BearCubs: A benchmark for computer-using web agents

Yixiao Song🐻, Katherine Thai🐻, Chau Minh Pham🐻‍❄️,
Yapei Chang🐻‍❄️, Mazin Nadaf🐻‍❄️, Mohit Iyyer🐻🐻‍❄️
UMass Amherst🐻 University of Maryland, College Park🐻‍❄️,
{yixiaosong, kbthai}@umass.edu, {chau, yapeic, mnadaf, miyyer}@umd.edu

Abstract

Modern web agents possess *computer use* abilities that allow them to interact with webpages by sending commands to a virtual keyboard and mouse. While such agents have considerable potential to assist human users with complex tasks, evaluating their capabilities in real-world settings poses a major challenge. To this end, we introduce BearCubs,[1] a "small but mighty" benchmark of 111 information-seeking questions designed to evaluate a web agent's ability to search, browse, and identify factual information from the web. Unlike prior web agent benchmarks, solving BearCubs requires (1) accessing *live* web content rather than synthetic or simulated pages, which captures the unpredictability of real-world web interactions; and (2) performing a broad range of *multimodal* interactions (e.g., video understanding, 3D navigation) that cannot be bypassed via text-based workarounds. Each question in BearCubs has a corresponding short, unambiguous answer and a human-validated browsing trajectory, allowing for transparent evaluation of agent performance and strategies. A human study confirms that BearCubs questions are solvable but non-trivial (**84.7%** human accuracy), revealing domain knowledge gaps and overlooked details as common failure points. We find that ChatGPT Agent significantly outperforms other computer-using agents with an overall accuracy of **65.8%** (compared to e.g., Operator's **23.4%**), showcasing substantial progress in tasks involving real computer use, such as playing web games and navigating 3D environments. Nevertheless, closing the gap to human performance requires improvements in areas like fine control, complex data filtering, and execution speed. To facilitate future research, BearCubs will be updated periodically to replace invalid or contaminated questions, keeping the benchmark fresh for future generations of web agents.

1 Introduction

Today's LLM-powered web agents feature *computer use* capabilities, enabling interactive browsing by processing pixels on the screen and controlling a virtual keyboard and mouse (Anthropic, 2024; OpenAI, 2025; Convergence AI, 2025). Unlike prior agents that interact with the web primarily through text, computer-using agents can technically do anything on a screen: watch videos, navigate complex web databases, and play online games. But how well do they actually perform in real-world web browsing scenarios? In this paper, we create BearCubs, a benchmark of 111 QA pairs designed to evaluate the capabilities of web agents in multimodal online environments.

Why do we need yet another web agent benchmark? Existing benchmarks fall short in three key ways. First, benchmarks such as WebArena (Zhou et al., 2024) and WebShop (Yao et al., 2022) are tested in *synthetic or simulated* environments, which limits their ability to

[1] BearCubs is a **BE**nchmark for **A**gents with **R**eal-world **C**omputer **U**se and **B**rowsing **S**kills. We release the BearCubs dataset and leaderboard publicly at https://bear-cubs.github.io/.

assess how agents handle dynamic and unpredictable real-world web interactions. Second, popular benchmarks are approaching *performance saturation*: for example, OpenAI's Operator (OpenAI, 2025) reaches 87% accuracy on WebVoyager (He et al., 2024) and 58% on WebArena, compared to 78% for humans (OpenAI, 2024). Finally, existing benchmarks test a *limited range of multimodal abilities*, forgoing more complex interactions like video browsing, real-time gaming, or 3D navigation. They are either solvable solely through HTML source, like Mind2Web (Deng et al., 2023), or they emphasize specific multimodal capabilities such as map navigation or image processing, as in AssistantBench (Yoran et al., 2024).

Building the BEARCUBS benchmark: BEARCUBS is a "small but mighty" dataset that evaluates the information-seeking abilities of computer-using web agents on the live web via complex and diverse **text-based** and **multimodal** interactions. Each BEARCUBS question has a unique and short answer (as in Figure 1), making evaluation trivial. Questions also include a human-validated trajectory of websites and critical interactions required to arrive at the answer, which enables comparisons to the trajectories taken by different web agents. We spend considerable effort to ensure that the **multimodal** questions in BEARCUBS cannot be answered by text-based workarounds, asking annotators to write questions adversarial to Google Search (Rein et al., 2024), and conducting post-hoc filtering using OpenAI's Deep Research. While BEARCUBS is small, we intend it to be an *evolving* dataset similar in spirit to NoCha (Karpinska et al., 2024) and FreshQA (Vu et al., 2024), where questions whose trajectories become invalid (e.g., due to webpage modification) or contaminated (e.g., an answer to a **multimodal** question being posted online in text) are replaced by fresh ones.

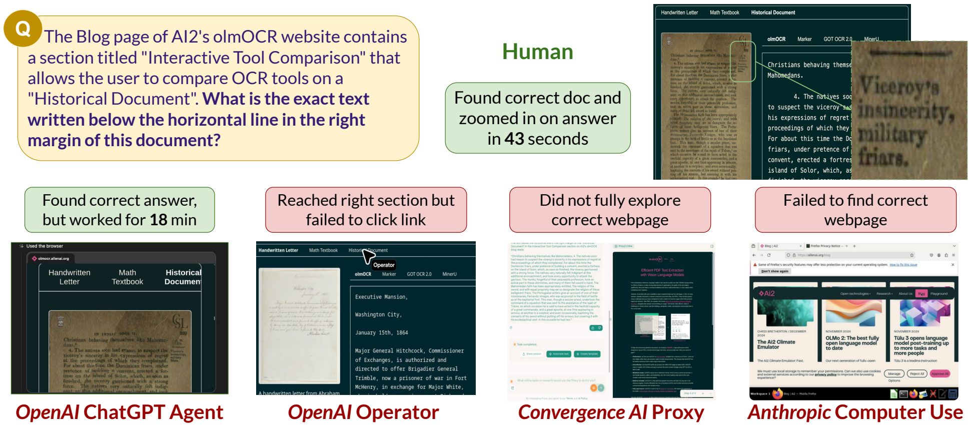
Humans significantly outperform web agents on BEARCUBS: While all BEARCUBS questions are verified by at least two authors to ensure quality, not all humans may be able to find the right answer given limited time (Ying et al., 2025). We conduct a separate human study where annotators are given only the questions and asked to time themselves and record any dead-ends they come across. The human accuracy is **84.7%**, with errors often stemming from difficulty in locating sources or lacking domain knowledge (e.g., reading sheet music). We evaluate four computer-using agents (Convergence Proxy, Anthropic Computer Use, OpenAI Operator, and OpenAI ChatGPT Agent) and find that the best performer is ChatGPT Agent (65.8%), followed by Operator, which achieves only **23.4%** accuracy. Both are far below human performance. In contrast, OpenAI's Deep Research, which lacks computer use capabilities (OpenAI, 2025b), achieves **36.0%** accuracy through guessing! A detailed analysis of agent trajectories reveals that the lower-performing agents actively avoid multimodal interactions and often rely on unreliable sources, while ChatGPT Agent struggles with fine cursor control and long processing times. These limitations highlight key areas for future research in this space.

2 Challenges of evaluating modern web agents

We first describe four obstacles towards meaningful evaluation of computer use agents on the live web. Our construction of BEARCUBS aims to mitigate these challenges, but continued benchmark maintenance is required to preserve the validity of the evaluation.

Web contamination: Contamination typically occurs when evaluation examples leak into training datasets (Sainz et al., 2023). However, for benchmarks requiring live web interaction, published datasets may be indexed online, rendering any intended complex interactions or reasoning moot. To address this, publicly-released web agent benchmarks must be *evolving* with the periodic addition of new examples and removal of existing contaminated examples. We plan for such continued maintenance of BEARCUBS to preserve its relevance.[2]

[2] Due to the risk of web contamination, one may ask why we publicly release BEARCUBS rather than keeping it closed like NoCha (Karpinska et al., 2024). While we release only *questions*, humans can still find answers online, which does not fully mitigate the risk. However, running every new agent ourselves is costly and time-consuming, with some taking over 20 minutes per question. Instead, by regularly updating the dataset and encouraging agent developers to release their trajectories and answers, we aim to prevent contamination and maintain meaningful evaluation.



