

# Computer Use

## CS6960 MultiModal LLM Agents

---

Kenneth Marino

# 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

## Computer Use Survey

A Visual Survey of Computer Use Agents

Kenneth Marino & Ana Marasović

<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<sup>1</sup>, Juliana Freire<sup>\*,1</sup>, Bharat Kumar<sup>1</sup>, Daniel Lieuwen<sup>1</sup>

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

# Record and replay user actions

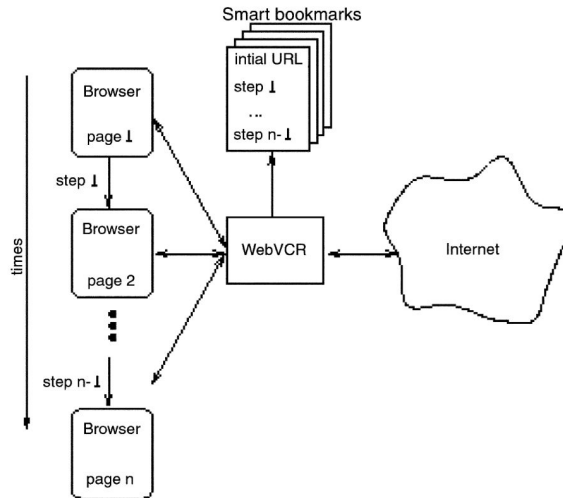


Fig. 6. Recording smart bookmarks.



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

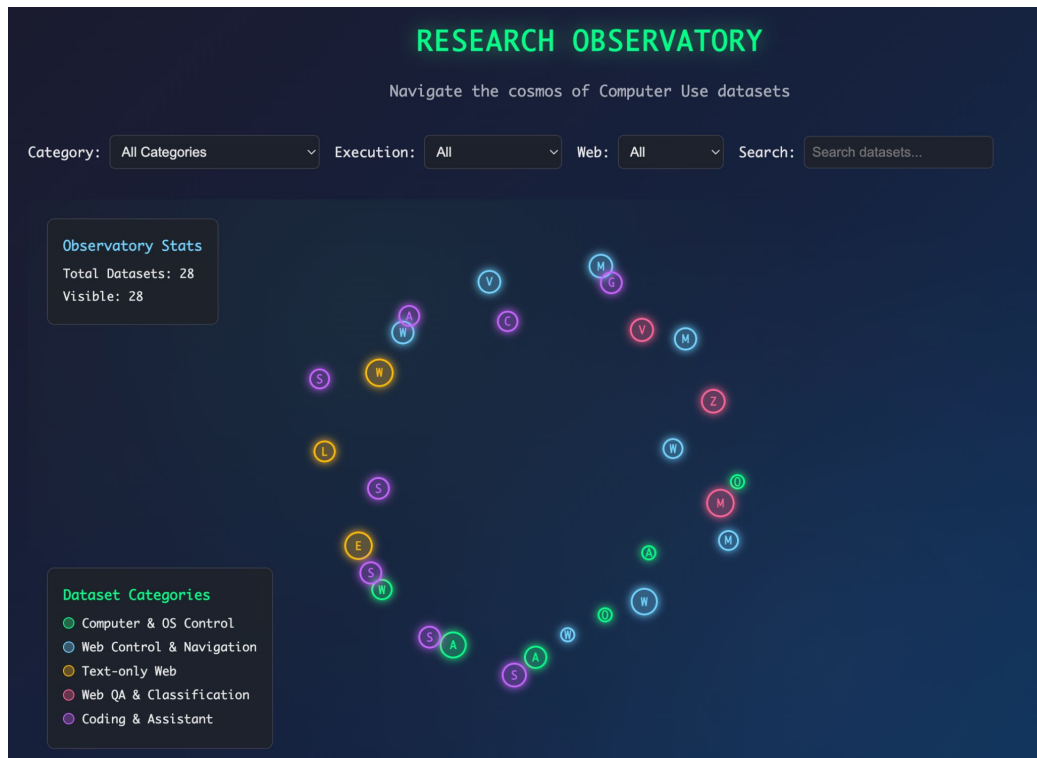
Learning to Navigate Wikipedia by Taking Random Walks

[23]

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

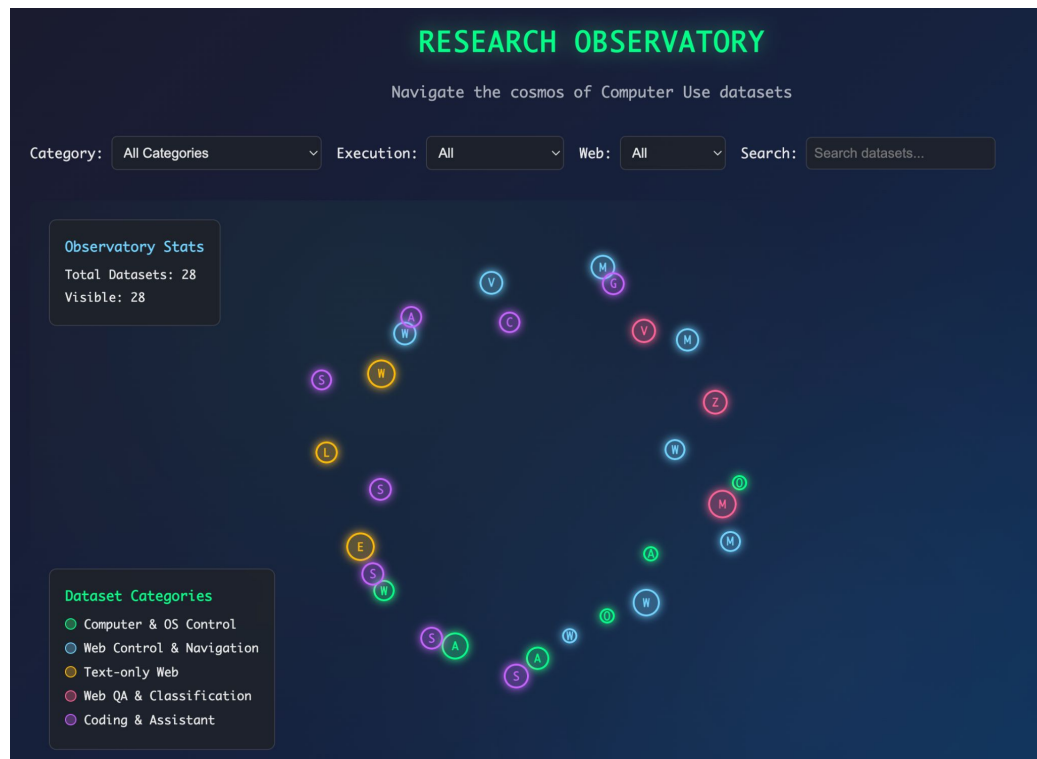
# Computer Use Environments

# Computer Use Environments

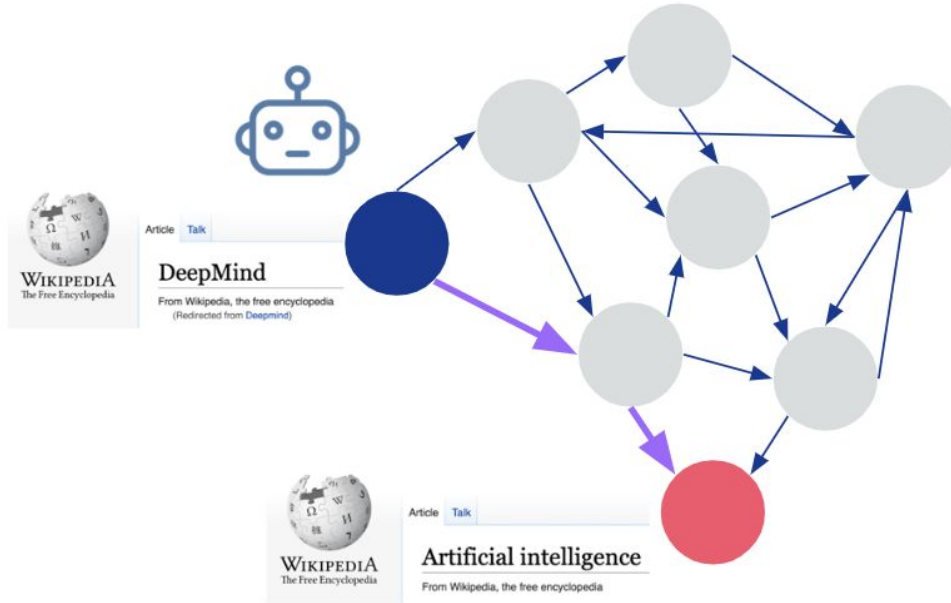


# Computer Use Environments

- 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

## Artificial intelligence

148 languages

Article Talk

Read View source View history Tools

From Wikipedia, the free encyclopedia

*"AI" redirects here. For other uses, see AI (disambiguation), Artificial intelligence (disambiguation), and Intelligent agent.*

**Artificial intelligence (AI)** is the intelligence of machines or software, as opposed to the intelligence of humans or animals. It is a field of study in computer science that develops and studies intelligent machines. Such machines may be called AIs.



