WebArXiv: Evaluating Multimodal Agents on Time-Invariant arXiv Tasks

Anonymous ACL submission

Abstract

Recent progress in large language models (LLMs) has enabled the development of autonomous web agents capable of navigating and interacting with real websites. However, evaluating such agents remains challenging due to the instability and inconsistency of existing benchmarks, which often rely on dynamic content or oversimplified simulations. In this work, we introduce WebArXiv, a static and time-invariant benchmark comprising 275 web-based tasks grounded in the arXiv platform. WebArXiv ensures reproducible and reliable evaluation by anchoring tasks in fixed web snapshots with deterministic ground truths and standardized action trajectories. Through behavioral analysis, we identify a common failure mode, Rigid History Reflection, where agents over-rely on fixed interaction histories. To address this, we propose a lightweight dynamic reflection mechanism that allows agents to selectively retrieve relevant past steps during decision-making. We evaluate ten state-of-the-art web agents on WebArXiv. Results demonstrate clear performance differences across agents and validate the effectiveness of our proposed reflection strategy. We release our open-sourced code at `https://anonymous.4open.science/r/74E4423BVNW`.

1 Introduction

The rapid advancement of large language models (LLMs), such as GPT-4 (OpenAI, 2023a) and Gemini (Georgiev et al., 2023), has led to the emergence of autonomous web agents capable of performing complex tasks on real-world websites (Garg et al., 2025). These agents combine vision-language reasoning with interactive decision-making to automate activities such as academic search (He et al., 2024a), job applications, and e-commerce navigation (Verma et al., 2024). As their applications expand across domains, the need for systematic evaluation protocols becomes increasingly critical (Yehudai et al., 2025). Reliable benchmarks are essential not only for measuring progress, but also for enabling reproducible research and supporting reinforcement learning-based agent training (Le Sellier De Chezelles et al., 2024; Song et al., 2025).

Despite recent efforts to develop frameworks for web agents, existing benchmarks face key limitations. Many tasks rely on real-time web content, which continuously evolves, resulting in volatile answers and unstable ground truths (Pan et al., 2024; Yoran et al., 2024). For example, benchmarks like WebVoyager (He et al., 2024b) operate on live websites, where answers to tasks such as "How many recent papers mention X?" or "What are the latest arXiv news" change frequently. Other benchmarks such as Mind2Web (Deng et al., 2023) and WebArena (Zhou et al., 2024) adopt simplified simulators or fixed action traces, which fail to reflect the dynamic complexity of real browsing environments. These limitations give rise to two major challenges for real-environment benchmarks: (1) Ground truth instability: Many tasks depend on live or frequently updated web content, leading to answer drift over time. This results in inconsistent or outdated labels, which hinders reproducible supervision and undermines the validity of benchmarks. (2) Evaluation inconsistency: Even with well-defined task objectives, dynamic web environments often cause unpredictable UI behaviors, shifting layouts, and content drift. These factors obscure the source of model failures, making it difficult to attribute errors and hindering fair and consistent comparisons across agents.

To address the aforementioned challenges, we present WebArXiv, a benchmark that supports static and consistent evaluation of web agents. WebArXiv comprises a suite of tasks sourced from the arXiv platform, all grounded in static and time-invariant webpage content. This ensures that task answers remain stable over time, mitigating noise caused by dynamic content drift. In addition,

WebArXiv provides standardized baseline, with prompts, reference action trajectories, and deterministic ground truths, enabling fair comparisons across diverse models in a consistent, real-world environment. All answers are precisely defined and machine-verifiable, eliminating the need for manual inspection and ensuring reliable evaluation unaffected by web drift or API changes.

In analyzing the behavioral patterns of existing web agents, we identified a common failure mode, Rigid History Reflection: most agents retain a fixed number of past interaction steps but fail to assess their relative importance. This often leads to agents attending irrelevant content or repeating previous actions. To investigate this issue, we introduce a lightweight reflection mechanism that enables agents to selectively retrieve the most relevant prior step before making each decision.

Finally, we evaluate ten state-of-the-art large multimodal web agents on the WebArXiv benchmark, such as GPT-4o OpenAI (2024a) and Gemini-2.0 DeepMind (2025). The evaluation results provide a clear view of baseline performance, provides well-aligned experimental comparisons across agents, and empirically demonstrates the effectiveness of our proposed reflection mechanism.

