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 present EconWebArena, a benchmark for evaluating autonomous agents on complex economic tasks in realistic web environments. It comprises 360 curated tasks from 82 authoritative websites across domains such as macroeconomics, labor, finance, trade, and policy. Each task requires agents to navigate live sites, interpret structured and visual content, and extract precise, time-sensitive data through multi-step workflows. Tasks are generated with large language models and refined via rigorous human curation to ensure clarity, feasibility, and source reliability. Unlike prior web agent benchmarks, EconWebArena emphasizes fidelity to official economic data and grounded reasoning. We evaluate state-of-the-art multimodal LLMs, analyze failure cases, and conduct ablations on visual grounding, planning, and interaction design. Results show large performance gaps and persistent challenges in navigation and multimodal understanding. EconWebArena thus offers a rigorous testbed for advancing domain-aware, interaction-capable agents in economic data retrieval and reasoning.

1 Introduction

Accurate and timely access to economic data [Einav and Levin, 2014a,b] is essential for research, policy, and financial decision-making. Such data are published by agencies, central banks, international organizations, and financial institutions through structured portals, but retrieval often requires navigating dynamic websites, interpreting charts and tables, and interacting with filters or forms [Edelman, 2012, Ferrara et al., 2014]. APIs exist but are inconsistent in availability, format, and coverage, making user-facing websites the most accessible and standardized sources. Reliance on search engines or pre-collected datasets risks errors in precision and interpretation.

Direct interaction with authoritative websites is therefore critical, yet existing benchmarks seldom reflect these demands. Most focus on general-purpose tasks such as shopping, email, or productivity tools [Zhou et al., 2023, Drouin et al., 2024, Yoran et al., 2024], emphasizing routine interactions while overlooking the reasoning, domain expertise, and precision required in economics.

We introduce EconWebArena[1], a benchmark of 360 tasks from 82 authoritative sources spanning macroeconomics, labor, trade, and public policy. Tasks require agents to navigate live sites, interpret diverse content, and extract verifiable economic data through multi-step workflows (Figure 1). By integrating domain-specific reasoning with realistic web interactions, EconWebArena provides a rigorous testbed for multimodal large language model (MLLM) agents [Zhang et al., 2024a]. Our contributions include a novel benchmark for web-based economic reasoning and retrieval, and also evaluation and failure analysis of state-of-the-art MLLMs with this benchmark.

*These authors contributed equally to this work.

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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

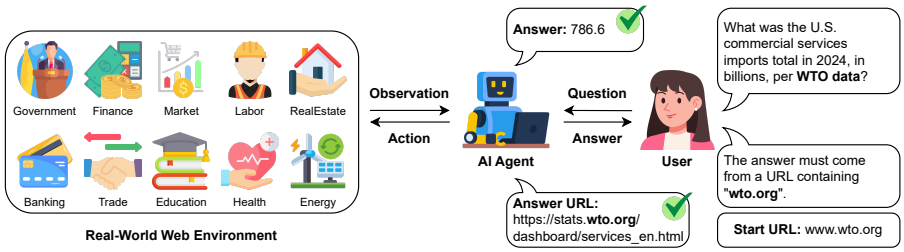


Figure 1: Overview of EconWebArena where agents solve realistic economic tasks by navigating real websites, interpreting content, and extracting grounded numeric answers.

2 EconWebArena

We describe the construction of **EconWebArena**, a benchmark for evaluating agents on realistic economic tasks in web environments. This includes task generation, human curation and annotation, and answer evaluation criteria.

2.1 Task Generation

We construct the benchmark by prompting four state-of-the-art 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 generates 50 tasks, yielding 200 candidates. The prompt enforces strict rules for clarity, specificity, and feasibility: tasks must be concise questions with a single verifiable answer, use fixed date ranges, and focus on meaningful economic content such as finance, labor, trade, or macroeconomic indicators. All tasks must require realistic web interactions (navigation, filtering, forms) and target distinct named websites. The full prompt is provided in Appendix B.1 Figure 3.

2.2 Task Curation and Annotation

We manually review the 200 generated tasks to ensure clarity, reliability, and feasibility. Only those pointing to verifiable economic data from reputable public sources are retained, discarding vague, redundant, subjective, subscription-based, or non-English cases. For each accepted task, we identify the exact webpage and record the correct numeric answer, yielding 120 high-quality seed tasks. Each task is assigned to one of ten categories (government, finance, markets, labor, banking, energy, trade, real estate, education, or health), covering a wide range of economic domains. The benchmark draws on 82 authoritative websites with definitions in Appendix B.2.

To expand coverage, we create two variants per seed task by altering elements such as time range, country, or indicator. Daily and monthly data are limited to early 2025, while quarterly and annual queries span 2022 to 2025. Each answer must be a single numeric value to enable consistent evaluation. Finalized tasks specify a question, start URL, expected answer format, and required domain. For multilingual sites, we use the English version. This process yields 360 tasks in total, with category distribution shown in Figure 2. Task examples are shown in Appendix B.3 Table 3.