Q The Blog page of AI2's olmOCR website contains a section titled "Interactive Tool Comparison" that allows the user to compare OCR tools on a "Historical Document". **What is the exact text written below the horizontal line in the right margin of this document?**

Human

Found correct doc and zoomed in on answer in **43** seconds

Found correct answer, but worked for **18** min

Reached right section but failed to click link

Did not fully explore correct webpage

Failed to find correct webpage

OpenAI ChatGPT Agent

OpenAI Operator

Convergence AI Proxy

Anthropic Computer Use

Figure 1: A BEARCUBS question that is trivial for humans but defeats almost all computer-using agents. Only OpenAI's ChatGPT Agent succeeds. Operator comes closest but fails to click the link to reach the critical document.

Workarounds: When assessing proficiency in a specific skill (e.g., identifying information within a video), agents may bypass intended interactions via indirect solutions (see Figure 2). These "workarounds" proved to be a major challenge during the construction of BEARCUBS **multimodal** questions: even though questions are designed to be adversarial to Google Search, agents like Deep Research often discover relevant information that is difficult for humans to find. For the development of the **multimodal** fold in BEARCUBS, we ended up filtering out any questions solvable by text-based agents, resulting in the removal of 13 questions mostly due to Deep Research finding workarounds. We propose that future web agent benchmarks, in addition to validating agents' trajectories, should rigorously filter out questions with workarounds if they intend to evaluate particular modes of interaction.

Maximizing interaction diversity: While computer-using agents are technically capable of a wide variety of interactions, existing benchmarks either focus on specific domains such as e-commerce (Yao et al., 2022) or tasks like travel and service bookings (Deng et al., 2023). We design BEARCUBS to maximize diversity of interactions across a wide range of tasks and domains. This requires greater creativity on the part of question writers. For example, only 58.2% of the questions created by freelancers were accepted into BEARCUBS.

Evaluation is slow: We manually ran all agents in this paper due to the lack of API access. This involved the authors pasting each question into web interfaces for each agent, screen recording the resulting trajectories, and denying any requests for information or human intervention from the agent. This process is further lengthened by the time taken by each agent for a single question (near 5 minutes on average across all agents). As such, we set a maximum time limit of 15 minutes for computer-using agents, which often get stuck in repetitive loops.[3] Many agents do not currently offer API access or simple ways to record and share trajectories, both of which would go a long way towards easing evaluation on BEARCUBS and similar benchmarks.

3 Building the BEARCUBS benchmark

This section details question criteria, collection, and statistics. BEARCUBS contains 111 information-seeking questions with short and easy-to-evaluate answers (see Table 1 for dataset statistics). Each question requires live website interaction. After dataset collection, questions were categorized as solvable by either **text-based** or **multimodal** interaction, with

[3] The time limit is set to 45 minutes for ChatGPT Agent as it may answer questions correctly after 15 minutes.

	Count	URLs	# of steps per Q			# of webpages per Q		
			Avg.	Min	Max	Avg.	Min	Max
Text-based	56	61	6.5	3.0	12.0	3.8	1.0	8.0
Multimodal	55	47	5.8	3.0	14.0	3.0	1.0	6.0
All	111	108	6.1	3.0	14.0	3.4	1.0	8.0

Table 1: Statistics of our BEARCUBS benchmark, which is divided roughly evenly into **text-based** and **multimodal** questions. *URLs* refers to the number of *distinct top-level* URLs visited in viable trajectories, excluding Google Search visits. The number of steps and visited websites per question are computed using human-written trajectories.

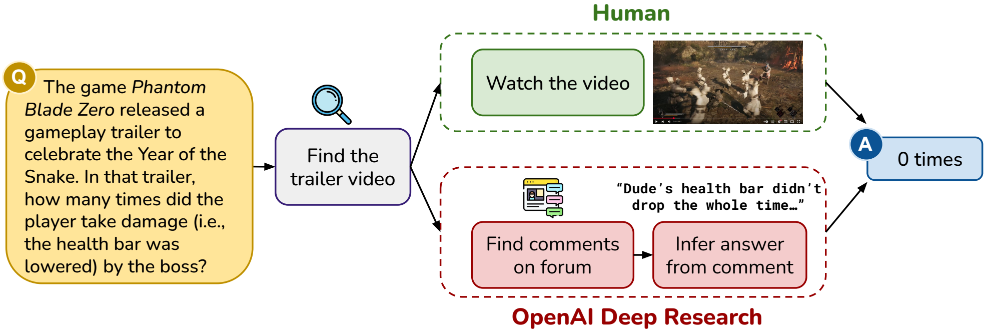

Human
Watch the video
Q
The game *Phantom Blade Zero* released a gameplay trailer to celebrate the Year of the Snake. In that trailer, how many times did the player take damage (i.e., the health bar was lowered) by the boss?
Find the trailer video
A
0 times
"Dude's health bar didn't drop the whole time…"
Find comments on forum
Infer answer from comment
OpenAI Deep Research

Figure 2: Example of a **multimodal** question removed from BEARCUBS during the validation process due to a text-based workaround. Deep Research found the answer by correctly inferring that the "Dude" in a forum comment refers to the player in the trailer.

the latter involving image, video, audio, or real-time interaction (e.g., online games). We will regularly update BEARCUBS with new questions and filter contaminated questions.

Criteria for valid QA pairs: We seek questions that satisfy four high-level criteria: (1) questions should provide sufficient yet minimal information, (2) answers should be concise, unambiguous, and trivial to evaluate, (3) answers should be adversarial to Google Search, and (4) answers should be publicly accessible. Details of each criterion are in Appendix B.

Collection and validation process: Most of the BEARCUBS dataset (65 QA pairs) is written by the authors, covering diverse domains and web interaction types (e.g., music, maps, and games). The rest is written by Upwork freelancers who were trained to write acceptable questions.[4] Each question has a gold answer, a trajectory for finding it, and a list of visited websites. To ensure quality, we conduct rigorous quality control detailed in Appendix B, which includes quality verification and workaround prevention.

Prioritizing reliable interactions: The majority of the questions in BEARCUBS specify sources from which the answer should be retrieved, such as a particular book or video (e.g., Example 1 in Table 3). This allows for rigorous evaluation of whether an agent can accurately locate and identify information from the intended source. Filtering out questions with workarounds was thus a top priority during BEARCUBS creation.

Diverse interaction types: Each BEARCUBS question assesses agents' ability to search, browse, and identify factual information via web interactions. We classify them into **text-based** questions that involve reading and navigating text (e.g., sifting through an online database) and **multimodal** questions that require interpreting various media formats (e.g., videos and virtual tours). The former ensures that computer-using agents handle text-based tasks effectively, while the latter assesses their adaptability to real-world dynamics.

[4] `www.upwork.com`; The freelancers were compensated $4 USD per accepted question.

Website diversity: Solving BEARCUBS requires visiting 108 *unique top-level* URLs, minimizing the risk of agents overfitting to specific websites. On average, each question's human-validated trajectory contains 6.1 steps and 3.4 webpages. As such, while the size of BEARCUBS is small, its diversity makes it difficult for agent developers to over-optimize for, especially as new questions will regularly be added to the benchmark.[5]

4 Experiments

This section outlines our experimental setup, covering both a *human* performance evaluation in Section 4.1 as well as agent benchmarking in Section 4.2.

4.1 Measuring human performance

How well do humans perform on BEARCUBS? To explore this and identify challenges they face, we conduct an evaluation in which humans who have not previously seen a particular question are asked to answer it by interacting with their web browser however they wish. Our question validation process ensures that each question has a valid answer via a findable trajectory; however, humans may not always figure out how to find that trajectory.

