Computer Use

CS6960 MultiModal LLM Agents

Kenneth Marino

Slide / Figure Credits: Farhan Ishmam, Ana Marasović

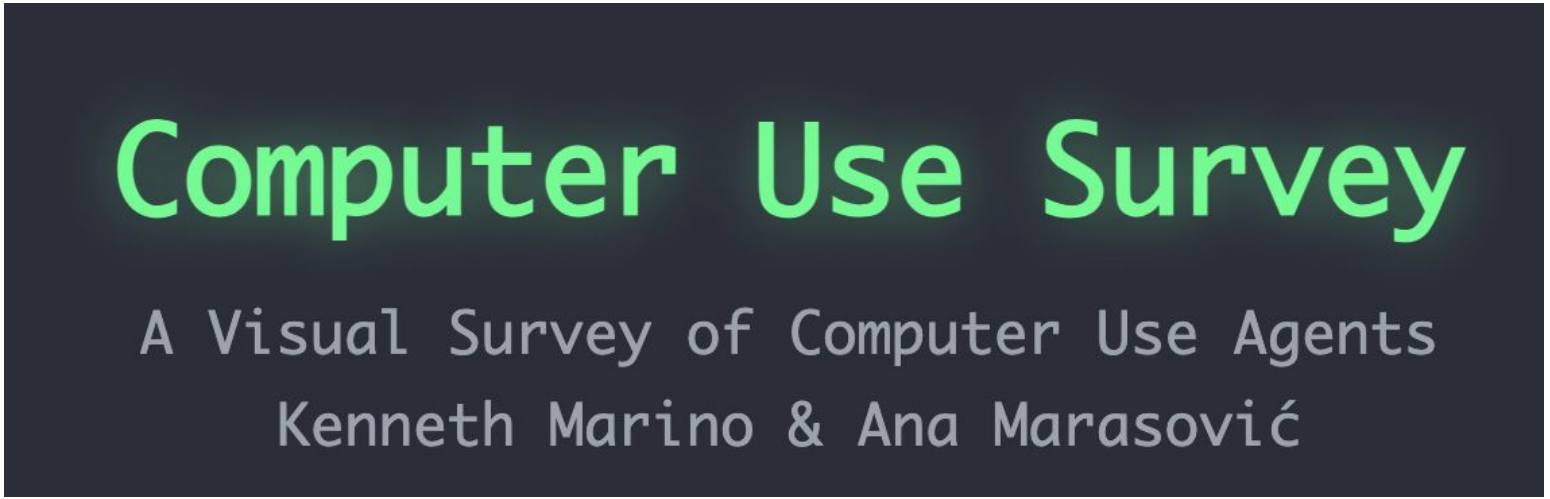
Announcements

- Projects
  - Working on Milestone 1 feedback
  - Don’t forget about Milestone 2, keep making progress on projects

Any Questions

What Is a “Computer Use” Agent

What Is a “Computer Use” Agent



https://kennethmarino.com/computeruse/computeruse.html

What Is a “Computer Use” Agent

Computer Use Survey - A Visual Survey of Computer Use Agents

In recent years, AI systems operating on the web and in computer environments have become a major topic of interest for both academia and industry. The goal of this blog is to provide an interesting and interactive survey of historical and recent works on computer use agents. We define key terms used in the literature, catalogue the expansive list of environments and datasets, discuss the evolution of the methodologies, and assess both today’s landscape and possible paths forward.

https://iclr-blogposts.github.io/2026/blog/2026/web-agent/

What Is a “Computer Use” Agent

- “By this we mean an agent that interacts dynamically with a computer system or web interface.”

History of Computer Use Agents

- We can essentially start with Web Automation

Automating Web navigation with the WebVCR

Vinod Anupam [1], Juliana Freire [*,1], Bharat Kumar [1], Daniel Lieuwen [1]

Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974, USA

Record and replay user actions

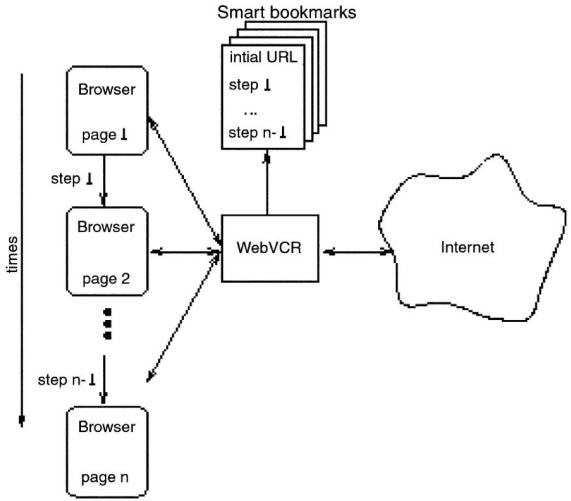


Fig. 6. Recording smart bookmarks.

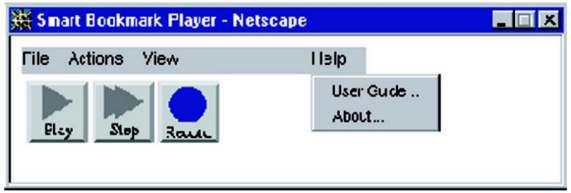


Fig. 7. Screenshot of WebVCR applet when first started up.

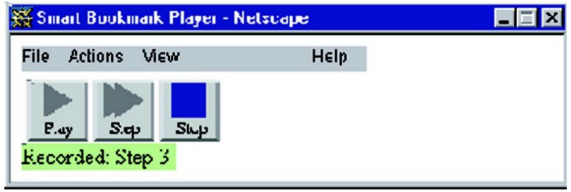


Fig. 8. Screenshot of WebVCR applet while recording steps.

Planning Era

- Some learning
- Some generalization

2000s: Classical Planning Era

The earliest works used classical planning techniques and cognitive architectures. Even from this early stage the problem was framed under an agent-environment paradigm [38].

2000	The User Interface as an Agent Environment	[36]
2002	Toward Automated Exploration of Interactive Systems	[37]

Key Technologies: STRIPS, Classical Planning, Cognitive Architectures, Symbolic AI

RL-Era

- Treat as an RL problem, train from scratch

2010s-2017: Early RL Revolution

Reinforcement Learning became the dominant paradigm. This line of work took automatic, self-contained reward environments and trained models from scratch to predict optimal actions.

2016	End-to-End Goal-Driven Web Navigation	[39]
2017	World of Bits: An Open-Domain Platform for Web-Based Agents	[18]
2017	AndroidEnv: A Reinforcement Learning Platform for Android	[40]

Key Technologies: Deep Q-Networks, Policy Gradients, OpenAI Universe, MiniWoB

More Recent

- Starting to treat this as an LLM agent (although the term didn't necessarily exist yet)

2020-2022: Behavior Cloning & Scale

Focus shifted to collecting human traces and using behavior cloning. Large-scale datasets emerged, and the field prepared for the LLM revolution that would follow.

2021	WebGPT: Browser-assisted Question-answering with Human Feedback	[24]
2022	A Data-Driven Approach for Learning to Control Computers	[43]
2022	Learning to Navigate Wikipedia by Taking Random Walks	[23]

Key Technologies: Human Feedback, Behavior Cloning, Large-Scale Datasets, Text-only Environments

