WebRetriever: A Large-Scale Comprehensive Benchmark for Efficient Web Agent Evaluation

Wei Dong[1*], Tianyu Fu[1*], Zhe Yu[1], Hanning Wang[1], Anyang Su[1], Zhizhou Fang[1], Yuyang Chen[1], Shuo Wang[1], Minghui Wu[1], Ping Jiang[1], Zhen Lei[2,3], and Chenxu Zhao[1†]

[1] Mininglamp Technology
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] MAIS, Institute of Automation, Chinese Academy of Sciences
zhaochenxu@mininglamp.com
https://github.com/Mininglamp-AI/WebRetriever

Abstract. As web agents increasingly demonstrate capabilities in automated task execution, the development of robust evaluation frameworks for assessing their navigation and task completion performance has emerged as a critical research priority. However, existing benchmarks exhibit several fundamental limitations. First, they suffer from insufficient scale and limited domain diversity, thereby constraining comprehensive evaluation of cross-domain generalization. Second, prevailing LLM-as-Judge evaluation methodologies inadequately capture fine-grained interaction semantics, particularly regarding precise query formulation and filtering operations. Third, current benchmarks predominantly emphasize navigation success metrics while neglecting critical requirements for real-world deployment scenarios. To address these limitations, we introduce WebRetriever, a large-scale benchmark encompassing 800 websites and 1,550 tasks across diverse domains, including consumer, professional, and enterprise sectors, with comprehensive coverage of user intent patterns. We propose NavEval (Navigation Evaluation), a novel LLM-as-Judge framework that leverages rich interaction context beyond visual screenshots, achieving state-of-the-art alignment with human judgment across multiple evaluation datasets. Furthermore, we establish three complementary evaluation protocols that collectively provide holistic assessment of web agent capabilities: navigation proficiency, knowledge-assisted interaction, and end-to-end task completion with information extraction. Extensive experimental analysis reveals substantial performance disparities across evaluation protocols, demonstrating that navigation success alone serves as an insufficient predictor of real-world application effectiveness. WebRetriever delivers fine-grained diagnostic insights into agent capabilities and establishes a rigorous foundation for advancing web agent research and development.

Keywords: Web Agent · Dataset and Benchmark · LLM-as-a-Judge · Large Language Models · Vision Language Action

*Equal contribution. †Corresponding author.

1 Introduction

Recent advances in large language models have significantly improved language understanding, reasoning, and decision making, driving multimodal web agents to emerge as a key paradigm for automating complex online tasks [3, 14, 17, 19, 24, 35]. By jointly perceiving the visual content, structural information, and textual semantics of web pages, these agents can interact with real websites and have demonstrated strong potential in e commerce [7, 42], customer service [10], and enterprise automation [5, 11]. As model capability and system complexity continue to grow, accurate, reliable, and scalable evaluation of web agents has become a critical bottleneck for further progress.

Web agent evaluation benchmarks are divided into offline and online settings. Offline benchmarks [6, 9, 16, 21, 23, 28–31, 42, 46] provide controlled, reproducible environments but suffer from limited fidelity to real-world complexity, creating gaps between benchmark scores and actual performance. Recent online benchmarks [5, 11, 12, 38, 41, 43] enable live website interaction for better real-world characterization. However, they remain limited in website scale, domain coverage, and intent diversity, failing to systematically capture the breadth required for practical deployment, leading to biased assessments and overly optimistic performance conclusions. Beyond benchmark limitations, evaluation scalability presents an equally critical challenge. Quality assessment relies heavily on costly, unscalable human annotation [23, 34, 36, 46], While automated methods [18, 33, 37, 40, 41] have been explored, they often exhibit insufficient accuracy on complex query conditions. Moreover, existing evaluation protocols assess limited dimensions. Real-world agents must navigate, leverage external knowledge, and execute end-to-end tasks like extracting information from complex websites. Current benchmarks overlook these requirements, failing to assess agents' ability

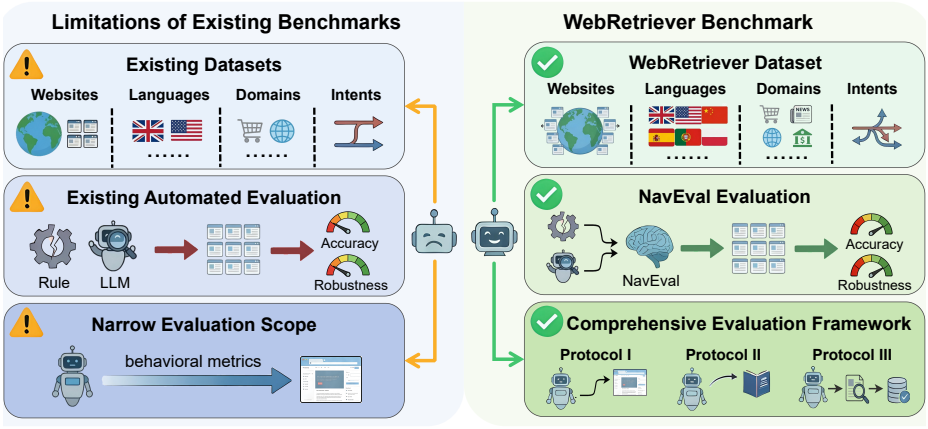


Fig. 1: Motivation for the WebRetriever benchmark. WebRetriever addresses key limitations of prior work from three aspects: dataset scale and diversity, automated evaluation reliability, and deployment-oriented evaluation protocols.

to utilize external knowledge and complete full task workflows. These limitations underscore the urgent need for comprehensive benchmarks, reliable automated evaluation, and deployment-oriented protocols.

