narrow success rate gap of just 3.8%, which is substantially higher than existing automated methods.

Models & Baselines. We benchmark AVENIR-WEB against the primary open-source baselines SEEACT (Zheng et al., 2024), BROWSER USE (use Team, 2024), and AGENT-E, alongside state-of-the-art proprietary agents such as OPERATOR (OpenAI, 2025a), GEMINI 2.5 COMPUTER USE (DeepMind, 2025), and ACT-1 (Enhans, 2025). We utilize Gemini 3 Pro (Google DeepMind, 2025) as our primary action backbone. Furthermore, our architecture leverages specialized models for auxiliary reasoning: the Experience-Imitation Planning (EIP) module employs *Claude 4.5 Sonnet* (Anthropic, 2025) for strategic roadmap synthesis, while the Task-Tracking Checklist is managed by a lightweight *Qwen-3-VL-8B* (Bai et al., 2025) model to ensure responsive state monitoring.

4.2. Main Results

Table 2 presents a comprehensive comparison on the ONLINE-MIND2WEB benchmark. AVENIR-WEB achieves a new open-source state-of-the-art performance on ONLINE-MIND2WEB, significantly outperforming existing baselines such as SEEACT (Zheng et al., 2024) (30.0%), AGENT-E (27.0%), and BROWSER USE (use Team, 2024) (26.0%).

Our primary configuration, AVENIR-WEB (Gemini 3 Pro), achieves a success rate of **53.7%**, representing an absolute improvement of at least **23.7%** over prior open-source systems. Notably, we also provide a fully open-source configuration using *Qwen-3-VL-8B* as the main action model. This version achieves a success rate of **25.7%**, demonstrating that our framework enables even a lightweight 8B model to reach performance levels comparable to existing open-source baselines that rely on much larger proprietary models. While it trails specialized commercial systems like *Yutori Navigator* (Yutori, 2025) (64.7%), OpenAI Operator (58.3%) (OpenAI, 2025a) and Google Gemini 2.5 Computer Use (57.3%) (DeepMind, 2025), it outperforms other major proprietary baselines including ACT-1-20250814 (Enhans, 2025) (52.7%) and Claude Computer Use 3.7 (Anthropic, 2025) (47.3%). This establishes AVENIR-WEB as the new open-source state-of-the-art on ONLINE-MIND2WEB, effectively bridging the performance gap with proprietary frontier models.

Table 2. Task Success Rate on ONLINE-MIND2WEB (Xue et al., 2025) across difficulty levels. Results for AVENIR-WEB (the new open-source state-of-the-art on ONLINE-MIND2WEB) are shown alongside industry baselines and specialized proprietary agents.

Agent	Main Model	Provider	Open Source	Easy	Med.	Hard	Overall
Proprietary Models							
Navigator (Yutori, 2025)	n1-preview-11-2025	Yutori	✗	84.0	62.2	48.7	64.7
Operator (OpenAI, 2025a)	OpenAI Computer-Using Agent	OpenAI	✗	73.5	59.4	39.2	58.3
Google Computer Use (DeepMind, 2025)	Gemini 2.5 Computer Use	Google DeepMind	✗	77.1	55.2	45.9	57.3
ACT-1-20250814 (Enhans, 2025)	o3-2025-04-16 and Claude-sonnet-4-20250514	Enhans	✗	71.1	52.4	32.4	52.7
Claude Computer Use 3.7 (Anthropic, 2025)	Claude-3.7-sonnet-20250219	Anthropic	✗	75.9	41.3	27.0	47.3
ACT-1-20250703 (Enhans, 2025)	o3-2025-04-16 and Claude-sonnet-4-20250514	Enhans	✗	53.7	39.2	24.3	39.5
Claude 3.5 (Anthropic, 2024)	Claude-3-5-sonnet-20241022	Anthropic	✗	51.8	16.1	8.1	24.0
Open-Source Baselines							
SeeAct (Zheng et al., 2024)	gpt-4o-2024-08-06	OSU	✓	51.8	28.0	9.5	30.0
Agent-E (Abuelsaad et al., 2024)	gpt-4o-2024-08-06	Emergence AI	✓	51.8	23.1	6.8	27.0
Browser Use (use Team, 2024)	gpt-4o-2024-08-06	Browser Use	✓	44.6	23.1	10.8	26.0
Avenir-Web (Ours)							
Avenir-Web	Gemini 3 Pro	Ours	✓	74.1	54.6	30.3	53.7
Avenir-Web	Qwen-3-VL-8B	Ours	✓	42.0	23.8	11.8	25.7

4.3. Ablation Studies

To evaluate the impact of the four core components of AVENIR-WEB, namely Experience-Imitation Planning (EIP), Mixture of Grounding Experts (MoGE), Task-Tracking Checklist, and Adaptive Memory, we perform ablation studies on a 50-tasks subset of Online-Mind2Web (Table 3).

Removing *Experience-Imitation Planning* drops the success rate from 48.0% to 36.0%, while disabling the *Task-Tracking Checklist* leads to a 4.0% decline, confirming the necessity of site-specific knowledge and verifiable state tracking. Disabling *MoGE* reduces performance to 40.0% due to struggles with fine-grained elements that visual-only models often misinterpret. Finally, replacing recursive *Adaptive Memory* with simple sliding windows ($W = 5$ or $W = \infty$) decreases performance; specifically, the $W = \infty$ setting causes significant hallucination as the context window saturates, validating that recursive distillation preserves situational awareness while preventing context-induced errors in long-horizon trajectories.

