VERIWEB: VERIFIABLE LONG-CHAIN WEB BENCHMARK FOR AGENTIC INFORMATION-SEEKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances have showcased the extraordinary capabilities of Large Language Model (LLM) agents in tackling web-based information-seeking tasks. However, existing efforts mainly focus on single-fact retrieval and rely on outcome-only verification, thereby limiting their scalability in realistic knowledge-intensive scenarios that involve long-horizon web tasks requiring large-scale retrieval and synthesis of information from diverse sources. In this work, we introduce VeriWeb, a novel verifiable long-chain web benchmark designed to facilitate the evaluation and development of web agents within realistic web environments. Our benchmark emphasizes two critical dimensions: (1) *long-chain complexity*, encompassing both breadth- and depth-oriented search tasks to assess how effectively web agents ensure comprehensive information coverage and consistent context tracking in multi-hop reasoning; and (2) *subtask-level verifiability*, where tasks are decomposed into a sequence of interdependent verifiable subtasks. This structure enables diverse exploration strategies within each subtask, while ensuring that each subtask-level answer remains unchanged and verifiable. The benchmark consists of *302* tasks across five real-world domains, each with a complete trajectory demonstration, *annotated by human experts*. Extensive experiments on VeriWeb using various agents powered by different foundation models reveal significant performance gaps in handling long-horizon web tasks, highlighting the need for more powerful agentic information-seeking capabilities[1].

1 INTRODUCTION

Autonomous web agents have recently demonstrated remarkable capabilities in complex information-seeking tasks by following high-level instructions (Jin et al., 2025; Li et al., 2025c; Song et al., 2025a; Wu et al., 2025a;c; Gao et al., 2025), supporting a wide range of deep research systems (Google, 2025; OpenAI, 2025; xAI, 2025; MoonshotAI, 2025; Li et al., 2025e;e; Zheng et al., 2025). Recent breakthroughs in Large Language Models (LLMs) (Anil et al., 2023; Achiam et al., 2023; Guo et al., 2025; Yang et al., 2025; Zhang et al., 2025) have enabled promising prototypes of such agents, capable of complex search and browsing without relying on hard-coded automation or domain-specific scripting (Huang et al., 2025; Xi et al., 2025). However, developing such general-purpose web agents involves multiple complex processes, as it requires the ability to retrieve context-specific knowledge (Kwiatkowski et al., 2019; Joshi et al., 2017; Mallen et al., 2022), perform multi-hop reasoning across diverse sources (Press et al., 2024; Trivedi et al., 2022), and synthesize large-scale information (Du et al., 2025; Li et al., 2025e). This also poses a new challenge: how to obtain high-quality datasets that capture diverse, realistic information-seeking tasks to evaluate these agents effectively (Li et al., 2025b; Tao et al., 2025).

To address this challenge, various datasets and benchmarks have been released to advance the development of autonomous web agents (Mallen et al., 2022; Wu et al., 2025b; Wei et al., 2025; 2024; Mialon et al., 2024; Phan et al., 2025). Despite encouraging results, existing web benchmarks still exhibit two major limitations. ***First***, most recent benchmarks focus on *single-fact retrieval* (Yang et al., 2018; Ho et al., 2020; Wei et al., 2025), where agents are tasked with retrieving an atomic fact, typically by performing shallow navigation and cross-page matching. For example, a task

[1]Anonymous code and data are available in the supplementary materials. Since the human demonstration data exceeds 100GB, only one sample is provided for reference due to the file size limit.

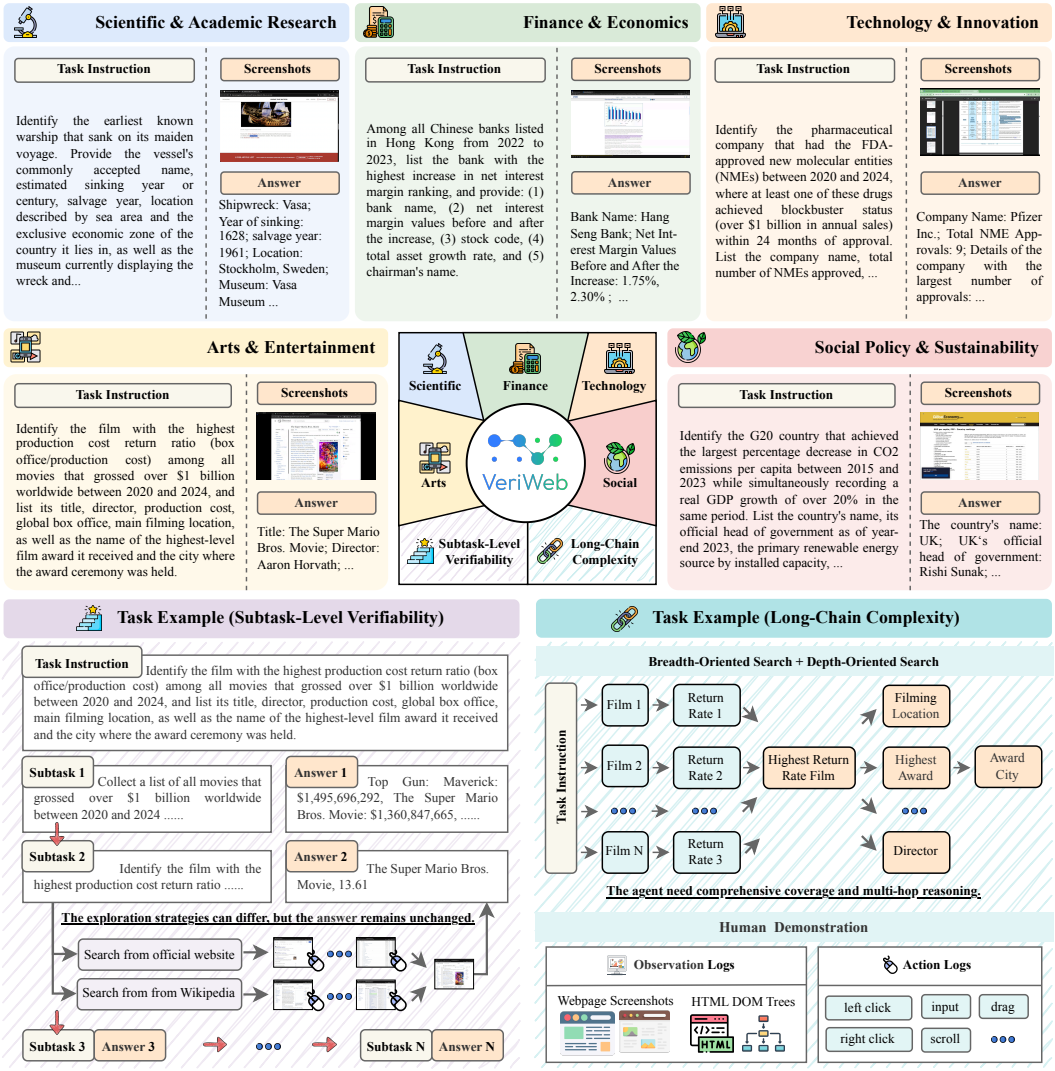


Figure 1: An overview of the VeriWeb benchmark across five domain-specific scenarios, which emphasizes (1) *long-chain complexity*, with tasks integrating both breadth- and depth-oriented search challenges, requiring comprehensive coverage and multi-hop reasoning. (2) *subtask-level verifiability*, where tasks are decomposed into interdependent subtasks with verifiable answers. Note that each task includes a complete human demonstration with detailed observation and action logs.

like "Which river that flows through Vienna also passes through Budapest?" can often be solved by simply navigating between two pages and extracting the answer "Danube". Such formulations rarely require large-scale retrieval and synthesis of information from diverse sources, both of which are essential for solving realistic information-seeking workflows that, most of the time, demand information-rich outputs rather than a single fact (Chatterji et al., 2025). ***Second***, existing evaluation protocols typically rely on *outcome-only validation*, checking only whether the final result matches the ground-truth answer. This coarse-grained supervision fails to capture the quality of intermediate steps, especially when tasks involve multiple interdependent subtasks. In such cases, when agents fail, it is often unclear where or why the failure occurred, thereby making it difficult to support improvements to agent capability.