To address these challenges, as illustrated in Fig. 1, we introduce WebRetriever, a large-scale benchmark for evaluating web agents in realistic online environments. It comprises 1,550 tasks on 800 websites, covering diverse domains and user intents, and provides substantially broader website scale, domain coverage, and intent diversity than existing benchmarks. By unifying evaluation across a wide spectrum of user-intent tasks, WebRetriever mitigates biases in prior benchmarks and enables a more representative and systematic assessment of web agents' overall capabilities. To enhance the scalability and accuracy of automated evaluation, we propose NavEval, an LLM-as-Judge method that integrates multi-source agent interaction information, achieving over 90% agreement with human judgments in large-scale experiments. We design three complementary evaluation protocols for comprehensive assessment: (1) Protocol I evaluates basic navigation ability to reach target pages; (2) Protocol II assesses navigation performance when provided with operational knowledge; (3) Protocol III measures end-to-end task completion by jointly evaluating navigation and information extraction, avoiding the limitation of equating page arrival with task success. Extensive experiments validate WebRetriever, NavEval, and our evaluation protocols, providing fine-grained capability insights and establishing a foundation for advancing web agent development.

In summary, the main contributions of this work are as follows:

1. **A large-scale, comprehensive benchmark for realistic web agent evaluation:** We curate 1,550 tasks across 800 real websites spanning diverse domains and user intents. Compared with prior benchmarks, WebRetriever provides unprecedented scale, diversity, and coverage, enabling more comprehensive and representative evaluation of web agents in realistic online environments.
2. **A general and high-precision automated evaluation method:** We propose NavEval, an automated evaluation method that attains approximately 90% human-level agreement in large-scale experiments, thereby enabling practical and reliable assessment of web agent performance at scale and in real-time.
3. **Comprehensive evaluation framework:** We propose three complementary evaluation protocols to systematically assess web agents, explicitly disentangling navigation success from answer correctness and characterizing behavioral reliability under injected operational knowledge, thereby providing diagnostic signals missing from prior benchmarks.

2 Related Work

2.1 Benchmarks for Web Agents

Web agent evaluation benchmarks are commonly divided into offline and online settings. Offline environments provide controlled and reproducible setups

for stable capability assessment. Early environments such as MiniWoB [36], MiniWoB++ [27], and CompWoB [15] provide controllable synthetic interactions, but the substantial domain gap from real websites limits their ability to reflect real-world generalization. Subsequent efforts, including WebShop [42], WebArena [46], VisualWebArena [23], ST-WebAgentBench [25] and Wonderbread [39], construct synthetic environments from real website snapshots, partially alleviating this issue, but their limited website coverage restricts robustness evaluation. Mind2Web [9] and WebLINX [31] adopt offline evaluation based on real HTML snapshots, enabling faster iteration and broader coverage; however, offline settings struggle to capture agents' exploration behavior and robustness to dynamic changes. Consequently, recent work has increasingly shifted toward online benchmarks. More recent benchmarks—including WorkArena [11], WorkArena++ [5], Mind2Web-Live [34], AssistantBench [43], WebVoyager [18], Bearcubs [37], and Online-Mind2Web [41]—support live evaluation and further mitigate the realism gap. However, existing online benchmarks remain limited in website scale, domain diversity, and coverage of user intent.

2.2 Automatic Evaluation for Web Agents

Offline environments provide stable, reproducible settings, while online evaluation faces challenges from dynamic websites [41]. SeeAct [44] pioneered human evaluation on live sites, but manual assessment does not scale with increasing task complexity. This motivates automated evaluation methods, which fall into rule-based [34, 46] and LLM-as-a-judge methods [4, 13, 26, 45]. Rule-based methods, such as Mind2Web-Live [34] and AssistantBench [43], suffer from brittleness to webpage dynamics and demand continuous maintenance, whereas LLM-as-judge approaches constrained to single-modality inputs demonstrate limited evaluation accuracy. Early approaches, such as Pan et al. [33], focus on final screenshot, providing only coarse-grained assessments and missing intermediate interactions. WebVoyager [18] extends evaluation to full trajectories, though at substantial token cost. AgentTrek [40] incorporates task descriptions and action traces to provide richer context, yet hallucinations still limit agreement with human judgments. WebJudge [41] generates key steps and extracts key screenshots for LLM judgment, but the quality of key step generation and matching precision directly constrain the final accuracy.

2.3 Evaluation Protocols for Web Agents

Current benchmarks predominantly assess web navigation success rates [5, 11, 23, 30, 34, 41, 43], step correctness [9, 21], answer accuracy [37], and execution efficiency [22] under settings where users provide only task objectives. However, in real-world scenarios, agents must leverage external knowledge and fulfill end-to-end user queries, encompassing both navigation and information extraction. Existing benchmarks fail to assess agents' navigation capabilities when augmented with external knowledge, nor can they evaluate whether end-to-end execution retrieves correct and complete data.

Table 1: Comparison between WebRetriever and related benchmarks. **Intent-Type**: task intent type (**G**: general, **P**: professional, **G&P**: both); **Setting**: the evaluation environment configuration; **Online**: whether online live connection evaluation is supported in real-world environments; **Interactive**: whether the environment allows interaction; **Websites**: number of websites; **Eval-Tasks**: number of evaluation tasks. Statistics are reported for the web-related evaluation subsets only.