4.4. Qualitative Analysis

To further investigate the impact of the proposed grounding and memory mechanisms, we perform a qualitative trajectory analysis and provide representative screenshot-based trajectories in Appendix A. As shown in Figure 8, AVENIR-WEB maintains stable progress through multi-step interfaces where DOM-centric baselines stall on iframe-heavy layouts, highlighting the practical benefits of MoGE grounding and memory-aware task tracking in real web settings.

5. Conclusion

In this work, we presented AVENIR-WEB, a new open-source state-of-the-art web agent on ONLINE-MIND2WEB designed to overcome the reliability bottlenecks inherent in existing web agents on live websites. By unifying *Experience-Imitation Planning* to incorporate site-specific

Table 3. Ablation study on MoGE, EIP, and Adaptive Memory on a Online-Mind2Web subset of 50 tasks (Backbone: Gemini 3 Flash).

Configuration	Success Rate (%)
Full Model (AVENIR-WEB)	48.0
w/o Task-Tracking Checklist	44.0
w/o Adaptive Memory ($W = 5$)	42.0
w/o Adaptive Memory ($W = \infty$)	36.0
w/o Experience-Imitation Planning (EIP)	36.0
w/o Mixture of Grounding Expert (MoGE)	40.0

procedural knowledge, a *Mixture of Grounding Experts (MoGE)* for precise cross-modal element interaction, and a *Task-Tracking Checklist* with *Adaptive Memory* for resilient state management, AVENIR-WEB effectively bridges the gap between high-level user intent and low-level execution on dynamic web interfaces.

Our extensive evaluation on the ONLINE-MIND2WEB (Xue et al., 2025) benchmark demonstrates that AVENIR-WEB sets a new open-source state-of-the-art on ONLINE-MIND2WEB, achieving a **53.7%** success rate and outperforming prior baselines by a significant margin while remaining competitive with top-tier proprietary models. Ablation results further confirm that each module contributes to reliability, with the largest drops observed when removing Experience-Imitation Planning or effective memory management (Table 3). These results validate the efficacy of our dual-layer reasoning architecture and hybrid grounding strategy in handling the complexity of modern web tasks. Furthermore, the integration of strategic planning and recursive memory distillation not only improves task success but also enhances operational efficiency by reducing redundant trial-and-error exploration and mitigating token-intensive context bloat. We posit that AVENIR-WEB provides a foundational step towards more autonomous, general-purpose digital assistants capable of navigating the open web with human-level reliability. Subsequent research will explore further optimization of agent latency and investigate the scalability of experience-guided planning across a broader range of digital applications.

Social Impact and Limitations. While AVENIR-WEB enhances productivity by automating routine workflows, its deployment involves critical safety and ethical considerations, including privacy risks and the potential for harmful actions. Technical limitations such as grounding accuracy and latency also persist.

References

- Abuelsaad, T., Akkil, D., Dey, P., Jagmohan, A., Vempaty, A., and Kokku, R. Agent-e: From autonomous web navigation to foundational design principles in agentic systems, 2024. URL <https://arxiv.org/abs/2407.13032>.
- Anthropic. Claude 3.5 sonnet, 2024. URL <https://www.anthropic.com/news/clause-3-5-sonnet>.
- Anthropic. Claude 4.5 sonnet, 2025. URL <https://anthropic.com/clause>.
- Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report, 2025. URL <https://arxiv.org/abs/2511.21631>.
- ByteDance Seed Team. Official release of seed 1.8: A generalized agentic model. https://seed/bytedance.com/en/seed1_8, 2025. Accessed Jan 28, 2026.
- Cheng, Y., Tang, L., Li, S., Huo, Y., Duan, T., Huang, K., Jing, Y., and Yan, Y. Evolving in tasks: Empowering the multi-modality large language model as the computer use agent, 2026. URL <https://arxiv.org/abs/2508.04037>.
- DeepMind, G. Introducing the gemini 2.5 computer use model. <https://blog.google/technology/google-deepmind/gemini-computer-use-model/>, 2025. Accessed: 2026-01-20.
- Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., and Su, Y. Mind2web: Towards a generalist agent for the web. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 28091–28114. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5950bf290a1570ea401bf98882128160-Paper-Datasets_and_Benchmarks.pdf.
- Dihan, M. L., Hashem, T., Ali, M. E., and Parvez, M. R. Weboperator: Action-aware tree search for autonomous agents in web environment, 2025. URL <https://arxiv.org/abs/2512.12692>.
- Enhans. Enhans: Empowering humanity to turn ideas into reality. <https://www.enhans.ai/>, 2025. Accessed: 2026-01-20.
- Furuta, H., Lee, K.-H., Nachum, O., Matsuo, Y., Faust, A., Gu, S. S., and Gur, I. Multimodal web navigation with instruction-finetuned foundation models, 2024. URL <https://arxiv.org/abs/2305.11854>.
- Google DeepMind. Gemini 3 pro. <https://deepmind.google/models/gemini/pro/>, 2025. Accessed Jan 28, 2026.
- Gou, B., Wang, R., Zheng, B., Xie, Y., Chang, C., Shu, Y., Sun, H., and Su, Y. Navigating the digital world as humans do: Universal visual grounding for GUI agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=kxnoqaisCT>.
- Gur, I., Furuta, H., Huang, A. V., Safdari, M., Matsuo, Y., Eck, D., and Faust, A. A real-world webagent with planning, long context understanding, and program synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=9JQtrumvg8>.
- H Company. Holo2: Cost-efficient models for cross-platform computer-use agents. <https://hcompany.ai/holo2>, 2025. Accessed Jan 28, 2026.
- He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., and Yu, D. Webvoyager: Building an end-to-end web agent with large multimodal models, 2024. URL <https://arxiv.org/abs/2401.13919>.
- Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Zhang, Y., Li, J., Xu, B., Dong, Y., Ding, M., and Tang, J. Cogagent: A visual language model for gui agents, 2024. URL <https://arxiv.org/abs/2312.08914>.
- Hsieh, Z., Wei, T.-J., and Yang, S. Zonui-3b: A lightweight vision-language model for cross-resolution gui grounding, 2025. URL <https://arxiv.org/abs/2506.23491>.
- Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A

- survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL <http://dx.doi.org/10.1145/3703155>.
- Lee, K., Joshi, M., Turc, I., Hu, H., Liu, F., Eisenschlos, J., Khandelwal, U., Shaw, P., Chang, M.-W., and Toutanova, K. Pix2struct: Screenshot parsing as pretraining for visual language understanding, 2023. URL <https://arxiv.org/abs/2210.03347>.
- Lei, B., Xu, N., Payani, A., Hong, M., Liao, C., Cao, Y., and Ding, C. GUI-SPOTLIGHT: Adaptive iterative focus refinement for enhanced gui visual grounding, 2025. URL <https://arxiv.org/abs/2510.04039>.
- Li, A. Y., Yu, B., Lei, D., Ren, T., and Liu, S. Chain-of-ground: Improving gui grounding via iterative reasoning and reference feedback, 2025. URL <https://arxiv.org/abs/2512.01979>.
- Lin, K. Q., Li, L., Gao, D., Yang, Z., Wu, S., Bai, Z., Lei, W., Wang, L., and Shou, M. Z. Showui: One vision-language-action model for gui visual agent, 2024. URL <https://arxiv.org/abs/2411.17465>.
- Liu, Z., Xie, J., Ding, Z., Li, Z., Yang, B., Wu, Z., Wang, X., Sun, Q., Liu, S., Wang, W., Ye, S., Li, Q., Dong, X., Yu, Y., Lu, C., Mo, Y., Yan, Y., Tian, Z., Zhang, X., Huang, Y., Liu, Y., Su, W., Luo, G., Yue, X., Qi, B., Chen, K., Zhou, B., Qiao, Y., Chen, Q., and Wang, W. Scalecua: Scaling open-source computer use agents with cross-platform data, 2025. URL <https://arxiv.org/abs/2509.15221>.
- Lu, Y., Yang, J., Shen, Y., and Awadallah, A. Omniparser for pure vision based gui agent, 2024. URL <https://arxiv.org/abs/2408.00203>.
- Marreed, S., Oved, A., Yaeli, A., Shlomov, S., Levy, I., Akrabi, O., Sela, A., Adi, A., and Mashkif, N. Towards enterprise-ready computer using generalist agent, 2025. URL <https://arxiv.org/abs/2503.01861>.
- OpenAI. Introducing operator. <https://openai.com/index/introducing-operator/>, 2025a. Accessed Jan 28, 2026.
- OpenAI. Computer-using agent. <https://openai.com/index/computer-using-agent/>, 2025b. OpenAI blog; accessed Jan 28, 2026.
- Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., Zhong, W., Li, K., Yang, J., Miao, Y., Lin, W., Liu, L., Jiang, X., Ma, Q., Li, J., Xiao, X., Cai, K., Li, C., Zheng, Y., Jin, C., Li, C., Zhou, X., Wang, M., Chen, H., Li, Z., Yang, H., Liu, H., Lin, F., Peng, T., Liu, X., and Shi, G. Ui-tars: Pioneering automated gui interaction with native agents, 2025. URL <https://arxiv.org/abs/2501.12326>.
- Shaw, P., Joshi, M., Cohan, J., Berant, J., Pasupat, P., Hu, H., Khandelwal, U., Lee, K., and Toutanova, K. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. In *Advances in Neural Information Processing Systems*, 2023. URL <https://arxiv.org/abs/2306.00245>.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. R., and Yao, S. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=vAE1hFcKw6>.
- Su, H., Sun, R., Yoon, J., Yin, P., Yu, T., and Arik, S. Ö. Learn-by-interact: A data-centric framework for self-adaptive agents in realistic environments, 2025. URL <https://arxiv.org/abs/2501.10893>.
- use Team, B. browser-use: Make websites accessible for ai agents. <https://github.com/browser-use/browser-use>, 2024. Accessed: 2026-01-17.
- Wang, H., Zou, H., Song, H., Feng, J., Fang, J., Lu, J., Liu, L., Luo, Q., Liang, S., Huang, S., Zhong, W., Ye, Y., Qin, Y., Xiong, Y., Song, Y., Wu, Z., Li, A., Li, B., Dun, C., Liu, C., Zan, D., Leng, F., Wang, H., Yu, H., Chen, H., Guo, H., Su, J., Huang, J., Shen, K., Shi, K., Yan, L., Zhao, P., Liu, P., Ye, Q., Zheng, R., Xin, S., Zhao, W. X., Heng, W., Huang, W., Wang, W., Qin, X., Lin, Y., Wu, Y., Chen, Z., Wang, Z., Zhong, B., Zhang, X., Li, X., Li, Y., Zhao, Z., Jiang, C., Wu, F., Zhou, H., Pang, J., Han, L., Liu, Q., Ma, Q., Liu, S., Cai, S., Fu, W., Liu, X., Wang, Y., Zhang, Z., Zhou, B., Li, G., Shi, J., Yang, J., Tang, J., Li, L., Han, Q., Lu, T., Lin, W., Tong, X., Li, X., Zhang, Y., Miao, Y., Jiang, Z., Li, Z., Zhao, Z., Li, C., Ma, D., Lin, F., Zhang, G., Yang, H., Guo, H., Zhu, H., Liu, J., Du, J., Cai, K., Li, K., Yuan, L., Han, M., Wang, M., Guo, S., Cheng, T., Ma, X., Xiao, X., Huang, X., Chen, X., Du, Y., Chen, Y., Wang, Y., Li, Z., Yang, Z., Zeng, Z., Jin, C., Li, C., Chen, H., Chen, H., Chen, J., Zhao, Q., and Shi, G. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning, 2025a. URL <https://arxiv.org/abs/2509.02544>.
- Wang, X., Wang, B., Lu, D., Yang, J., Xie, T., Wang, J., Deng, J., Guo, X., Xu, Y., Wu, C. H., Shen, Z., Li, Z., Li, R., Li, X., Chen, J., Zheng, B., Li, P., Lei, F., Cao, R., Fu, Y., Shin, D., Shin, M., Hu, J., Wang, Y., Chen, J., Ye, Y., Zhang, D., Du, D., Hu, H., Chen, H., Zhou, Z., Yao, H., Chen, Z., Gu, Q., Wang, Y., Wang, H., Yang, D., Zhong, V., Sung, F., Charles, Y., Yang, Z., and Yu, T. Opencua:

