

# Web Agents

October 30th

# Project goal recap

- An agent on top of a LLM or VLM can solve tasks by navigating the web.
  - But simple agents like this will struggle with complex or long horizon tasks
- RL might make agents better
  - But it's inefficient and relies on a reasonable initial performance to improve from
- Supervised fine-tuning has the potential to significantly improve web agent performance on challenging tasks
  - But where is the data?

**Solution:** watch people complete tasks.



# Browser Agent Benchmarks

- WebVoyager ( real world websites)
- Web Bench
- BrowseComp
- Mind2Web
- WEBARENA
- GAIA
- WebDS

# Enterprise Benchmarks

- Work Bench( real world websites)
- Spreadsheet Bench
- DocBench
- MailBench
- SlideBench
- CalendarBench

# Potential opening

- There was no cross App interaction across the Gsuite , M365, slack.

# Project Update

- We're planning focus to Work Arena as our main environment
- And to try this out before trying any of the cross App
- Using our custom Chrome extension to record real human trajectories
  - HTML snapshots
  - Accessibility tree
  - Video timestamps
  - Action logs
- Goal: Build a high-quality dataset to fine-tune and ev



# WorkArena Tasks

## Task Levels and Their Characteristics

- L1 Tasks (WorkArena):
  - ~200 tasks
  - Single-step, simple actions (e.g., filling a form, clicking a button, simple queries)
  - Designed to test basic perception, action grounding, navigation.
- L2 Tasks (WorkArena++):
  - ~300 tasks
  - Multi-step procedures with intermediate planning (e.g., navigating lists, combining filter and retrieval, submitting forms in sequence)
  - Tests for short-horizon reasoning, data extraction, and manipulation.
- L3 Tasks (WorkArena++):
  - ~182 tasks
  - Require complex reasoning, compositionality (e.g., planning multi-step workflows, handling conditional logic, integrating information from multiple locations within ServiceNow)
  - Test long-horizon planning & generalization.

# Task Diversity

## Task Types Include:

- Catalog ordering (e.g., order an iPad Pro, submit an IT service request)
- Filtering and sorting lists (users, incidents, hardware)
- Dashboards (retrieving and combining information from multiple visualizations)
- Multi-field form filling
- Approval routing and basic workflow management
- Information retrieval, search, and report submission

# Evaluation Protocol

Automatic validation:

- Each task has a reference solution and validation logic.

Seed diversity:

- For each task, they will have multiple seed variations.

# BrowserGym Leaderboard

WorkArena-L1 <span>⌵</span> <span>○ Ascending</span> <span>● Descending</span>								
Agent	WebArena	WorkArena-L1	WorkArena-L2	WorkArena-L3	MiniWoB	WebLINX	VisualWebArena	AssistantBench
<a href="#">GenericAgent-GPT-5</a>	-	79.10	69.40	11.50	71.50	-	-	-
<a href="#">GenericAgent-Claude-4-Sonnet</a>	-	63.30	40.40	-	70.70	-	-	-
<a href="#">GenericAgent-GPT-5-mini</a>	-	60.60	47.70	-	71.00	-	-	-
<a href="#">GenericAgent-GPT-o1-mini</a>	28.60	56.70	14.90	0.00	67.80	12.50	-	6.90
<a href="#">GenericAgent-Claude-3.5-Sonnet</a>	36.20	56.40	39.10	0.40	69.80	13.70	21.00	5.20
<a href="#">GenericAgent-GPT-oss-120b</a>	-	50.90	11.50	-	66.40	-	-	-
<a href="#">GenericAgent-o3-mini</a>	-	48.20	-	-	-	-	-	-
<a href="#">GenericAgent-GPT-4o</a>	31.40	45.50	8.50	0.00	63.80	12.50	26.70	4.80
<a href="#">GenericAgent-Llama-3.1-405b</a>	24.00	43.30	7.20	0.00	64.60	7.90	-	3.90
<a href="#">GenericAgent-GPT-5-nano</a>	-	40.60	3.40	-	64.80	-	-	-
<a href="#">GenericAgent-GPT-oss-20b</a>	-	38.50	2.60	-	64.00	-	-	-
<a href="#">GenericAgent-AgentTrek-1.0-32b</a>	22.40	38.29	2.98	0.00	60.00	-	-	-
<a href="#">GenericAgent-Llama-3.1-70b</a>	18.40	27.90	2.10	0.00	57.60	8.90	-	2.80
<a href="#">GenericAgent-GPT-4o-mini</a>	17.40	27.00	1.30	0.00	56.60	11.60	16.90	2.10

