WebCanvas

Benchmarking Web Agents in Online Environments

Carnegie Mellon University

iMean.AI
Web Canvas

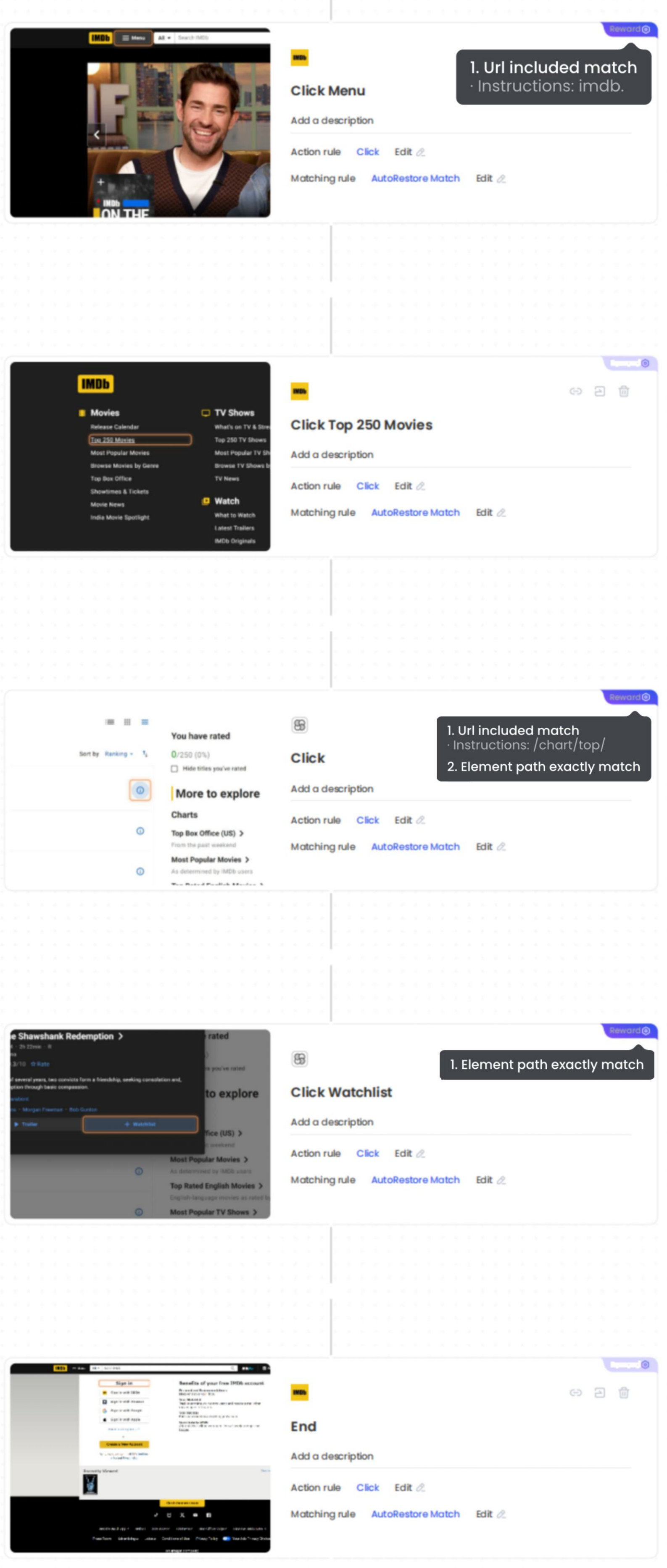
Yichen Pan*♠, Dehan Kong*♠, Sida Zhou*♠, Cheng Cui*♠, Yifei Leng♠, Bing Jiang♠, Hangyu Liu♠, Yanyi Shang♠, Shuyan Zhou♣, Tongshuang Wu♣, Zhengyang Wu♠
♠iMean AI, ♣Carnegie Mellon University mail: dehan@imean.ai

- Web agent evaluations are either static or operate within a limited environment.
- We eventually need to connect the agent with the real world with Total Web Dynamics.
- We craft and host your agentic data for you to connect your offline agent with the online environment!

