

WorkArena++

Compositional Planning and Reasoning-based Knowledge Work Tasks



Massimo



Alex L



Max



Thibault



Léo



Megh



Alex D



Nicolas

Léo Boisvert
ServiceNow Research



Making a basic Web UI Agent



Prompt

- Task Description
- Web Page as text
- Action Space



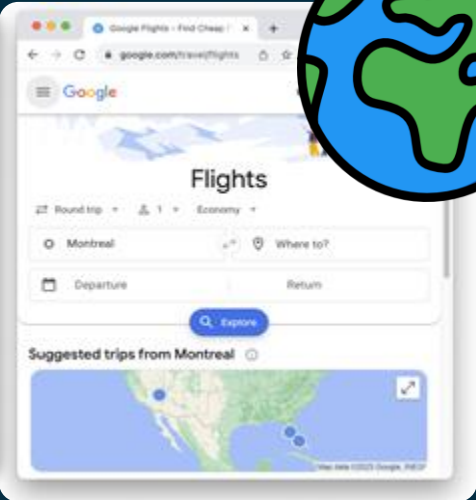
Answer

- Action 1
- Action 2



Task

Fly me to Yellowstone for the next long weekend



Execute actions

- Python + Playwright

You can do this by prompting GPT-4

Example prompt (simplified):

```
Task:
- Enter "Enola" into the text field and press Submit.

DOM (Web Page):
<html>
<body>
...
</body>
</html>

Action space:

# Fill out a form field
fill(backend_id: str, value: str)

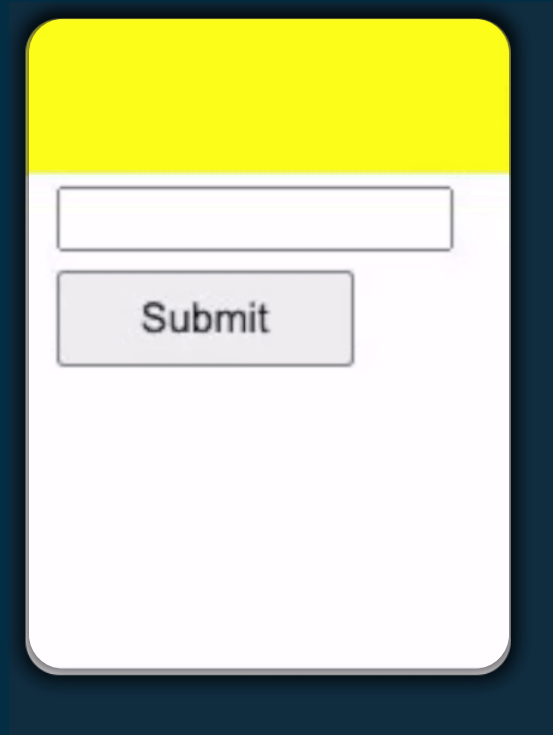
# Click an element
click(backend_id: str)

# Move the mouse to a location
mouse_move(x: float, y: float)



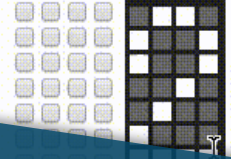




Answer Format:
<action>
Your actions
</action>
```

LLM response:

```
<action>
fill('14', 'Enola')
click('15')
</action>
```



How do we evaluate web agents?

