EconWebArena: Benchmarking Autonomous Agents on Economic Tasks in Realistic Web Environments

Zefang Liu*
Capital One
San Jose, CA, USA
zefang.liu@capitalone.com

Yinzhu Quan*
Georgia Institute of Technology
Atlanta, GA, USA
yquan9@gatech.edu

Abstract

We introduce EconWebArena, a benchmark for evaluating autonomous agents on complex, multimodal economic tasks in realistic web environments. The benchmark comprises 360 curated tasks from 82 authoritative websites spanning domains such as macroeconomics, labor, finance, trade, and public policy. Each task challenges agents to navigate live websites, interpret structured and visual content, interact with real interfaces, and extract precise, time-sensitive data through multi-step workflows. We construct the benchmark by prompting multiple large language models (LLMs) to generate candidate tasks, followed by rigorous human curation to ensure clarity, feasibility, and source reliability. Unlike prior work, EconWebArena emphasizes fidelity to authoritative data sources and the need for grounded web-based economic reasoning. We evaluate a diverse set of state-of-the-art multimodal LLMs as web agents, analyze failure cases, and conduct ablation studies to assess the impact of visual grounding, plan-based reasoning, and interaction design. Our results reveal substantial performance gaps and highlight persistent challenges in grounding, navigation, and multimodal understanding, positioning EconWebArena as a rigorous testbed for economic web intelligence.

1 Introduction

Accurate and timely access to economic data [Einav and Levin, 2014a,b] is essential for research, policy analysis, and financial decision-making. Such data are typically published by government agencies, central banks, international organizations, and financial institutions through structured web portals. Retrieving this information often requires navigating dynamic websites, interpreting charts and tables, and interacting with elements like filters, dropdowns, and forms [Edelman, 2012, Ferrara et al., 2014]. Although some platforms offer application programming interfaces (APIs) for direct data access, such interfaces are not consistently available. Many official sources do not support APIs, and those that do often differ significantly in format, coverage, and usability across countries and institutions. In practice, user-facing websites are more common and standardized, making them a more accessible and reliable target for autonomous agents. Nevertheless, many existing approaches rely on general-purpose search engines or pre-collected datasets, which often point to secondary or less reliable sources. This indirect access can introduce errors in precision, units, or interpretation, reducing the reliability of downstream analysis.

Direct interaction with authoritative online sources is critical for high-fidelity economic data acquisition, but it poses unique challenges for autonomous agents. These agents must operate in live web environments, reason over both structured and visual content, and execute multi-step procedures to obtain specific, verifiable information. In practice, such tasks mirror typical workflows in applied

*These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Bridging Language, Agent, and World Models for Reasoning and Planning (LAW).

economics, for example retrieving official Consumer Price Index (CPI) releases for inflation analysis, collecting central bank interest rate data for policy evaluation, or accessing trade and labor statistics for empirical research. However, existing benchmarks rarely reflect these demands. Most web agent benchmarks [Zhou et al., 2023, Drouin et al., 2024, Yoran et al., 2024] focus on general-purpose tasks such as shopping, email handling, or navigating productivity tools. These emphasize routine interactions but overlook the structured reasoning, domain expertise, and precision needed in economic contexts.

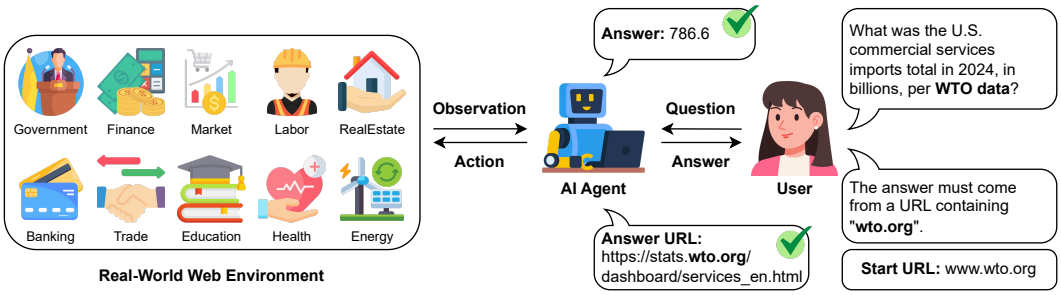
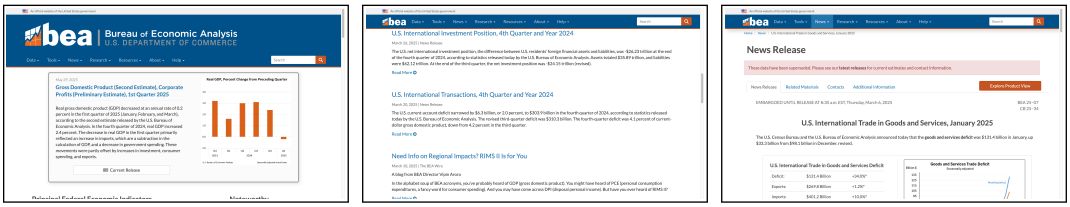


Figure 1: Overview of EconWebArena where agents solve realistic economic tasks by navigating real websites, interpreting content, and extracting grounded numeric answers.



(a) Open an economic website (b) Navigate to a relevant report (c) Extract the required value

Figure 2: Illustrative example of a typical EconWebArena task. Solving a task involves multi-step navigation across economic data websites, accessing the appropriate report or database, and extracting a target numeric value from tables, diagrams, or databases.

To address this gap, we introduce EconWebArena[2], a benchmark for evaluating autonomous agents on realistic economic tasks embedded in real-world websites. EconWebArena comprises 360 tasks based on 82 high-quality data sources across domains such as macroeconomics, labor, trade, and public policy. Each task is designed to test an agent's ability to navigate, interpret, and extract accurate economic data through live browser interactions. As illustrated in Figure 1 and Figure 2, EconWebArena covers a wide range of economic domains and requires agents to navigate authoritative websites, interpret diverse content types, and follow multi-step workflows grounded in real data. By combining domain-specific reasoning with realistic web workflows, EconWebArena offers a rigorous evaluation setting for next-generation multimodal large language model (MLLM) [Zhang et al., 2024a] agents in applied economic contexts. Our contributions are threefold: 1) we propose a novel benchmark that captures the real-world demands of web-based economic reasoning and data retrieval, 2) we evaluate a diverse set of state-of-the-art MLLM agents and analyze their failure modes, and 3) we conduct ablation studies to assess the effect of visual grounding, reasoning strategies, and interaction design on task performance.

2 Related Work

Recent benchmarks for autonomous agents have primarily focused on general-purpose tasks in simulated or real web environments, with growing attention to multimodal grounding, interface interaction, and domain-specific reasoning.

[2] https://econwebarena.github.io/

2.1 Web-Based Benchmarks

Web-based agent benchmarks have primarily focused on general-purpose tasks such as navigation, data entry, and e-commerce, with foundational platforms like WebArena [Zhou et al., 2023] and BrowserGym [Chezelles et al., 2024] enabling large-scale or modular evaluation. Specialized benchmarks include WebShop [Yao et al., 2022] for online shopping and WorkArena [Drouin et al., 2024, Boisvert et al., 2024] for productivity tasks. More recent efforts such as AssistantBench [Yoran et al., 2024], Mind2Web [Deng et al., 2023], WebCanvas [Pan et al.], RealWebAssist [Ye et al., 2025], and WebLINX [Lù et al., 2024] aim to test broader agent capabilities in cross-task generalization, long-horizon reasoning, and interactive assistance. To support multimodal grounding, benchmarks like VisualWebArena [Koh et al., 2024], VideoWebArena [Jang et al., 2024], and MMINA [Zhang et al., 2024c] incorporate image and video inputs, while VisualAgentBench [Liu et al., 2024c] and VisualWebBench [Liu et al., 2024b] assess agents' ability to align language with UI elements. However, none of these settings focus on the specialized challenges of economic data retrieval and domain-specific reasoning. While they could in principle be adapted to economics, such modifications would not ensure grounding in authoritative sources or the numeric precision required in financial contexts, which EconWebArena is designed to support. Unlike prior work that emphasizes scale or multimodal grounding, our benchmark targets domain-specific workflows that require careful handling of reporting periods and reliable numeric accuracy.

