

Multimodal Autonomous AI Agents

Russ Salakhutdinov

Machine Learning Department
Carnegie Mellon University



Carnegie
Mellon
University

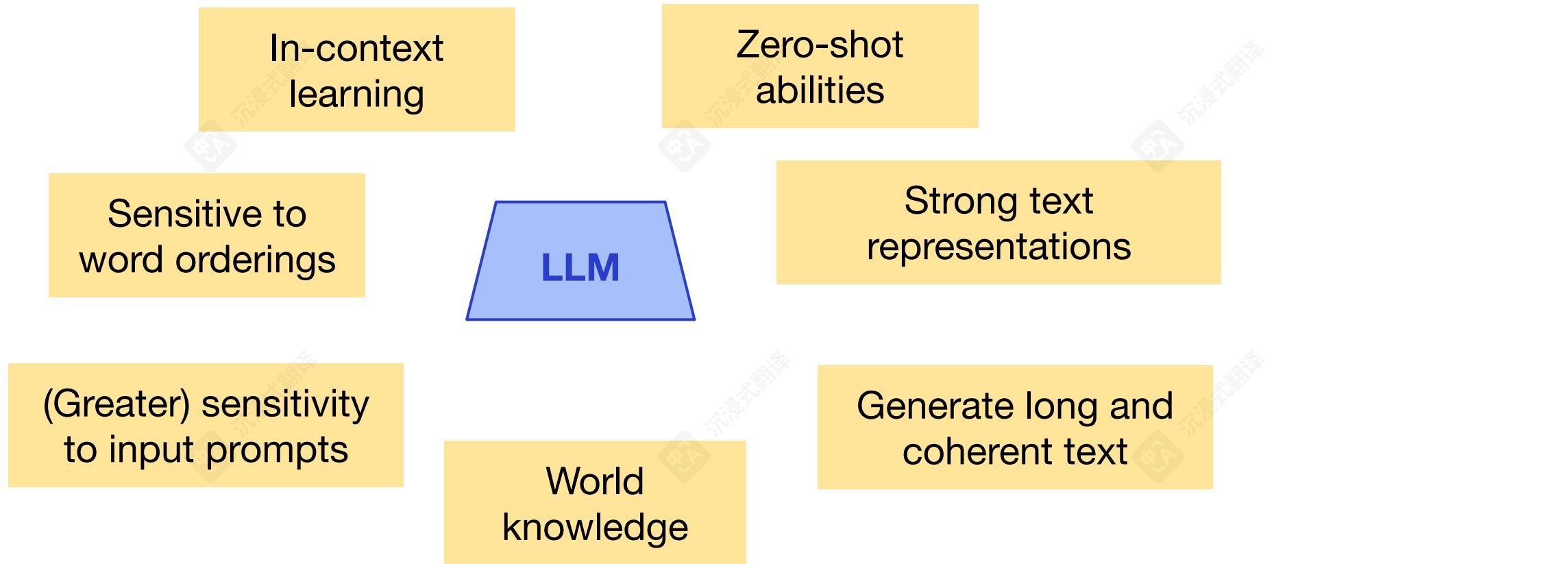
Multimodal Autonomous AI Agents

Russ Salakhutdinov

