MultiOn Interview Project

Goal: Create an end-to-end AI application that takes in an input image screenshot and translates that to browser actions to complete a task using the MultiOn API.

Requirements:

- Img to text generation for the correct commands to pass to MultiOn using a visual language model (VLMs) such as Instruct-BLIP, Qwen-VL, etc. with in-context learning and prompting
- Convert commands to browser actions using MultiOn API
- Verify that the task was completed correctly and terminate, if there were any errors or issues successfully trigger retries.
- Should work end to end successfully for variety of practical use cases (min 3)
- Use of a library like Langchain or LlamaIndex is optional but might be useful and both integrate with MultiOn API

Bonus:

- Compare various SOTA VLMs and benchmark them on different use cases with pros/cons of each for end-to-end accuracies
- Use chaining, reflexion and critique feedback approaches to improve performance
- Use fine-tuning to improve the performance of VLM on a specific task
- Build an iOS mobile App that can take as an input and send it to the FastAPI server
- Build tenacity and multiple retry mechanism to handle failures when Agent does something wrong
- Go crazy, the goal is to impress us to your best ability! Sky is just the lower limit

Deliverables:

- Working FastAPI backend that takes in an image as input and triggers MultiOn Chrome Extension Client to complete the task using the API
- Jupyter notebook or Google Collab showing how to use the App and ability to send the screenshots to FastAPI backend
- (Optional but preferred) Deploy the FastAPI server on the cloud and have a live working link that can be tested with Google Colab/jupyter notebook
- Writeup detailing approach, what worked and not, and future improvements
- Code with proper documentation, readme, good code quality and testing
- Videos showcasing successful working on at least three different use cases
- End-to-end success and other accuracy metrics on different use cases

Getting started:

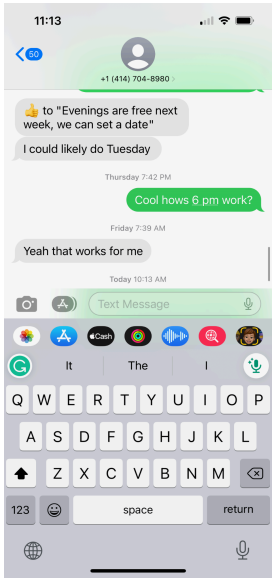
- Colab with example getting started code: https://colab.research.google.com/drive/1xamTVo4NVK_IjQ9Olt5i5FIcrytLWaDe?usp=sharing
- Youtube demo on the expected working: https://www.youtube.com/watch?v=wkWVnjv8dxo&list=PLz9R-6Sz08pA-NEU7G0a8NFVphZpc0wFX&index=4

Timeline:

- We have tested and found that the project can be fully finished in **less than a day or 24 hours**
- Nevertheless, what matters is the final quality of the deliverable and showing a well thought methodology, proper working and task completion accuracy
- You have **three days to submit the project from when you start** (please send an email to hiring@multion.ai when you start the project to start tracking)

Examples:

- Schedule meetings based on screenshots



Example flow:

image -> VLM -> generate command “book a meeting on Tuesday at 6pm using google calendar and invite …” -> send cmd to MultiOn API -> take actions -> verify task is completed successfully -> tell user that the meeting is booked

- Order a food dish based on a screenshot

- Buying clothing / furniture based on a photo