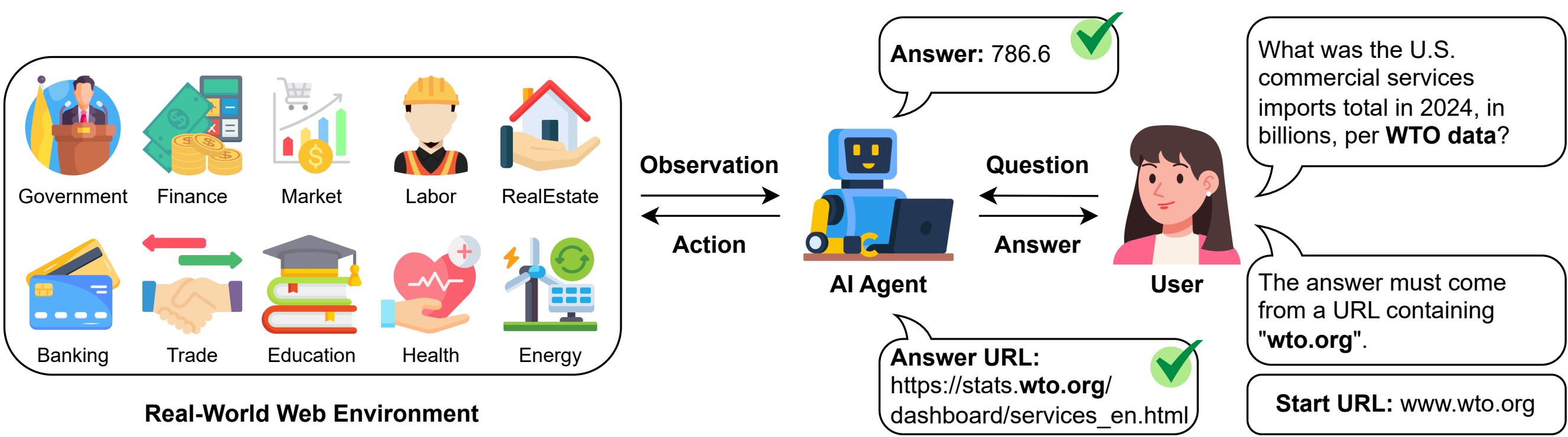


Motivation

- Economic questions often require navigating **complex, authoritative websites** with **tables, charts, and interactive reports**.
- Existing benchmarks are mostly **task-agnostic** and miss the **domain-specific workflows** of real economic analysis.
- EconWebArena** tests whether LLM-based **web agents** can **find, interpret, and extract** accurate data from the **live economic websites**.

Framework



Benchmark

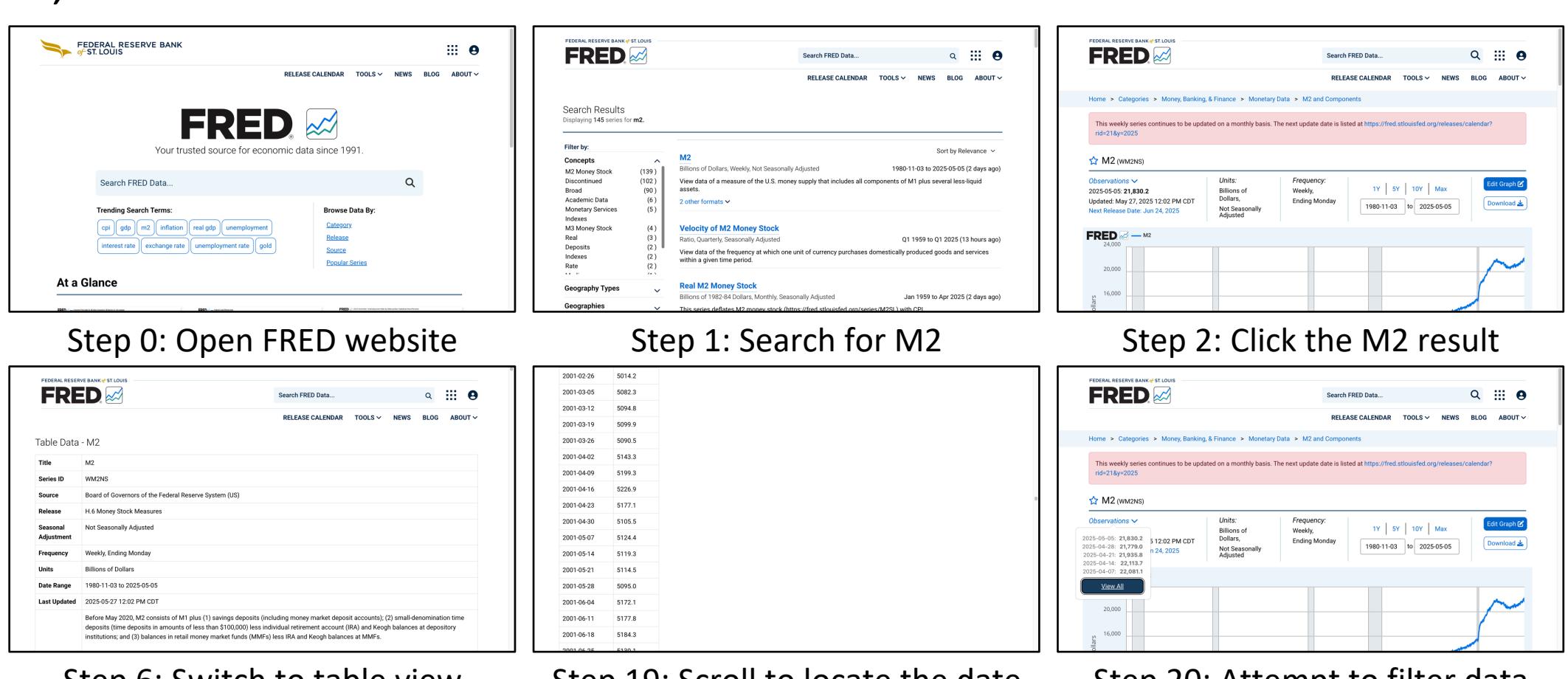
- Includes **360 curated tasks** from **82 real economic websites** across major domains like **finance, labor, and markets**.
- Each task requires navigating **live webpages** and extracting a **verified numeric value** from tables, charts, or documents.
- Tasks are created through a combined **LLM-generated and human-curated process** to ensure realism and diversity.