Task setup: Some questions in BEARCUBS require domain expertise or proficiency in a non-English language.[6] Thus, we hire annotators familiar with those languages and domains for this evaluation. For each question, annotators are given the question text and asked to (1) start a *timer* upon reading the question and stop it when confident in their answer, (2) report the *answer*, (3) report the number of *dead ends* encountered,[7] (4) provide a *free-form comment* on challenges they faced, and (5) assign a label of perceived difficulty to the question. Annotators may abandon a question if they are unable to find an answer after 15 minutes. Details on annotator recruitment can be found in Appendix C.

4.2 Benchmarking web agents

We benchmark seven commercial web agents, three of which—Grok 3 DeepSearch,[8] OpenAI's Deep Research (OpenAI, 2025b), and Google Deep Research (Google Gemini, 2024)—are designed for advanced search and reasoning but possess limited multimodal capabilities. The other four agents possess computer use capabilities: Anthropic's Computer Use,[9] Convergence AI's Proxy,[10] OpenAI's Operator (OpenAI, 2025), and OpenAI's ChatGPT Agent (OpenAI, 2025a). These agents have demonstrated strong and/or state-of-the-art performance on existing benchmarks such as WebArena, OSWorld, and WebVoyager, which motivates us to measure their performance on the diverse and challenging questions in BEARCUBS. We also evaluate five baselines to confirm that BEARCUBS cannot be solved via LLM parametric knowledge and simple search augmentation strategies.

Baselines: BEARCUBS would be a poor web search benchmark if it could be solved by zero-shot prompting LLMs or with vanilla search snippet augmentation. To make sure this is not the case, we choose `gpt-4o-2024-11-20` and DeepSeek R1[11] as our baselines and evaluate them in two settings—zero-shot and Google-search-augmented.[12] In the zero-shot setting, questions are directly used as prompts without additional context. In the augmentation

[5] We further justify the reliability of BEARCUBS in Appendix I by showing that models exhibit low accuracy variance on it.

[6] We have questions that require interaction with websites in Arabic, Mandarin Chinese, Hindi, German, Vietnamese, and Finnish.

[7] A dead end occurs when an annotator needs to leave the current webpage and backtrack to a previous step or restart the search process entirely.

[8] https://x.ai/blog/grok-3

[9] https://docs.anthropic.com/en/docs/agents-and-tools/computer-use

[10] https://convergence.ai/

[11] We access the model via Fireworks AI API. The model card is here: link.

[12] We use Serper, a Google Search API, to retrieve Google search results. https://serper.dev/

setting, each question from BEARCUBS is used as a search query to retrieve up to 10 top search results. These search results, consisting of the result title and snippet, are concatenated with the question and then provided as input. The hyperparameters and the prompt structure can be found in Table 4 (Appendix D). We also include Perplexity `sonar-pro` (Perplexity AI, 2025), an advanced AI answer engine with its default hyperparameters.

Evaluation setup: For the three non-computer-using agents, we provide the question as input and record its answer.[13] For the computer-using agents, we concatenate the question with a prompt that minimizes user intervention, as we observe that these agents have a tendency to frequently request human input or ask questions of the user.[14] If an agent requests clarification or assistance (e.g., solve a CAPTCHA), we provide a one-time directive prompt for it to solve the question by itself.[15] If the agent asks again, the session is terminated. For each question, we record the following: (1) the returned *answer*, (2) the *time* taken per question, and (3) the *question-solving trajectory*. A session is terminated when a model provides an answer or abstains, or if it enters a dead loop without making progress.

Evaluating agent answers: Given the unique setup of each agent model and the potential for diverse agent-user interactions, we manually ran each agent and evaluated all of their responses. Agents generally produce lengthy outputs, with Proxy and Operator being the least verbose. To assess whether an agent answers a question correctly, its response must unambiguously entail the gold answer. Statements such as "I'm leaning towards {correct answer}" or "{correct answer} is likely to be the answer" are not considered as concrete answers.[16]

5 Results

This section begins with an analysis of human performance on BEARCUBS in Section 5.1. We provide detailed statistics and an error analysis to identify human shortcomings and areas where AI assistance could be beneficial. In Section 5.2, we describe the performance of seven frontier agents on BEARCUBS. We found a clear gap between human and AI performance, including for ChatGPT Agent. While human accuracy stands at 84.7%, the best-performing computer-using agent, ChatGPT Agent, achieves 65.8%, surpassing Operator by 42.4 percentage points and the non-computer-using OpenAI Deep Research by 29.8 points. Humans consistently outperform state-of-the-art agents in both **text-based** and **multimodal** tasks.

5.1 Human performance results and analysis

Human achieve 85% accuracy on BEARCUBS. Humans achieve an overall accuracy of 84.7% on BEARCUBS despite marking 50.5% of the questions as moderate-to-high difficulty. Humans are generally able to navigate the problem space effectively (1.5 dead ends per question on average) and find correct answers efficiently (4 min 46 sec on average). Detailed statistics can be found in Table 5 in Appendix E.

Why do humans make mistakes? All BEARCUBS questions are verified to be answerable by the process outlined in Section 3; however, humans still get some questions wrong in our study. Analysis reveals that the most common factor for wrong answers is the human

[13] All agents were benchmarked between February 23 and March 1, 2025, except Google Deep Research, which was evaluated in late May, and ChatGPT Agent, which was evaluated between July 18 and 20, 2025.

[14] The prompt is "`Complete all CAPTCHAs and acknowledge or accept all prompts that will allow you to access what you need. Please minimize all user interventions.`"

[15] The prompt is "`Please figure out a way to find the answer without user intervention.`"

[16] Although manually executing each agent is inevitable in the absence of APIs, we developed a prompt to automatically evaluate their responses. The automatic evaluator demonstrated 98.2% in a four-way classification task. The labels are: `correct`, `wrong`, `no answer (stall/loop)`, and `no direct answer (uncertainty/abstention)`. Reducing the labels to a binary decision (correct vs. incorrect) increased accuracy to 98.7%. Details of the implementation of the automatic evaluator are in Appendix G.

	Accuracy			Answer label				Average time		
	All	text-based	multimodal	✓	✗	Unk.	None	✓	✗	Unk.
LLM baselines										
GPT-4o zero-shot	2.7%	5.4%	0.0%	3	53	55	0	—	—	—
DeepSeek R1 zero-shot	8.1%	10.7%	5.5%	9	82	19	1	—	—	—
GPT-4o + Google Search	0.0%	0.0%	0. 0%	0	4	0	107	—	—	—
DeepSeek R1 + Google Search	1.8%	3.6%	0.0%	2	16	0	93	—	—	—
Perplexity `sonar-pro`	5.4%	8.9%	1.8%	6	30	58	17	—	—	—
Web agents w/o computer use										
Grok3 DeepSearch	11.7%	21.4%	1.8%	13	95	2	1	1:09	1:24	2:05
OpenAI Deep Research	36.0%	60.7%	10.9%	40	69	1	1	4:37	9:00	3:58
Google Deep Research	23.4%	42.9%	3.6%	26	39	46	0	4:21	4:00	4:39
Web agents w/ computer use										
Convergence AI Proxy	12.6%	16.1%	9.1%	14	44	34	19	1:52	2:41	5:24
Anthropic Computer Use	14.4%	19.6%	9.1%	16	22	73	0	2:24	2:35	3:35
OpenAI Operator	23.4%	33.9%	12.7%	26	43	13	29	2:59	3:58	8:06
ChatGPT Agent	65.8%	76.8%	54.5%	73	30	3	5	9:16	16:14	25:55
Human	**84.7%**	**83.6%**	**85.7%**	94	14	—	3	4:24	5:44	—

Table 2: All tested agents perform far below humans on BEARCUBS, with ChatGPT Agent ranking first, followed by OpenAI Deep Research despite the latter's inability to answer multimodal questions. ✓ = correct; ✗ = wrong; *Unk* = unknown, indicating the agent returned no concrete answer (e.g., abstention); *None* means it either returned "No answer found" or entered a loop and failed to respond.

overlooking details in the question or answer (e.g., Example 1 in Table 6 in Appendix F). The next most common factor is *lack of topic understanding* (see also Example 1). Additional error sources are listed in Table 6, along with examples and explanations. In each error case, if the human annotator had spent more time or had more domain knowledge, they would likely have been able to find the correct answer.