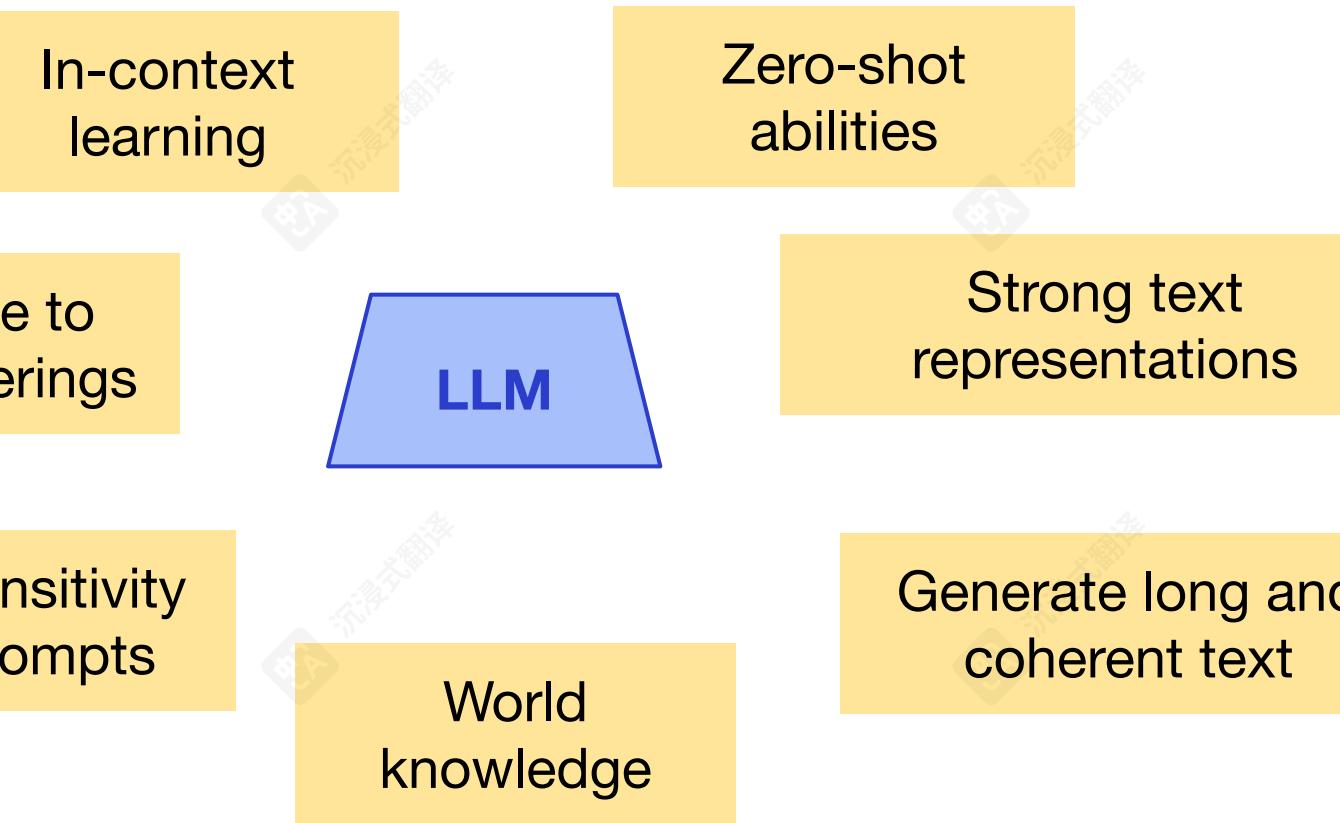
机器学习部门
卡内基梅隆大学



Large Language Models



大型语言模型



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

自主 AI 代理

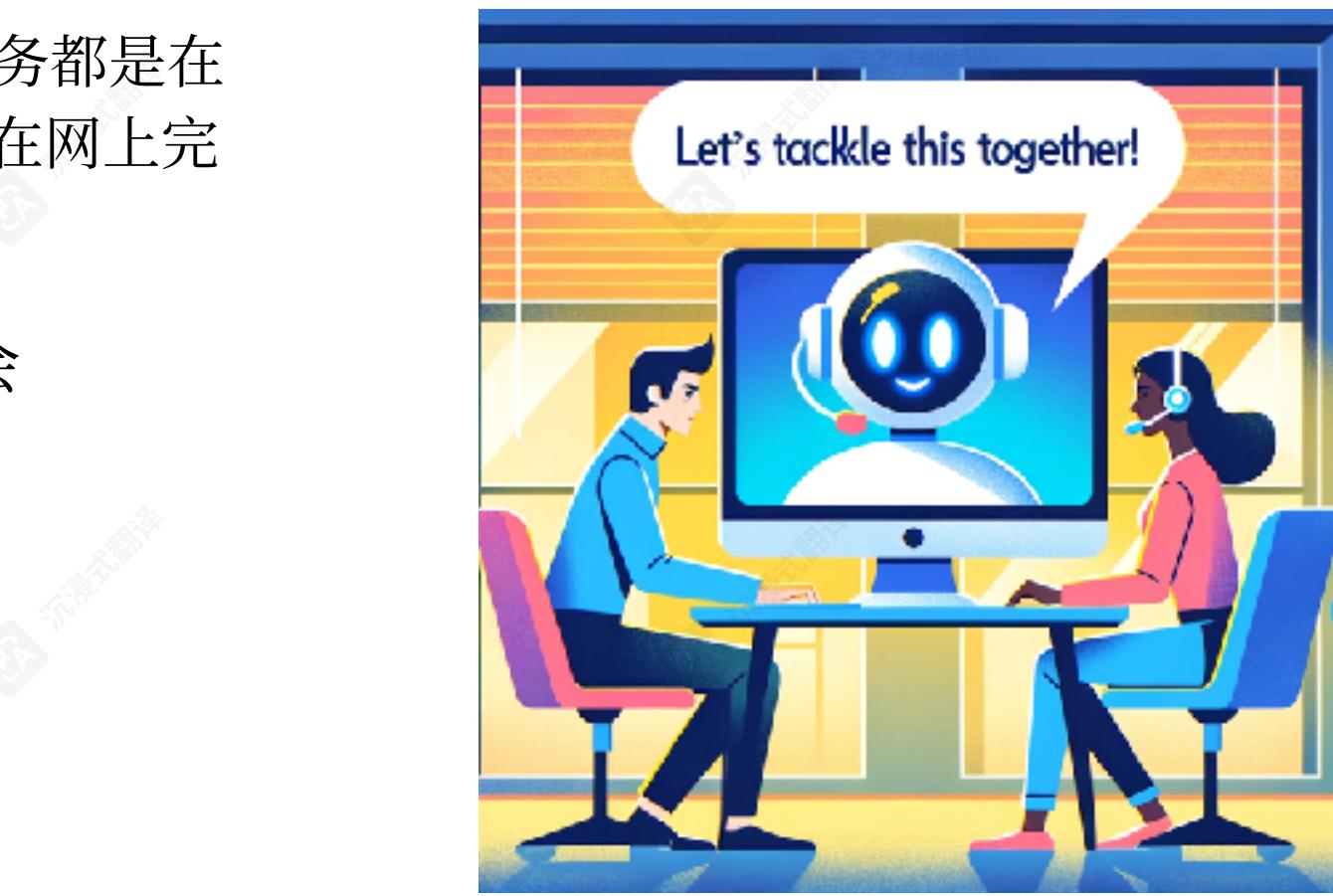
- 我们今天完成的许多生产性任务都是在计算机上完成的 - 而其中许多是在网上完成的