Our contributions are summarized as follows:

- We introduce WebArXiv, a static and time-invariant benchmark for evaluating multimodal web agents.
- We propose a lightweight dynamic reflection mechanism to to improve upon rigid history usage in web agent decision-making.
- We conduct a comprehensive evaluation of ten state-of-the-art web agents on WebArXiv, demonstrating clear baseline performance and validating the effectiveness of our method.

2 Related Work

Large language models (LLMs) have continued to demonstrate strong capabilities in reasoning, problem-solving, and natural language understanding (Touvron et al., 2023; Luo et al., 2025). This progress has spurred the development of autonomous LLM-powered agents for complex web navigation tasks, which involve interpreting open-ended instructions and executing multi-step interactions (Gravitas, 2023; Schick et al., 2024; Lutfi Eren Erdogan, 2025). While earlier work focused on controlled or simulated web environments (Chae et al., 2024), recent efforts have shifted toward real-world interfaces, exemplified by benchmarks like Mind2Web (Deng et al., 2023) and WebArena (Zhou et al., 2023).

Emerging agent architectures include text-finetuned agents like WebGPT (Nakano et al., 2023), HTML-pretrained agents such as WebAgent (Iong et al., 2024), and instruction-following agents using lightweight prompting methods for zero-shot decision-making (Yao et al., 2023; Shinn et al., 2023). In multimodal web settings, agents like Pix2Act (Shaw et al., 2023) and WebGUM (Furuta et al., 2024) operate directly on screenshots, while SeeAct (Boyuan Zheng, 2024) further combines visual grounding with tool-enhanced candidate selection.

3 WebArxiv

WebArXiv is a static and time-invariant benchmark with 275 tasks aimed to evaluate web agents' ability to retrieve reliable information from the arXiv platform, covering site info, submission rules, search features, paper metadata, and navigation.

3.1 Benchmark Construction

To construct the WebArXiv dataset, we adopted a hybrid data creation process that combines self-instruct (Kim et al., 2025) with expert-guided refinement. Inspired by WebVoyager, we defined five distinct and temporally stable categories for WebArXiv: (1) Website Information & Organizational Details, (2) Rules, Licensing, and User Account Management, (3) Research Paper Discovery & Retrieval, (4) Advanced Search & Filtering, and (5) Deep Paper Content Extraction.

Human experts drafted 100 candidate tasks for each category with the assistance of LLM-generated exemplars, simulating realistic user queries and task intents. To ensure diversity and minimize semantic overlap, we conducted sentence-level semantic similarity analysis using the `all-mpnet-base-v2` model, followed by manual inspection. After filtering out redundant or overly similar items, 55 high-quality tasks were retained per category.

All final task answers were manually verified by three independent annotators to ensure uniqueness, clarity, and temporal invariance. The resulting dataset provides a reliable and reproducible benchmark for evaluating web agents in a stable academic domain.

Figure 1: WebArXiv task benchmark creation pipeline, illustrating the stages of task generation, LLM filtering, and expert annotation.

3.2 Annotation

For each task, annotators review the agent's full action trajectory, including screenshots and interaction steps to make a binary judgment on task success. To ensure reliability, all tasks are independently reviewed by three annotators to assess inter-annotator agreement.

Task outcomes are labeled as: **Correct**: The retrieved information exactly matches the gold-standard answer. **Incorrect**: The agent provides an incorrect answer or fails to retrieve the required content. **Partial Correct**: The agent's trajectories show that the agent failed is on the right track and almost approaching the last step to find out the answer.

3.3 Dynamic Reflection

Most webagents handles navigation context by retaining the last three interaction steps, capturing recent visual observations and associated text. However, it treats all steps equally, without assessing which is most relevant. This leads to two key issues: in advanced search tasks, the agent often stalls amid dense UI elements; in content-heavy pages, it relies on truncated visible text and overlooks useful prior views—resulting in loops or incomplete answers.

To guide the agent's decision-making at each interaction step, we implement a dynamic reflection mechanism. The model first identifies the most relevant of the last three visual observations for reasoning, then combines this with the current view to form a context for action generation. The selected action is executed, and the interaction history is updated accordingly.

4 Experiment

4.1 Experiment setup

Web Agents We evaluate two categories of web agents: (1) LLM-driven agents, implemented through our developed web agent framework that interacts with general-purpose APIs such as GPT-4o, GPT-4 Turbo, and Gemini-2.5 (OpenAI, 2024b, 2023b; DeepMind, 2024), and (2) specialized web agents, which are explicitly designed for structured web interaction (e.g., SeeAct, LiteWebAgent, OpenWebAgent) (et al., 2023b, 2024, 2023a). Detailed descriptions of these web agents are provided in the Appendix A.