AI technology is widely used throughout industry, government, and science. Some high-profile applications are: advanced web search engines (e.g., Google Search), recommendation systems (used by YouTube, Amazon, and Netflix), understanding human speech (such as Google Assistant, Siri, and Alexa), self-driving cars (e.g., Waymo), generative and creative tools (ChatGPT and AI art), and superhuman play and analysis in strategy games (such as chess and Go).

Alan Turing was the first person to conduct substantial research in the field that he called Machine Intelligence.<sup>[2]</sup> Artificial intelligence was founded as an academic discipline in 1956.<sup>[3]</sup> The field went through multiple cycles of optimism<sup>[4]</sup> followed by disappointment and loss of funding.<sup>[6][7]</sup> Funding and interest vastly increased after 2012 when deep learning surpassed all previous AI techniques,<sup>[8]</sup> and after 2017 with the transformer architecture.<sup>[9]</sup> This led to the AI spring of the 2020s, with companies, universities, and laboratories overwhelmingly based in the United States pioneering significant advances in artificial intelligence.<sup>[10]</sup>

The various sub-fields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception, and support for robotics.<sup>[4]</sup> General intelligence (the ability to complete any task performable by a human) is among the field's long-term goals.<sup>[11]</sup>

To solve these problems, AI researchers have adapted and integrated a wide range of problem-solving techniques, including search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, operations research, and economics.<sup>[6]</sup> AI also draws upon psychology, linguistics, philosophy, neuroscience and other fields.<sup>[12]</sup>

### Goals

The general problem of simulating (or creating) intelligence has been broken into sub-problems. These consist of particular traits or capabilities that researchers expect an intelligent system to display. The traits described below have received the most attention and cover the scope of AI research.<sup>[4]</sup>

## Google Search

Article Talk

From Wikipedia, the free encyclopedia

"Google.com" redirects here. For the company itself, see Google. **Google Search** (also known simply as **Google** or **Google.com**) is a search engine owned and operated by Google. Handling more than 3.5 billion searches per day,<sup>[2]</sup> it has a 90% share of the global search engine market.<sup>[14]</sup> It is the most-visited website in the world. Approximately 26.75% of Google's monthly global traffic comes from the United States, 4.44% from India, 4.4% from Brazil, 3.92% from the United Kingdom and 3.84% from Japan according to data provided by Similarweb.<sup>[2]</sup>

## Waymo

Article Talk

From Wikipedia, the free encyclopedia

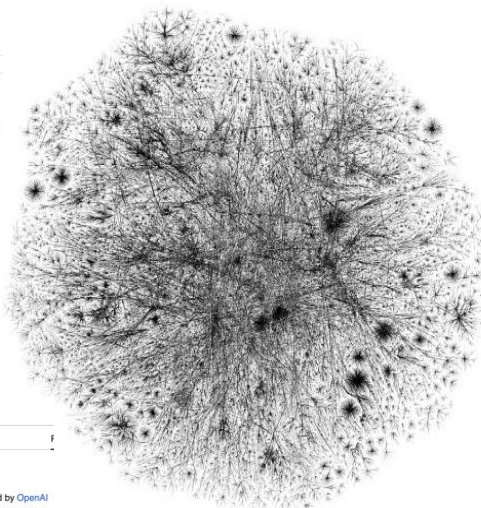
**Waymo LLC**, formerly known as the **Google Self-Driving Car Project**, is an American autonomous driving technology company headquartered in Mountain View, California. It is a subsidiary of Alphabet Inc, the parent company of Google.

## ChatGPT

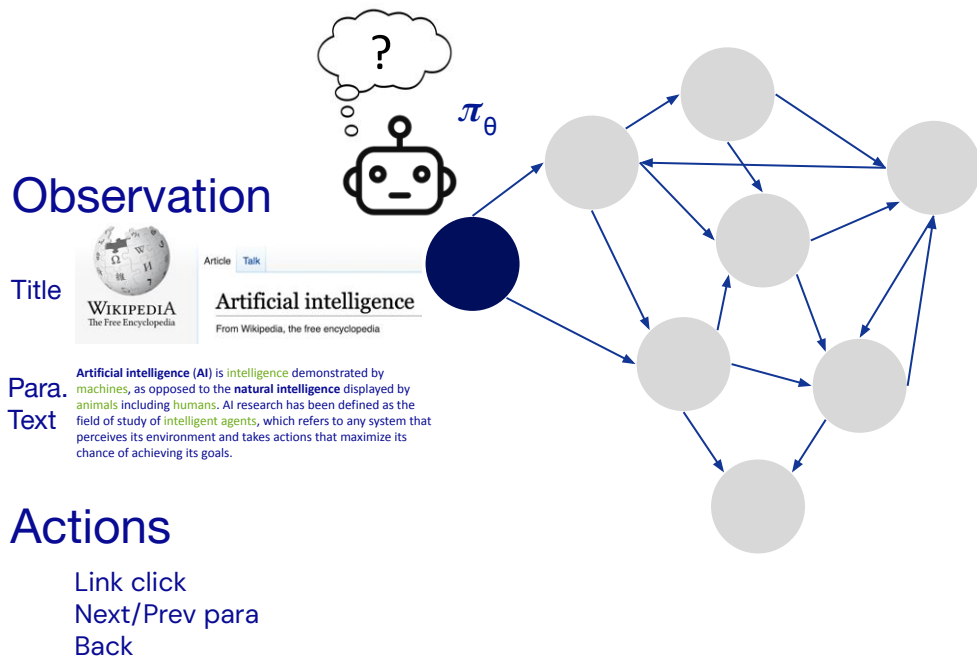
Article Talk

From Wikipedia, the free encyclopedia

**ChatGPT (Chat Generative Pre-trained Transformer)** is a chatbot developed by OpenAI and launched on November 30, 2022. Based on a large language model, it enables users to refine and steer a conversation towards a desired length, format, style, level of detail, and language. Successive prompts and replies, known as prompt engineering, are considered at each conversation stage as a context.<sup>[2]</sup>