- 许多自动化琐碎任务的机会



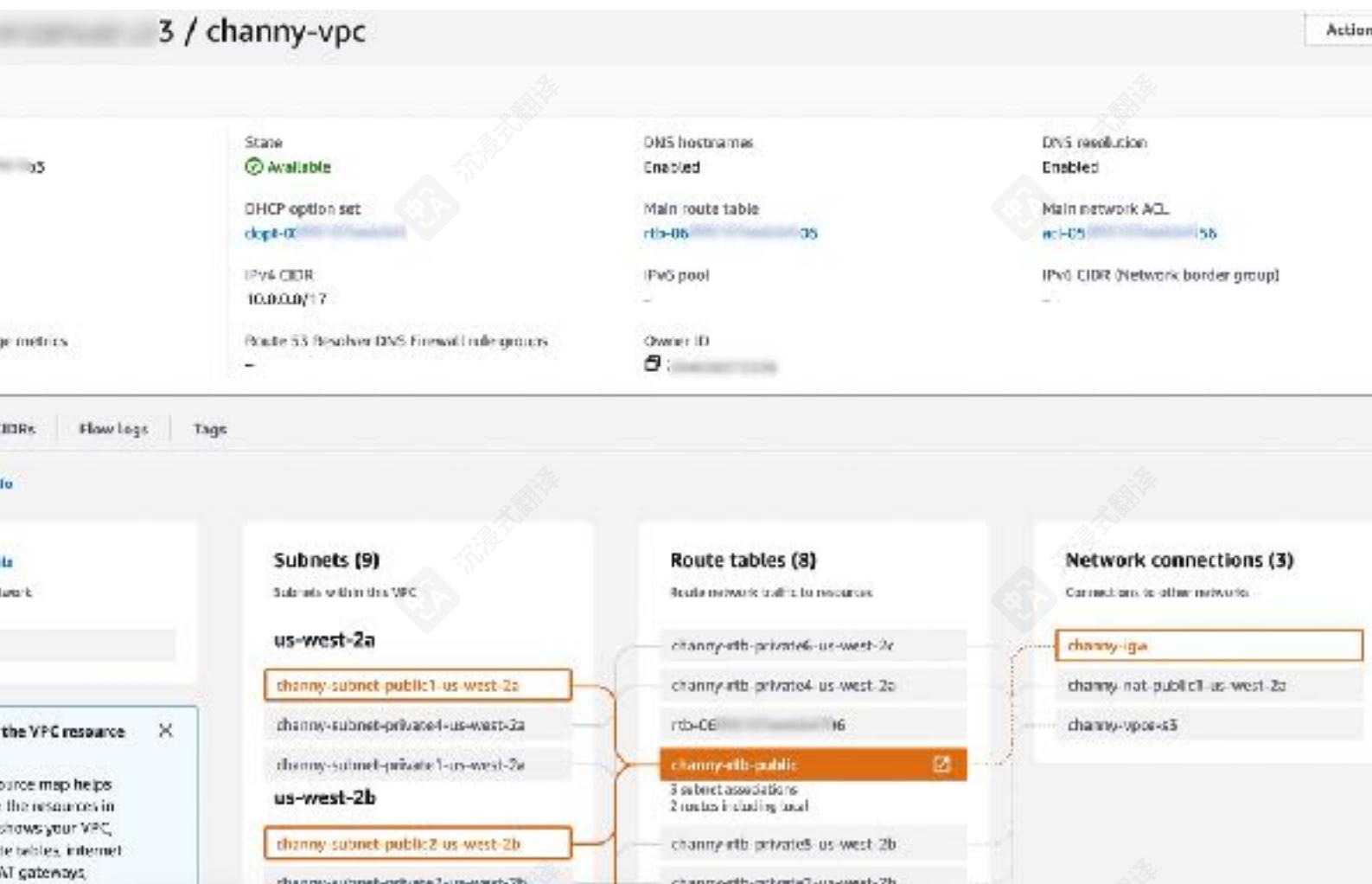