Evaluation Protocol We adopt task success rate as the primary evaluation metric, which measures the proportion of tasks the agent retrieves the correct final answer. Each agent is evaluated on all tasks in the WebArXiv benchmark, and success is determined by comparing the agent's final response with the verified gold-standard answer. The evaluation is conducted under a strict matching criterion to ensure answer accuracy.

We performed each task three times and report the averaged results for ten web agents across five task categories in the WebArXiv benchmark.

4.2 Main Results

WebArXiv provides a fair comparison across varies models with time-invariant arXiv tasks. Experiment shows that performance across categories varied significantly. GPT-o1 achieved the highest scores in Platform Information (72.7%) and Paper Retrieval (65.5%), while Gemini-2.5 excelled in Rules & Accounts (57.3%) and Advanced Search & Filters (47.3%). LiteWebAgent led in Deep Paper Extraction (45.5%). However, Advanced Search & Filters continued to be the most challenging category overall, with only one model exceeding the 45% mark.

These findings further demonstrate that model size alone does not determine performance on WebArXiv. In the controlled setting (static, and time-invariant tasks), the ability to interpret prompts and navigate structured content becomes particularly important. GPT-o1 and Gemini-2.5 likely benefited from more effective prompting and reasoning strategies, while even smaller models like GPT o4-mini achieved competitive results. This highlights that success in structured, knowledge-centric environments depends more on prompt sensitivity and reasoning efficiency than on sheer model scale.

Web Agents	Platform & Org Info	Rules & Accounts	Paper Retrieval	Adv. Search & Filters	Deep Paper Extraction	Total (%)
GPT-4-Turbo	43.6%	34.5%	47.3%	25.8%	30.9%	36.4%
GPT-4o	36.1%	29.6%	34.5%	25.7%	38.2%	32.7%
GPT-o1	**72.7%**	50.3%	**65.5%**	43.2%	44.5%	**56.7%**
GPT-o4-mini	52.7%	48.2%	56.4%	29.1%	32.7%	43.8%
Gemini-1.5-pro	47.3%	42.2%	52.7%	34.0%	37.8%	42.9%
Gemini-2.0	34.5%	29.1%	34.8%	25.2%	27.3%	30.6%
Gemini-2.5	65.2%	**57.3%**	52.7%	**47.3%**	35.4%	51.1%
SeeAct	28.2%	20.0%	25.7%	20.8%	24.9%	23.6%
LiteWebAgent	43.7%	47.3%	43.4%	32.3%	**45.5%**	44.0%
OpenWebAgent	34.5%	38.9%	43.6%	34.5%	18.2%	33.8%

Table 1: Task success rates across five arXiv task categories for webagent models.

Web Agents	Platform & Org Info	Rules & Accounts	Paper Retrieval	Adv. Search & Filters	Deep Paper Extraction	Total (%)
GPT-4-Turbo	43.6%	34.5%	47.3%	25.8%	30.9%	36.4%
GPT-4-Turbo + dynamic reflection	52.6%	42.7%	46.4%	30.0%	29.1%	40.2%
GPT-4o	36.1%	29.6%	34.5%	25.7%	38.2%	32.7%
GPT-4o + dynamic reflection	63.6%	60.0%	38.2%	34.5%	52.7%	38.4%
GPT-o1	72.7%	50.3%	65.5%	43.2%	44.5%	56.7%
GPT-o1 + dynamic reflection	73.3%	55.5%	64.5%	52.7%	60.2%	61.8%
GPT-o4-mini	52.7%	48.2%	56.4%	29.1%	32.7%	43.8%
GPT-o4-mini + dynamic reflection	57.3%	31.8%	52.7%	30.9%	35.5%	41.6%
Gemini-1.5-pro	47.3%	42.2%	52.7%	34.0%	37.8%	42.9%
Gemini-1.5-pro + dynamic reflection	59.7%	59.1%	51.8%	38.2%	45.5%	50.9%
Gemini-2.5	65.2%	57.3%	52.7%	47.3%	35.4%	51.1%
Gemini-2.5 + dynamic reflection	81.8%	72.7%	56.4%	43.6%	41.1%	60.0%

Table 2: Comparison of base models and their dynamic reflection enhanced models across five task categories.