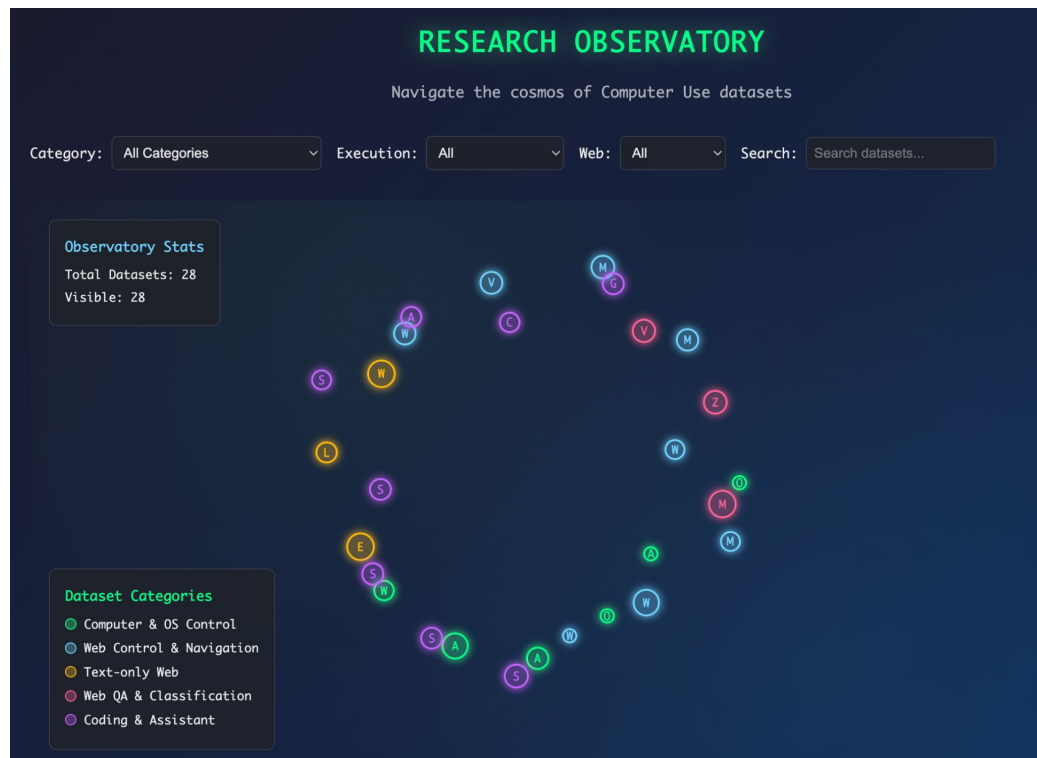


# Learn to Navigate Wikipedia Links



# Computer Use Environments

- Text-only Web Envs
- Full HTML web

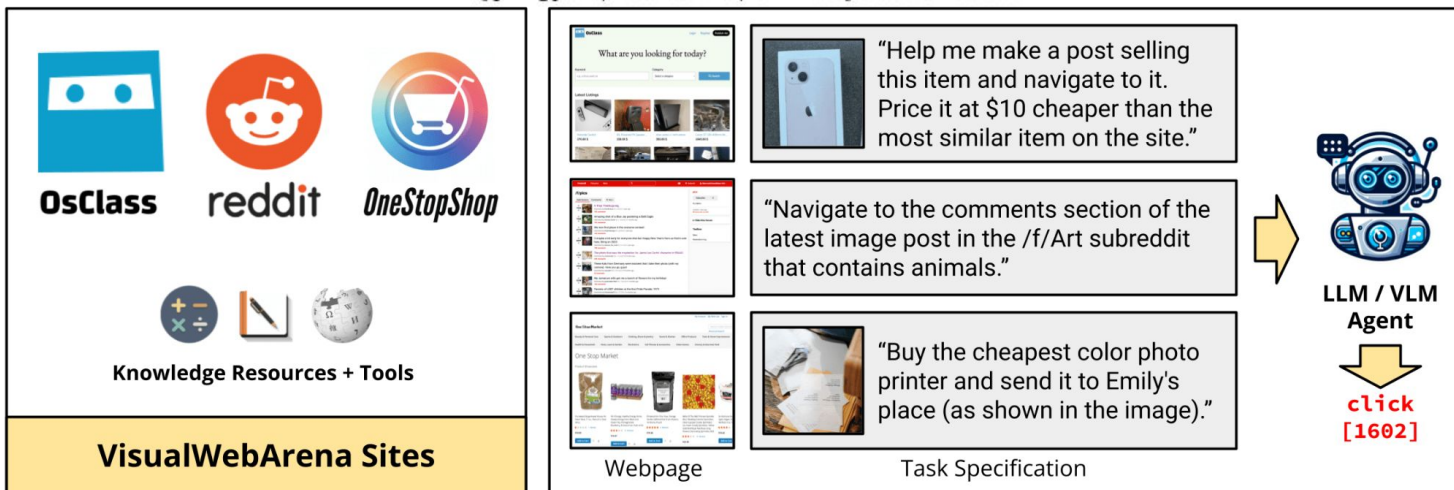


# VisualWebArena: Evaluating Multimodal Agents on Realistic Visually Grounded Web Tasks

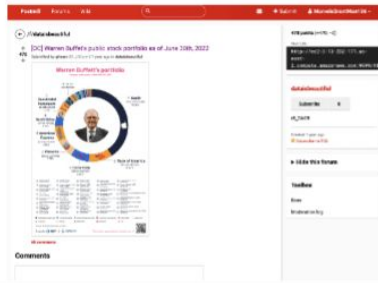
Jing Yu Koh Robert Lo\* Lawrence Jang\* Vikram Duvvur\*  
Ming Chong Lim\* Po-Yu Huang\* Graham Neubig Shuyan Zhou  
Ruslan Salakhutdinov Daniel Fried

Carnegie Mellon University

{jingyuk,rsalaku,dfried}@cs.cmu.edu



Visual







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.

ally



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



  
**OsClass**    **redc**



  
**Knowledge Resour**

**VisualWebAI**



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


**OsClass**

Pristine 2021 Toyota 86 - Low Miles, Factory Warranty
 25000.00 \$

Released date: 20231024  
 Modified date: 20231025  
 Location: Pittsburgh, Pennsylvania, United States






Contact publisher: 

**Shared information**

- Available for purchase on monthly and yearly
- Item was sold through Open Market
- Description of other information displayed on this page
- Description of other information displayed on this page
- This site is owned by the user

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


  
**LLM / VLM Agent**


  

  
**click**  
**[1602]**



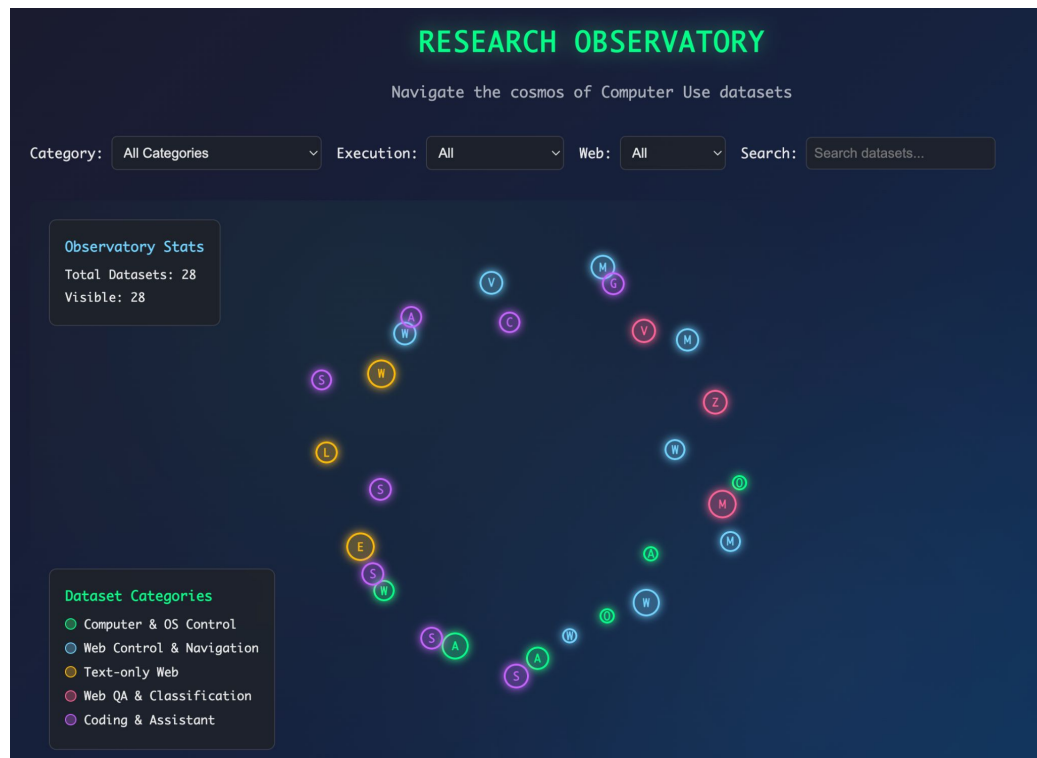
Led by Gabriel Sarch - now Princeton Postdoc

The screenshot shows a web browser window displaying the One Stop Market website. The website features a search bar, navigation tabs for various product categories, and a grid of product listings. The terminal window on the right displays a series of AI-generated text blocks, including "Predicted Next State" and "Action" sections, which appear to be part of a larger program or script. The text in the terminal is partially obscured by a dark background, but some key phrases like "stop", "click", and "summary" are visible.

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

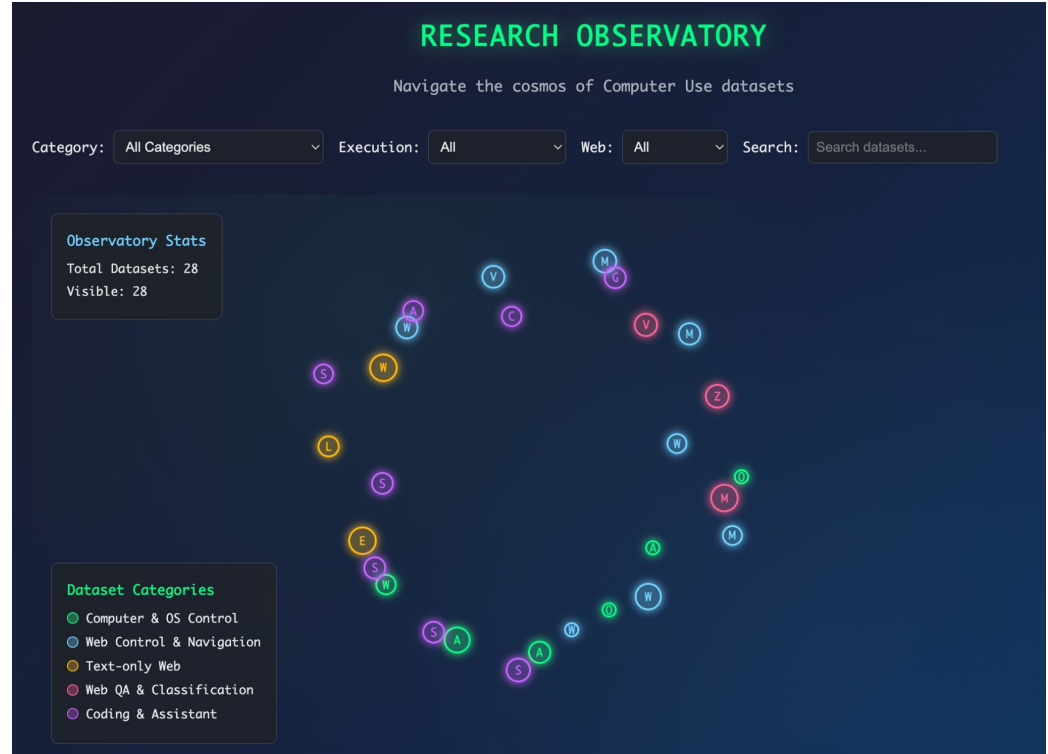
- Text-only Web Envs
- Full HTML web
- Specific tools (code)