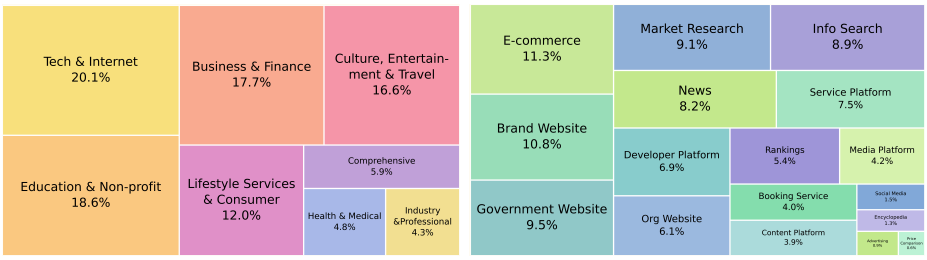
Benchmarks	#Intent-Type	#Setting	#Online	#Interactive	#Websites	#Eval-Tasks
MiniWoB [36]	G	Synthetic	✗	✓	100	100
MiniWoB++ [27]	G	Synthetic	✗	✓	100	100
WebArena [46]	G	Semi-real	✗	✓	4	812
VisualWebArena [23]	G	Semi-real	✗	✓	3	910
REAL [16]	G	Semi-real	✗	✓	11	112
WebLINX [31]	G	Real	✗	✗	155	1368
Mind2Web [9]	G	Real	✗	✗	137	1341
Mind2Web-Live [34]	G	Real	✓	✓	46	104
MMInA [38]	G	Real	✓	✓	14	1050
AssistantBench [43]	G	Real	✓	✓	258	214
WebVoyager [18]	G	Real	✓	✓	15	643
Bearcubs [37]	G	Real	✓	✓	108	111
Online-Mind2Web [41]	G	Real	✓	✓	136	300
WebShop [42]	P	Semi-real	✗	✓	1	500
ST-WebAgentBench [25]	P	Semi-real	✗	✓	3	222
Wonderbread [39]	P	Semi-real	✗	✓	4	598
WorkArena [11]	P	Real	✓	✓	5	33
WorkArena++ [5]	P	Real	✓	✓	5	682
OmniACT [21]	P	Real	✓	✓	27	736
WebRetriever (Ours)	**G&P**	**Real**	✓	✓	**800**	**1550**

3 WebRetriever

To address the limitations discussed in Sec. 2, this chapter presents WebRetriever's construction methodology and evaluation framework. We first describe our large-scale benchmark dataset composition. Next, we detail NavEval, our automated evaluation framework that integrates multi-modal interaction signals, applies rule-based filtering, and leverages LLMs for final assessment. Finally, we introduce three complementary evaluation protocols that emulate real-world deployment scenarios while providing multi-dimensional capability assessment.

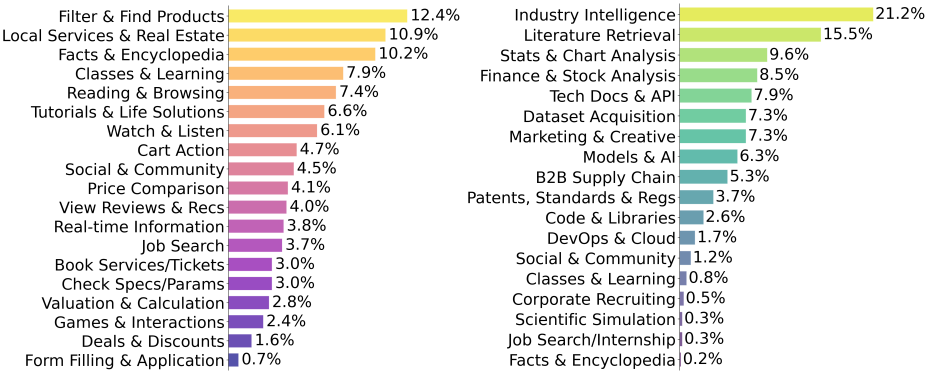
3.1 Dataset Construction

Given the limitations of existing benchmarks in website scale, domain coverage, and intent diversity, our WebRetriever design ensures comprehensive dataset diversity and coverage. As shown in Tab. 1, compared with prior benchmarks, WebRetriever encompasses 1,550 cross-industry tasks and over 800 carefully curated high-quality active websites. To capture the fundamental landscape of mainstream internet behavior, we leverage SimilarWeb traffic data as our baseline, targeting eight core sectors including Technology & Internet, Business &



(a) Domain Coverage **(b)** Website Type

Fig. 2: Visualization of WebRetriever's website coverage: (a) distribution across industry domains, (b) distribution of website types.



(a) General User Intents **(b)** Professional User Intents

Fig. 3: Visualization of user intent distributions in WebRetriever tasks: (a) distribution of general user intents, (b) distribution of professional user intents.

Finance, Education & Research, Culture, Entertainment & Travel, Life Services, Healthcare, Industrial Manufacturing, and Public Services, as illustrated in Fig. 2(a), WebRetriever spans eight core industry sectors closely related to everyday life and work. We select the top 30 sites by traffic within each sector, and additionally incorporate specialized vertical sites from authoritative internet research institutions (e.g., 199it) that maintain comprehensive data navigation repositories. This approach particularly strengthens our dataset's evaluation capabilities in business intelligence domains while ensuring broad website type distribution (Fig. 2(b)). At the task design level, a diverse team comprising industry experts, mid-level managers, senior data analysts, and university students contributes tasks for websites within their domains of expertise, grounded in authentic user intents. Building upon this foundation, we categorize task intents into two orthogonal dimensions: "general intents" and "professional intents" (Fig. 3). By combining these intent types with domain-specific "Fact Elements," we generate numerous tasks with distinct scenario characteristics. This approach evaluates not only whether agents can navigate websites, but crucially, whether

they understand the underlying business logic. To ensure rigorous executability and business authenticity for each generated task, we conduct multiple rounds of cross-validation during the task refinement phase, guided by three core principles: uniqueness, stability, and logical authenticity. While traditional datasets often feature flat task structures, WebRetriever tasks must align with expert operational intuition. For instance, in the financial domain, when a task involves low-frequency trading in over-the-counter markets, expert validation led us to eliminate unrealistic requirements for "intraday high-frequency data" queries, instead directing agents toward long-term quarterly or annual reports that reflect actual financial data disclosure patterns. Furthermore, to systematically characterize agent performance across varying difficulty levels, we categorize tasks based on the number of action steps $Step_n$ required by human annotators: tasks with $n < 6$ are defined as easy, $6 \leq n \leq 15$ as medium, and $n > 15$ as hard.

Considering the dynamic nature of web content, WebRetriever will be continuously maintained: when a task becomes obsolete or unreproducible due to page updates, it will be replaced with a new task of matching difficulty, ensuring comparability and long-term validity across dataset versions.

