

# Multimodal Autonomous AI Agents

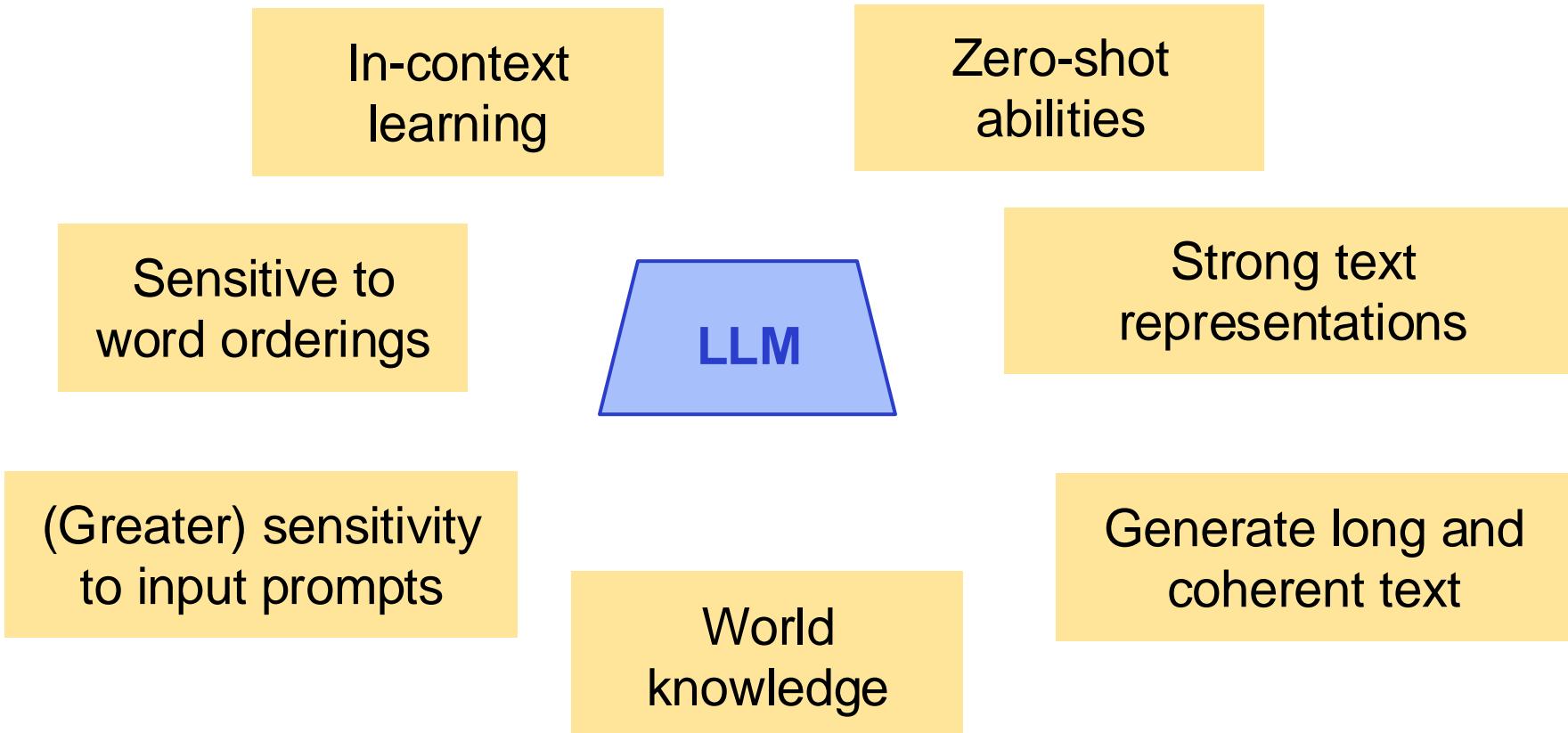
Russ Salakhutdinov

Machine Learning Department  
Carnegie Mellon University

Carnegie  
Mellon  
University

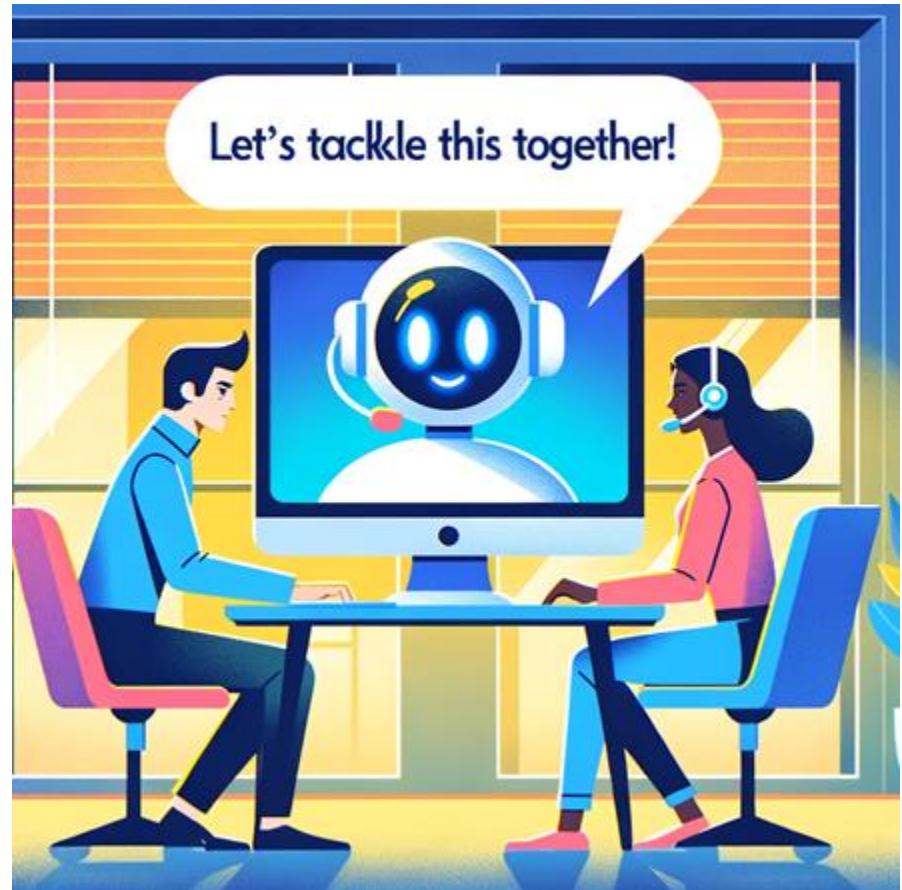


# Large Language Models



# Autonomous AI Agents

- Many productive tasks we perform today are done on the computer
  - And many of these are on the web
- Many opportunities to automate menial tasks
- Augment human capabilities



Generated with DALLE

# Autonomous Agents

vpc-01 3 / channy-vpc Actions ▾

**Details** Info

|   |   |                               |                                       |
|---|---|-------------------------------|---------------------------------------|
| VPC ID<br>vpc-01-03                       | State<br><span>Available</span>                 | DNS hostnames<br>Enabled      | DNS resolution<br>Enabled             |
| Tenancy<br>Default                        | DHCP option set<br>dopt-01                      | Main route table<br>rtb-06-06 | Main network ACL<br>acl-05-56         |
| Default VPC<br>No                         | IPv4 CIDR<br>10.0.0.0/17                        | IPv6 pool<br>-                | IPv6 CIDR (Network border group)<br>- |
| Network Address Usage metrics<br>Disabled | Route 53 Resolver DNS Firewall rule groups<br>- | Owner ID<br>channy            |                                       |

[Resource map](#) | [CIDRs](#) | [Flow logs](#) | [Tags](#)

**Resource map** Info

The diagram illustrates the relationships between VPC resources. A central orange box labeled "channy-rtb-public" is connected to three subnets: "channy-subnet-public1-us-west-2a", "channy-subnet-private4-us-west-2a", and "channy-subnet-private1-us-west-2a". It is also connected to two route tables: "channy-rtb-private6-us-west-2c" and "channy-rtb-private4-us-west-2a". Finally, it connects to three network connections: "channy-igw", "channy-nat-public1-us-west-2a", and "channy-vpce-s3".

**VPC** [Show details](#)  
Your AWS virtual network  
channy-vpc

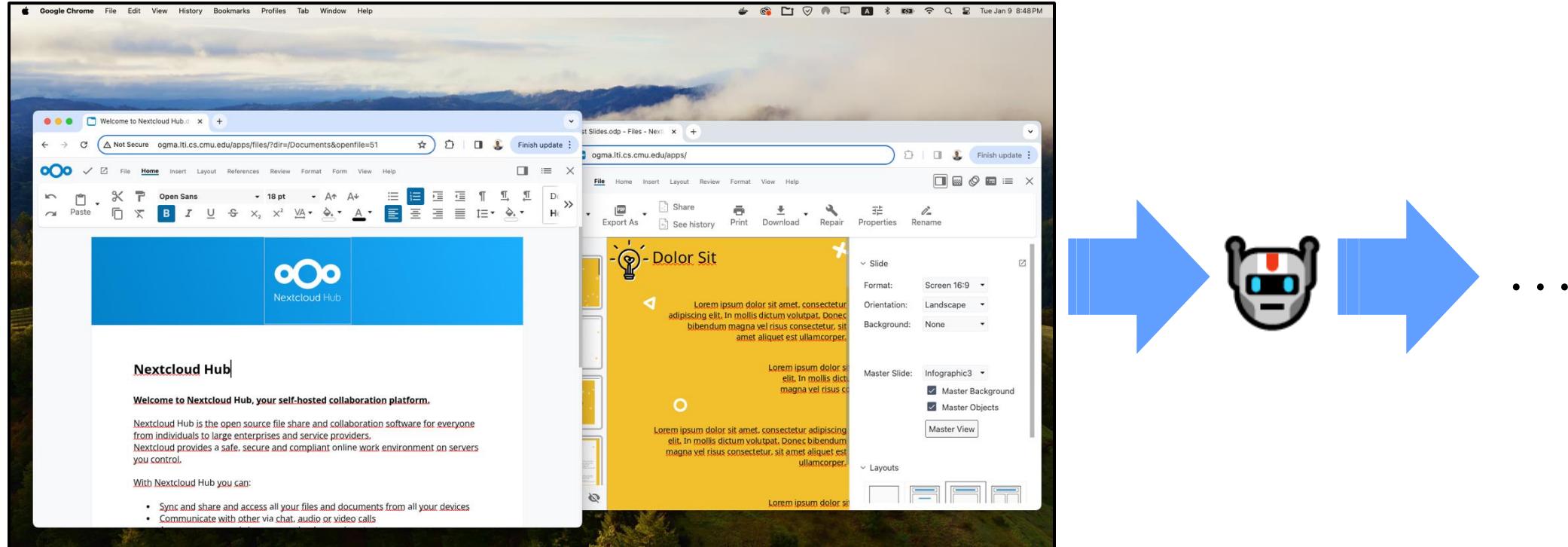
**Subnets (9)**  
Subnets within this VPC

**Route tables (8)**  
Route network traffic to resources

**Network connections (3)**  
Connections to other networks

**Introducing the VPC resource map**  
The new resource map helps you visualize the resources in your VPC. It shows your VPC, subnets, route tables, internet gateways, NAT gateways,

# Autonomous Agents



**Task:** “Create a set of PowerPoint slides to present the content in this paper.”

# Autonomous Agents

Training scores

File Edit View Insert Format Data Tools Extensions Help Last edit was seconds ago

A1:C17 Employee

|    | A                | B          | C     | D | E | F | G | H | I |
|----|------------------|------------|-------|---|---|---|---|---|---|
| 1  | Employee         | Department | Score |   |   |   |   |   |   |
| 2  | Bob Jones        | HR         | 89    |   |   |   |   |   |   |
| 3  | Sarah Smith      | Marketing  | 93    |   |   |   |   |   |   |
| 4  | Julia Kane       |            |       |   |   |   |   |   |   |
| 5  | Christina Graham |            |       |   |   |   |   |   |   |
| 6  | Mike Beck        |            |       |   |   |   |   |   |   |
| 7  | Alison Adams     |            |       |   |   |   |   |   |   |
| 8  | Josh White       |            |       |   |   |   |   |   |   |
| 9  | Zoey Clark       |            |       |   |   |   |   |   |   |
| 10 | Robert Jackson   |            |       |   |   |   |   |   |   |
| 11 | Sam Johnson      |            |       |   |   |   |   |   |   |
| 12 | Mary Brown       |            |       |   |   |   |   |   |   |
| 13 | Chris Williams   |            |       |   |   |   |   |   |   |
| 14 | Emily Anderson   |            |       |   |   |   |   |   |   |
| 15 | John Lee         |            |       |   |   |   |   |   |   |
| 16 | Tina Thompson    |            |       |   |   |   |   |   |   |
| 17 | Katie Allen      |            |       |   |   |   |   |   |   |
| 18 |                  |            |       |   |   |   |   |   |   |
| 19 |                  |            |       |   |   |   |   |   |   |
| 20 |                  |            |       |   |   |   |   |   |   |
| 21 |                  |            |       |   |   |   |   |   |   |
| 22 |                  |            |       |   |   |   |   |   |   |
| 23 |                  |            |       |   |   |   |   |   |   |
| 24 |                  |            |       |   |   |   |   |   |   |
| 25 |                  |            |       |   |   |   |   |   |   |

Department and Score

Employee

Chart editor

Setup

Customize

Chart type

Column chart

Stacking

None

Data range

A1:C17

X-axis

Employee

Department

Aggregate

Series

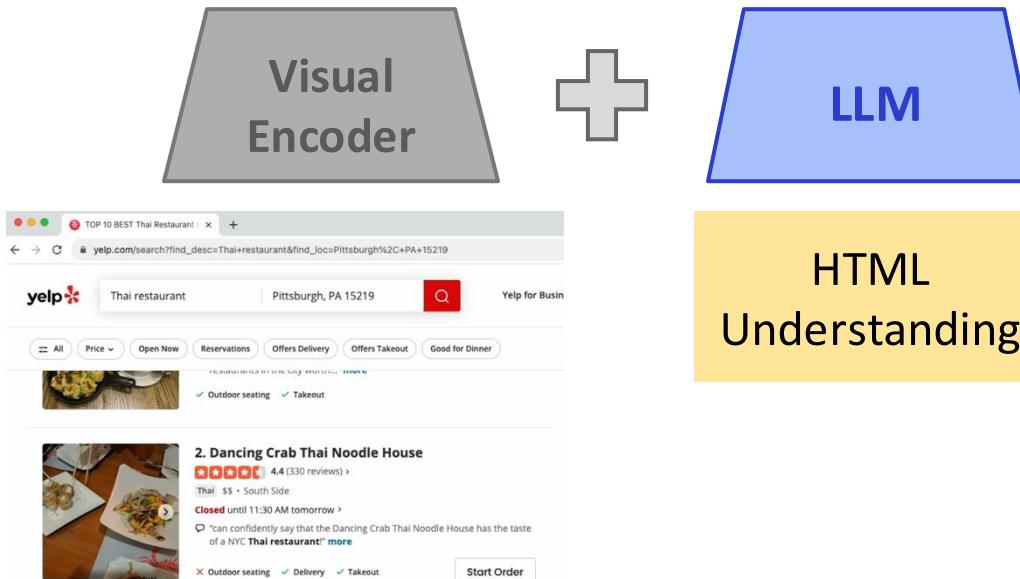
Sum: 1259

Explore

Sheet1

# Web Agents

Web  
Grounding



# Web Agents

## Web

Shunyu Yao, REACT Synergizing Reasoning and Acting in Language Models, 2023

Jason Wei et al, Chain of Thought Prompting Elicits Reasoning in Large Language Models, 2022

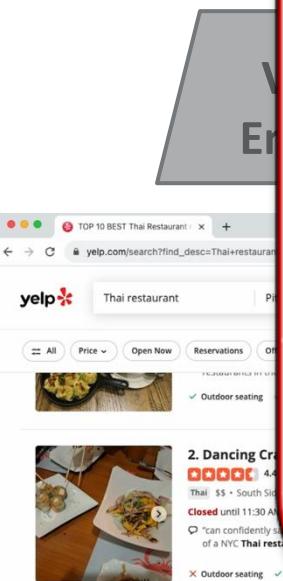
Reiichiro Nakano et al, WebGPT: Browser-assisted Question-Answering with Human Feedback, 2021.

Xiang Deng et al, MIND2WEB: Towards a Generalist Agent for the Web, 2023

Timo Schick et al, Toolformer: Language Models can Teach Themselves to Use Tools, 2023

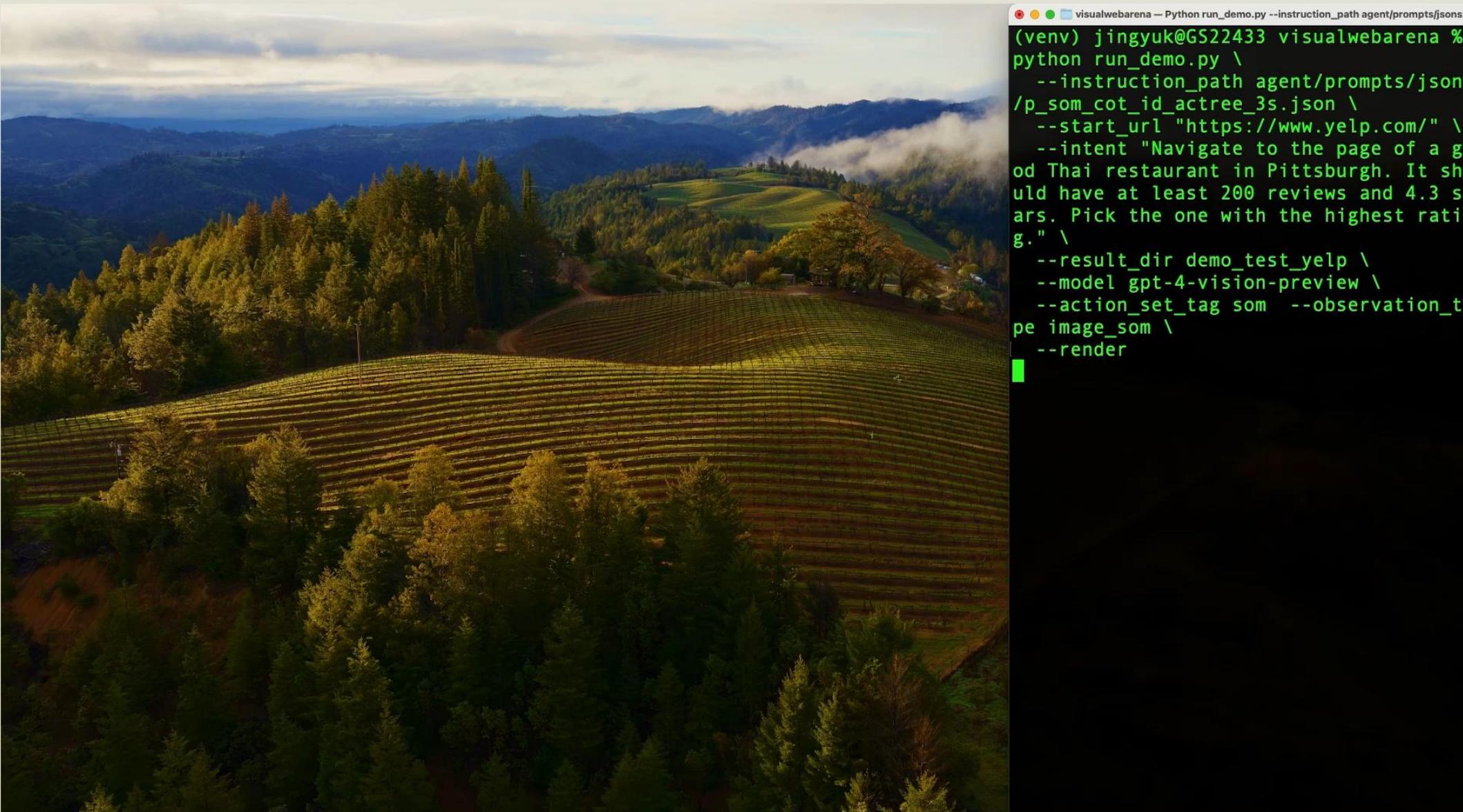
Shibo Hao et al, ToolkenGPT: Augmenting Frozen Language Models with Massive Tools via Tool Embeddings, 2023