2.2 Economic Benchmarks

Benchmarks in economics and finance largely emphasize static inputs such as question answering, document analysis, or language modeling. EconQA [Van Patten, 2023], FinanceBench [Islam et al., 2023], and EconLogicQA [Quan and Liu, 2024a] evaluate LLMs on domain-specific factual and reasoning questions, while FLUE/FLANG [Shah et al., 2022] target financial language modeling. Other efforts explore tool-augmented agents in finance [Zhang et al., 2024b], CRM tasks [Huang et al., 2024], or multi-agent simulations such as InvAgent [Quan and Liu, 2024b] and EconArena [Guo et al.]. Several benchmarks also extend to non-English financial domains, including BBT-Fin [Lu et al., 2023] and FinEval [Zhang et al., 2023]. However, these settings mostly rely on static content and do not capture the web-based workflows necessary for real-world economic data retrieval.

3 EconWebArena

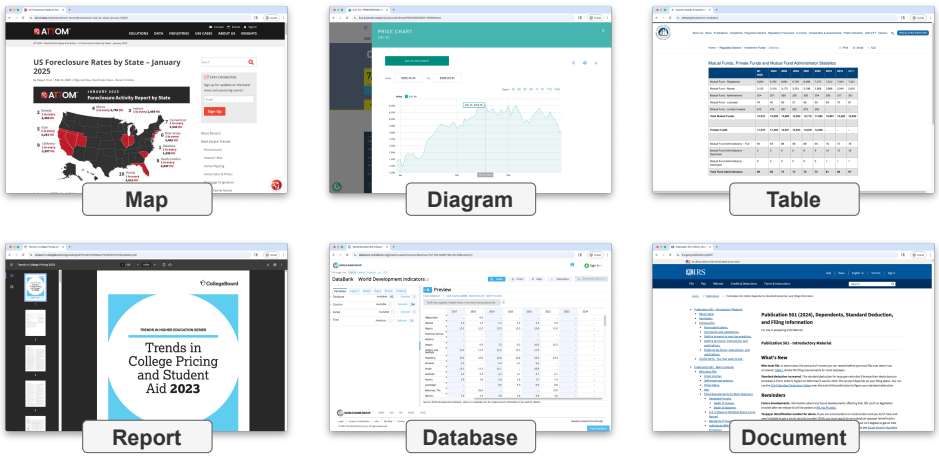


Figure 3: Representative EconWebArena tasks involve interacting with diverse web elements, including maps, diagrams, tables, reports, databases, documents, and so on.

In this section, we describe the construction of EconWebArena, a benchmark for evaluating agents on realistic economic tasks in web environments. We outline the process of task generation, manual curation and annotation, and the criteria used for answer evaluation.

3.1 Task Generation

To construct the benchmark, we prompt four state-of-the-art large language models (LLMs): GPT-4o [OpenAI, 2024a], Claude-3.7-Sonnet [Anthropic, 2025b], DeepSeek-V3 [Liu et al., 2024a], and Gemini-2.0-Flash [Google DeepMind, 2024]. Each model generates 50 candidate tasks, resulting in 200 initial questions. The prompt is carefully designed to elicit realistic, high-level tasks that autonomous agents can perform on real-world economic websites. It enforces strict guidelines for clarity, specificity, and feasibility, requiring each task to be a concise question with a single verifiable answer, using only fixed date ranges and objective language. Tasks must involve meaningful economic content, such as finance, labor, trade, or macroeconomic indicators, and be executable via realistic web interactions like navigating, filtering, or completing forms. To ensure diversity and coverage, each task must target a different named website. The full prompt is provided in Figure 6 in Appendix A.1. As illustrated in Figure 3, the resulting tasks span a wide range of interactive web elements, including maps, tables, diagrams, documents, and databases.

3.2 Task Curation and Annotation

Following task generation, we manually review, filter, and revise all 200 candidate tasks to ensure clarity, reliability, and feasibility. We retain only those that point to verifiable economic data from reputable public sources, discarding any that are vague, redundant, rely on subjective interpretation, or require access to subscription-based or non-English websites. For each accepted task, we identify the exact webpage containing the correct answer and record the corresponding numeric value. This process results in 120 high-quality seed tasks, each clearly stated, yielding a single numeric answer, and grounded in authoritative data.

Each task is assigned to one of ten high-level categories based on the domain of its source website: government, finance, markets, labor, banking, energy, trade, real estate, education, or health. These categories capture a wide spectrum of economic subject areas, ranging from macroeconomic indicators and central banking to employment, housing, and healthcare. The benchmark draws from 82 authoritative websites with the complete list detailed in Appendix A.3 Table 6. Definitions for each category are provided in Appendix A.2.

To expand the dataset, we generate two variants per seed task by modifying elements such as time range, country, or indicator. Tasks involving daily or monthly data are limited to early 2025, while quarterly and annual queries span 2022 to 2025 to ensure temporal relevance. We enforce strict output constraints so that each answer is a single numeric value, enabling consistent automated evaluation. Each finalized task includes a question with a start URL, an expected answer format, and a required domain in the answer URL. For multilingual websites, we use the English-language version as the start URL. This process yields 360 tasks in total, with category distribution shown in Figure 4. Task examples are shown in Table 1, and the full dataset is publicly available[3].

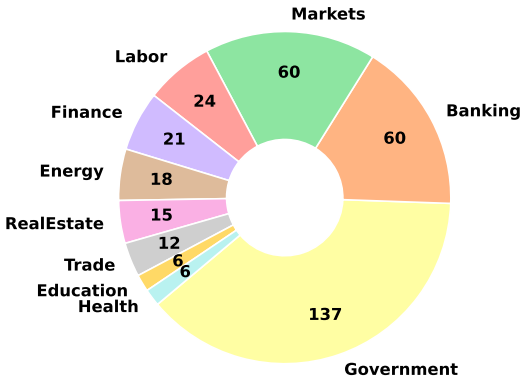


Figure 4: Distribution of task categories in EconWebArena, reflecting the diversity of economic domains represented in the benchmark.

[3] https://huggingface.co/datasets/EconWebArena/EconWebArena

Table 1: Representative EconWebArena tasks illustrating diverse economic domains, answer formats, and real-world data sources.

Category	Task Description	Start URL	Answer	Domain
Government	As published by the Office for National Statistics, what was the CPIH annual inflation rate for all items (2015=100) in the United Kingdom in March 2025? Provide only the number as a decimal with one digit after the decimal point, without percent symbols or other units.	`https://www.ons.gov.uk/`	3.4	ons.gov.uk
Energy	As reported by the U.S. Energy Information Administration, what was the average retail price of regular gasoline in California during the week of March 24, 2025, in dollars per gallon? Provide only the number as a decimal with three digits after the decimal point, without currency symbols, commas, or other units.	`https://www.eia.gov/`	4.418	eia.gov
Markets	As reported by Cox Automotive, what was the total number of unsold used vehicles in the United States as of March 31, 2025? Provide only the number as a decimal with two digits, in millions, without commas or other units.	`https://www.coxautoinc.com/`	2.14	coxautoinc.com
Banking	As reported by the Federal Reserve Bank of New York, what was the effective federal funds rate on January 10, 2025? Provide only the number as a decimal with two digits, without percent symbols or other units.	`https://www.newyorkfed.org/`	4.33	newyorkfed.org

3.3 Observation and Action Spaces

EconWebArena extends the BrowserGym [Chezelles et al., 2024] framework to support economic tasks in realistic web environments. Each task provides the agent with structured observations, including the accessibility tree (AXTree), webpage screenshots, and contextual metadata such as focused elements, error logs, and action history. These inputs enable both language-based and multimodal reasoning. The action space includes fine-grained browser control primitives at both the element and coordinate levels, supporting operations like mouse movement, text input, tab switching, and form submission. This setup allows agents to perform the low-level interactions required for complex economic data retrieval. Further details are provided in Section 4.1.