3.2 NavEval

Evaluating web agents in real-world environments is crucial yet inherently challenging due to the continuous evolution of web content. Human assessment is

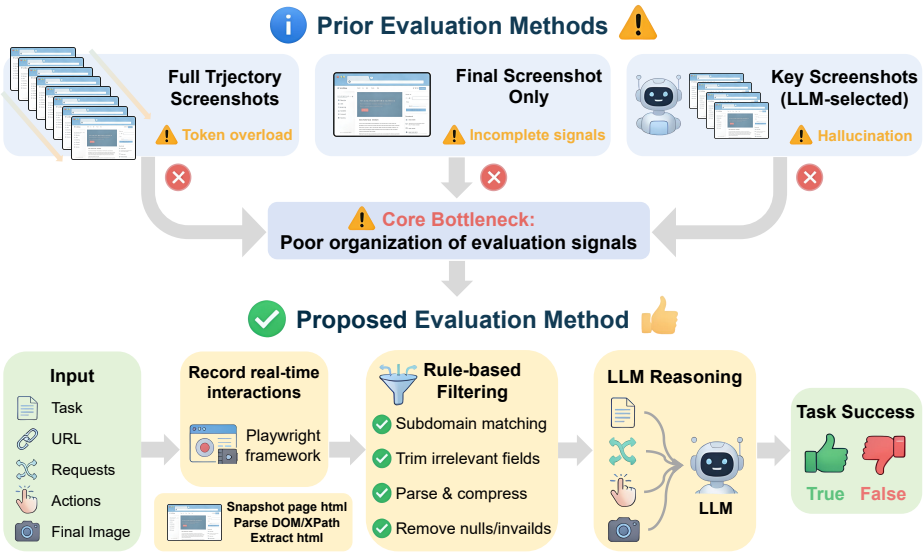


Fig. 4: Workflow of NavEval. Compared to existing methods, NavEval applies rule-based filtering to extract fine-grained intermediate signals, which are then jointly reasoned with the task description, action trajectory, and final screenshot by an LLM to determine task success, enabling robust evaluation with higher human agreement rates.

costly and difficult to scale, while existing automated approaches—whether rule-based or LLM-based—struggle to simultaneously achieve high accuracy and robustness in dynamic online evaluation settings. Current LLM-based evaluation pipelines exhibit limitations in organizing and utilizing evaluation signals. As illustrated in Fig. 4, existing methods employ three screenshot strategies: (1) full-trajectory screenshots, which provide comprehensive coverage but suffer from redundancy and high computational costs; (2) final screenshot only, which captures task completion status but misses critical intermediate process information; and (3) LLM-selected key screenshots, which reduce redundancy but lack sufficient accuracy in key step selection and matching. These limitations reveal the core challenge of automated evaluation: how to extract critical information to enhance LLM judgment accuracy.

To address these limitations, we propose NavEval (Fig. 4), which integrates more diverse information from agent-browser interactions. Beyond actions and final screenshots, NavEval leverages all navigation URLs and network requests generated during interactions. Through rule-based filtering to eliminate substantial noise, these signals are restructured into organized information that enhances LLM judgment accuracy. Formally, given an input consisting of a task description T, a website URL U, a sequence of web requests $R = (r_1, r_2, ..., r_n)$, a sequence of executed actions $A = (a_1, a_2, ..., a_n)$, and the final screenshot I, NavEval produces a binary classification output indicating task success or failure:

$$P = \text{NavEval}(T, U, R, A, I), \tag{1}$$

where $P \in \{\text{True}, \text{False}\}$ denotes whether the task was successfully completed. To robustly extract intermediate interaction information, we develop an online task evaluation framework based on Playwright, which enables real-time assessment of web agents in live environments. During task execution, the system records the web page screenshots, executed actions, and triggered requests at each step. Formally, the process can be expressed as:

$$\mathcal{R}'_l = \mathcal{F}_{rule}(R, U), \tag{2}$$

where $\mathcal{F}_{rule}(\cdot)$ denotes the rule-based filtering function, and $\mathcal{R}'_l$ represents the filtered request sequence. Specifically, NavEval first matches the triggered request set R against task URL U at the subdomain level to retain relevant requests. Rule-based filtering then removes task-irrelevant or random fields, normalizes structured payloads, and eliminates invalid entries, yielding a compact intermediate representation $\mathcal{R}'_l$. This representation captures critical query and filtering information that images alone cannot convey. Finally, NavEval integrates the task description, action sequence, filtered intermediate signals, and final screenshot to generate a comprehensive task completion assessment.

3.3 Evaluation Protocols

As discussed in Sec. 2.3, existing benchmarks evaluate agents within narrow scopes, overlooking their ability to leverage external knowledge or execute critical

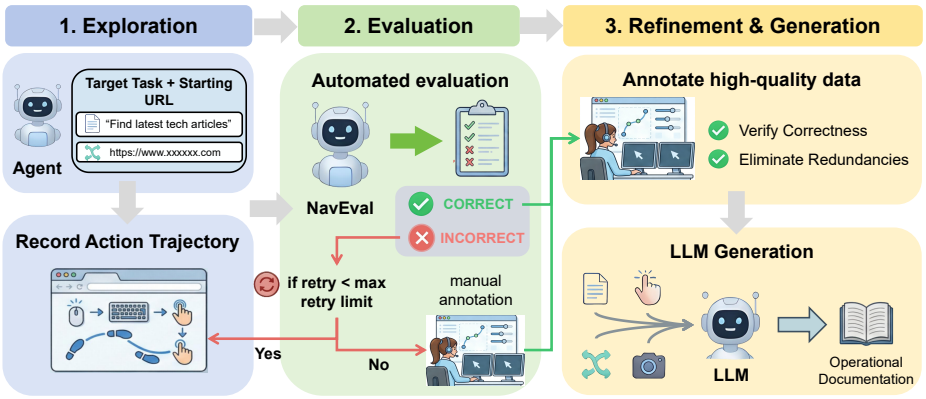


Fig. 5: Workflow of the semi-automated pipeline for constructing operational documentation in Protocol II. The process integrates automated exploration, evaluation, manual refinement, and LLM-based generation to produce high-quality operational documentation.

operations after reaching target pages in realistic scenarios. Consequently, strong benchmark performance fails to translate into practical capability, revealing a significant gap between evaluation metrics and real-world utility.