- Open foundations for computer-use agents, 2025b. URL <https://arxiv.org/abs/2508.09123>.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.1903>.
- Wu, H., Chen, H., Cai, Y., Liu, C., Ye, Q., Yang, M.-H., and Wang, Y. DiMo-GUI: Advancing test-time scaling in GUI grounding via modality-aware visual reasoning. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26246–26256, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1334. URL <https://aclanthology.org/2025.emnlp-main.1334/>.
- Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., and Lin, J. Qwen3-omni technical report, 2025. URL <https://arxiv.org/abs/2509.17765>.
- Xue, T., Qi, W., Shi, T., Song, C. H., Gou, B., Song, D., Sun, H., and Su, Y. An illusion of progress? assessing the current state of web agents. In *Second Conference on Language Modeling*, 2025. URL <https://openreview.net/forum?id=6jZi4HSs6o>.
- Xue, T., Peng, C., Huang, M., Guo, L., Han, T., Wang, H., Wang, J., Zhang, X., Yang, X., Zhao, D., Ding, J., Ma, X., Xie, Y., Pei, P., Cai, X., and Qiu, X. Evocua: Evolving computer use agents via learning from scalable synthetic experience, 2026. URL <https://arxiv.org/abs/2601.15876>.
- Yan, A., Yang, Z., Zhu, W., Lin, K., Li, L., Wang, J., Yang, J., Zhong, Y., McAuley, J., Gao, J., Liu, Z., and Wang, L. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation, 2023. URL <https://arxiv.org/abs/2311.07562>.
- Yang, J., Zhang, H., Li, F., Zou, X., Li, C., and Gao, J. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v, 2023. URL <https://arxiv.org/abs/2310.11441>.
- Yang, Y., Li, D., Dai, Y., Yang, Y., Luo, Z., Zhao, Z., Hu, Z., Huang, J., Saha, A., Chen, Z., Xu, R., Pan, L., Savarese, S., Xiong, C., and Li, J. Gta1: Gui test-time scaling agent, 2025. URL <https://arxiv.org/abs/2507.05791>.
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. React: Synergizing reasoning and acting in language models, 2023. URL <https://arxiv.org/abs/2210.03629>.
- Yutori. Introducing navigator — yutori. <https://yutori.com/blog/introducing-navigator>, 2025. Accessed: 2026-01-20.
- Zhang, R., Qiu, M., Tan, Z., Zhang, M., Lu, V., Peng, J., Xu, K., Agudelo, L. Z., Qian, P., and Chen, T. Symbiotic cooperation for web agents: Harnessing complementary strengths of large and small llms, 2025. URL <https://arxiv.org/abs/2502.07942>.
- Zhang, Y., Ma, Z., Ma, Y., Han, Z., Wu, Y., and Tresp, V. Webpilot: A versatile and autonomous multi-agent system for web task execution with strategic exploration, 2024. URL <https://arxiv.org/abs/2408.15978>.
- Zheng, B., Gou, B., Kil, J., Sun, H., and Su, Y. Gpt-4v(ision) is a generalist web agent, if grounded. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=piecKJ2D1B>.
- Zhou, H., Zhang, X., Tong, P., Zhang, J., Chen, L., Kong, Q., Cai, C., Liu, C., Wang, Y., Zhou, J., and Hoi, S. Maiui technical report: Real-world centric foundation gui agents, 2025. URL <https://arxiv.org/abs/2512.22047>.
- Zhou, J., Wang, J., Zhang, W., Liu, W., Zhang, Z., Lou, X., Zhang, W., Deng, H., and Wang, J. Colorbrowseragent: An intelligent gui agent for complex long-horizon web automation, 2026. URL <https://arxiv.org/abs/2601.07262>.