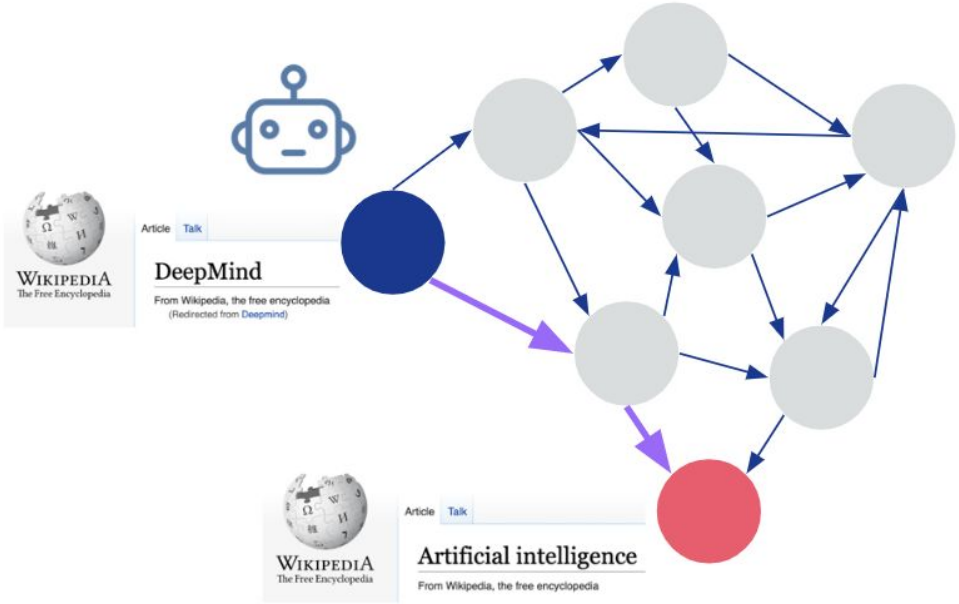
Computer Use Environments

Computer Use Environments

Computer Use Environments

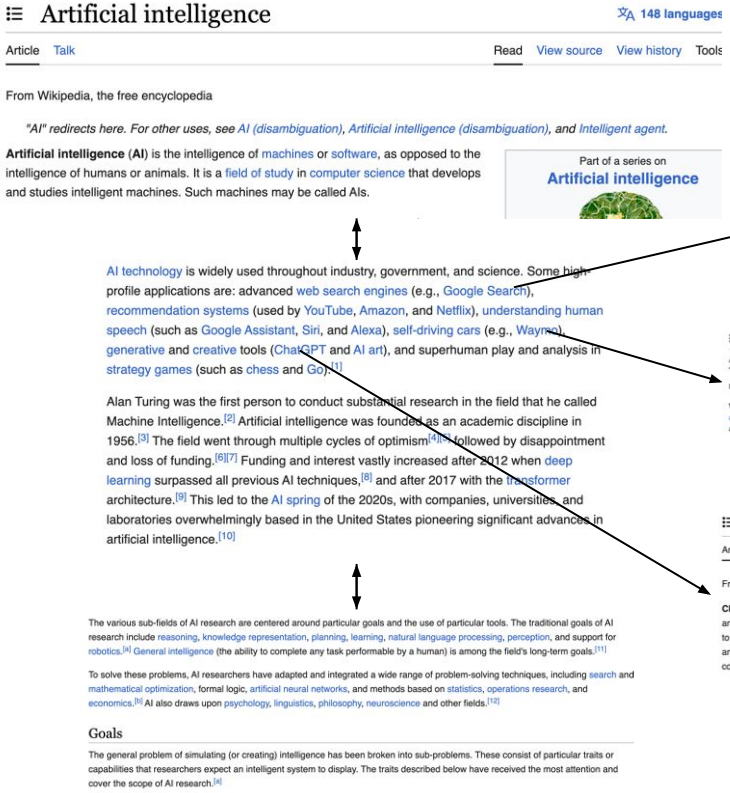
- Text-only Web Envs

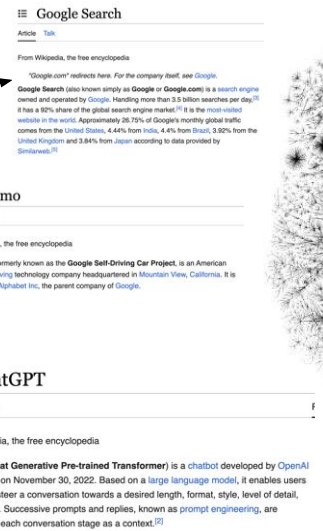
Computer Use Environments

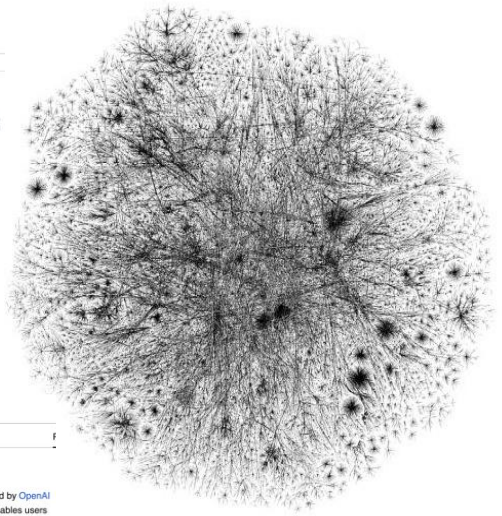


Manzil Zaheer*, **Kenneth Marino***, Will Grathwohl*, John Schultz*, Wendy Shang, Sheila Babayan, Arun Ajua, Ishita Dasgupta, Christine Kaeser-Chen, Rob Fergus. “Learning to Navigate Wikipedia by Taking Random Walks" *NeurIPS* 2022.

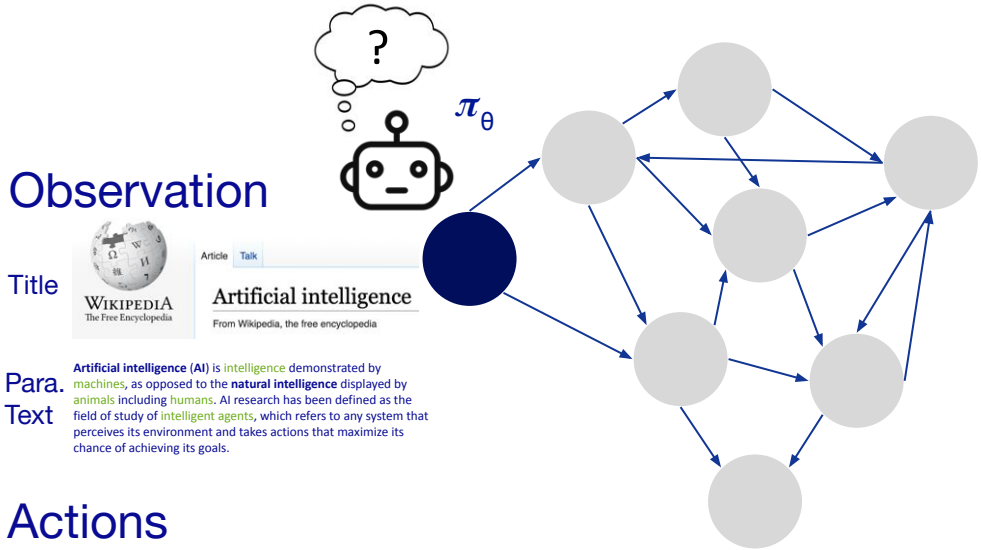
Wikipedia text/links as a graph env







Learn to Navigate Wikipedia Links

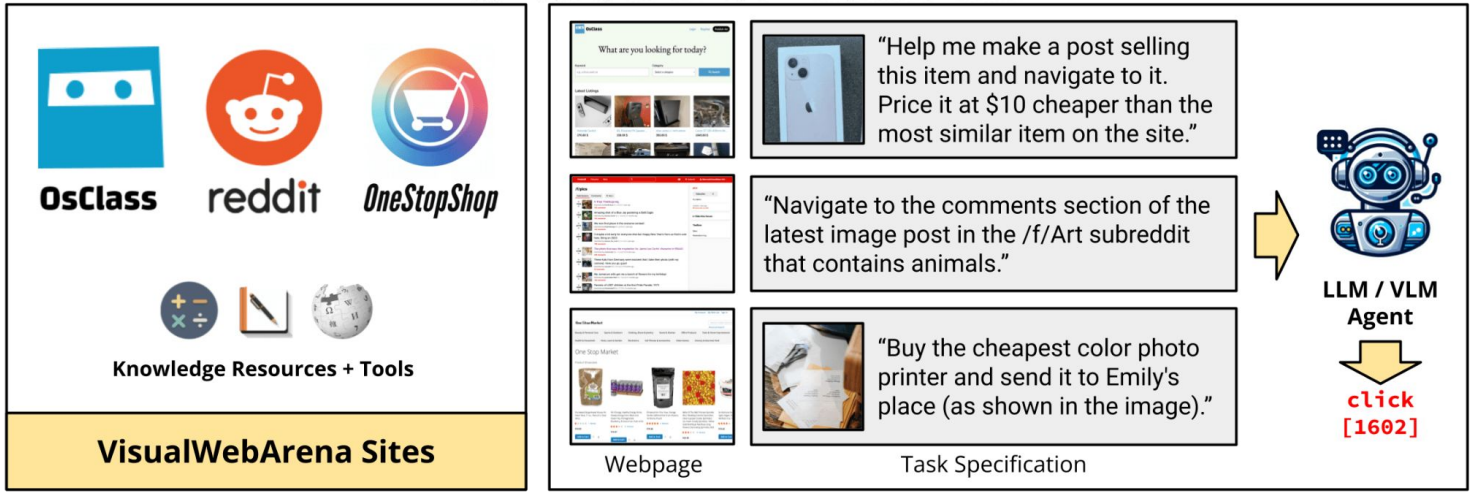


Computer Use Environments

- Text-only Web Envs
- Full HTML web

VisualWebArena: Evaluating Multimodal Agents on Realistic Visually Grounded Web Tasks

Jing Yu Koh **Robert Lo*** **Lawrence Jang*** **Vikram Duvvur***
Ming Chong Lim* **Po-Yu Huang*** **Graham Neubig** **Shuyan Zhou**
Ruslan Salakhutdinov **Daniel Fried**
Carnegie Mellon University
{jingyuk,rsalakhu,dfried}@cs.cmu.edu

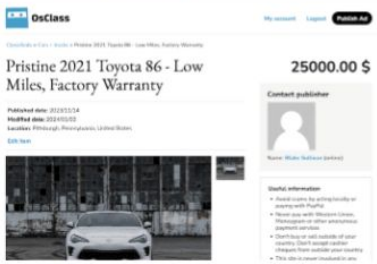


Visua ally

What is the ISIN of the company that occupies the largest portion in Warren Buffet's portfolio? Answer using the information from the Wikipedia site in the second tab.

Add something like what the man is wearing to my wish list.

Create a post for each of the following images in the most related forums.



Navigate to my listing of the white car and change the price to $25000. Update the price in the description as well.

OsClass

redc

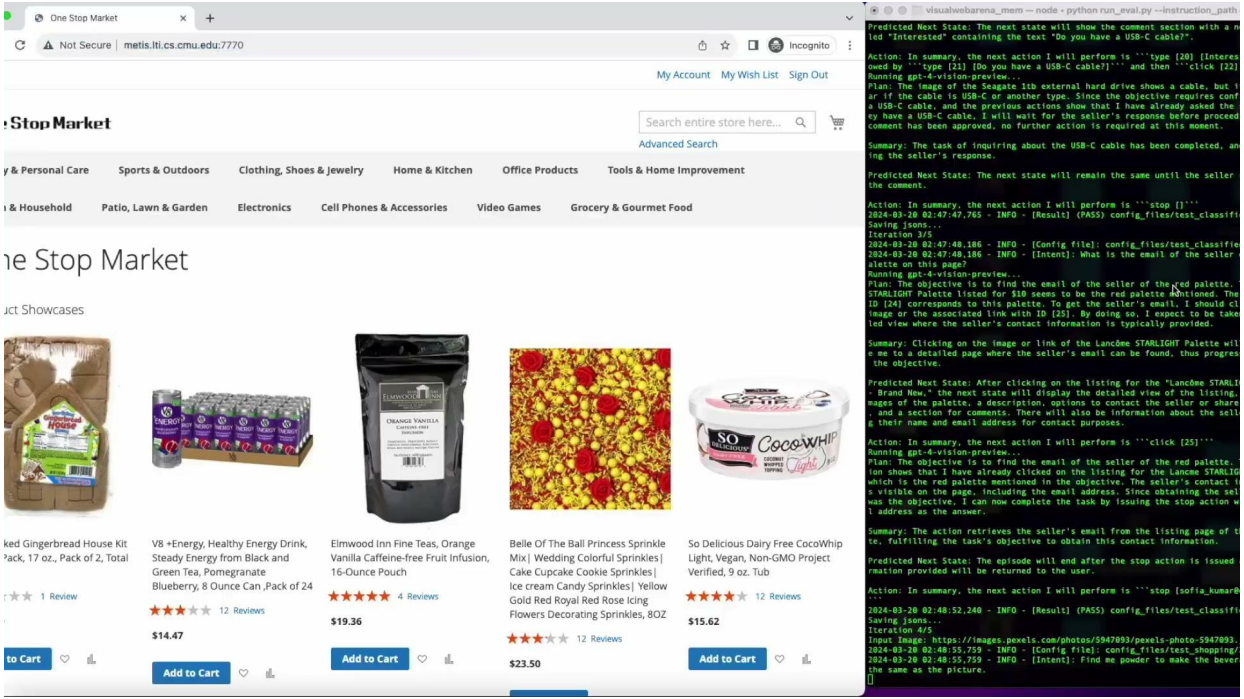
Knowledge Resou

VisualWebAr

LLM / VLM Agent

click [1602]

Led by Gabriel Sarch - now Princeton Postdoc



Gabriel Sarch, Lawrence Jang, Michael Tarr, William W Cohen, Kenneth Marino, Katerina Fragkiadaki. “VLM agents generate their own memories: Distilling experience into embodied programs of thought." *NeurIPS* 2024.