3.4 Answer Evaluation

We evaluate each agent response using two criteria. First, the predicted answer must contain the correct numeric value, typically an integer or decimal that represents an economic measure such as a price, rate, or index. The value must match the gold annotation exactly, even if it appears within a longer sentence. Second, the response must include a valid URL that contains the required domain name, confirming the information was retrieved from the intended authoritative source. A task is considered correct only if both conditions are met. This evaluation supports automated scoring while ensuring factual accuracy and source reliability. We report success rate as the percentage of correctly answered tasks within each category and also include the average number of steps taken on successful tasks.

4 Experiments

In this section, we evaluate a range of agents on EconWebArena, examining their overall performance, common failure patterns, and the effects of different design choices through ablation studies.

4.1 Experimental Setup

We implement the EconWebArena tasks using BrowserGym[4] [Chezelles et al., 2024], a lightweight browser simulation environment that enables fine-grained agent control in web-based settings. The action space includes a broad range of primitives for common browser operations, such as clicking, typing, filling input fields, hovering, and scrolling. Agents can also manage tabs and navigate across pages. Interactions are supported at both the element level (via structured identifiers) and the coordinate level (via pixel positions), allowing flexible strategies for handling diverse web layouts. Additional actions include sending messages to the user and performing no-op delays. A full list of supported actions is shown in Table 2.

Table 2: Action space in EconWebArena, organized by primitive category based on BrowserGym.

Category	Primitive	Description
`bid`	`fill`	Input text
	`click`	Click element
	`hover`	Hover on element
	`press`	Press key combination
	`focus`	Focus element
	`clear`	Clear input
	`select_option`	Select dropdown option
`coord`	`mouse_move`	Move mouse
	`mouse_click`	Click by position
	`mouse_drag_and_drop`	Drag and drop
	`keyboard_press`	Press key(s)
	`keyboard_type`	Type with keyboard
`tab`	`new_tab`	Open new tab
	`tab_close`	Close current tab
	`tab_focus`	Focus on tab
`nav`	`go_back`	Navigate back
	`go_forward`	Navigate forward
	`goto`	Go to URL
`misc`	`send_msg_to_user`	Send message
	`scroll`	Scroll page
	`noop`	Wait without action

The web agents are built using AgentLab[5] [Chezelles et al., 2024], a modular framework for prompting and executing LLM-based agents. Each agent receives structured observations, including the AXTree, a webpage screenshot, metadata on the focused element, recent errors, and the full history of prior actions and thoughts. We enable key prompt features such as chain-of-thought reasoning, contextual hints, and real-world examples, while disabling multi-action execution and memory retrieval. Additional settings control element extraction, coordinate formatting, and response style. These configurations are tuned to balance informativeness and efficiency within token limits. The full setup is listed in Appendix B.1 Table 7.

For evaluation, we select a range of multimodal large language models (MLLMs), covering both proprietary and open-weight models: GPT-4o [Achiam et al., 2023, OpenAI, 2024a], GPT-4.1 [OpenAI, 2025a], o4-mini [OpenAI, 2025b], Claude Sonnet 4 [Anthropic, 2025a], Gemini 2.5 Flash [Gemini Team, 2023, Google DeepMind, 2025], and Llama 4 Maverick [Grattafiori et al., 2024, Meta AI, 2025]. Each model is allowed up to 30 steps per task to prevent looping or excessive action chains. Due to API credit constraints, we run a single trial per model. Since each seed task has two

[4] `https://github.com/ServiceNow/BrowserGym`

[5] `https://github.com/ServiceNow/AgentLab`

variants, the reported results reflect an average over three instances per seed task. In addition, we conduct human evaluation by dividing the benchmark among the authors (graduate-level researchers with general knowledge of economics), ensuring that each annotator reviews only tasks they had not previously seen. All experiments were conducted during the final week of May 2025 to ensure consistency across models and reduce variance due to changes in live website content.

4.2 Experimental Results

Table 3: Average task success rates (SR) and average steps (on successful tasks) on EconWebArena by major category for models (o4-mini, GPT-4.1, GPT-4o, Claude Sonnet 4, Gemini 2.5 Flash, Llama 4 Maverick) and human. (Other: Energy, RealEstate, Trade, Education, and Health.)

Category	Tasks	o4-mini	GPT-4.1	GPT-4o	Claude-4	Gemini-2.5	Llama-4	Human
Banking	60	41.7%	23.3%	18.3%	38.3%	28.3%	21.7%	95.0%
Finance	21	33.3%	14.3%	14.3%	23.8%	33.3%	9.5%	95.2%
Government	138	57.2%	45.7%	35.5%	47.1%	39.1%	26.1%	91.3%
Labor	24	20.8%	0.0%	8.3%	12.5%	4.2%	4.2%	91.7%
Markets	60	48.3%	35.0%	33.3%	41.7%	33.3%	15.0%	96.7%
Other	57	42.1%	24.6%	21.1%	31.6%	22.8%	12.3%	93.0%
All SR (↑)	360	**46.9%**	31.9%	26.9%	38.6%	31.1%	18.9%	93.3%
Steps (↓)	-	8.99	**7.23**	7.77	11.77	9.29	9.54	-

To better understand model capabilities on EconWebArena, we organize our analysis around three core research questions (RQs) concerning performance, failure patterns, and potential improvements. Table 3 reports overall accuracy across major economic categories, while Table 8 in Appendix B.2 provides more detailed, category-specific breakdowns.

4.2.1 RQ1: How well can LLM agents solve economic tasks in realistic web environments?

As shown in Table 3, current LLM agents achieve only partial success on EconWebArena. The best-performing model, o4-mini, reaches an average success rate of 46.9%, while other proprietary models like GPT-4.1 (31.9%), Claude Sonnet 4 (38.6%), and Gemini 2.5 Flash (31.1%) show mixed results across categories. Open-weight model Llama 4 Maverick performs considerably lower, with an overall success rate of 18.9%. Government and market domains yield higher scores across most models, whereas labor, finance, and other specialized categories remain particularly challenging. Human performance remains consistently high at 93.3%, underscoring the gap between current agents and expert-level competence. In terms of efficiency, GPT-4.1 completes successful tasks with the fewest steps on average, while Claude Sonnet 4 tends to follow longer action sequences. These results highlight the limitations of existing agents in handling complex, real-world economic workflows. Representative success cases are provided in Appendix B.3.

4.2.2 RQ2: What causes LLM agents to fail on EconWebArena tasks?

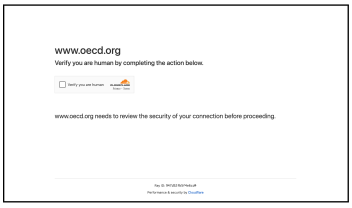
Table 4: Distribution of failure cases by error type in o4-mini's performance on EconWebArena seed tasks.

Error Type	Count	Percentage
Access Issues	16	25.0%
Data Extraction Errors	16	25.0%
Interaction Failures	8	12.5%
Navigation Failures	15	23.4%
Visual Understanding Failures	9	14.1%
All	64	100.0%

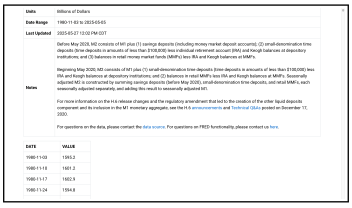
To understand why LLM agents fail on EconWebArena tasks, we perform a detailed error analysis on the o4-mini model. Out of 360 benchmark tasks, o4-mini fails 193, among which 64 are seed tasks. We manually review these 64 failures to identify common error patterns and categorize them into five

distinct types: access issues, data extraction errors, navigation failures, visual understanding failures, and interaction failures. Table 4 summarizes the distribution of these error types, and Figure 5 illustrates representative examples for each category (see Appendix B.4 for full task visualizations).