We build better interface for you to talk to your data

50 ↓ %
in annotation

80 ↓ %
in maintainance



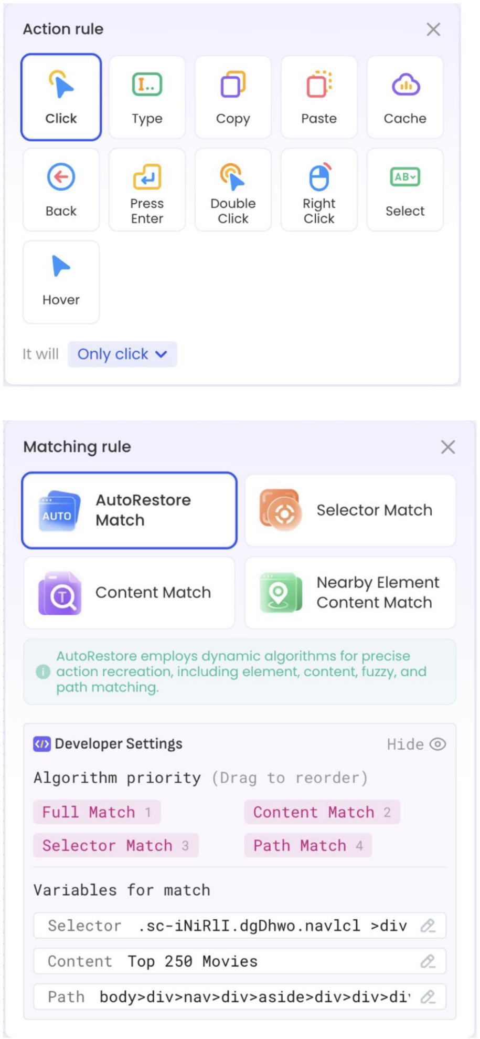
Click Menu
Add a description
Action rule Click Edit
Matching rule AutoRestore Match Edit
1. Url included match
· Instructions: imdb.

Click Top 250 Movies
Add a description
Action rule Click Edit
Matching rule AutoRestore Match Edit

Click
Add a description
Action rule Click Edit
Matching rule AutoRestore Match Edit
1. Url included match
· Instructions: /chart/top/
2. Element path exactly match

Click Watchlist
Add a description
Action rule Click Edit
Matching rule AutoRestore Match Edit
1. Element path exactly match

End
Add a description
Action rule Click Edit
Matching rule AutoRestore Match Edit



It will Only click

AutoRestore employs dynamic algorithms for precise action recreation, including element, content, fuzzy, and path matching.

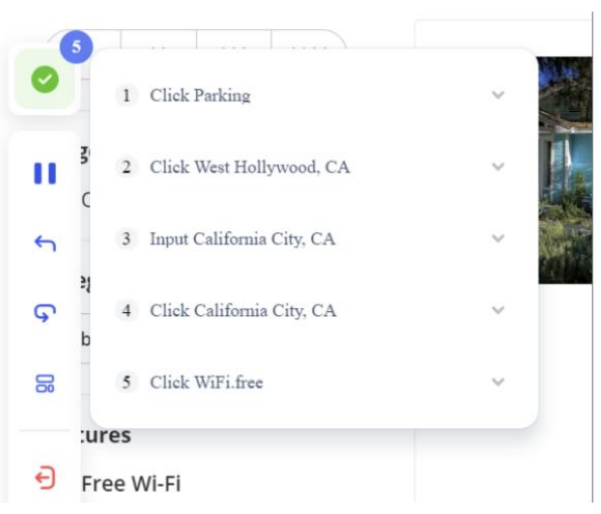
```
Algorithm priority (Drag to reorder)
Full Match 1          Content Match 2
Selector Match 3      Path Match 4

Variables for match
Selector  .sc-iNiRlI.dgDhwo.navlcl >div
Content  Top 250 Movies
Path  body>div>nav>div>aside>div>div>di
```

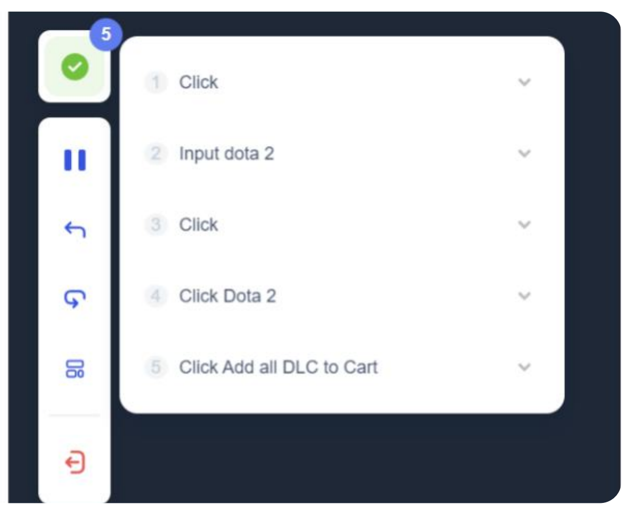
Debug your trajectory

Procedures to create your own benchmark online

- Step1: Define your tasks.
- Step2: Record your trajectories using iMean Builder (We support a broad range of action space, even with data operations).