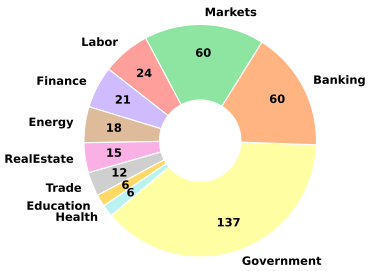


Figure 2: Distribution of task categories in EconWebArena, reflecting the diversity of economic domains represented in the benchmark.

2.3 Spaces and Evaluation

EconWebArena builds on the BrowserGym [Chezelles et al., 2024] framework to support economic tasks in realistic web environments. Each task provides 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 consists of fine-grained browser controls at the element and coordinate levels, covering operations such as mouse movement, text input, tab switching, and form submission. This design supports the low-level interactions needed for complex economic data retrieval (see Section 3.1). A full list of supported actions is shown in Appendix C.1 Table 4.

Agent responses are evaluated on two criteria. First, the answer must include the correct numeric value, exactly matching the gold annotation. Second, the response must contain a valid URL from the required domain, verifying retrieval from the intended authoritative source. A task is correct only if both conditions are satisfied. This procedure enables automated scoring while ensuring factual accuracy and reliability. We report success rate by category and the average number of steps on successful tasks.

3 Experiments

In this section, we evaluate a range of agents on EconWebArena, examining their overall performance and common failure cases.

3.1 Experimental Setup

EconWebArena tasks are implemented in BrowserGym[2] [Chezelles et al., 2024], a lightweight browser simulator supporting fine-grained control of operations such as clicking, typing, filling inputs, hovering, scrolling, tab management, and navigation. Actions are available at both element (via structured identifiers) and coordinate (via pixels) levels, with additional options for user messages and no-op delays. Agents are built with AgentLab[3] [Chezelles et al., 2024], a modular framework for prompting and executing LLM-based agents. Inputs include the AXTree, webpage screenshot, focused element, error logs, and full action history. Prompts enable chain-of-thought reasoning, contextual hints, and real-world examples, while disabling multi-action execution and memory retrieval. Further settings manage element extraction, coordinate formatting, and response style. Full configurations are listed in Appendix C.1 Table 5.

We evaluate multimodal large language models (MLLMs) including 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 limited to 30 steps per task to avoid loops. Due to API credit limits, we run one trial per model. Since each seed task has two variants, results are averaged over three instances per seed task. We also conduct human evaluation, with graduate-level researchers reviewing tasks they had not previously seen. All experiments were run in late May 2025 to ensure consistency across models and live website content.

3.2 Experimental Results

We evaluate model performance on EconWebArena through three research questions (RQs): overall accuracy, sources of failure, and directions for improvement. Table 1 reports accuracy across major categories, while Table 6 (Appendix C.2) provides detailed breakdowns.

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

As shown in Table 1, current LLM agents achieve only partial success. The best-performing model, o4-mini, reaches 46.9%, while GPT-4.1 (31.9%), Claude Sonnet 4 (38.6%), and Gemini 2.5 Flash (31.1%) show mixed results. The open-weight Llama 4 Maverick performs considerably lower at 18.9%. Government and market tasks are relatively easier, while finance, labor, and other specialized

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

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

Table 1: 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	-

domains remain challenging. Human accuracy is 93.3%, underscoring the large gap with expert-level competence. Efficiency also varies: GPT-4.1 completes successful tasks with the fewest steps on average, while Claude Sonnet 4 often uses longer action chains. These results highlight both progress and persistent limitations in handling complex workflows.

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

Table 2: 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 better understand errors, we analyze o4-mini's 193 failures, focusing on 64 seed tasks. Manual review reveals five categories: access issues, data extraction errors, navigation failures, visual understanding failures, and interaction failures. Table 2 shows their distribution, while Appendix C.3 provides examples and details. These errors highlight the difficulty of robustly navigating dynamic websites and interpreting structured or visual economic content.

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

Our analysis points to several promising directions. Robust visual grounding and multimodal reasoning are essential for interpreting charts, tables, and diagrams common in official portals. Combining structured inputs (AXTree, screenshots, coordinates) improves perception and interaction. Planning and self-criticism prompts support multi-step navigation. Hybrid approaches that integrate GUI-based exploration with structured API calls, when available, can boost reliability and efficiency. Finally, domain-adaptive tuning and retrieval-augmented prompting with economic examples may further enhance performance on specialized tasks.