In this work, we introduce VeriWeb, a new verifiable long-chain benchmark tailored for the evaluation and development of web agents. VeriWeb comprises various web task trajectories across five domain-specific scenarios. All trajectories are carefully curated and annotated by human experts, ensuring long-chain complexity and subtask-level verifiability, as shown in Fig. 1. (1) The *long-chain complexity* of VeriWeb features tasks that require agents to perform both breadth-oriented search

(*e.g.*, listing all films with a box office over $1 billion) and depth-oriented search (*e.g.*, identifying the film with the highest return rate → its highest award → the award city). To succeed, agents must engage in multi-hop reasoning, maintain coherent context across pages, and fuse fragmentary evidence into a well-supported synthesis. This design mirrors real-world information-seeking workflows, where relevant information is scattered across sources and must be reliably linked along a long reasoning chain. (2) The *subtask-level verifiability* of VeriWeb enables a fine-grained assessment of intermediate results at every subtask rather than solely at the final outcome. Notably, each subtask is designed to serve as a valid starting point, supporting agent evaluation at different task stages. A subtask consists of multiple steps involving realistic search and browsing operations. Instead of verifying each low-level action, the benchmark focuses on evaluating whether the goal of each subtask has been correctly achieved, providing a more informative supervision signal. This design supports open-ended interaction within each subtask, encouraging agents to explore diverse strategies to accomplish the subtask goal rather than adhering to a fixed action sequence. Our core contributions are summarized as follows:

- We curate a high-cost, human-annotated dataset of *302* verifiable long-chain task trajectories across five real-world domains, capturing both *long-chain complexity* and *subtask-level verifiability*. Each task is decomposed into interdependent subtasks with fixed, verifiable answers, emphasizing information-rich synthesis rather than single-fact retrieval.
- On top of this dataset, we design the VeriWeb benchmark, supporting multiple levels of evaluation, including task success rate, task completion rate, and action efficiency. This enables fine-grained analysis of agent capabilities across different stages of task execution and provides deeper insights into failure modes and information-seeking bottlenecks.
- Extensive experiments with a range of various agents using state-of-the-art foundation models show consistently poor performance on long-horizon information-seeking tasks, underscoring current limitations of complex retrieval and synthesis in web agents.

2 Related Works

2.1 Information-Seeking Web Benchmarks

Existing web benchmarks broadly fall into two categories: interaction-centric and information-seeking tasks. The former focuses on UI-grounded action execution on the web, such as online shopping or emailing (Yao et al., 2022; Deng et al., 2023; Zhou et al., 2023). The latter targets search and browsing behaviors (Li et al., 2025b; Gao et al., 2025; Tao et al., 2025; Du et al., 2025)[2]. Early information-seeking benchmarks emphasize simple question answering over static corpora, typically solvable with single-hop queries (Kwiatkowski et al., 2019; Joshi et al., 2017; Mallen et al., 2022). Subsequent multi-hop benchmarks introduce compositional reasoning across multiple pages (Yang et al., 2018; Ho et al., 2020; Trivedi et al., 2022; Press et al., 2024; Wu et al., 2025b), but remain limited to Wikipedia-like closed environments. Recent efforts shift toward open-web scenarios (Wei et al., 2025; Zhou et al., 2025; Chen et al., 2025c; Wei et al., 2024; Chen et al., 2025a), though many still target single-fact retrieval, falling short of realistic tasks that demand information-rich, synthesized outputs. To bridge this gap, Du et al. (2025) benchmarks agents on report-level generation, aligning better with deep research workflows. However, such reports often include subjective or time-sensitive content, complicating direct ground-truth evaluation. Moreover, most benchmarks rely on outcome-only verification, neglecting challenges like error localization in long-horizon tasks. In contrast, VeriWeb is designed to reflect the complexity of real-world information-seeking tasks, supporting long-chain complexity and subtask-level verifiability.

2.2 Information-Seeking Web Agents

Information-seeking web agents have evolved from prompt-engineered browsing to reinforcement-learned "reason-with-search" behaviors (Sun et al., 2025; Jin et al., 2025; Li et al., 2025c; Song et al., 2025a;b; Chen et al., 2025b). These agents are now capable of deciding when and what to query, and of integrating retrieved evidence during multi-step reasoning (Wu et al., 2025a; Zheng et al., 2025; Li et al., 2025d; Geng et al., 2025). A growing line of work explores deep research agents (Li

[2]Note that this work focuses on information-seeking tasks, while interaction-centric tasks are out of scope.

et al., 2025e;a; Qiao et al., 2025), which aim to perform end-to-end evidence gathering and report synthesis, mirrored by production features like OpenAI Deep Research (OpenAI, 2025) and Gemini Deep Research (Google, 2025). Despite this momentum, our experiments show that current agents struggle with comprehensive coverage and multi-hop reasoning with consistent context tracking in complex information-seeking workflows, underscoring the need for benchmarks like VeriWeb that explicitly test long-horizon web tasks requiring large-scale retrieval and synthesis of information from diverse sources.

3 VERIWEB BENCHMARK

In this section, we present the task formulation, data collection procedure, and statistical analysis of the VeriWeb benchmark. VeriWeb is a carefully designed and human-curated benchmark featuring rigorous task formulation, expert annotation, and multi-stage review, yielding a challenging suite that targets real-world information-seeking tasks.

3.1 TASK FORMULATION

We formulate information-seeking tasks in VeriWeb as a Partially Observable Markov Decision Process (POMDP), defined by the tuple $\langle \mathcal{S}, \mathcal{O}, \mathcal{A}, P, O, R \rangle$, where $\mathcal{S}$ is the set of environment states, representing the full underlying web system. $\mathcal{O}$ is the observation space, and $O : \mathcal{S} \to \mathcal{O}$ is the observation function, which models the partial observations agents/humans receive from the environment. For web agents, the action space $\mathcal{A}$ consists of different tools (*e.g.*, search queries, webpage browsing), while the corresponding observations are the tool-call feedback. For human demonstrations, the action space $\mathcal{A}$ instead corresponds to user interactions such as mouse clicks or keyboard inputs, while the observations include webpage screenshots and the HTML DOM tree. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition function, modeling the dynamics of the web environment in response to actions. R is the reward function, which is defined through verifiable answers.

For each information-seeking task in VeriWeb with an instruction Q, we collect a complete trajectory demonstration $\tau = (o_0, a_0, o_1, a_1, \ldots, o_T)$ from human annotators, where T denotes the number of steps in the trajectory. To capture intermediate results and provide dense supervision, we decompose τ into a sequence of K subtasks $\tau^{(1)}, \tau^{(2)}, \ldots, \tau^{(K)}$, such that $\tau = \tau^{(1)} \circ \tau^{(2)} \circ \cdots \circ \tau^{(K)}$, where $\circ$ denotes trajectory concatenation. The subtask $\tau^{(k)} = (o_{t_k}, a_{t_k}, \ldots, a_{t_{k+1}-1}, o_{t_{k+1}})$ corresponds to a contiguous segment of the full trajectory, where t_k and t_{k+1} denote the start and end timesteps. Each subtask $\tau^{(k)}$ is associated with a sub-instruction $Q^{(k)}$ and a subtask-level ground-truth answer $Y^{(k)}$. The task instruction Q also has a task-level ground-truth answer Y.

3.2 DATA COLLECTION.

Data Source. The VeriWeb dataset is constructed from a wide range of real-world web environments. We specifically focus on deep-research-like scenarios involving large-scale information retrieval and synthesis. Thus, we curate data from publicly accessible and authoritative sources as shown in Fig. 2, including official websites of government agencies, academic institutions, online encyclopedias, financial databases, and news portals. These tasks cover five primary thematic domains: (1) scientific and academic research, (2) finance and economics, (3) technology and innovation, (4) arts and entertainment, and (5) social policy and sustainability. This categorization ensures diverse topical coverage and reflects realistic user intentions in complex information-seeking tasks.

Task Instruction Generation. To generate realistic and executable instructions, we develop a multi-stage pipeline combining human curation with language model generation, as shown in the left part of Fig. 2. Initially, a small batch of seed instructions is manually selected for each topical domain. These seed instructions, representing high-level user intents, are input to a language model to generate a large number of candidate tasks. Human annotators then review these outputs, selecting only those that are grammatically clear, semantically meaningful, and practically feasible. Once a vetted pool of main tasks is established, the language model is prompted to perform subtask decomposition to obtain complete task instructions, including detailed sub-instructions of each subtask. This process is guided by seed instructions and strict formatting constraints. After generation, each batch of instructions undergoes automated filtering, followed by a second, stricter verification phase in-

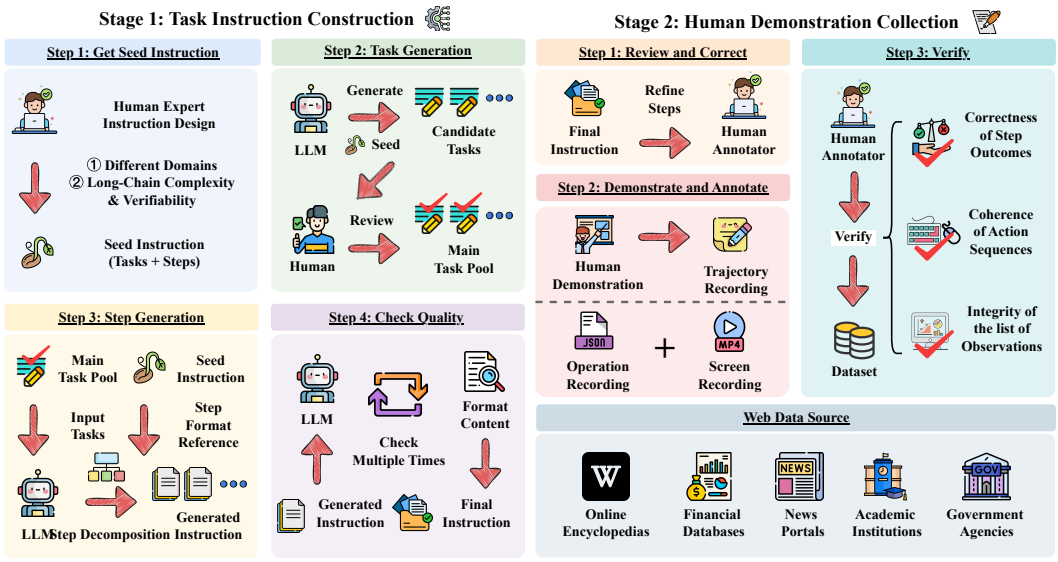


Figure 2: An overview of the proposed VeriWeb framework, consisting of two stages: task instruction construction and human demonstration collection. The framework combines LLM-based generation with human annotation to ensure realistic, high-quality web tasks and demonstrations.

volving multiple passes of model-based evaluation. Only those tasks that pass all verification rounds are retained. This procedure enables efficient instruction generation while maintaining the factual correctness, diversity, and task feasibility necessary for web datasets.