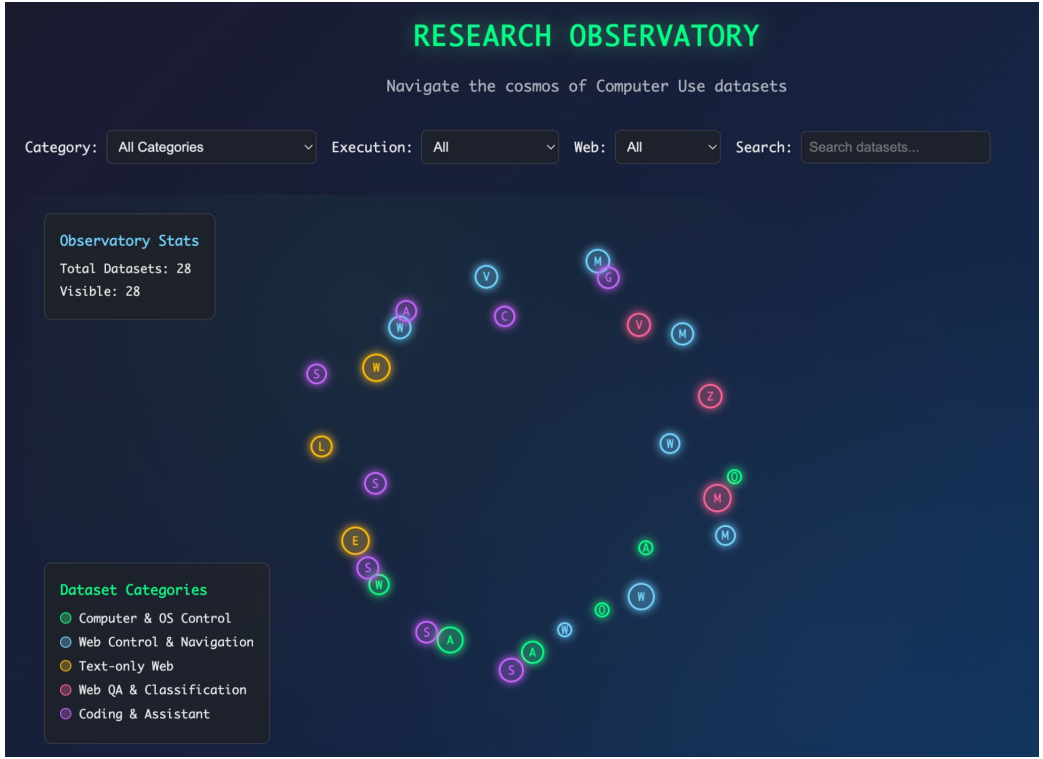
Computer Use Environments

- Text-only Web Envs
- Full HTML web
- Specific tools (code)

Computer Use Environments

- Text-only Web Envs
- Full HTML web
- Specific tools (code)

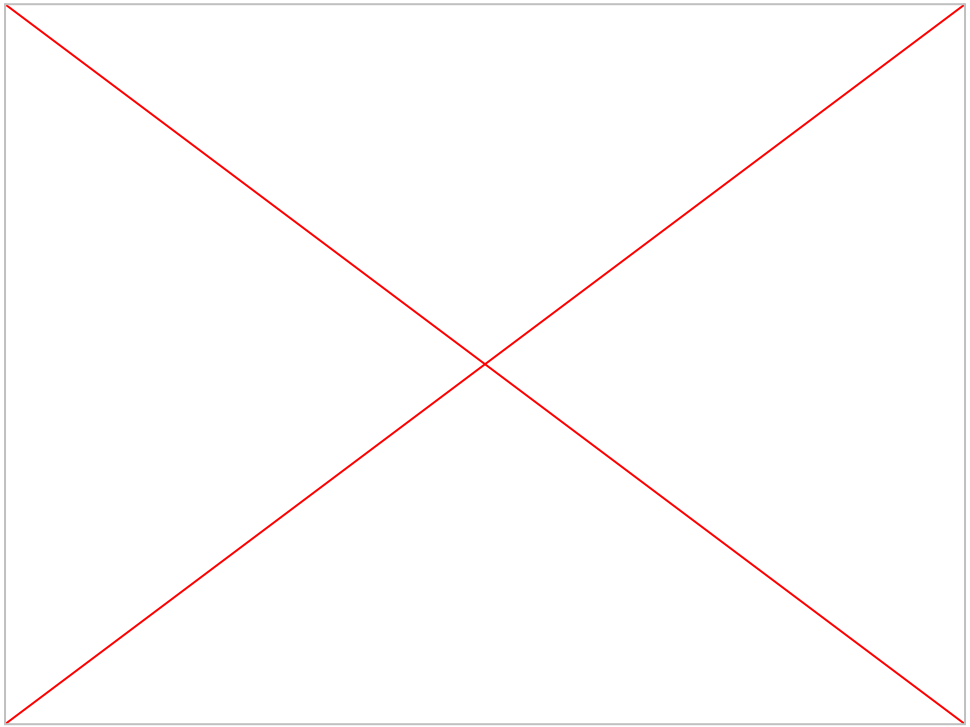
Is an agent that navigates a GitHub codebase and adds files a coding agent or a Computer Use Agent?



Computer Use Environments

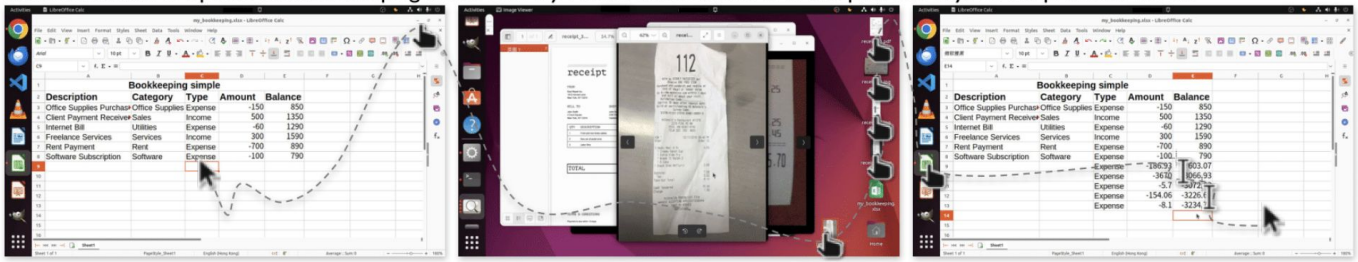
- Text-only Web Envs
- Full HTML web
- Specific tools (code)
- Full OS Simulation

OSWorld

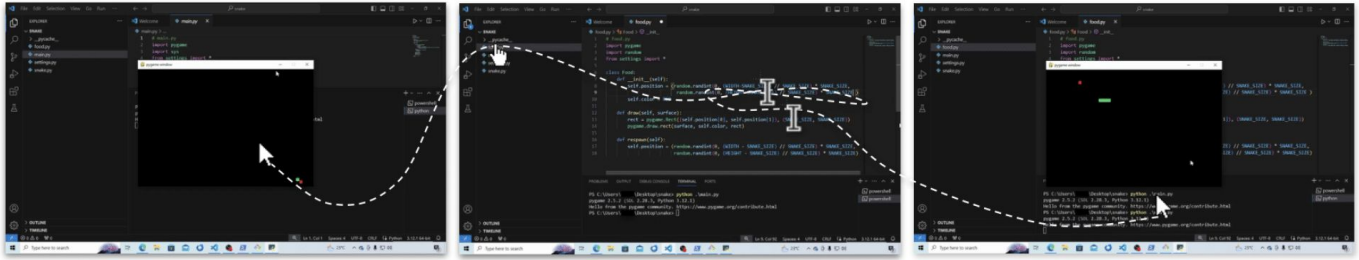


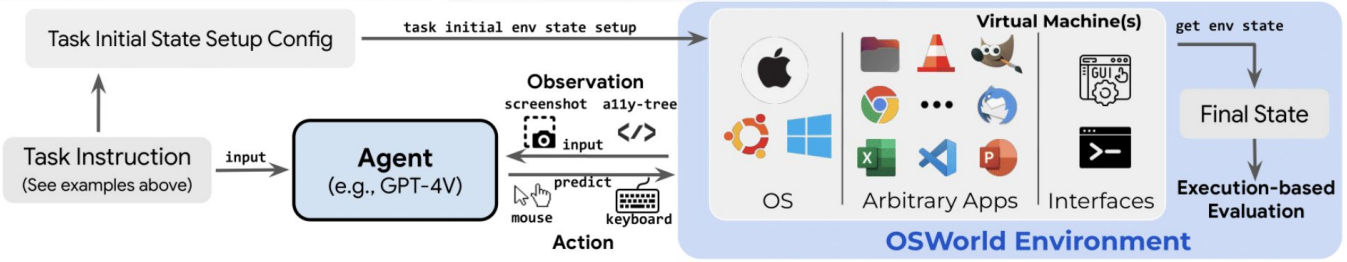
OSWorld

Task instruction 1: Update the bookkeeping sheet with my recent transactions over the past few days in the provided folder.