Experiments

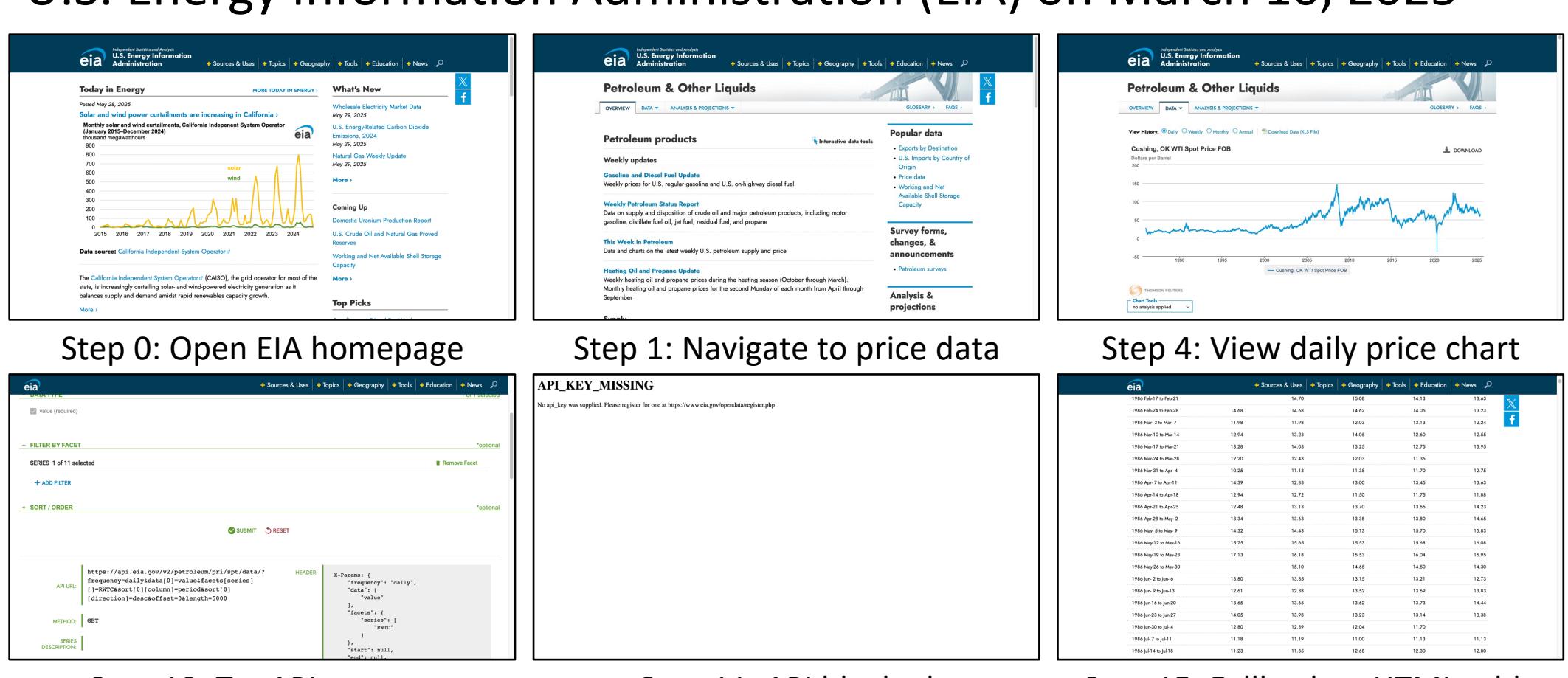
- Agents receive a rich **state** that includes the webpage **screenshot**, **AXTree**, the currently **focused element**, and their **past actions**.
- They choose from a high-level **action space** with operations such as **click, type, scroll, select, navigate, and tab control**.
- Models are evaluated in a **real browser environment**, requiring them to interpret multimodal content and interact with live webpages.
- Each episode is limited to **30 steps**, and success depends on reaching the **target page** and returning the **correct numeric value**.

