fellowship.ai

Demo and Status Report: Autonomous Web Agent

- 07.03.2025

Overview

Recent progress

- Transitioning from UI-Tars to Proxy-Lite
- Why we chose Proxy-lite and our progress so far
- Benchmarking

Why Transition from UI-Tars?

We decided that moving away from UI-Tars would be the better move since it proved to be too complex for our needs; while it is powerful, its level exceeds what are required from the project.

UI-Tars is able to navigate the web and also navigate through the OS that it is running on. Additionally, its 2B parameters model did not perform well when compared to the 7B parameters, the 7B parameters model although performed well, required end points and a GPU to run.

Why Proxy-Lite (3B)

Proxy-lite offers us a lighter and more efficient alternative when compared with UI-Tars and it better aligns with our project's final goal. While UI-Tars is better in the sense that it can navigate both the web and through the OS it would require endpoints and a GPU to run, Proxy lite offers us what we need which is web interactions and its ability to run on a CPU. This makes it more focused and less resource intensive; meaning it performs the required tasks without any unnecessary complexity.

Progress with Proxy-Lite

We are currently working on integrating the 3B parameters model into the agents built by Iremide and Rahul.

This process would still involve fine-tuning the model so that it would align with the agents' and the project's requirements so that it would be able to better handle tasks and complete them with high accuracy.

It also seems like it would be receptive to fine-tuning which would mean it would perform better for our task.

Errors we ran into

While integrating the model into the agents we ran into an “Image-Processor” error. We have noticed that this error mainly occurs when we try loading Proxy-lite as a VLM where it fails to process images correctly.

However, when it is used for text generation the model function as expected without running into any errors.

The next step from here would be to dive deep into the error and try to resolve it by next week.

Benchmarking

When it comes to benchmarking we focused on comparing Proxy-lite with the previous Cohort's agent. To evaluate both of them we decided to use the WebVoyager dataset which consists of 15 websites for each one there are ~40 prompts/tasks. The proxy-lite model has already been benchmarked on this dataset by its creators and it achieved strong results (they measured the success rate, the finish rate and the avg. number of steps, the agent achieved an overall success rate of 72.4%, the exact results can be found here in the evaluation section).

Benchmarking

When testing and benchmarking the previous Cohort’s agent, we ran it through the first ten tasks on Allrecipes. However, due to the extended time it took to complete these tasks, we decided to stop testing after the initial ten. Out of these, the agent was able to finish two tasks, and of those, only one was successfully completed as intended. The results highlighted inefficiencies in both time and success rate, leading us to discontinue further testing at that point.

Goals for next meeting

- Successfully integrate the 3b parameter Proxy-lite into the agents.