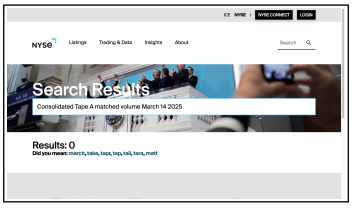
Access issues (Figure 5a) include problems such as blocked or restricted websites, loading errors, or access denials that prevent task execution entirely. Data extraction errors (Figure 5b) occur when the agent navigates to the correct page but extracts the wrong value, often due to confusion among dense statistics or approximate retrieval. Navigation failures (Figure 5c) reflect difficulty locating the correct webpage or section, including getting stuck or choosing incorrect paths. Visual understanding failures (Figure 5d) are due to an inability to interpret charts, diagrams, or other non-textual representations. Interaction failures (Figure 5e) stem from execution errors in UI manipulation, such as failing to trigger dropdowns or input fields. These results suggest that agent failure often stems from combined weaknesses in reasoning, perception, and interface fluency when dealing with complex economic content online.



(a) Access Issue

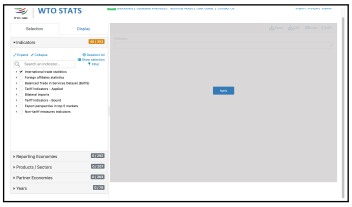


(b) Data Extraction Error



(c) Navigation Failure

(d) Visual Understanding Failure



(e) Interaction Failure

Figure 5: Representative examples of common error types from o4-mini in EconWebArena: access issues, data extraction errors, navigation failures, visual understanding failures, and interaction failures.

4.2.3 RQ3: What directions can improve LLM agent performance on these tasks?

Our error analysis suggests several directions to improve LLM agents on EconWebArena tasks. First, robust visual grounding and multimodal reasoning are essential for interpreting economic charts, tables, and diagrams, which frequently appear in official data portals. Structured input representations, such as combining AXTree with screenshots and coordinates, can better support perception and interaction. Planning and self-criticism prompting also improve performance in multi-step navigation tasks. In addition, hybrid methods that integrate GUI-based exploration with structured API calls, when available, can enhance both reliability and efficiency. Finally, domain-adaptive tuning and retrieval-augmented prompting using realistic economic examples may further strengthen agent accuracy in complex cases.

4.3 Ablation Studies

To assess the contribution of each prompt configuration, we conduct ablation experiments on the o4-mini agent, varying components related to observation, action, and reasoning. The initial setup is described in Appendix B.1 Table 7, and results are reported in Table 5. Each variant alters a single feature based on AgentLab conventions [Chezelles et al., 2024].

Removing the structured accessibility tree and using only raw HTML (`-use_ax_tree +use_html`) leads to a large performance drop (from 46.9% to 36.7%), confirming the critical role of clean DOM representations. Excluding screenshots (`-use_screenshot`) causes a modest decline (44.7%), but

Table 5: Ablation results for the o4-mini model, showing the impact of prompt configuration options on EconWebArena success rate (SR).

Category	Configuration	SR (↑)	Steps (↓)
-	Initial configuration	46.9%	8.99
observation	- use_ax_tree + use_html	36.7%	9.61
	- use_screenshot	44.7%	8.88
	+ use_som	47.2%	8.71
	- use_history	45.8%	7.76
	- extract_coords	43.6%	8.61
action	+ multiaction	41.9%	7.37
reasoning	- use_thinking	46.9%	8.93
	+ use_plan	49.4%	9.86
	+ use_criticise	47.5%	8.61

visual information remains essential for tasks requiring interpretation of charts and diagrams. Removing history context (-use_history) slightly reduces success rate (45.8%) but leads to noticeably fewer steps (7.76), suggesting that brief context improves efficiency but is not always necessary for correctness. Turning off coordinate extraction (-extract_coords) also hurts performance (43.6%), as many UI components require position-based interaction that AXTree alone may not cover. Enabling set-of-mark prompting (+use_som) [Yang et al., 2023] slightly improves accuracy (47.2%), indicating its benefit for spatial referencing.

In the action section, enabling multiple action execution (+multiaction) leads to a lower average step count (7.37) but reduces accuracy to 41.9%, indicating a trade-off between efficiency and reliability. For reasoning, removing explicit thinking traces (-use_thinking) has no measurable effect on success rate (46.9%), suggesting that o4-mini is already capable of internal step-by-step reasoning. Adding high-level planning (+use_plan) yields the largest improvement in success rate (49.4%) but also increases the average number of steps, reflecting its benefit for long-horizon tasks. Enabling self-critique (+use_criticise) slightly improves success rate (47.5%) and reduces steps (8.61), suggesting it helps refine decision-making. Overall, these results emphasize the importance of structured observations, visual grounding, coordinate awareness, and global planning in complex economic web tasks.

5 Conclusion

We present EconWebArena, a benchmark for evaluating autonomous agents on complex economic tasks grounded in real-world web environments. The benchmark includes 360 manually curated tasks across 82 authoritative websites and ten economic domains, requiring agents to navigate live webpages, interpret structured and visual content, and extract precise, time-sensitive data through multi-step interactions. Our evaluation of state-of-the-art multimodal large language models reveals significant limitations in grounding, navigation, and multimodal reasoning. Error analysis highlights persistent failure modes, while ablation studies identify key configuration choices that impact performance. EconWebArena provides a challenging and realistic testbed for advancing domain-aware, interaction-capable agents for economic data retrieval and reasoning.

Limitations

EconWebArena uses a strict evaluation metric based on exact numeric matching and source URL verification, which ensures high precision but may not account for partially correct answers or valid intermediate reasoning steps. While the benchmark operates on live web content, it is limited to publicly accessible pages and excludes websites requiring login, subscription, or region-specific access. All tasks are presented in English, which restricts the benchmark's applicability in multilingual or localized settings. Moreover, EconWebArena focuses on single-turn, goal-oriented tasks and does not capture longer-term planning or interactive dialogue scenarios.

Ethical Considerations

This work follows principles of transparency, reproducibility, and responsible AI development. All tasks are sourced from publicly available, authoritative economic websites, with no use of personal or sensitive data. A thorough human review process filters out ambiguous, unverifiable, or inappropriate content to ensure quality and integrity. Since the benchmark does not involve human subjects or private information, it presents minimal ethical risk. We openly release the dataset and tools to support further research, with an emphasis on factual accuracy and source accountability.

Acknowledgments

This research was supported in part by API credits provided through the OpenAI Researcher Access Program.

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.

Anthropic. Introducing claude 4, May 2025a. URL `https://www.anthropic.com/news/claude-4`.

Anthropic. Claude 3.7 sonnet and claude code, February 2025b. URL `https://www.anthropic.com/news/claude-3-7-sonnet`.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault de Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks. *Advances in Neural Information Processing Systems*, 37:5996–6051, 2024.

De Chezelles, Thibault Le Sellier, Maxime Gasse, Alexandre Lacoste, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, et al. The browsergym ecosystem for web agent research. *arXiv preprint arXiv:2412.05467*, 2024.

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. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, and Alexandre Lacoste. Workarena: How capable are web agents at solving common knowledge work tasks? In *International Conference on Machine Learning*, pages 11642–11662. PMLR, 2024.

Benjamin Edelman. Using internet data for economic research. *Journal of Economic Perspectives*, 26(2):189–206, 2012.

Liran Einav and Jonathan Levin. The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1):1–24, 2014a.

Liran Einav and Jonathan Levin. Economics in the age of big data. *Science*, 346(6210):1243089, 2014b.

Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: A survey. *Knowledge-based systems*, 70:301–323, 2014.

Gemini Team. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, December 2024. URL `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/`.

Google DeepMind. Gemini 2.5 flash, 2025. URL https://deepmind.google/models/gemini/flash/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shangmin Guo, Haochuan Wang, Haoran Bu, Yi Ren, Dianbo Sui, Yu-Ming Shang, and Siting Estee Lu. Economics arena for large language models. In *Language Gamification-NeurIPS 2024 Workshop*.

Kung-Hsiang Huang, Akshara Prabhakar, Sidharth Dhawan, Yixin Mao, Huan Wang, Silvio Savarese, Caiming Xiong, Philippe Laban, and Chien-Sheng Wu. Crmarena: Understanding the capacity of llm agents to perform professional crm tasks in realistic environments. *arXiv preprint arXiv:2411.02305*, 2024.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Lawrence Keunho Jang, Yinheng Li, Charles Ding, Justin Lin, Paul Pu Liang, Dan Zhao, Rogerio Bonatti, and Kazuhito Koishida. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Russ Salakhutdinov, and Daniel Fried. Visualwebarena: Evaluating multimodal agents on realistic visual web tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 881–905, 2024.