Reflection Mechanism	Successful (↑)	Partial (↓)	Failed (↓)
GPT-4-Turbo last 3 steps	36.4%	18.2%	45.5%
GPT-4-Turbo last 2 steps	34.5%	20.4%	45.2%
GPT-4-Turbo last step	**43.6%**	**14.5%**	**41.8%**
GPT-4-Turbo + dynamic reflection	40.2%	16.3%	43.6%
GPT-o1 last 3 steps	56.7%	16.2%	27.0%
GPT-o1 last 2 steps	58.2%	15.1%	26.8%
GPT-o1 last step	60.0%	14.4%	25.7%
GPT-o1 + dynamic reflection	**61.8%**	**12.7%**	**25.5%**

Table 3: Task success rates of GPT-o1 and GPT-4 Turbo models under different reflection strategies. The baseline uses the last 3 steps to make decisions, while dynamic reflection only use the most relevant steps to make decision.

4.3 Ablation Study

Performance of Dynamic Reflection In Table 2, we compare each base model with its dynamic reflection tuned variant across five task categories. Notably, dynamic reflection o1 achieved the highest overall success rate at 61.8%, outperforming its base version (56.7%) and setting a new benchmark across Platform Information (73.3%) and Deep Paper Extraction (60.2%). Similarly, dynamic reflection Gemini-2.5 reached 60.0%, an 8.9-point improvement over its base (51.1%), with particularly strong gains in Platform Information (81.8%) and Rules & Accounts (72.7%). These improvements show the effectiveness of our dynamic reflection mechanism. The strong and standardized baselines established by WebArXiv enable a fair and transparent comparison, through which we clearly observe the superior robustness and consistency of our approach over a wide range of existing web agents.

Rigid Reflection vs. Dynamic Reflection In Table 3, empirically, dynamic reflection GPT-o1 with dynamic reflection achieved a 61.8% success rate, outperforming simpler baselines using only the last step (60.0%) or uniform three-step memory (56.7%). Similarly, reflection improved dynamic reflection 4-turbo from 36.4% to 40.2%, validating its effectiveness in dynamic decisions under complex UI conditions.

5 Conclusion

We introduced WebArXiv, a static and time-invariant benchmark tailored for evaluating web agents on the arXiv platform. WebArXiv enables consistent, reproducible assessment across models and settings. To further enhance model's decision-making, we proposed a lightweight dynamic reflection mechanism to improve agent performance. Our findings underscore the importance of stable benchmarks and adaptive reflection in advancing real-world, multimodal web agents.

6 Limitation

One limitation of our benchmark is its exclusive focus on the English-language interface of the arXiv platform. This design choice overlooks multilingual versions of the site, which may present different navigation behaviors for non-English users. As a result, the benchmark may not fully capture the challenges faced by web agents operating in multilingual or international contexts. Expanding the benchmark to include tasks in other languages or region-specific interfaces would improve the generalizability of the benchmark and support more inclusive evaluation of web agents designed for a global user base.

7 Ethics Statement

This work introduces a benchmark for evaluating multimodal web agents on static, time-invariant tasks derived from the arXiv platform. All experiments were conducted on publicly available webpages without requiring user authentication or access to private data. No personal, sensitive, or user-generated information was collected or processed during the study. The benchmark tasks are carefully designed to avoid topics that could be ethically sensitive or controversial.

Our dynamic reflection mechanism operates solely on public UI elements and visual context, and does not involve training or fine-tuning on human data beyond publicly released LLMs. Human annotators involved in verifying task outcomes were fully informed of the study's goals and provided explicit consent. Annotations were limited to factual assessments of agent performance and did not require subjective judgments about individuals or user behavior.

References

Jihyung Kil Huan Sun Yu Su Boyuan Zheng, Boyu Gou. 2024. Gpt-4v(ision) is a generalist web agent, if grounded. *arXiv preprint arxiv.org/abs/2401.01614*.

Hyungjoo Chae, Namyoung Kim, Minju Gwak, Gwanwoo Song, Jihoon Kim, Kai Ong, Seonghwan Kim, Dongha Lee, and Jinyoung Yeo. 2024. Web agents with world models: Learning and leveraging environment dynamics in web navigation. In *NeurIPS Workshop on System-2 Reasoning at Scale*.

Google DeepMind. 2024. Gemini 1.5: Technical overview.

Google DeepMind. 2025. Gemini 1.5 and gemini flash: Multimodal models with 1 million token context. https://deepmind.google/technologies/gemini. Accessed: May 2025.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. https://arxiv.org/abs/2306.06070. ArXiv:2306.06070.

Author et al. 2023a. Openwebagent: Open-domain web interaction with llms.

Author et al. 2023b. Seeact: A multi-modal agent for web navigation with visual perception and action. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Author et al. 2024. Litewebagent: A lightweight agent for structured web interaction.