Yang et al., SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering, 2024



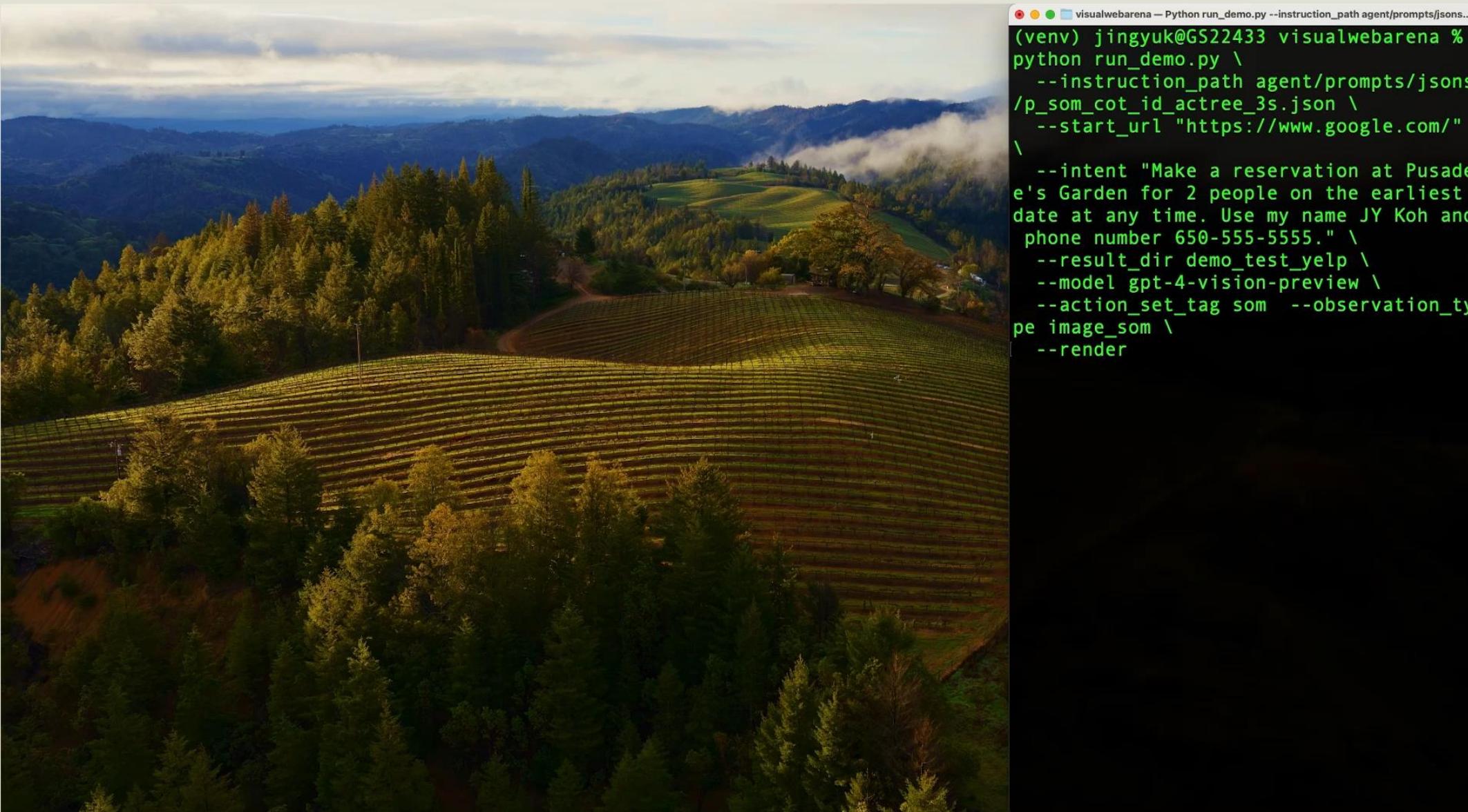
Task: Navigate to a page of a good Thai restaurant in Pittsburgh. It should have at least 200 reviews and 4.3 stars. Pick the one with the highest rating

**Task:** Navigate to the page of a good Thai restaurant in Pittsburgh. It should have at least 200 reviews and 4.3 stars. Pick the one with the highest rating.



```
visualwebarena — Python run_demo.py --instruction_path agent/prompts/jsons...
(venv) jingyuk@GS22433 visualwebarena %
python run_demo.py \
--instruction_path agent/prompts/jsons
/p_som_cot_id_actree_3s.json \
--start_url "https://www.yelp.com/" \
--intent "Navigate to the page of a go
od Thai restaurant in Pittsburgh. It sho
uld have at least 200 reviews and 4.3 st
ars. Pick the one with the highest ratin
g." \
--result_dir demo_test_yelp \
--model gpt-4-vision-preview \
--action_set_tag som --observation_ty
pe image_som \
--render
```

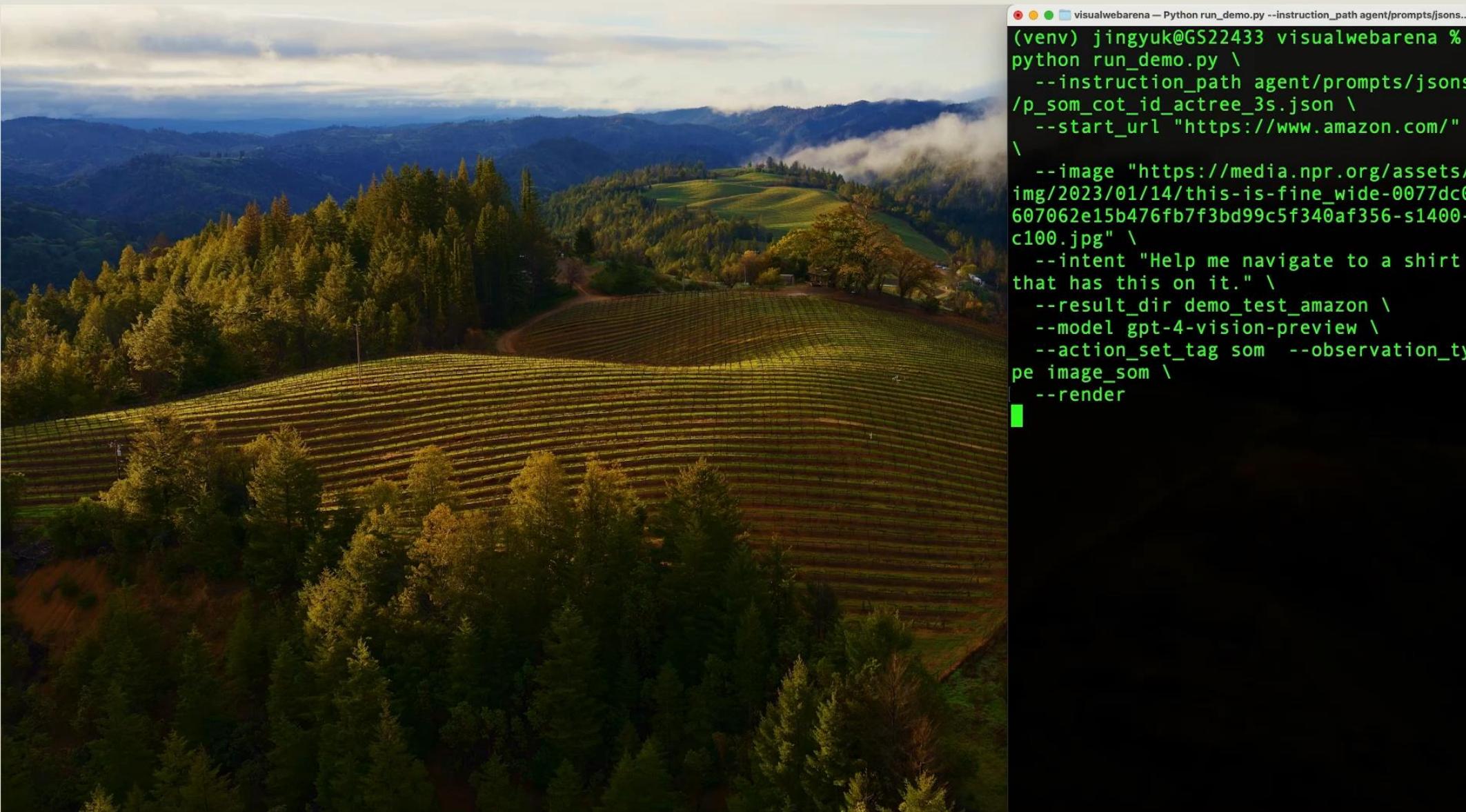
Task: Make a reservation at Pusadee's Garden for 2 people on the earliest date for dinner. Use my name JY Koh and phone number 650-555-5555.



```
visualwebarena — Python run_demo.py --instruction_path agent/prompts/jsons...
(venv) jingyuk@GS22433 visualwebarena %
python run_demo.py \
--instruction_path agent/prompts/jsons
/p_som_cot_id_actree_3s.json \
--start_url "https://www.google.com/"
\
--intent "Make a reservation at Pusade
e's Garden for 2 people on the earliest
date at any time. Use my name JY Koh and
phone number 650-555-5555." \
--result_dir demo_test_yelp \
--model gpt-4-vision-preview \
--action_set_tag som --observation_ty
pe image_som \
--render
```



Task: Help me navigate to a shirt that has this on it.



```
visualwebarena — Python run_demo.py --instruction_path agent/prompts/jsons...
(venv) jingyuk@GS22433 visualwebarena %
python run_demo.py \
--instruction_path agent/prompts/jsons
/p_som_cot_id_actree_3s.json \
--start_url "https://www.amazon.com/" \
\
--image "https://media.npr.org/assets/
img/2023/01/14/this-is-fine_wide-0077dc0
607062e15b476fb7f3bd99c5f340af356-s1400-
c100.jpg" \
--intent "Help me navigate to a shirt
that has this on it." \
--result_dir demo_test_amazon \
--model gpt-4-vision-preview \
--action_set_tag som --observation_ty
pe image_som \
--render
```

# WebArena

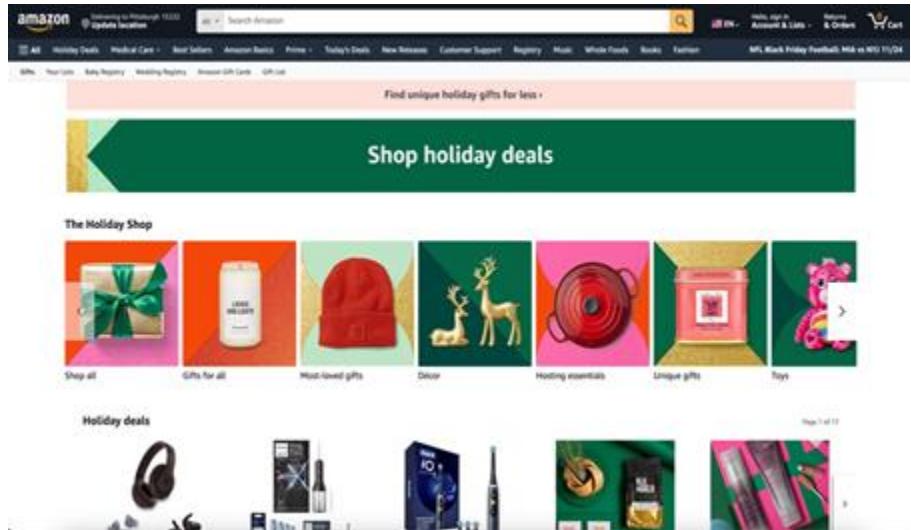


Shuyan Zhou

Frank Xu

- Most realistic web environment at the moment
- Websites from popular categories (shopping, Reddit, GitLab)
  - Self-hosted open source re-implementations
  - Data from real websites (Amazon, Reddit, GitHub)
- Tasks are easy for humans (78% success rate) but difficult for language model agents (14%)
- **But:** Tasks are designed to use just text and HTML source code
- Messy HTML, JavaScript: usually minified or compressed for efficiency
- Interactive elements don't display correctly in HTML
  - e.g., JavaScript/CSS code that moves objects after the page is loaded
- Context length: HTML pages are complex, easily filling up > 100k tokens

# HTML is insufficient



```

<head>
  <script type="text/javascript">
    var ue_08 = new Date();
    ue_08.setUTCFullYear(ue_08.getFullYear() - 1);
    ue_08.setUTCHours(0);
    ue_08.setUTCMinutes(0);
    ue_08.setUTCSeconds(0);
    ue_08.setUTCMilliseconds(0);
    document.cookie = "lastseen=" + ue_08.toUTCString();
  </script>
</head>
<body>
  <script type="text/javascript">
    var ue_09 = new Date();
    ue_09.setUTCFullYear(ue_09.getFullYear() - 1);
    ue_09.setUTCHours(0);
    ue_09.setUTCMinutes(0);
    ue_09.setUTCSeconds(0);
    ue_09.setUTCMilliseconds(0);
    document.cookie = "lastseen=" + ue_09.toUTCString();
  </script>
</body>
<script type="text/javascript">
  var ue_10 = new Date();
  ue_10.setUTCFullYear(ue_10.getFullYear() - 1);
  ue_10.setUTCHours(0);
  ue_10.setUTCMinutes(0);
  ue_10.setUTCSeconds(0);
  ue_10.setUTCMilliseconds(0);
  document.cookie = "lastseen=" + ue_10.toUTCString();
</script>
</html>

```

- Messy HTML, JavaScript: usually minified or compressed for efficiency
  - Interactive elements don't display correctly in HTML
    - e.g., JavaScript/CSS code that moves objects after the page is loaded
    - Spatial layout is also usually not conveyed well
  - Context length: HTML pages are complex, easily filling up > 100k tokens

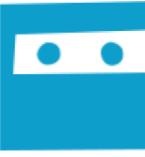
# VisualWebArena

Jing Yu  
Koh

Shuyan Zhou

Frank Xu

- Build and track the progress of **multimodal agents**
- We design visually grounded tasks to test these abilities
- Visual inputs (and outputs) allow for unique, interesting, and realistic tasks

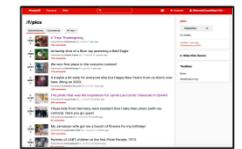
 OneStopShop
 reddit
 OsClass


Knowledge Resources + Tools

VisualWebArena Sites



“Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).”



“Navigate to the comments section of the latest image post in the /r/fArt subreddit that contains animals.”



“Help me make a post selling this item and navigate to it. Price it at \$10 cheaper than the most similar item on the site.”

→

→

click [1602]

LLM / VLM Agent

# VisualWebArena: Classifields



**Task:** Find this exact bike that's listed for \$300-500 and post a comment offering \$10 less than their asking price.

OsClass

My account Logout Publish Ad

## What are you looking for today?

Keyword  Category

Latest Listings

|   |  |   |   |
|---|--|---|---|
| <br>Nintendo Switch<br>270.00 \$ | <br>JBL Powered PA Speaker ...<br>150.00 \$ | <br>xbox series x / with extras<br>350.00 \$ | <br>Canon EF 100-400mm f/4....<br>1645.00 \$ |
|                                |   |    |    |

# VisualWebArena: Shopping



**Task:** Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).

My Account My Wish List Sign Out Welcome to One Stop Market

Search entire store here...  Advanced Search

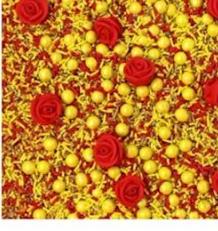
**One Stop Market**

Beauty & Personal Care - Sports & Outdoors - Clothing, Shoes & Jewelry - Home & Kitchen - Office Products - Tools & Home Improvement -

Health & Household - Patio, Lawn & Garden - Electronics - Cell Phones & Accessories - Video Games - Grocery & Gourmet Food -

One Stop Market

Product Showcases

|   |   |  |  |   |
|---|---|--|--|---|
| <br>Pre-baked Gingerbread House Kit Value Pack, 17 oz., Pack of 2, Total 34 oz.<br>★★★★★ 1 Review<br>\$19.99<br><a href="#">Add to Cart</a> | <br>V8 +Energy, Healthy Energy Drink, Steady Energy from Black and Green Tea, Pomegranate Blueberry, 8 Ounce Can ,Pack of 24<br>★★★★★ 12 Reviews<br>\$14.47<br><a href="#">Add to Cart</a> | <br>Elmwood Inn Fine Teas, Orange Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch<br>★★★★★ 4 Reviews<br>\$19.36<br><a href="#">Add to Cart</a> | <br>Belle Of The Ball Princess Sprinkle Mix  Wedding Colorful Sprinkles  Cake Cupcake Cookie Sprinkles  Ice cream Candy Sprinkles  Yellow Gold Red Royal Red Rose Icing Flowers Decorating Sprinkles, 8OZ<br>★★★★★ 12 Reviews<br>\$23.50<br><a href="#">Add to Cart</a> | <br>So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project Verified, 9 oz. Tub<br>★★★★★ 12 Reviews<br>\$15.62<br><a href="#">Add to Cart</a> |
|---|---|--|--|---|

# VisualWebArena: Reddit



**Task:** What is the 2022 total nominal GDP of the area that produces most sugarcane in the year of 2021? (in billion)?

[OC] Sugarcane was first introduced to Brazil in 1532. Half a millennium later, the country produces over 700M tonnes yearly (roughly the same amount as all of Asia, and 7x the amount produced by Africa)

Submitted by latinometrics [t3\_10z0y3g] 11 months ago in dataisbeautiful

1,163 points (+1163, -0)

Short URL:  
<http://ec2-3-13-232-171.us-east-2.compute.amazonaws.com:9999/f/dataisbeautiful/103854>

dataisbeautiful

t5\_2tk95

Created 1 year ago

Subscribe via RSS

Toolbox

Bans

Moderation log

Comments

You must log in or register to comment.

# VisualWebArena

POMDP environment:  $\mathcal{E} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T} \rangle$

- Observations  $\mathcal{O}$

The figure displays three screenshots of a web browser window for 'webarena.onestopshop.com'. The first screenshot shows the main 'Patio, Lawn & Garden' page with various filters and a search bar. The second screenshot shows a detailed product page for an 'Outdoor Patio ...' set, featuring an image, a price of \$49.99, a rating of 82%, and 12 reviews. The third screenshot shows the raw HTML code for the same product listing, highlighting the structure of the web page.

- Actions  $\mathcal{A}$

| 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 optional output. |

- Deterministic transition function

$$\mathcal{T} : \mathcal{S} \times \mathcal{A} \longrightarrow \mathcal{S}$$

- Reward function:  $r(\mathbf{a}, \mathbf{s})$

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: ...

# Image Inputs:



**Task:** “Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).”

## One Stop Market

Beauty & Personal Care · Sports & Outdoors · Clothing, Shoes & Jewelry · Home & Kitchen · Office Products · Tools & Home Improvement ·

Health & Household · Patio, Lawn & Garden · Electronics · Cell Phones & Accessories · Video Games · Grocery & Gourmet Food ·

Home > Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier

### Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier

IN STOCK SKU B005TI2Q50

Be the first to review this product

\$2.56

Qty

1

Add to Cart

Add to Wish List Add to Compare



## Shopping



**Task: “Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).”**

My Account My Wish List Sign Out Welcome, Emma Lopez!

Search entire store here... Advanced Search

**One Stop Market**

Beauty & Personal Care Sports & Outdoors Clothing, Shoes & Jewelry Home & Kitchen Office Products Tools & Home Improvement

Health & Household Patio, Lawn & Garden Electronics Cell Phones & Accessories Video Games Grocery & Gourmet Food

One Stop Market

Product Showcases

Pre-baked Gingerbread House Kit  
Value Pack, 17 oz., Pack of 2, Total 34 oz.  
 1 Review  
\$19.99

V8 +Energy, Healthy Energy Drink, Steady Energy from Black and Green Tea, Pomegranate Blueberry, 8 Ounce Can, Pack of 24  
 12 Reviews  
\$14.47

Elmwood Inn Fine Teas, Orange Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch  
 4 Reviews  
\$19.36

Beile Of The Ball Princess Sprinkle Mix | Wedding Colorful Sprinkles | Cake Cupcake Cookie Sprinkles | Ice cream Candy Sprinkles | Yellow Gold Red Royal Red Rose Icing Flowers Decorating Sprinkles, 8OZ.  
 12 Reviews  
\$15.62

So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project Verified, 9 oz. Tub  
 12 Reviews  
\$15.62

My Account My Wish List Sign Out Welcome, Emma Lopez!

Search entire store here... Advanced Search

**One Stop Market**

Beauty & Personal Care Sports & Outdoors Clothing, Shoes & Jewelry Home & Kitchen **Office Products** Tools & Home Improvement

Health & Household Patio, Lawn & Garden Electronics Cell Phones & Accessories Video Games Grocery & Gourmet Food

Home > Office Products > Office Electronics > Printers & Accessories

Printers & Accessories

Shop By Items 1-12 of 331 Sort By Position

Shopping Options

Price  
\$0.00 - \$9,999.99 (330)  
\$10,000.00 and above (1)

Compare Products  
You have no items to compare.

Recently Ordered

- Nintendo Joy-Con (L/R) Fortnite Fleet Force Bundle - Nintendo Switch
- OEM HTC USB Travel Charger Adapter U250 / CRH6300 / 59H00095-14M
- MacBook Charger Case

Epson WorkForce WF-3620 WiFi Direct All-in-One Color Inkjet Printer, Copier, Scanner, Amazon Dash Replenishment Ready

Digital Check TS240 Check Scanner - 50 DPM, No Inkjet Printer (Renewed)

Canon All-in-One Color Inkjet Wired Printer, Print Scan Copy for Home Office, up to 60 Sheets, 600 x 1200 dpi, Portability, Lightweight, PIXMA

HP M225DW Mono Laserjet Pro MFP

**Step 0:** Start on the homepage of OneStopMarket.

**Step 1:** Navigate to the printers category.



**Task:** “Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).”

**Printers & Accessories**

Shop By Items 1-12 of 331

Sort By Price ↑

| Image | Name   | Price  | Actions     |
|-------|--|--------|-------------|
|       | MUNBYN USB Upgrade Label Printer, Thermal Printer for Barcodes-Labels Labeling with MUNBYN Thermal Direct Shipping Label (Pack of 500 4x6 Per Roll Labels) | \$2.56 | Add to Cart |
|       | Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier   | \$4.16 | Add to Cart |
|       | WPFYI XP-C300H High Speed 300mm/s Printing Speed 80mm USB POS Receipt Printer Support Wall Hanging   | \$6.06 | Add to Cart |

**Step 2:** Sort by descending price.

**One Stop Market**

Search entire store here... Advanced Search

Beauty & Personal Care · Sports & Outdoors · Clothing, Shoes & Jewelry · Home & Kitchen · **Office Products** · Tools & Home Improvement ·

Health & Household · Patio, Lawn & Garden · Electronics · Cell Phones & Accessories · Video Games · Grocery & Gourmet Food ·

Home > Office Products > Office Electronics > Printers & Accessories > Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier

**Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier**

IN STOCK SKU: BOOST12Q60 Be the first to review this product.

**\$2.56**

Qty  Add to Cart

Add to Wish List Add to Compare

**Step 3:** Click on the cheapest color photo printer.



**Task:** “Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).”

**One Stop Market**

Search entire store here... Advanced Search

Beauty & Personal Care · Sports & Outdoors · Clothing, Shoes & Jewelry · Home & Kitchen · Office Products · Tools & Home Improvement ·

Health & Household · Patio, Lawn & Garden · Electronics · Cell Phones & Accessories · Video Games · Grocery & Gourmet Food ·

**Shopping Cart**

| Item   | Price  | Qty | Subtotal |
|--|--------|-----|----------|
| Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier | \$2.56 | 1   | \$2.56   |

Move to Wishlist Edit Remove Item

< Continue Shopping Update Shopping Cart

Privacy and Cookie Policy  Enter your email address Subscribe

This screenshot shows the One Stop Market shopping cart page. It displays a single item: a Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier, priced at \$2.56. The quantity is set to 1, and the subtotal is also \$2.56. Below the cart, there are buttons for 'Move to Wishlist', 'Edit', and 'Remove Item'. At the bottom, there are links for 'Continue Shopping' and 'Update Shopping Cart', along with a newsletter sign-up form.