# Computer Use Environments

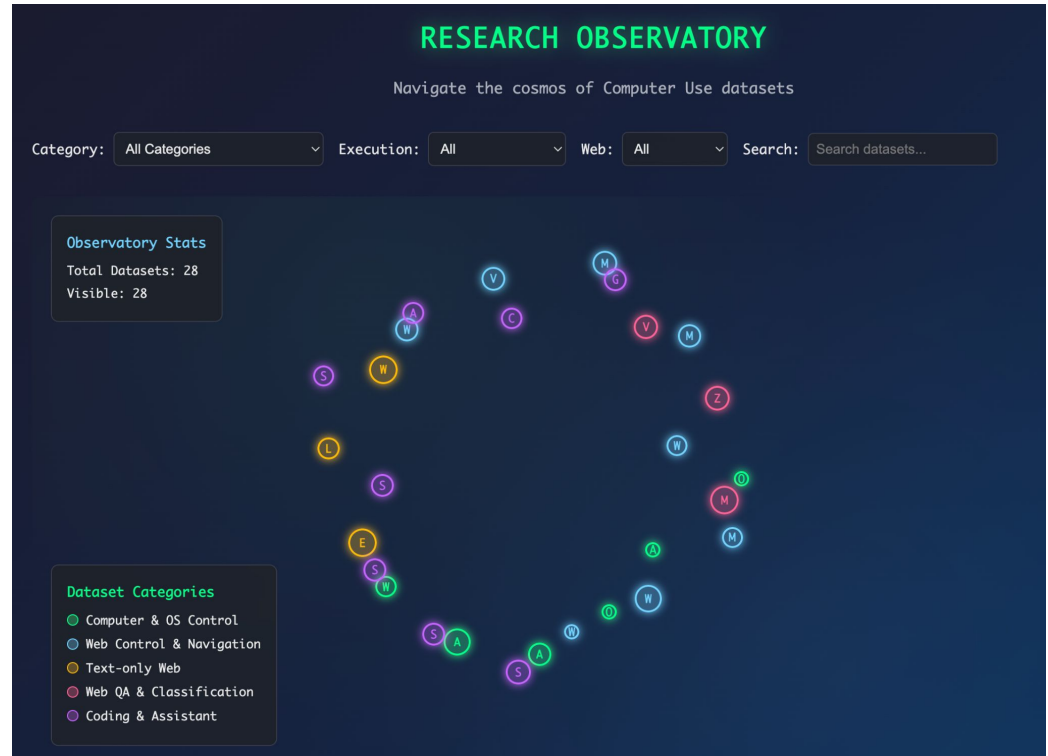
- Text-only Web Envs
- Full HTML web
- Specific tools (code)

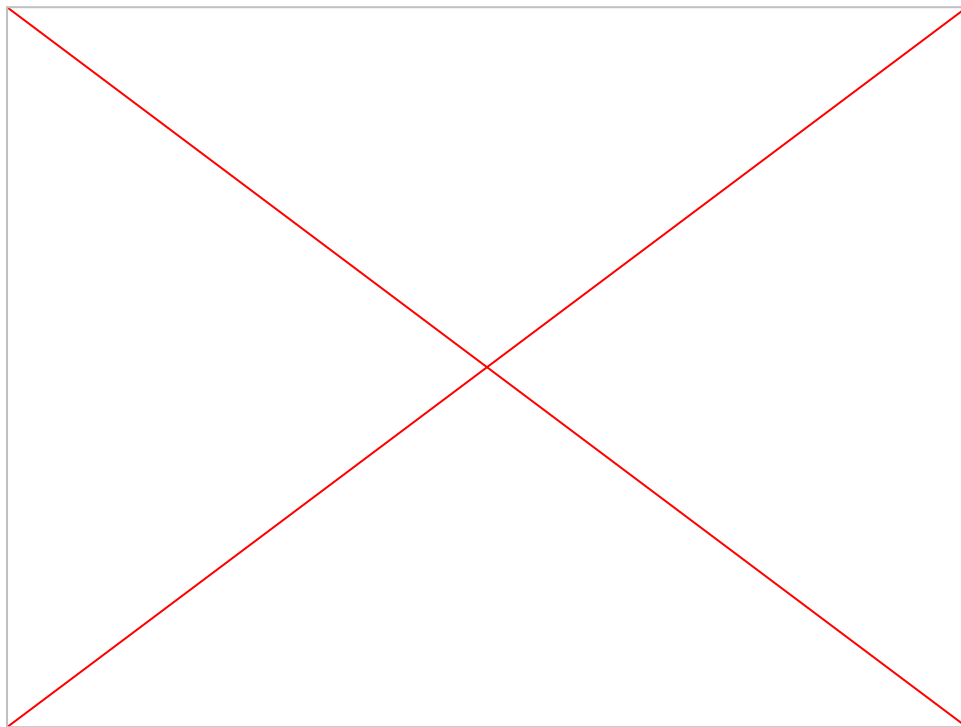
Is an agent that navigates a GitHub codebase and adds files a coding agent or a Computer Use Agent?