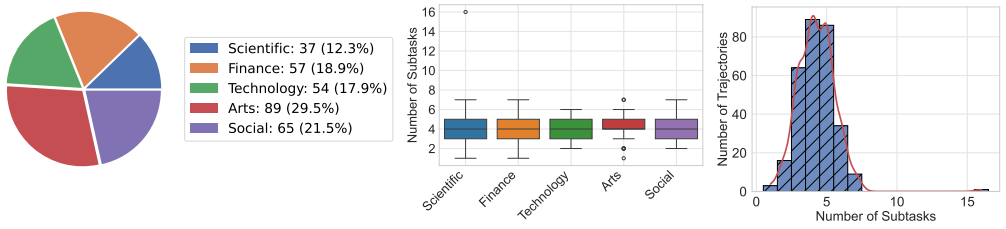
Human Demonstration Collection. Human annotators manually execute each task based on the given final instruction and record the complete trajectory demonstration, as shown in the right part of Fig. 2. Before execution, human annotators refine the subtask sequence to ensure feasibility and smooth operation, allowing adjustments as needed during interaction. Demonstrations are recorded using screen capture tools, with detailed annotations including action logs, observation logs, and subtask-level goals. To ensure high-quality supervision and accurate benchmarking, all trajectory demonstrations undergo strict quality control. This includes both automatic checks and manual review to verify the correctness of subtask outcomes, coherence of action sequences, and integrity of observations. Only demonstrations that meet all criteria are retained. This guarantees that VeriWeb provides reliable and verifiable supervision for long-horizon web agents. The detailed collection document can be found in Appendix A.

3.3 Data Statistics

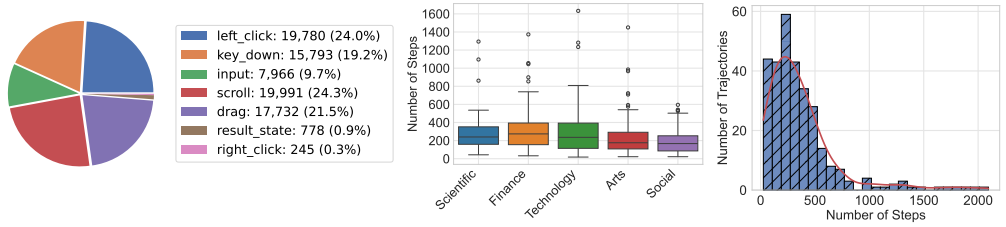
To better understand the characteristics of VeriWeb, Figure 3 and Table 1 present statistical summaries of the collected human demonstrations for all tasks. The domain distribution of tasks in Fig. 3a demonstrates that the dataset covers a wide range of domains, ensuring broad coverage and diversity across real-world tasks. Each task is decomposed into a sequence of subtasks, with each subtask associated with a verifiable answer. Figure 3c illustrates the distribution of subtasks per task, with an average of four subtasks per task. This subtask-level structure enables intermediate supervision, supporting more fine-grained evaluation and training. VeriWeb further emphasizes long-chain complexity. As shown in Fig. 3f, many tasks require executing hundreds of steps before completion. The large number of subtask answer items in Tab. 1 further highlights the need for large-scale retrieval and synthesis. Overall, these statistics demonstrate that VeriWeb provides both subtask-verifiable and long-chain tasks, offering a realistic and challenging benchmark for long-horizon reasoning and information-seeking in the web environment.

Table 1: The overall data statistics of collected human demonstrations in VeriWeb.

Statistic	Value
# Tasks	302
Avg. # Subtasks per Task	4.3
Avg. # Steps per Task	272.5
Avg. # Steps per Subtask	63.5
Avg. # Items of Task Answers	5.7
Max. # Items of Task Answers	28
Max. # Items of Subtask Answers	94



(a) Task Domain Distribution (b) Subtask Count by Domain (c) Subtask Count Distribution



(d) Action Distribution (e) Action Count by Domain (f) Step Count Distribution

Figure 3: The detailed data statistics of collected human demonstrations in VeriWeb.

4 Experiments

4.1 Experimental Settings

Baselines. To demonstrate the effectiveness of VeriWeb, we compare four agent paradigms with different foundation models: (1) Deep research agents: closed-source systems with built-in search, including OpenAI Deep Research (OpenAI, 2025) and Gemini Deep Research (Google, 2025). (2) Search engine agents: models combined with an open-source search tool[3] via the model context protocol[4]. (3) Browser-use agents: models using the Browser-Use framework (Müller & Žunič, 2024). (4) Multi-agent systems: models using the Camel OWL framework (Hu et al., 2025). All used foundation models (Yang et al., 2025; Bai et al., 2025; Liu et al., 2024; Anil et al., 2023; Anthropic, 2024; Hurst et al., 2024) are shown in Tab. 2. Note that except for the closed-source deep research agents, search engine agents retrieve text-only context without page interaction, while browser-use and OWL agents operate on web elements and handle both text and visual input.

Evaluation Metrics. We evaluate agent performance using three metrics: (1) The *task Success Rate (SR)* measures whether the agent completes the overall task. (2) The *task Completion Rate (CR)* measures the extent to which the agent achieves the overall task goal. Since our tasks often involve multiple subtasks, CR estimates the completion level by calculating the proportion of correct items in the output. (3) The *Action Count (AC)* quantifies the planning effectiveness of agents by measuring the number of steps required to arrive at the final answer. For both the SR and the CR, we use the LLM-as-a-Judge score (Gu et al., 2024) based on GPT-4.1 to evaluate the correctness of the agents' final answers. Detailed prompts are provided in Appendix B. Due to the high cost of API calls, each experiment is conducted once to obtain the final evaluation results.

Evaluation Setting. Our task involves long-chain complexity, encompassing both breadth- and depth-oriented search. We observe that the first subtask emphasizes breadth search, while subsequent ones focus on depth search. Therefore, besides evaluating the overall task, we also report results under two specific settings: (1) Breadth-oriented search, which evaluates only the first subtask. (2) Depth-oriented search, which evaluates the overall task given the result of the first subtask.

4.2 Main Results

Table 2 reports the agent performance on VeriWeb across five domains, measuring both task success rate and completion rate. Overall, the results highlight the difficulty of VeriWeb: no single

[3]https://github.com/searxng/searxng-docker
[4]https://modelcontextprotocol.io/

Table 2: Comparison of different agents on VeriWeb across five domains.