**Step 4:** Add it to the shopping cart.

**One Stop Market**

Shipping Review & Payments

**Shipping Address**

Emma Lopez  
101 S San Mateo Dr  
San Mateo, California 94010  
United States  
650551212

+ New Address

**Order Summary**  
1 Item in Cart

**Shipping Methods**

\$5.00 Fixed Flat Rate Next

This screenshot shows the One Stop Market checkout process at the shipping address step. The shipping address is listed: Emma Lopez, 101 S San Mateo Dr, San Mateo, California 94010, United States, 650551212. A red box highlights this address information. Below the address, there is a '+ New Address' button. To the right, an 'Order Summary' section shows '1 Item in Cart'. Further down, the 'Shipping Methods' section offers '\$5.00', 'Fixed', and 'Flat Rate' options. A blue 'Next' button is located at the bottom right.

**Step 5:** Proceed to checkout



**Task:** “Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).”

**One Stop Market**

Shipping

**Shipping Address**

First Name \*

Last Name \*

Company

Street Address \*

Country \*

State/Province \*

City \*

Zip/Postal Code \*

Phone Number \*

Save in address book

[Cancel](#) [Ship Here](#)

**One Stop Market**

[My Account](#) [My Wish List](#) [Logout](#) [EVERGREEN EXTRAS SUPPORT](#)

[Advanced Search](#)

[Print receipt](#)

Beauty & Personal Care · Sports & Outdoors · Clothing, Shoes & Jewelry · Home & Kitchen · Office Products · Tools & Home Improvement ·

Health & Household · Patio, Lawn & Garden · Electronics · Cell Phones & Accessories · Video Games · Grocery & Gourmet Food ·

**Thank you for your purchase!**

Your order number is: **000000198**.

We'll email you an order confirmation with details and tracking info.

[Continue Shopping](#)

[Privacy and Cookie Policy](#) [Search Terms](#) [Advanced Search](#) [Contact Us](#)

[Subscribe](#)

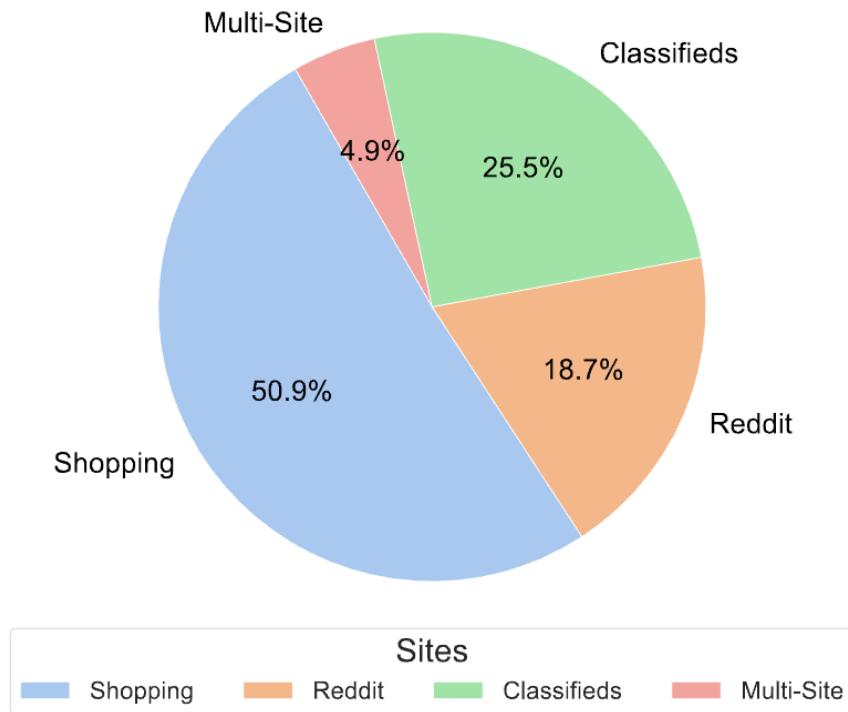
Copyright © 2013 OneStopMarket, Inc. All rights reserved.

**Step 6:** Edit address to that of Emily's place.

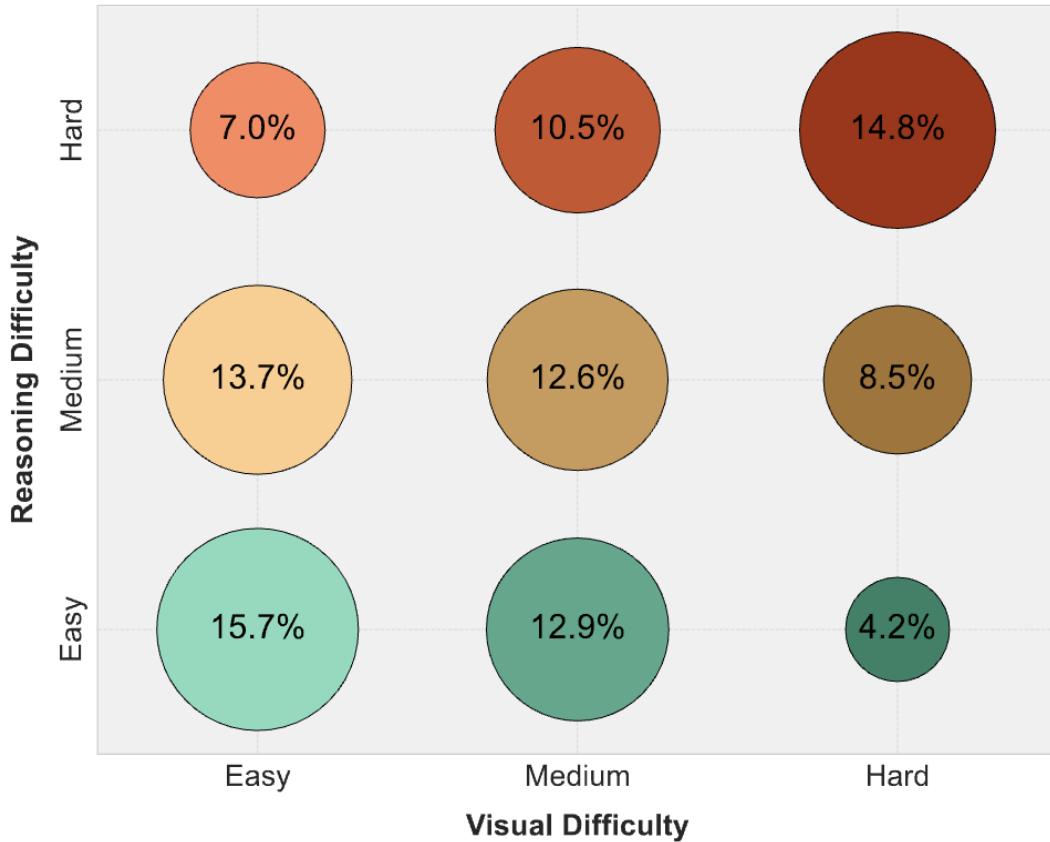
**Step 7:** Place the order

# VisualWebArena

Distribution of Tasks Across Sites



Distribution of Tasks by Difficulty



# Execution Based Evaluation

| Webpage / Input Image(s)   | Example Intent  | Reward Function $r(s, a)$ Implementation  |
|--|---|---|
|   | 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. | exact_match( $\hat{a}$ , "US0378331005")  |
|   | Add something like what the man is wearing to my wish list.   | <pre>url="/wishlist" locator(".wishlist .product-image-photo") eval_vqa(s, "Is this a polo shirt? (yes/no)", "yes") eval_vqa(s, "Is this shirt green? (yes/no)", "yes")</pre> |
|   | Create a post for each of the following images in the most related forums.  | eval_fuzzy_image_match( $s, a^*$ )  |
|  | Navigate to my listing of the white car and change the price to \$25000. Update the price in the description as well.   | <pre>url="/index.php?page=item&amp;id=84144" must_include(<math>\hat{a}</math>, "\$25000  OR  \$25,000") must_exclude(<math>\hat{a}</math>, "\$30000  OR  \$30,000")</pre>    |

# LLM and VLM Agents

# Visual Language Models as Agents

```
Tab 0 (current): Search results for: 'hp inkjet'

[1] RootWebArea "Search results for: 'hp inkjet'" focused: True
  [81] link 'My Account'
  [82] link 'My Wish List'
  [83] link 'Sign Out'
  [4086] StaticText 'Welcome to One Stop Market'
  [37] link 'Skip to Content'
  [23] link 'store logo'
    [39] img 'one_stop_market_logo'
  [40] link '\ue611 My Cart'
  [278] StaticText 'Search'
  [163] combobox 'ue615 Search' autocomplete: both hasPopup: listbox required: False expanded: False
    [426] StaticText 'hp inkjet'
  [281] link 'Advanced Search'
  [120] button 'Search' disabled: True
  [4080] tablist '' multiselectable: False orientation: horizontal
    [4082] tabpanel ''
      [2326] menu '' orientation: vertical
        [3077] menuitem 'ue622 Beauty & Personal Care' hasPopup: menu
        [3142] menuitem 'ue622 Sports & Outdoors' hasPopup: menu
        [3152] menuitem 'ue622 Clothing, Shoes & Jewelry' hasPopup: menu
        [3166] menuitem 'ue622 Home & Kitchen' hasPopup: menu
        [3203] menuitem 'ue622 Office Products' hasPopup: menu
        [3211] menuitem 'ue622 Tools & Home Improvement' hasPopup: menu
        [3216] menuitem 'ue622 Health & Household' hasPopup: menu
        [3222] menuitem 'ue622 Patio, Lawn & Garden' hasPopup: menu
        [3227] menuitem 'ue622 Electronics' hasPopup: menu
        [3288] menuitem 'ue622 Cell Phones & Accessories' hasPopup: menu
        [3303] menuitem 'ue622 Video Games' hasPopup: menu
        [3316] menuitem 'ue622 Grocery & Gourmet Food' hasPopup: menu
  [47] link 'Home'
  [12] main ''
    [32] heading "Search results for: 'hp inkjet'"
    [264] StaticText 'View as'
    [146] strong 'Grid'
    [147] link 'View as \ue0b List'
    [148] StaticText 'Items'
    [151] StaticText '-'
    [153] StaticText '12'
    [154] StaticText 'of '
    [156] StaticText '687'
    [269] StaticText 'Sort By'
    [158] combobox 'Sort By' hasPopup: menu expanded: False
    [159] link '\ue614 Set Ascending Direction'
    [424] link 'Image'
      [1010] img 'Image'
    [1011] link 'HP Business Inkjet 2800 Wide Format Printer (C8174A#A2L)'
    [720] LayoutTable ''
      [1451] StaticText 'Rating:'
      [1232] generic '47'
      [1869] link '12 \xa0Reviews'
    [1871] StaticText '$37.64'
    [1600] link 'Image'
      [17cc1] img 'Image'

Tab 1 (background): Welcome to One Stop Market
  [1] My Account
  [2] My Wish List
  [3] Sign In
  [4] Create an Account
  [5] Welcome to One Stop Market
  [6] One Stop Market
  [7] Advanced Search
  [8] hp inkjet
  [9] Advanced Search
  [10] Search
  [11] Cart
  [12] Beauty & Personal Care
  [13] Sports & Outdoors
  [14] Clothing, Shoes & Jewelry
  [15] Home & Kitchen
  [16] Office Products
  [17] Tools & Home Improvement
  [18] Health & Household
  [19] Patio, Lawn & Garden
  [20] Electronics
  [21] Cell Phones & Accessories
  [22] Video Games
  [23] Grocery & Gourmet Food
  [24] Home > Search results for: 'hp inkjet'
```

## Accessibility tree / HTML representations:

Cluttered with unnecessary information, long and confusing context.



**VLM + SoM:** Simplified representation with Set-of-Marks (SoM) prompting over interactable elements.

# Visual Language Models as Agents

The screenshot shows a webpage with a red header bar containing 'Postmill', 'Forums', and 'Wiki' buttons. Below the header is a search bar and 'Log in' / 'Sign up' buttons. The main content area has a title '/f/food'. It displays a list of posts with small thumbnail images, titles, and interaction counts (upvotes and comments). A sidebar on the right is titled 'Toolbox' and includes 'Bans' and 'Moderation log' buttons.

Original Webpage

A large yellow arrow points from the 'Original Webpage' screenshot to this one. The sidebar on the right now includes a 'Hot' button above the 'Toolbox' section. The main content area is labeled 'Webpage with SoM of Interactable Elements' and contains a list of numbered elements corresponding to the visual components:

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

Below this is another label 'SoM Elements and Text Content'.



# Visual Language Models as Agents

User goal:



*I'm trying to find this post. Navigate to the comment section for it.*

Multimodal LLM

Observations

$o_t$ :

The screenshot shows a list of posts in the /r/food subreddit. Post 34 is highlighted, showing a picture of a pumpkin and a comment link. The interface includes a navigation bar with 'Postroll', 'Forums', 'Wiki', and a search bar. A sidebar on the right contains a 'Toolbox' with 'Bans' and 'Moderation log' options.

...

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

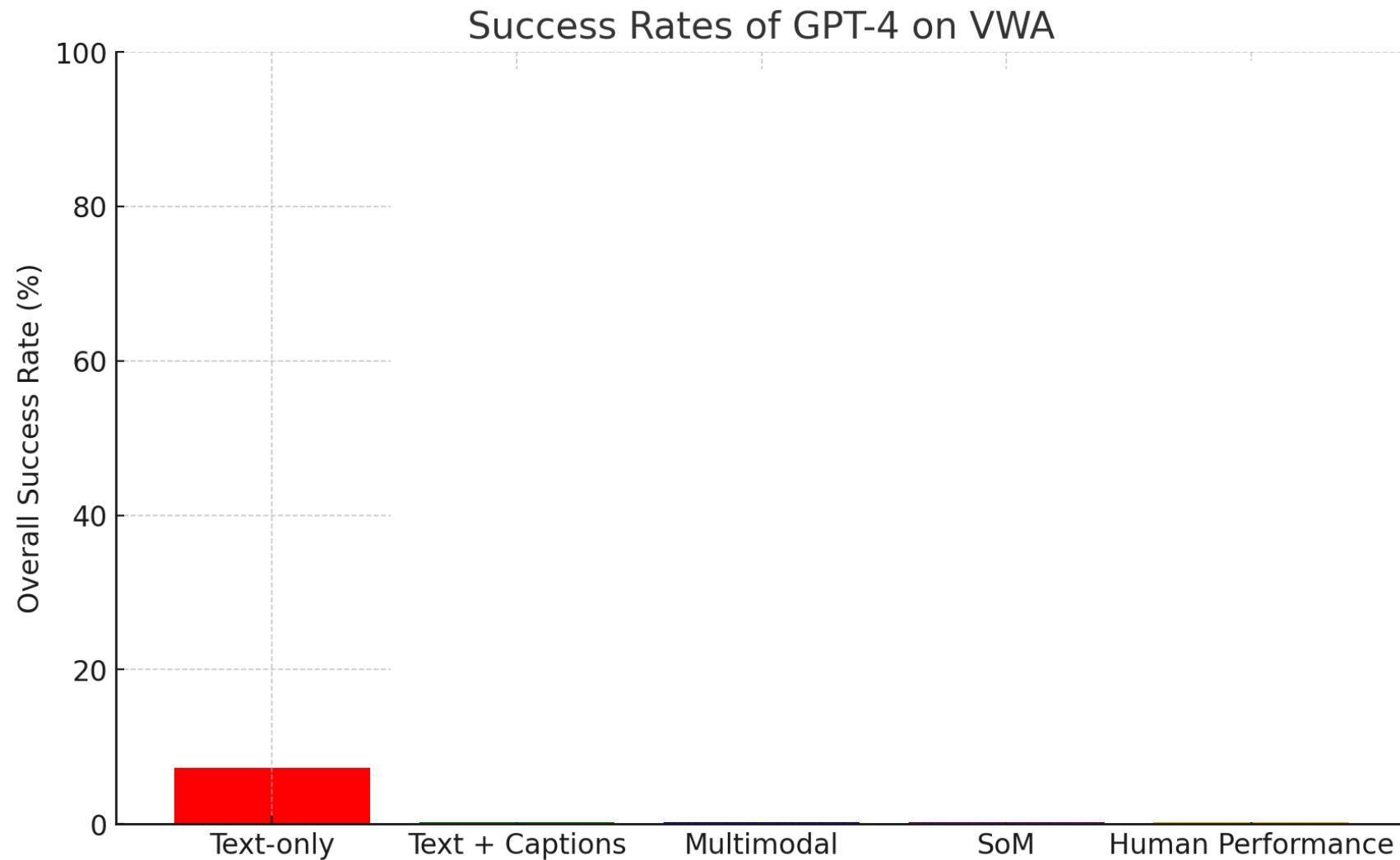
...

Let's think step-by-step... The objective is to navigate to the find the post and navigate to the comment section for it. From the observation, I can see... To navigate to this listing, I need to click on the comment link associated with the sushi. In summary, the next action I will perform is ``click [34]``

Action  $a_t$ : click [34]

**VLM + SoM:** Simplified representation with Set-of-Marks (SoM) prompting over interactable elements.

# Baseline Agents

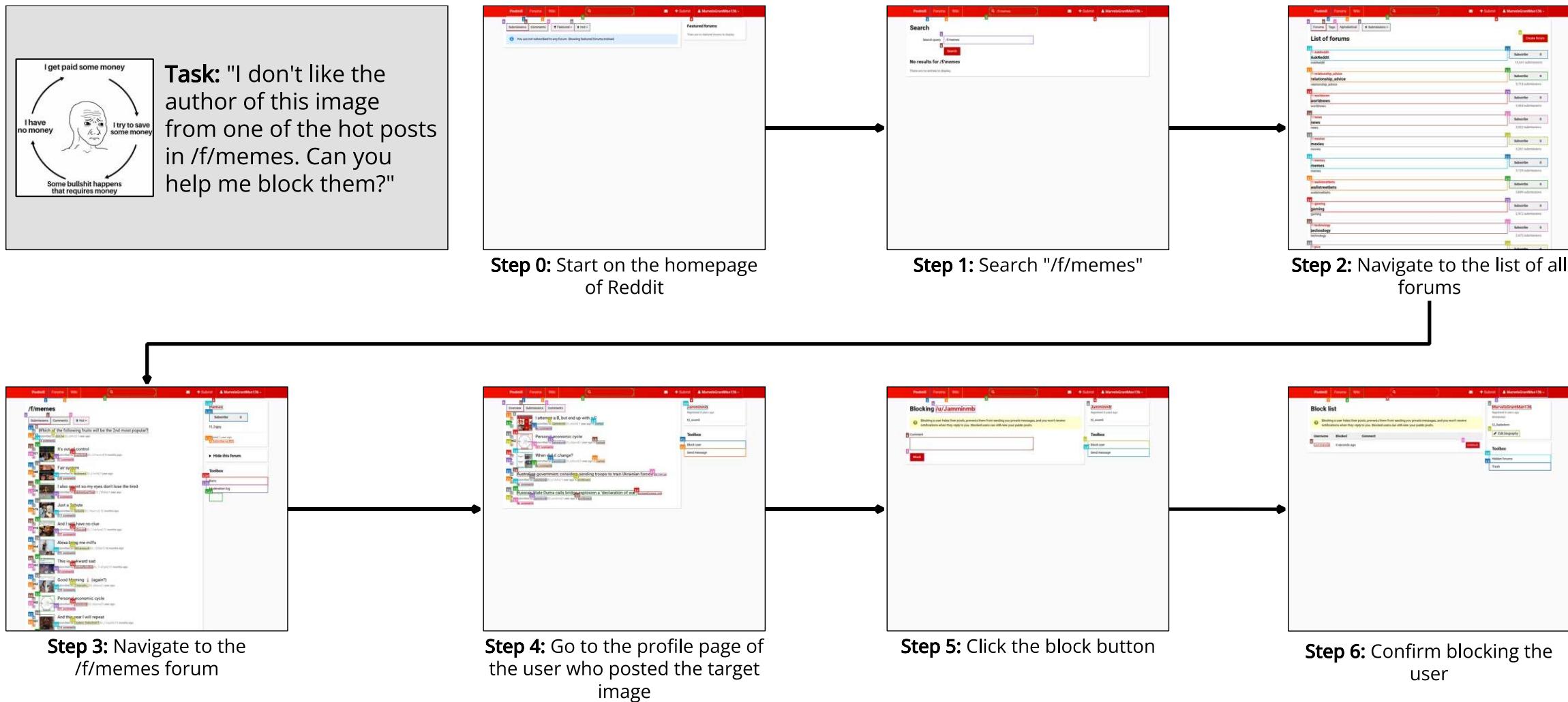


# Baseline Agents: Text-based LLMs

| Model Type        | LLM Backbone | Visual Backbone | Inputs                        | Success Rate (↑) |
|-------------------|--------------|-----------------|-------------------------------|------------------|
| Text-only         | LLaMA-2-70B  | -               | Accessibility Tree            | 1.10%            |
|                   | Mixtral-8x7B |                 |                               | 1.76%            |
|                   | Gemini-Pro   |                 |                               | 2.20%            |
|                   | GPT-3.5      |                 |                               | 2.20%            |
|                   | GPT-4        |                 |                               | 7.25%            |
| Caption-augmented | LLaMA-2-70B  | BLIP-2-T5XL     | Accessibility Tree + Captions | 0.66%            |
|                   | Mixtral-8x7B | BLIP-2-T5XL     |                               | 1.87%            |
|                   | GPT-3.5      | LLaVA-7B        |                               | 2.75%            |
|                   | GPT-3.5      | BLIP-2-T5XL     |                               | 2.97%            |
|                   | Gemini-Pro   | BLIP-2-T5XL     |                               | 3.85%            |
|                   | GPT-4        | BLIP-2-T5XL     |                               | 12.75%           |