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*, 2024a.

Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visualwebbench: How far have multimodal llms evolved in web page understanding and grounding? In *First Conference on Language Modeling*, 2024b.

Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Yifan Xu, Xixuan Song, Shudan Zhang, Hanyu Lai, Xinyi Liu, Hanlin Zhao, et al. Visualagentbench: Towards large multimodal models as visual foundation agents. *arXiv preprint arXiv:2408.06327*, 2024c.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*, 2023.

Xing Han Lù, Zdeněk Kasner, and Siva Reddy. Weblinx: Real-world website navigation with multi-turn dialogue. *arXiv preprint arXiv:2402.05930*, 2024.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

OpenAI. Hello gpt-4o, May 2024a. URL https://openai.com/index/hello-gpt-4o/.

OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence, July 2024b. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. Introducing gpt-4.1 in the api, April 2025a. URL https://openai.com/index/gpt-4-1/.

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

Yichen Pan, Dehan Kong, Sida Zhou, Cheng Cui, Yifei Leng, Bing Jiang, Hangyu Liu, Yanyi Shang, Shuyan Zhou, Tongshuang Wu, et al. Webcanvas: Benchmarking web agents in online environments. In *Agentic Markets Workshop at ICML 2024*.

Yinzhu Quan and Zefang Liu. Econlogicqa: A question-answering benchmark for evaluating large language models in economic sequential reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2273–2282, 2024a.

Yinzhu Quan and Zefang Liu. Invagent: A large language model based multi-agent system for inventory management in supply chains. *arXiv preprint arXiv:2407.11384*, 2024b.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335, 2022.

Tate Van Patten. Evaluating domain specific llm performance within economics using the novel econqa dataset. 2023.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757, 2022.

Suyu Ye, Haojun Shi, Darren Shih, Hyokun Yun, Tanya Roosta, and Tianmin Shu. Realwebassist: A benchmark for long-horizon web assistance with real-world users. *arXiv preprint arXiv:2504.10445*, 2025.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. Assistantbench: Can web agents solve realistic and time-consuming tasks? In *EMNLP*, 2024.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024a.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*, 2023.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4314–4325, 2024b.

Ziniu Zhang, Shulin Tian, Liangyu Chen, and Ziwei Liu. Mmina: Benchmarking multihop multimodal internet agents. *arXiv preprint arXiv:2404.09992*, 2024c.

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.

A Benchmark Details

This appendix provides additional details on the task generation process, prompt design, and the categorization of websites used in EconWebArena.

A.1 Task Generation Prompt

To ensure the quality, diversity, and feasibility of the benchmark tasks, we use a carefully designed prompt to guide task generation by large language models. The prompt instructs the models to produce concise, goal-driven questions grounded in real economic scenarios. It specifies formatting rules, data constraints, and task requirements to guarantee that each example can be realistically completed by an autonomous web agent. The full prompt is shown in Figure 6.

A.2 Website Categories

The benchmark tasks are grouped into ten economic categories, each reflecting a distinct type of data source. These categories span a broad range of content, from macroeconomic indicators to labor statistics and market data. Each task is assigned a single category based on its data source, as listed below.

- **Government:** National or international public-sector institutions that publish official statistics, policy data, tax rules, economic indicators, or demographic reports across sectors (e.g., World Bank, U.S. Census Bureau, OECD, Statistics Canada, IRS, IMF).
- **Finance:** Organizations or entities focused on investment products, corporate financials, fund performance, insurance costs, or regulatory financial filings (e.g., Vanguard, JPMorgan Chase, Amazon, ATTOM Data, SEC).
- **Markets:** Sources that report real-time or historical prices of financial instruments, commodities, indices, or cryptocurrencies, typically from exchanges or market aggregators (e.g., NASDAQ, Yahoo Finance, CoinMarketCap, London Stock Exchange, USAGOLD).
- **Labor:** Institutions that publish data on employment, unemployment, wages, workforce demographics, or job benefits at national or regional levels (e.g., U.S. Bureau of Labor Statistics, ILO, Texas Workforce Commission, U.S. Department of Labor).
- **Banking:** Central banks and monetary authorities that control interest rates, currency policy, money supply, and other macro-financial instruments (e.g., Federal Reserve, European Central Bank, Bank of England, Central Bank of Brazil).
- **Energy:** Organizations reporting on production, pricing, or consumption of energy sources such as oil, gas, or electricity (e.g., U.S. Energy Information Administration, OPEC).
- **Trade:** Sources focused on international commerce including import/export data, tariffs, and trade balances between nations (e.g., World Trade Organization, UN Comtrade, Observatory of Economic Complexity).
- **RealEstate:** Entities that report on housing prices, rents, real estate markets, or commercial property metrics (e.g., Zillow, CBRE, Cushman & Wakefield, National Association of Realtors).
- **Education:** Sources providing data on tuition, school fees, academic costs, or university-affiliated expenses (e.g., College Board, Harvard Business School).
- **Health:** Agencies or organizations reporting health-related economic data such as premiums, public healthcare costs, or Medicare rates (e.g., U.S. Medicare, Kaiser Family Foundation).

A.3 Website List

This appendix lists all websites included in the EconWebArena benchmark. For each site, we provide its URL and the associated category, as shown in Table 6.

Prompt for Economic Task Generation

Generate a list of 50 high-level, realistic, and unambiguous tasks designed to evaluate the performance of autonomous web-based agents in economics-related scenarios. Each task must be phrased as a concise question and presented as a plain-text bullet point.

All tasks must follow these criteria:

• Task format:

  ○ Each task must be phrased as a concise, clearly worded question.
  ○ Each question must have a single, unique, and verifiable answer.
  ○ The answer to each task must be expressible in plain text.
  ○ Present the list of tasks as plain-text bullet points using a dash ("-") for each item. Do not include numbering or additional formatting.

• Clarity and specificity:

  ○ Use only absolute dates or fixed date ranges. Do not use vague or relative time expressions such as "current", "latest", "recent", or "as of now".
  ○ Do not include any dates beyond April 2025.
  ○ Avoid vague or subjective language. Each task must be clearly defined and objectively measurable.
  ○ Do not include URLs, domain names, or any form of web address in the task descriptions.

• Relevance:

  ○ Each task must relate to meaningful economic content, such as markets, trade, labor, finance, regulation, pricing, taxation, or macroeconomic indicators.
  ○ Refer only to specific, named websites where the task should be performed. Do not list alternatives or refer to general categories of websites.

• Web-based execution:

  ○ Each task must be realistically executable on a real-world website using standard browser interactions.
  ○ Tasks must involve interactive actions such as navigating, clicking, typing, selecting, filtering, or filling forms.
  ○ Tasks should generally require multiple user actions and may involve visiting multiple webpages within or across websites.

• Diversity and complexity:

  ○ Avoid simple fact lookups (e.g., a price or rate on a date) unless they are part of a broader, practical objective.
  ○ Include a range of task types and difficulty levels, emphasizing realistic, multi-step, goal-oriented scenarios that reflect real-world economic research, analysis, or decision-making.
  ○ Ensure that each task involves a different website to increase coverage and variety.

Generate exactly 50 tasks that meet all of the above requirements.

Figure 6: Prompt used to generate a diverse and rigorous task set for evaluating autonomous web-based agents.

B Experimental Details

This appendix provides supplementary details on the experimental setup, evaluation configurations, and result breakdowns for EconWebArena.

Table 6: List of websites used in EconWebArena, along with their URLs and categories.