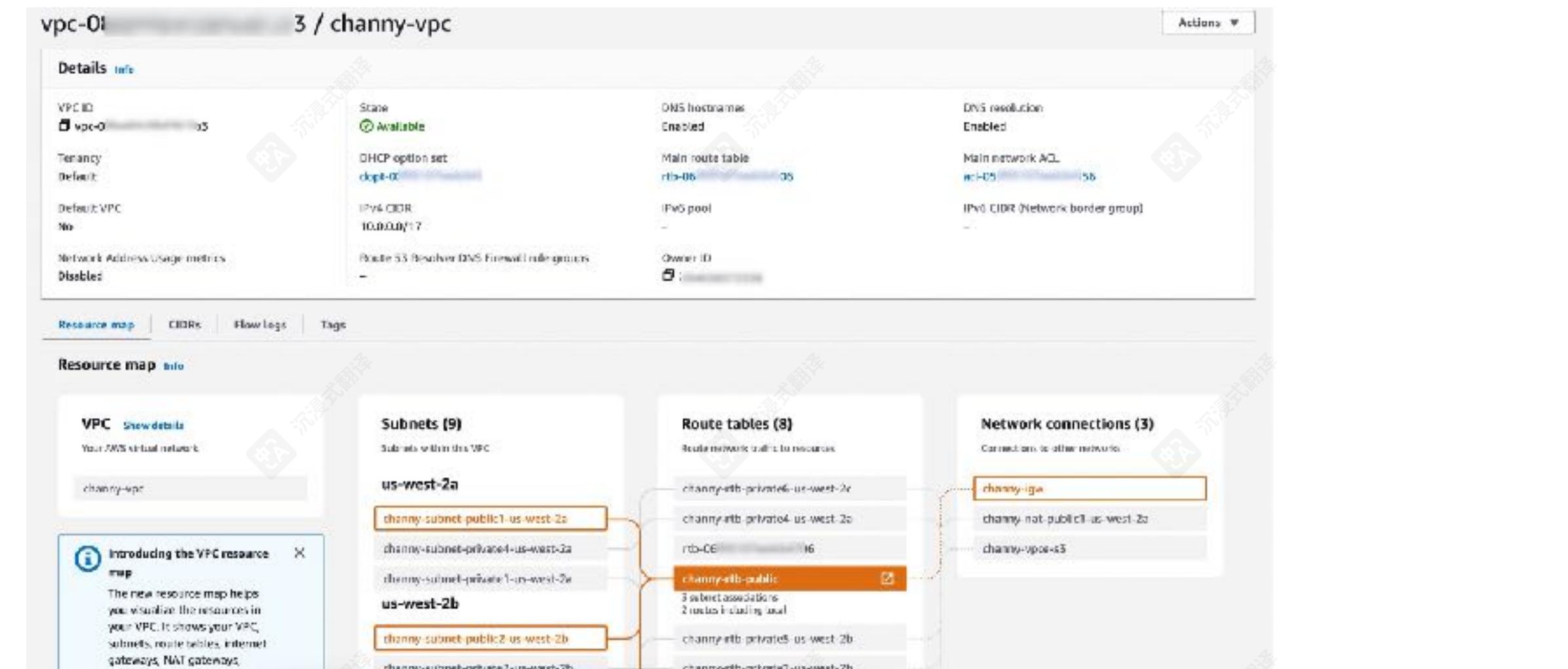
- 增强人类能力