To comprehensively evaluate web agent capabilities, we design three protocols reflecting realistic deployment scenarios. **Protocol I** assesses fundamental navigation ability by measuring whether agents can reach target pages given user tasks. **Protocol II** evaluates navigation performance when agents are provided with operational documentation. For Protocol II, operational knowledge is consolidated into structured documentation through a closed-loop framework (Fig. 5). Web agents first generate action trajectories for target tasks, which NavEval evaluates automatically. Correct trajectories proceed to annotation, while incorrect ones are regenerated until reaching retry limits, after which human annotators refine them. The annotation platform further verifies accepted trajectories and streamlines redundant operations. Finally, LLMs generate operation manuals from these refined trajectories, completing the documentation pipeline.

While Protocols I and II focus on navigation, realistic web environments require capabilities beyond page navigation. Agents must demonstrate deep page understanding, multi-source information integration, and precise extraction capabilities critical for holistic evaluation. **Protocol III** mirrors realistic deployment by assessing whether agents can accurately retrieve target information across multiple modalities (text, documents, charts). Task construction follows three principles: (1) Authoritativeness: information from professional platforms ensures reliability; (2) Interaction Necessity: answers require browser-based interactions, not simple search; (3) Determinism: explicit queries yield unique, fact-based answers that remain stable for reproducible evaluation. Detailed specifications are in the supplementary material.

Based on these designs, we collect 1,000, 1,000, and 100 tasks for Protocols I, II, and III, respectively. Notably, Protocols I and II share 550 overlapping tasks, with the sole difference being the availability of operational documentation.

3.4 Evaluation Metrics

To rigorously evaluate WebRetriever and NavEval, we report two metrics using human annotations as reference: Success Rate (SR) measures web agent performance across the three protocols, while Human Agreement Rate (AR) evaluates NavEval's reliability. For task set $\mathcal{T}$, with human annotation $y_t \in \{0, 1\}$ and automated prediction $\hat{y}_t$:

$$\mathrm{SR} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \hat{y}_t. \tag{3}$$

$$\mathrm{AR} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \begin{cases} 1, & \text{if } \hat{y}_t = y_t, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Higher AR indicates stronger alignment with human judgments, reflecting evaluation fidelity.

4 Experiments and Results

4.1 Experimental Setup

To evaluate the effectiveness of the proposed benchmark, we develop an online task testing framework based on Playwright, enabling real-time assessment of web agents in live environments. Using this framework, we systematically evaluate a suite of state-of-the-art agents on WebRetriever across the three proposed evaluation protocols: Protocol I, Protocol II, and Protocol III. The evaluated agents include SeeAct [44], Browser-Use [47], UI-TARS-1.5 [35], Agent-E [1], as well as Claude-4.5 [2] and Gemini-2.5-Pro [8] in their Computer-Use mode. To more precisely assess agent navigation and task execution in realistic web settings, we adopt a controlled experimental setup. Each task starts from a predefined website entry URL, with search engine access explicitly restricted to prevent reliance on real-time retrieval shortcuts. This design ensures the evaluation faithfully reflects agents' intrinsic decision-making and interaction capabilities.

Based on the execution trajectories generated by the web agents, we further assess the reliability of the automated evaluation stage by systematically comparing NavEval with representative existing methods, including Autonomous Eval [33], AgentTrek Eval [40], WebVoyager [18], and WebJudge [41]. For a comprehensive comparison, we instantiate multiple LLM-as-a-Judge backbones, including GPT-4o [20], O4-mini [32], and Claude-4.5-Sonnet [2]. In our implementation, NavEval's judgment module adopts Claude-4.5-Sonnet to ensure stable semantic understanding and reasoning. Further implementation details and experimental settings are provided in the supplementary material.

Table 2: Task Success Rate (SR) of web agent trajectories on WebRetriever across the three proposed evaluation protocols, assessed using NavEval and human annotation, respectively. All values are reported as percentages (%).

Agent	Protocol I		Protocol II		Protocol III
	NavEval	Human	NavEval	Human	Human
SeeAct	11.3	9.2	18.9	17.1	6.0
Agent-E	12.9	11.6	22.5	20.4	9.0
UI-TARS-1.5	18.3	16.5	27.0	24.8	8.0
Browser-Use	26.6	24.0	35.2	31.6	11.0
Gemini-2.5-Pro (Computer-Use)	40.9	37.1	50.1	45.2	21.0
Claude-4.5 (Computer-Use)	31.3	28.1	40.1	36.3	16.0
Avg SR	**23.6**	**21.1**	**32.3**	**29.2**	**11.8**

4.2 Results and Analysis

As shown in Tab. 2, human evaluation indicates that web agents perform poorly on WebRetriever across all three protocols. This sharply contrasts with their high scores on existing benchmarks and highlights the comprehensiveness and realism of WebRetriever: unlike prior benchmarks, it exposes agents to challenges closer to real-world scenarios, including a large number of websites spanning multiple geographic regions and industry domains, and diverse user-intent tasks. Under Protocol I, which evaluates basic navigation skills, agents achieve an average human-assessed success rate of only 21.1%. The scale, domain diversity, geographic breadth, and variety of user intents in WebRetriever make even seemingly straightforward path-planning tasks highly demanding. Compared with existing benchmarks, WebRetriever encompasses both general intent and professional intent tasks, with the latter requiring agents to navigate domain-specific websites at a level comparable to skilled human operators. This dual coverage allows for realistic evaluation of both types of capabilities within a single framework, providing critical insights into agents' practical effectiveness. Building on this, Protocol II introduces operational documentation to simulate agents' ability to integrate external knowledge for automated web tasks in realistic scenarios. With access to this operational guidance, the average success rate increases to 29.2%, with Gemini-2.5-Pro in Computer-Use mode achieving the highest score of 45.2%. These results indicate that operational manuals can effectively support agents in navigating unfamiliar websites, reducing hallucinations and improving task completion. Finally, Protocol III further increases the challenge by requiring end-to-end task execution, combining navigation with downstream information extraction. As shown in Tab. 2, the average human-assessed success drops to 11.8%, indicating that even when agents reach the target pages, they struggle to accurately extract and integrate the required information. Most existing benchmarks do not evaluate this capability, which is critical for realistic end-to-end deployment. Collectively, these results demonstrate that current agents remain far from reliably handling practical end-to-end web tasks, revealing a significant gap between traditional benchmark performance and real-world applicability.