Website	URL	Category
ATTOM Data	https://www.attomdata.com	Finance
Amazon	https://ir.aboutamazon.com	Finance
Australian Bureau of Statistics	https://www.abs.gov.au	Government
Bank of England	https://www.bankofengland.co.uk	Banking
Brazilian Institute of Geography and Statistics	https://www.ibge.gov.br	Government
California Franchise Tax Board	https://www.ftb.ca.gov	Government
Canadian Real Estate Association	https://stats.crea.ca	RealEstate
Cayman Islands Monetary Authority	https://www.cima.ky	Finance
Central Bank of Argentina	https://www.bcra.gob.ar	Banking
Central Bank of Brazil	https://www.bcb.gov.br	Banking
Central Bank of Egypt	https://www.cbe.org.eg	Banking
Central Bank of Mexico	https://www.banxico.org.mx	Banking
Central Bank of Nigeria	https://www.cbn.gov.ng	Banking
Central Bank of the Russian Federation	https://www.cbr.ru	Banking
China National Bureau of Statistics	https://www.stats.gov.cn	Government
CoinMarketCap	https://coinmarketcap.com	Markets
Coldwell Banker Richard Ellis	https://www.cbre.com	RealEstate
College Board	https://www.collegeboard.org	Education
Cox Automotive	https://www.coxautoinc.com	Markets
Cushman & Wakefield	https://www.cushmanwakefield.com	RealEstate
Euronext	https://live.euronext.com	Markets
European Central Bank	https://www.ecb.europa.eu	Banking
Eurostat	https://ec.europa.eu	Government
Federal Reserve	https://www.federalreserve.gov	Banking
Federal Reserve Bank of New York	https://www.newyorkfed.org	Banking
Federal Reserve Economic Data	https://fred.stlouisfed.org	Banking
France National Institute of Statistics and Economic Studies	https://www.insee.fr	Government
Frankfurt Stock Exchange	https://www.boerse-frankfurt.de	Markets
German Federal Statistical Office	https://www.destatis.de	Government
Harvard Business School	https://www.hbs.edu	Education
Hellenic Statistical Authority	https://www.statistics.gr	Government
India Brand Equity Foundation	https://www.ibef.org	Government
Inland Revenue Authority of Singapore	https://www.iras.gov.sg	Government
International Labour Organization	https://www.ilo.org	Labor
International Monetary Fund	https://www.imf.org	Government
JPMorgan Chase	https://www.jpmorganchase.com	Finance
Kaiser Family Foundation	https://www.kff.org	Health
London Bullion Market Association	https://www.lbma.org.uk	Markets
London Stock Exchange	https://www.londonstockexchange.com	Markets
NASDAQ	https://www.nasdaq.com	Markets
National Association of Realtors	https://www.nar.realtor	RealEstate
New York Department of Labor	https://dol.ny.gov	Labor
New York Stock Exchange	https://www.nyse.com	Markets
Observatory of Economic Complexity	https://oec.world	Trade
Organization for Economic Co-operation and Development	https://www.oecd.org	Government
Organization of the Petroleum Exporting Countries	https://www.opec.org	Energy
Pew Research Center	https://www.pewresearch.org	Government
Philippine Statistics Authority	https://psa.gov.ph	Government
Reserve Bank of New Zealand	https://www.rbnz.govt.nz	Banking
Saudi Arabia General Authority for Statistics	https://www.stats.gov.sa	Government
Saudi Central Bank	https://www.sama.gov.sa	Banking
South African Reserve Bank	https://www.resbank.co.za	Banking
Spanish Statistical Office	https://www.ine.es	Government
State Bank of Pakistan	https://www.sbp.org.pk	Banking
Statistics Canada	https://www.statcan.gc.ca	Government
Statistics Korea	https://kostat.go.kr	Government
Statistics Sweden	https://www.scb.se	Government
The Wall Street Journal	https://www.wsj.com	Markets
Turkish Statistical Institute	https://www.tuik.gov.tr	Government
U.K. Office for National Statistics	https://www.ons.gov.uk	Government
U.S. Bureau of Economic Analysis	https://www.bea.gov	Government
U.S. Bureau of Labor Statistics	https://www.bls.gov	Labor
U.S. Census Bureau	https://www.census.gov	Government
U.S. Citizenship and Immigration Services	https://www.uscis.gov	Government
U.S. Department of Housing and Urban Development	https://www.huduser.gov	Government
U.S. Department of Labor	https://www.dol.gov	Labor
U.S. Department of the Treasury	https://home.treasury.gov	Government
U.S. Energy Information Administration	https://www.eia.gov	Energy
U.S. Internal Revenue Service	https://www.irs.gov	Government
U.S. Medicare	https://www.medicare.gov	Health
U.S. Securities and Exchange Commission	https://www.sec.gov	Finance
U.S. Small Business Administration	https://www.sba.gov	Finance
U.S. Social Security Administration	https://www.ssa.gov	Government
USAGOLD	https://www.usagold.com	Markets
United Nations Comtrade Database	https://comtradeplus.un.org	Trade
Vanguard	https://investor.vanguard.com	Finance
Westmetall	https://www.westmetall.com	Markets
World Bank	https://www.worldbank.org	Government
World Inequality Database	https://wid.world	Government
World Trade Organization	https://www.wto.org	Trade
Yahoo Finance	https://finance.yahoo.com	Markets
Zillow	https://www.zillow.com	RealEstate

B.1 More Experimental Settings

In this section, we provide additional configuration details for all experiments conducted on EconWebArena. Table 7 summarizes the prompt-level settings used by agents, adapted from the AgentLab [Chezelles et al., 2024] interface. Unless otherwise specified, these settings were fixed across all model evaluations. We evaluate a comprehensive set of LLMs, including GPT-4o[6] [Achiam et al., 2023, OpenAI, 2024a], GPT-4o mini[7] [OpenAI, 2024b], GPT-4.1[8] [OpenAI, 2025a], GPT-4.1 mini[9] [OpenAI, 2025a], o4-mini[10] [OpenAI, 2025b], Claude Sonnet 4[11] [Anthropic, 2025a], Gemini 2.5 Flash[12] [Gemini Team, 2023, Google DeepMind, 2025], and Llama 4 Maverick[13] [Grattafiori et al., 2024, Meta AI, 2025].

Table 7: Prompt configuration options used in EconWebArena, based on available flags and settings from AgentLab.

Category	Flag	Setting	Description
observation	use_html	✗	Include raw HTML content in the input prompt
	use_ax_tree	✓	Incorporate AXTree structure into the prompt
	use_focused_element	✓	Indicate the currently focused element
	use_error_logs	✓	Attach the last encountered error
	use_past_error_logs	✗	Append historical error messages
	use_history	✓	Add contextual history from the past
	use_action_history	✓	Provide a timeline of previous actions taken
	use_think_history	✓	Show prior reasoning steps (chain-of-thought)
	use_diff	✗	Use image-based differences to represent changes
	use_screenshot	✓	Use visual input by including a screenshot
	use_som	✗	Replace screenshots with a Set-of-Marks format
	extract_visible_tag	✓	Tag elements that are visually present
	extract_clickable_tag	✓	Mark elements that are clickable
	extract_coords	✓	Provide location data for each element
	filter_visible_elements_only	✗	Restrict to only visible elements
action	multiaction	✗	Enable execution of multiple actions at once
	long_description	✗	Include full documentation for each action
	individual_examples	✗	Add individual usage examples per action
reasoning	use_thinking	✓	Enable chain-of-thought reasoning
	use_plan	✗	Let the agent draft and refine a plan at each step
	use_criticise	✗	Let the agent to critique its own action
	use_memory	✗	Retrieve and apply long-term memory
	use_concrete_example	✓	Utilize specific real-world examples for guidance
	use_abstract_example	✓	Use generalized, descriptive examples
	use_hints	✓	Provide additional contextual hints to the model
	enable_chat	✗	Allow multi-turn conversational interactions
	be_cautious	✓	Promote a more conservative response style
	extra_instructions	✗	Supplement with extra task-specific guidance

B.2 More Experimental Results

We provide detailed performance metrics by category and model in Table 8, complementing the summary results reported in the main paper.