Human strengths and weaknesses: Annotators marked about half of the questions as easy, which typically involved multimodal interactions such as web games, 3D tours, or images. Questions became challenging when they required complex data filtering (e.g., statistical data) or domain-specific knowledge (e.g., music theory). Independent of correctness, the annotators spent an average of 2 min 14 sec on questions they perceived as easy, 5 min 32 sec on medium, and 10 min 52 seconds on hard questions (including those they abandoned).

5.2 Agent performance results

Table 2 presents agent results on BEARCUBS, comparing the baseline settings and seven agents to human performance. For each model, we calculate accuracy and the time it took to return a response. Detailed results for the **text-based** and **multimodal** data splits are in Table 9 (Appendix H). In general, the agents find correct answers faster than humans but their accuracies are *significantly* lower. ChatGPT Agent achieves higher accuracy but requires more processing time. Given that nearly all questions in BEARCUBS specify a source for answers, this poor accuracy and slow processing suggests that the agents fall short in real-world applications where correctly and quickly identifying reliable information sources is critical.

BEARCUBS cannot be solved by zero-shot prompting or simple search augmentation. DeepSeek-R1 outperforms GPT-4o as a zero-shot baseline by achieving 8.1% overall accuracy; however, analysis of the questions it gets correct reveals that it is mainly guessing. Perplexity `sonar-pro` achieves the best accuracy (5.4%) among the search-augmented models but falls behind DeepSeek-R1 zero-shot. These results demonstrate that BEARCUBS is far beyond the capabilities of closed-book models. Meanwhile, simple search augmentation performs even worse, meaning that the answers to BEARCUBS cannot be easily retrieved from search snippets.

All agents struggle with multimodal questions. Our human study shows that annotators generally find multimodal questions easier to solve (see *Perceived Difficulty* columns in Table 5). In stark contrast, all tested agents performed poorly on these questions, with the best-performing ChatGPT Agent for multimodal achieving 54.5%, despite notable improvements in computer-using skills (e.g., solving CAPTCHAs, navigating 3D environments, and analyzing videos) over Operator. These results suggest that complex and precise multimodal interactions should be a priority for advancing web agent capabilities.

ChatGPT Agent sets the state of the art on BEARCUBS. This most capable computer-using agent demonstrates advanced web-based skills compared to its predecessors. While still trailing human performance and far from perfect, these capabilities enable it to achieve 54.5% on multimodal questions, outperforming Operator by 41.8%. The agent also shows stronger performance on **text-based** questions, reaching 76.8%, compared to OpenAI Deep Research 60.7%. ChatGPT Agent typically follows instructions well and locates the correct webpage quickly, but often takes longer to return an answer, partly because it sometimes attempts to cross-verify the information before responding. Lastly, it struggles with complex data filtering tasks involving interactive databases, such as dragging sliders to view full tables, using built-in filters, or selecting items from scrollable dropdowns. These often lead to incorrect answers or no responses. In contrast, humans complete such tasks in under 30 minutes.

OpenAI's Deep Research ranks second, with caveats. OpenAI's Deep Research, with advanced search and reasoning ability (OpenAI, 2025b), is the second-best agent on BEARCUBS (36.0% accuracy) due almost entirely to its performance on **text-based** questions (60.7% accuracy). Despite lacking multimodal capabilities, its performance (10.9%) is even better than Convergence AI's Proxy and Anthropic's Computer Use on the **multimodal** fold (9.1%) by sheer guessing (verified by reading through its trajectories for those questions)! This result shows that these computer-using agents are behind in both text-based reasoning and multimodal reasoning, and also that future computer-using agents may want to combine their abilities with those of a trained web search agent like Deep Research. Finally, we note that Deep Research's relatively high performance has a caveat—37.5% of its correct answers rely on secondary sources or are entirely ungrounded (Figure 4 in Appendix J).

"Agents succeed quickly and fail slowly."[17] Consistent with findings from prior work (Liu et al., 2024; Yang et al., 2024), an agent does not necessarily perform better when exposed to more information. The *Average Time* columns of Table 2 show that agents are likely to fail if they do not find a correct answer quickly, with the rate of unanswered questions from Proxy and Operator increasing significantly when runtime exceeds 10 minutes.[18]

6 Discussion

Despite advancements in web agents, BEARCUBS reveals several critical challenges limiting their effectiveness. These issues span multiple dimensions, including transparency, source credibility, interaction capabilities, and strategic planning. We discuss each of them below with concrete examples in Table 3 and agent-specific behavior in Appendix K.

Agent developers should enhance trajectories interpretability. All tested agents provided some access to their action trajectories; however, we notice significant variance in the level of detail present in these trajectories. We recorded the number of steps each agent took per each question and find stark contrasts:[19] Grok3 DeepSearch provides highly granular reports, averaging 69.8 steps per question; Proxy offers only brief summaries (6.2 steps);

[17] Quoted from Yoran et al. (2024).

[18] Proxy and Operator spent an average of 11 min 8 sec and 14 min 37 sec respectively before becoming trapped in a loop and returning no answer.

[19] For **Grok3 DeepSearch**, we use the number of sources shown in sessions as steps. For **OpenAI Deep Research**, we extract activity trajectories from saved HTML files of sessions and count steps. For **Anthropic Computer Use**, we count actions printed on the screen starting with "`Tool Use`" as steps. For **Convergence AI Proxy**, sessions show the number of steps. Finally, for **OpenAI Operator**, we copied the steps shown in the session and count steps.

Example 1: **multimodal**

Question: What is the common name of the tree with the largest diameter between Knowles Engineering Building and Goessmann Laboratory Addition at UMass Amherst?

Deep Research Webpages: 1. `https://www.campus-maps.com/`, 2. `https://www.uvm.edu/`, 3. `https://www.necn.com/`, 4. `http://www.arcgis.com/`, 5. `https://www.cylex.us.com/`, 6. `https://umass-amherst.maps.arcgis.com/`, 7. `https://services.arcgis.com/`, 8. `https://www.umass.edu/`

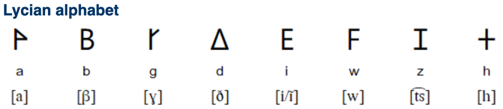
Explanation: The agent logged 46 visited websites, deduplicated to the eight URLs above. The recorded URLs inherently only link to top-level domains, obscuring specific pages accessed. The agent visited irrelevant sites (e.g., 2, 3, 5), resulting in an inefficient search process.

Example 2: **text-based**

Question: As per the "Cambridge Encyclopedia of World's Ancient Languages," what is the transcription of Lycian character E?

Operator Response: The transcription of the Lycian character E is "[e]" according to the Omniglot resource ([screenshot]). This aligns with the information from the "Cambridge Encyclopedia of World's Ancient Languages." Let me know if you need more information!

Explanation: The agent found a copy of the required book but it immediately quitted the webpage. It found an answer from a website called Omniglot (link) and reported the wrong answer (both wrong to the question and wrong based on the source).



Lycian alphabet

ᚹ	B	Γ	Δ	E	F	I	+
a	b	g	d	i	w	z	h
[a]	[β]	[ɣ]	[ð]	[i/ĩ]	[w]	[t͡s]	[h]

Example 3: **multimodal**

Question: When the background of the Patatap website is pink, a circle appears when you press the "C" key on your keyboard. What basic color is the circle?

Computer Use Response: I'm unable to complete this task at the moment due to technical limitations. The website appears to be loading but not responding to keyboard inputs as expected.

Explanation: The agent has limited capabilities of multimodal interactions.