# Baseline Agents: Multimodal LLMs

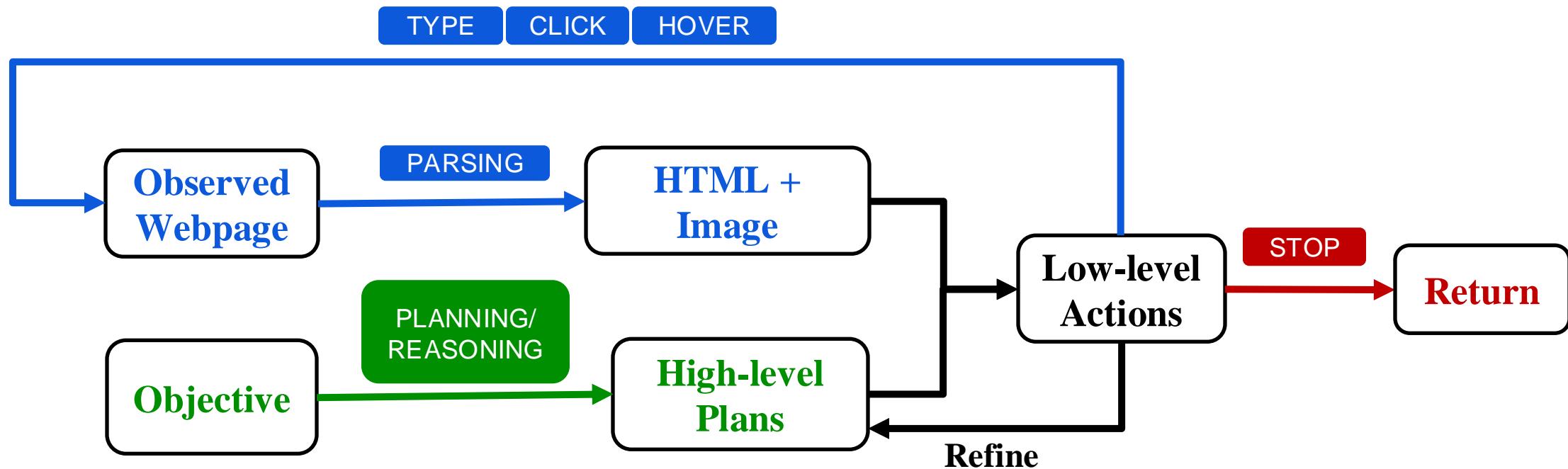
| Model Type        | Multimodal Model     | Inputs                                | Success Rate (↑) |
|-------------------|----------------------|---------------------------------------|------------------|
| Multimodal        | IDEFICS-80B-Instruct | Image + Captions + Accessibility Tree | 0.77%            |
|                   | CogVLM               |                                       | 0.33%            |
|                   | Gemini-Pro           |                                       | 6.04%            |
|                   | GPT-4V               |                                       | 15.05%           |
| Multimodal (SoM)  | IDEFICS-80B-Instruct | Image + Captions + SoM                | 0.99%            |
|                   | CogVLM               |                                       | 0.33%            |
|                   | Gemini-Pro           |                                       | 5.71%            |
|                   | GPT-4V               |                                       | 16.37%           |
| Human Performance | -                    | Webpage                               | 88.70%           |



Successful execution trajectory of the GPT-4V + SoM agent on the task for blocking a user that posted a certain picture

# Web Agent Architecture

- Model architecture of our interactive agent:
  - High-level Planning and Reasoning
  - Observation Parsing
  - Low-level Action Generation



# Planning

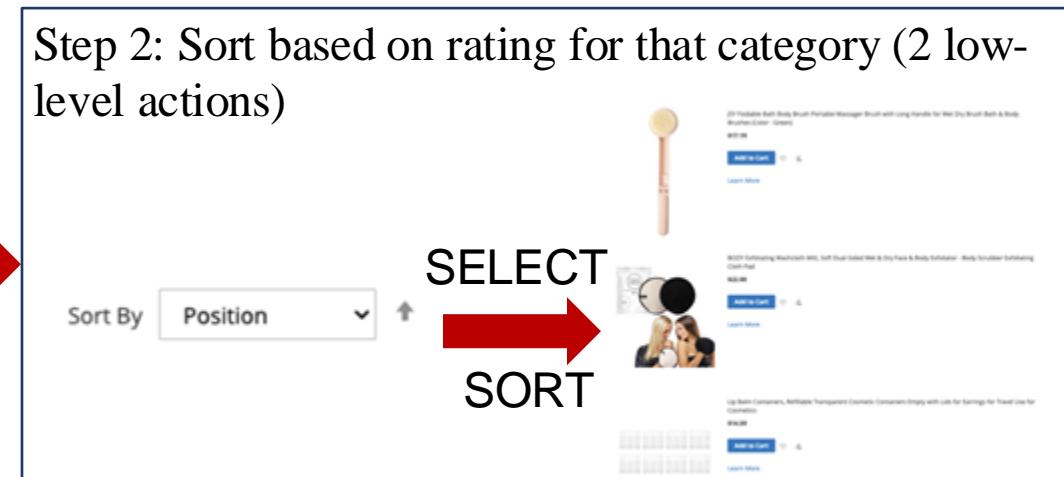
High-level plans are important for long-sequence and complex objectives.

Task: Buy the highest rated product from the Beauty & Personal Care category within a budget under 20.

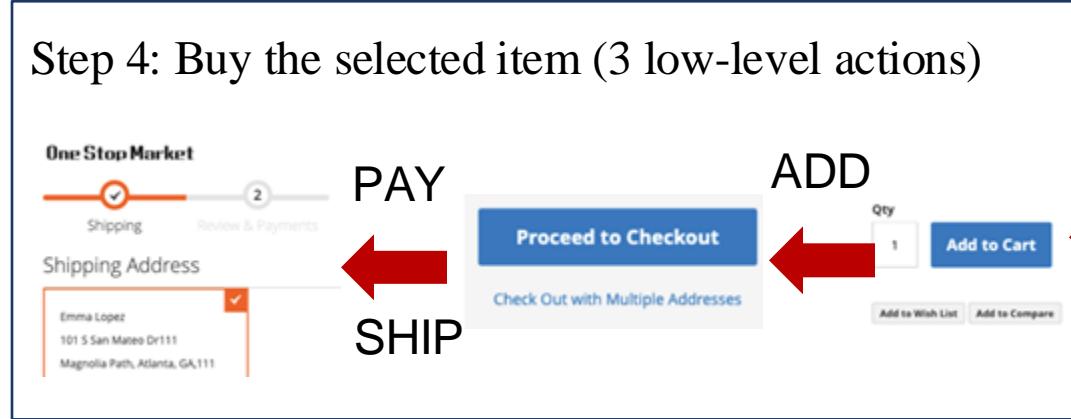
Step 1: Navigate to the Beauty & Personal Care Category (1 low-level action)



Step 2: Sort based on rating for that category (2 low-level actions)



Step 4: Buy the selected item (3 low-level actions)



Step 3: Select one item under 20 dollars (1 low-level action)

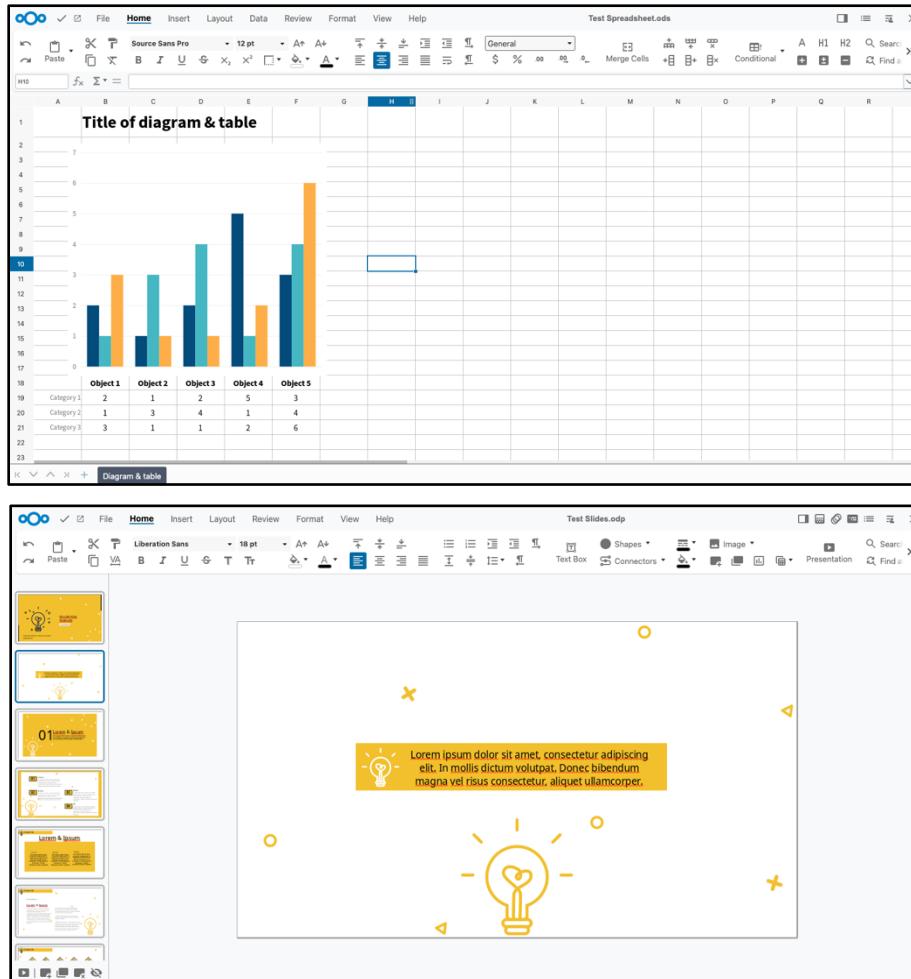


# Measuring Productive Tasks

VisualWebArena is a step towards building general purpose agents. But:

- Tasks are not very **consequential**: do not represent significant economic value
- Tasks are simpler, as current LLM agents do not even do well on these problems

**Long term:** Automate productive, economically valuable tasks



Examples from [Collabora Online](#) / LibreOffice.

# Common Failure Modes

- Long horizon reasoning and planning:
  - Models oscillate between two webpages, or get stuck in a loop
  - Correctly performing tasks but undoing them
  - Agents tend to stop exploration / execution too early

# What is Missing?

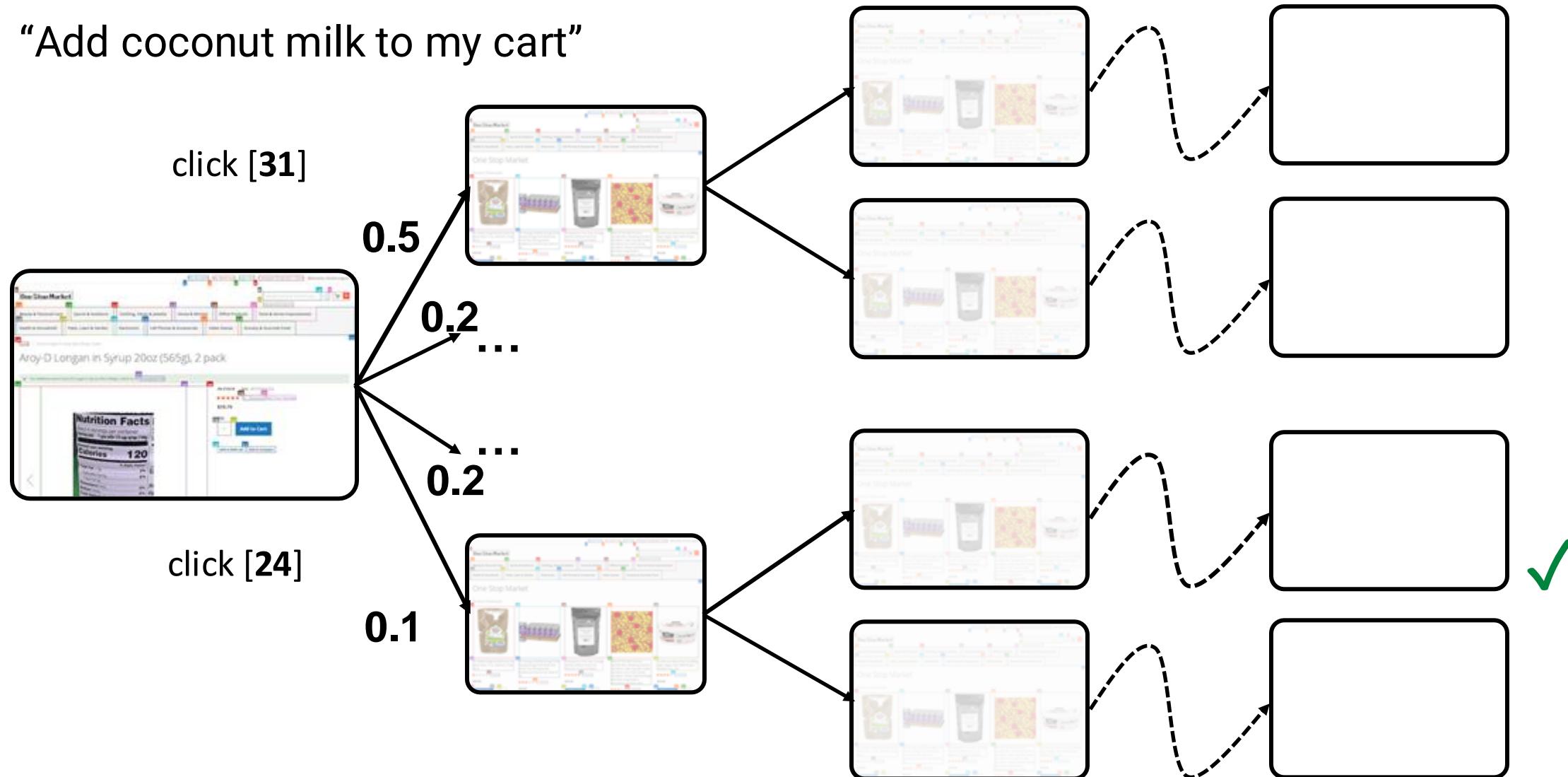
- We need to do a lot more to close the gap:
  - **Reasoning and Planning** over long horizons
  - Allow agent to **Search**, execute and coordinate multiple instances in parallel and ask for clarifications/confirmations
  - Strong vision-language-code models
  - Identifying the appropriate level of abstraction for agents (HTML/screenshots/APIs)
- **Multimodal models:** Many real-world tasks require visual grounding to effectively solve (e.g., every task involving PowerPoint, Excel, Photoshop). To develop strong general agents, we will need to train and build strong vision-language models.

# Exponential Error Compounding in Agents

| <b>Accuracy @ k steps:</b> |          |           |           |           |
|----------------------------|----------|-----------|-----------|-----------|
| <b>1 (single step)</b>     | <b>5</b> | <b>10</b> | <b>30</b> | <b>50</b> |
| 90%                        | 59.05%   | 34.87%    | 4.24%     | 0.52%     |
| 95%                        | 77.38%   | 59.87%    | 21.46%    | 7.69%     |
| 99%                        | 95.10%   | 90.44%    | 73.97%    | 60.50%    |
| 99.9%                      | 99.50%   | 99.00%    | 97.04%    | 95.12%    |
| 99.99%                     | 99.95%   | 99.90%    | 99.70%    | 99.50%    |

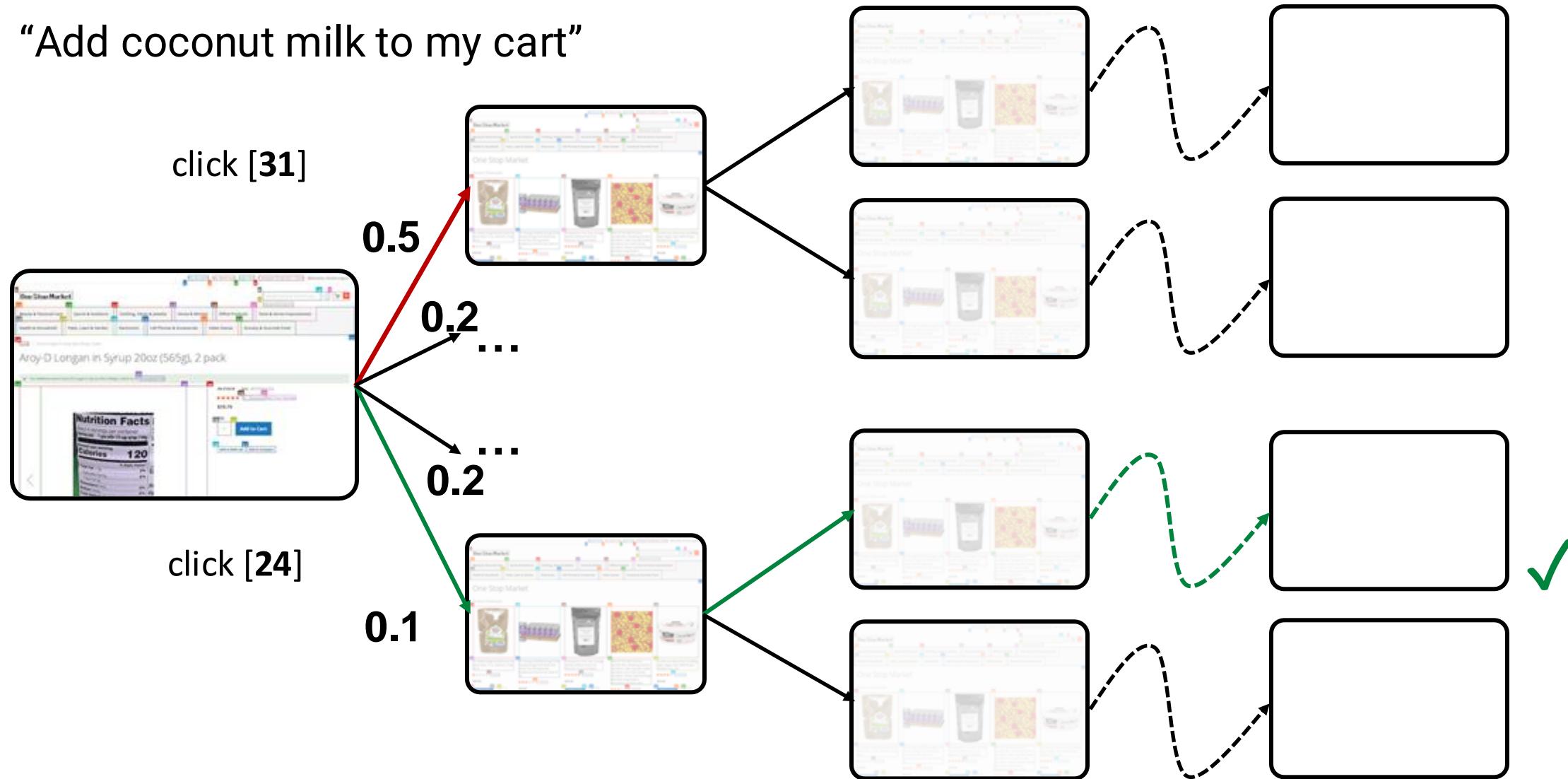
# Local Decisions; Global Consequences

"Add coconut milk to my cart"



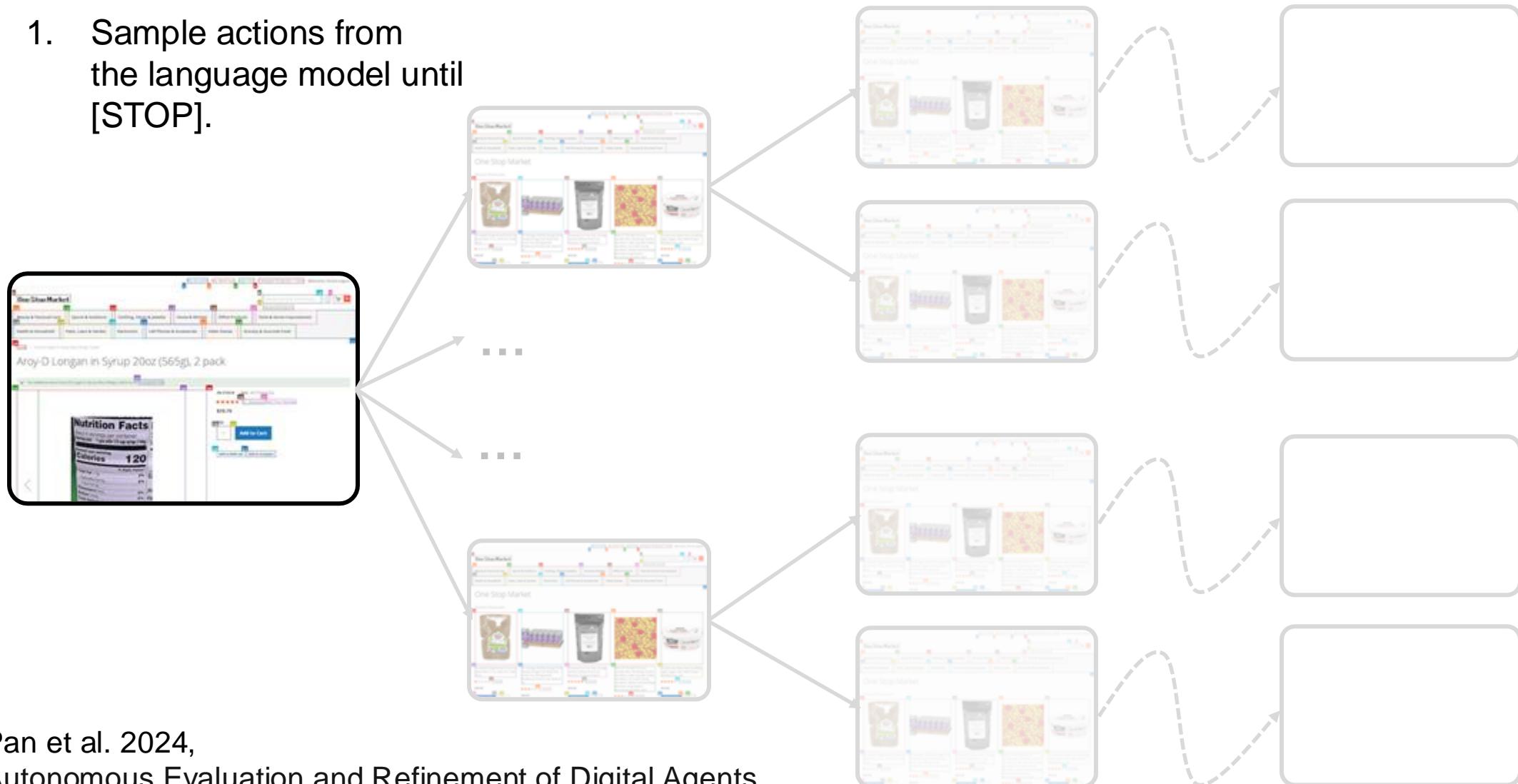
# Local Decisions; Global Consequences

"Add coconut milk to my cart"