Appendix: Avenir-Web Supplemental Material

A. Case Study and Qualitative Analysis

To further investigate the impact of the proposed grounding and memory mechanisms, we perform a qualitative trajectory analysis. Figure 8 illustrates a representative failure mode for the SEEACT (Zheng et al., 2024) baseline, which lacks the MoGE module. In this scenario, the baseline agent fails to navigate the interface due to its inability to operate within nested iframes, resulting in repeated non-responsive actions and eventual task timeout. Conversely, AVENIR-WEB demonstrates superior navigational resilience by synergizing visual coordinate prediction with semantic fallbacks.

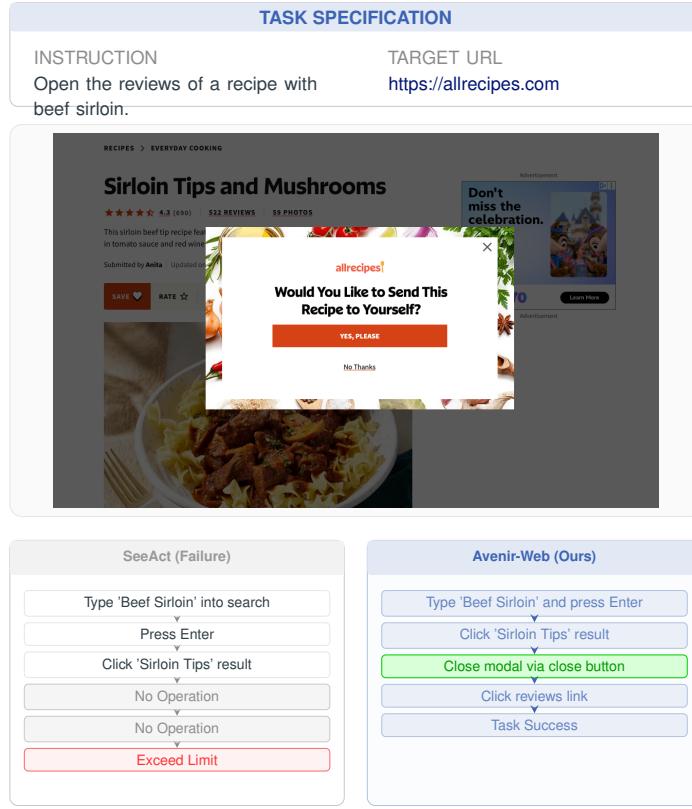


Figure 8. Comparative trajectory analysis on allrecipes.com. The SeeAct baseline fails to operate within nested iframes, resulting in a sequence of non-responsive actions. In contrast, AVENIR-WEB successfully navigates the complex interface by leveraging the MoGE grounding module and modal recovery.

As shown in the success trajectory on allrecipes.com (Figure 8), the agent effectively manages a multi-stage workflow involving: (i) an integrated text entry and search step where it types “Beef Sirloin” and immediately executes the submission, (ii) precise interaction to navigate to the recipe details, (iii) detection and dismissal of an obstructing modal window that appears prior to the target interaction, and (iv) successful execution of the review link interaction following modal recovery. This execution highlights how the MoGE module resolves spatial ambiguities, while the Adaptive Memory and Task-Tracking Checklist prevent the agent from entering the repetitive stalling loops characteristic of purely visual or DOM-centric baselines (Gur et al., 2024; He et al., 2024).

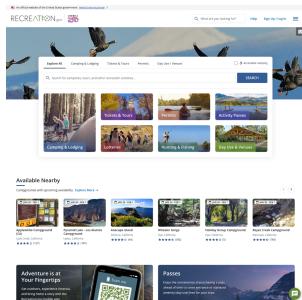
B. Detailed Case Study: Recreation.gov

To further illustrate the operational resilience of AVENIR-WEB on live websites, we provide a step-by-step execution trace for a task on recreation.gov. The metadata for this task is summarized below:

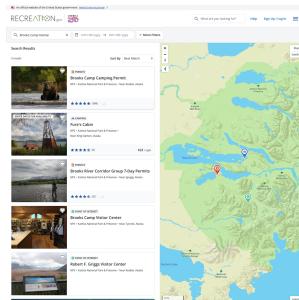
Task Instruction: Check permit availability for a group of 4 in Brooks Camp, Katmai National Park for next Saturday.

Target Website: <https://www.recreation.gov/>

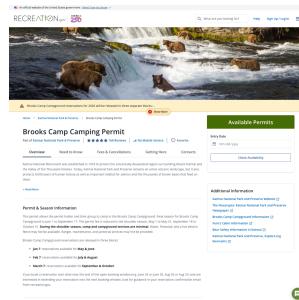
Task ID: 502e864440283214e0180645015f568b_110325



(a) Step 1: TYPE “Brooks Camp Katmai” into the main search bar.



(b) Step 2: CLICK the “Brooks Camp Camping Permit” result link.