Hiroki Furuta, Kuang-Huei Lee, Ofir Nachum, Yutaka Matsuo, Aleksandra Faust, Shixiang Shane Gu, and Izzeddin Gur. 2024. Multimodal web navigation with instruction-finetuned foundation models. In *International Conference on Learning Representations (ICLR)*.

Siddhant Garg, Harshita Bansal, Yihan Wang, Daniel Khashabi, and Ashish Sabharwal. 2025. Real: Benchmarking autonomous agents on deterministic simulations of real websites. *arXiv preprint arXiv:2504.11543*.

Petko Georgiev, Rohan Anil, and et al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Significant Gravitas. 2023. Auto-gpt: Self-improving ai agent using gpt-4. GitHub repository. https://github.com/Torantulino/Auto-GPT.

Hongjin He, Ning Ding, Xu Han, Zhiyuan Liu, Rui Jiang, Jiawei Yan, and Maosong Sun. 2024a. Pasa: A paper searching agent with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. 2024b. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*.

Iat Long Iong, Xiao Liu, Yuxuan Chen, Hanyu Lai, Shuntian Yao, Pengbo Shen, Hao Yu, Yuxiao Dong, and Jie Tang. 2024. Openwebagent: An open toolkit to enable web agents. In *ACL Demo Track*.

Jungwoo Kim, Minsang Kim, and Sungjin Lee. 2025. Sedi-instruct: Enhancing alignment of language models through self-directed instruction generation. *arXiv preprint arXiv:2502.04774*.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. 2024. The browsergym ecosystem for web agent research. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. 2025. Gfm-rag: Graph foundation model for retrieval augmented generation. *arXiv preprint arXiv:2502.01113*.

Sehoon Kim Suhong Moon Hiroki Furuta Gopala Anumanchipalli Kurt Keutzer Amir Gholami Lutfi Eren Erdogan, Nicholas Lee. 2025. Plan-and-act: Improving planning of agents for long-horizon tasks. *arXiv preprint arXiv:2503.09572*.

Reiichiro Nakano and 1 others. 2023. Webgpt: Browser-assisted question-answering with human feedback. In *ICLR*.

OpenAI. 2023a. Gpt-4 technical report. https://arxiv.org/abs/2303.08774. ArXiv:2303.08774.

OpenAI. 2023b. Gpt-4 turbo overview.

OpenAI. 2024a. Gpt-4o technical report. https://openai.com/index/gpt-4o. Accessed: May 2024.

OpenAI. 2024b. Gpt-4o technical report.

Yuxiang Pan, Difei Kong, Shuyan Zhou, Chuan Cui, Yizhou Leng, Bing Jiang, Haoran Liu, Yujie Shang, Shuchang Zhou, Tong Wu, and Zhaojun Wu. 2024. Webcanvas: Benchmarking web agents in online environments. *arXiv preprint arXiv:2406.12373*.

Timo Schick and 1 others. 2024. Toolformer: Language models can teach themselves to use tools. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Peter Shaw, Mandar Joshi, James Cohan, Jonathan Berant, Panupong Pasupat, Hexiang Hu, Urvashi Khandelwal, Kenton Lee, and Kristina Toutanova. 2023. From pixels to ui actions: Learning to follow instructions via graphical user interfaces. *arXiv preprint arXiv:2306.00245*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yixiao Song, Katherine Thai, Chau Minh Pham, Yapei Chang, Mazin Nadaf, and Mohit Iyyer. 2025. Bearcubs: A benchmark for computer-using web agents. In *Proceedings of the 2025 Conference on Web Intelligence and Autonomous Systems*.

Hugo Touvron and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Gaurav Verma, Rachneet Kaur, Nishan Srishankar, Zhen Zeng, Tucker Balch, and Manuela Veloso. 2024. Adapting multimodal web agents with few-shot learning from human demonstrations. In *Proceedings of the 2024 Conference on Neural Information Processing Systems (NeurIPS)*.

Shinn Yao and 1 others. 2023. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

Ori Yoran, Shoval J Amouyal, Chaitanya Malaviya, Ben Bogin, Omer Press, and Jonathan Berant. 2024. Assistantbench: Can web agents solve realistic and time-consuming tasks? *arXiv preprint arXiv:2407.15711*.

Shuyan Zhou, Frank F Xu, Haozhe Li, Hang Lv, Amanpreet Singh, Alexander Ratner, Anca Dragan, and Chelsea Finn. 2024. Webarena: A realistic web environment for building autonomous agents. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and 1 others. 2023. Webarena: A realistic web environment for building autonomous agents. https://arxiv.org/abs/2307.13854. ArXiv:2307.13854.