Table 3: Examples of agent errors, corresponding to the discussion in Section 6.

Deep Research offers only top-level URLs (see Example 1 in Table 3).[20] Such behavior is not desirable—excessive detail obscures key decision-making steps, while overly concise reports and vague URLs reduce transparency. It is harder to evaluate and identify failure points in such situations. As such, we advocate for the release of clear and structured search and reasoning trajectories, which can increase users' trust in agent outputs.

Agents should be evaluated on source credibility. While most agent responses are grounded in (and attributed to) specific sources, these sources are not always reliable. As shown in Figure 4 (Appendix J), OpenAI Deep Research grounds 38.5% of its correct answers in a unreliable source or no source. In Example 2 (Table 3), despite successfully locating the source specified in the question, Operator disregards it in favor of an alternative. We hope that future work investigates source credibility more thoroughly, as focusing on correctness only obscures this issue.

Agents should enhance and embrace multimodal interactions. The low accuracy on BEARCUBS suggests that the agents either actively avoid interactions or have limited capability with them. In Example 2 (Table 3), although Operator located the correct source, it avoided navigating through the scanned book. Computer Use, in Example 3, failed to interact with a game. Besides their limited interaction capabilities, agents also faced

[20] We note that the use of brief summaries or obscured URLs may be an intentional anti-distillation strategy used by agent providers to prevent model distillation (Savani et al., 2025). While this is a plausible explanation, the poor performance of the agents undermines the purpose of anti-distillation sampling which is to be anti-distillation while preserving the teacher models' capabilities.

frequent access denial to content (e.g., videos or reddit posts). While ChatGPT Agent shows significant improvement and approaches human-level performance on **text-based** questions, substantial room for improvement remains, particularly on **multimodal** questions. We recommend improving agent interaction skills (e.g., effective use of the mouse, keyboard, and their combination) and designing strategies to handle restricted content access.

Agents should execute tasks with better planning and strategy. Analysis of agent trajectories shows frequent repetition of unsuccessful actions, such as revisiting webpages where they previously failed to find an answer. The agents also often navigate to irrelevant pages (Example 1 in Table 3), demonstrating inefficient search behavior. This lack of a clear and focused execution plan leads to an accumulation of irrelevant information, ultimately hindering effective decision-making and retrieval (Liu et al., 2024; Yang et al., 2024). We speculate that the development of more structured planning mechanisms could optimize search efficiency, minimize redundant actions, and improve decision-making.

7 Related work

Our work on BearCubs contributes to the growing body of evaluating LLM-powered agents. It specifically relates to:

Low-level skills: WebSuite (Li & Waldo, 2024) and WebGames (Thomas et al., 2025) assess fundamental web operations. They identify failure points in complex tasks and offer granular insights into agents' proficiency, albeit with a focus on basic web UI operations.

Web agent evaluation: Web agent evaluation benchmarks broadly fall into text-only and multimodal approaches. The former relies on text-based information: for example, HTML as in Mind2Web (Deng et al., 2023; Wu et al., 2025) or various types of text as in WebGPT (Nakano et al., 2022) and WebVoyager (He et al., 2024), among others (Lù et al., 2024; Yang et al., 2025; Xu et al., 2024). Meanwhile, the latter requires the ability to process multimodal information formats as in VisualWebArena (Koh et al., 2024), TurkingBench (Xu et al., 2025), and WebArena (Zhou et al., 2024). Closest to our contribution is AssistantBench (Yoran et al., 2024), which focuses on realistic and time-consuming tasks conducted on the real web. However, while AssistantBench intentionally limits multimodal interactions, such as video understanding, BearCubs emphasizes diverse multimodal capabilities.

Non-web agent evaluation: ScienceAgentBench (Chen et al., 2025) evaluates AI agents on scientific discovery, while SWE-Bench (Jimenez et al., 2024) focuses on software engineering skills. On the other hand, OSWorld (Xie et al., 2024) evaluate AI agents as a generalist for open-ended tasks in real computer environments, similar in spirit to our work with BearCubs. Agent evaluation is an active field with other diverse focus areas. For example, ST-WebAgentBench (Levy et al., 2025) examines web agent safety, CowPilot (Huq et al., 2025) explores human-agent interactions, and Wei et al. (2025) measures the ability of AI agents to locate hard-to-find information.

8 Conclusion

We introduce BearCubs, a dataset designed to evaluate the ability of a web agent to identify factual information from the real web through multimodal interactions. Through careful dataset creation and curation, we identify and mitigate key challenges in evaluating web agents, including web contamination, agent workarounds, interaction diversity, and slow evaluation. We find that agents lag significantly behind human performance, particularly on multimodal interactions that humans find trivial to perform. Finally, we highlight impactful directions for future agent development, including enhancing trajectory transparency, source credibility, multimodal capabilities, and planning.

Acknowledgments

We sincerely thank Meiyu Li and Ali Nirheche for volunteering in the Chinese and Arabic human study. We thank Ben Glickenhaus for helping with the BEARCUBS website. We extend gratitude to Kalpesh Krishna, who shared question design ideas with us at the beginning of the project. We thank the Upwork annotators for their hard work, and the members from the UMass NLP lab and UMD CLIP lab for their feedback. This project was partially supported by awards IIS2046248, IIS-2312949, and IIS-2202506 from the National Science Foundation (NSF).

References

Anthropic. Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku, 2024. URL https://www.anthropic.com/news/3-5-models-and-computer-use. Accessed: March 4, 2025.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6z4YKr0GK6.

Convergence AI. Proxy: Your AI assistant for your daily tasks, 2025. URL https://convergence.ai/. Accessed: March 4, 2025.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Thirty-seventh*

Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL https://openreview.net/forum?id=kiYqbO3wqw.

Google Gemini. Gemini deep research. https://gemini.google/overview/deep-research/, December 2024. Accessed July 2025.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. WebVoyager: Building an end-to-end web agent with large multimodal models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6864–6890, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.371. URL https://aclanthology.org/2024.acl-long.371/.

Faria Huq, Zora Zhiruo Wang, Frank F. Xu, Tianyue Ou, Shuyan Zhou, Jeffrey P. Bigham, and Graham Neubig. CowPilot: A framework for autonomous and human-agent collaborative web navigation, 2025. URL https://arxiv.org/abs/2501.16609.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A "novel" challenge for long-context language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17048–17085, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.948. URL https://aclanthology.org/2024.emnlp-main.948/.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 881–905, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.50. URL https://aclanthology.org/2024.acl-long.50/.

Ido Levy, Ben wiesel, Sami Marreed, Alon Oved, Avi Yaeli, and Segev Shlomov. ST-WebAgentBench: A benchmark for evaluating safety and trustworthiness in web agents, 2025. URL https://openreview.net/forum?id=IIzehISTBe.

Eric Li and Jim Waldo. WebSuite: Systematically evaluating why web agents fail, 2024. URL https://arxiv.org/abs/2406.01623.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with multi-turn dialogue, 2024. URL https://arxiv.org/abs/2402.05930.

MAA. American invitational mathematics examination - aime, February 2024. URL https://maa.org/math-competitions/american-invitational-mathematics-examination-aime, https://huggingface.co/datasets/Maxwell-Jia/AIME_2024.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback, 2022. URL https://arxiv.org/abs/2112.09332.

OpenAI. Computer-using agent, 2024. URL https://openai.com/index/computer-using-agent/. Accessed: 2025-03-06.

OpenAI. Introducing chatgpt agent: bridging research and action, 2025a. URL https://openai.com/index/introducing-chatgpt-agent//. Accessed: 2025-07-21.

OpenAI. Deep research system card, February 2025b. URL https://cdn.openai.com/deep-research-system-card.pdf. Accessed: 2025-02-27.

OpenAI. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, April 2025. Accessed July 2025.

OpenAI. Operator system card, January 2025. URL https://cdn.openai.com/operator_system_card.pdf. Accessed: 2025-03-01.

Perplexity AI. Build with the best ai answer engine. https://sonar.perplexity.ai/, April 2025. Accessed July 2025.