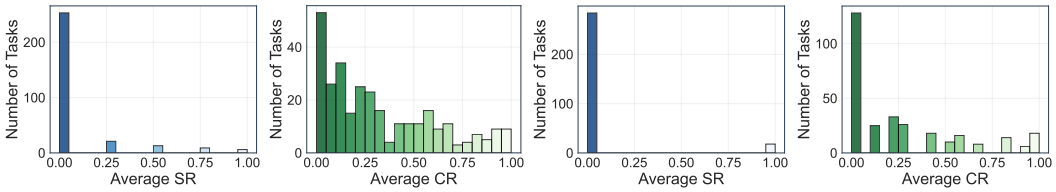
Model	Scientific		Finance		Technology		Arts		Social		Overall	
	SR (%)	CR (%)	SR (%)	CR (%)	SR (%)	CR (%)	SR (%)	CR (%)	SR (%)	CR (%)	SR (%)	CR (%)
Deep Research Agents												
Gemini-2.5-Flash	0.0	19.7	1.8	19.6	0.0	28.3	19.1	48.7	4.6	26.2	6.9	31.2
Gemini-2.5-Pro	5.4	28.1	0.0	15.6	9.3	35.6	21.3	52.9	7.7	32.2	10.3	35.3
OpenAI-o4-mini	2.7	18.1	0.0	12.6	3.7	18.0	14.6	41.3	6.2	27.2	6.6	25.9
OpenAI-o3	5.4	24.9	0.0	23.5	3.7	30.0	15.7	51.7	12.3	37.8	8.6	36.2
Search Engine Agents												
OpenAI-o3	0.0	17.6	0.0	18.2	7.4	28.0	13.5	37.6	3.1	20.6	6.0	26.1
Browser-Use Agents												
Qwen-VL-Max	0.0	3.0	1.8	2.1	0.0	3.7	1.1	10.3	0.0	10.8	0.7	6.8
DeepSeek-V3.1	0.0	2.4	0.0	3.5	1.9	8.0	3.4	16.4	3.1	6.6	2.0	8.7
Gemini-2.5-Flash	0.0	0.8	0.0	1.6	0.0	0.9	0.0	5.6	0.0	3.5	0.0	3.0
Gemini-2.5-Pro	2.7	10.0	1.8	5.8	0.0	6.1	7.9	23.3	1.5	20.8	3.3	14.7
Claude-3.7-Sonnet	0.0	12.2	0.0	5.4	1.9	12.4	10.1	30.0	1.5	7.5	3.6	15.2
Claude-4.0-Sonnet	0.0	10.0	0.0	2.6	0.0	8.3	9.0	24.2	1.5	9.7	3.0	12.4
OpenAI-o3	0.0	13.0	0.0	4.7	3.7	11.7	13.5	31.5	1.5	16.2	5.0	17.3
GPT-4.1	0.0	14.6	1.8	0.7	0.0	14.3	11.2	36.9	1.5	19.5	4.0	20.8
GPT-5	0.0	8.9	0.0	1.8	1.9	4.8	6.7	15.4	3.1	15.1	3.0	10.1
Multi-Agent Systems												
Qwen3-235B	0.0	8.4	0.0	3.3	0.0	6.7	2.2	19.0	0.0	12.3	0.6	11.1
DeepSeek-V3.1	0.0	7.3	0.0	1.4	0.0	3.7	5.6	12.6	3.1	11.7	2.3	8.0
OpenAI-o3	2.7	21.6	1.8	12.6	3.7	20.0	16.9	44.5	6.2	25.2	7.6	27.2
GPT-5	2.7	14.6	7.0	9.5	0.0	5.6	5.6	23.3	9.2	18.3	5.3	15.4

configuration achieves more than a 15% success rate or a 40% completion rate. This underscores the challenging nature of VeriWeb, which involves large-scale retrieval, multi-hop reasoning, and complex information synthesis. We analyze the results from three perspectives: foundation model capability, agent paradigm, and domain-specific behavior.

Foundation Model Comparison. Performance differences across foundation models are substantial. Within deep research agents, OpenAI-o3 and Gemini-2.5-Pro stand out, demonstrating relatively strong reasoning and task generalization, while OpenAI-o4-mini lags behind. For search engine agents, our results include only OpenAI-o3; its absolute performance is modest. In browser-use agents, Gemini-2.5-Flash and Qwen-VL-Max demonstrate limited effectiveness. Among multi-agent systems, OpenAI-o3 again yields the best results, while Qwen3-235B and DeepSeek-V3.1 struggle significantly. Interestingly, GPT-5 shows only moderate gains, pointing to open challenges in translating foundation model strength into effective agentic performance.

Impact of Agent Paradigms. The paradigm adopted by the agent strongly influences performance. Search engine agents, constrained to passive retrieval, typically achieve the lowest success rates. For example, OpenAI-o3 as a search engine agent underperforms its deep research and multi-agent counterparts. Browser-use agents leverage webpage structure and simulated interactions, leading to somewhat better CR, though gains are uneven across models. Deep research agents demonstrate the highest overall CR, benefiting from stronger retrieval and summarization pipelines, with models like Gemini-2.5-Pro and OpenAI-o3 maintaining relatively balanced performance. Multi-agent systems show potential, as collaborative reasoning boosts robustness in certain domains.

Performance Across Domains. We also examine performance across the five domains to understand how content type affects agent effectiveness. Tasks in *arts and entertainment* generally saw the highest success and completion rates, likely due to the relatively clear and concrete nature of the information required. In contrast, domains like *finance and economics* and *social policy and sustainability* were more challenging, often requiring the agent to process fragmented, abstract information from less standardized content. Most models performed poorly in these areas. The *scientific and academic research* and *technology and innovation* domains presented intermediate difficulty, involving complex technical descriptions or multi-attribute reasoning. These patterns indicate that the complexity of information presentation plays a crucial role in information-seeking tasks.



(a) Deep Research SR (b) Deep Research CR (c) Search Engine SR (d) Search Engine CR

(e) Browser-Use SR (f) Browser-Use CR (g) Multi-Agent SR (h) Multi-Agent CR

Figure 4: Distribution of task success rate (SR) and completion rate (CR) on VeriWeb.

Table 3: Comparison of different settings on VeriWeb

Model	Breadth-Oriented		Depth-Oriented		Overall	
	SR (%)	CR (%)	SR (%)	CR (%)	SR (%)	CR (%)
Deep Research Agents						
Gemini-2.5-Flash	17.5	43.1	7.6	36.1	6.9	31.2
OpenAI-o4-mini	17.2	37.4	8.3	43.4	6.6	25.9
Search Engine Agents						
OpenAI-o3	18.2	38.7	10.6	44.3	6.0	26.1
Browser-Use Agents						
GPT-4.1	17.2	20.3	12.3	43.1	4.0	20.8

Table 3 shows the performance comparison between breadth-oriented and depth-oriented settings. Breadth-oriented tasks focus on large-scale information retrieval, where the primary goal is to gather a wide range of relevant information efficiently. These tasks are generally less demanding in reasoning capabilities, as they emphasize efficiency in collecting broad data without requiring extensive multi-step analysis. In this particular setting, models achieve relatively higher SR, though their overall performance remains notably unsatisfactory. By contrast, depth-oriented tasks are significantly more complex, requiring deeper reasoning and sophisticated multi-step processes. Here, agents must not only retrieve information but also understand and synthesize it across several sequential steps, which leads to consistently lower SR across all models. Taken together, the experimental results highlight that breadth-oriented tasks prioritize large-scale retrieval, whereas depth-oriented tasks demand handling more intricate, context-dependent information. Both settings remain highly challenging for current agents.

4.3 Analysis

Analysis of Task Difficulty. To better understand the intrinsic difficulty of tasks in VeriWeb, we conduct a fine-grained statistical analysis of SR and CR distributions across all tasks, comparing results from different agent frameworks. The distribution curves in Fig. 4 reveal that for both agent types, the majority of tasks yield low SR and CR values, with a long tail of near-zero success, underscoring the challenge of VeriWeb's long-chain requirements. To systematically categorize task difficulty, we define five levels based on the average SR and CR across all models and agents: (1) Level 1 includes tasks with SR above 0%, indicating they are relatively tractable for current agents. (2) Level 2 includes tasks with zero SR but CR above 20%. (3) Level 3 includes tasks with zero SR but CR between 5% and 20%. (4) Level 4 includes tasks with zero SR but CR between 0% and 5%. (5) Level 5 includes tasks where both SR and CR are zero, indicating no model was able to make progress. The results in Fig. 5 show that most tasks fall into Levels 2-5 with zero SR, indicating high-complexity tasks. Only a minority of tasks fall into Level 1, suggesting that relatively few tasks are easily solvable. This categorization provides a practical framework for benchmarking future agent progress.

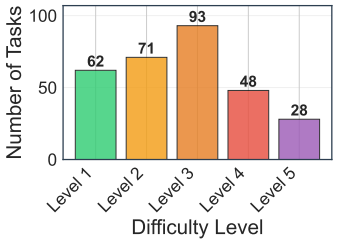


Figure 5: Task Difficulty Level.

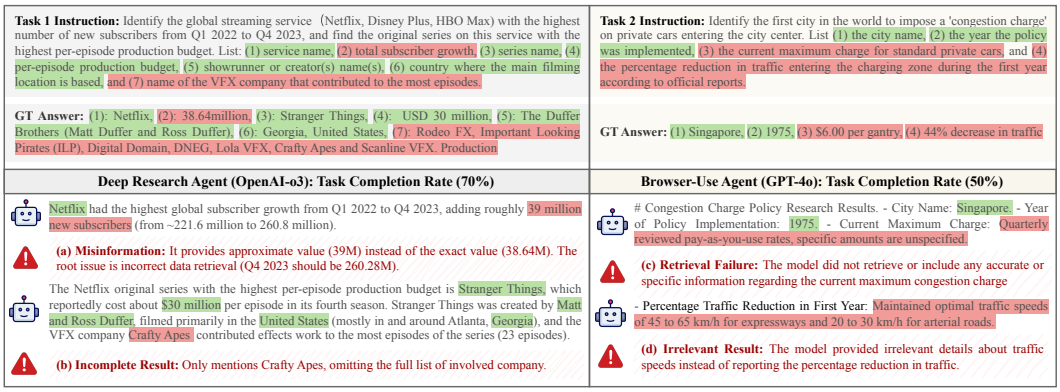


Figure 6: Case studies of agent performance on two information-seeking tasks in VeriWeb.