Table 3: Human Agreement Rate (AR) of web agent trajectories on WebRetriever across automated evaluation methods with different LLM-as-a-Judge models. Avg AR denotes the average human agreement rate. All values are reported as percentages (%).

Model	Auto-Eval	SeeAct	Agent-E	Browser-Use	Gemini-2.5-Pro (Computer-Use)	Claude-4.5 (Computer-Use)	Avg AR
GPT-4o	Autonomous Eval	69.8	67.0	66.6	65.4	65.9	66.9
	AgentTrek Eval	60.2	52.5	55.1	55.2	54.6	55.5
	WebVoyager	69.1	66.7	62.2	64.7	63.1	65.2
	WebJudge	71.9	70.5	72.6	70.8	70.3	71.2
O4-mini	Autonomous Eval	70.3	75.1	69.8	67.1	69.1	70.3
	WebVoyager	71.6	75.9	70.5	67.3	68.6	70.8
	WebJudge	76.5	77.6	75.9	72.5	73.1	75.1
Claude-4.5 -Sonnet	Autonomous Eval	79.5	76.5	75.2	74.1	74.4	75.9
	AgentTrek Eval	66.8	60.7	61.7	62.4	61.7	62.7
	WebVoyager	78.7	80.4	74.6	75.4	76.5	77.1
	WebJudge	80.9	81.1	81.7	80.7	80.8	81.0
Claude-4.5 -Sonnet	**NavEval (Ours)**	**92.2**	**91.3**	**90.9**	**91.4**	**90.1**	**91.2**

Given the substantial challenges revealed by WebRetriever, a reliable and fine-grained evaluation framework is crucial. As Tab. 2 shows, the task success rates evaluated by NavEval closely match those from human assessment across Protocols I and II, demonstrating both the accuracy of NavEval and its robustness across diverse task scenarios. As further illustrated in Tab. 3, NavEval consistently achieves human agreement rates above 90% across all web agents, outperforming existing methods—including Autonomous Eval, AgentTrek Eval, WebVoyager, and WebJudge—which exhibit wide variability, with average ARs ranging from 55% to 81%. This improvement stems not only from leveraging LLM reasoning via Claude-4.5-Sonnet, but also from incorporating fine-grained interaction data: beyond screenshots, NavEval analyzes the request sequences generated during task execution on web pages, allowing precise assessment of query execution and filtering. By combining these detailed trajectories with rule-based constraints, NavEval detects subtle behavioral differences that conventional automated methods often miss, closely approximating human judgment. These results establish NavEval as a reliable, discriminative, and fine-grained framework for evaluating web agents in realistic, instruction-guided scenarios. To further demonstrate the generalizability and effectiveness of NavEval, we conduct additional evaluation on Online-Mind2Web [41], a comprehensive benchmark comprising 300 high-quality real-world tasks across 136 popular websites from diverse domains. In addition to the previously used LLM-based judge models, we include WebJudge-7B, proposed in Online-Mind2Web, for a more comprehensive comparison. As shown in Tab. 4, NavEval consistently outperforms existing automated evaluation methods by a clear margin on this external benchmark. Prior approaches, including Autonomous Eval, AgentTrek Eval, WebVoyager, and WebJudge, exhibit noticeable variability across evaluator backbones and agent trajectories, with average agreement rates generally below 88%. In contrast, NavEval achieves a substantially higher Avg AR of 97%,

Table 4: Human Agreement Rate (AR) of web agent trajectories on Online-Mind2Web across automated evaluation methods with different LLM-as-a-Judge models. Avg AR denotes the average human agreement rate. All values are reported as percentages (%).

Model	Auto-Eval	SeeAct	Agent-E	Browser-Use	Avg AR
GPT-4o	Autonomous Eval	84.7	85.0	76.0	81.9
	AgentTrek Eval	73.0	64.3	63.3	66.9
	WebVoyager	–	75.3	71.3	–
	WebJudge	86.7	86.0	81.4	84.7
O4-mini	Autonomous Eval	79.7	85.7	86.0	83.8
	WebVoyager	–	80.3	79.0	–
	WebJudge	85.3	86.3	89.3	87.0
WebJudge-7B [41]	WebJudge	86.0	87.3	88.3	87.2
Claude-4.5 -Sonnet	**NavEval (Ours)**	**96.5**	**97.4**	**97.1**	**97.0**

demonstrating strong robustness and cross-benchmark stability. This advantage stems from NavEval's fine-grained design. By jointly leveraging rule-based constraints and structured interaction signals, particularly the request sequences generated during webpage execution, NavEval more accurately verifies query execution and filtering correctness, thereby reducing the ambiguity that commonly affects screenshot-only evaluators. Overall, the results on Online-Mind2Web confirm that NavEval generalizes effectively beyond WebRetriever and provides a reliable, high-fidelity automated evaluation framework for realistic web agent assessment.

4.3 Ablation Analysis

Table 5: Ablation study on operational documentation (Doc). Protocol I is originally defined without Doc, while Protocol II includes Doc. Reported values are task Success Rates (SR, %) of web agent trajectories on WebRetriever under different protocol settings. Settings indicate whether Doc is provided (w/ or w/o Doc).