# Search By Repeated Sampling

1. Sample actions from the language model until [STOP].



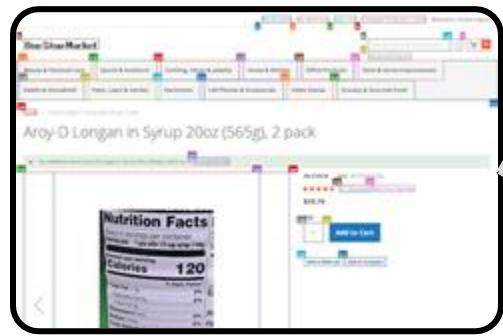
# Search By Repeated Sampling

1. Sample actions from the language model until [STOP].
2. Evaluate the resulting trajectory
3. Repeat?



# Search By Repeated Sampling

1. Sample actions from the language model until [STOP].
2. Evaluate the resulting trajectory
3. Repeat?

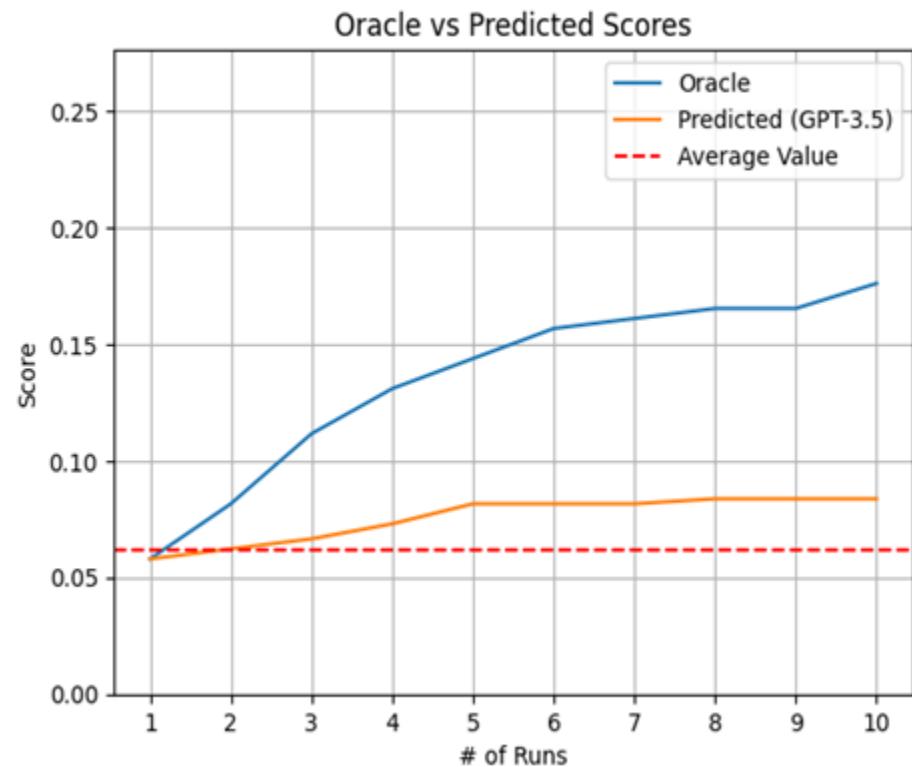


# Search By Repeated Sampling

1. Sample actions from the language model until [STOP].
2. Evaluate the resulting trajectory
3. Repeat?



# Search By Repeated Sampling



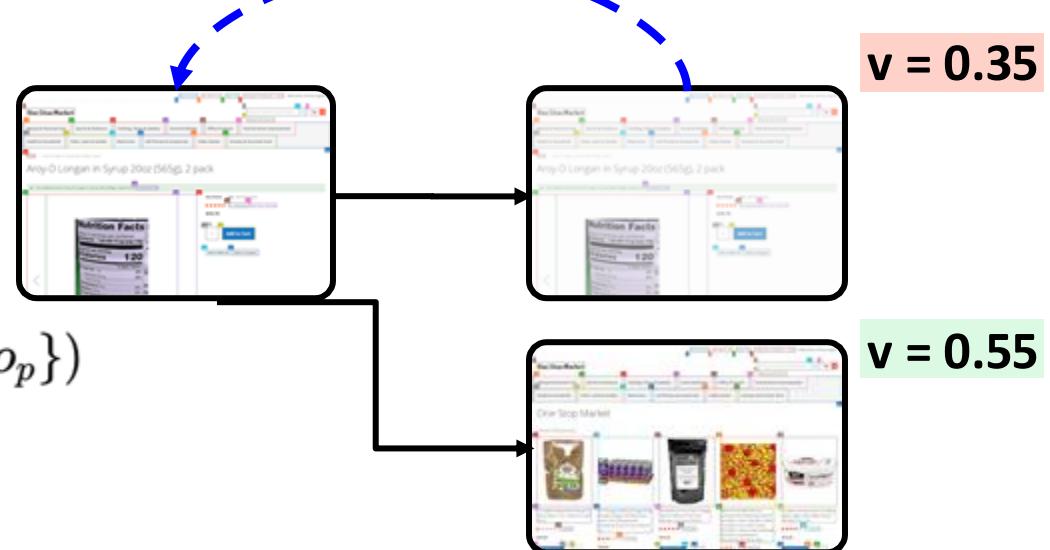
Repeated sampling helps!

- But the space is exponentially large. Can we guide exploration?
- Key idea: apply value function to intermediate nodes.



# Our Method: Tree Search

- Best-first search algorithm
- Ingredients:
  - Baseline agent to propose actions.
  - Way to backtrack in the environment.
  - A **value function**  $v_p = f_v(I, \{o_1, \dots, o_p\})$  to score and rerank candidate states.



In this work, we prompt a multimodal LLM (GPT-4o) to act as an evaluator.



**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

1 Step sequence

v = 1.0

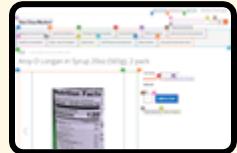
State values

► Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search



Starting State



**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

1 Step sequence

v = 1.0

State values

► Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search



Starting State



**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

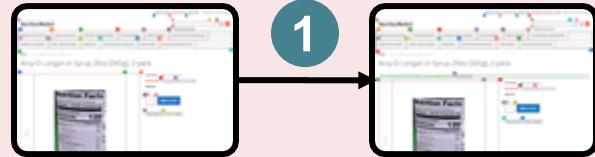
### Legend

1 Step sequence

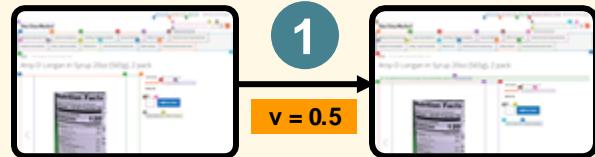
v = 1.0 State values

► Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search



Starting State



**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

1 Step sequence

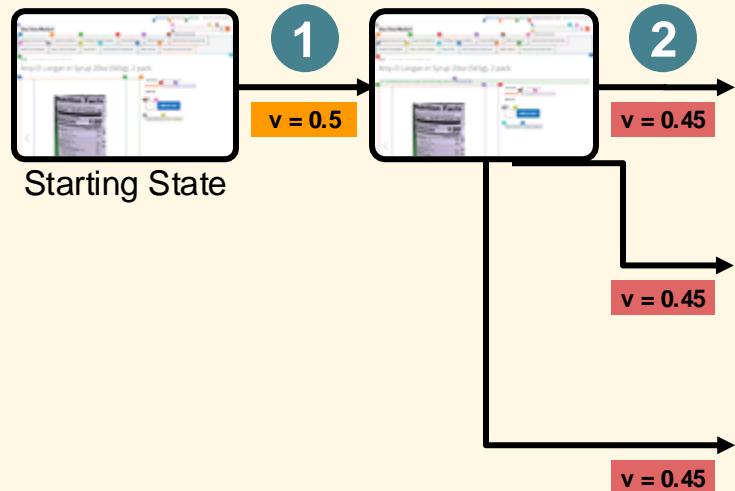
v = 1.0 State values

► Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

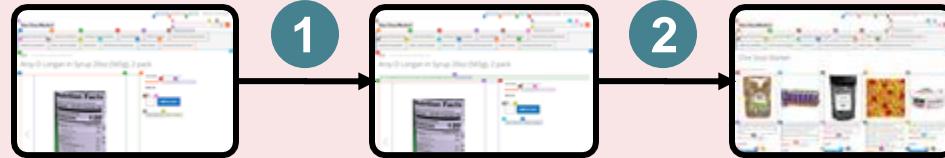
### Legend

1 Step sequence

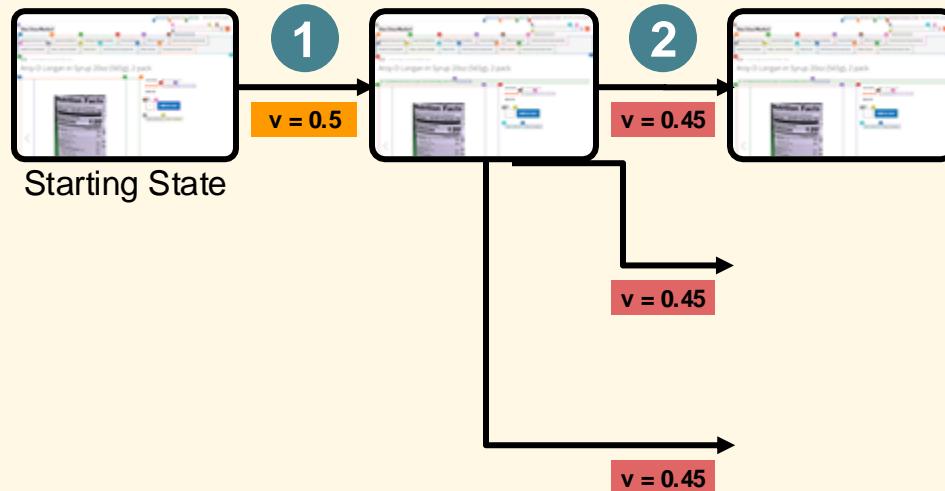
v = 1.0 State values

► Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

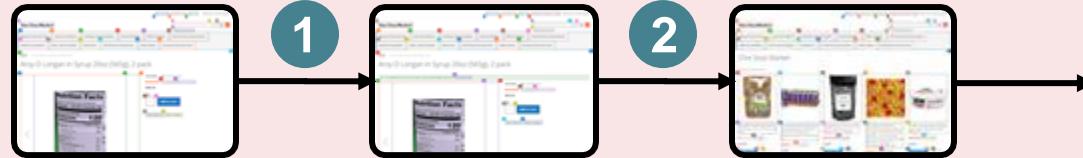
### Legend

1 Step sequence

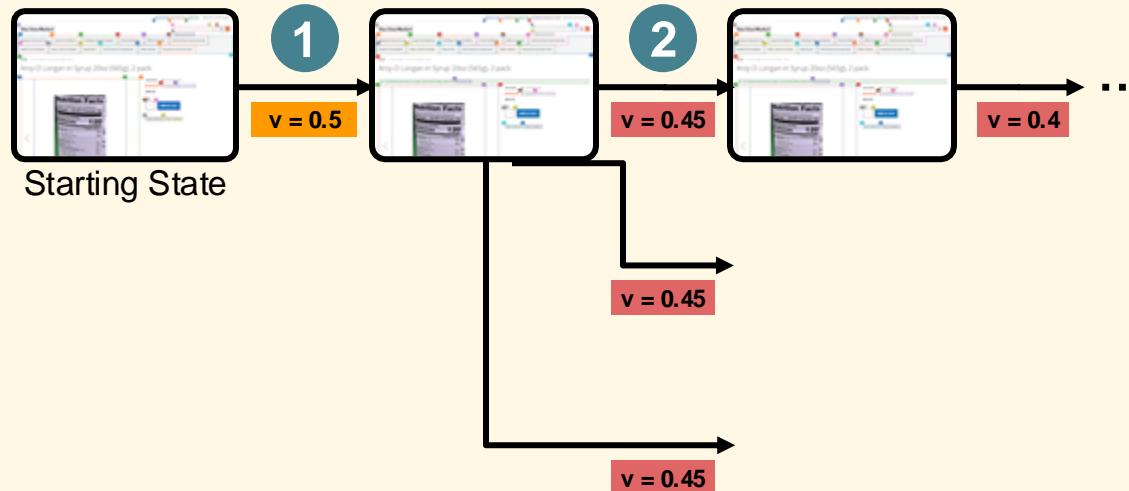
v = 1.0 State values

► Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

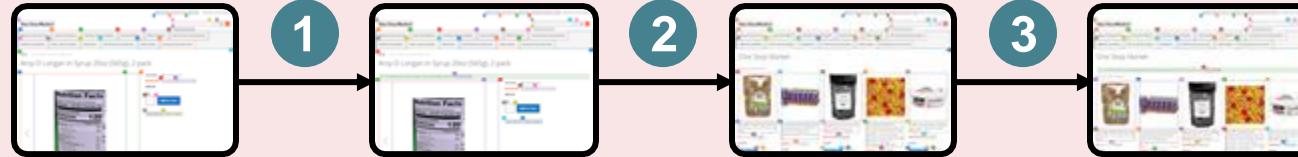
### Legend

1 Step sequence

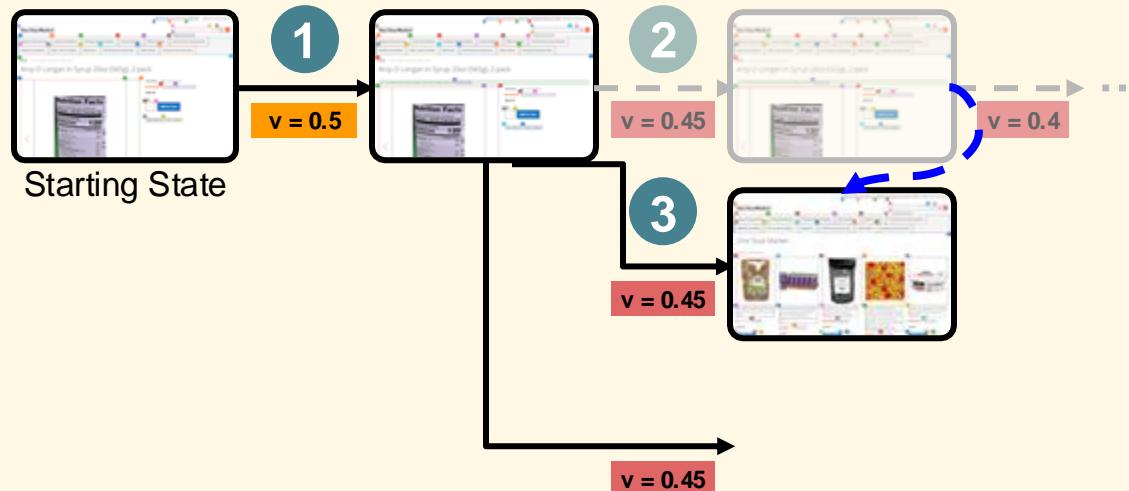
v = 1.0 State values

► Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

1 Step sequence

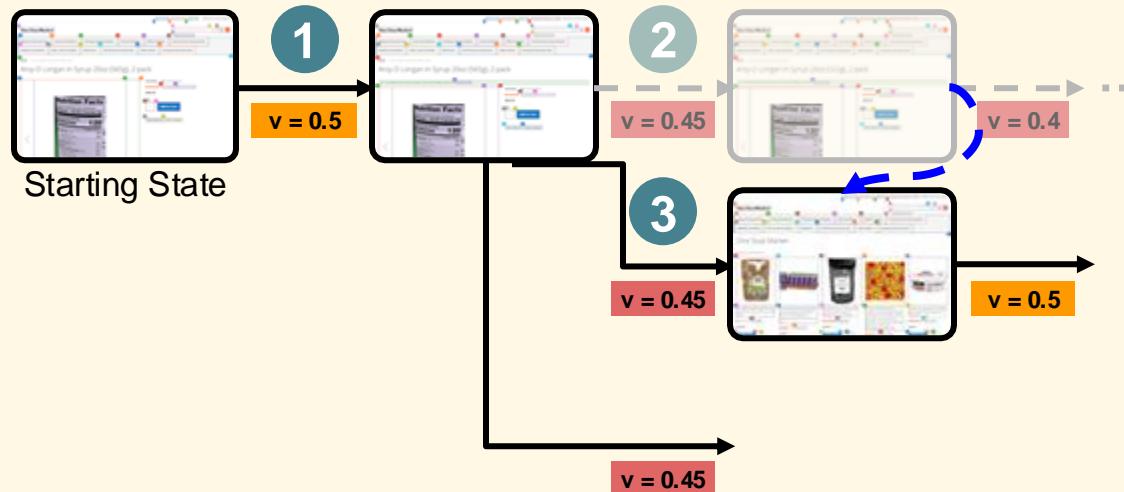
v = 1.0 State values

→ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

1 Step sequence

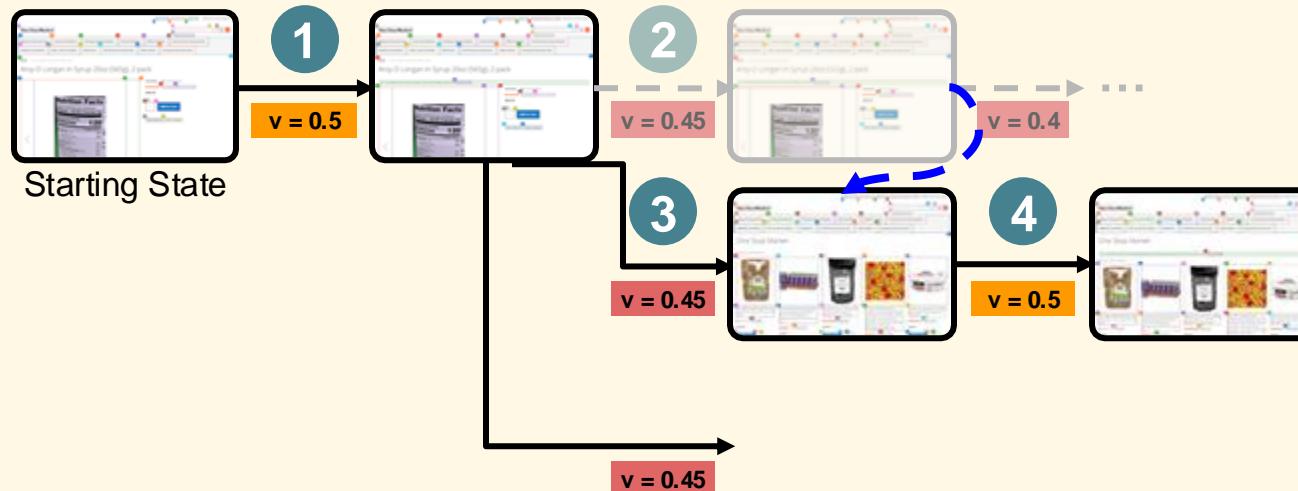
v = 1.0 State values

→ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

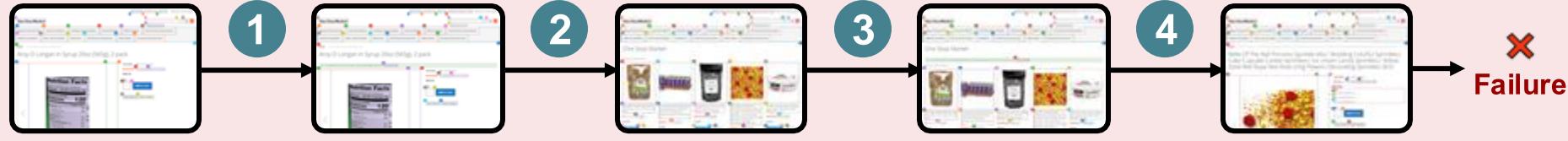
### Legend

1 Step sequence

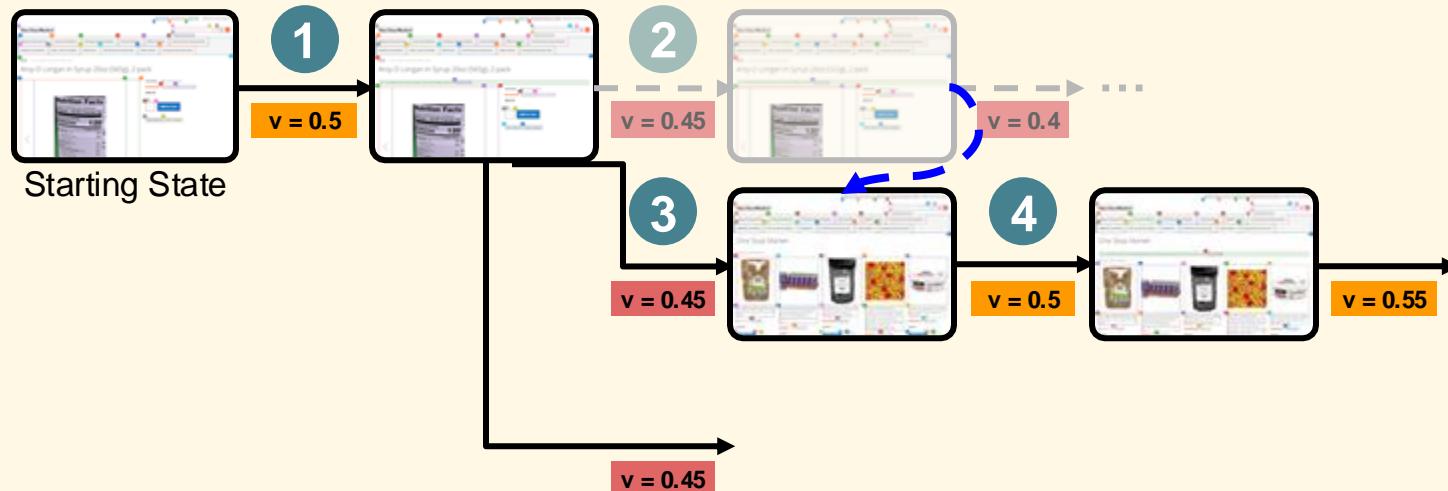
v = 1.0 State values

→ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

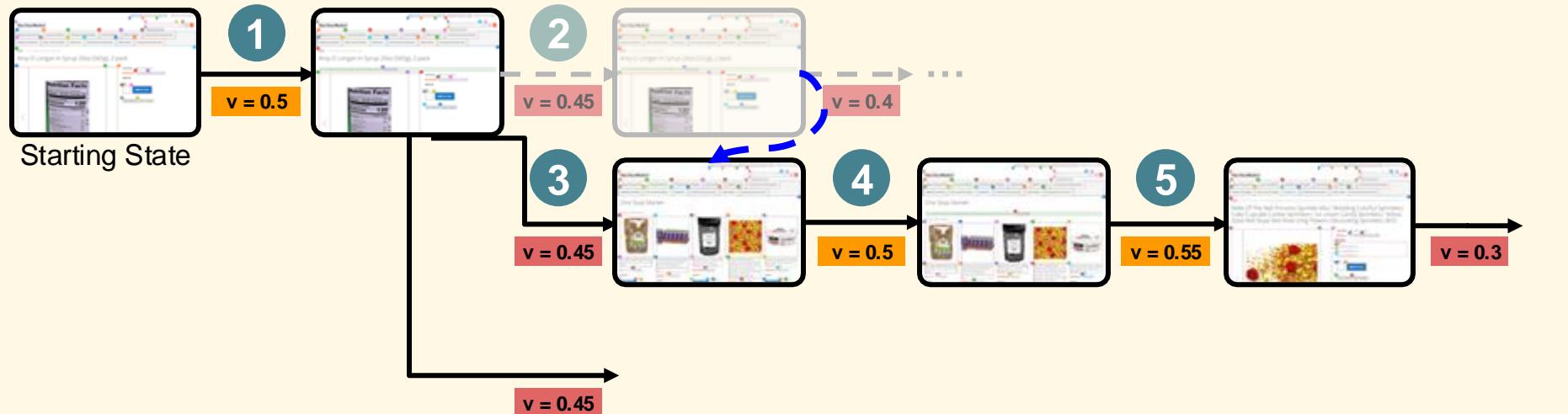
### Legend

- 1 Step sequence
- v = 1.0 State values
- Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