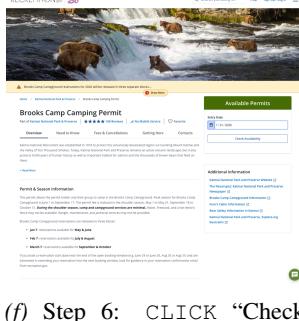
(c) Step 3: CLICK the “Entry Date” input field to open the calendar picker.



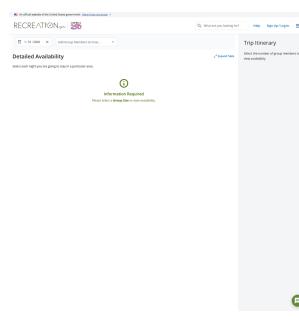
(d) Step 4: CLICK the calendar icon inside the “Entry Date” field.



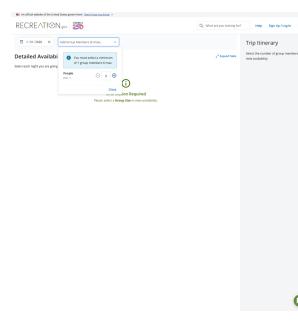
(e) Step 5: CLICK “31” to select the target Saturday date.



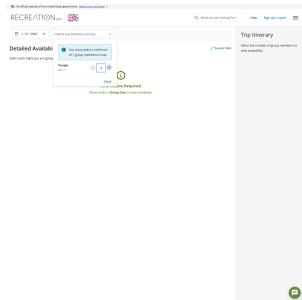
(f) Step 6: CLICK “Check Availability” to view permits for the selected date.



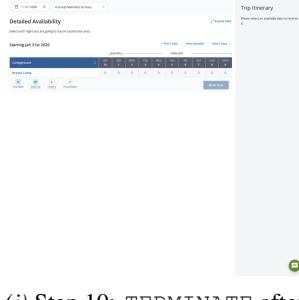
(g) Step 7: CLICK “Add Group Members” to set group size.



(h) Step 8: TYPE “4” into the guest counter input field.



(i) Step 9: CLICK “Close” to apply the group size selection.



(j) Step 10: TERMINATE after confirming zero availability for the selected date.

Figure 9. Full 10-step success trajectory of AVENIR-WEB on [recreation.gov](https://www.recreation.gov/). The agent successfully interacts with dynamic UI components including a calendar picker and a guest counter dropdown. The task concludes with an accurate termination based on the updated availability grid.

As shown in Figure 9, the agent follows a systematic roadmap:

- **Navigation and Search (Steps 1–2):** The agent identifies the search bar, types the target destination, and selects the correct permit link from the auto-suggested results.
- **Dynamic Component Interaction (Steps 3–6):** The agent navigates the complex date-picking interface by first focusing the input field and then using the visual calendar picker to select the specific Saturday. This illustrates the precision of the MoGE module in coordinate-based grounding.

- **Parameter Configuration (Steps 7–9):** The agent opens the guest configuration menu, inputs the numerical group size (“4”), and applies the selection to update the availability grid.
- **Termination (Step 10):** Upon confirming that the availability grid has updated (showing zero spots for the selected parameters), the agent terminates the session with a success status.

Complexity and Performance Analysis. The `recreation.gov` trajectory serves as a rigorous stress test for web agents due to the inherent logical complexity of its interface. Unlike static benchmarks, this live task requires navigating a deeply nested UI where information is gated by state-dependent widgets. The difficulty is primarily three-fold: First, the search phase requires disambiguating between various national park services to find the specific “Brooks Camp Camping Permit,” a process that often leads to navigational drift in less capable agents. Second, the interaction with the calendar picker represents a significant grounding challenge; the agent must not only locate the grid but also reason over the spatial arrangement of dates to select a specific day (“next Saturday”) while ignoring unavailable or disabled slots. Finally, the task requires maintaining a stable internal state while transitioning from the global search context to the specific parameter configuration menu for group size.

AVENIR-WEB overcomes these challenges through its dual-layer reasoning architecture. The high-precision visual grounding of the MoGE module allows it to treat the calendar and guest counters as direct visual targets, bypassing the brittle DOM hierarchies that frequently paralyze structural agents. Simultaneously, the synergy between the *Task-Tracking Checklist* and *Adaptive Memory* ensures that the agent remains logically anchored to the user’s specific constraints (group of 4, specific date) across multiple state transitions. By completing this high-entropy workflow in a concise 10-step sequence, AVENIR-WEB demonstrates a level of execution reliability and strategic consistency that bridges the gap between research prototypes and production-ready digital assistants.

C. Prompts

C.1. System Prompt

The system prompt defines the agent’s persona, capabilities (tools), and operational rules.