<p>Move the cube around so that "5" is the active side facing the user.</p>  <p>Submit</p>	<p>Set the sliders to the combination [13,20,13] and submit.</p>  <p>Submit</p>	<p>Draw the number "2" in the checkboxes using the example on the right and press Submit when finished.</p>  <p>Submit</p>	<p>Select 5 with the spinner and hit Submit.</p> <p>Select a value: 0</p> <p>Submit</p>	<p>Keep your mouse inside the circle as it moves around.</p> 	<p>Move the cube around so that "4" is the active side facing the user.</p>  <p>Submit</p>
<p>Copy the text in the text area below and paste it into the text box.</p> <p>Enim. Elementum, tortor ullamcor</p> <p>Submit</p>	<p>Select 09/23/2016 as the date and hit Submit.</p> <p>09/23/2016</p> <p>Submit</p>	<p>Drag all rectangles into the text field and hit Submit.</p> <p>qu</p> <p>Submit</p>	<p>Select all the shades of blue and press Submit.</p>  <p>Submit</p>	<p>Find the 4th word in the paragraph, type that into the textbox and press "Submit".</p> <p>Non arcu ut ultricies est. Gravida gravida. Porta erat nulla eget condimentum posuere a</p> <p>Submit</p>	
<p>Enter an item that starts with "Tuni".</p> <p>Tags:</p> <p>Submit</p>	<p>Enter "Vb8" into the text field and press Submit.</p> <p>Vb8</p> <p>Submit</p>	<p>Focus into the 1st input textbox.</p> <p>Submit</p>	<p>Focus into the text box.</p> <p>Submit</p>	<p>Move the cube around so that "2" is the active side facing the user.</p>  <p>Submit</p>	<p>Select 5Gi and click Submit.</p> <p>5Gi sPUT</p> <p>Submit</p>

This is **UNREALISTIC**

More realistic benchmarks have been proposed

WebArena: A Realistic Web Environment for Building Autonomous Agents

Shuyan Zhou^{1*}, Frank F. Xu^{1*},
Hao Zhu¹⁺, Xuhui Zhou¹⁺, Robert Lo¹⁺, Abishek Sridhar¹⁺,
Xianyi Cheng¹, Tianyue Ou¹, Yonatan Bisk¹, Daniel Fried¹, Uri Alon¹, Graham Neubig^{1,2}.

¹Carnegie Mellon University, ²Inspired Cognition
*Lead contributors. +Equal contribution.
{shuyanzh, fangzhex, gneubig}@cs.cmu.edu

[Paper](#) [Code](#) [Data](#) [Docker Environment](#) [Leaderboard](#)

The diagram illustrates the WebArena environment. On the left, a box labeled 'WebArena' contains 'Self-hosted fully functional web applications' (OneStopShop, CMS, reddit, GitLab), a 'Toolbox' (calculator, graph, map, etc.), and 'Knowledge resources' (handbook, globe, etc.). An 'Agent' (a colorful robot head) is shown in the center, performing 'Action' on the applications and receiving 'Feedback'. Two example prompts are shown: 'Tell me how much I spent on food purchase in March 2023' and 'Create a 'NolanFans' repo, listing Nolan's Oscar-winning films in a README file'. Below the agent, a terminal window shows commands like 'check_repo', 'check_readme', and 'check_answer', with a green checkmark for 'Functional Success' and a red X for 'Functional Failure'.

Benchmark Explosion



- MiniWoB++ (Shi et al., 2017; Liu et al., 2018) **125 tasks**
- WebShop (Yao, Chen et al., 2022) **12 087 tasks**
- WebArena (Zhou et al., 2023) **812 tasks**
- VisualWebArena (Koh et al., 2024) **910 tasks**
- WebLINX (Lù et al., 2024) **2 300 tasks**
- WebCanvas (Pan et al., 2024) **438 tasks**
- WebVoyager (He et al., 2024) **643 tasks**
- AssistantBench (Yoran et al., 2024) **214 tasks**

But!

....none are specifically designed for enterprise workflows





WorkArena



`pip install browsergym-workarena`

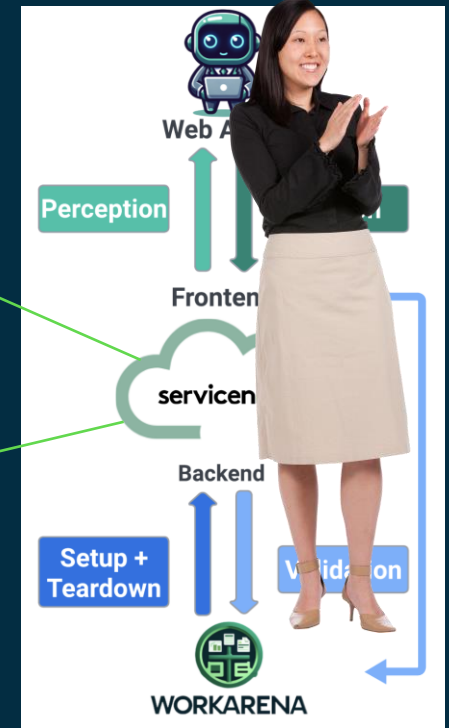
An open-source benchmark of 682 work-related tasks built on the ServiceNow platform