# Computer Use Environments

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







# Computer Use Agents

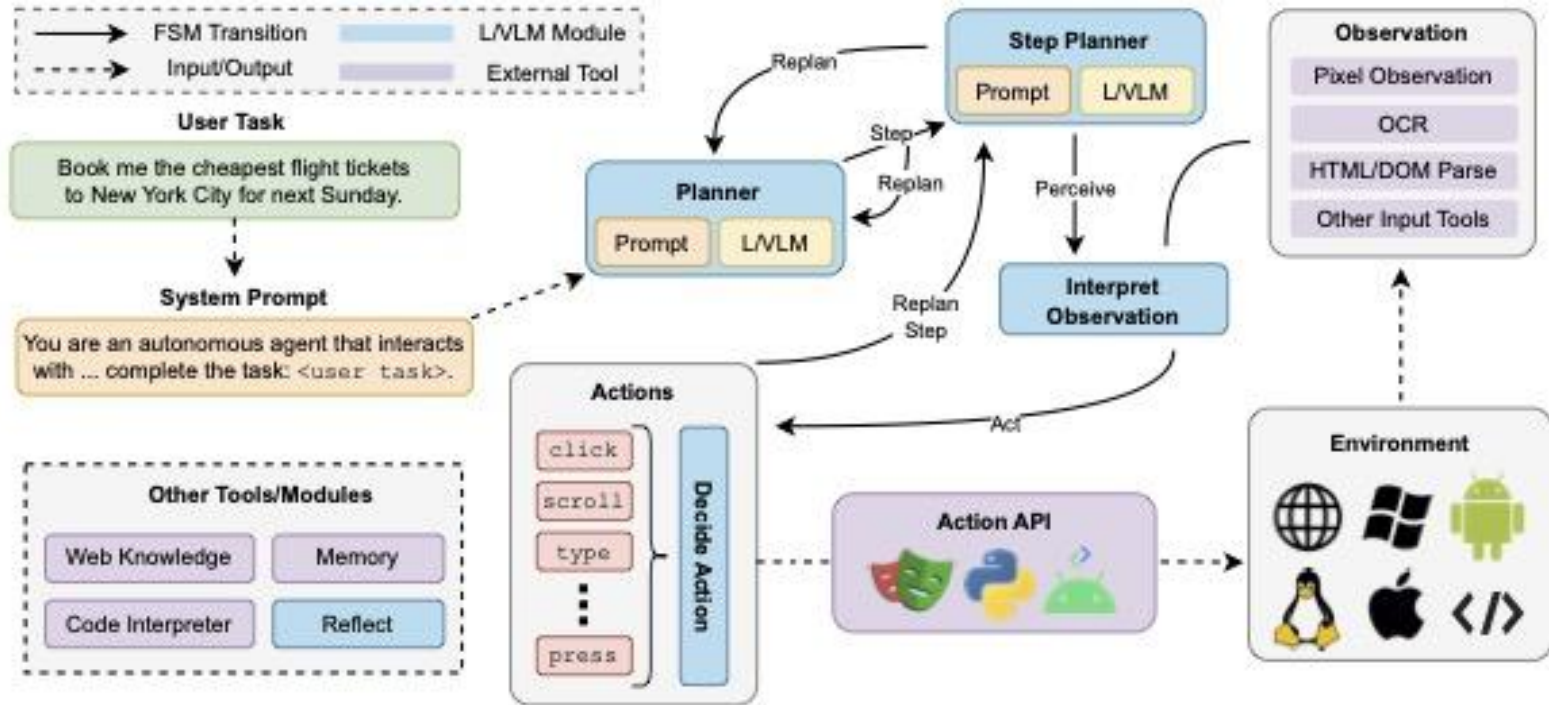


Image credit: Farhan Ishmam

