What Web Agents Need Isn't More Data – It's High Quality Data

Authors: Judith Clymo, Karthik Raja Anandan, Shubham Gaur, Siddharth Suresh, Yuchia Chang

Mentor: Yilin Shen

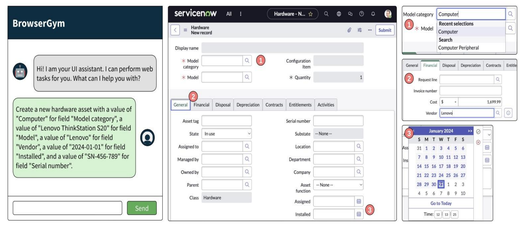
Background

Autonomous web agents accept natural language instructions and operate browser interfaces to find information and complete tasks.

The value of static training datasets is limited because

- Websites evolve rapidly so data becomes outdated.
- Enterprise systems are not publicly available.

Instead: **continually create custom training data**.



Above: Example **WorkArena** tasks as they are presented within **BrowserGym** agent environment. The agent is provided with a task description and a way to respond in natural language, observations that describe the current state of the webpage, and a set of possible actions to interact with the webpage. Tasks are divided by difficulty level and mimic typical tasks undertaken by knowledge workers.

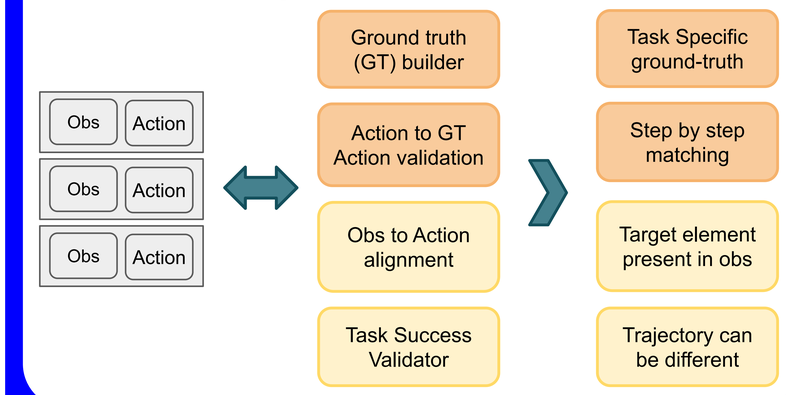
Our analysis of 33 **WorkArena** tasks shows **systematic failure modes** across task types:

Navigation: wrong page or menu selection
Forms: missing required fields, invalid entries
Filtering/Sorting: inability to detect or apply UI controls
Data Retrieval: incorrect chart/table value extraction
Impersonation: user search and selection failures

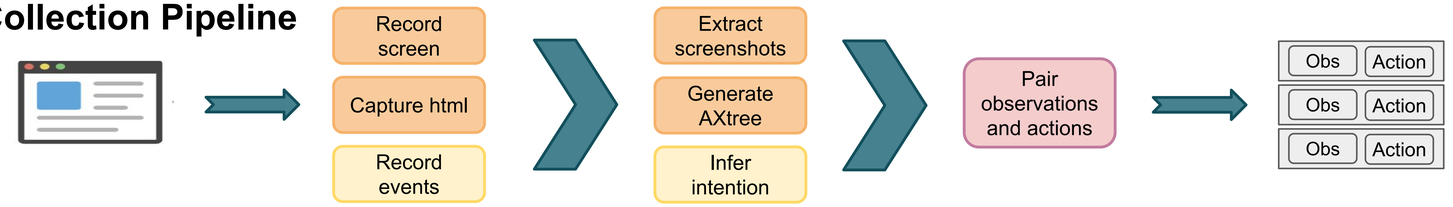
These errors highlight a lack of high-quality human demonstrations that teach agents how to interact with complex workplace UIs.

To address this, we present a customizable browser extension that captures real human interactions and converts them to **BrowserGym**-compatible training samples. We demonstrate its use by collecting data relevant to WorkArena tasks.