# Baseline Experiment

Agents	WorkAreana Tasks		
	L1	L2	L3
GenericAgent GPT5	79.1	69.4	11.5
Generic Agent gpt-4o-mini	27	1.3	0
<b>Generic Agent gpt-4o-mini</b>	<b>31.8</b>	<b>1.8</b>	<b>0</b>

# Methodology

- Collect → Human demonstrations using the browser extension
- Post-process → (State, Action, Next State) triplets from HTML, AxTree, and video.
- Train → Fine-tune the base LLM with supervised learning.
- Evaluate → Benchmark on WorkArena
- Analyze → Check generalization and overfitting effects.

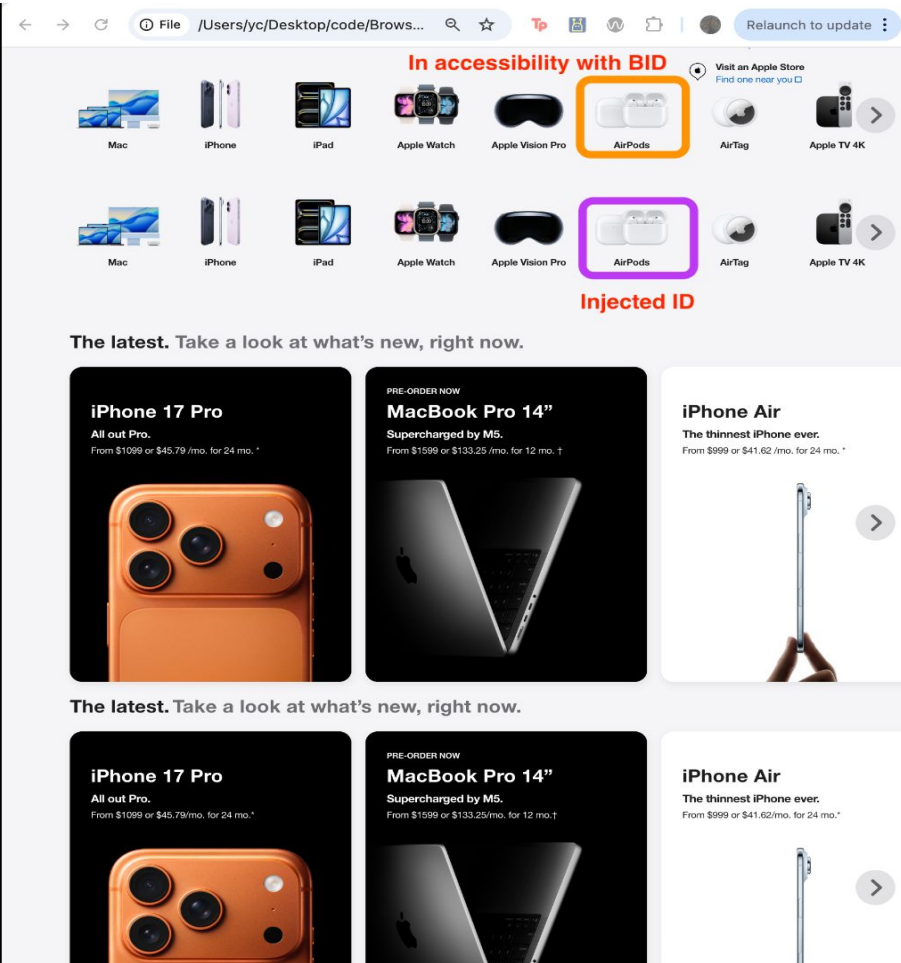
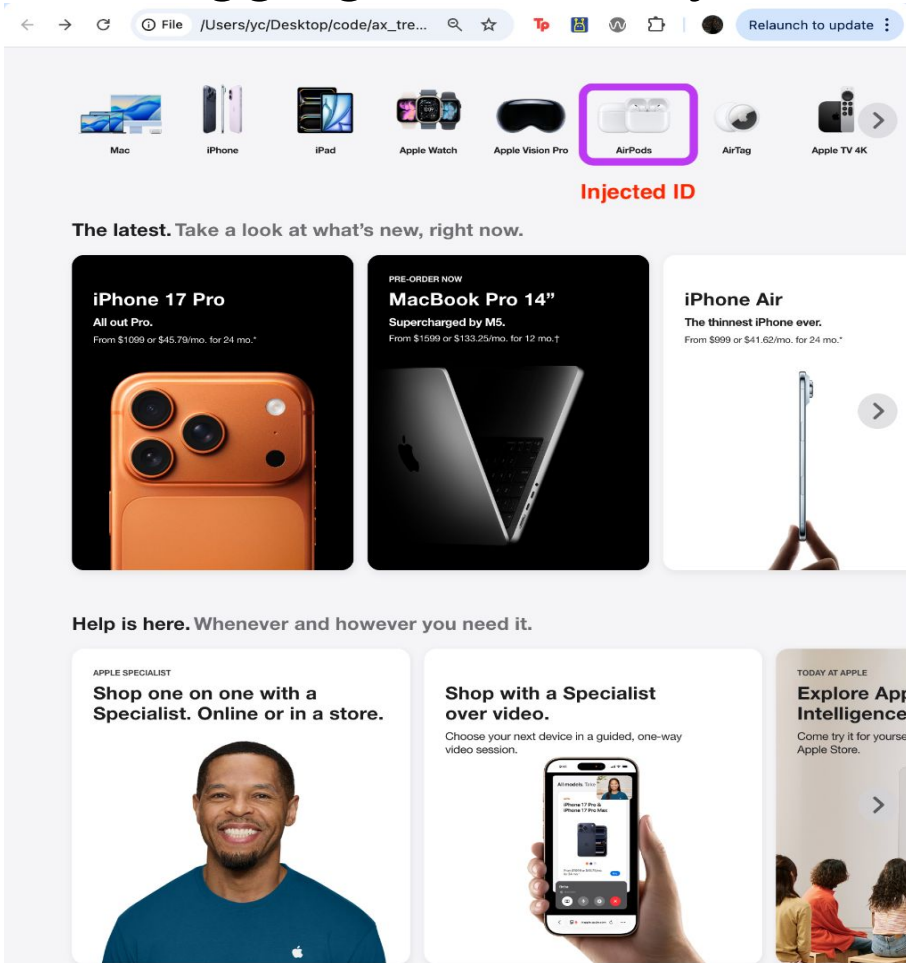
# Why This Is Novel