Task instruction 2: ...some details about snake game omitted… Could you help me tweak the code so the snake can actually eat the food?





Computer Use Agents

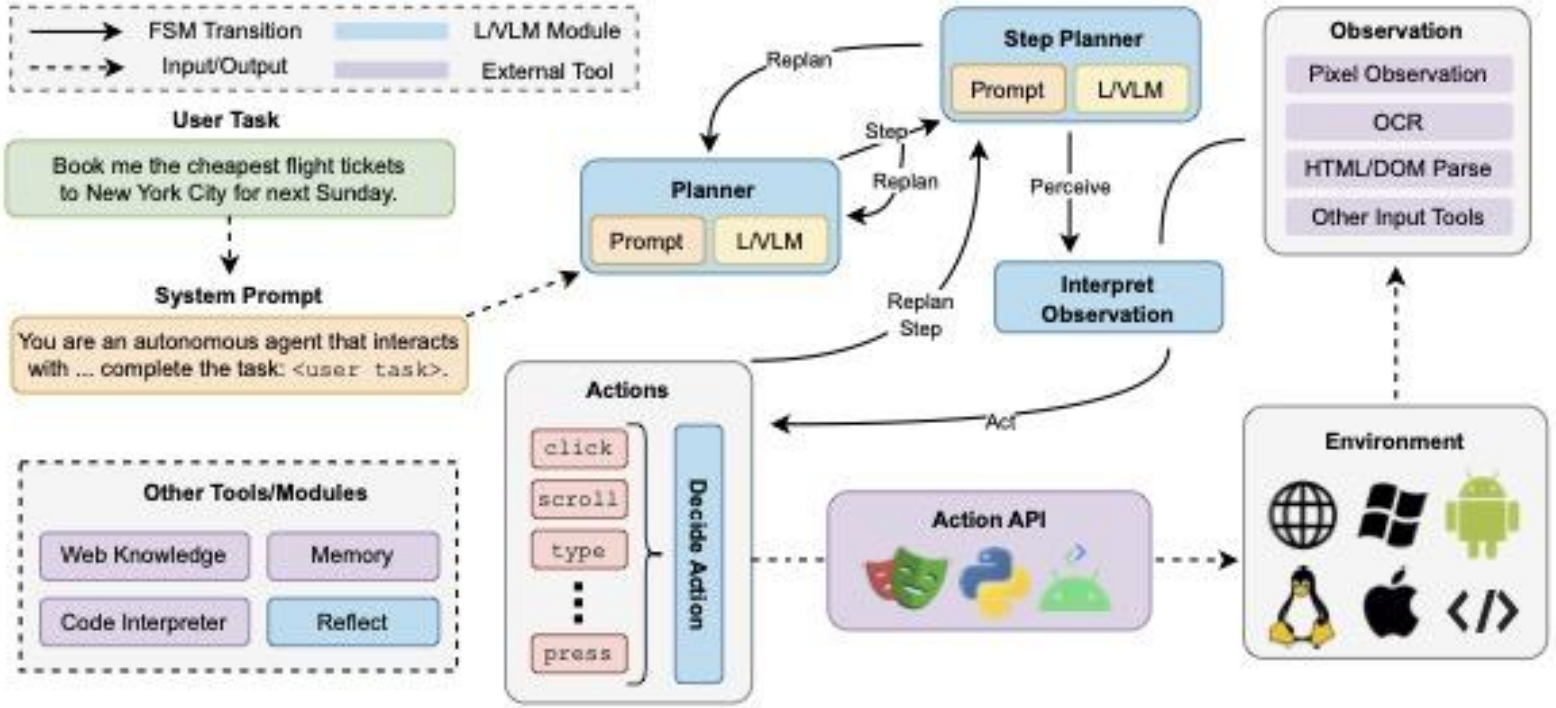


Image credit: Farhan Ishmam

Computer Use Agents

Hand-crafted observations useful for VLM

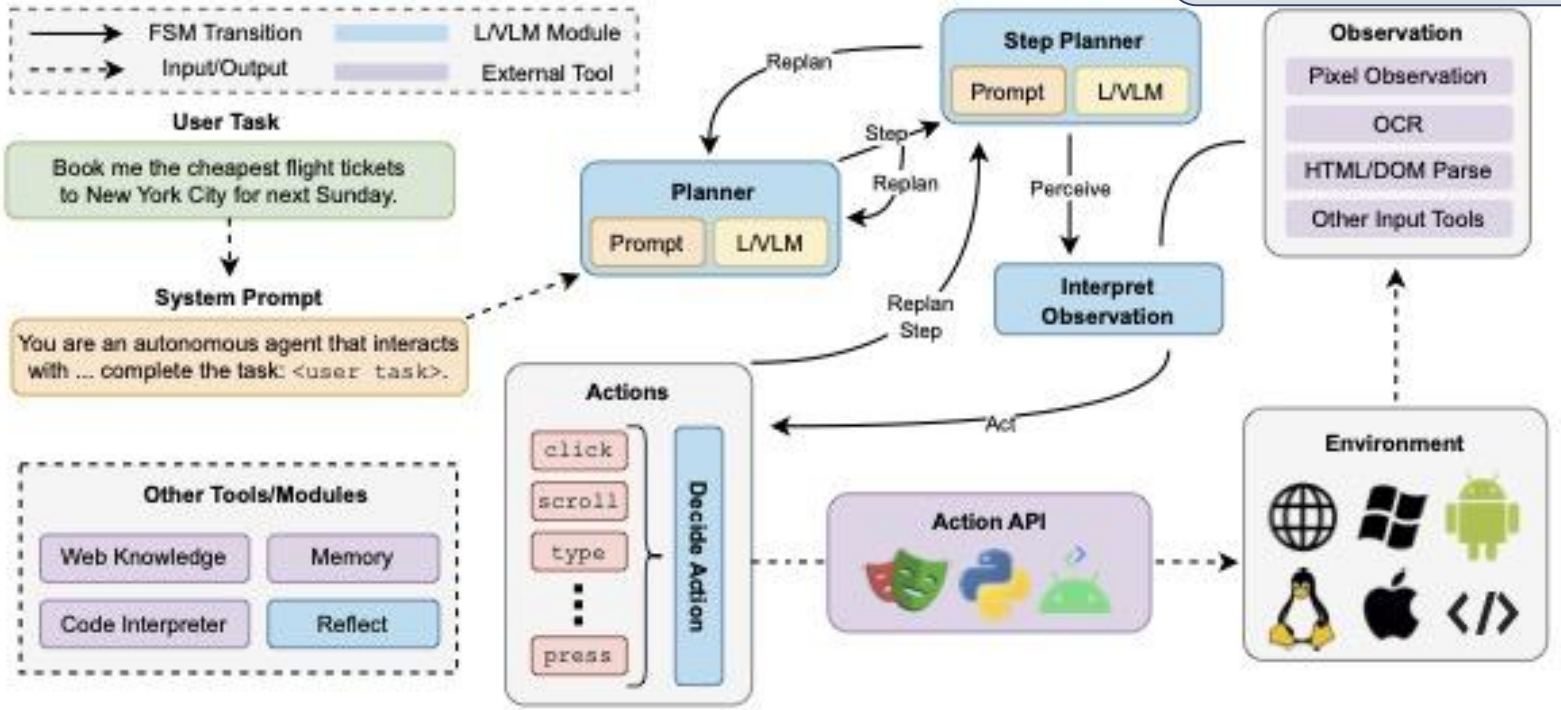
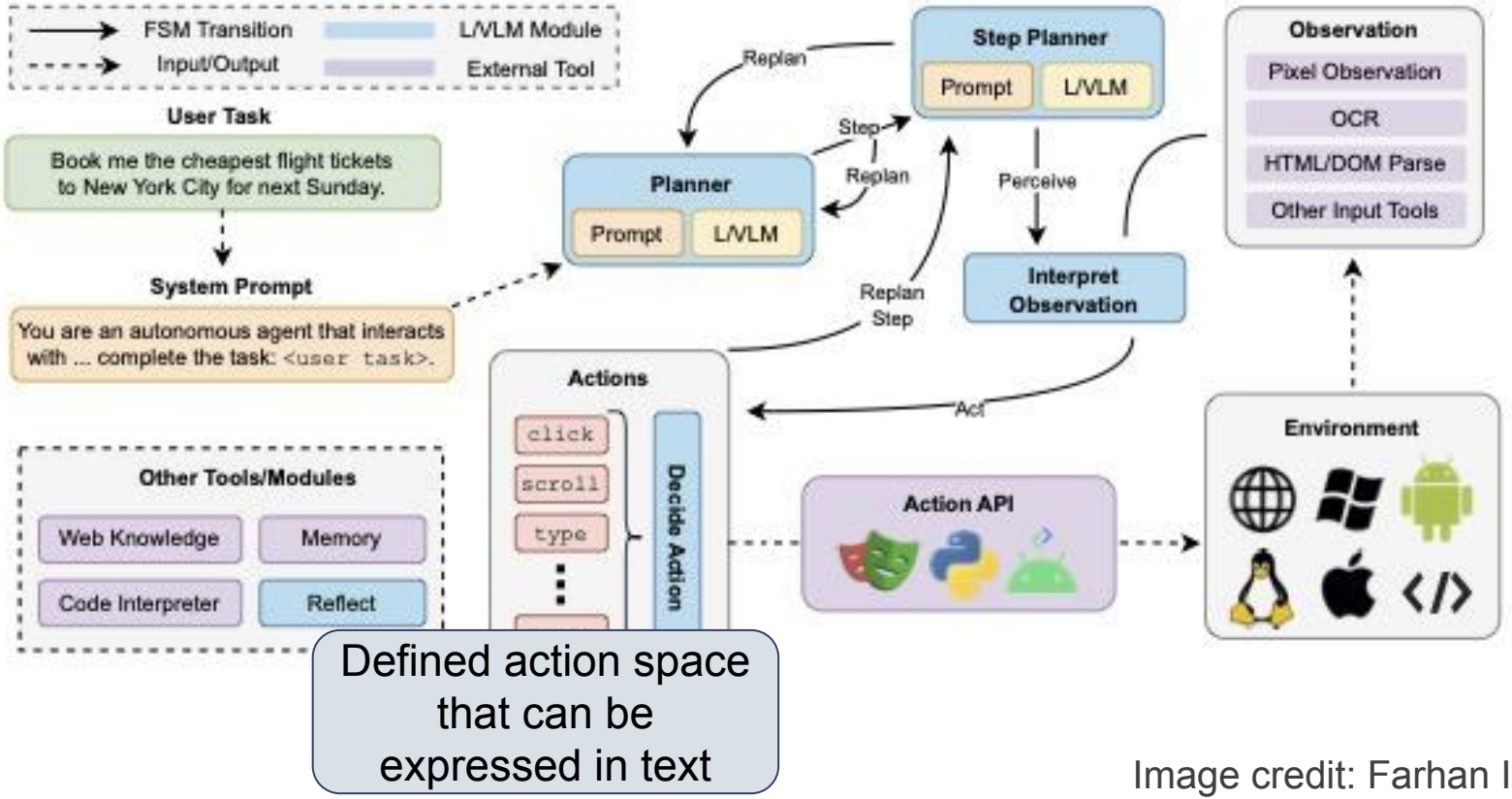


