WebArena-Pro: Extending web agent environments to more real-world websites

Introduction

WebArena is a benchmark for evaluating autonomous *web agents* on complex, multi-step tasks in realistic web environments. It provides a standalone, self-hostable web environment with fully functional websites across four domains (e-commerce, social forums, collaborative development, and content management), closely mimicking real-world websites. The environment is enriched with tools (e.g. maps) and knowledge bases (e.g. Wikipedia) to emulate human-like problem solving. Each environment is self-hosted inside a Docker container created by the authors, with data specifically populated for the tasks in the benchmark.

While humans succeed at these tasks nearly 80% of the time, recent modular designs and training techniques have improved success rates to ~60%. However, the tasks in WebArena are limited to 5 common websites, which does not cover the diversity and complexity of the real internet. The proposed project is to extend the WebArena benchmark to new real-world websites.

The project will fall under a collective initiative called **WebArena-Pro**, which we hope will eventually become the successor to the WebArena benchmark for evaluating web agents**.** Multiple teams can participate in this initiative, as long as each team has a separate project with different environments, tasks and agents.

Project Overview

Teams of **3 students** will work together to develop an extension of the WebArena, with new environments, tasks and agents. Each team will **select or create two web environments** in the same style as WebArena (which contains platforms like online shopping sites and a wiki or forum) and **define realistic user tasks** for each. These tasks should reflect plausible user goals (e.g. “find a product under $20 and add it to the cart” or “create a new wiki page and link it on the homepage”) that require multi-step interactions. Students will then **develop an autonomous agent** using the Agentlab framework to complete these tasks. This involves leveraging the provided tools: using Docker-based environment and WebArena’s JSON task specifications (e.g. *test.raw.json*) for configuration. The agent can utilize advanced NLP techniques – for instance, prompt-based planning or fine-tuned language models – to navigate the websites, interact with forms or buttons, and gather information to achieve each goal. Throughout development, teams will have access to **Agentlab, BrowserGym and WebArena**’s open-source resources (code, data, and a reproducible Docker environment are available), allowing them to easily run the simulated websites and test their agent’s performance in a controlled setting.

Timeline & Effort Breakdown

Plan to allocate effort in roughly a **2/3 – 1/3 split** between environment design and agent modeling:

- **Environment (2/3):** Building or customizing the two web environments and designing a suite of realistic tasks. This includes populating the environments with content/data and setting up the WebArena Docker instances.
- **Task design and modeling (1/3):** Developing the agent and performing evaluation. First, you will need to craft tasks, initial states and evaluation rules (JSON) to ensure meaningful evaluation. Then, you will need to implement the agent's decision-making (possibly iterating on prompts or model choices), debugging its interactions, and evaluating success on the defined tasks. Teams will run experiments to measure the agent's task completion rates and analyze failure cases.

Tools & Deliverables

Teams are expected to use **GitHub** for version control and collaboration on code. The final web environments (and possibly a demo of the agent) can be hosted on **Hugging Face Hub**. The culmination of WebArena-Pro will be an **6 to 8 page research-style report** (following a typical conference paper format). The exact sections are to be determined by the teaching team, but the report would likely include the standard sections – *Introduction*, *Related Work*, *Data/Environment*, *Methods (Agent Design)*, *Experiments*, *Results*, *Discussion*, and *Conclusion* – detailing the project findings. This write-up is an opportunity to discuss design decisions, present quantitative results (e.g. task success rates), and reflect on challenges and insights gained. All code should be submitted via the GitHub repository, and any custom datasets or environment files should be documented (with a link, if hosted on Hugging Face or similar).

Next Steps

Interested in WebArena-Pro? Please reach out and **meet with the mentor (Xing Han Lu)** to discuss your ideas and questions. This meeting will help ensure the project's scope is well-defined and that you have the resources needed to succeed. WebArena-Pro offers a chance to tackle a **novel, engaging NLP challenge**: building an agent that can intelligently browse and act on the web. The mentors look forward to guiding any team excited to pursue this project option. We will meet in a group and discuss any questions or concerns you may have.

If you are interested, please sign up here: https://forms.gle/2aRscTPHfu8b3n1r6