Analysis of Action Efficiency. The analysis of action efficiency reveals notable differences in the searching strategies of browser-use agents powered by various foundation models. As shown in Tab. 4, models such as Gemini-2.5-Flash and Qwen-VL-Max generally required more actions, suggesting a more exploratory style. In contrast, models like GPT-5 and GPT-4.1 tended to accomplish tasks with fewer steps, indicating a more direct strategy. However, lower action counts did not necessarily correlate with higher success rates. Conversely, higher action counts sometimes reflected more thorough exploration, which proved advantageous in tasks involving complex or ambiguous objectives.

Table 4: Comparison of the action count for browser-use agents on VeriWeb.

Model	Average AC
Qwen-VL-Max	56.4
DeepSeek-V3.1	45.5
Gemini-2.5-Flash	65.7
Gemini-2.5-Pro	42.0
Claude-3.7-Sonnet	23.5
Claude-4.0-Sonnet	16.5
OpenAI-o3	18.0
GPT-4.1	28.6
GPT-5	15.3

4.4 Case Studies

To better understand agent behaviors and limitations in information-seeking tasks, we present two representative cases from VeriWeb in Fig. 6. These examples illustrate retrieval fidelity, multi-hop reasoning quality, and four typical failure modes: (a) misinformation, (b) incomplete result, (c) retrieval failure, and (d) irrelevant result. For Task 1, the agent performed relatively well, correctly identifying the service, series, and most metadata. However, it introduced *misinformation* by reporting an approximate subscriber growth figure instead of the exact value, and gave an *incomplete result* by listing only one VFX company. For Task 2, the agent identified the correct city and year but failed in two key areas. It suffered a *retrieval failure* by not providing a specific congestion charge, and produced an *irrelevant result* by reporting traffic speeds rather than the required percentage reduction. Beyond individual examples, our experiments also reveal several systemic limitations. First, the agents often demonstrate shallow search behavior, invoking tools only a few times and stopping early. This limits their ability to perform comprehensive, multi-hop retrieval. Second, the agents often formulate web queries using full sentences rather than concise keywords, leading to suboptimal results and reduced accuracy.

5 Conclusion

In this work, we introduce VeriWeb, a carefully designed, human-annotated benchmark created to address the growing need for verifiable, long-chain web benchmarks in information-seeking agents. Unlike prior benchmarks that focus on single-fact retrieval and outcome-only validation, VeriWeb emphasizes *long-chain complexity* and *subtask-level verifiability*, supporting the development and evaluation of agent capabilities in real-world search workflows. Extensive experiments across a range of leading agent models highlight persistent challenges in large-scale retrieval and synthesis, underscoring the importance of benchmarks like VeriWeb in pushing the frontier of generalist agent intelligence. Future work will focus on scaling VeriWeb with broader data coverage and exploring its role as a training dataset for more robust agent models. We hope VeriWeb serves as a valuable resource for the community, fostering further research into agentic information-seeking.

ETHICS STATEMENT

This paper introduces the VeriWeb benchmark to advance the evaluation and development of information-seeking web agents. While such technologies may have broader societal impacts, including potential misuse for biased retrieval or disinformation, our work is limited to building a human-curated evaluation dataset with a focus on factuality and verifiability. We therefore see no foreseeable ethical concerns or violations of the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we clearly present the experimental setting in Section 4.1. Besides, we provide a detailed collection document for human demonstration in Appendix A. Anonymous code and data are available in the supplementary materials. Since the human demonstration data exceeds 100GB, only one sample is provided for reference due to the file size limit.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www.anthropic.com.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, et al. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *arXiv preprint arXiv:2506.13651*, 2025a.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*, 2025b.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025c.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. In *NeurIPS*, volume 36, pp. 28091–28114, 2023.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025.

Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuyi He, Zhiyu Mei, Banghua Zhu, and Yi Wu. Beyond ten turns: Unlocking long-horizon agentic search with large-scale asynchronous rl. *arXiv preprint arXiv:2508.07976*, 2025.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.

Google. Gemini deep research, 2025. URL https://gemini.google/overview/deep-research.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *COLING*, pp. 6609–6625, 2020.

Mengkang Hu, Yuhang Zhou, Wendong Fan, Yuzhou Nie, Bowei Xia, Tao Sun, Ziyu Ye, Zhaoxuan Jin, Yingru Li, Qiguang Chen, et al. Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation. *arXiv preprint arXiv:2505.23885*, 2025.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, Jianye Hao, et al. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, pp. 1601–1611, 2017.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litu Ou, Dingchu Zhang, Xixi Wu, Jialong Wu, et al. Websailor-v2: Bridging the chasm to proprietary agents via synthetic data and scalable reinforcement learning. *arXiv preprint arXiv:2509.13305*, 2025a.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, et al. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*, 2025b.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*, 2025c.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*, 2025d.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, et al. Webweaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*, 2025e.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 2022.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *ICLR*, 2024.

MoonshotAI. Kimi-researcher, 2025. URL https://moonshotai.github.io/Kimi-Researcher.

Magnus Müller and Gregor Žunič. Browser use: Enable ai to control your browser, 2024. URL https://github.com/browser-use/browser-use.

OpenAI. Deep research system card, 2025. URL https://cdn.openai.com/deep-research-system-card.pdf.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *EMNLP Findings*, 2024.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, et al. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *arXiv preprint arXiv:2509.13309*, 2025.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*, 2025a.

Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. R1-searcher++: Incentivizing the dynamic knowledge acquisition of llms via reinforcement learning. *arXiv preprint arXiv:2505.17005*, 2025b.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*, 2025.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, et al. Webshaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

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. *arXiv preprint arXiv:2504.12516*, 2025.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, et al. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*, 2025a.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*, 2025b.

Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Yong Jiang, Pengjun Xie, Fei Huang, et al. Resum: Unlocking long-horizon search intelligence via context summarization. *arXiv preprint arXiv:2509.13313*, 2025c.

xAI. Grok 4, 2025. URL https://x.ai/news/grok-4.

Yunjia Xi, Jianghao Lin, Yongzhao Xiao, Zheli Zhou, Rong Shan, Te Gao, Jiachen Zhu, Weiwen Liu, Yong Yu, and Weinan Zhang. A survey of llm-based deep search agents: Paradigm, optimization, evaluation, and challenges. *arXiv preprint arXiv:2508.05668*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pp. 2369–2380, 2018.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *NeurIPS*, volume 35, pp. 20744–20757, 2022.

Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. The landscape of agentic reinforcement learning for llms: A survey. *arXiv preprint arXiv:2509.02547*, 2025.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*, 2025.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, et al. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*, 2025.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.

Appendix

TABLE OF CONTENTS

A HUMAN DEMONSTRATION COLLECTION

A.1 ACTION SPACE

For human demonstrations, the action space $\mathcal{A}$ defines a unified set of web operations applicable across web tasks, as shown in Tab. 5. These actions cover common interaction modalities such as clicks, input, and key events. During execution, the agent selects one action per step from this action set. In some cases, the `result_state()` action is used by the model to output the final result. The specific mapping between the actions recorded during data collection and the web actions is provided as follows.

Table 5: Human Action in VeriWeb

Action Type	**Description**
`left_click([x, y])`	Left-click at given coordinates
`right_click([x, y])`	Right-click at given coordinates
`drag([x, y])`	Drag at given coordinates
`scroll()`	Scroll vertically
`input(text)`	Type a string input
`key_down(key)`	Press down a single key
`result_state()`	Output result statement

We develop a screen capture tool to support human annotators in collecting detailed task trajectories. Each recorded trajectory logs all mouse and keyboard events, which can be systematically mapped to the predefined action space. The mapping is as follows:

- Scroll wheel events (`WHEEL`) are mapped to the `scroll` action.
- Key press events (`KEY_DOWN`) are mapped to the `key_down` action.
- Text input events (`INPUT`) are mapped to the `input` action.
- Text output events (`RESULT_STATE`) are mapped to the `result_state` action.
- Right-click context menu events (`CONTEXT_MENU`) are mapped to the `right_click` action.
- Tab switching events (`TAB_CHANGE`) are interpreted to the `left_click` action at the corresponding coordinates.
- Mouse drag actions (`MOUSE_DRAG`) are mapped to the `drag` action.
- If a `MOUSE_DOWN` event is not followed by a `MOUSE_DRAG` event, it is interpreted as the `left_click` action.
- Additionally, `MOUSE_UP` events are recorded to help determine the end of drag actions or validate click completions, although they are not directly mapped to any action in the defined space.

This mapping ensures consistency between the raw recorded interactions and the unified action space $\mathcal{A}$, enabling accurate interpretation and reproduction of user behaviors by the model during both training and inference.