# Computer Use Agents

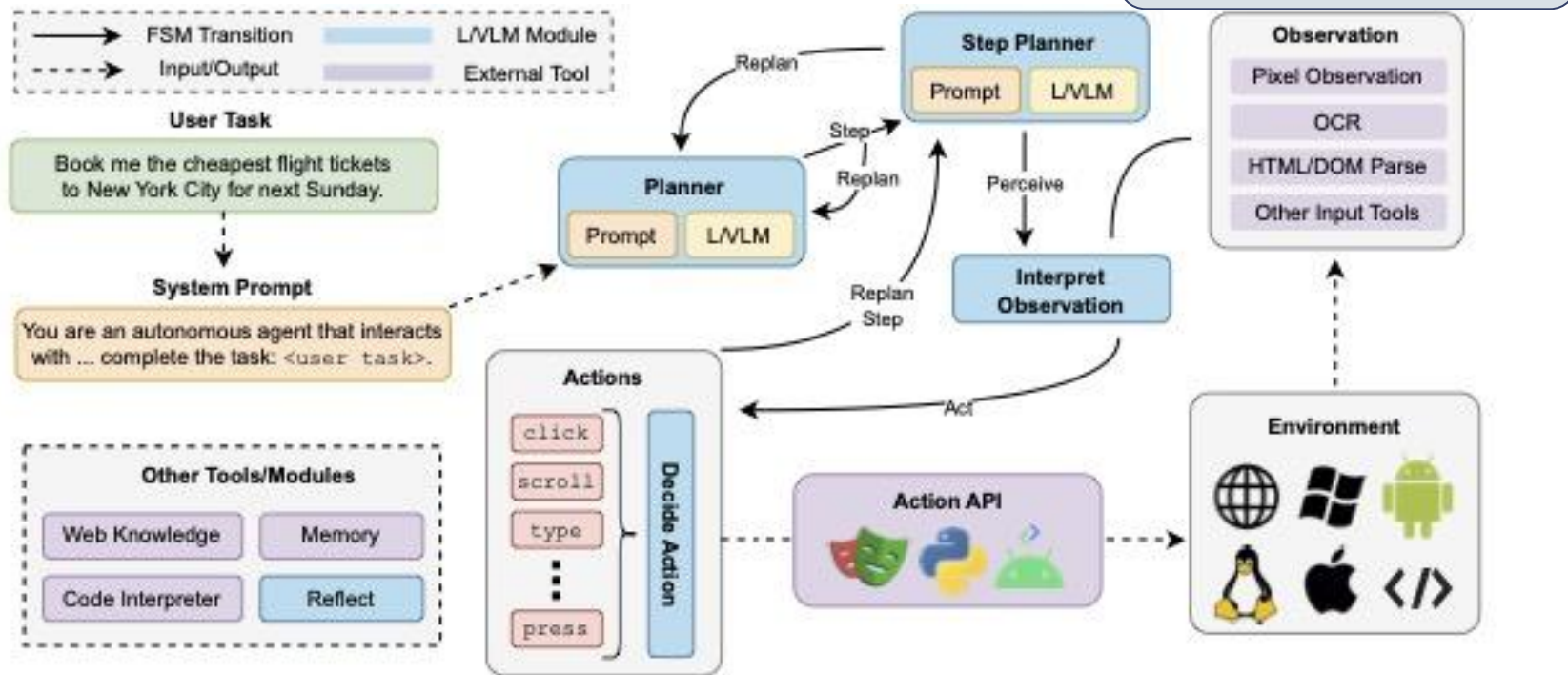


Image credit: Farhan Ishmam

# Computer Use Agents

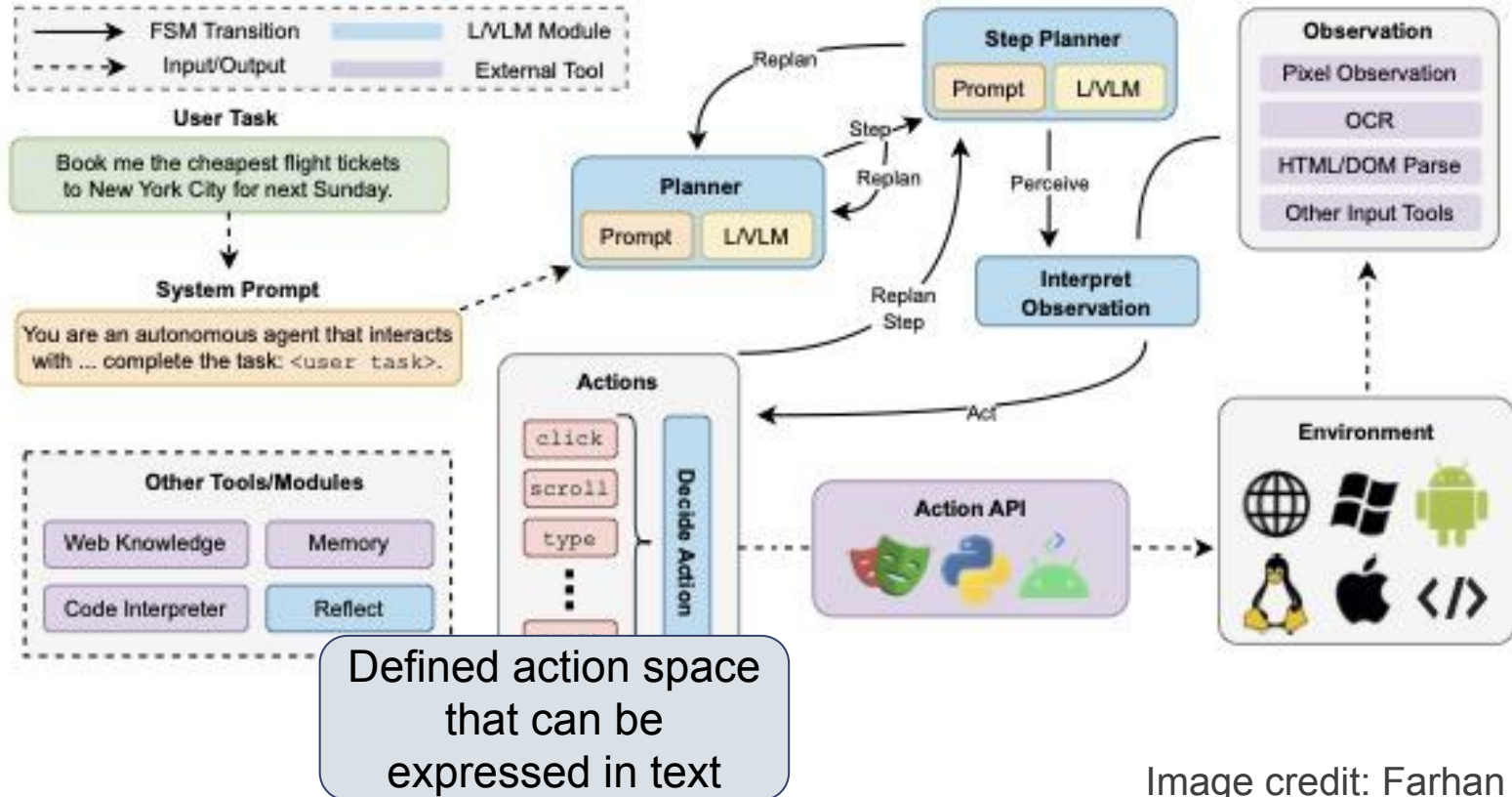
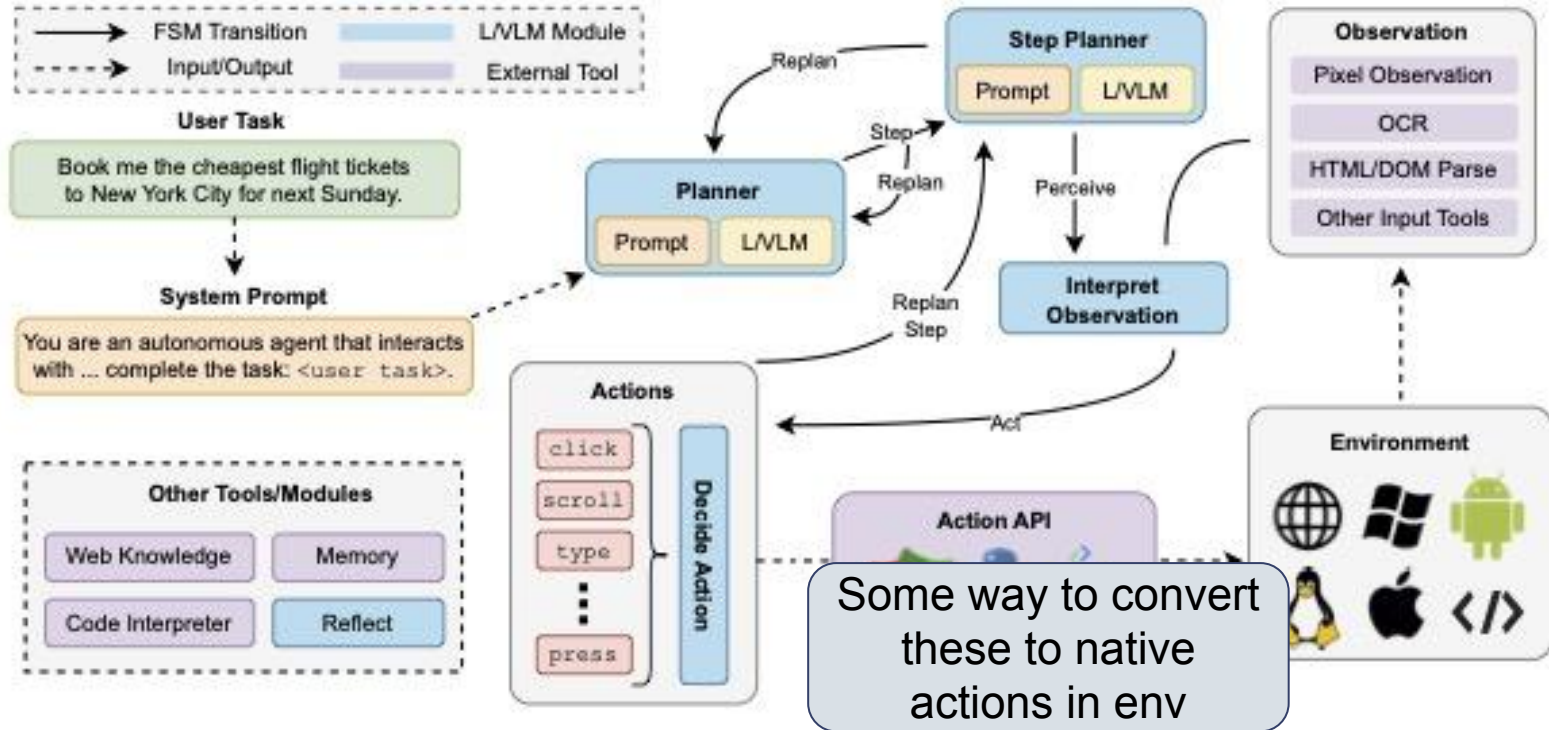
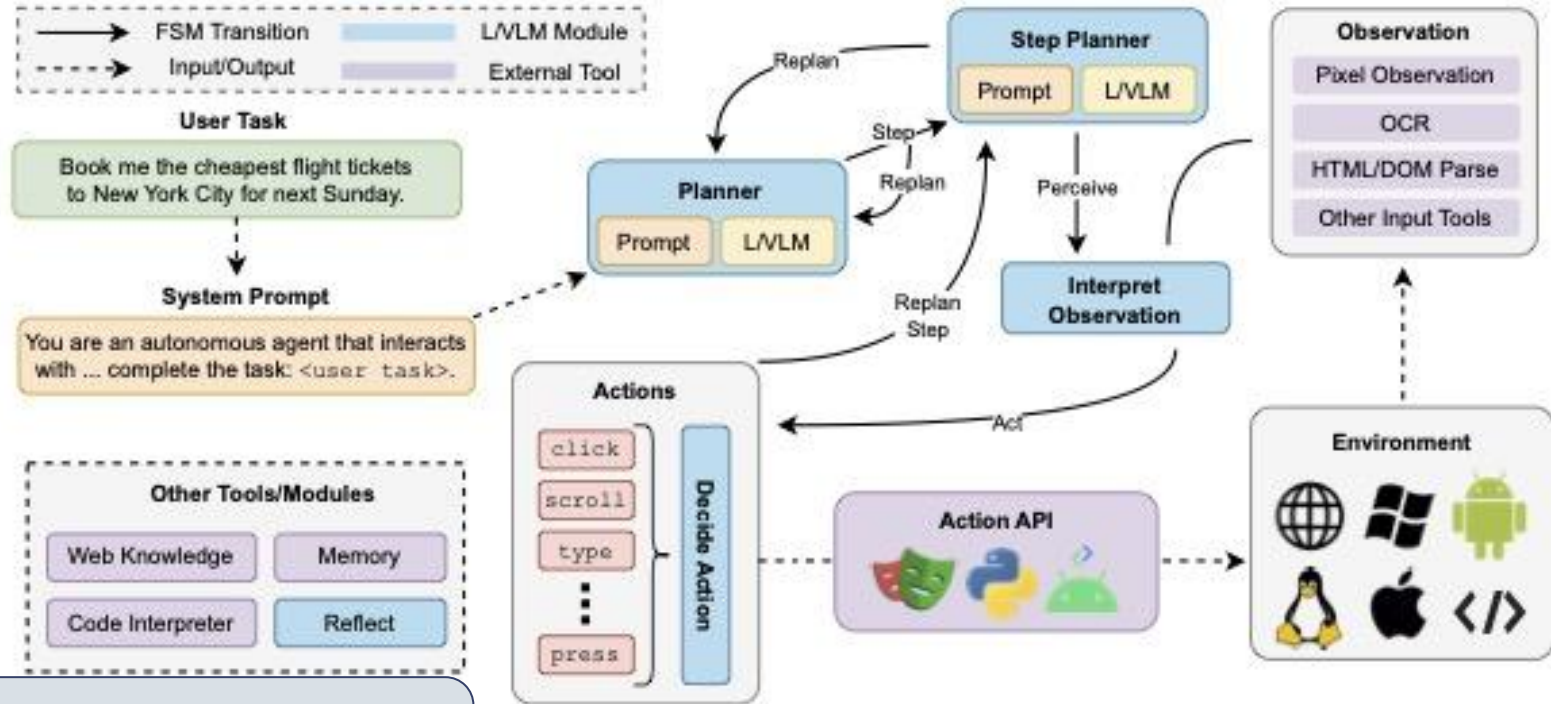