Protocol	Setting	Gemini-2.5-Pro (Computer-Use)	Claude-4.5 (Computer-Use)
Protocol I	–	40.9	31.3
	w/ Doc	**49.2 (+8.3)**	**39.7 (+8.4)**
Protocol II	–	50.1	40.1
	w/o Doc	**41.4 (-8.7)**	**31.9 (-8.2)**

To evaluate the effect of operational documentation, we conduct an ablation study under Protocols I and II (Tab. 5). Adding documentation to Protocol I, which was originally defined without it, consistently improves performance, with success rates increasing by 8.3% for Gemini-2.5-Pro and 8.4% for Claude-4.5 in Computer-Use mode. Conversely, removing documentation from Protocol

Table 6: Ablation on end-to-end task completion in Protocol III, reporting task Success Rates (SR, %) of web agent trajectories on WebRetriever under different protocol settings. Settings indicate whether exact information extraction from webpages is required (w/ or w/o Extract).

Protocol	Setting	Gemini-2.5-Pro (Computer-Use)	Claude-4.5 (Computer-Use)
Protocol III	–	21.0	16.0
	w/o Extract	**43.0 (+22.0)**	**34.0 (+18.0)**

II, originally defined with it, leads to substantial drops of 8.7% and 8.2%, respectively. These results demonstrate that operational documentation provides crucial guidance on unfamiliar websites, helping agents reduce hallucinations and achieve more reliable task completion. At the same time, the moderate magnitude of improvement indicates that agents still face significant challenges in fully understanding and leveraging external knowledge, emphasizing the need for further advances in knowledge integration. We further conduct an ablation study on Protocol III, which requires end-to-end navigation and information extraction (see Tab. 6). Strikingly, even when agents successfully reach the target pages, their ability to extract the required information remains far from reliable: success rates for Gemini-2.5-Pro and Claude-4.5 in Computer-Use mode are only 43% and 34%, respectively. When evaluating full end-to-end task completion—combining navigation with information extraction—success rates drop almost by half, to 21% and 16%. This sharp decline reveals a critical blind spot in current agents: reaching the correct page does not guarantee successful task execution. Protocol III thus exposes the true difficulty of realistic end-to-end tasks, highlighting limitations largely overlooked by existing benchmarks and underscoring the importance of evaluating both navigation and actionable information processing capabilities. In addition to the protocol-level ablations, we provide further analyses of NavEval, including judge backbone self-bias and rule-based filtering ablations, in the supplementary material.

5 Conclusion

In this paper, we address the limitations of existing benchmarks for web agent evaluation, including insufficient scale, limited domain coverage, and lack of task diversity. To overcome these challenges, we introduce WebRetriever, a large-scale benchmark for realistic online evaluation, and NavEval, a scalable automated evaluation method that reduces human effort while maintaining high fidelity with human judgments. Building upon this foundation, we further propose three deployment-oriented evaluation protocols, namely Protocol I, Protocol II, and Protocol III, to systematically assess agents' core navigation abilities, ability to leverage external knowledge, and end-to-end task execution capabilities, respectively. Extensive experiments demonstrate that our benchmark, evaluation method, and protocols provide fine-grained insights into agent performance, re-

veal capability gaps overlooked by conventional evaluations, and establish a solid foundation for the development of more capable and reliable web agents in practical settings.

Acknowledgements

This work was supported by Mininglamp Technology. We thank the annotation and engineering teams for their contributions to dataset construction and the design of the evaluation framework. We also thank Han Lin and Yuting Liao for their valuable support throughout this project.

References

1. Abuelsaad, T., Akkil, D., Dey, P., Jagmohan, A., Vempaty, A., Kokku, R.: Agent-e: From autonomous web navigation to foundational design principles in agentic systems. arXiv preprint arXiv:2407.13032 (2024)
2. Anthropic: Introducing claude sonnet 4.5. Official Product Announcement (2025), https://www.anthropic.com/news/claude-sonnet-4-5
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bai, Y., Ying, J., Cao, Y., Lv, X., He, Y., Wang, X., Yu, J., Zeng, K., Xiao, Y., Lyu, H., et al.: Benchmarking foundation models with language-model-as-an-examiner. Advances in Neural Information Processing Systems **36**, 78142–78167 (2023)
5. Boisvert, L., Thakkar, M., Gasse, M., Caccia, M., De Chezelles, T.L., Cappart, Q., Chapados, N., Lacoste, A., Drouin, A.: Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. Advances in Neural Information Processing Systems **37**, 5996–6051 (2024)
6. Chen, Q., Pitawela, D., Zhao, C., Zhou, G., Chen, H.T., Wu, Q.: Webvln: Vision-and-language navigation on websites. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1165–1173 (2024)
7. Chen, S., Wiseman, S., Dhingra, B.: Chatshop: Interactive information seeking with language agents. arXiv preprint arXiv:2404.09911 (2024)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., Su, Y.: Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems **36**, 28091–28114 (2023)
10. Deng, Y., Zhang, X., Zhang, W., Yuan, Y., Ng, S.K., Chua, T.S.: On the multi-turn instruction following for conversational web agents. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8795–8812 (2024)
11. Drouin, A., Gasse, M., Caccia, M., Laradji, I.H., Del Verme, M., Marty, T., Boisvert, L., Thakkar, M., Cappart, Q., Vazquez, D., et al.: Workarena: How capable are web agents at solving common knowledge work tasks? arXiv preprint arXiv:2403.07718 (2024)