A.2 DATA COLLECTION DOCUMENT

This section outlines the Standard Operating Procedure (SOP) for data collection tasks.

A.2.1 TASK CONTENT

- Based on the instruction tasks provided by the platform, find relevant results and complete a step-by-step recording of all search processes.
- We have provided subtask steps that need to guide the retrieval operations according to these steps.

A.2.2 TASK WORKFLOW

- Claim tasks within the platform.

- Answer sub-instruction questions step by step, ensuring that answers are obtained from reliable data sources found through keyword searches related to the sub-instructions. Appropriate modification/merging of sub-instruction content is allowed to adjust search logic (modified sub-instructions must be complete sentences).
- Summarize all sub-instruction answers and fill in the required answer content for the instruction.
- Open the plugin to start recording the task retrieval process. No AI tools or large language models are allowed during recording
- After completing the recording, fill in the corresponding results and sub-results in English.

A.2.3 COLLECTION REQUIREMENTS

Content Information Requirements

- Ensure that retrieved answer information corresponds to instruction/sub-instruction content.
- When modifying instructions/sub-instructions, the modified content must be a complete sentence.
- The recorded screen must show relevant data required by instructions/sub-instructions. For data analysis, the integration process must be demonstrated:
  - When relevant data is scattered across different tables, all table data screens need to be retrieved and scrolled through for recording.
  - Extract relevant page information and place it in an online Excel spreadsheet, recording the data processing procedure.

Recording Detail Requirements

- No large language models are allowed for finding answers during recording.
- Ensure the recording interface is entirely in English, including browser language and input content.
- Ensure normal retrieval flow during recording without extraneous irrelevant search operations.
- Content that needs to be input during recording can be directly copied and pasted.
- Before recording each sub-instruction, create a browser with only one page at google.com. When searching questions, enter google.com in the browser's URL bar to access the Google interface for searching. Direct searching in the top red URL bar is prohibited.
- When relevant answer information is found during recording, it must be highlighted with the mouse.
- Ensure clean and concise recording screen (shortcut key: Ctrl+Shift+N). Recording screen cannot have other plugin pop-ups (e.g., translation software pop-ups, though small amounts of advertising are allowed).
- Ctrl+F search operations are not allowed during recording.

Search Source Credibility Hierarchy

Data retrieval logic priority decreases from top to bottom:

- Directly related enterprise/brand official websites.
- Global organization/government official websites and database websites.
- Wikipedia.
- Niche database official websites.
- News reports.
- Other websites.

A.3 DATA REVIEW DOCUMENT

This section outlines the Standard Operating Procedure (SOP) for data review tasks.

A.3.1 REVIEW CONTENT

Review Principles

- Ensure data answer accuracy.
- Ensure no illogical search processes appear.
- Ensure no Chinese appears (excluding cases where data itself requires searching for Chinese answers, even so, the search process must only use English).

Review Recording Logic

- Cannot directly search for answers to reach conclusions; the recording task environment involves searching and recording under unknown conditions.
- Ensure recorded answers correspond to instruction/subtask content; data serves instructions rather than randomly recording irrelevant information.
- Information in corresponding web pages browsed in the video contradicts the given results (e.g., task requires counting occurrences of certain data, but results are inconsistent with comprehensive statistical data on web pages).
- If video cannot confirm whether browsed web page information can support results, verification through links to corresponding web pages is required.

Review Video Content

- Unless required by task, recorded web pages cannot contain Chinese (except when instructions require querying Chinese results, but search recording must still use English).
- No operations in the top address bar.
- Cannot directly search for answer results (recording without logic).
- Cannot show large amounts of small advertisements (more than 3 advertisements on the same page is considered large amounts).
- Cannot expose bookmark bars, other irrelevant plugins (including translation plugins), etc.
- Interface may show large language models, but cannot reference/use large language model answers during recording.
- Key information found must be highlighted with mouse selection (if any sub-answer that doesn't require calculation is not highlighted, this item fails).

Review Text Content

- Replace corresponding demonstrative pronouns with proper nouns in sub-instructions.
- All groups must be merged; results must be properly grouped.
- Answers must be accurately grouped according to quantity.
- Confirm whether answers requiring counting are correct.
- Confirm whether calculated growth, ratios, differences, and other data are correct.

A.3.2 ADDITIONAL NOTES

- Top URL bar only allows entering google.com. All other searches must be conducted within the Google search box.
- AI summary content (Google's built-in Gemini) and directly clicking related web pages are not allowed.
- Based on actual data information obtained and search logic, modify and adjust instruction/subtask field content. For example:
  - When pronouns involve one or two pieces of information, directly replace the pronoun "the country" with the specific country found.

  - When pronouns involve large amounts of information, directly add English parentheses after the pronoun "the country" and fill in corresponding information.

- If instructions require finding "bond name, issuance amount, underwriter, issuing company, use of proceeds, bond rating, and issuance date" for certain content, results can be filled in 7 lines, with each answer separated by "English comma + space".

B Detailed Experimental Settings

B.1 Agent Prompt

The agent prompts for different agents shown below:

Deep Research Agent Prompt

{question}

Search Engine Agent Prompt

You are the EXECUTOR agent. You will receive one task description at a time. Your role is to complete the task efficiently, using available tools via function calls when necessary.

Guidelines:
- Always think step by step before responding.
- Provide concise answers.
- If a tool is needed, respond only with the function call — no extra text.
- When the task is complete, respond with: FINAL ANSWER: [your answer here]

Browser-Use Agent Prompt

We follow the official agent prompt from Browser-Use (Müller & Žunič, 2024).

OWL Agent Prompt

You are a helpful assistant that can search the web, extract webpage content, simulate browser actions, and provide relevant information to solve the given task.

You are now working in 'working_dir'. All your work related to local operations should be done in that directory.

Mandatory Instructions
1. **Take Detailed Notes**: You MUST use the 'append_note' tool to record your findings. Ensure notes are detailed, well-organized, and include source URLs. Do not overwrite notes unless summarizing; append new information. Your notes are crucial for the Document Agent.

Web Search Workflow
1. **Initial Search**: Start with a search engine like 'search_google' or 'search_bing' to get a list of relevant URLs for your research if available.
2. **Browser-Based Exploration**: Use the rich browser toolset to investigate websites.
- **Navigation**: Use 'visit_page' to open a URL. Navigate with 'click', 'back', and 'forward'. Manage multiple pages with 'switch_tab'.
- **Analysis**: Use 'get_som_screenshot' to understand the page layout and identify interactive elements. Since this is a heavy operation, only use it when visual analysis is necessary.
- **Interaction**: Use 'type' to fill out forms and 'enter' to submit.

3. **Detailed Content Extraction**: Prioritize using the scraping tools from 'Crawl4AIToolkit' for in-depth information gathering from awebpage.

Guidelines and Best Practices
- **URL Integrity**: You MUST only use URLs from trusted sources (e.g., search engine results or links on visited pages). NEVER invent or guess URLs.
- **Thoroughness**: If a search query is complex, break it down. If a snippet is unhelpful but the URL seems authoritative, visit the page. Check subpages for more information.
- **Persistence**: If one method fails, try another. Combine search, scraper, and browser tools for comprehensive information gathering.
- **Collaboration**: Communicate with other agents using 'send_message' when you need help. Use 'list_available_agents' to see who is available.
- **Clarity**: In your response, you should mention the URLs you have visited and processed.

B.2 LLM-AS-A-JUDGE PROMPT

For web tasks, the goal is defined as obtaining a correct textual answer through multi-turn information retrieval and reasoning. Thus, we use GPT-4.1 as a judge to semantically evaluate the correctness of agents' final answers based on the question, ground truth, and model response, and report the LLM-as-a-Judge score. The detailed evaluation prompt is provided as follows:

LLM-as-a-Judge Prompt for Web Task

You are a strict evaluator assessing answer correctness. You must score the model's prediction on a scale from 0 to 10, where 0 represents an entirely incorrect answer and 10 indicates a highly correct answer.

Input
Question:
```
{question}
```
Ground Truth Answer:
```
{answer}
```
Model Prediction:
```
{pred}
```

Evaluation Rules
- The model prediction may contain the reasoning process, you should spot the final answer from it.
- Assign a high score if the prediction matches the answer semantically, considering variations in format.
- Deduct points for partially correct answers or those with incorrect additional information.
- Ignore minor differences in formatting, capitalization, or spacing since the model may explain in a different way.
- Treat numerical answers as correct if they match within reasonable precision
- For questions requiring units, both value and unit must be correct

Scoring Guide
Provide a single integer from 0 to 10 to reflect your judgment of the answer's correctness.

Strict Output format example

4

C Detailed Task Analysis

C.1 Task Example

The web tasks focus on deep research requiring multi-turn information retrieval and reasoning. In VeriWeb, these tasks span five key thematic domains: scientific and academic research; finance and economics; technology and innovation; arts and entertainment; and social policy and sustainability. Below are some examples of web tasks.