Turning the stream of data into meaningful training samples

- Most prior work uses synthetic or scripted web trajectories.
- We collect natural, human trajectories directly from task executions.
- Few papers explore fine-tuning on domain-specific enterprise web data.



# Debugging BrowserGym



# Amazon.com only undergo minor changes

File /Users/jc/Desktop/code/Brows... Relaunch to update

amazon Delivering to Santa Clara 95050 Update location All chocolate EN Hello, sign in Account & Lists Returns & Orders 0 Cart

All Amazon Haul Medical Care Best Sellers Books Amazon Basics New Releases Registry Today's Deals NBA on Prime Friday 6:30PM ET

1-48 of over 100,000 results for "chocolate" Sort by: Featured

### Popular Shopping Ideas

Vegan  
Sugar Free  
Sea Salt  
Organic  
See more

### Eligible for Free Shipping

Free Shipping by Amazon  
Get FREE Shipping on eligible orders shipped by Amazon

### Delivery Day

Get It Today  
Get It by Tomorrow

### Price

\$0 - \$1,450+

Up to \$10  
\$10 to \$15  
\$15 to \$20  
\$20 to \$30  
\$30 & above

### Deals & Discounts

All Discounts  
Today's Deals

### From Our Brands

Amazon Brands

### SNAP EBT

SNAP EBT Eligible

### Customer Reviews

★ ★ ★ ★ ★ & Up

### All Top Brands


Top Brands

### Brands


Lindt  
HERSHEY'S  
Cadbury  
Dove  
M&M's  
Reese's

### HERSHEY'S


Everyday is sweeter with Hershey's  
Shop Hershey's Chocolate



HERSHEY'S NUGGETS Assorted Chocolate... ★ ★ ★ ★ ★ 23,228 prime




HERSHEY'S, KIT KAT and REESE'S Miniatures Assorted... ★ ★ ★ ★ ★ 36,585 prime




HERSHEY'S and REESE'S Miniatures Assorted... ★ ★ ★ ★ ★ 638 prime

### Results


Check each product page for other buying options.



Sponsored  
Feastables by MrBeast Milk Chocolate KING Size Chocolate Bar, 2.1oz (60g), 10 count  
Milk Chocolate  
10 Count (Pack of 1)  
Options: 7 flavors  
4.1 ★ ★ ★ ★ ★ (1.8K)  
800+ bought in past month  
\$26<sup>15</sup> (\$1.23/ounce)  
\$24.84 with Subscribe & Save discount  
SNAP EBT eligible  
FREE delivery Sun, Nov 2 on \$35 of items shipped by Amazon  
Or fastest delivery Thu, Oct 30 Arrives before Halloween  
Add to cart



Sponsored  
Feastables MrBeast Milk Chocolate Bar, Full Size Bar for Adults & Kids, Candy Bars for...  
Milk Chocolate 24 Count  
Options: 2 sizes, 3 flavors  
4.0 ★ ★ ★ ★ ★ (292)  
200+ bought in past month  
\$34<sup>99</sup> (\$28.22/ounce)  
SNAP EBT eligible  
FREE delivery Sun, Nov 2 on \$35 of items shipped by Amazon  
Or fastest delivery Tomorrow, Oct 29 Arrives before Halloween  
Add to cart



Sponsored  
Michel et Augustin Chocolate Cookie Squares, Halloween Chocolate Gift Box, Dark Chocol...  
4.4 ★ ★ ★ ★ ★ (57)  
3K+ bought in past month  
\$13<sup>49</sup> (\$0.45/count) Typical: \$14.99  
FREE delivery Oct 30 - 31  
Or FREE pickup  
Add to cart

File /Users/jc/Desktop/code/Brows... Relaunch to update

### Popular Shopping Ideas

Vegan  
Sugar Free  
Sea Salt  
Organic  
See more

### Eligible for Free Shipping

Free Shipping by Amazon  
Get FREE Shipping on eligible orders shipped by Amazon

### Delivery Day

Get It Today  
Get It by Tomorrow

### Price

\$0 - \$1,450+

Up to \$10  
\$10 to \$15  
\$15 to \$20  
\$20 to \$30  
\$30 & above

### Deals & Discounts

All Discounts  
Today's Deals

### From Our Brands

Amazon Brands

### SNAP EBT

SNAP EBT Eligible

### Customer Reviews

★ ★ ★ ★ ★ & Up

### All Top Brands


Top Brands

### Brands


Lindt  
HERSHEY'S  
Cadbury  
Dove  
M&M's  
Reese's  
VALRHONA

### HERSHEY'S


Everyday is sweeter with Hershey's  
Shop Hershey's Chocolate



HERSHEY'S NUGGETS... ★ ★ ★ ★ ★ 23,228 prime




HERSHEY'S, KIT KAT... ★ ★ ★ ★ ★ 36,585 prime




HERSHEY'S... ★ ★ ★ ★ ★ prime

### Results


Check each product page for other buying options.