Ivo Petrov, Jasper Dekoninck, Lyuben Baltadzhiev, Maria Drencheva, Kristian Minchev, Mislav Balunović, Nikola Jovanović, and Martin Vechev. Proof or bluff? evaluating llms on 2025 usa math olympiad, 2025. URL https://arxiv.org/abs/2503.21934.

Shanghaoran Quan, Jiaxi Yang, Bowen Yu, Bo Zheng, Dayiheng Liu, An Yang, Xuancheng Ren, Bofei Gao, Yibo Miao, Yunlong Feng, et al. Codeelo: Benchmarking competition-level code generation of llms with human-comparable elo ratings. *arXiv preprint arXiv:2501.01257*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, December 2023.

Yash Savani, Asher Trockman, Zhili Feng, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J. Zico Kolter. Antidistillation sampling, 2025. URL https://arxiv.org/abs/2504.13146.

George Thomas, Alex J. Chan, Jikun Kang, Wenqi Wu, Filippos Christianos, Fraser Greenlee, Andy Toulis, and Marvin Purtorab. WebGames: Challenging general-purpose web-browsing ai agents, 2025. URL https://arxiv.org/abs/2502.18356.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.813. URL https://aclanthology.org/2024.findings-acl.813/.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents, 2025. URL https://arxiv.org/abs/2504.12516.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking llms in web traversal, 2025. URL https://arxiv.org/abs/2501.07572.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments, 2024. URL https://arxiv.org/abs/2404.07972.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. TheAgentCompany: Benchmarking llm agents on consequential real world tasks, 2024. URL https://arxiv.org/abs/2412.14161.

Kevin Xu, Yeganeh Kordi, Tanay Nayak, Adi Asija, Yizhong Wang, Kate Sanders, Adam Byerly, Jingyu Zhang, Benjamin Van Durme, and Daniel Khashabi. Tur[k]ingBench: A challenge benchmark for web agents, 2025. URL https://arxiv.org/abs/2403.11905.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik R Narasimhan, and Ofir Press. SWE-agent: Agent-computer interfaces enable automated software engineering. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=mXpq6ut8J3.

Ke Yang, Yao Liu, Sapana Chaudhary, Rasool Fakoor, Pratik Chaudhari, George Karypis, and Huzefa Rangwala. AgentOccam: A simple yet strong baseline for LLM-based web agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oWdzUpOlkX.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. WebShop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20744–20757. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/82ad13ec01f9fe44c01cb91814fd7b8c-Paper-Conference.pdf.

Lance Ying, Katherine M. Collins, Lionel Wong, Ilia Sucholutsky, Ryan Liu, Adrian Weller, Tianmin Shu, Thomas L. Griffiths, and Joshua B. Tenenbaum. On benchmarking human-like intelligence in machines, 2025. URL https://arxiv.org/abs/2502.20502.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. AssistantBench: Can web agents solve realistic and time-consuming tasks? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8938–8968, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.505. URL https://aclanthology.org/2024.emnlp-main.505/.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

A Limitations

While we ensure high-quality of BEARCUBS via rigorous data revision and filtering, we identify the following limitations of the benchmark and hope future work will improve on these aspects to develop more robust and versatile web agents. First, every question in BEARCUBS has a single short answer, whereas in a more realistic setting, some questions may not have an answer at all or may have multiple or even long-form answers. For such questions, agents should provide credible sources for each possible answer and should be evaluated on source quality. Second, while BEARCUBS includes questions that are multilingual, testing the ability of agents to handle multilingual queries is not the primary goal of our benchmark due to its limited size. We encourage future research to conduct systematic studies on agents' performance across different cultures and languages. Third, directly comparing agents based on their action trajectories is challenging due to inconsistencies in the level of detail individual agents provide. We encourage future advancements in agent transparency to enable more straightforward and meaningful comparisons.

B Data creation workflow

We seek questions that satisfy the following four high-level criteria. The concrete data creation workflow can be found in Figure 3.

(1) *Questions should be short but unambiguous*: questions should provide sufficient yet minimal information to unambiguously lead to correct answers.
(2) *Answers should be trivial to evaluate*: answers must be correct, unique, and concise. Answers cannot be lists or sets, unlike, for example, AssistantBench (Yoran et al., 2024), and paraphrases of answers cannot be considered correct.
(3) *Answers should be adversarial to Google Search*: answers must not appear in Google Search snippets or top-ranked results when the question or fragments of the question are used as the query. Furthermore, **multimodal** questions must not be solvable by methods that only operate on text (e.g., Deep Research).
(4) *Answers must be publicly accessible*: answers must be available on non-paywalled websites, without requiring any account creation or login actions.

Figure 3 details the data validation process discussed in Section 3. Each question is verified by at least two authors who review each question carefully against the four criteria, along with its viable trajectory and visited links. We remove questions whose trajectories involve **multimodal** interaction but were found to have a text-only workaround by non-multimodal agents.

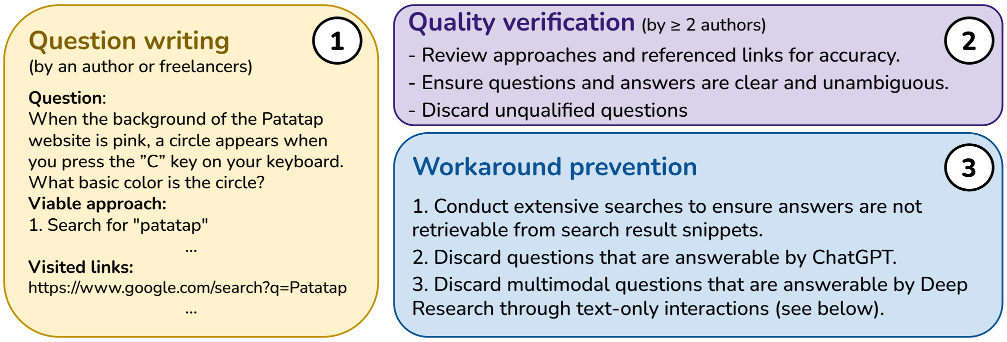

Question writing
(by an author or freelancers)
1
Question:
When the background of the Patatap website is pink, a circle appears when you press the "C" key on your keyboard. What basic color is the circle?
Viable approach:
1. Search for "patatap"
...
Visited links:
https://www.google.com/search?q=Patatap
...
Quality verification (by ≥ 2 authors)
2
- Review approaches and referenced links for accuracy.
- Ensure questions and answers are clear and unambiguous.
- Discard unqualified questions
Workaround prevention
3
1. Conduct extensive searches to ensure answers are not retrievable from search result snippets.
2. Discard questions that are answerable by ChatGPT.
3. Discard multimodal questions that are answerable by Deep Research through text-only interactions (see below).

Figure 3: Workflow of creating BEARCUBS. Each question and its viable trajectory and visited links are verified by at least two authors. Each **multimodal** question is verified to be only solvable via multi-modal interactions.

C Human annotator recruitment

This section provides details on human annotator recruitment for the human study in Section 4.1. We organize BEARCUBS questions into sets, each containing all questions from a specific non-English language and some English questions. We recruit native speaker volunteers for Arabic and Chinese questions, and we hire three annotators via Upwork to handle the Hindi, German, and Finnish questions, respectively. An English-only set is attempted by annotators and authors who did not write or validate them. The hired annotators receive $2.5 USD per question, with an additional $1 USD bonus for each correct answer to incentivize accuracy.

D Baseline prompt

We provide the baseline hyperparameters and prompt for the Google-search-augmented baselines in Table 4

Model: `gpt-4o-2024-11-20` **max_tokens**: 518 **temperature**: 0 # for both baseline settings
Model: `DeepSeek R1` **max_tokens**: 8000 **temperature**: 0 # for both baseline settings
Prompt: # for the Google-search-augmented setting
Use the provided context to answer the question accurately and concisely. Do not use your own knowledge.
- If the context contains a direct answer, provide it in a precise and straightforward manner.
- If the context does not provide a clear answer, reply with 'No answer found'.
- Avoid unnecessary elaboration.