Web Task Example - Scientific and Academic Research

Task Instruction
Identify the earliest known warship that sank on its maiden voyage. Provide the vessel's commonly accepted name, estimated sinking year or century, salvage year, location described by sea area and the exclusive economic zone of the country it lies in, as well as the museum currently displaying the wreck and the official name of any anchor-related artifacts from it in the museum's collection.

Task Answer
Shipwreck: Vasa
Year of sinking: 1628
Salvage year: 1961
Location: Stockholm, Sweden
Museum: Vasa Museum
Artifact: Ankarstock

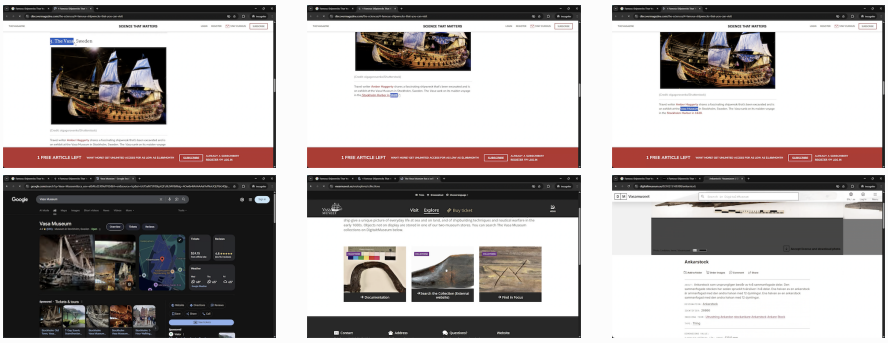


Figure 7: Human demonstration screenshots for the scientific and academic research task example.

Subtask 1 Instruction
Compile a list of major shipwreck discoveries where the primary search and identification technology used was multibeam sonar.

Subtask 1 Answer

Name	Year
USS Kittiwake	2011
C-50 Naufragio Vicente Palacio Riva Ship	2000
The Vasa	1961
The Lusitania	1915

Subtask 2 Instruction
For each shipwreck on the list, find its estimated sinking date (year or century). Identify the wreck with the earliest sinking date.

Subtask 2 Answer

Name	Year
USS Kittiwake	1994
C-50 Naufragio Vicente Palacio Riva Ship	2000
The Vasa	1628 (oldest)
The Lusitania	1906

Subtask 3 Instruction
Confirm the full name of the organization, or institution responsible for the discovery of the Vasa.

Subtask 3 Answer
Organization: Vasa Museum

Subtask 4 Instruction
Determine The Vasa's location, specifying the sea or ocean body and the Exclusive Economic Zone (EEZ) of the relevant coastal nation.

Subtask 4 Answer
Location: Stockholm, Sweden

Subtask 5 Instruction
Search museum databases and archaeological reports to find a museum that currently exhibits The Vasa.

Subtask 5 Answer
Museum: Vasa Museum

Subtask 6 Instruction
From the Vasa Museum's official collection catalog or website, find the official name of the specific artifact related to the anchor stock in the museum's collection.

Subtask 6 Answer
Name: Ankarstock

Web Task Example - Finance and Economics

Task Instruction
Among all Chinese banks listed in Hong Kong from 2022 to 2023, list the bank with the highest increase in net interest margin ranking, and provide: (1) bank name, (2) net interest margin values before and after the increase, (3) stock code, (4) total asset growth rate, and (5) chairman's name.

Task Answer
Bank Name: Hang Seng Bank
Net Interest Margin Values Before and After the Increase: 1.75%, 2.30%
Stock Code: 0011.HK
Total Asset Growth Rate: -8.75%
Chairman's Name: Irene Lee

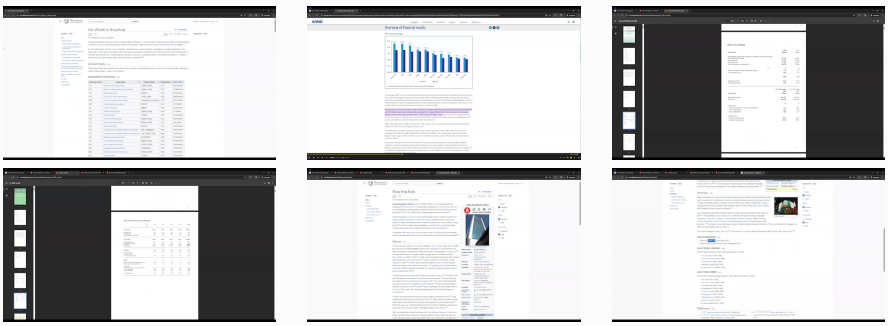


Figure 8: Human demonstration screenshots for the finance and economics task example.

Subtask 1 Instruction
Collect the list of the top 10 licensed banks in Hong Kong by total assets in 2024 and their respective annual net interest margin (NIM) data.

Subtask 1 Answer
Hongkong and Shanghai Banking Corporation Limited (The): 10,500,393 HK$ million
Bank of China (Hong Kong) Limited: 3,685,578 HK$ million
Standard Chartered Bank (Hong Kong) Limited: 2,534,695 HK$ million
Hang Seng Bank, Limited: 1,692,094 HK$ million
Industrial and Commercial Bank of China (Asia) Limited: 915,960 HK$ million
Bank of East Asia, Limited (The): 860,361 HK$ million
Nanyang Commercial Bank, Limited: 555,149 HK$ million
China Construction Bank (Asia) Corporation Limited: 493,858 HK$ million
China CITIC Bank International Limited: 470,387 HK$ million
DBS Bank (Hong Kong) Limited: 467,621 HK$ million

Subtask 2 Instruction
Collect the list of the top 10 licensed banks in Hong Kong from 2022 to 2023 and their annual net interest margin data, and calculate the increase in net interest margin for each bank and identify the bank with the largest increase.

Subtask 2 Answer
Bank: Hang Seng Bank
NIM Increase: 55bp

Subtask 3 Instruction
Find the following for the bank: (1) name, (2) specific net interest margin values for 2022 and 2023, (3) stock code.

Subtask 3 Answer
Name: Hang Seng Bank
2022 NIM: 1.75%
2023 NIM: 2.30%
Stock Code: 0011.HK

Subtask 4 Instruction
Find the bank's(Hang Seng Bank) total asset data for 2022-2024 and calculate the total asset growth rate.

Subtask 4 Answer
2022 asset data: 1,854.4 HK$bn
2023 asset data: 1,692.1 HK$bn
Growth rate: -8.75%

Subtask 5 Instruction
Find the current chairman's name of the bank(Hang Seng Bank).

Subtask 5 Answer
Chairman: Irene Lee

Web Task Example - Technology and Innovation

Task Instruction
Identify the pharmaceutical company that had the FDA-approved new molecular entities (NMEs) between 2020 and 2024, where at least one of these drugs achieved blockbuster status (over $1 billion in annual sales) within 24 months of approval. List the company name, total number of NMEs approved, the name and indication of the fastest blockbuster drug, its peak annual sales figure, and the name and specialization of the lead scientist credited with its discovery.

Task Answer
Company Name: Pfizer Inc.
Total NME Approvals: 9
Details of the company with the largest number of approvals: Approval date drug trade name drug generic name 2021-11-05 Paxlovid™ nirmatrelvir/ritonavir, 2022-05-25 Cibinqo™ abrocitinib, 2023-01-30 Zavzpret® zavegepant, 2023-05-25 Paxlovid nirmatrelvir/ritonavir, 2023-06-05 Litfulo ritlegepitinib, 2023-08-22 Penbraya™ pentavalent meningococcal, 2023-10-12 Velsipity™ etrasimod, 2024-03-14 Rezdiffra* resmetirom, 2023-03-09 Zavzepant* zavegepant
Fastest Blockbuster Drug: Paxlovid (nirmatrelvir/ritonavir)
Indication: treatment of mild-to-moderate COVID-19 in adults and pediatric patients (12 years of age and older weighing at least 40 kg) who are at high risk for progression to severe COVID-19
Peak Annual Sales: $18.933 billion (2022)
Lead Scientist: Dafydd Owen
Specialization: medicinal chemist in the design and synthesis of drug-like molecules

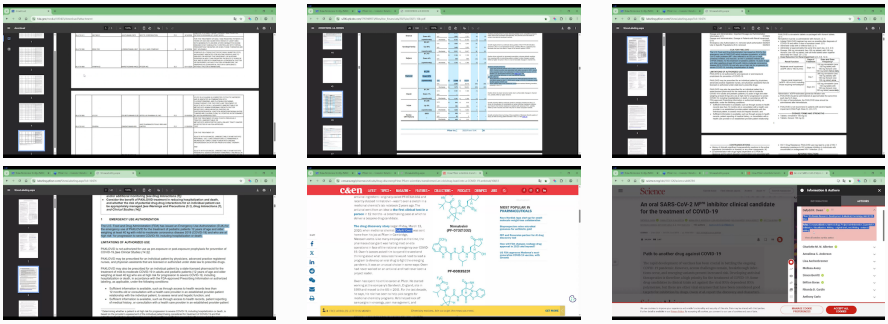


Figure 9: Human demonstration screenshots for the technology and innovation task example.