Sponsored  
Feastables by MrBeast Milk Chocolate KING Size Chocolate Bar, 2.1oz (60g), 10 count  
Milk Chocolate  
10 Count (Pack of 1)  
Options: 7 flavors  
4.1 ★ ★ ★ ★ ★ (1.8K)  
800+ bought in past month  
\$26<sup>15</sup> (\$1.23/ounce)  
\$24.84 with Subscribe & Save discount  
SNAP EBT eligible  
FREE delivery Sun, Nov 2 on \$35 of items shipped by Amazon  
Or fastest delivery Tomorrow, Oct 29 Arrives before Halloween  
Add to cart



Sponsored  
Feastables MrBeast Milk Chocolate Bar, Full Size Bar for Adults & Kids, Candy Bars for...  
Milk Chocolate  
24 Count  
Options: 2 sizes, 3 flavors  
4.0 ★ ★ ★ ★ ★ (292)  
200+ bought in past month  
\$34<sup>99</sup> (\$28.22/ounce)  
SNAP EBT eligible  
FREE delivery Sun, Nov 2 on \$35 of items shipped by Amazon  
Or fastest delivery Tomorrow, Oct 29 Arrives before Halloween  
Add to cart



Sponsored  
Michel et Augustin Chocolate Cookie Squares, Halloween Chocolate Gift Box, Dark Chocol...  
4.4 ★ ★ ★ ★ ★ (57)  
3K+ bought in past month  
\$13<sup>49</sup> (\$0.45/count) Typical: \$14.99  
FREE delivery Oct 30 - 31  
Or FREE pickup  
Add to cart

# How it look like on our target website

empmassimo23.service-now.com/... Relaunch to update

**MIGHTY CAPITAL** All : **Loaner Laptop** ☆

< [Service Catalog](#) > [Hardware](#) > [Loaner Laptop](#) Search catalog

**Short term, while computer is repaired/imaged. Waiting for computer order, special projects, etc. Training, special events, check-in process**

Did you break your laptop? Maybe lost it? Need a temporary loaner? We can help.

In order to take advantage of a loaner notebook computer, you must meet company eligibility requirements per the Notebook Loaner Policy

Loaner laptops will be provided based on what devices are available.

When do you need it ?

How long do you need it for ?

1 day

**Order this Item**

Quantity 1

Delivery time 2 Days

**Order Now**

**Add to Cart**

**Shopping Cart**

Empty

Live ServiceNow Admin Page



# Target elements are in iFrame

empmassimo23.service-now.com/... Relaunch to update

MIGHTY CAPITAL All : **Loaner Laptop**

< Service Catalog > Hardware > Loaner Laptop ...

**Short term, while computer is repaired/imaged. Waiting for computer order, special projects, etc. Training, special events, check-in process**

Did you break your laptop? Maybe lost it? Need a temporary loaner? We can help.

In order to take advantage of a loaner notebook computer, you must meet company eligibility requirements per the Notebook Loaner Policy

Loaner laptops will be provided based on what devices are available.

When do you need it ?

How long do you need it for ?

**Order this Item**

Quantity

Delivery time 2 Days

**Shopping Cart**

Empty

File /Users/yc/Desktop/code/Brows... Relaunch to update

Back 

- [Service Catalog](#)
- [Hardware](#)
- [Loaner Laptop](#)

+ -3

**Short term, while computer is repaired/imaged. Waiting for computer order, special projects, etc. Training, special events, check-in process**

Did you break your laptop? Maybe lost it? Need a temporary loaner? We can help.

In order to take advantage of a loaner notebook computer, you must meet company eligibility requirements per the Notebook Loaner Policy

Loaner laptops will be provided based on what devices are available.

When do you need it ?

How long do you need it for ?

**Order this Item**

Quantity

Delivery time 2 Days

**Shopping Cart**

Empty

# Prompt to generate similar task data

I will provide you original screenshot with the interacted element highlighted in a red box, raw **HTML**, **original task**, and **correct response**. Your job is to generate a response based on the same screenshot and HTML, but with a **different task**.

You need to generate a JSON object inside "" "" ". The object has two key "element" and "action".

1. Element: the element's nodeId
2. Action: an action within action space defined below.

Action Space:

- `click`: This action clicks on an element with a specific id on the webpage.
- `type [content] [press\_enter\_after=0|1]`: Use this to type the content into the field. By default, the "Enter" key is pressed after typing unless press\_enter\_after is set to 0.
- `hover`: Hover over an element.
- `press [key\_comb]`: Simulates the pressing of a key combination on the keyboard (e.g., Ctrl+v).
- `scroll [down/up]`: Scroll the page up or down. You need to output the command like scroll [down] to scroll down.
- `goto [url]`: Navigate to a specific URL.
- `go\_back`: Navigate to the previously viewed page.
- `go\_forward`: Navigate to the next page (if a previous 'go\_back' action was performed).
- `stop [answer]`: Issue this action when you believe the task is complete. If the objective is to find a text-based answer, provide the answer in the bracket. If you believe the task is impossible to complete, provide the answer as "N/A" in the bracket.