A Appendix

A.1 LLM-Driven Agents

These agents use general-purpose large language models (LLMs) capable of processing both textual and visual inputs to interact with web interfaces. They typically operate in an instruction-following manner without explicit environment modeling.

- **GPT-o1:** A state-of-the-art multimodal model developed by OpenAI that accepts both image and text input. We use screenshots of the webpage and natural language instructions as input. Actions are selected via few-shot prompting.

- **GPT-4-Turbo:** A high-efficiency variant of GPT-4 with similar reasoning capabilities but optimized inference latency.

- **Gemini 1.5 / 2.0 / 2.5:** Google DeepMind's multimodal models supporting vision-language understanding. Used in a similar prompting setup as GPT-4o, with instruction + screenshot as input.

- **GPT-4o-mini / GPT-4o:** Versions of GPT 4 models with reduced parameters. Used to test whether compact models can maintain reasonable task performance.

These models do not explicitly track interaction history or webpage state beyond the current screenshot unless specified in the prompt.

A.2 Specialized Web Agents

These models are explicitly designed to operate in structured web environments. They typically rely on DOM parsing, fine-grained action spaces (e.g., click, type), and internal state tracking for reasoning.

- **SeeAct:** A vision-based web agent that combines a perception module (CLIP) with an action decoder. It uses a global planning strategy and allows step-wise interaction with screenshots.

- **LiteWebAgent:** A lightweight web automation agent that parses DOM structures and uses language models to predict high-level actions. It is optimized for speed and interpretability.

- **OpenWebAgent:** A modular web agent architecture with DOM-based environment modeling, visual grounding, and tool-use support. It supports both retrieval-augmented inputs and explicit memory of previous steps.

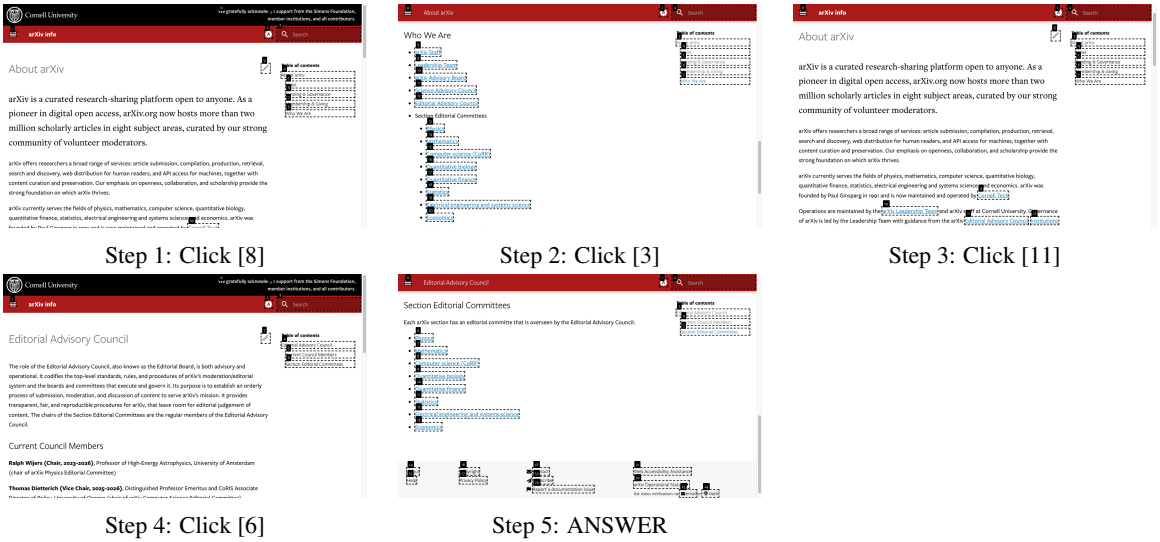


Figure 2: An organizational information retrieval case for arXiv. Given the task: “On arXiv’s About page, find the categories of the Section Editorial Committees.” The agent successfully retrieves the answer: “Physics, Mathematics, Computer science (CoRR), Quantitative biology, Quantitative finance, Statistics, Electrical engineering and systems science, Economics,” correctly identifying all eight top-level research domains designed by the platform’s editorial structure.