Subtask 1 Instruction
Compile a statistical summary of all FDA-approved NMEs (2020-2024), identify the company with the highest number of approvals, and report its approved drugs with both brand and generic names.

Subtask 1 Answer
The ranking of the number of NME approvals of the company:

	Company name	Approved quantity
1	Pfizer Inc.	9
2	Novartis Pharmaceuticals	7
3	Bristol Myers Squibb	6
4	Merck Sharp & Dohme	5
5	Takeda Pharmaceuticals	4
6	Eli Lilly and Company	3

Company Name: Pfizer Inc.
Total NME Approvals: 9
Details of the company with the largest number of approvals:

Approval date	drug Trade Name	Generic Name
2021-11-05	Paxlovid™	nirmatrelvir/ritonavir
2022-05-25	Cibinqo™	abrocitinib
2023-01-30	Zavzpret®	zavegepant
2023-05-25	Paxlovid	nirmatrelvir/ritonavir
2023-06-05	Litfulo	ritlegepitinib
2023-08-22	Penbraya™	pentavalent meningococcal
2023-10-12	Velsipity™	etrasimod
2024-03-14	Rezdiffra*	resmetirom
2023-03-09	Zavzepant*	zavegepant

Subtask 2 Instruction
Among qualifying companies, identify Pfizer Inc. with the most FDA-approved NMEs and find which of their drugs reached blockbuster status fastest after approval and its peak annual sales figure.

Subtask 2 Answer
Fastest Blockbuster Drug: Paxlovid (nirmatrelvir/ritonavir)
Peak Annual Sales: $18.933 billion (2022)

Subtask 3 Instruction
Find the primary indication for Paxlovid (nirmatrelvir/ritonavir).

Subtask 3 Answer
Indication: treatment of mild-to-moderate COVID-19 in adults and pediatric patients (12 years of age and older weighing at least 40 kg) who are at high risk for progression to severe COVID-19

Subtask 4 Instruction
Search for the lead scientist or principal investigator credited with discovering Paxlovid (nirmatrelvir/ritonavir), including their full name and area of specialization.

Subtask 4 Answer
Lead Scientist: Dafydd Owen
Specialization: medicinal chemist in the design and synthesis of drug-like molecules

Web Task Example - Arts and Entertainment

Task Instruction
Identify the film with the highest production cost return ratio (box office/production cost) among all movies that grossed over $1 billion worldwide between 2020 and 2024, and list its title, director, production cost, global box office, main filming location, as well as the name of the highest-level film award it received and the city where the award ceremony was

held.

Task Answer
Title: The Super Mario Bros. Movie
Director: Aaron Horvath
Production cost: $100,000,000
Global box office: $1,360,847,665
Main filming location: Paris, France
The name of the highest-level film award it received and the city where the award ceremony was held: Festival Film Bandung - Film Impor Terpuji / Commendable Imported Film, Bandung, Indonesia

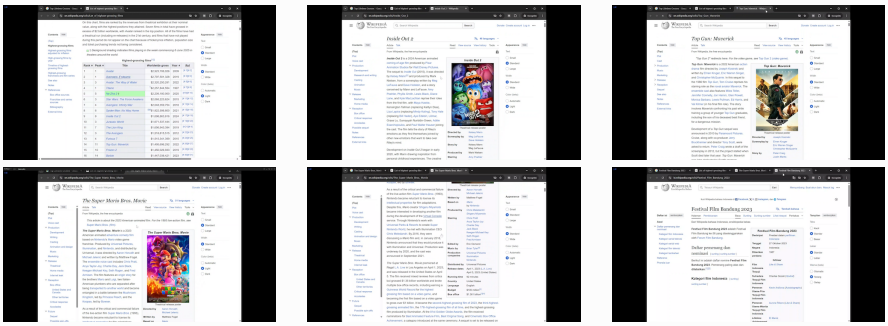


Figure 10: Human demonstration screenshots for the arts and entertainment task example.

Subtask 1 Instruction
Collect a list of all films worldwide with box office earnings exceeding $1 billion from 2020 to 2024, along with their box office data.

Subtask 1 Answer
Avatar: The Way of Water: $2,320,250,281
Inside Out 2: $1,698,863,816
Spider-Man: No Way Home: $1,922,598,800
Top Gun: Maverick: $1,495,696,292
Barbie: $1,447,038,421
The Super Mario Bros. Movie: $1,360,847,665
Deadpool & Wolverine: $1,338,073,645
Moana 2: $1,059,242,164

Subtask 2 Instruction
Search the production cost of each film, and calculate the ratio of box office to production cost to identify the film with the highest return on investment. List only the highest-rated movies and their ratios.

Subtask 2 Answer
The Super Mario Bros. Movie, 13.61

Subtask 3 Instruction
Find the director's name of The Super Mario Bros. Movie, the specific production cost, and the exact global box office revenue.

Subtask 3 Answer
Aaron Horvath, $100,000,000, $1,360,847,665

Subtask 4 Instruction
Search for the main filming locations of The Super Mario Bros. Movie.

Subtask 4 Answer
Paris, France

Subtask 5 Instruction
Find all the film awards that The Super Mario Bros. Movie has received, identify the highest-level award among them, and find the host city of the corresponding award ceremony.

Subtask 5 Answer
Festival Film Bandung - Film Impor Terpuji / Commendable Imported Film, Bandung, Indonesia

Web Task Example - Social Policy and Sustainability

Task Instruction
Identify the G20 country that achieved the largest percentage decrease in CO2 emissions per capita between 2015 and 2023 while simultaneously recording a real GDP growth of over 20% in the same period. List the country's name, its official head of government as of year-end 2023, the primary renewable energy source by installed capacity, and the official title of its most recent Nationally Determined Contribution (NDC) report submitted to the UNFCCC.

Task Answer
The country's name: UK
UK's official head of government as of year-end 2023: Rishi Sunak
The primary renewable energy source by installed capacity: wind sources
The official title of its most recent Nationally Determined Contribution (NDC) report submitted to the UNFCCC: United Kingdom of Great Britain and Northern Ireland's 2035 Nationally Determined Contribution

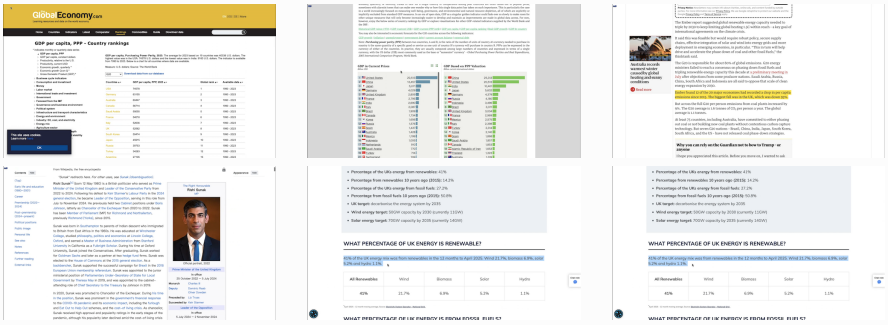


Figure 11: Human demonstration screenshots for the social policy and sustainability task example.

Subtask 1 Instruction
Identify the G20 country that achieved a real GDP growth of over 20% between 2015 and 2023.

Subtask 1 Answer
USA, China, india, UK, Brazil, Russia, Canada, Mexico, Indonesia, Turkey, Saudi Arabia, Argentina

---

Subtask 2 Instruction
Identify the country with the largest percentage decrease in CO2 emissions per capita.

Subtask 2 Answer
The country name: UK

Subtask 3 Instruction
Find the full name of the head of government for UK, who was in office on December 31, 2023.

Subtask 3 Answer
The full name: Rishi Sunak

Subtask 4 Instruction
Research the energy profile of UK to determine its primary renewable energy source based on the latest available data for installed capacity (in MW or GW).

Subtask 4 Answer
UK‘s primary renewable energy source: wind sources

Subtask 5 Instruction
Search the official UNFCCC registry or UK’s national environmental ministry website to find its most recently submitted Nationally Determined Contribution (NDC) report. Record its full official title and the year it was published/submitted.

Subtask 5 Answer
The official title of its most recent Nationally Determined Contribution (NDC) report submitted to the UNFCCC: United Kingdom of Great Britain and Northern Ireland’s 2035 Nationally Determined Contribution

D THE USE OF LLMS

In this work, LLMs were used exclusively for two auxiliary purposes: (1) LLMs assisted with minor language polishing and stylistic refinement. (2) LLMs were used as auxiliary tools in task construction and preliminary review, while humans designed the framework, curated the data, and performed the final validation. Note that LLMs did not contribute to research ideation, methodology design, experimental design, or data analysis. The authors take full responsibility for all content.