PROMPT: System Prompt

```

One action per turn with pixel coordinates.
CLICK: provide 'coordinate' or visible 'text'.
Close/accept blocking modals, overlays, cookie banners first.
Do not repeat actions unless page state visibly changed.
TYPE/SELECT only when target field/dropdown is visible.
KEYBOARD: use 'code' for keys; 'text' for typing; 'CLEAR' to clear active field.
SCROLL: omit coordinates to scroll page; include [x,y] to scroll a container.
If you see potential <select> elements, MUST use 'select' action directly. DO NOT use 'click' to open
dropdowns.
When objectives are achieved, TERMINATE with status 'success'.

Strategic guidance:
{strategic_reasoning}

<tools>
{
  "type": "function",
  "function": {
    "name": "browser_use",
    "description": "Single-step browser interaction using pixel coordinates or visible text.",
    "parameters": {
      "type": "object",
      "required": ["action"],
      "properties": {
        "action": {
          "type": "string",
          "enum": ["left_click", "hover", "keyboard", "type", "select", "press_enter", "scroll_up", "scroll_down", "scroll_top", "scroll_bottom", "new_tab", "close_tab", "go_back", "go_forward", "wait", "terminate"]
        },
        "coordinate": {"type": "array", "description": "Normalized [x,y] in 0-1000. REQUIRED for all actions
except scroll; omit for scroll only. Include to target a container."},
        "text": {"type": "string", "description": "Visible label or input text. Use 'CLEAR' for keyboard.", "maxLength": 200}
      }
    }
  }
}
```

```

"code": {"type": "string", "description": "KeyboardEvent.code (e.g., 'PageDown', 'ArrowDown', 'Enter')", "maxLength": 50},
"clear_first": {"type": "boolean", "description": "Clear active field before typing (type/keyboard)" },
"press_enter_after": {"type": "boolean", "description": "Press Enter after typing (action=type)" },
"field": {"type": "string", "description": "Semantic field name (email/search/password/country)", "maxLength": 100},
"time": {"type": "number", "description": "Seconds to wait" },
"status": {"type": "string", "enum": ["success", "failure"], "description": "Task status for terminate"}, "description": {"type": "string", "description": "Short action description (<=200 chars). REQUIRED.", "maxLength": 200}
}
}
}
</tools>

Screen: 1000x1000, origin (0,0) top-left.

Rules:
- Do not use GOTO for URL navigation.
- For <select> elements, YOU MUST use 'select' action directly. DO NOT use 'click' to open dropdowns.
- For all actions except scroll actions (scroll_up, scroll_down, scroll_top, scroll_bottom), YOU MUST provide the 'coordinate' parameter with normalized [x,y] values in 0-1000.
- keyboard: use 'code' for keys; 'text' for typing; 'CLEAR' clears the active field.
- **IMPORTANT**: You MUST provide 'coordinate' [x,y] for every CLICK, HOVER, or TYPE action. Do NOT rely on 'text' alone.

Return strictly in <tool_call> tags:
<tool_call>
{"name": "browser_use", "arguments": {"action": "...", ...}}
</tool_call>

```

C.2. User Prompt (Per-Step Input)

This prompt is constructed at each step to provide the current state and task context.

PROMPT: User Prompt

```

Task:
{task}

Pre-step:
Close or accept any cookie/consent banner before other actions.

Strategic guidance:
{strategic_reasoning}

Constraints:
{policy_constraints}

Previous actions:
{previous_actions}

Task progress:
{checklist_context}

```

C.3. Checklist Generation Prompt

Used to decompose the high-level task into atomic requirements.

PROMPT: Checklist Generation Prompt

Create 2-6 atomic outcome states based STRICTLY on the task description.

Task: {task_description}

Rules:

- 1) Each item is an observable goal state (not an action)
- 2) Max 10 words; short and specific
- 3) IDs: "requirement_1", "requirement_2", ...
- 4) Examples: "Size 'blue'", "T-shirt page", "Year: 2022-2023"
- 5) Status must be lowercase: pending, in_progress, completed, failed
- 6) DO NOT invent requirements not explicitly mentioned in the task.

```
Output JSON:
{
  "checklist": [
    {"id": "requirement_1", "description": "First outcome state", "status": "pending"},
    {"id": "requirement_2", "description": "Second outcome state", "status": "pending"}
  ]
}
```

C.4. Checklist Update Prompt

Used to update the status of checklist items based on the agent's actions.

PROMPT: Checklist Update Prompt

```
Update the checklist based on this action:

Action: {action_type} | Success: {success} | Error: {error}

Recent actions:
{history_text}
Page:
{page_state_text}...
Checklist:
{checklist_text}

Update rules:
* completed = fully satisfied
* in_progress = partially done
* pending = not started/reset
* failed = action failed
* Update exactly ONE item per action (most directly affected)
* new_status must be one of: pending, in_progress, completed, failed (lowercase)

Output JSON:
{
  "updates": [
    {"item_id": "requirement_X", "new_status": "pending", "reason": "Brief reason"}
  ]
}
```

C.5. Task Constraints

Standard safety/policy constraints injected into the User Prompt.

PROMPT: Task Constraints

```
Task-specific soft constraints:
- Do NOT attempt to log in, sign in, sign up, or provide credentials.
- If a login/sign-in UI is detected (password fields, 'Sign in', 'Log in', 'Create account'), TERMINATE
immediately with status 'failure' and reason 'login prohibited'.
```

D. Experience-Imitation Planning (EIP) Implementation

The *Experience-Imitation Planning (EIP)* module is responsible for generating site-specific roadmaps. Rather than relying on static prompts, the system leverages its integrated search capability to gather real-world knowledge about the target website.

D.1. Strategic Search and Synthesis

Before beginning any task, the agent performs a targeted search for the website's help documentation, community forums, or user guides. This allows it to identify specific workflows and interaction patterns that are unique to that site. These findings are then synthesized into a concise, imperative plan that guides subsequent tool calls. This process ensures that the agent is not just guessing, but is following established "best practices" for the specific interface it is interacting with.