Image credit: Farhan Ishmam

Computer Use Agents



Defined action space that can be expressed in text

Image credit: Farhan Ishmam

Computer Use Agents

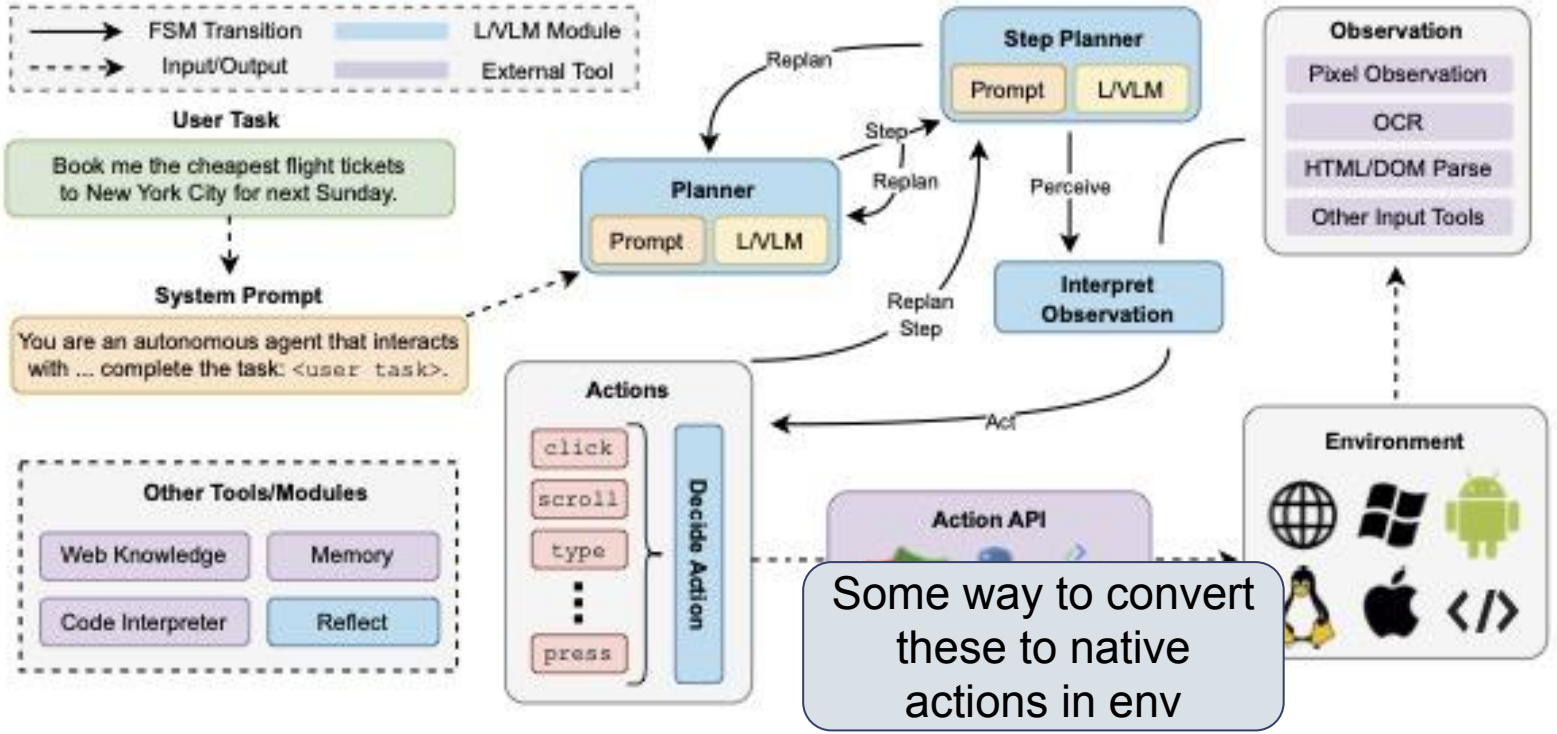
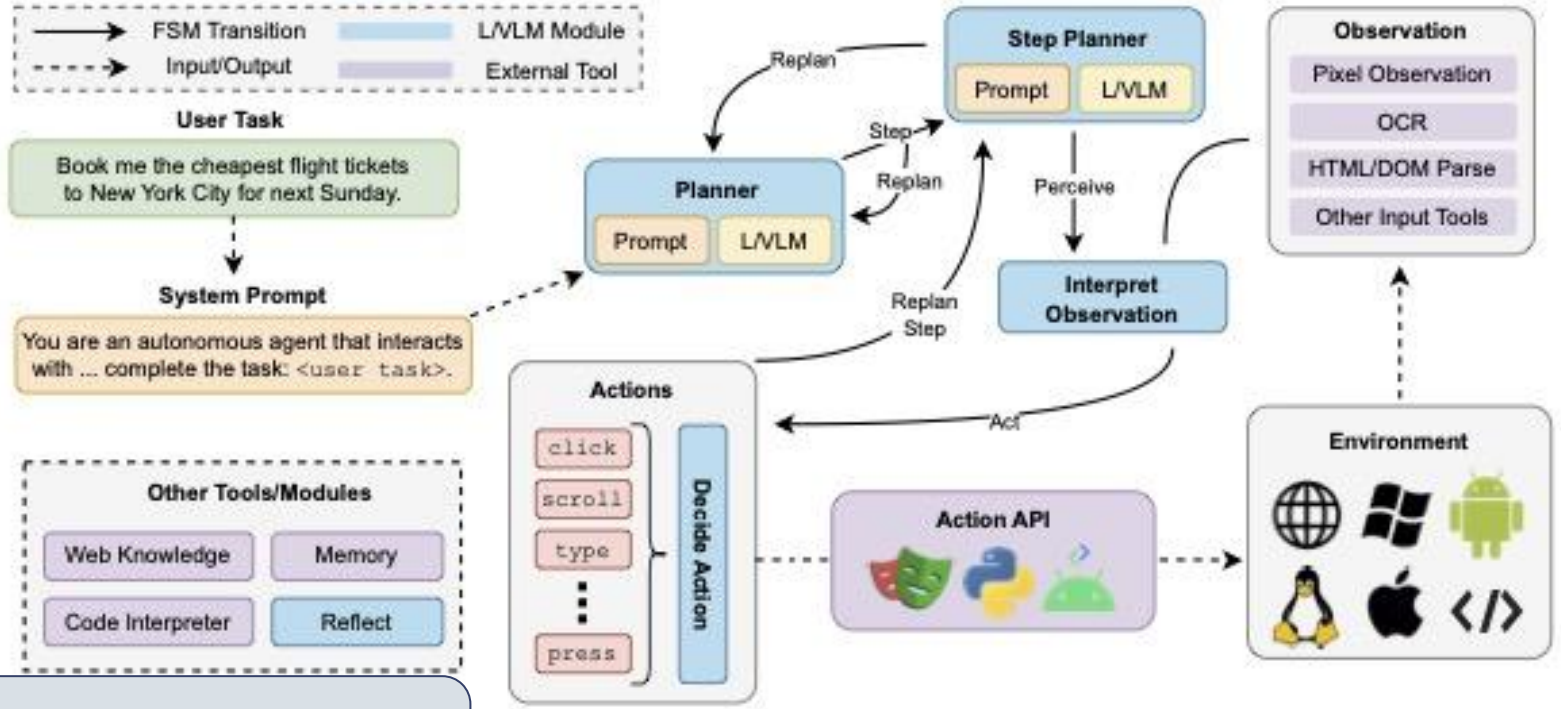