4 Conclusion

We introduce EconWebArena, a benchmark of 360 curated tasks from 82 authoritative websites across ten economic domains. Tasks require agents to navigate live webpages, interpret structured and visual content, and extract precise data through multi-step interactions. Evaluations of state-of-the-art multimodal LLMs reveal major challenges in grounding, navigation, and multimodal reasoning. Error analysis highlights recurring failure modes. EconWebArena thus offers a rigorous and realistic testbed for advancing domain-aware agents for economic data retrieval and reasoning.

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 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.

A.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.

A.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.

B Benchmark Details

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

B.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 3.

B.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).

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

- **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).

B.3 Task Examples

Representative task examples are provided in Table 3. These illustrate the diversity of economic domains, answer formats, and authoritative data sources included in EconWebArena.

Table 3: 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

C Experimental Details

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

C.1 More Experimental Settings

In this section, we provide additional configuration details for all experiments conducted on EconWebArena. Table 5 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[4] [Achiam et al., 2023, OpenAI, 2024a], GPT-4o mini[5] [OpenAI, 2024b], GPT-4.1[6] [OpenAI, 2025a], GPT-4.1 mini[7] [OpenAI, 2025a], o4-mini[8] [OpenAI, 2025b], Claude Sonnet 4[9] [Anthropic, 2025a], Gemini 2.5

[4] `gpt-4o-2024-08-06`: `https://platform.openai.com/docs/models/gpt-4o`
[5] `gpt-4o-mini-2024-07-18`: `https://platform.openai.com/docs/models/gpt-4o-mini`
[6] `gpt-4.1-2025-04-14`: `https://platform.openai.com/docs/models/gpt-4.1`
[7] `gpt-4.1-mini-2025-04-14`: `https://platform.openai.com/docs/models/gpt-4.1-mini`
[8] `o4-mini-2025-04-16`: `https://platform.openai.com/docs/models/o4-mini`
[9] `claude-sonnet-4-20250514`: `https://openrouter.ai/anthropic/claude-sonnet-4`

Flash[10] [Gemini Team, 2023, Google DeepMind, 2025], and Llama 4 Maverick[11] [Grattafiori et al., 2024, Meta AI, 2025].

Table 4: 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

C.2 More Experimental Results

We provide detailed performance metrics by category and model in Table 6, complementing the summary results reported in the main paper.

C.3 Failure Case Definitions

Access issues (Figure 4a) include problems such as blocked or restricted websites, loading errors, or access denials that prevent task execution entirely. Data extraction errors (Figure 4b) 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 4c) reflect difficulty locating the correct webpage or section, including getting stuck or choosing incorrect paths. Visual understanding failures (Figure 4d) are due to an inability to interpret charts, diagrams, or other non-textual representations. Interaction failures (Figure 4e) 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.

C.4 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 C.1 Table 5, and results are reported in Table 7. Each variant alters a single feature based on AgentLab conventions [Chezelles et al., 2024].

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

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

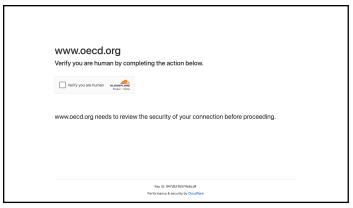
Table 5: Prompt configuration options used in EconWebArena, based on available flags and settings from AgentLab.

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

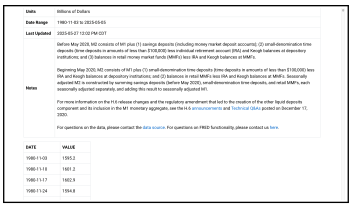
Table 6: Detailed success rates (SR) and average steps (on successful tasks) on EconWebArena by category for models (o4-mini, GPT-4.1, GPT-4.1 mini, GPT-4o, GPT-4o mini, Claude Sonnet 4, Gemini 2.5 Flash, Llama 4 Maverick) and human.

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	-

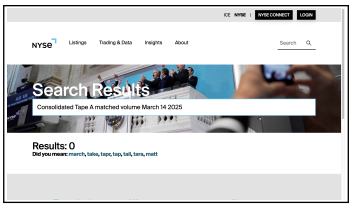
Removing the structured accessibility tree and using only raw HTML (`-use_ax_tree +use_html`) leads to a large performance drop (from 46.9% to 36.7%), confirming the critical role of clean DOM representations. Excluding screenshots (`-use_screenshot`) causes a modest decline (44.7%), but visual information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



(a) Access Issue

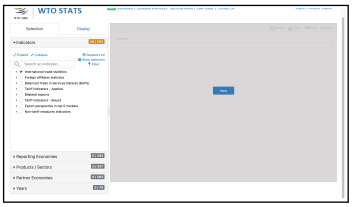


(b) Data Extraction Error



(c) Navigation Failure

(d) Visual Understanding Failure



(e) Interaction Failure

Figure 4: Representative examples of common error types from o4-mini in EconWebArena: access issues, data extraction errors, navigation failures, visual understanding failures, and interaction failures.

Table 7: 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

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.