Question: {question}
Context: {context} # consists of a list of result titles and snippets
Answer:

Table 4: Baseline hyperparameters and the prompt in the Google-search-augmented setting.

E Human performance results

Table 5 complements the discussion in Section 5.1 by providing detailed results on the number of dead ends encountered, time taken, and perceived difficulty experienced by the human annotators.

	Correctness				Dead End		Time (min:sec)			Perceived Difficulty		
	Accuracy	Correct	Wrong	None	Avg.	Max	Avg.	Min	Max	Easy	Medium	Hard
Text-based	83.6%	46	9	1	1.83	12	5:19	0:43	27:24	23	20	13
Multimodal	85.7%	48	5	2	1.22	14	4:23	0:26	24:14	32	16	7
All	84.7%	94	14	3	1.50	14	4:46	0:26	27:24	55	36	20

Table 5: Humans achieve 84.7% accuracy on BEARCUBS and were generally able to find answers smoothly (1.5 dead ends on average) and quickly (4 min 46 sec). The *None* label marks questions abandoned after 15 minutes. *Perceived difficulty* reflects annotators' subjective assessment of question difficulty. The minimum number of dead ends was zero.

F Human error examples

Table 6 lists six key causes of human errors mentioned in Section 5.1, along with corresponding examples and explanations.

G Details of automatic agent answer evaluator

This section provides the implementation details of the automatic evaluator in Section 4.2.

Prompt of the automatic evaluator An example of the prompt used for automatic evaluation is shown in Table 7. Because the prompt contains 17 in-context examples, the full text is omitted here. The complete prompt and the code of the evaluator are available at `https://bear-cubs.github.io/`.

Modal and hyperparameters The evaluator is built on `gpt-4o-2024-11-20` as its base LLM, with the temperature fixed at 0. Running the evaluator on 111 questions takes approximately 1 minute and 30 seconds, costing around $0.80 USD.

Evaluator performance As shown in the prompt in Table 7, the evaluator performs a four-way classification task. We also report results with the four labels collapsed into two

Error type & question	Explanation
Type: Missing details in the question or answer ; Lack of topic knowledge ; Suboptimal source selection . **Question**: Which bird(s) in the Wingspan bird cards from the Asia expansion have the word "Great" (but not "Greater") in their name and can reside exclusively in wetlands?	The annotator (1) was unfamiliar with the game, (2) found a Wingspan card website lacking expansion and habitat details, then identified and searched bird with "great" individually, and (3) overlooked the Asia expansion requirement.
Type: Obvious oversight **Questions**: What does the October 2024 edition (817) of Tinkle magazine teach readers to make on page 36?	The annotator accessed a relevant magazine page but overlooked the answer on it.
Type: Complexity of task **Question**: On February 10, 2016, at 2:00 AM, a tagged Altai Snow Leopard was observed in West Mongolia, according to the data from USGS. What was the straight-line distance it traveled by the time it was observed again at 6:00 AM on the same day?	The annotator easily found the website but gave up after failing to identify the correct map markers representing the Snow Leopard's location at the specified date and time.

Table 6: Examples of human errors. The annotators answered 14 questions incorrectly and gave up on 3, primarily due to five key reasons listed in the table. We provide three examples, each with attributed error reasons and detailed explanation. The two more frequent error reasons are missing details in the question or answer and lack of topic knowledge .

categories, `correct` and `incorrect`. The evaluator was applied to the outputs of five agents: Grok3 DeepSearch, OpenAI Deep Research, Anthropic Computer Use, Convergence AI Proxy, and OpenAI Operator. The results for both the four-way and binary settings are presented in Table 8. To verify the reliability of the evaluator, we ran it three times on outputs from Anthropic and Proxy, obtaining identical results in all runs.

H Detailed agent performance results

Table 9 provides detailed agent results, grouped by all questions, **text-based**, and **multimodal** questions.

I Justifying size of BEARCUBS

Small size benchmarks are not rare. BEARCUBS currently contains 111 questions. Such a benchmark size is not uncommon in prior popular benchmarks. The widely used AIME dataset (MAA, 2024) contains only 30 questions and is still adopted in both evaluation and model development efforts such as by OpenAI (OpenAI, 2025) and DeepSeek (DeepSeek-AI et al., 2025). Similarly, CodeElo (Quan et al., 2025) includes just 387 problems, and the Proof or Bluff dataset (Petrov et al., 2025) uses only six USAMO questions for testing.

Creating high-quality BEARCUBS questions is non-trivial. In our experience, freelance annotators were only able to write a small number of accepted questions before quality declined noticeably. Additionally, OpenAI's Deep Research agent frequently identified unintended shortcuts, especially in multimodal setting, which led us to discard many questions to preserve benchmark integrity.

BEARCUBS remains a meaningful benchmark. This is due to two key factors. First, the search space for BEARCUBS questions spans the entire real-world web, unlike, for example, constrained multiple-choice formats. This makes guessing the correct answer difficult and reduces accuracy variance. Second, solving even a single BEARCUBS question typically requires a complex mix of retrieval, reasoning, and synthesis, making each example informative for evaluating model capabilities. To support the first claim, we evaluated model accuracy variance on BEARCUBS by running DeepSeek R1 with and without Serper across

Prompt of the automatic agent response evaluator

Given a question, its gold answer, and an AI generated answer, your task is to check for the following four cases:

1. no_answer: AI generated answer is empty.
2. no_direct_answer: Agent abstains or hedges with modal verbs (e.g., “may”, “could be”) or phrases expressing uncertainty such as “according to common sense” or “I am leaning toward.”
3. correct: AI returns a direct and unambiguous answer which entails the gold answer.
4. wrong: AI returns a direct and unambiguous answer which, however, does not entail the gold answer.

Every question has an answer. If the AI answer rejects the premise of a question, it should be labeled as “wrong.”

The gold answers are always concise and never presented as a list. They may be a number, a price, a word, or a phrase, among other types. In contrast, AI answers can be long, such as a paragraph, with an answer embedded within. The embedded answer may be correct but differ in format (e.g., how a number is written). Nonetheless, they should express the same information (i.e., 3 and four are different, 303kg and 302kg are different.) Below, I provide examples of how to decide which of the four labels to assign to the AI answer along with explanations. However, in your evaluation, you do not need to explain your decisions.

Example 1:
Question: What is the name of the bird(s) in the Wingspan bird cards from the Asia expansion have the word ”Great” (but not ”Greater”) in their name and can reside exclusively in wetlands?
Gold Answer: Great Cormorant
AI Answer:
Label: no_answer
Explanation: The AI answer is empty. Hence, the label is “no_answer.”

Example 2:
Question: In one episode of Wissen Macht Ah! in April 2023, the hosts demonstrated how to make a DIY water filter in the wild. What was the second filtering material that they placed into the plastic bottle? Answer the question with the exact word used in the episode?
Gold Answer: Activkohle
AI Answer: Sorry, I don’t have access to specific episodes of *Wissen Macht Ah!* or their content, including the exact word used in the April 2023 episode. You might want to check the episode directly or refer to its official resources for the precise information.
Label: no_direct_answer
Explanation: The AI answer does not provide a direct answer to the question and instead states that it does not have access to the specific episode’s content.

Further examples omitted for the sake of space.

Now it is your turn to label the following AI answers. You do not need to explain your decisions.

Question: {question}
Gold Answer: {gold_answer}
AI Answer: {ai_output}
Label:

Table 7: Example of the prompt used for the automatic agent response evaluator. Examples 3 to 17 are omitted for the sake of space.

four independent trials. The results, summarized below, show minimal standard deviation in accuracy across runs.[21]

J Model correct answer source proportions

Figure 4 presents the proportion of correct answers within each agent categorized by reliance on primary sources, secondary sources, and ungrounded responses. The analysis is provided in Section 6.