Image credit: Farhan Ishmam

Computer Use Agents



Tools VLM can invoke that do different things

Image credit: Farhan Ishmam

Computer Use Agents

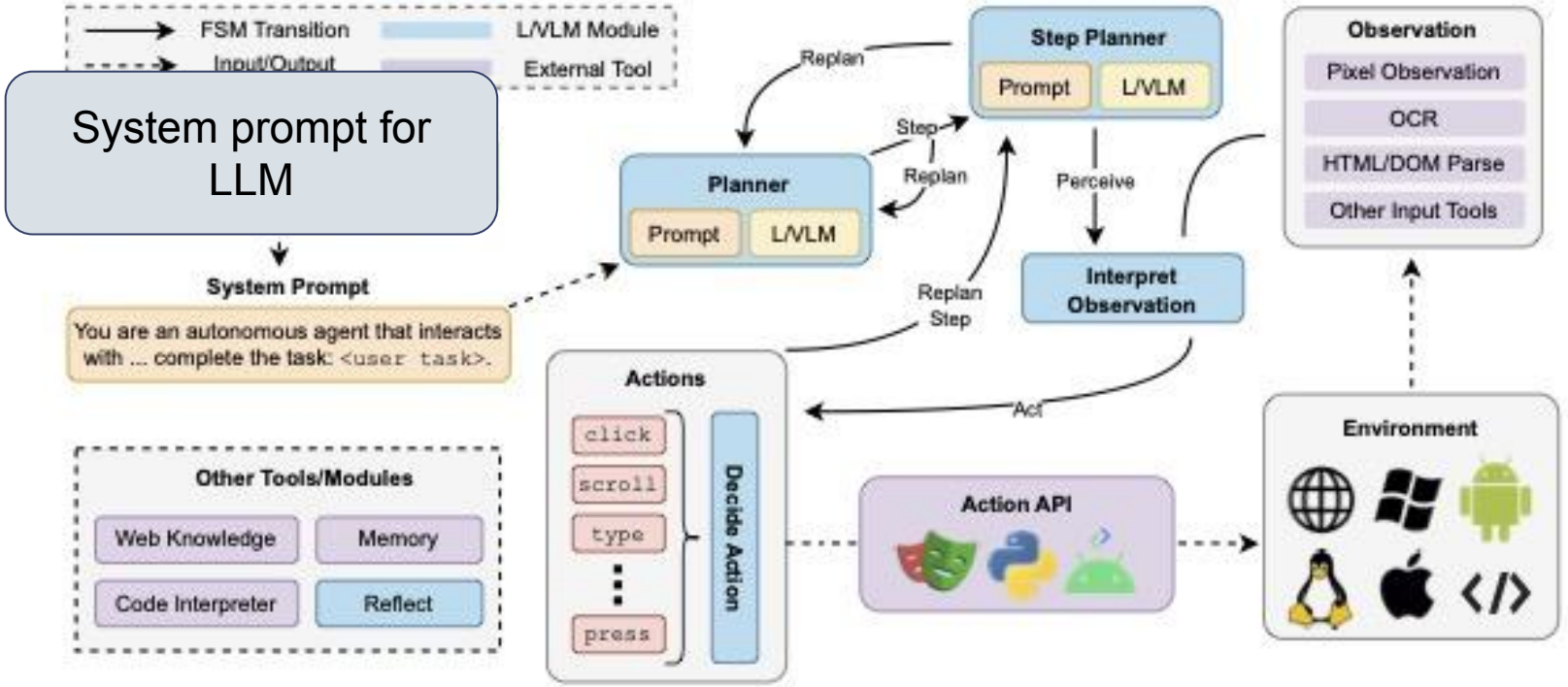


Image credit: Farhan Ishmam

Computer Use Ag

Frameworks for LLMs to be better at actions (reflect, plan, replan, etc)

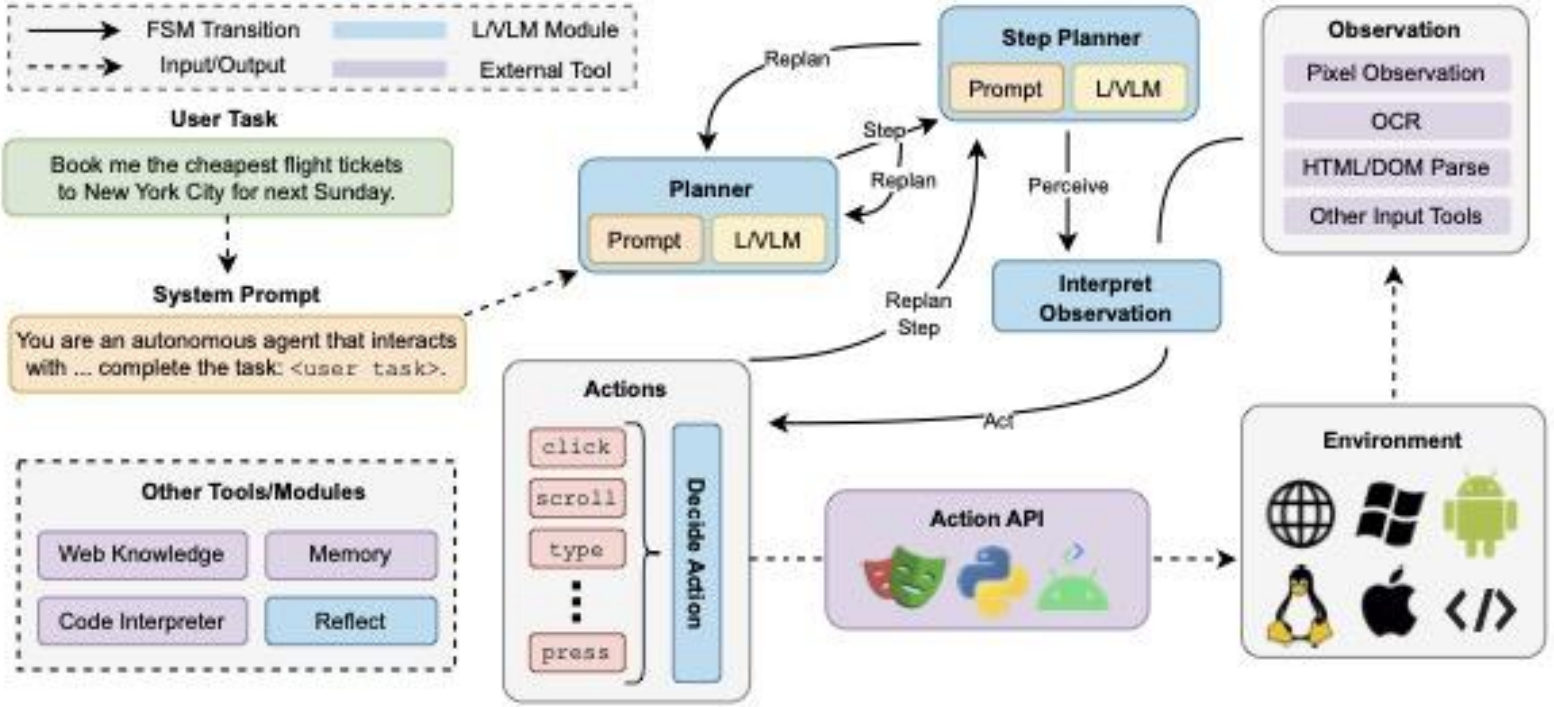
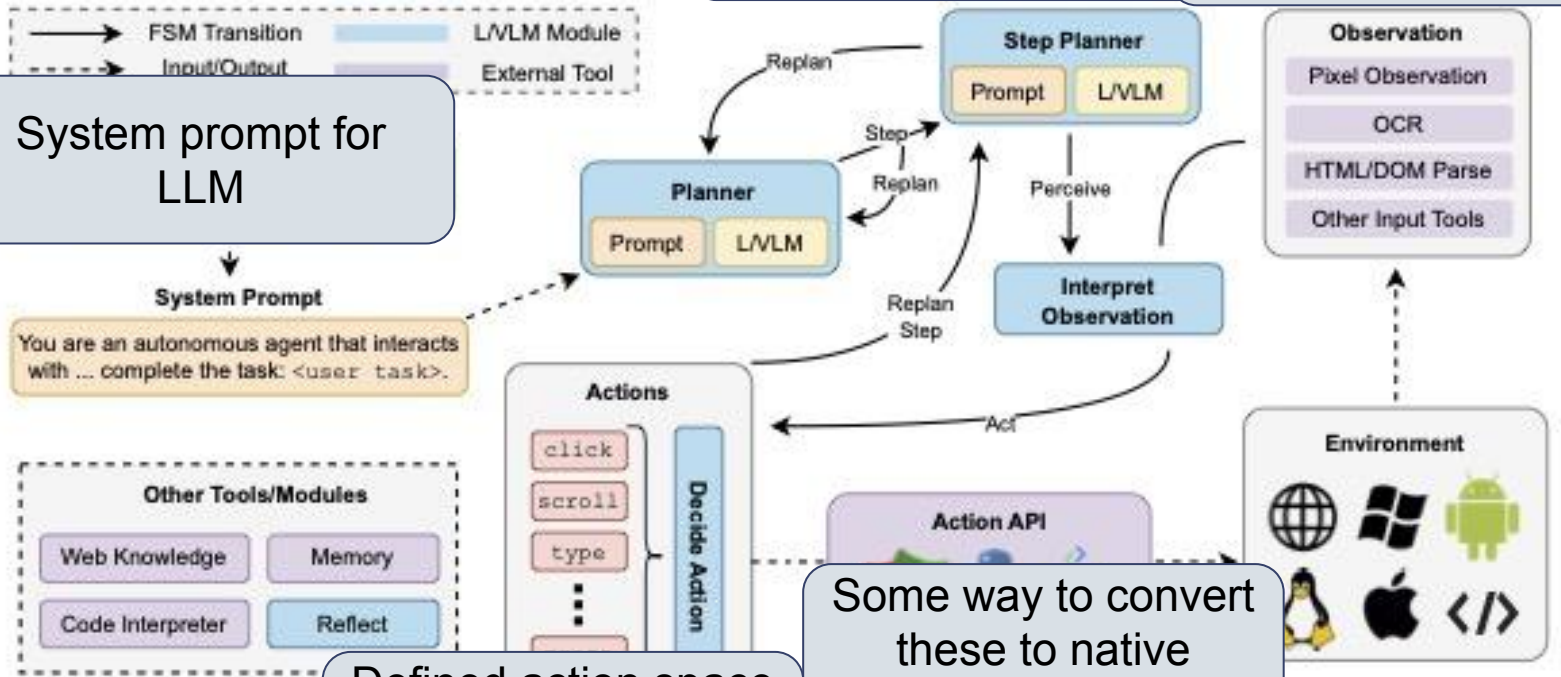


Image credit: Farhan Ishmam

Computer Use Ag

Frameworks for LLMs be better at actions (reflect, plan, replan, e

Hand-crafted observations useful for VLM



System prompt for LLM

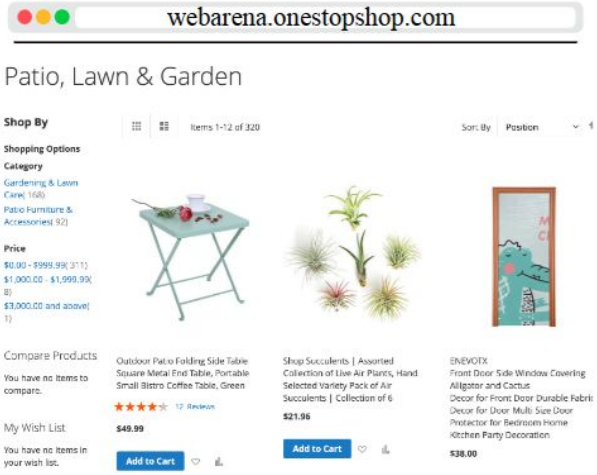
Some way to convert these to native actions in env