Dashboard

Knowledge Base

Service Catalog

servicenow.com
Your instance **Free**
Your instance URL: [https://instance-name.service-now.com](#)
Username: admin
Current password:
Keep your new instance active by developing on the instance or logging into the Developer Site. If you do not log in for 10 days, it will be reclaimed and released for other developers to use.
Return to the Developer Site



Tasks span basic UI interactions and complex realistic workflows

Powered by PDIs



BrowserGym



`pip install browsergym`

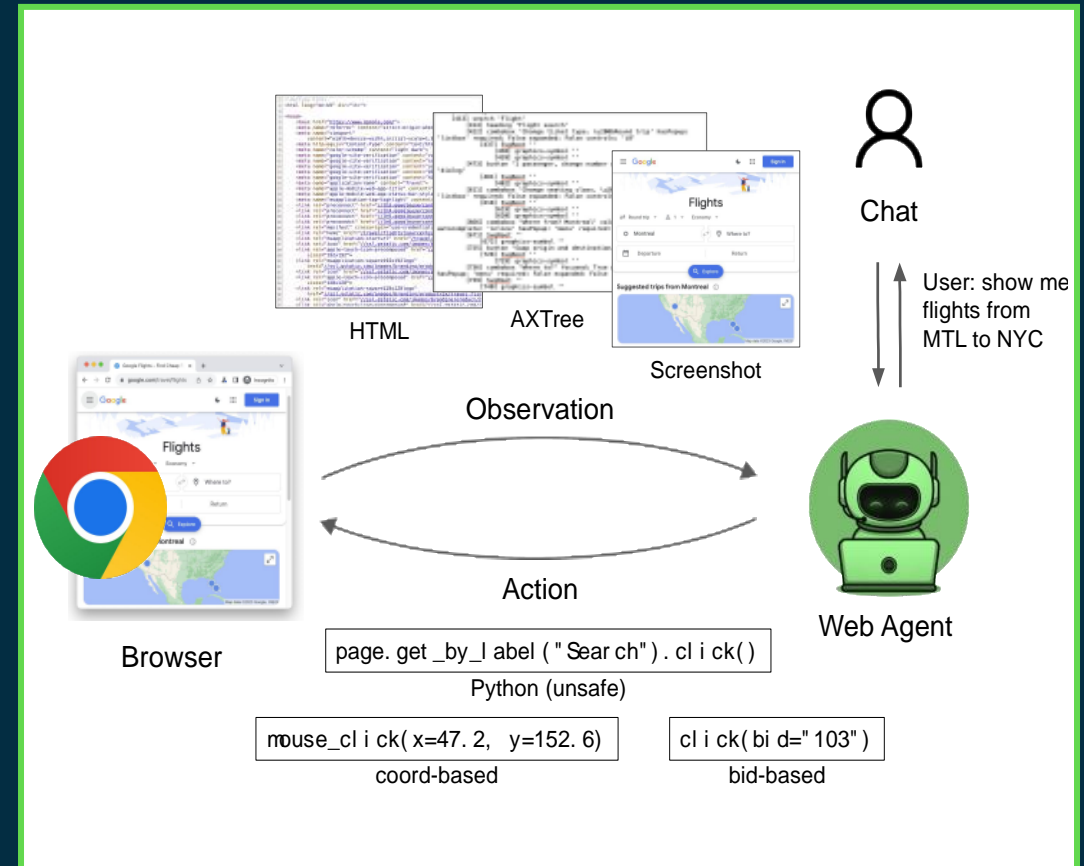
A unified evaluation platform

> Standard Observation Space

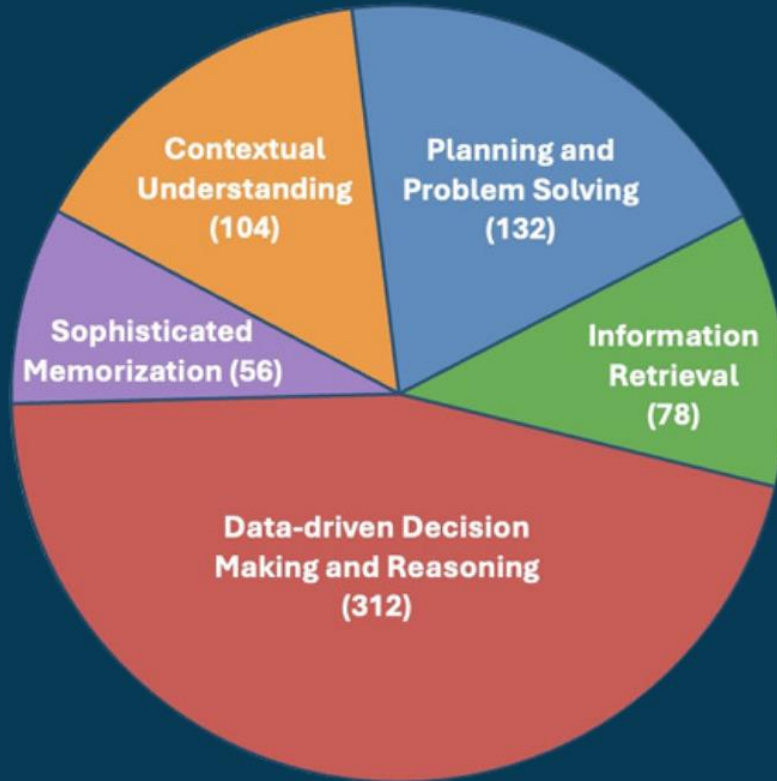
- HTML
- Screenshots
- Accessibility Tree
- And more

> Standard Action Space

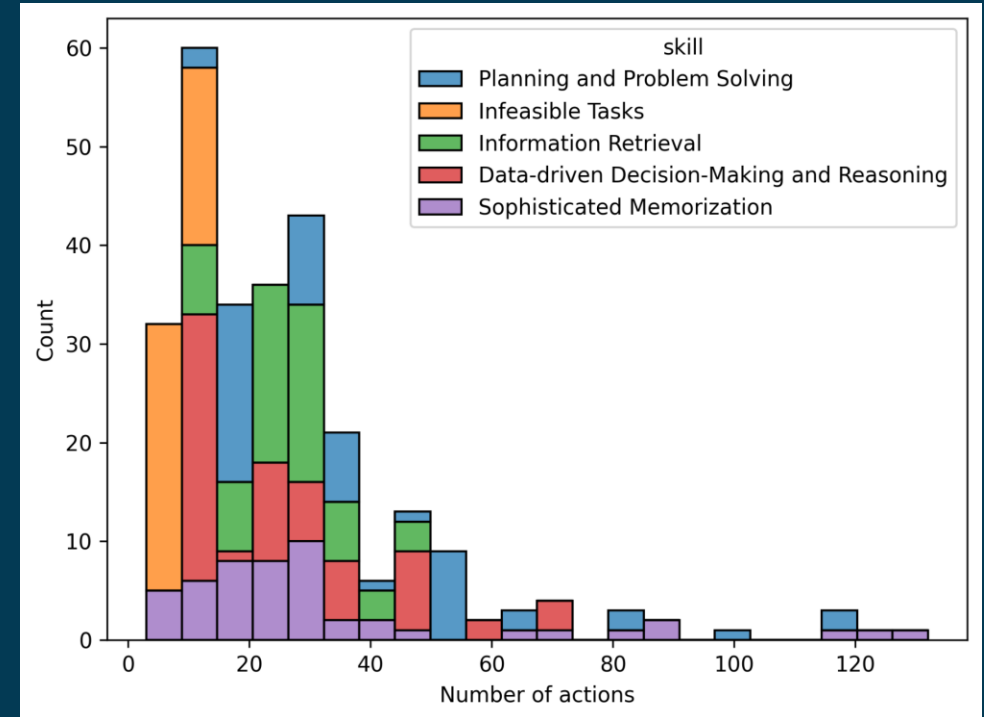
> Regroups all major benchmarks (thousands of realistic tasks)