D.2. Execution Flow

The EIP process follows a strictly defined narrative flow:

1. **Exploration:** The agent starts by searching for the target website's official documentation or relevant community-sourced guidance.
2. **Roadmap Generation:** It summarizes the search results into 2–4 actionable sentences, prioritizing visible labels and concrete interaction steps.
3. **Strategic Injection:** This high-level roadmap is injected into the main reasoning context, providing a strategic anchor for every subsequent action taken on the page.

E. Social Impact and Limitations

E.1. Ethical Constraints and Anti-Bot Challenges

In the development and evaluation of AVENIR-WEB, we maintain a strict ethical stance regarding web interaction and transparency. While proprietary agents such as ACT-1 may utilize more aggressive stealth or obfuscation strategies, our framework deliberately avoids the incorporation of CAPTCHA bypass services or header-masking mechanisms. We believe that autonomous agents should operate within the boundaries of a website's intended security protocols. This adherence to transparency is reflected in our results: in our most successful run, approximately 10% of tasks (31 out of 300) were blocked by the host infrastructure before any actions could be performed, a result that highlights the inherent friction encountered by non-stealthy agents in dynamic web environments.

However, this ethical commitment introduced substantial operational difficulties during our real-world online evaluation:

- **Cloudflare and WAF Challenges:** A significant portion of the websites in the ONLINE-MIND2WEB benchmark are protected by advanced Web Application Firewalls (WAFs) like Cloudflare. Without stealth capabilities, our agents were frequently flagged as automated traffic, resulting in persistent "Under Attack" mode challenges or JS-based browser integrity checks that the agent could not resolve autonomously.
- **IP Blocking and Rate Limiting:** Due to the live nature of our evaluation, which often requires repeated visits to the same websites for debugging and trajectory verification, our testing IP addresses were frequently flagged by anti-bot mechanisms (e.g., Cloudflare, Akamai). To ensure consistent evaluation, we utilized a rotating residential proxy service to bypass rate limits.
- **CAPTCHA Barriers:** We encountered numerous instances where legitimate user flows (e.g., account creation or form submission) were gated by CAPTCHAs. Adhering to our policy of not bypassing these mechanisms meant that the agent would terminate the task, negatively impacting our success rate but preserving our commitment to non-evasive research.

These challenges highlight a fundamental tension in web agent research: the need for realistic evaluation on the live web versus the necessity of maintaining high ethical and transparency standards. Future work may need to explore "white-listed" evaluation environments or cooperative protocols between agents and website operators.

E.2. Other Ethical and Safety Considerations

Beyond anti-bot challenges, the real-world deployment of autonomous web agents raises other significant concerns:

- **Privacy Risks:** Unauthorized access to personal profiles or sensitive user data during automated browsing.

- **Security of Sensitive Operations:** Risks associated with the automation of financial transactions or official form submissions.
- **Harmful Actions:** During our evaluation, we observed that agents could potentially generate harmful actions, necessitating manual safety validation.

We release our code strictly for research purposes and firmly oppose any harmful use of this technology.

E.3. Technical Limitations

Several technical constraints remain to be addressed in future work:

- **Computational Costs:** Reliance on large-scale MLLMs introduces latency and computational overhead unsuitable for sub-second response applications.
- **Grounding Accuracy:** Hybrid grounding accuracy is bounded by current multimodal backbones (e.g., ~74% on SCREENSPOT-PRO). Even sophisticated strategies like CHAIN-OF-GROUND (CoG) utilizing dual-layer GEMINI 3 PRO only achieve 75.3% accuracy, underscoring the difficulty of precise visual interaction.

F. Failure Detection and Recovery Implementation

The detection of operation failure in AVENIR-WEB is a multi-layered process occurring at four distinct stages, ensuring that tactical errors are caught and propagated for strategic reflection.

F.1. Trace Analysis of Detection Logic

The following trace illustrates the internal tool invocations used to determine how action success or failure is detected and recorded within the agent logic.

```
PROMPT: Failure Detection Trace

toolName: search_codebase
query: Find how action success or failure is detected and recorded in agent.py

toolName: search_codebase
query: Find where action['success'] or success variable is set in agent.py

toolName: view_files
filePath: SeeReAct/src/seeact/agent.py

toolName: search_by_regex
query: def.*action

toolName: search_by_regex
query: def\s+\w+
```

F.2. Multi-Layered Detection Mechanism

1. Execution-Level Detection First, the agent attempts to execute the command in the browser. If the automation fails directly, such as when an element isn't found, a timeout occurs, or a click is intercepted, the error is caught immediately and reported as a failure.

2. State-Change Verification Even if the command runs without error, the agent verifies if it actually changed the page by comparing snapshots before and after the action. It looks for changes in visible text, interactive elements, focus, the URL, scroll position, or modal popups. For actions like clicking or typing, if zero changes are detected, the action is marked as a failure even if no technical error occurred.

3. Action-Specific Checks Certain actions have deeper verification. For example, after typing, the agent reads the value back from the browser to confirm it matches the intended text. If it doesn't match, it is flagged as a failure.

4. Strategic Pattern Analysis Finally, the system analyzes the agent's history of actions to detect potential stalling. If several consecutive identical failures or a high failure rate occur in recent steps, it triggers a warning to reconsider the strategy or terminate the task.