- 1 Step sequence
- v = 1.0 State values
- ▶ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

- 1 Step sequence
- v = 1.0 State values
- Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

- 1 Step sequence
- v = 1.0 State values
- ▶ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

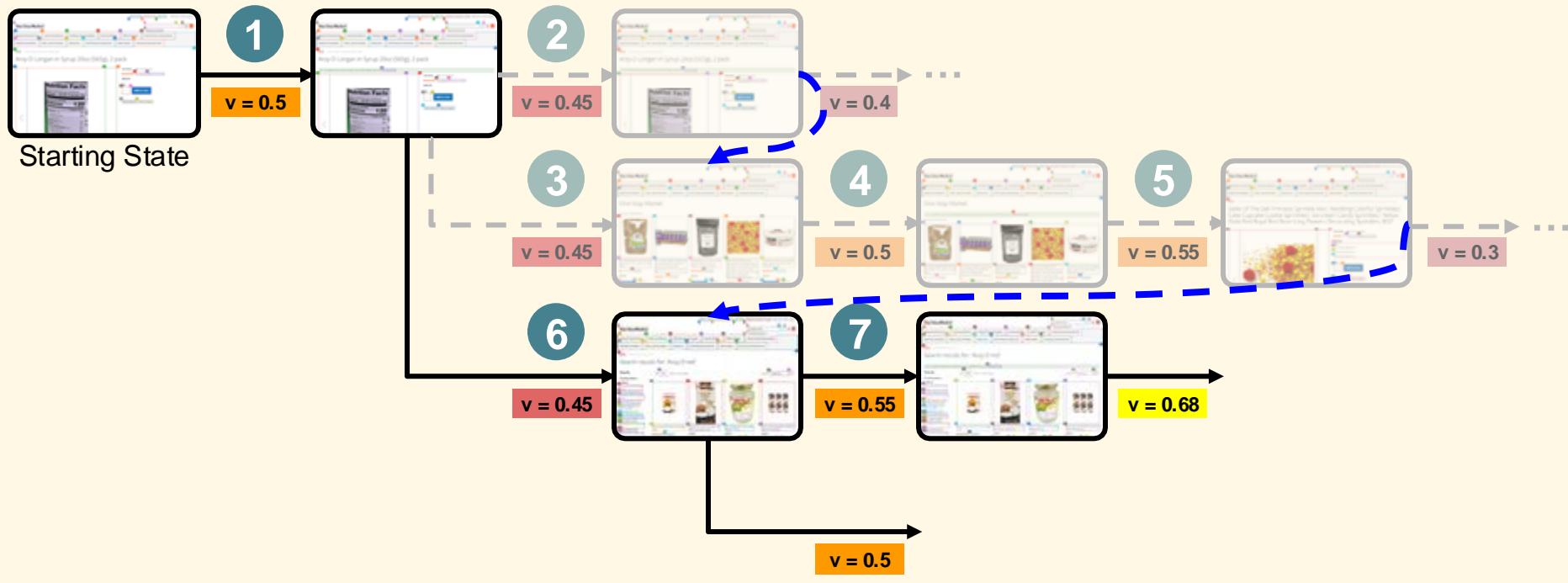
### Legend

- 1 Step sequence
- v = 1.0 State values
- ▶ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search



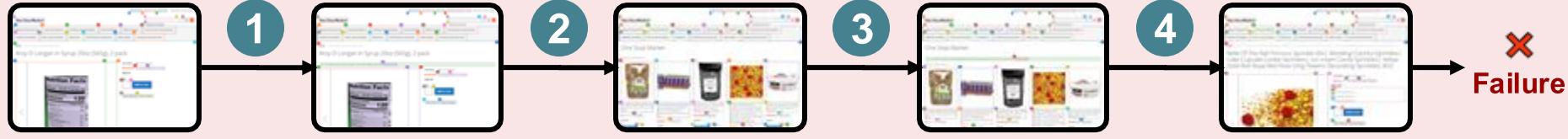


**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

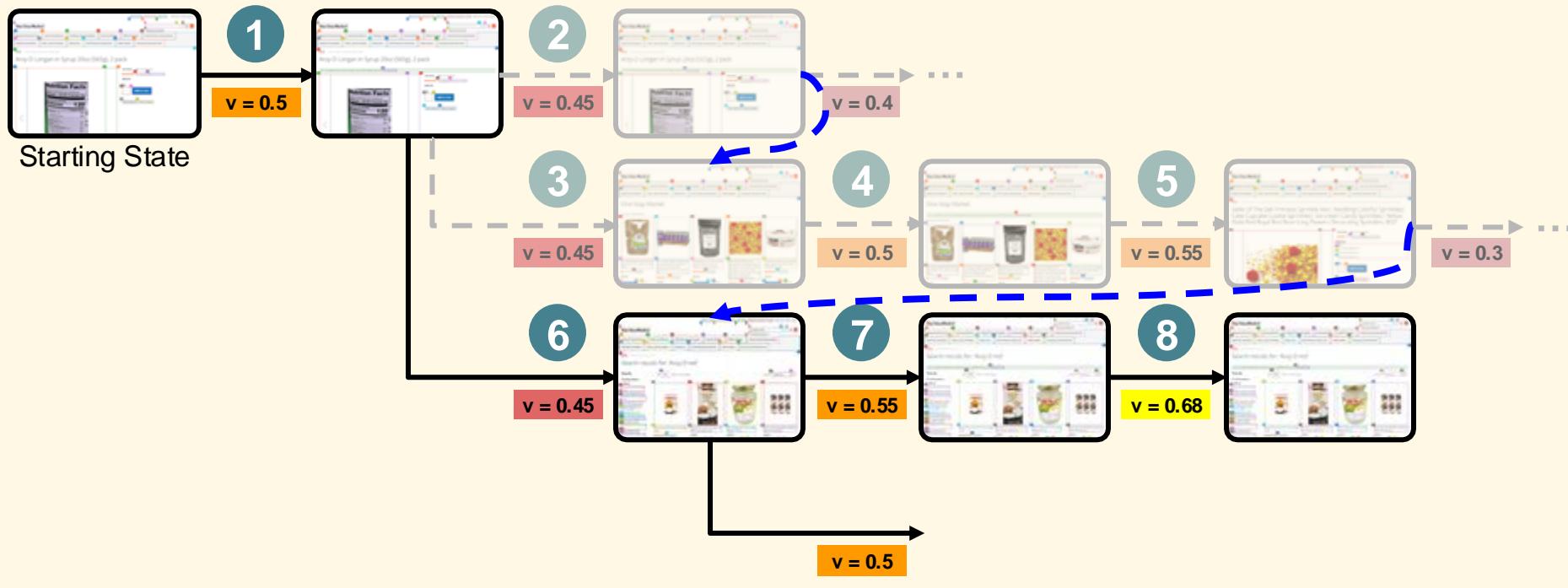
### Legend

- 1 Step sequence
- v = 1.0 State values
- ▶ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

- 1 Step sequence
- v = 1.0 State values
- ▶ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search





**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

- 1 Step sequence
- v = 1.0 State values
- ▶ Backtracking

### GPT-4o Agent



### GPT-4o Agent + Search



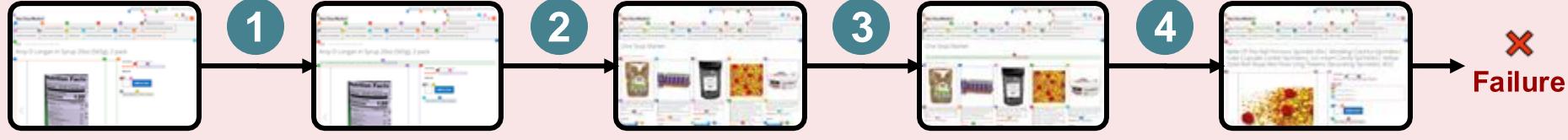


**Task Instruction (I):** “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

### Legend

- 1 Step sequence
- v = 1.0 State values
- ▶ Backtracking

### GPT-4o Agent

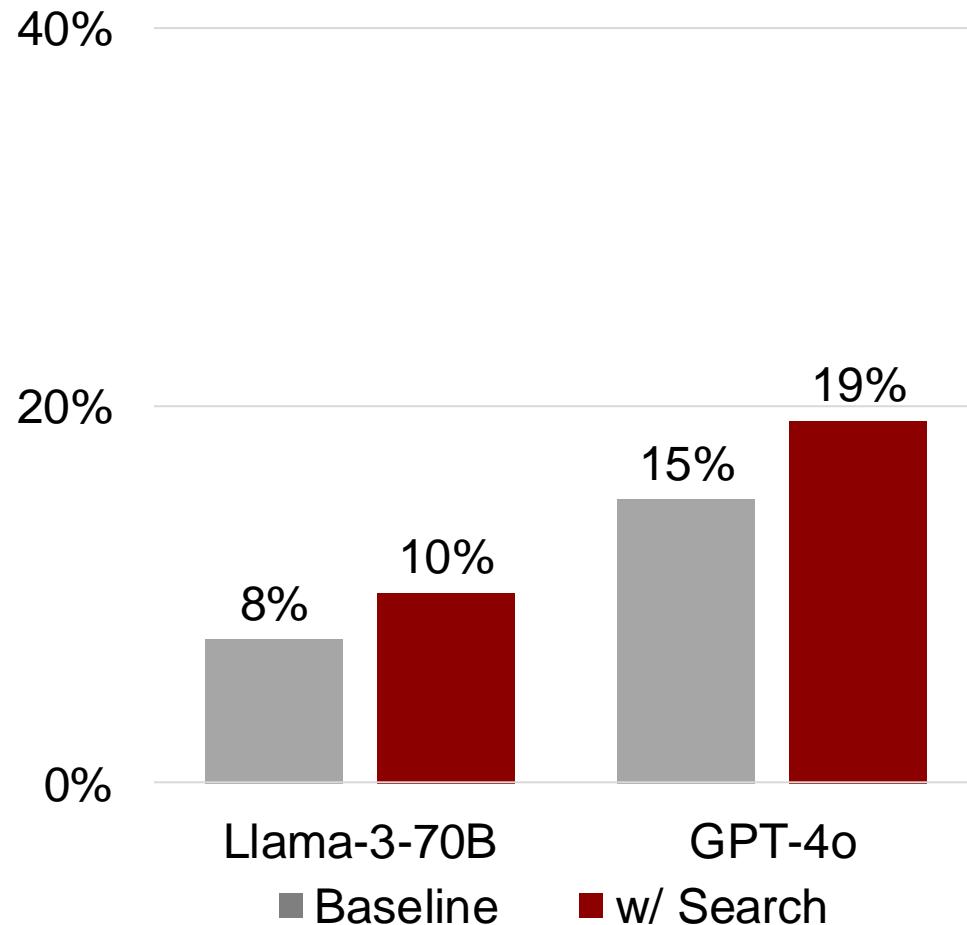


### GPT-4o Agent + Search

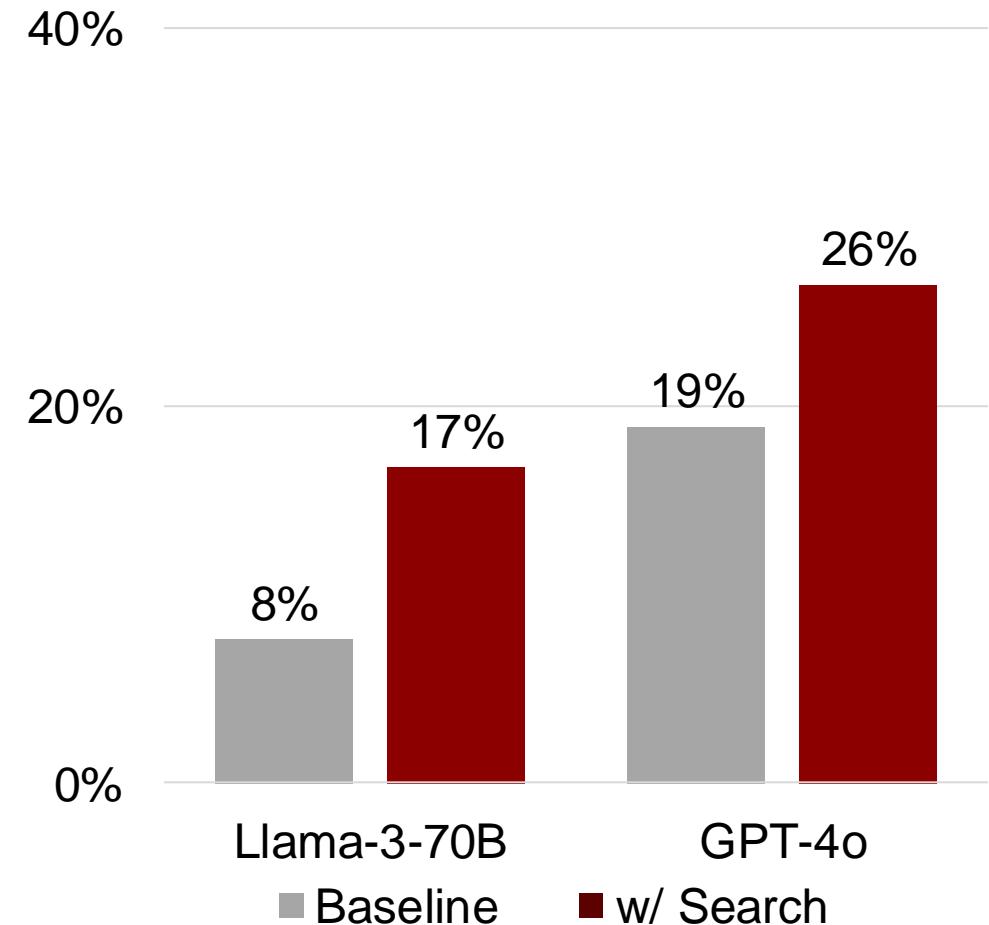


# Results

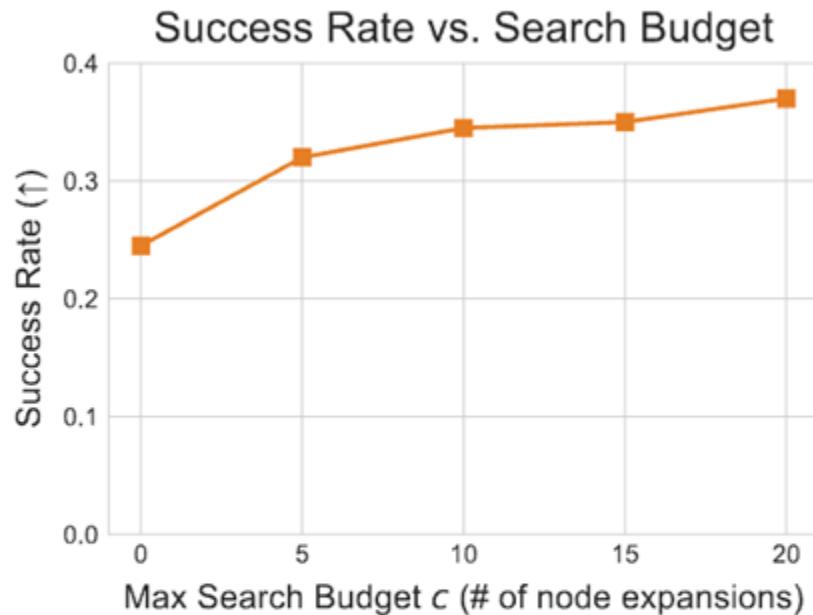
WebArena



VisualWebArena



# Ablations



Success rate on a subset of 200 VWA tasks with search budget  $c$ .  $c = 0$  indicates no search is performed. Success rate generally increases as  $c$  increases.

| Depth $d$ | Branch $b$ | SR ( $\uparrow$ ) | $\Delta$ |
|-----------|------------|-------------------|----------|
| 0         | 1          | 24.5%             | 0%       |
|           | 3          | 26.0%             | +6%      |
| 1         | 5          | 32.0%             | +31%     |
|           | 3          | 31.5%             | +29%     |
| 2         | 5          | 35.0%             | +43%     |
|           | 3          | 35.5%             | +45%     |
| 3         | 5          | <b>37.0%</b>      | +51%     |

Success rate (SR) and relative change over the baseline ( $\Delta$ ) on a subset of 200 VWA tasks with varying search depth ( $d$ ) and branching factor ( $b$ ).  $d = 0$  indicates no search is performed. All methods use a max search budget  $c = 20$ .

# Ablations

- Having a good value function is essential.
- There is still a lot of headroom for improving both the base agent policy, and the value function.

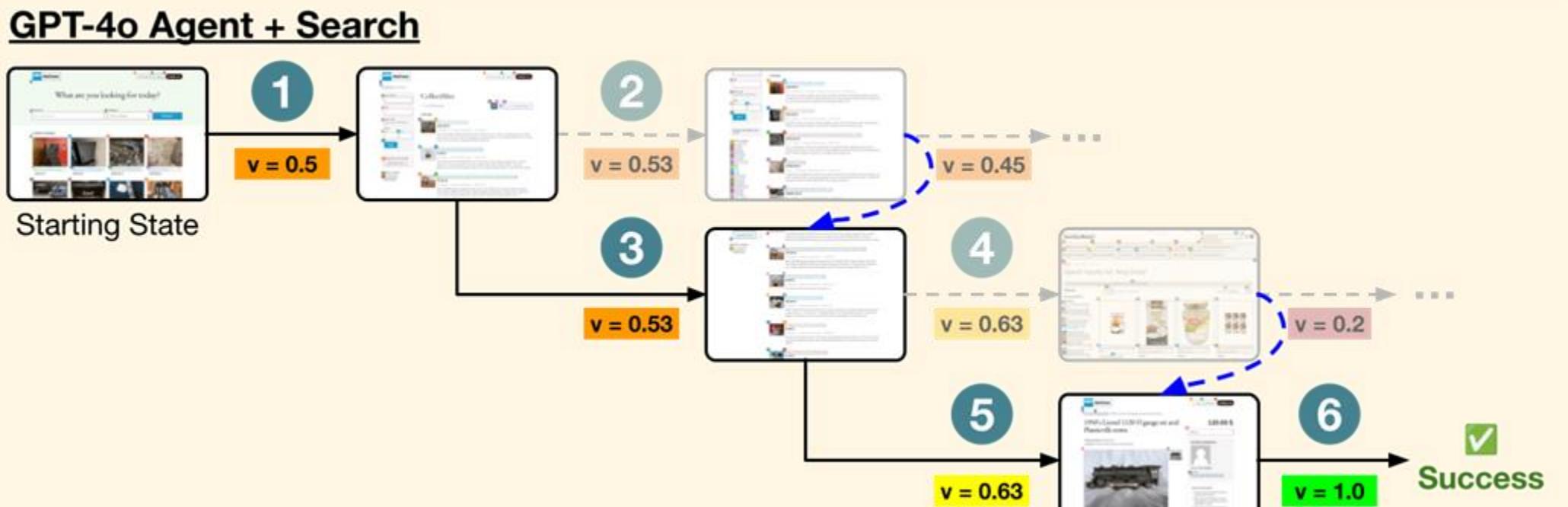
| <b>Value Function</b>     | <b>SR (<math>\uparrow</math>)</b> |
|---------------------------|-----------------------------------|
| None (no search)          | 24.5%                             |
| LLaVA (w/ SC, $n = 20$ )  | 30.0%                             |
| GPT-4o (no SC)            | 28.5%                             |
| GPT-4o (w/ SC, $n = 5$ )  | 32.5%                             |
| GPT-4o (w/ SC, $n = 20$ ) | 37.0%                             |

Table 3: Success rate of the GPT-4o agent with different value functions.