Tools VLM can i that do different t

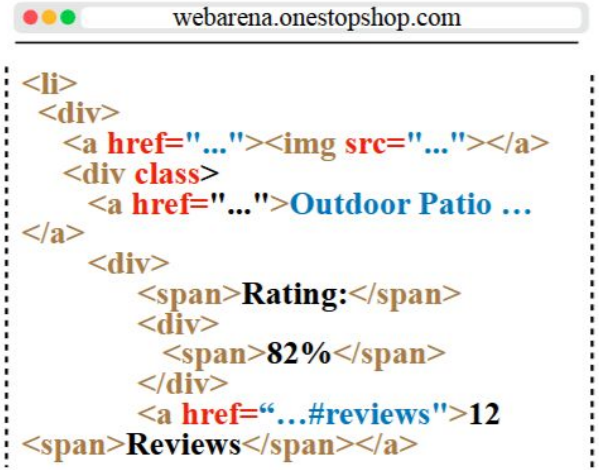
Defined action space that can be expressed in text

Image credit: Farhan Ishmam

Observation Space



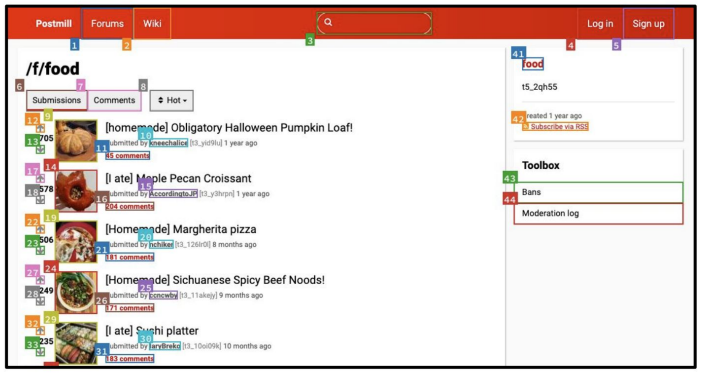
(a) Screenshot of the webpage (included in the observation space of a Vision Language Model (VLM) agent)



(b) HTML DOM Tree

(c) Accessibility Tree

Action Space



Webpage with SoM of Interactable Elements

```
...
[7] [A] [Comments]
[8] [BUTTON] [Hot]
[9] [IMG] [description: picture of a pumpkin]
[10] [A] [kneechalice]
...
```

SoM Elements and Text Content

Action Type a	**Description**
`click [elem]`	Click on element `elem`.
`hover [elem]`	Hover on element `elem`.
`type [elem] [text]`	Type `text` on element `elem`.
`press [key_comb]`	Press a key combination.
`new_tab`	Open a new tab.
`tab_focus [index]`	Focus on the i-th tab.
`tab_close`	Close current tab.
`goto [url]`	Open `url`.
`go_back`	Click the back button.
`go_forward`	Click the forward button.
`scroll [up|down]`	Scroll up or down the page.
`stop [answer]`	End the task with an output.

Methods

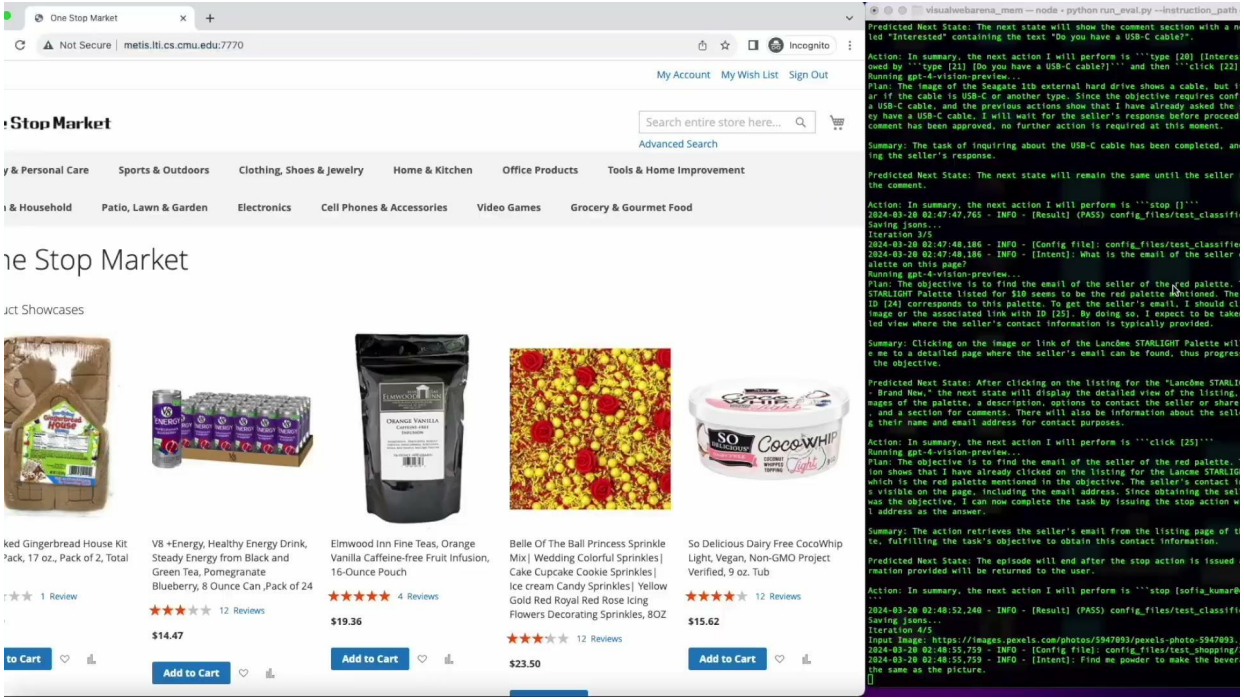
Prompting LLMs

- Take general purpose LLM/VLM
- Clever prompting, observation/action space modification
- See how far it can go

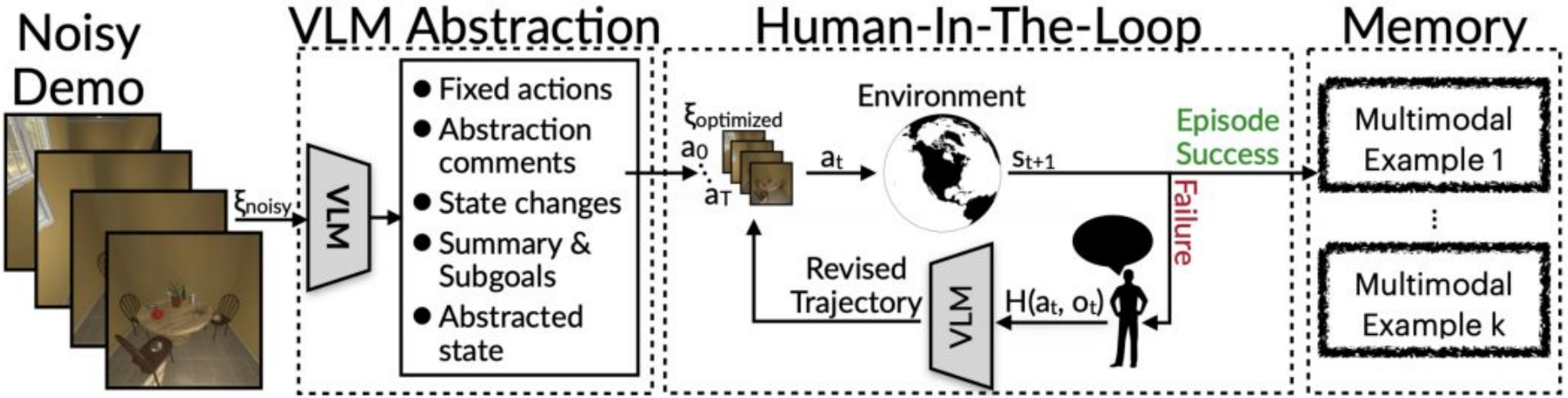
Fixed Base Models

The first category of work does not train the LLM backbone, but relies on innovations in the firmament components or adding new modules such as OCR. Much of the early LLM agent work on web agents uses prompted only GPT-4(V) [47] (e.g [53] [56]) as at the time of its release it was one of the most capable models, was easily callable through an API and had visual input. Later much of the non-training work switched over to GPT-4o [57] due to greater affordability and performance [54]. Claude-3 and Claude-3.5 [58] (e.g. in [54]) and GPT-4-Turbo (in [59]) are other popular choices.