Verification Pipeline

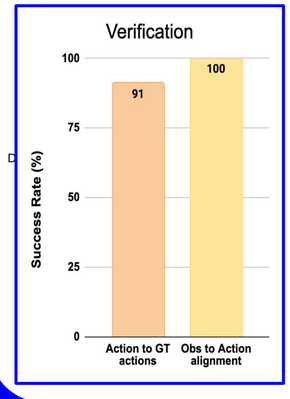


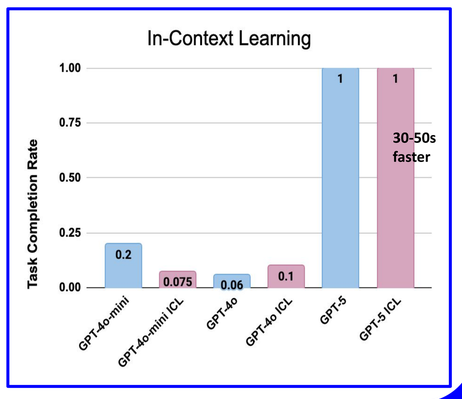
Data Collection Pipeline



Results

- 50 human trajectories collected for Work Arena - L1 tasks





Browser Extension

A chrome extension attaches to the user's browser to observe how they complete various tasks online. The extension collects observations of the environment and relevant javascript interactions.

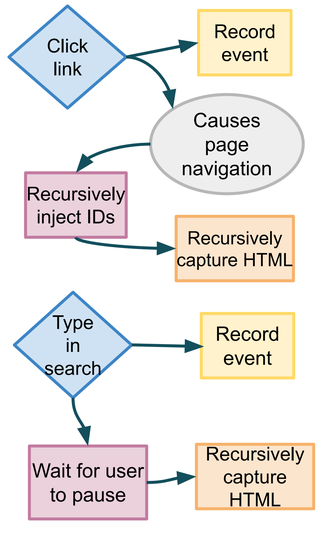
Identify relevant actions and changes
Javascript listeners respond to

- Javascript events (caused by user interactions)
- DOM changes in dynamic websites
- Page navigation.

Unambiguous event records
Unique IDs are injected into the webpage so events can be easily matched to their targets.

Full HTML and event capture
User-configurable set of events will be recorded - flexible to balance coverage with relevance. Recursive code injection and message passing ensures interactions within iframes are logged correctly.

Multimodal observations
Screen recording offers additional context. The currently active tab is continually captured.



Post Processing

We infer the users intention from the record of Javascript events.

Identify and remove unintended (not meaningful) events.

~~`mouseDown`~~ `click` ~~`mouseUp`~~

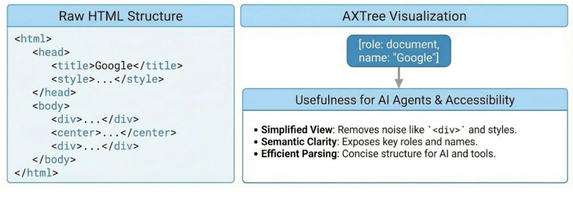
Combine related events

`input 'a' input 'ab' input 'abc' change 'abc'` → `input('abc')`

Map Javascript events to BrowserGym actions

`click (dropdown) click (item)` → `selectOption(item)`

Convert HTML to accessibility tree - easier for LLMs to parse



Match actions with immediately prior observations
each group becomes a training sample



Future Work

- Having proved that our tool can collect human browsing data and infer high quality training samples we would like to demonstrate the downstream benefit of this data on agent performance. Since we cannot access the large proprietary models we will use a range of open source models.
- We want to collect datasets for a range of benchmarks to show the flexibility and general applicability of our tool. This will also help us to further improve the data collection pipeline and find edge cases.

References

- **WorkArena: How Capable are Web Agents at Solving Common Knowledge Work Tasks?**, Alexandre Drouin, Maxime Gasse, Massimo Caccia, Issam H. Laradji, Manuel Del Verme, Tom Marty, David Vazquez, Nicolas Chapados, Alexandre Lacoste Proceedings of the 41st International Conference on Machine Learning, PMLR 235:11642-11662, 2024
- **WorkArena++: Towards Compositional Planning and Reasoning-based Common Knowledge Work Tasks**, Léo Boisvert · Megh Thakkar · Maxime Gasse · Massimo Caccia · Thibault de Chezelles · Quentin Cappart · Nicolas Chapados · Alexandre Lacoste · Alexandre Drouin, NeurIPS 2024 poster