Image credit: Farhan Ishmam

# Computer Use Agents



# Computer Use Agents



Tools VLM can invoke that do different things

Image credit: Farhan Ishmam

# Computer Use Agents

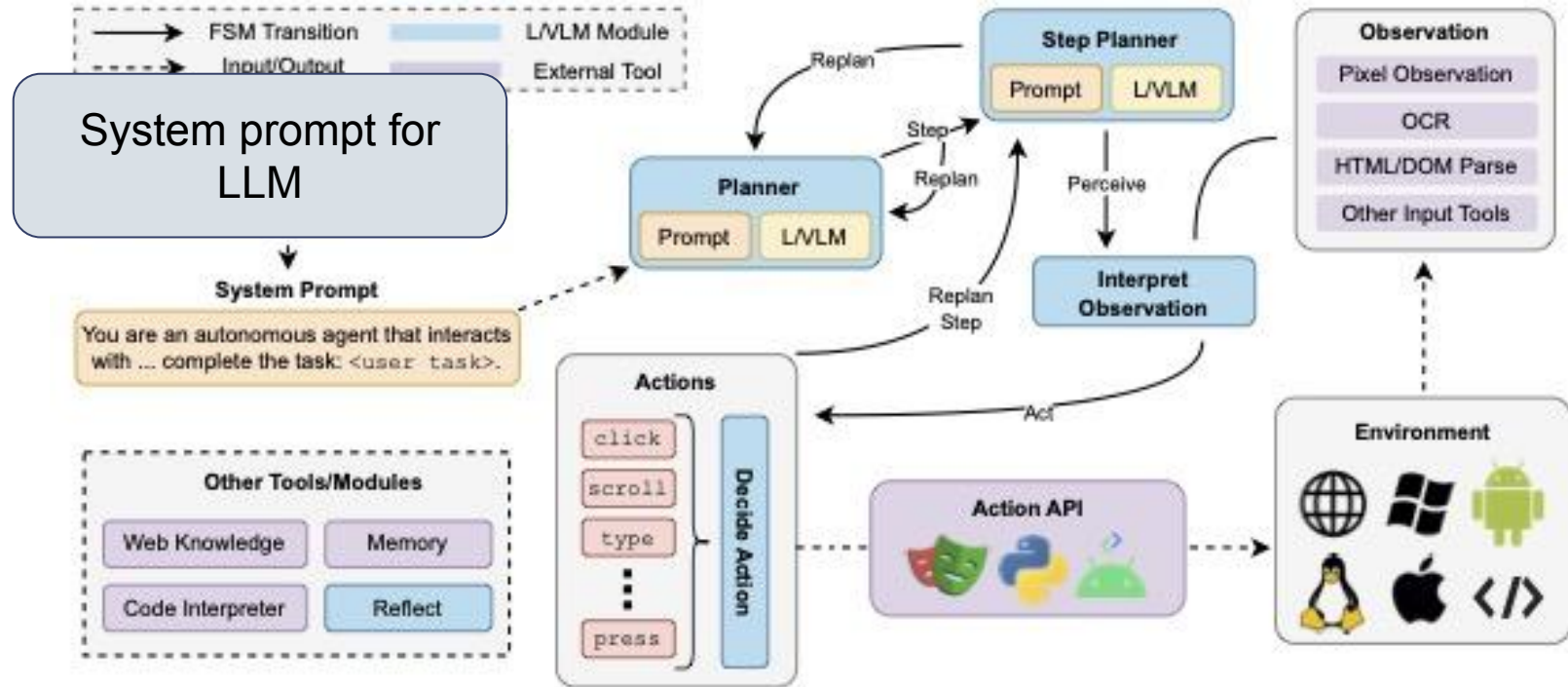


Image credit: Farhan Ishmam

# Computer Use Agent

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

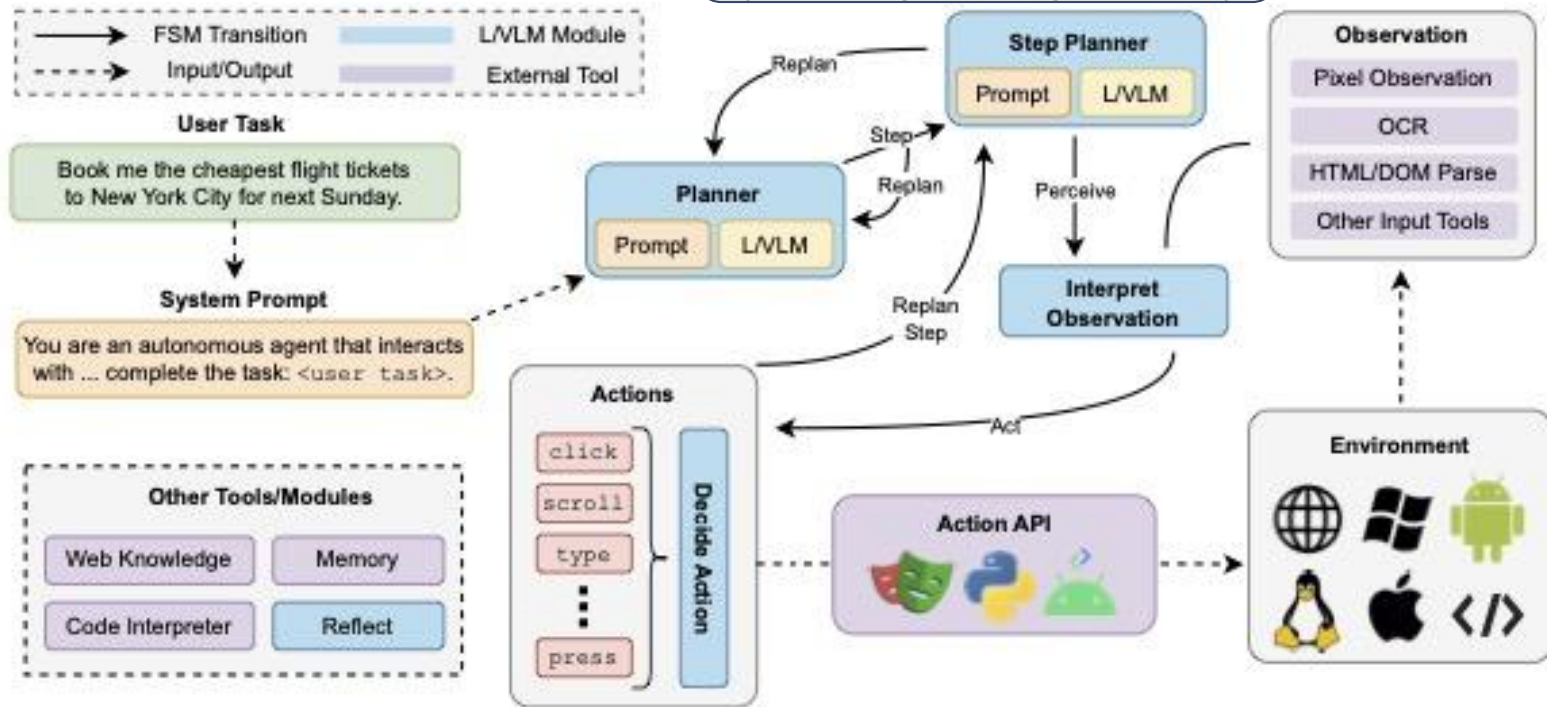
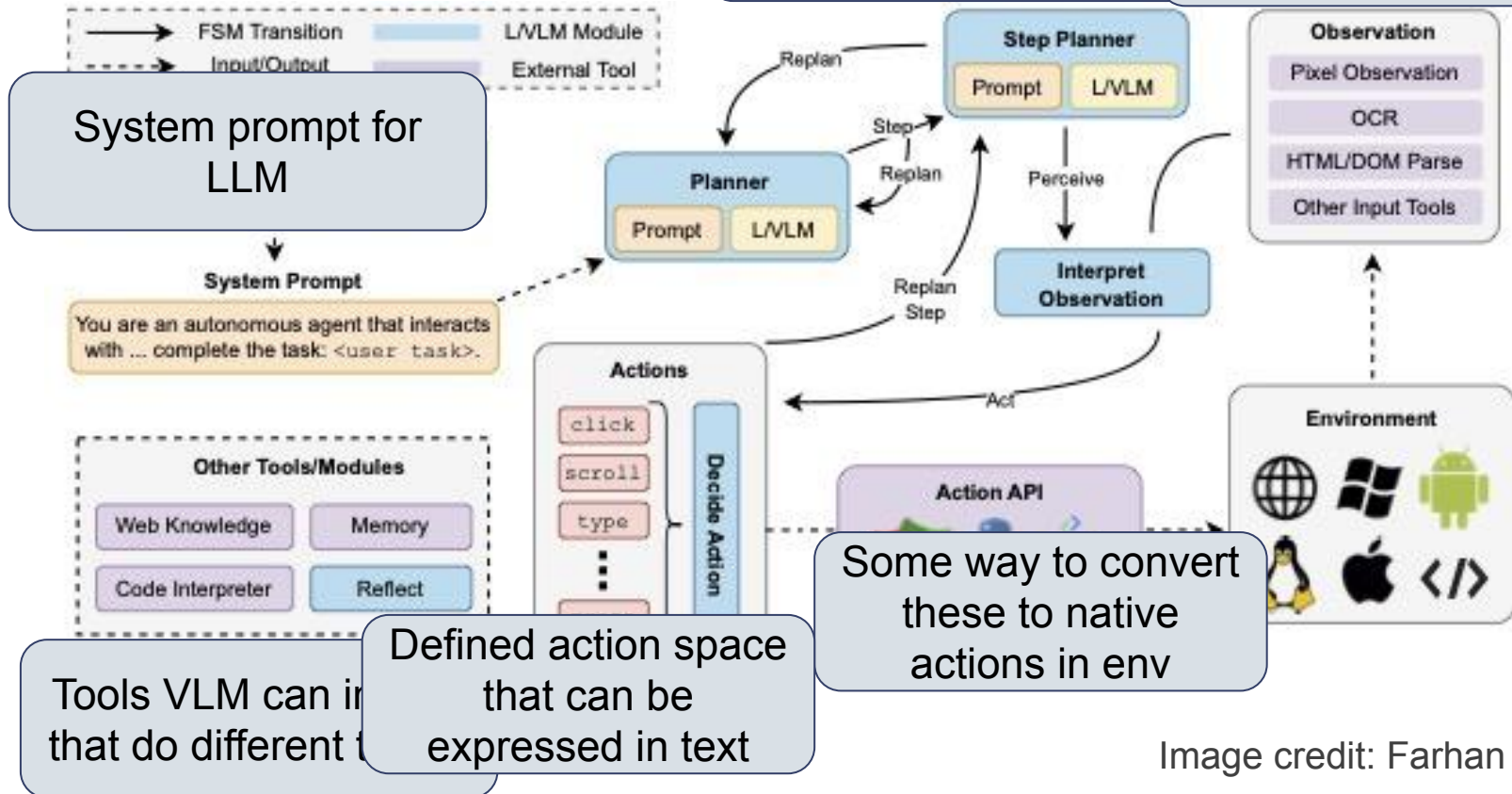


Image credit: Farhan Ishmam

# Computer Use Agent

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

Hand-crafted  
observations useful  
for VLM



System prompt for LLM

You are an autonomous agent that interacts with ... complete the task: <user task>.