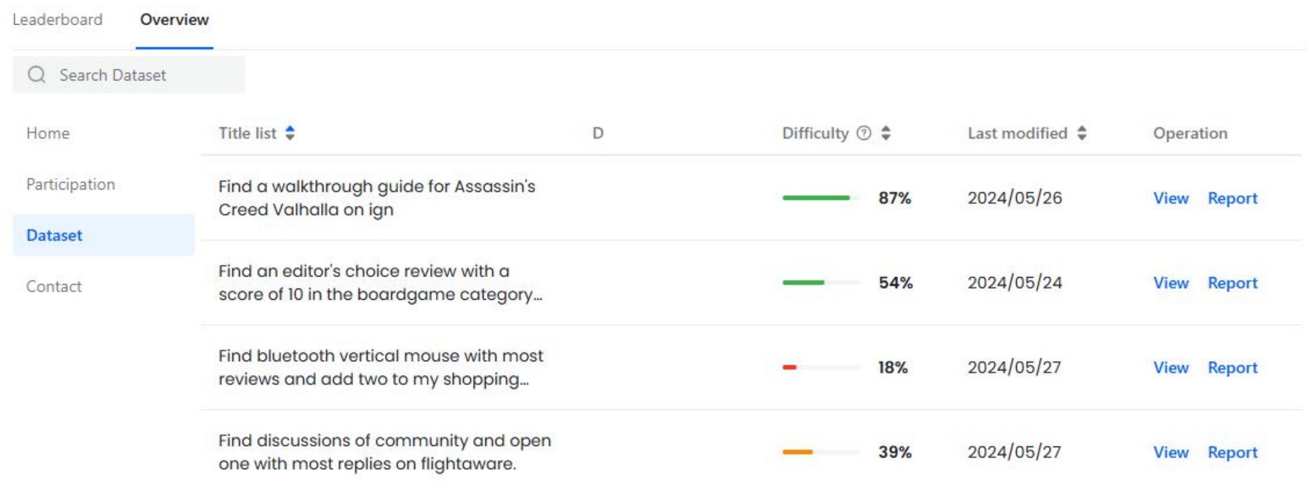


1 Click Parking
2 Click West Hollywood, CA
3 Input California City, CA
4 Click California City, CA
5 Click WiFi.free



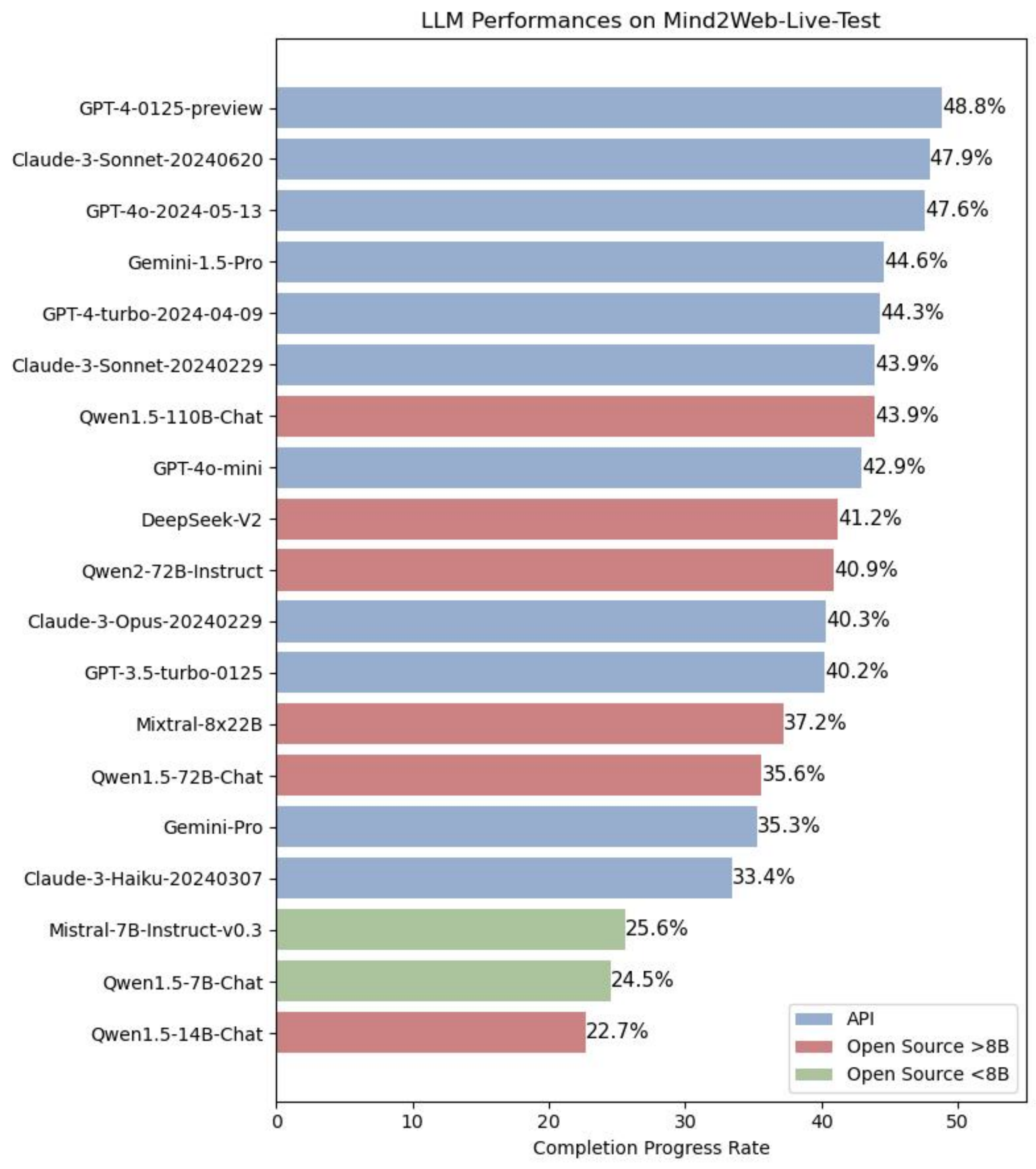
1 Click
2 Input dota 2
3 Click
4 Click Dota 2
5 Click Add all DLC to Cart

- Step3: Annotate the Key Nodes (static rule based evaluation for now).
- Step4: Connect the data to WebCanvas by creating a challenge for the community (You can make it private during construction).



Title list	D	Difficulty	Last modified	Operation
Find a walkthrough guide for Assassin's Creed Valhalla on ign		87%	2024/05/26	View Report
Find an editor's choice review with a score of 10 in the boardgame category...		54%	2024/05/24	View Report
Find bluetooth vertical mouse with most reviews and add two to my shopping...		18%	2024/05/27	View Report
Find discussions of community and open one with most replies on flightaware.		39%	2024/05/27	View Report

- Step5: Evaluate your web agents online by easily integrating WebCanvas and accumulate insights of web agent performances as a community.



Takeaways

- Better toolkits save you time in agentic data management.
- Key-node based metrics is essential for evaluating agents in the wild.
- Models finetuned on static datasets struggle to generalize in online environments a year later.
- Self reward doesn't help, but with human-labeled rewards as a reference, agents improve.
- Less capable models don't benefit from memory and ReAct in web tasks.
- Agent performance varies by domain, website and physical environment.

Step Forward

	Completion Rate	Task Success Rate	USD / Key Node Score
GPT-3.5-turbo	42.5%	17.3%	0.092
GPT-4o	51.4%	28.8%	0.142
GPT-4o mini	42.9%	21.5%	0.004

- Better observation: accurate and fast to compute
- Better conversion: human interface -> agent interface
- Dynamic evaluation functions
- Cloud environment: increase reliability
- Secured action of web agents
- Error handling and authorization

Web agent efficiency should gain more attention!
GPT-4o mini is >30 times more cost efficient than GPT-4o

OUR VISON IS TO CREATE THE ULTIMATE LIVE SANDBOX FOR WEB AGENT ADVANCEMENTS

↓ ↓ ↓ WE WANT YOU TO JOIN WITH US AS A COMMUNITY! ↓ ↓ ↓

Paper

Github

LiveDemo