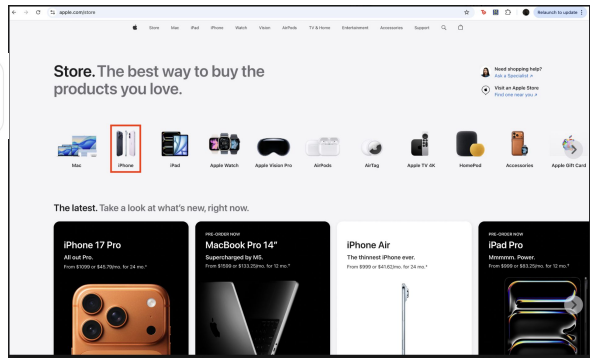
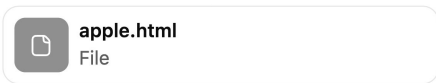
Example response "" ""

```
{
  "element": "document.querySelector("body > 123").querySelector("#item").querySelector("button")"
  "action": "click"
}
```

The original task is "buy an iPhone".  
The response is "" ""

```
{
  "element": "document.querySelector("body > 123").querySelector("#item").querySelector("button")" ,
  "action": "click"
}
```

Now, generate a response with new task "Buy an iPad".



# Verification

# Why do we need to verify?

Verification ensures that both **data and actions** produced by the agent are **accurate, interpretable, and correctable** - across *training* and *deployment*.

**Garbage in  $\Leftrightarrow$  Garbage out**

# When do we verify?

## **During Data Collection / Training:**

1. Validate event correctness
2. Filter “non consequential events”
3. Grounded learning/ Action - Consequence - (state, action, next\_state)

## **During Inference / Runtime**

1. Real-time action validation - 200 OK
2. Self-correction and recovery
3. State tracking and awareness
4. Reference Replay Comparison 🙌

# How do we get started on Verification?

## **Step Verification:**

Validation of each **atomic browser action** against an **immediate, observable ground truth**.

## **Flow-Level Verification:**

Evaluation of the entire **multi-step trajectory** against a clearly defined **task goal and reference outcome**

Do we need both? - Ideally yes

Step-level = **local accuracy** → catches micro-failures early

Flow-level = **global correctness** → ensures end-to-end success

# But what exactly is getting verified?

## 1. Action preconditions

If element is present in the DOM, visible, interactable (`element.clickable == true`).

## 2. Action intent match

The selected element's semantic role, label, or text aligns with the expected action description (from the plan or prompt).

## 3. Action execution result

measurable post-state change (e.g., DOM mutation, network request, navigation event, or updated field value).

# But what are we verifying against?

## **Source of truth**

1. Browser runtime observations (DOM state, URL, network logs).
2. Deterministic expected conditions in the prompt/ terminal states
3. Reference Replay

# In Summary

Signal Type	What It Measures	Verification Purpose
<b>Selector validity</b>	Existence, Visibility, Interactability	Detects broken or outdated selectors
<b>Event return code</b>	Whether browser API (e.g., <code>click()</code> ) returned <b>OK</b>	Confirms the action executed without exceptions
<b>HTTP status</b>	Status of triggered network requests (e.g., 200 OK)	Validates that backend accepted the operation
<b>DOM diff</b>	Change in DOM tree before vs after the action	Confirms a meaningful state change occurred
<b>Screenshot hash</b>	Pixel-level checksum difference	Detects visible page updates or failures

# Discussion

- Help us design some experiments
  - How to verify that the extension is collecting what we want it to and aligns with BrowserGym observations
  - How to determine good inputs for the underlying model(s) without overwhelming ourselves with a billion different possibilities



Spot the difference?