[6] gpt-4o-2024-08-06: https://platform.openai.com/docs/models/gpt-4o

[7] gpt-4o-mini-2024-07-18: https://platform.openai.com/docs/models/gpt-4o-mini

[8] gpt-4.1-2025-04-14: https://platform.openai.com/docs/models/gpt-4.1

[9] gpt-4.1-mini-2025-04-14: https://platform.openai.com/docs/models/gpt-4.1-mini

[10] o4-mini-2025-04-16: https://platform.openai.com/docs/models/o4-mini

[11] claude-sonnet-4-20250514: https://openrouter.ai/anthropic/claude-sonnet-4

[12] gemini-2.5-flash-preview-05-20: https://openrouter.ai/google/gemini-2.5-flash-preview-05-20

[13] llama-4-maverick-17b-128e-instruct-fp8: https://openrouter.ai/meta-llama/llama-4-maverick

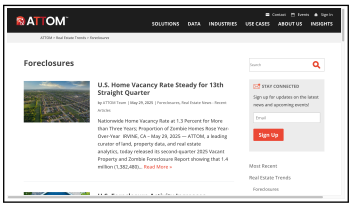
Table 8: Detailed success rates (SR) and average steps (on successful tasks) on EconWebArena by category for models (o4-mini, GPT-4.1, GPT-4.1 mini, GPT-4o, GPT-4o mini, Claude Sonnet 4, Gemini 2.5 Flash, Llama 4 Maverick) and human.

Category	Tasks	o4-mini	GPT-4.1	-mini	GPT-4o	-mini	Claude	Gemini	Llama	Human
Banking	60	41.7%	23.3%	15.0%	18.3%	15.0%	38.3%	28.3%	21.7%	95.0%
Education	6	50.0%	50.0%	50.0%	50.0%	33.3%	50.0%	50.0%	0.0%	100.0%
Energy	18	27.8%	5.6%	27.8%	5.6%	0.0%	44.4%	11.1%	11.1%	100.0%
Finance	21	33.3%	14.3%	19.0%	14.3%	19.0%	23.8%	33.3%	9.5%	95.2%
Government	138	57.2%	45.7%	34.1%	35.5%	12.3%	47.1%	39.1%	26.1%	91.3%
Health	6	100.0%	100.0%	66.7%	100.0%	16.7%	16.7%	50.0%	33.3%	100.0%
Labor	24	20.8%	0.0%	4.2%	8.3%	0.0%	12.5%	4.2%	4.2%	91.7%
Markets	60	48.3%	35.0%	35.0%	33.3%	6.7%	41.7%	33.3%	15.0%	96.7%
RealEstate	15	13.3%	0.0%	6.7%	0.0%	0.0%	0.0%	0.0%	0.0%	93.3%
Trade	12	66.7%	33.3%	33.3%	16.7%	0.0%	50.0%	41.7%	25.0%	75.0%
All SR (↑)	360	46.9%	31.9%	27.5%	26.9%	10.3%	38.6%	31.1%	18.9%	93.3%
Steps (↓)	-	8.99	7.23	8.7	7.77	8.7	11.77	9.29	9.54	-

B.3 Successful Cases

To illustrate how LLM agents can succeed on complex economic web tasks, we present four representative success cases completed by the o4-mini model. These examples span diverse categories and interaction styles, including document navigation, structured table retrieval, search-based report access, and chart-driven queries. In each case, the agent accurately interprets the task, locates the appropriate webpage, handles user interface components such as dropdowns and filters, and extracts the correct numeric value. Figures 7-10 demonstrate that while the benchmark poses many challenges, properly configured agents are capable of executing realistic workflows across a wide variety of sources.

(a) Step 0: Open ATTOM website



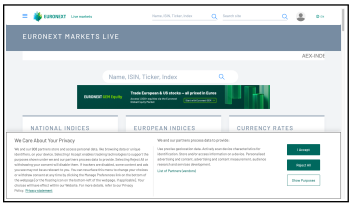
(b) Step 4: Navigate foreclosure reports

(c) Step 5: Extract the answer

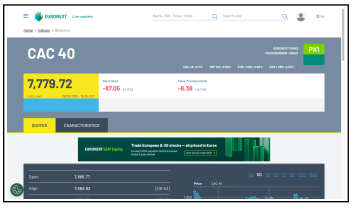
Figure 7: Successful example of Task 5 with o4-mini, which retrieves Delaware's foreclosure rate from ATTOM Data Solutions for January 2025. The agent uses search, navigates through blog categories, and correctly locates the relevant report.

B.4 Error Cases

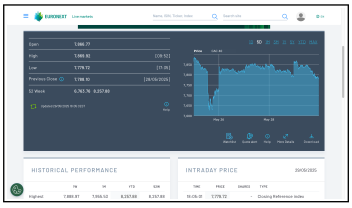
To supplement our quantitative error analysis, this appendix presents four representative examples that showcase the diverse failure modes observed in o4-mini's behavior. These include data extraction failures due to difficulty parsing large tables (Figure 11), navigation breakdowns across complex multi-page layouts (Figure 12), visual understanding issues in interpreting charts and diagrams (Figure 13), and interaction failures in dynamic UIs with layered filtering and selection logic (Figure 14). Together, these examples highlight the need for stronger grounding, multimodal reasoning, and interface robustness in economic web agents.



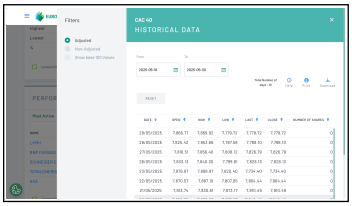
(a) Step 0: Open Euronext website



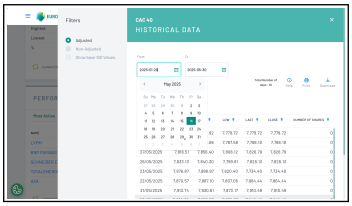
(b) Step 2: Navigate to CAC 40 page



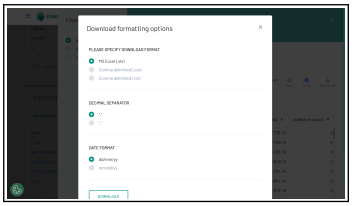
(c) Step 3: Browse the index page



(d) Step 20: Locate historical data section



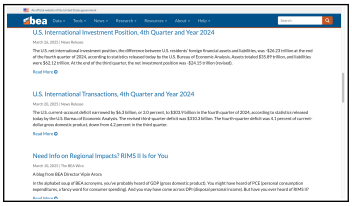
(e) Step 21: Set the target date



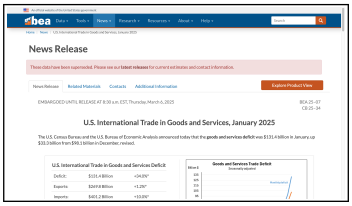
(f) Step 23: Output the closing value

Figure 8: Successful example of Task 66 with o4-mini, retrieving the CAC 40 closing value from Euronext on January 28, 2025. The agent correctly navigates to the index page, accesses historical data, inputs the date, and extracts the required number.

(a) Step 0: Navigate to recent releases

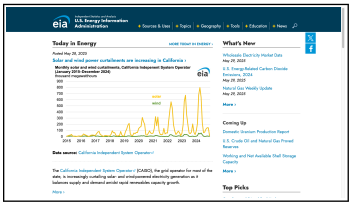


(b) Step 1: Open January trade report

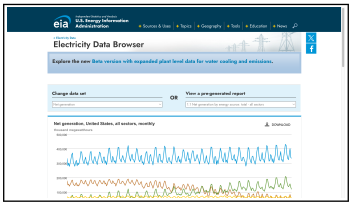


(c) Step 2: Extract the import value

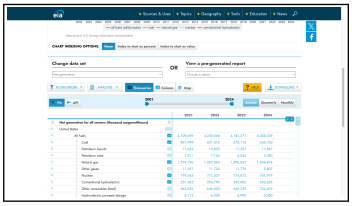
Figure 9: Successful example of Task 237 with o4-mini, retrieving U.S. imports for January 2025 from the Bureau of Economic Analysis. The agent correctly navigates to older releases, opens the January report, and extracts the reported trade imports value.