# Qualitative Results



**Task Instruction (I):** "I recall seeing this exact item on the site, help me find the most recent post of it. I recall seeing it in either the Collectibles or Antiques section."



**Legend:**



Search sequence

- → Backtracking



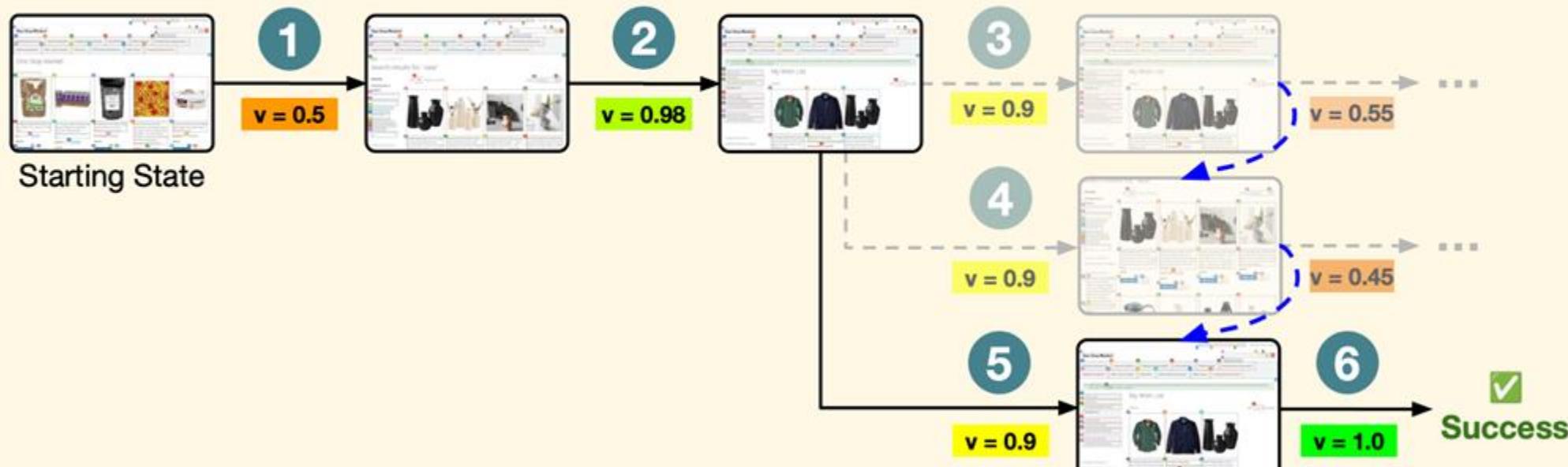
State values

# Qualitative Results



**Task Instruction (I):** “I need something like this for my apartment. Can you add one to my wishlist?”

## GPT-4o Agent + Search

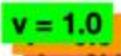


**Legend:**



Search sequence

-> Backtracking



State values

# Limitations

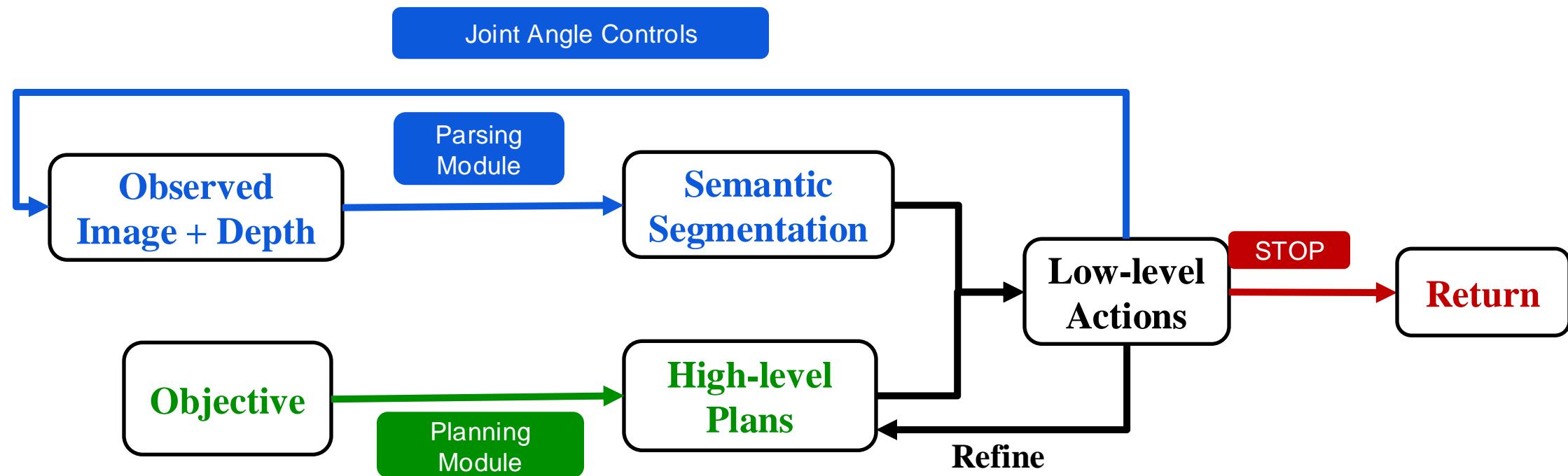
- Search is slow
  - We implemented backtracking in a relatively naive way (store actions in a queue, take them again to get to the original state)
- Dealing with destructive actions
  - Some things on the web are very difficult to undo, e.g., ordering an item

# Current Work

- Search as a policy improvement function
- Improving Value Function by fine-tuning instead of prompting
- Explore compute tradeoff between improving baseline agent vs. doing **more search at inference time**
- What if we don't have a perfect simulator – **how can we collect data at scale?**

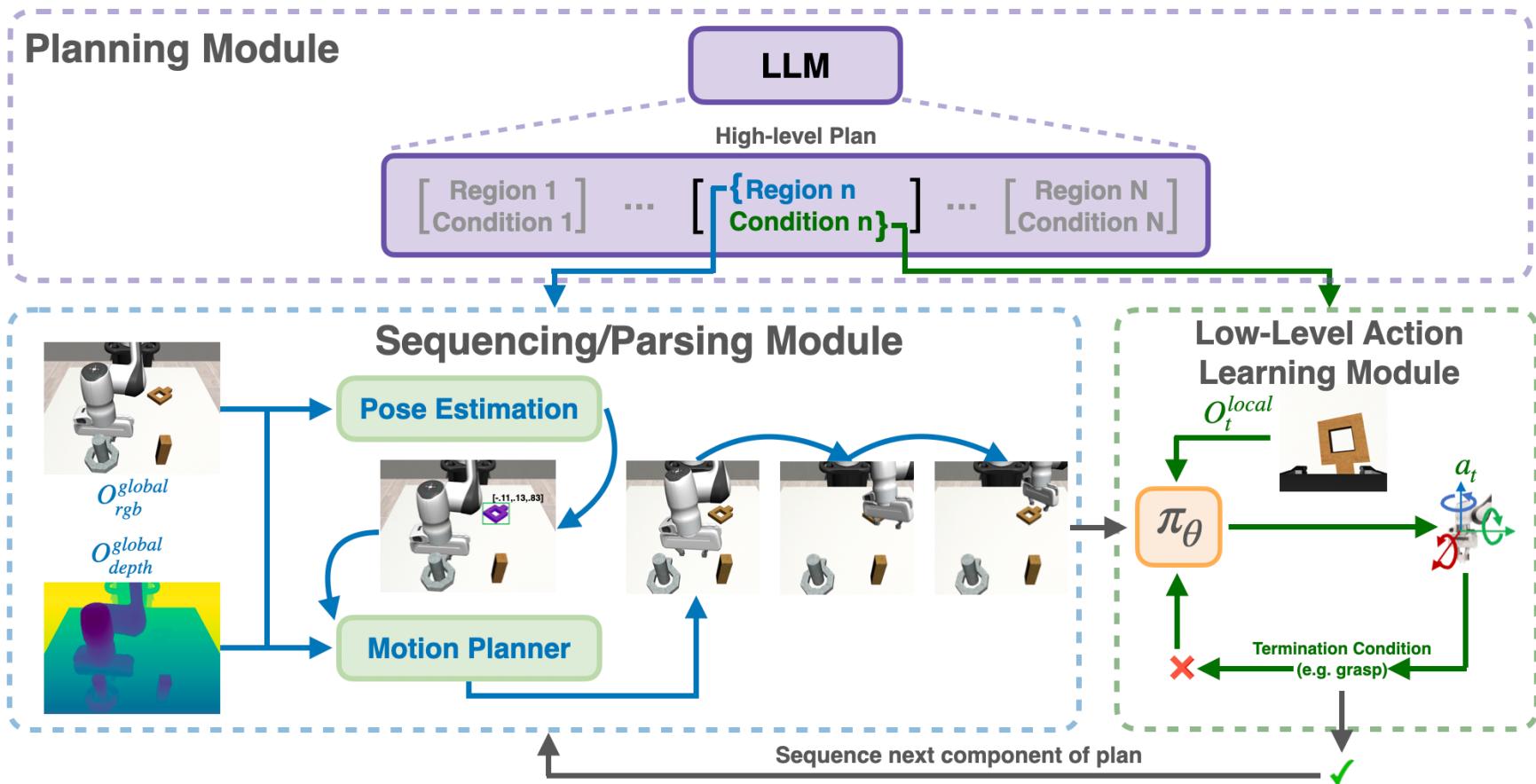
# Physical Agent: Long-horizon Robotic Manipulation Task

- Model architecture of our interactive agent:
  - High-level Planning
  - Observation Parsing
  - Low-level Action Generation





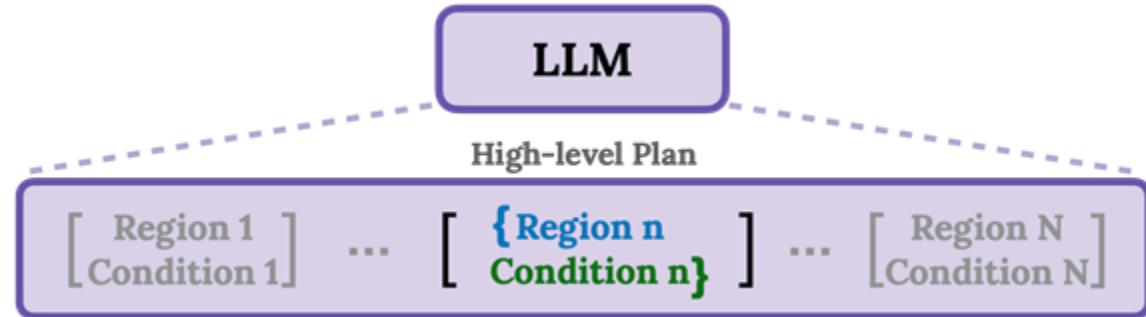
# Plan-Sequence-Learn



Plan-Seq-Learn (PSL): Language Model Guided RL for Solving Long Horizon Robotics, M Dalal, T Chiruvolu, D Chapat, R Salakhutdinov, ICLR 2024

# Planning Module

- Structured language plans: (object, condition)
- Prompt: Task description, conditions, objects, formatting



**Stage termination conditions:** (grasp, place).

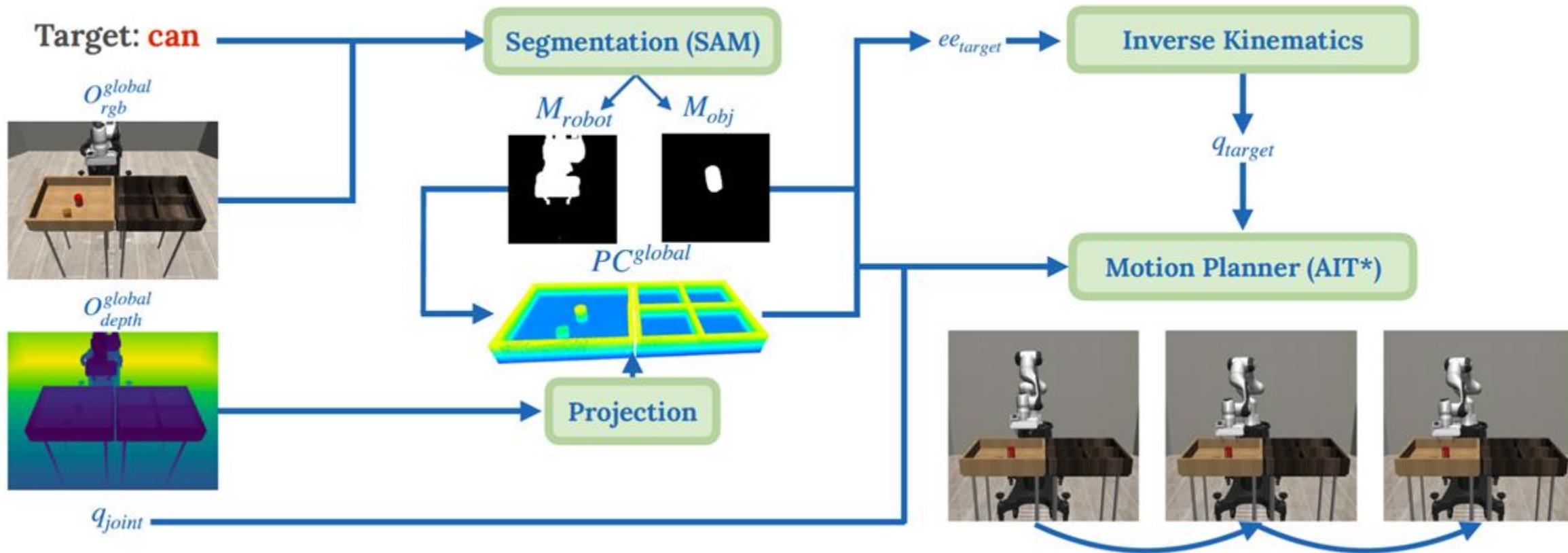
**Task description:** The silver nut goes on the silver peg and the gold nut goes on the gold peg. Give me a simple plan to solve the task using only the stage termination conditions. Make sure the plan follows the formatting specified below and make sure to take into account object geometry.

**Formatting of output:** a list in which each element looks like: (<object/region>, <stage termination condition>). Don't output anything else.

**Output:**

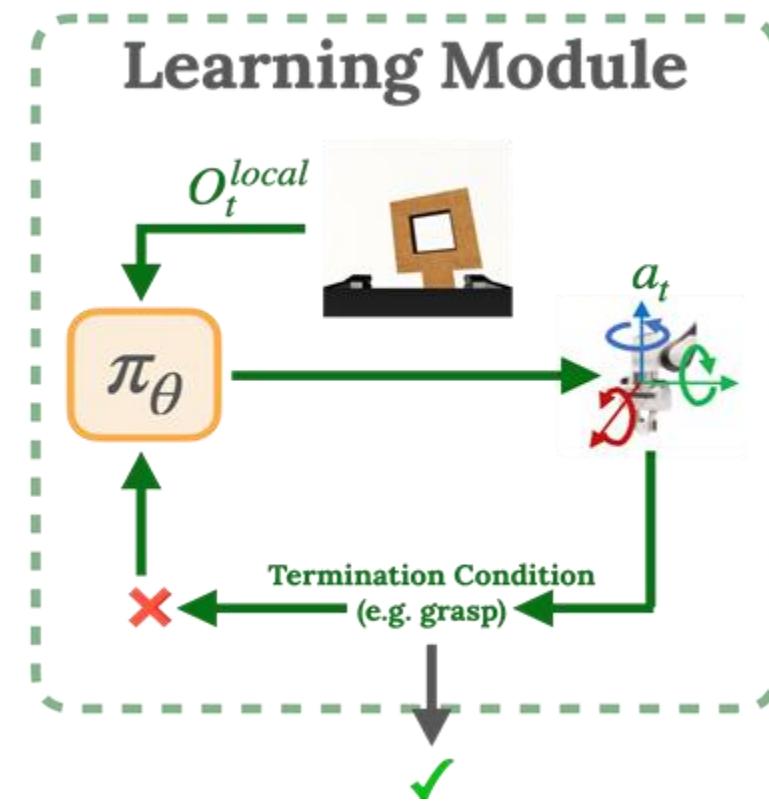
[("silver nut", "grasp"), ("silver peg", "place"), ("gold nut", "grasp"), ("gold peg", "place")]

# Sequencing/Parsing Module: Grounding Language Plans in the Scene



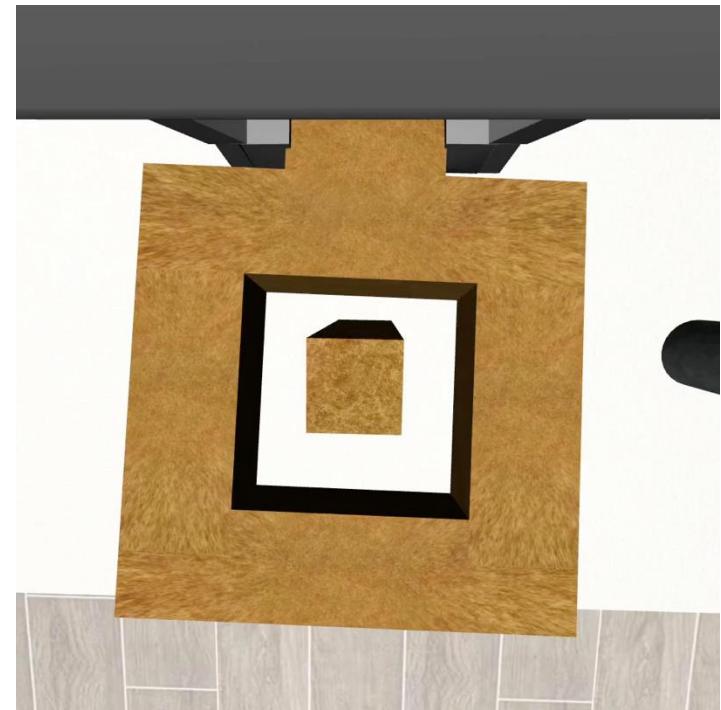
# Learning Low-level Actions Module: Learning Local Control

- Learned RL policies for interaction
- Trained with task reward
- Single RL model instead of separate per stage
- Local instead of global observations

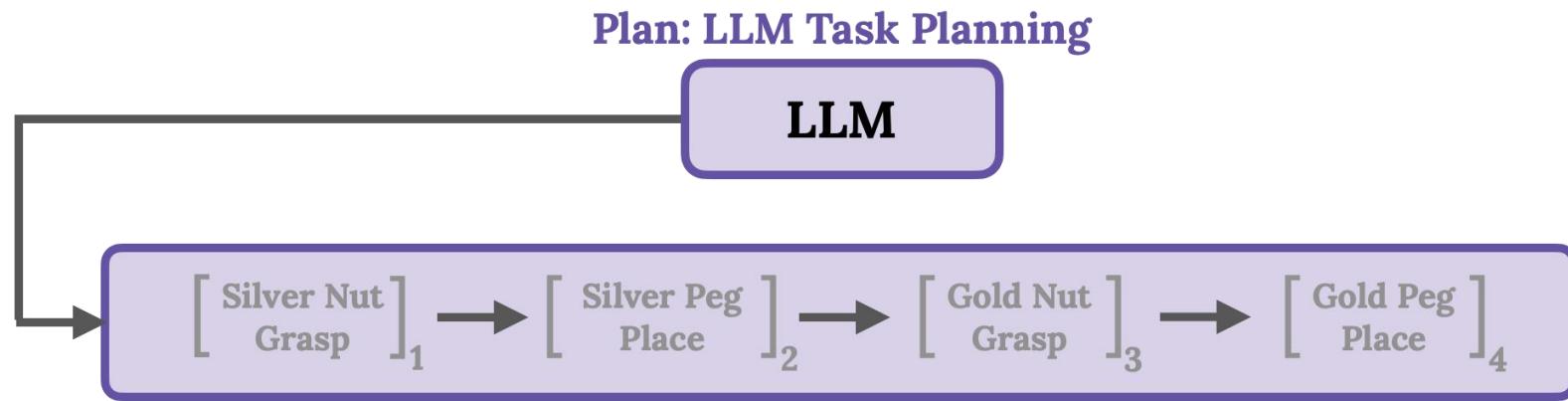


# Learning Low-level Actions Module: Learning Local Control

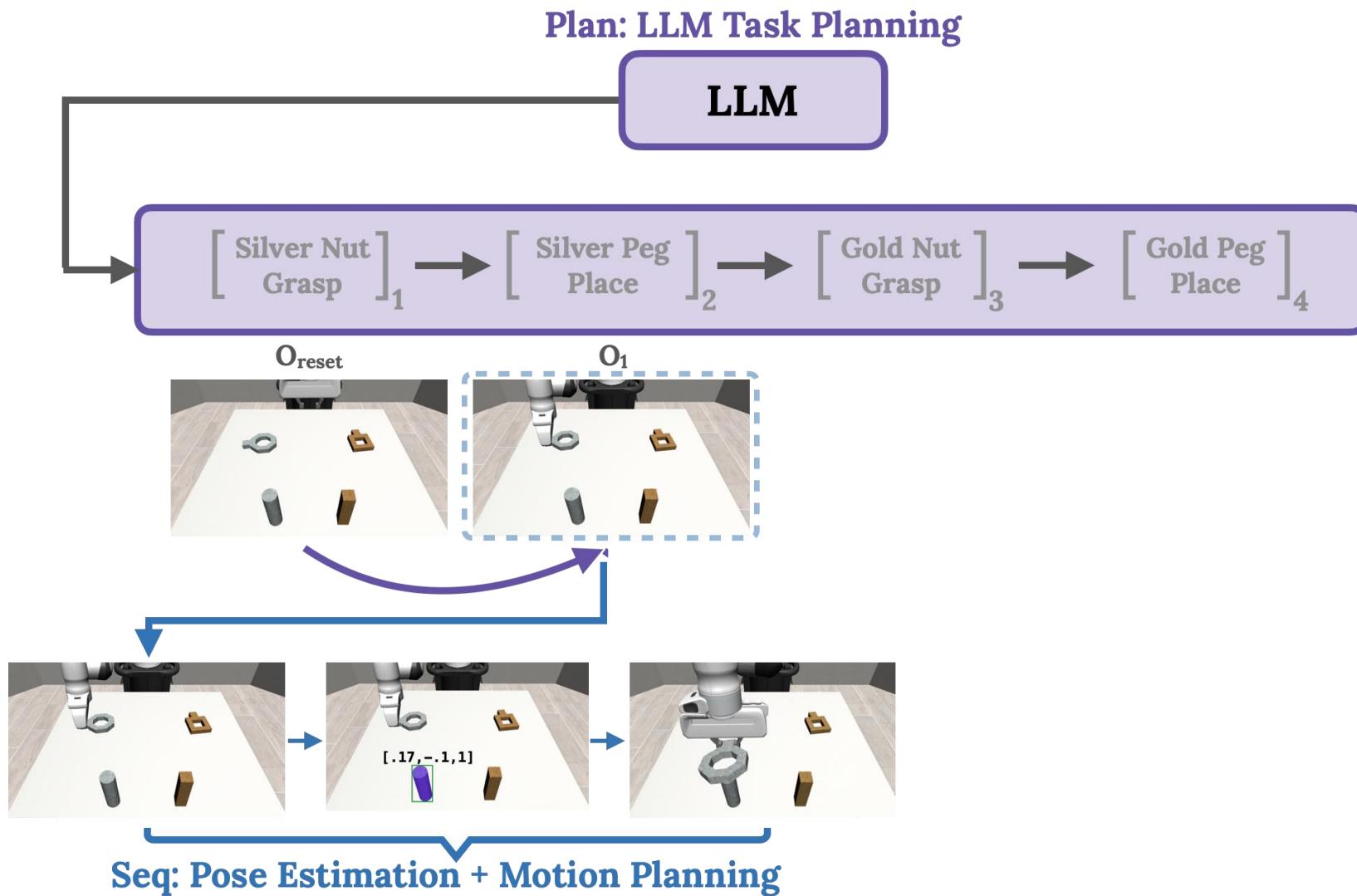
- Learned RL policies for interaction
- Trained with task reward
- Single RL model instead of separate per stage
- Local instead of global observations