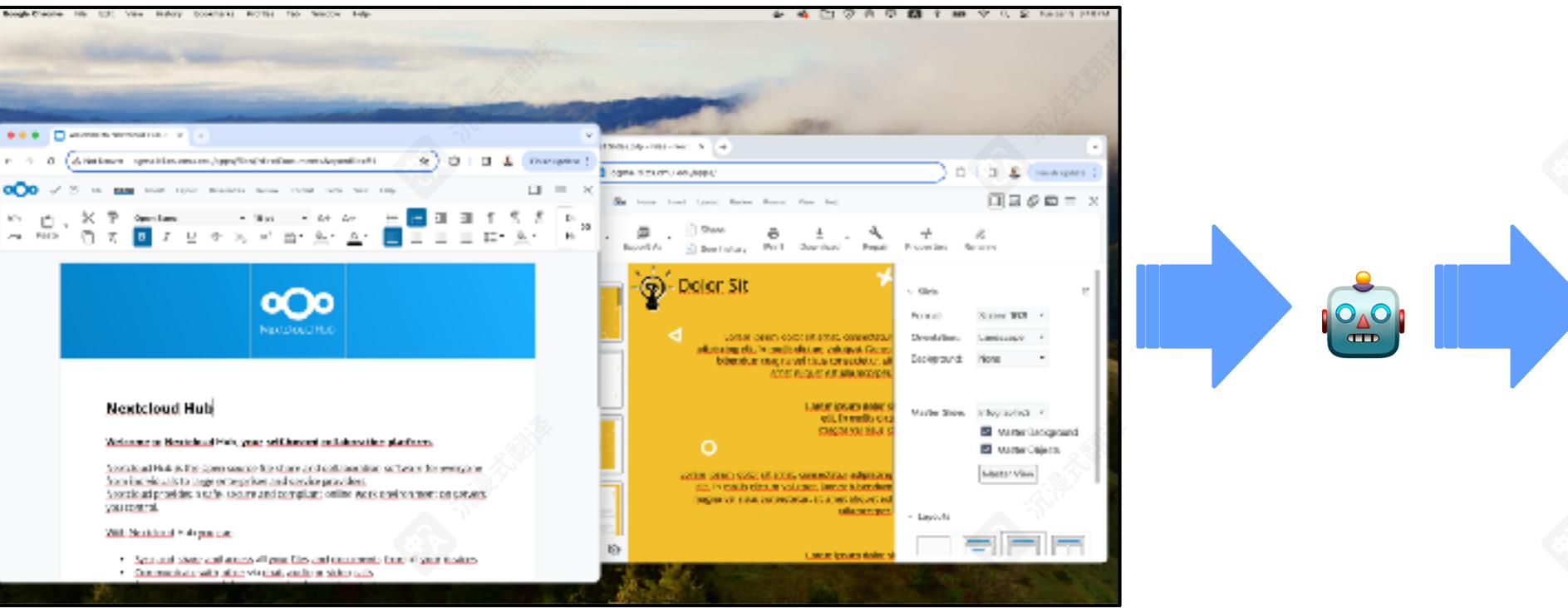
使用 DALLE 生成

Autonomous Agents

自主代理