[21] The first run was conducted in March 2025 using the Fireworks AI API; subsequent runs were carried out via OpenRouter.

Agent	4-Way Accuracy	Binary Accuracy
Grok3 Deep Search	99.1%	99.1%
OpenAI Deep Research	96.4%	96.4%
Anthropic Computer Use	99.1%	100%
Convergence AI Proxy	99.1%	100%
OpenAI Operator	97.3%	98.2%

Table 8: Evaluator accuracy by agent in the four-way classification setting. Binary accuracy is derived by collapsing the four labels into two categories: `correct` and `incorrect`.

Model	Accuracy	Answer label				Average time			Correct answer source attribution		
		Correct	Wrong	Uncertain	None	Correct	Wrong	Uncertain	Ungrounded	Primary	Secondary
All Questions											
GPT-4o zero-shot	2.7%	3	53	55	0	—	—	—	—	—	—
DeepSeek R1 zero-shot	8.1%	9	82	19	1	—	—	—	—	—	—
GPT-4o + Google Search	0.0%	0	4	0	107	—	—	—	0	0	0
DeepSeek R1 + Google Search	1.8%	2	16	0	93	—	—	—	1	0	1
Perplexity `sonar-pro`	5.4%	6	30	58	17	—	—	—	2	4	0
Grok3 DeepSearch	11.7%	13	95	2	1	1:09	1:24	2:05	3	8	2
OpenAI Deep Research	36.0%	40	69	1	1	4:37	9:00	3:58	5	25	10
Google Deep Research	23.4%	26	39	46	0	4:21	4:00	4:39	1	25	0
Convergence AI Proxy	12.6%	14	44	34	19	1:52	2:41	5:24	0	13	1
Anthropic Computer Use	14.4%	16	22	73	0	2:24	2:35	3:35	0	14	2
OpenAI Operator	23.4%	26	43	13	29	2:59	3:58	8:06	1	24	1
OpenAI ChatGPT Agent	65.8%%	73	30	3	5	9:16	16:14	25:55	1	72	0
Human	84.7%	94	14	—	3	4:24	5:44	—	—	—	—
Text-based Questions											
GPT-4o zero-shot	5.4%	3	30	23	0	—	—	—	—	—	—
DeepSeek R1 zero-shot	10.7%	6	38	12	0	—	—	—	—	—	—
GPT-4o + Google Search	0.0%	0	4	0	52	—	—	—	0	0	0
DeepSeek R1 + Google Search	3.6%	2	8	0	46	—	—	—	1	0	1
Perplexity `sonar-pro`	8.9%	5	19	26	6	—	—	—	1	4	0
Grok3 DeepSearch	21.4%	12	42	2	0	1:07	1:30	—	2	8	2
OpenAI Deep Research	60.7%	34	21	0	1	4:12	8:32	—	0	24	10
Google Deep Research	42.9%	24	13	19	0	4:22	4:13	5:13	0	24	0
Convergence AI Proxy	16.1%	9	26	14	7	2:05	2:36	4:11	0	8	1
Anthropic Computer Use	19.6%	11	12	33	0	2:49	2:27	3:21	0	9	2
OpenAI Operator	33.9%	19	24	3	10	3:13	3:51	7:24	0	18	1
OpenAI ChatGPT Agent	76.8%%	43	10	1	2	6:29	12:06	43:49	0	43	0
Human	83.6%	46	8	—	1	5:09	5:07	—	—	—	—
Multimodal Questions											
GPT-4o zero-shot	0.0%	0	23	32	0	—	—	—	—	—	—
DeepSeek R1 zero-shot	5.5%	3	44	7	1	—	—	—	—	—	—
GPT-4o + Google Search	0.0%	0	0	0	55	—	—	—	0	0	0
DeepSeek R1 + Google Search	0.0%	0	8	0	47	—	—	—	0	0	0
Perplexity `sonar-pro`	1.8%	1	11	32	11	—	—	—	1	0	0
Grok3 DeepSearch	1.8%	1	53	0	1	1:25	1:20	—	1	0	0
OpenAI Deep Research	10.9%	6	48	1	0	6:58	9:12	3:58	5	1	0
Google Deep Research	3.6%	2	26	27	0	4:15	3:53	4:16	1	1	0
Convergence AI Proxy	9.1%	5	18	20	12	1:29	2:48	10:42	0	5	0
Anthropic Computer Use	9.1%	5	10	40	0	1:29	2:45	3:47	0	5	0
OpenAI Operator	12.7%	7	19	10	19	2:21	4:06	8:19	1	6	0
OpenAI ChatGPT Agent	54.5%%	30	20	2	3	13:15	18:19	16.58	1	29	0
Human	85.7%	48	6	—	2	3:41	6:41	—	—	—	—

Table 9: Model performance on BEARCUBS with the baselines `gpt-4o-2024-11-20` and DeepSeek R1 (with/without Google Search), Perplexity `sonar-pro`, deep research agents , computer-use agents, and human performance . 'Uncertain' indicates that the agent did not return a concrete answer (e.g., abstention), while 'None' means either a baseline model returned "No answer found" or an agent entered a dead loop and failed to provide any response. 'Ungrounded' answers are those not based on a source but rather on the agent's internal knowledge or reasoning. 'Primary' denotes an answer derived from a reliable source or the source specified in the question, whereas 'Secondary' refers to an answer obtained from an unreliable source or a source not mentioned in the question.

K Agent-specific behavior

We continue the discussion in Section 6 and present interesting agent-specific behavior and the challenges those agents face that hinder their utility.

Setup	1st Run	2nd Run	3rd Run	4th Run	Std. Dev.
DeepSeek R1 w/o Serper	7.2%	5.4%	5.4%	5.4%	0.8%
DeepSeek R1 w/ Serper	1.8%	0.9%	2.7%	2.7%	0.7%

Table 10: Accuracy standard deviation across four trials of DeepSeek R1 on BEARCUBS, with and without Serper. Results show low standard deviation, indicating stable performance.

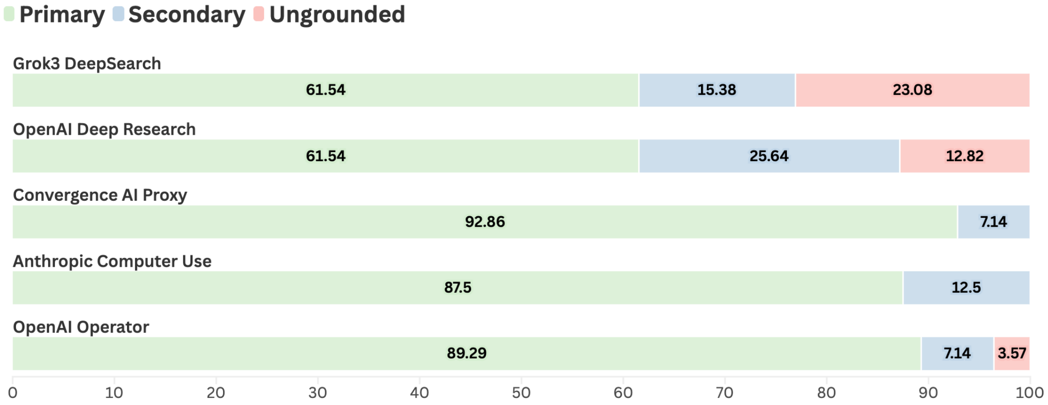


Figure 4: Proportion of correct answers within each agent categorized by reliance on primary sources, secondary sources, and ungrounded responses.

We observe that Grok 3 returns its search and reasoning trajectory in a non-English language if it appears in the input, which can be unhelpful for users seeking assistance for that language. Grok 3 also never abstains, always generating a response (see Table 2)—a behavior that needs user study to determine its desirability. Meanwhile, Computer Use sometimes deems a task impossible without attempting it, despite users expecting an effort and justification. These issues, along with the broader challenges discussed above, highlight the need for improved adaptability, transparency, and user-centered refinement in LLM-based computer use agents.