# Demo and Status Report: Autonomous Web Agent

- 05.05.2023

# Overview

## Recent progress

- Made progress with UI-Tars and explored midscene.js
- Worked on fine tuning

## Goals for next week:

- Decide on a model and keep improving it

# Progress

At the start of this week Feras suggested that we work on Midscene.js and experiment with it as it is simpler than UI-Tars, and in reply we divided ourselves into three groups

1. Experiment with Midscene.js
2. Getting UI-Tars to run
3. Fine tuning

# Midscene.js

Midscene, which was suggested to us by Feras at the beginning of the week, seems like what we’ll use from now on to run our webagent.

At first when we started using it we ran into errors like “401 unauthorized error” and “503 service error”. Once we got openAI’s credits we were able to run Midscene on the chrome extension which allowed us to make solid progress in the testing phase (the credits helped with the  “401 unauthorized error”). The “503 service error” was caused by inactivity; the servers shut down after 15 minutes of inactivity to save resources.

# Midscene.js

Although we have made a lot of progress there is still that still needs to be made. The model currently is not capable of executing actions, it only recommends it, so for next week we need to figure out how to make it execute actions rather than suggesting them.

# UI-Tars Desktop

The team working on UI-Tars Desktop used Google Colab GPUs and following the steps of a youtube tutorial which was found on Monday. Many of the errors we were facing previously were resolved such as using an API key rather than ngrok Authtoken, after resolving this the connection between Google Colab and UI-Tars desktop interface seems to be working fine.

However, while the connection is established the Desktop agent does not respond to prompts, while google colab returns a “GET / HTTP/1.1 404 Not Found error” which might be because of an incorrect ngrok tunnel configuration or that the model was updated since the video was uploaded.

# Fine tuning progress

We have also made significant progress in Fine tuning this week, primarily we have categorized our approach into two types DPO (direct Preference Optimization) and a SFT (Supervised fine tuning), each one has four subcategories under them:  QnA, Reasoning, Summarizing, and Math.

QnA -> SQuAD

Reasoning -> Hellswag

Math -> GSM8K

Summarization ->CNN_Dailymail

# Fine tuning

Although the model seems to answer a lot of the questions correctly, it still runs into some errors such as it falling into a loop or answering incorrectly. These issues aren't a major concern, as they are expected since the complexity of working with both visual and textual data in a Vision-Language Model (VLM), as opposed to a purely Language Model (LLM). The primary focus of a VLM is integrating and interpreting multimodal inputs, so occasional errors related to these aspects are expected.

# Goals for next meeting

1. Decide on a model and keep improving it.