Diverse Skills



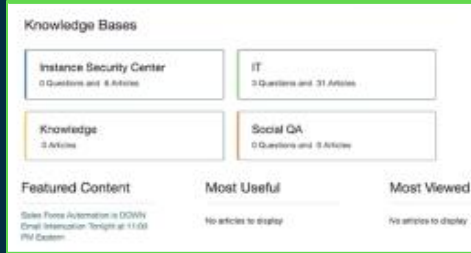
Number of Steps



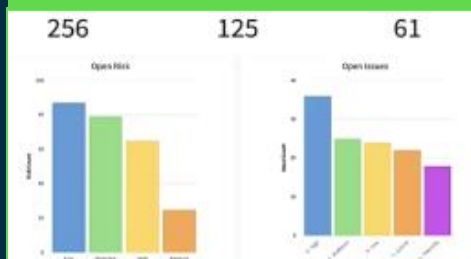
Benchmark for long-range tasks

WorkArena++ Towards realistic enterprise workflows

1 Knowledge base



2 Dashboard



3 Service Catalog



Example: The agent is assigned a ticket and instruction: "Please solve this."

The screenshot shows a 'Private Task' in ServiceNow. The ticket number is PTSK47711968, priority is 4 - Low, and state is Open. The owner and assigned to are Sandy Martinez. The description includes a task instruction: "Retrieve information from the chart with the title #CAT044377552 and perform the mentioned task. For calculations, please round off to the..." and a detailed protocol instruction: "Referring to the company protocol 'Dashboard Retrieve Information and Perform Task' (located in the 'Company Protocols' knowledge base), complete the dashboard retrieval task." The protocol includes steps like "Please retrieve the 'greatest' value of all the items in stock." and "Task: Place an order for the least available item in stock." The ticket also has buttons for 'Discuss', 'Follow', and 'Update'.

WorkArena: the benchmark is far from being solved

Realistic Workflows

Task Category (task count)	Agent Curriculum (full benchmark)					Human
	GPT-3.5	GPT-4o	GPT-4o-v	Llama3	Mixtral	
WorkArena L3 (235)						93.9 ±3.4
Contextual Understanding (32)						87.5 ±11.7
Data-driven Decision-Making (55)						100.0 ±0.0
Planning and Problem Solving (44)						87.5 ±11.7
Information Retrieval (56)						100.0 ±0.0
Sophisticated Memorization (48)						91.7 ±8.0
WorkArena L2 (235)						93.9 ±3.4
Contextual Understanding (32)						100.0 ±0.0
Data-driven Decision-Making (55)						84.6 ±10.0
Planning and Problem Solving (44)						100.0 ±0.0
Information Retrieval (56)						100.0 ±0.0
Sophisticated Memorization (48)						91.7 ±8.0
WorkArena L1 (33 × 10 seeds)						
MiniWoB (125 × 5 seeds)						
WebArena (812)						

Future Directions – Road to deployment

★ Security, robustness (e.g., hijacking, jailbreaking)

★ Long-term planning

- Large-scale data collection
- Dealing with sparse rewards
- LLM fine tuning strategies

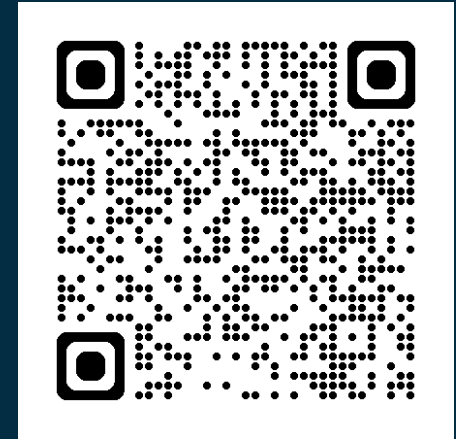
★ Cost and speed

- Shrinking huge observation spaces (RAG, etc.)
- Multi-agent architectures with specialized agents

★ Societal Impact of agents



Thank you!



Try our demo

[github.com/ServiceNow/](https://github.com/ServiceNow/AgentLab) **AgentLab**

[github.com/ServiceNow/](https://github.com/ServiceNow/BrowserGym) **BrowserGym**

[github.com/ServiceNow/](https://github.com/ServiceNow/WorkArena) **WorkArena**