Error Analysis

Data Extraction Error: Retrieve the M2 money supply for February 1, 2025 from FRED



Visual Understanding Failure: Retrieve the WTI spot price from the U.S. Energy Information Administration (EIA) on March 10, 2025



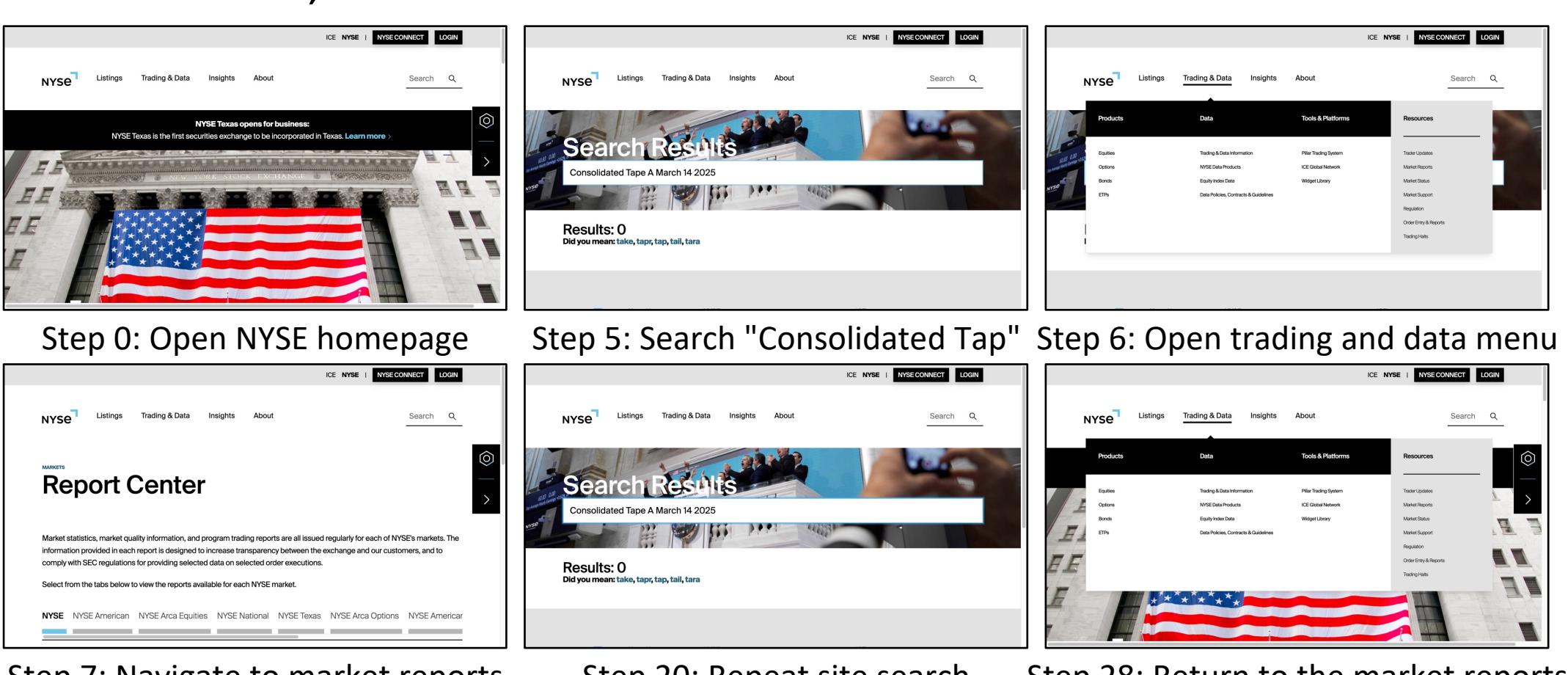
Error Type	Count	Percentage
Access Issue	16	25.0%
Data Extraction Error	16	25.0%
Interaction Failure	8	12.5%
Navigation Failure	15	23.4%
Visual Understanding Failure	9	14.1%
All	64	100.0%