(a) Step 0: Open EIA homepage

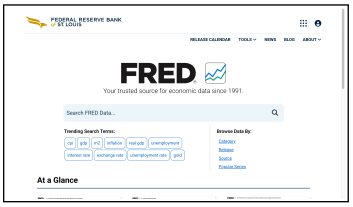


(b) Step 4: Navigate to electricity data

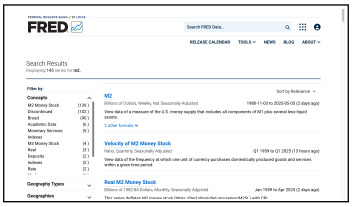


(c) Step 5: Extract the 2024 value

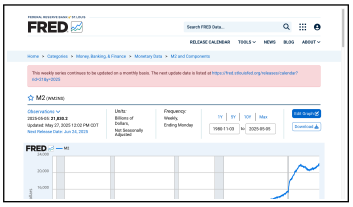
Figure 10: Successful example of Task 280 with o4-mini, which retrieves the total net electricity generation from natural gas in the United States for 2024 using the EIA Electricity Data Browser.



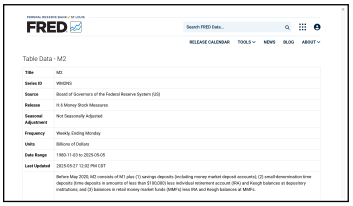
(a) Step 0: Open FRED website



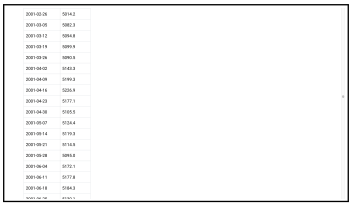
(b) Step 1: Search for M2



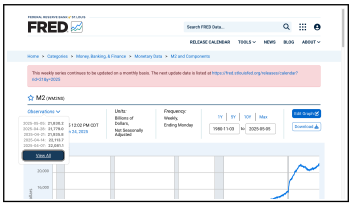
(c) Step 2: Click the M2 result



(d) Step 6: Switch to table view

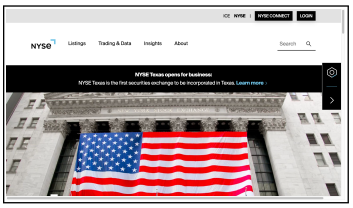


(e) Step 19: Scroll to locate the date

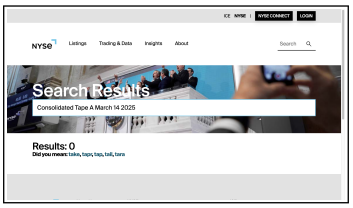


(f) Step 20: Attempt to filter data

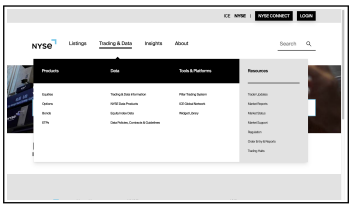
Figure 11: Example of a data extraction error in Task 88 with o4-mini on retrieving the M2 money supply for February 1, 2025 from FRED. The agent reaches the correct page but fails to extract the target value despite repeated scrolling and filtering.



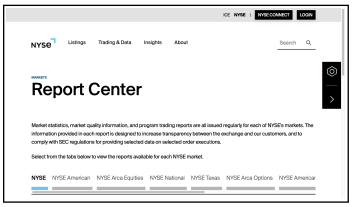
(a) Step 0: Open NYSE homepage



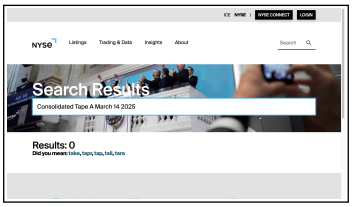
(b) Step 5: Search for "Consolidated Tap"



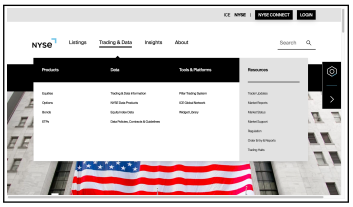
(c) Step 6: Open trading and data menu



(d) Step 7: Navigate to market reports

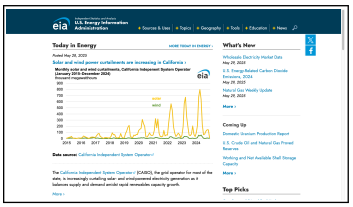


(e) Step 20: Repeat site search

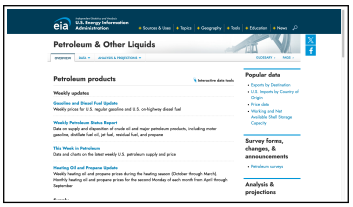


(f) Step 28: Return to the market reports

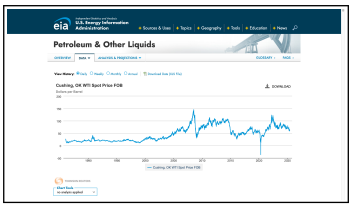
Figure 12: Example of a navigation failure in Task 157 with o4-mini on retrieving Consolidated Tape A trading volume for March 14, 2025 from the NYSE website. The agent repeatedly loops through menus and search pages but fails to reach the report containing the target value.



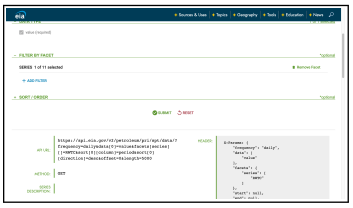
(a) Step 0: Open EIA homepage



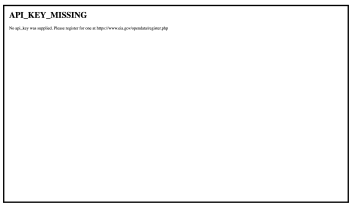
(b) Step 1: Navigate to price data



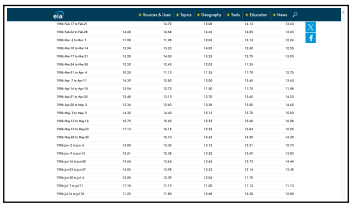
(c) Step 4: View daily price chart



(d) Step 10: Try API request

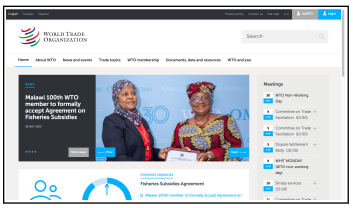


(e) Step 11: API blocked

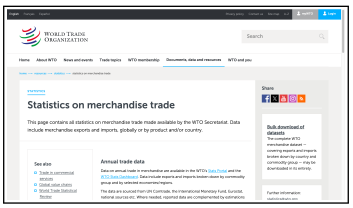


(f) Step 15: Fallback to HTML table

Figure 13: Example of a visual understanding failure in Task 283 with o4-mini for retrieving the WTI spot price from the U.S. Energy Information Administration (EIA) on March 10, 2025. The agent encounters multiple layout views, attempts API access, and struggles to correctly extract the value from the diagram or fallback table.

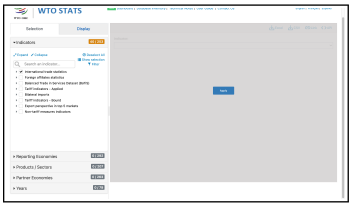


(a) Step 0: Open WTO homepage

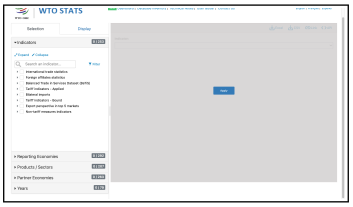


(b) Step 1: Navigate to statistics

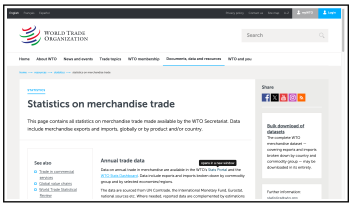
(c) Step 4: Go to WTO Stats portal



(d) Step 10: Expand indicators list



(e) Step 11: Struggles with selection



(f) Step 15: Returns to static page

Figure 14: Example of an interaction failure in Task 337 with o4-mini. The agent attempts to retrieve Egypt's 2024 merchandise export value from the WTO Stats portal but struggles to operate the interface, repeatedly failing to select the correct indicator due to the dynamic layout and nested checkbox menus.