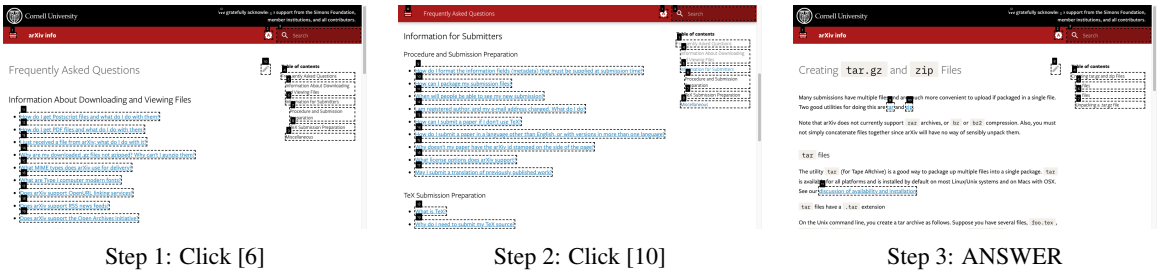


Figure 3: A user account management task on arXiv. Given the task: “How can I package my submission files?” The agent correctly returns the instruction: “Create tar.gz and zip Files,” accurately capturing the recommended submission packaging methods outlined in the official arXiv help documentation for authors preparing their papers.

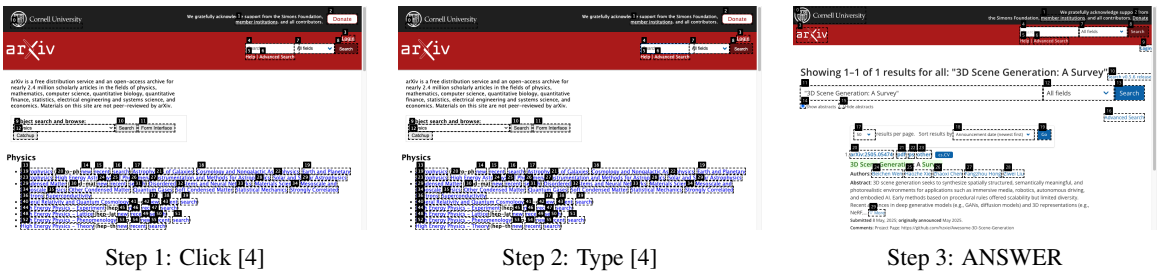
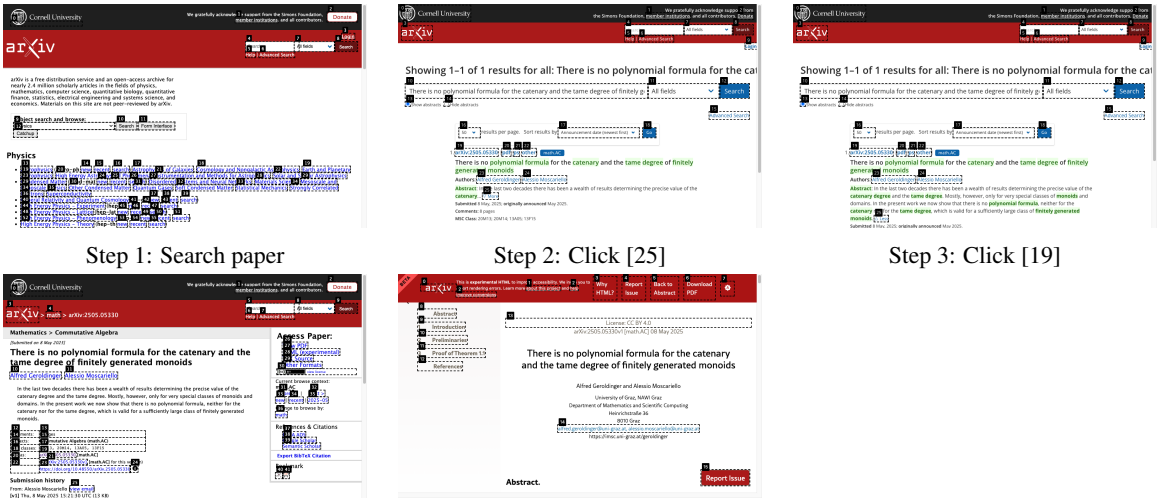


Figure 4: A paper discovery task on arXiv. Given the task: “State the name of the second Author of this paper: 3D Scene Generation: A Survey.” The agent successfully identifies the second listed author as “Haozhe Xie,” confirming the correct retrieval of metadata related to the specified research paper.

Step 1: Scroll [WINDOW] Step 2: Scroll [WINDOW] Step 3: ANSWER

Figure 5: A search interaction task on arXiv's advanced search page. Given the task: "Tell me how to search within a subcategory." The agent correctly interprets the search interface and returns the instruction: "Select All fields," demonstrating its ability to navigate and extract advanced search instructions from the user interface.