Tools VLM can interact with that do different things

Defined action space that can be expressed in text

Some way to convert these to native actions in env

# Observation Space

webarena.onestopshop.com

Patio, Lawn & Garden

Shop By

Shopping Options

Category

Gardening & Lawn Care (165)

Patio Furniture & Accessories (92)

Price

\$0.00 - \$999.99 (311)

\$1,000.00 - \$1,999.99 (8)

\$3,000.00 and above (1)

Compare Products

You have no items to compare.

My Wish List

You have no items in your wish list.

Outdoor Patio Folding Side Table

Square Metal End Table, Portable

Small Bistro Coffee Table, Green

★★★★★ 12 Reviews

\$49.99

Add to Cart

Shop Succulents | Assorted Collection of Live Air Plants, Hand Succulents | Collection of 6

\$21.99

Add to Cart

ENEVOTE Front Door Side Window Covering Alligator and Cactus Decor for Front Door Durable Fabric Decor for Door Multi-Sure Door Protector for Bedroom Home Kitchen Party Decoration

\$38.00

webarena.onestopshop.com

```
<li>
  <div>
    <a href="..."></a>
    <div class>
      <a href="...">Outdoor Patio ...
    </a>
    <div>
      <span>Rating:</span>
      <div>
        <span>82%</span>
      </div>
      <a href="...#reviews">12
    </a>
    <span>Reviews</span></a>
  </div>
</li>
```

webarena.onestopshop.com

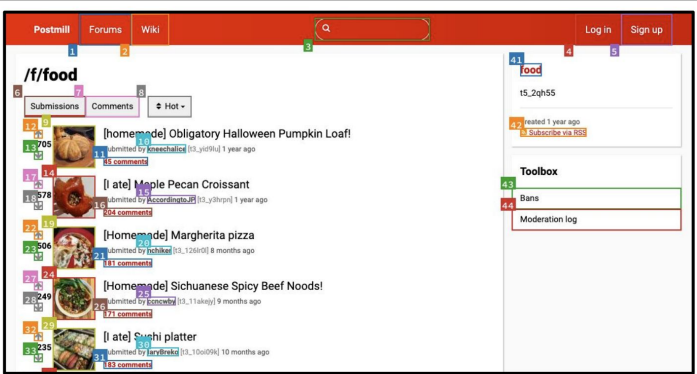
```
RootWebArea 'Patio, Lawn ..'
  link 'Image'
  img 'Image'
  link 'Outdoor Patio..'
  LayoutTable ''
    StaticText 'Rating:'
    generic '82%'
    link '12 Reviews'
  StaticText '$49.99'
  button 'Add to Cart' focusable: True
  button 'Wish List' focusable: ...
  button 'Compare' focusable: ...
```

(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



The screenshot shows a forum page with several elements annotated with numbers 1 through 10. The annotations are: 1. Search bar, 2. Forum navigation tabs, 3. Submissions/Comments/Hot dropdown, 4. Log in/Sign up buttons, 5. User profile, 6. Post title, 7. Post content, 8. Comment button, 9. Image, 10. Comment button.

**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

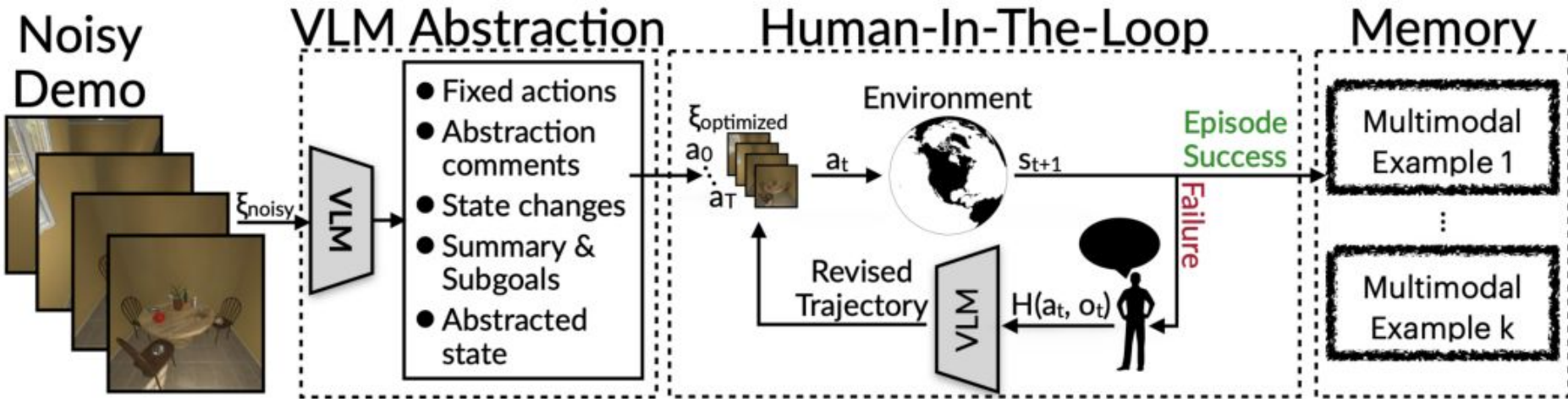
The first category of work does not train the LLM backbone, but relies on innovations in the firmament components or adding new modules such as OCR. Much of the early LLM agent work on web agents uses prompted only GPT-4(V) [47] (e.g [53] [56]) as at the time of its release it was one of the most capable 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

The screenshot shows a web browser window displaying the One Stop Market website. The website has a navigation bar with categories like Personal Care, Sports & Outdoors, Clothing, Shoes & Jewelry, Home & Kitchen, Office Products, and Tools & Home Improvement. Below the navigation bar, there are product listings with images and descriptions. The terminal window on the right side of the browser shows a series of AI-generated text blocks, including "Predicted Next State", "Action: In summary, the next action I will perform is...", "Plan: The image of the package...", and "Summary: The task of inquiring about the USB-C cable has been completed...". The terminal also shows some code snippets like "config\_files/test\_classification.py" and "config\_files/test\_classification.py".

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.



## New instruction

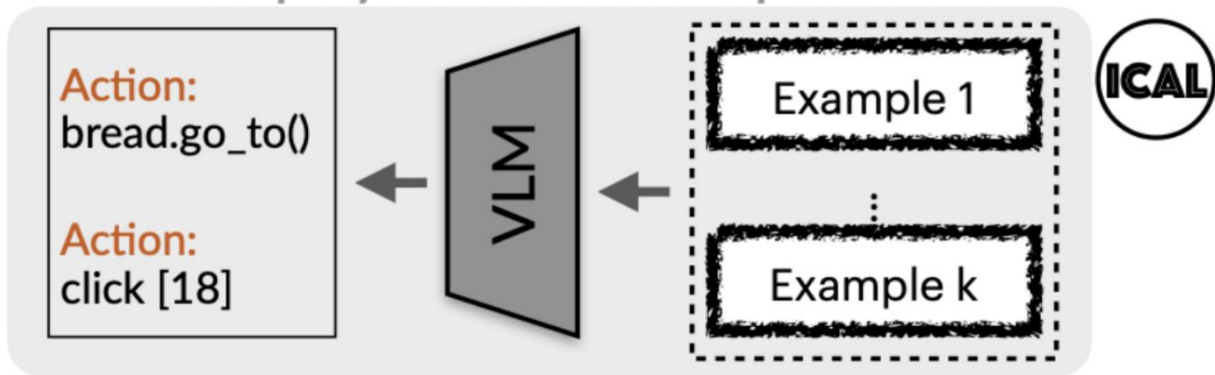


Today we make a sandwich. 2 slices of bread need to be toasted.



Buy the cheapest color photo printer and send it to Emily's place

## Deploy w/ ICAL examples learned



# 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



Worth  
Interrogating a bit

# 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



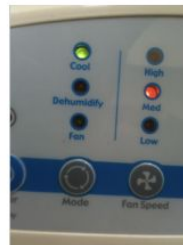
**Q:** What color is this?  
**A:** green



**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



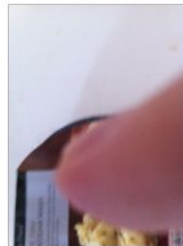
**Q:** What type of soup is this?  
**A:** unsuitable image



**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)

## Computer Use Survey

A Visual Survey of Computer Use Agents

Kenneth Marino & Ana Marasović

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

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

# Any Questions



Questions

**Now for the presentations!**