Results

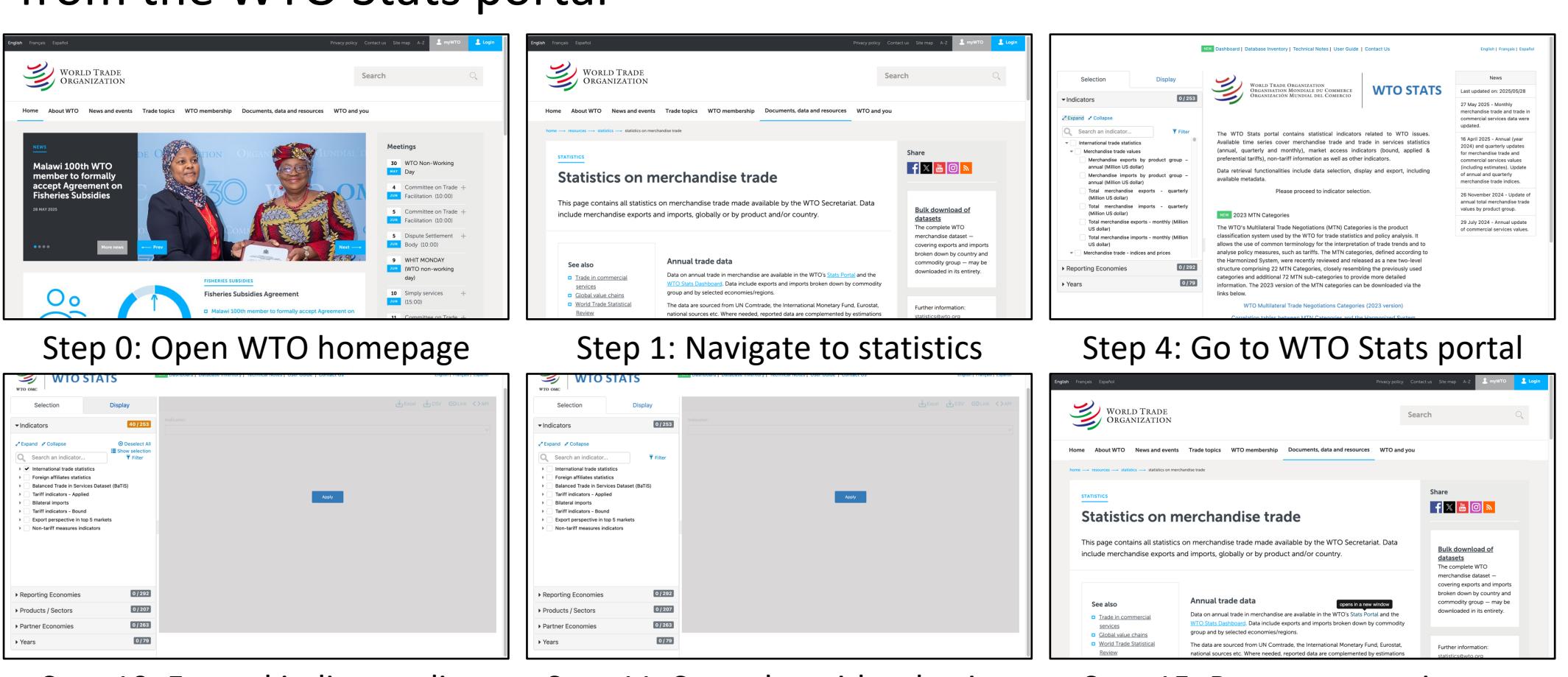
- Across all domains, the best-performing model (**o4-mini**) reaches only **46.9%** success, far below the **93.3%** achieved by humans.
- Performance varies significantly by category, with models doing best on **Government** and **Markets** tasks and struggling most on **Labor** tasks.

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	-

Navigation Failure: Retrieve Consolidated Tape A trading volume for March 14, 2025 from the NYSE website



Interaction Failure: Retrieve Egypt's 2024 merchandise export value from the WTO Stats portal



Conclusion

- EconWebArena** offers a realistic testbed for evaluating **web agents** on **economic data retrieval** across **live, multimodal websites**.
- Results reveal **substantial gaps** from human performance, highlighting the need for stronger **navigation, visual grounding, and numeric accuracy** in future agents.