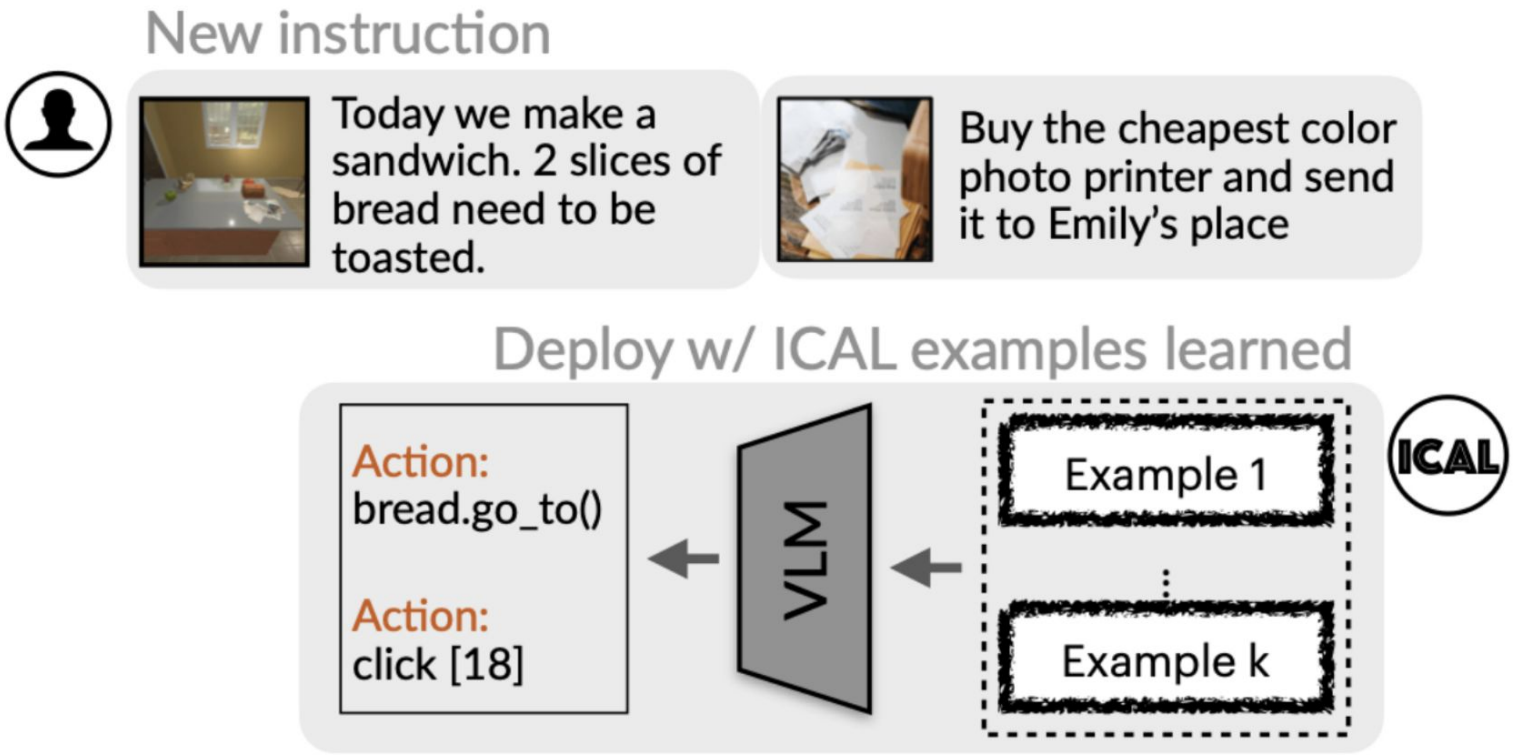
Led by Gabriel Sarch - now Princeton Postdoc



Gabriel Sarch, Lawrence Jang, Michael Tarr, William W Cohen, Kenneth Marino, Katerina Fragkiadaki. “VLM agents generate their own memories: Distilling experience into embodied programs of thought." *NeurIPS* 2024.



Gabriel Sarch, Lawrence Jang, Michael Tarr, William W Cohen, Kenneth Marino, Katerina Fragkiadaki. “VLM agents generate their own memories: Distilling experience into embodied programs of thought." *NeurIPS* 2024.



Gabriel Sarch, Lawrence Jang, Michael Tarr, William W Cohen, Kenneth Marino, Katerina Fragkiadaki. "VLM agents generate their own memories: Distilling experience into embodied programs of thought." *NeurIPS* 2024.

Finetuning LLMs/VLMs

- Fintune for better observation / actions
- Align better with CU problems

Supervised Finetuning

Because base VLMs and LLMs are trained on typical web images and text, certain text and images may be out of distribution and models will struggle on these tasks [60]. This is often the case for web tasks as well, so several works have looked to train models for this task (e.g xLAM [61]). CogAgent [62] for example adds a high-resolution image stack to CogVLM [55] and finetunes on a large-scale dataset of GUI and OCR tasks. Similarly, OS-ATLAS [63] create a large GUI grounding corpus and finetune Qwen2-VL [64] and InternVL-2 [65] to better perform on these tasks.

RL Finetuning

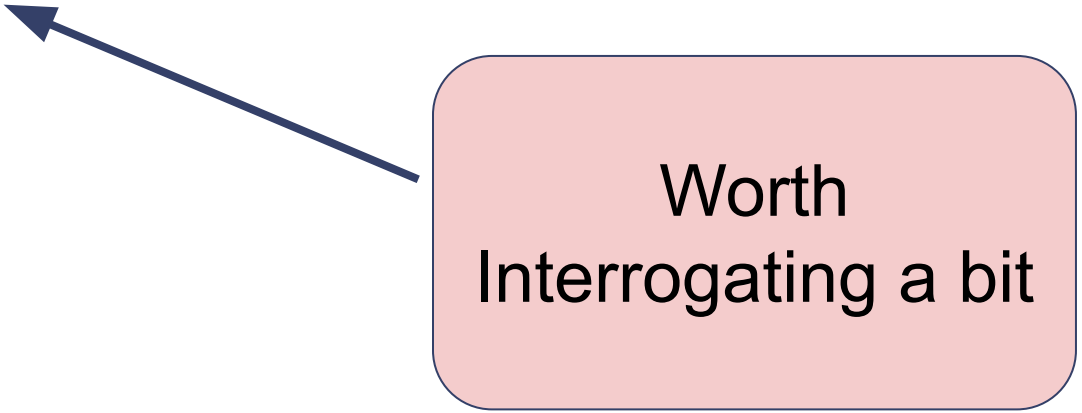
Similarly, there have been many works looking at automatic exploration or web and OS environments for finetuning [66] [67]. These methods interact with the OS or Web environment directly to then update the model. In [67], traces in the environment are created synthetically by either using the dataset training set or prompting an LLM for a task, using the base LLM to generate an actions in the environment, then using another LLM to self-critique to determine if the task was accomplished successfully. Similarly [68] generates synthetic traces by first starting with an environment trace and then labeling the task in hindsight. In both of these works, the synthetic datasets are used for supervised finetuning of a base LLM model for the task.

But why web agents?

- Automation of tedious tasks
- Accessibility

But why web agents?

- Automation of tedious tasks
- Accessibility



But why web agents?

- Automation of tedious tasks
- Accessibility

Real Question: if you were blind, how would you want to use the Internet?

VizWiz: What do blind users actually want?

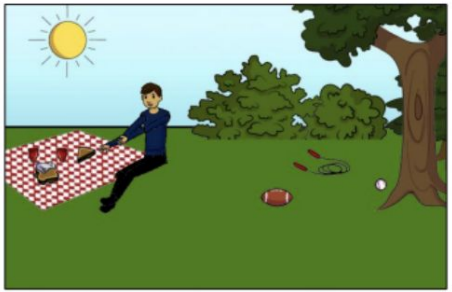
- In VQA, accessibility was often cited as a main motivator
- But is this true?

VizWiz: What do blind users actually want?

Typical VQA Questions

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

VizWiz: What do blind users actually want?

Actual blind user questions

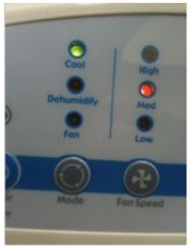
Q: Does this foundation have any sunscreen?
A: yes

Q: What is this?
A: 10 euros

Q: What color is this?
A: green

Q: Please can you tell me what this item is?
A: butternut squash red pepper soup

Q: Is it sunny outside?
A: yes



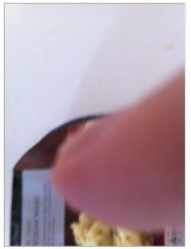
Q: Is this air conditioner on fan, dehumidifier, or air conditioning?
A: air conditioning

Q: What type of pills are these?
A: unsuitable image

Q: What type of soup is this?
A: unsuitable image

Q: Who is this mail for?
A: unanswerable

Q: When is the expiration date?
A: unanswerable



Q: What is this?
A: unanswerable

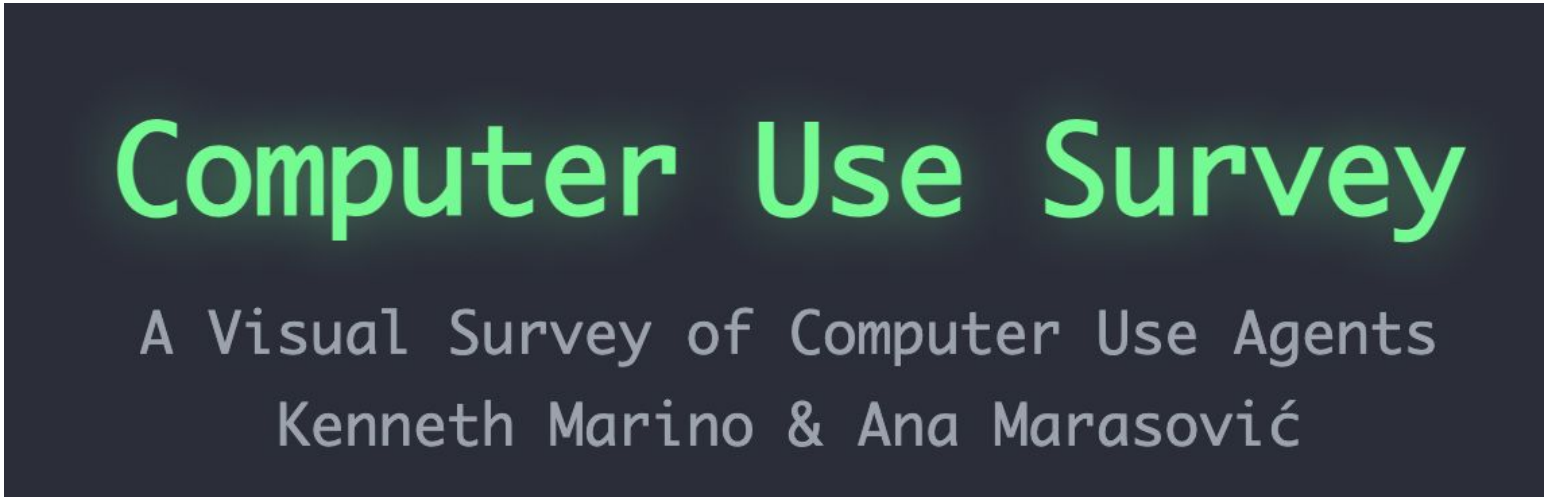
Q: Can you please tell me what the oven temperature is set to?
A: unanswerable

But why web agents?

- Automation of tedious tasks
- Accessibility

How can AI be used to empower users (with and without disabilities)?

Go Check out the Survey(s)



https://kennethmarino.com/computeruse/computeruse.html
https://iclr-blogposts.github.io/2026/blog/2026/web-agent/

Any Questions

Questions

Now for the presentations!