Step 1: Search paper Step 2: Click [25] Step 3: Click [19]

Step 4: Click [27] Step 5: ANSWER

Figure 6: A publication detail retrieval task on arXiv. Given the task: "Provide the name of the university publishing in this paper: There is no polynomial formula for the catenary and the tame degree of finitely generated monoids." The agent correctly extracts the affiliation information and returns: "University of Graz," confirming successful deep content extraction from the publication metadata.

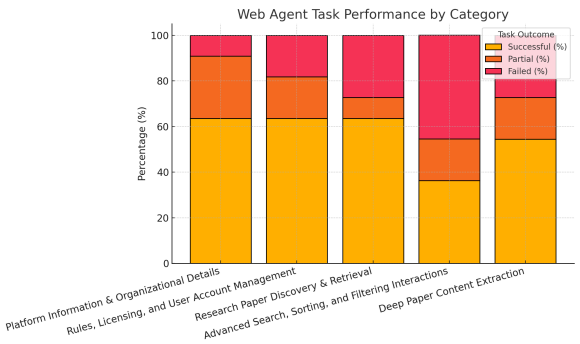


Figure 7: Stacked bar chart of GPT-o1, showing task completion rates across five arXiv-specific categories, where higher values indicate better performance.

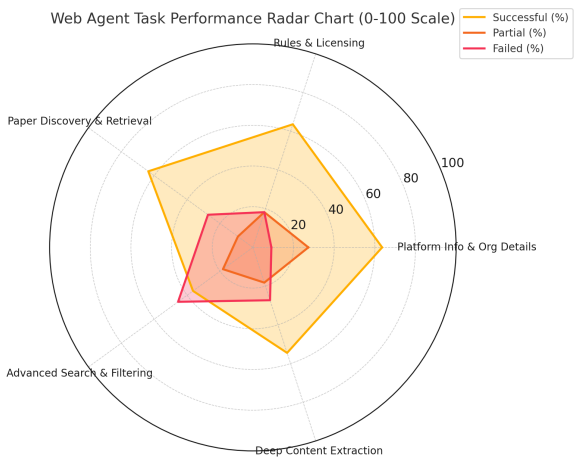


Figure 8: Radar chart of GPT-o1, visualizing success, partial, and failure rates across five arXiv-specific categories. Larger coverage indicates stronger task completion.

For each interaction step t:

1. Retrieve the last 3 visual observations and their associated element texts:
   `last_3_steps = get_last_3_steps()`
2. Ask the model which of these steps is most useful for reasoning:
   `reflection_prompt = format_reflection_prompt(last_3_steps)`
   `important_step_index = model.respond(reflection_prompt)`
3. Construct the reasoning context:
   Reasoning source: `last_3_steps[important_step_index]`
   Current view: `last_3_steps[-1]`
4. Ask the model to generate the next action using both reasoning and current view:
   `action_prompt = format_action_prompt(task, guidance_context, history)`
   `next_action = model.respond(action_prompt)`
5. Execute the action and update history:
   `result = execute(next_action)`
   `update_history(next_action, result)`

Figure 9: Pseudocode for dynamic reflective agent across the last 3 steps in WebArXiv.

Step	Screenshot	Thought (with dynamic reflection)	Action	Action target (button/text)
1	screenshot1.png	To find the university, I should search the paper by its title. (Initial step — no prior context to reflect on.)	Type [4]	Search bar [4]; input: "There is no polynomial formula for the catenary and the tame degree of finitely generated monoids"
2	screenshot2.png	Search results are shown. From Step 1, I confirmed this is the correct paper. Let's open more details.	Click [25]	More [25]; expands metadata and abstract
3	screenshot3.png	(Dynamic reflection: Step 2 showed author names, but no affiliations.) To get structured metadata, we should access the article detail page.	Click [19]	arXiv:2505.05330 [19]; opens paper's detail page
4	screenshot4.png	(Dynamic reflection: Step 3 showed abstract, still no affiliations.) The HTML version might include full author info and contact details.	Click [27]	HTML (experimental) [27]; opens the full paper in HTML format
5	screenshot5.png	(Dynamic reflection: Step 4 gave full text. University found in both email domain and author line.)	ANSWER; University of Graz	From author info: alfred.geroldinger@uni-graz.at; full text shows: University of Graz

Table 4: Interaction trajectory for Task ArXiv–51 with dynamic reflection. The agent uses targeted recall to improve navigation and stability across dense UI structures.