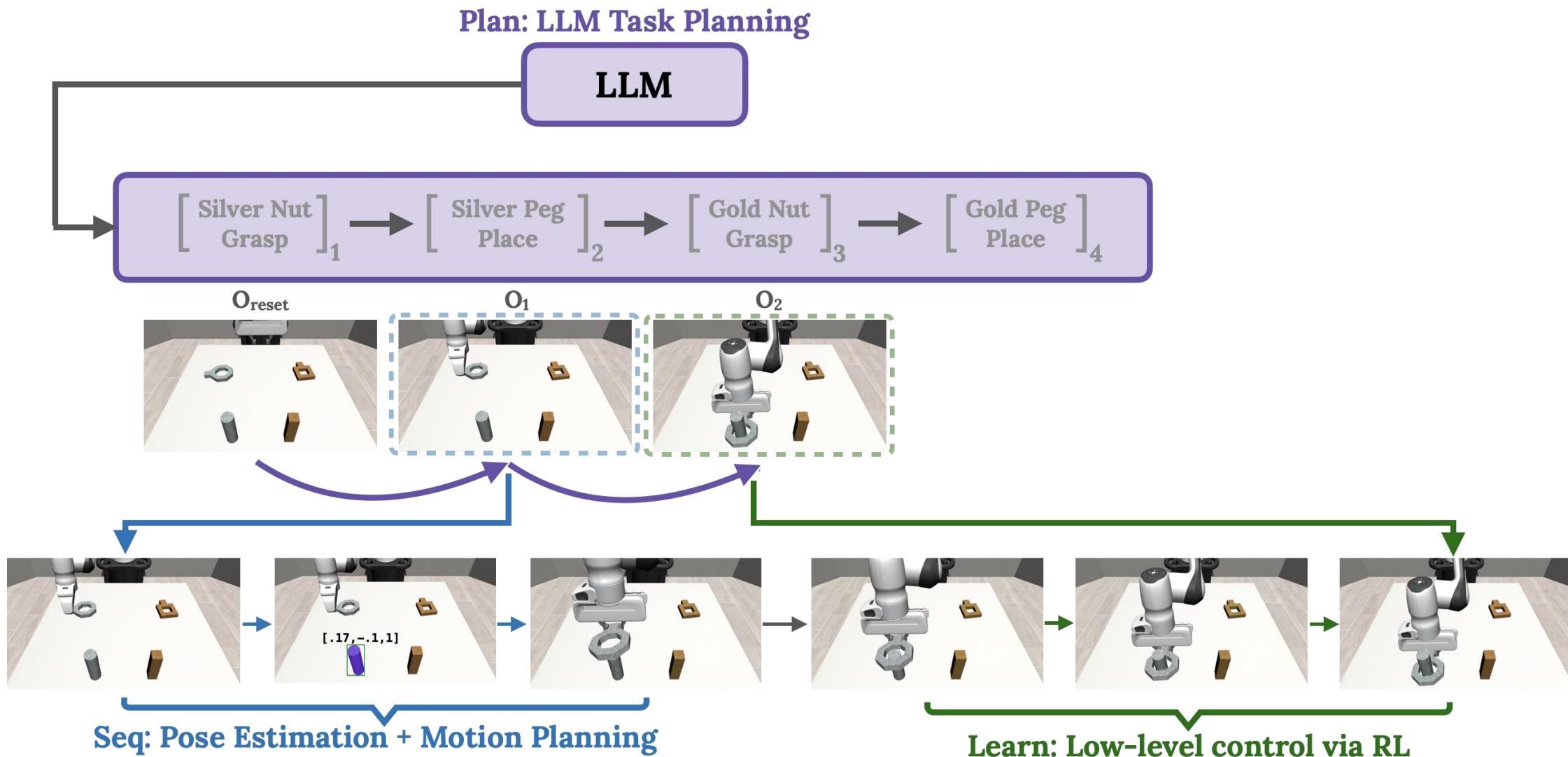
# Full Pipeline Example



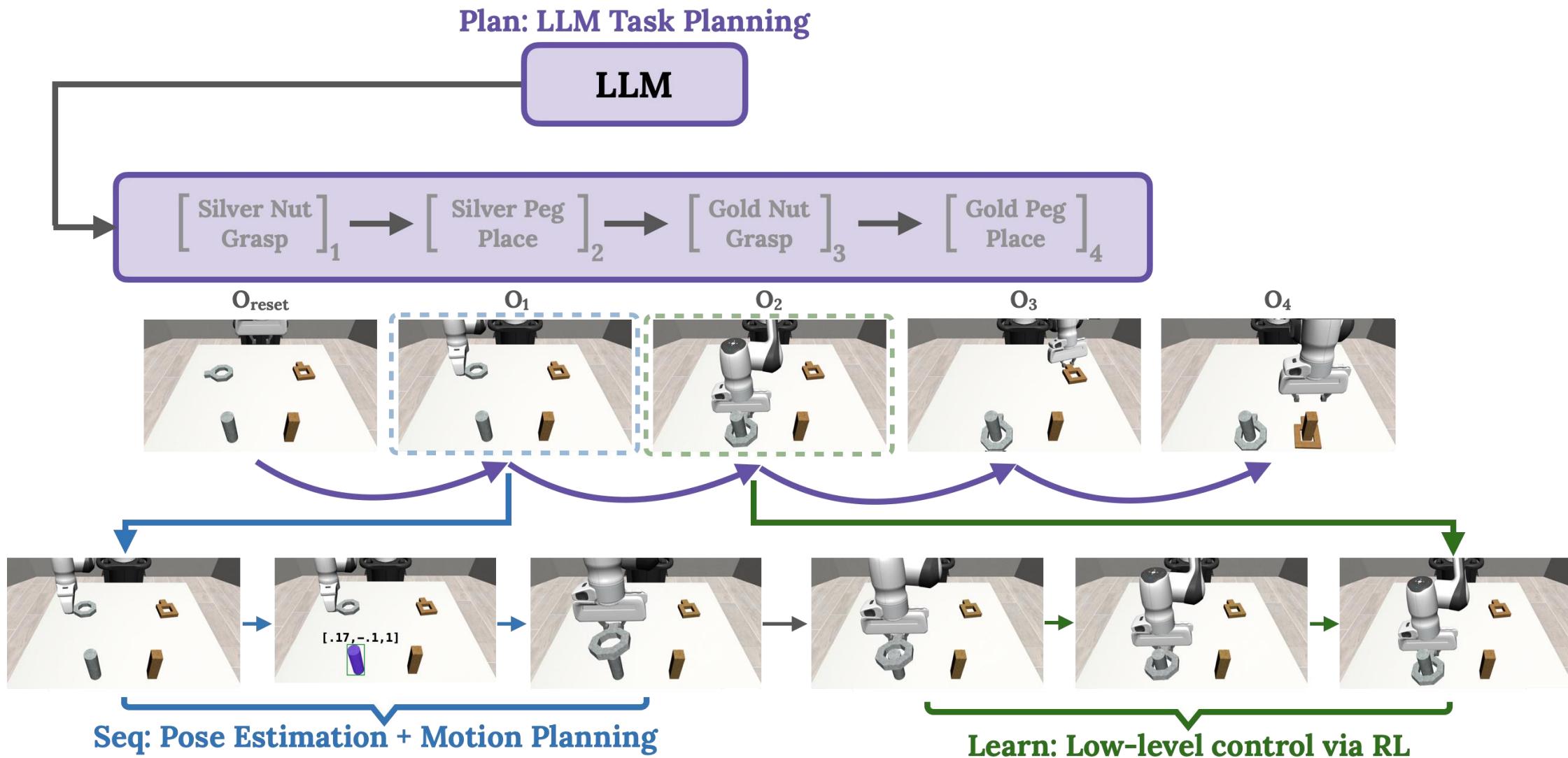
# Full Pipeline Example

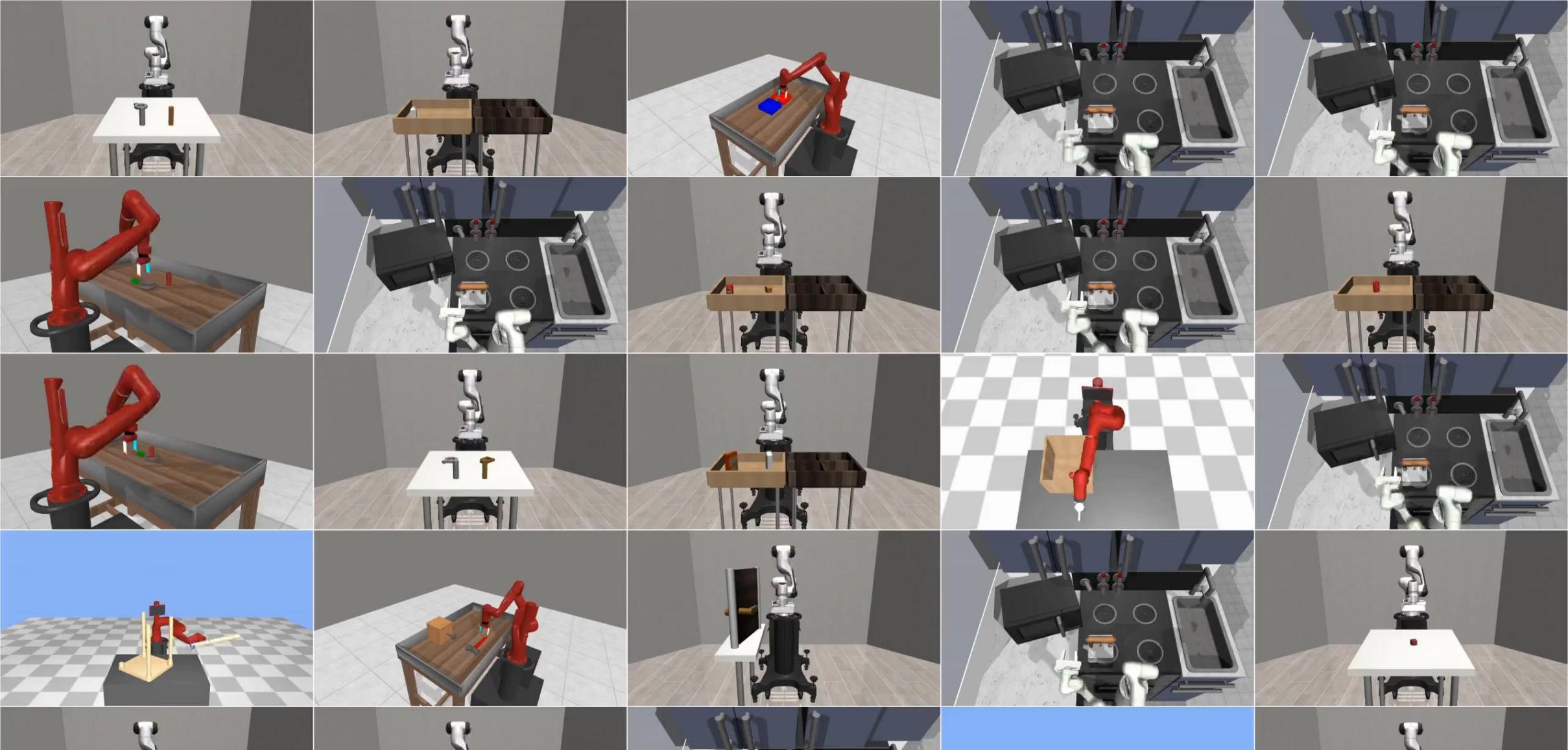


# Full Pipeline Example



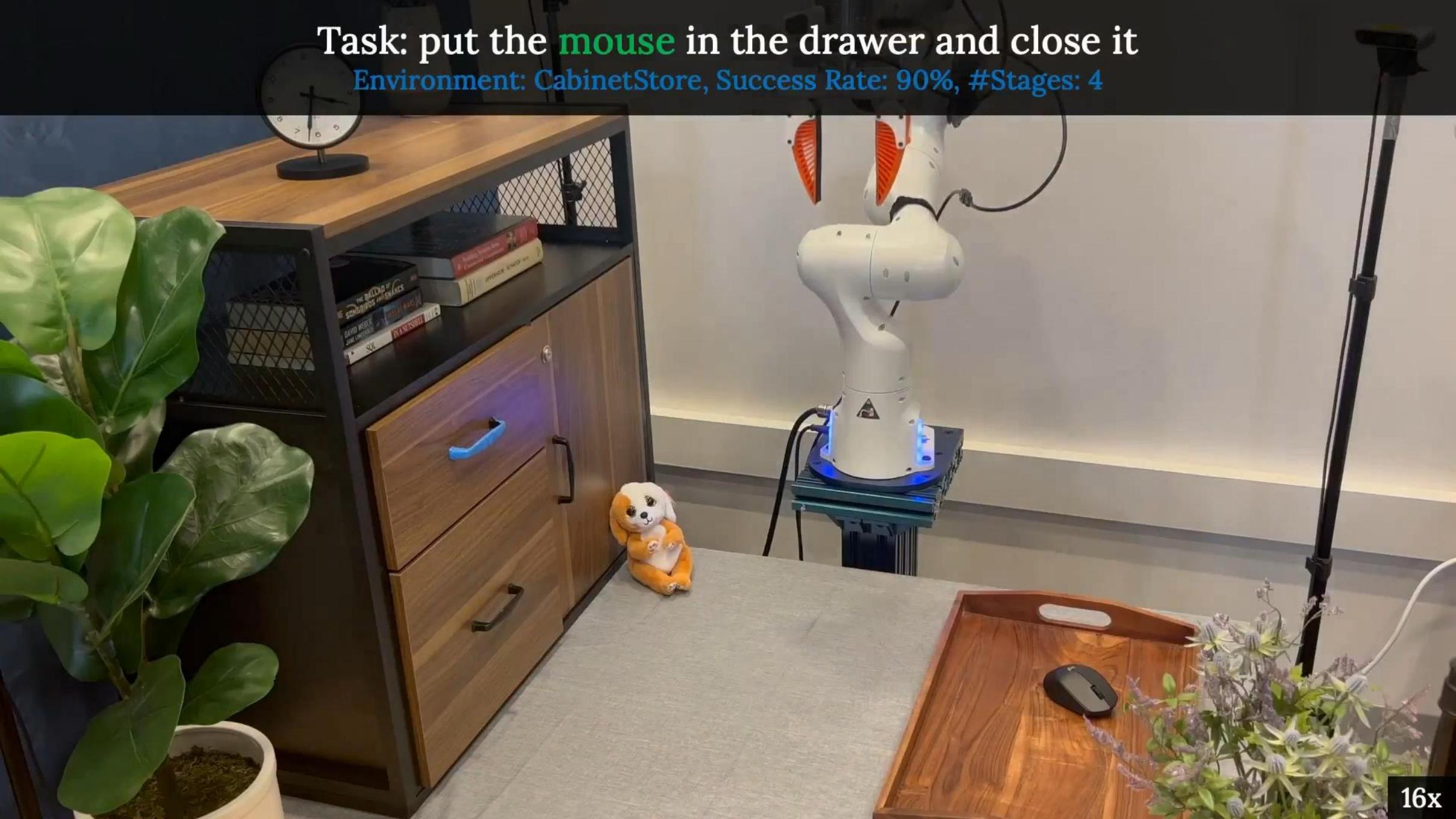
# Full Pipeline Example





PSL solves 25+ long-horizon robotics tasks across four benchmark environment suites with greater than 85% success rates

Task: put the mouse in the drawer and close it  
Environment: CabinetStore, Success Rate: 90%, #Stages: 4



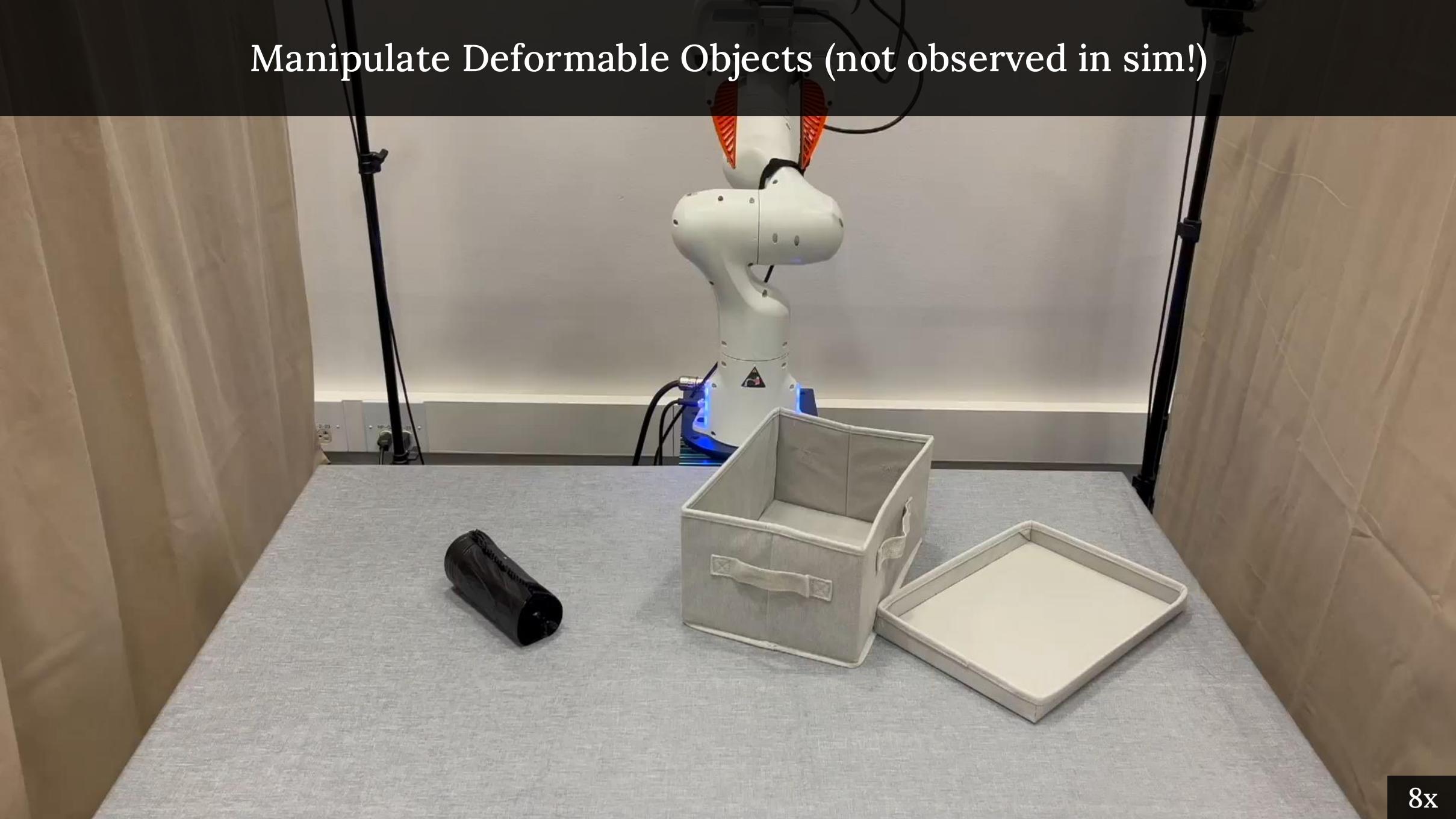
# Generalizes to Novel Object Geometries/Categories



Manipulate novel objects with unseen receptacles



Manipulate Deformable Objects (not observed in sim!)



# Summary

- VisualWebArena: a benchmark of realistic tasks designed to rigorously evaluate and advance the capabilities of autonomous multimodal web agents
- Inference-time search algorithm designed to enhance the capabilities of language model agents on realistic web tasks
- **AI Safety and robustness, especially in the age of autonomous systems.**

# Adversarial Attacks on Multimodal Agents

**Chen Henry Wu, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, Aditi Raghunathan**

Carnegie Mellon University

{chenwu2,jingyuk,rsalakhu,dfried,aditirag}@cs.cmu.edu