12. Fan, Y., Ma, X., Wu, R., Du, Y., Li, J., Gao, Z., Li, Q.: Videoagent: A memory-augmented multimodal agent for video understanding. In: European Conference on Computer Vision. pp. 75–92. Springer (2024)
13. Fernandes, P., Deutsch, D., Finkelstein, M., Riley, P., Martins, A.F., Neubig, G., Garg, A., Clark, J.H., Freitag, M., Firat, O.: The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. arXiv preprint arXiv:2308.07286 (2023)
14. Fu, T., Su, A., Zhao, C., Wang, H., Wu, M., Yu, Z., Hu, F., Shi, M., Dong, W., Wang, J., et al.: Mano technical report. arXiv preprint arXiv:2509.17336 (2025)
15. Furuta, H., Matsuo, Y., Faust, A., Gur, I.: Language model agents suffer from compositional generalization in web automation. In: NeurIPS 2023 Foundation Models for Decision Making Workshop (2023)
16. Garg, D., VanWeelden, S., Caples, D., Draguns, A., Ravi, N., Putta, P., Garg, N., Abraham, T., Lara, M., Lopez, F., et al.: Real: Benchmarking autonomous agents on deterministic simulations of real websites. arXiv preprint arXiv:2504.11543 (2025)
17. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
18. He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., Yu, D.: Webvoyager: Building an end-to-end web agent with large multimodal models. arXiv preprint arXiv:2401.13919 (2024)
19. Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., et al.: Cogagent: A visual language model for gui agents. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14281–14290 (2024)
20. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
21. Kapoor, R., Butala, Y.P., Russak, M., Koh, J.Y., Kamble, K., AlShikh, W., Salakhutdinov, R.: Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In: European Conference on Computer Vision. pp. 161–178. Springer (2024)
22. Kara, S., Faisal, F., Nath, S.: Waber: Evaluating reliability and efficiency of web agents with existing benchmarks. In: ICLR 2025 Workshop on Foundation Models in the Wild (2025)
23. Koh, J.Y., Lo, R., Jang, L., Duvvur, V., Lim, M.C., Huang, P.Y., Neubig, G., Zhou, S., Salakhutdinov, R., Fried, D.: Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. arXiv preprint arXiv:2401.13649 (2024)
24. Lai, H., Liu, X., Iong, I.L., Yao, S., Chen, Y., Shen, P., Yu, H., Zhang, H., Zhang, X., Dong, Y., et al.: Autowebglm: A large language model-based web navigating agent. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 5295–5306 (2024)
25. Levy, I., Wiesel, B., Marreed, S., Oved, A., Yaeli, A., Shlomov, S.: St-webagentbench: A benchmark for evaluating safety and trustworthiness in web agents. arXiv preprint arXiv:2410.06703 (2024)
26. Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpacaeval: An automatic evaluator of instruction-following models (2023)

27. Liu, E.Z., Guu, K., Pasupat, P., Shi, T., Liang, P.: Reinforcement learning on web interfaces using workflow-guided exploration. arXiv preprint arXiv:1802.08802 (2018)
28. Liu, J., Song, Y., Lin, B.Y., Lam, W., Neubig, G., Li, Y., Yue, X.: Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? arXiv preprint arXiv:2404.05955 (2024)
29. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al.: Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688 (2023)
30. Liu, X., Zhang, T., Gu, Y., Iong, I.L., Xu, Y., Song, X., Zhang, S., Lai, H., Liu, X., Zhao, H., et al.: Visualagentbench: Towards large multimodal models as visual foundation agents. arXiv preprint arXiv:2408.06327 (2024)
31. Lù, X.H., Kasner, Z., Reddy, S.: Weblinx: Real-world website navigation with multi-turn dialogue (2024)
32. OpenAI: o4-mini: Fast and cost-efficient reasoning model. Official Product Announcement (2025), https://openai.com/index/introducing-o3-and-o4-mini/
33. Pan, J., Zhang, Y., Tomlin, N., Zhou, Y., Levine, S., Suhr, A.: Autonomous evaluation and refinement of digital agents. arXiv preprint arXiv:2404.06474 (2024)
34. Pan, Y., Kong, D., Zhou, S., Cui, C., Leng, Y., Jiang, B., Liu, H., Shang, Y., Zhou, S., Wu, T., et al.: Webcanvas: Benchmarking web agents in online environments. arXiv preprint arXiv:2406.12373 (2024)
35. Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., et al.: Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326 (2025)
36. Shi, T., Karpathy, A., Fan, L., Hernandez, J., Liang, P.: World of bits: An open-domain platform for web-based agents. In: International Conference on Machine Learning. pp. 3135–3144. PMLR (2017)
37. Song, Y., Thai, K., Pham, C.M., Chang, Y., Nadaf, M., Iyyer, M.: Bearcubs: A benchmark for computer-using web agents. arXiv preprint arXiv:2503.07919 (2025)
38. Tian, S., Zhang, Z., Chen, L.Y., Liu, Z.: Mmina: Benchmarking multihop multimodal internet agents. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 13682–13697 (2025)
39. Wornow, M., Narayan, A., Viggiano, B., Khare, I., Verma, T., Thompson, T., Hernandez, M., Sundar, S., Trujillo, C., Chawla, K., et al.: Wonderbread: A benchmark for evaluating multimodal foundation models on business process management tasks. Advances in Neural Information Processing Systems **37**, 115963–116021 (2024)
40. Xu, Y., Lu, D., Shen, Z., Wang, J., Wang, Z., Mao, Y., Xiong, C., Yu, T.: Agenttrek: Agent trajectory synthesis via guiding replay with web tutorials. arXiv preprint arXiv:2412.09605 (2024)
41. Xue, T., Qi, W., Shi, T., Song, C.H., Gou, B., Song, D., Sun, H., Su, Y.: An illusion of progress? assessing the current state of web agents. In: Second Conference on Language Modeling (2025), https://openreview.net/forum?id=6jZi4HSs6o
42. Yao, S., Chen, H., Yang, J., Narasimhan, K.: Webshop: Towards scalable real-world web interaction with grounded language agents. In: ArXiv (preprint)
43. Yoran, O., Amouyal, S.J., Malaviya, C., Bogin, B., Press, O., Berant, J.: Assistantbench: Can web agents solve realistic and time-consuming tasks? arXiv preprint arXiv:2407.15711 (2024)
44. Zheng, B., Gou, B., Kil, J., Sun, H., Su, Y.: Gpt-4v (ision) is a generalist web agent, if grounded. arXiv preprint arXiv:2401.01614 (2024)

45. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems **36**, 46595–46623 (2023)
46. Zhou, S., Xu, F.F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Bisk, Y., Fried, D., Alon, U., et al.: Webarena: A realistic web environment for building autonomous agents. ICLR (2024)
47. Zunic, G.: Browser use = state of the art web agent. Blog Technical Report (2024), `https://browser-use.com/posts/sota-technical-report`