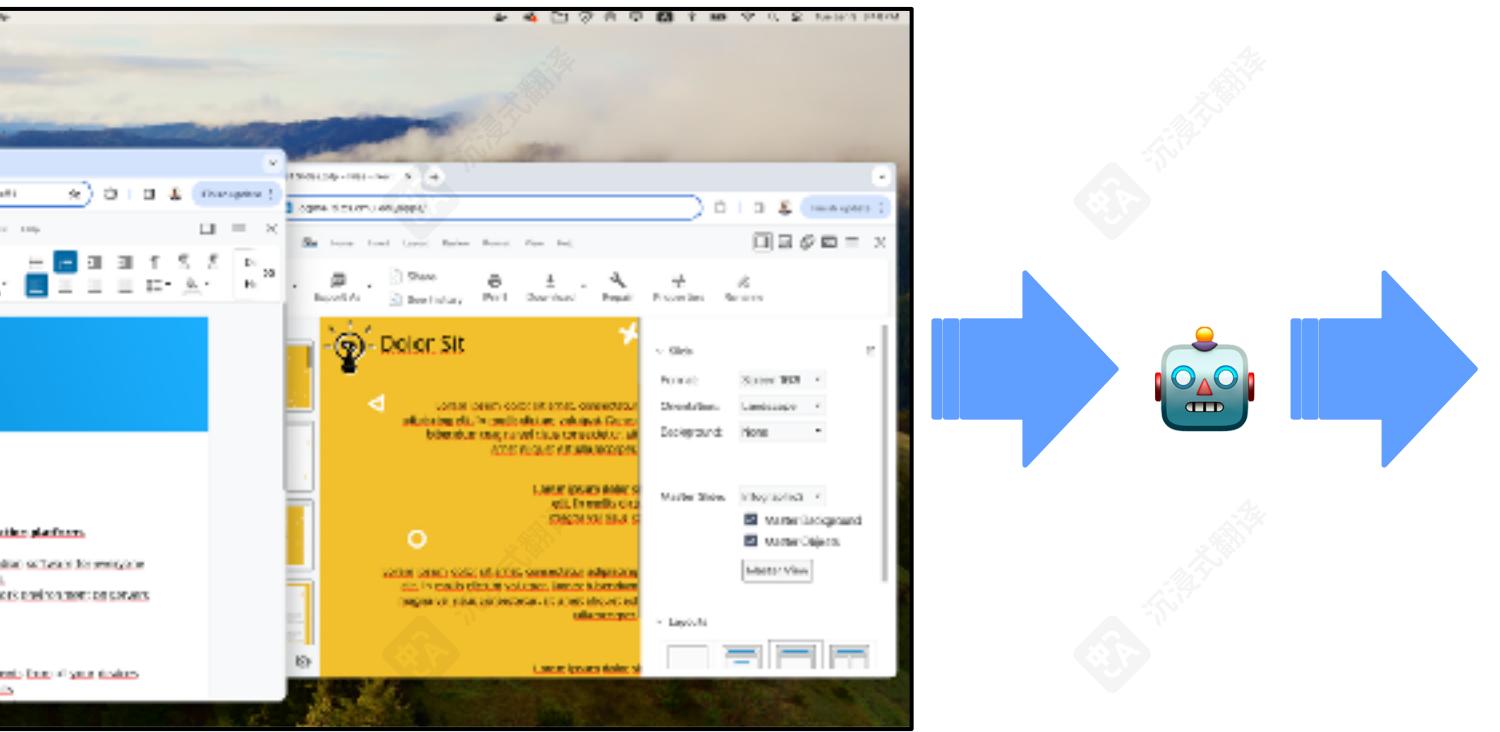


Autonomous Agents



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

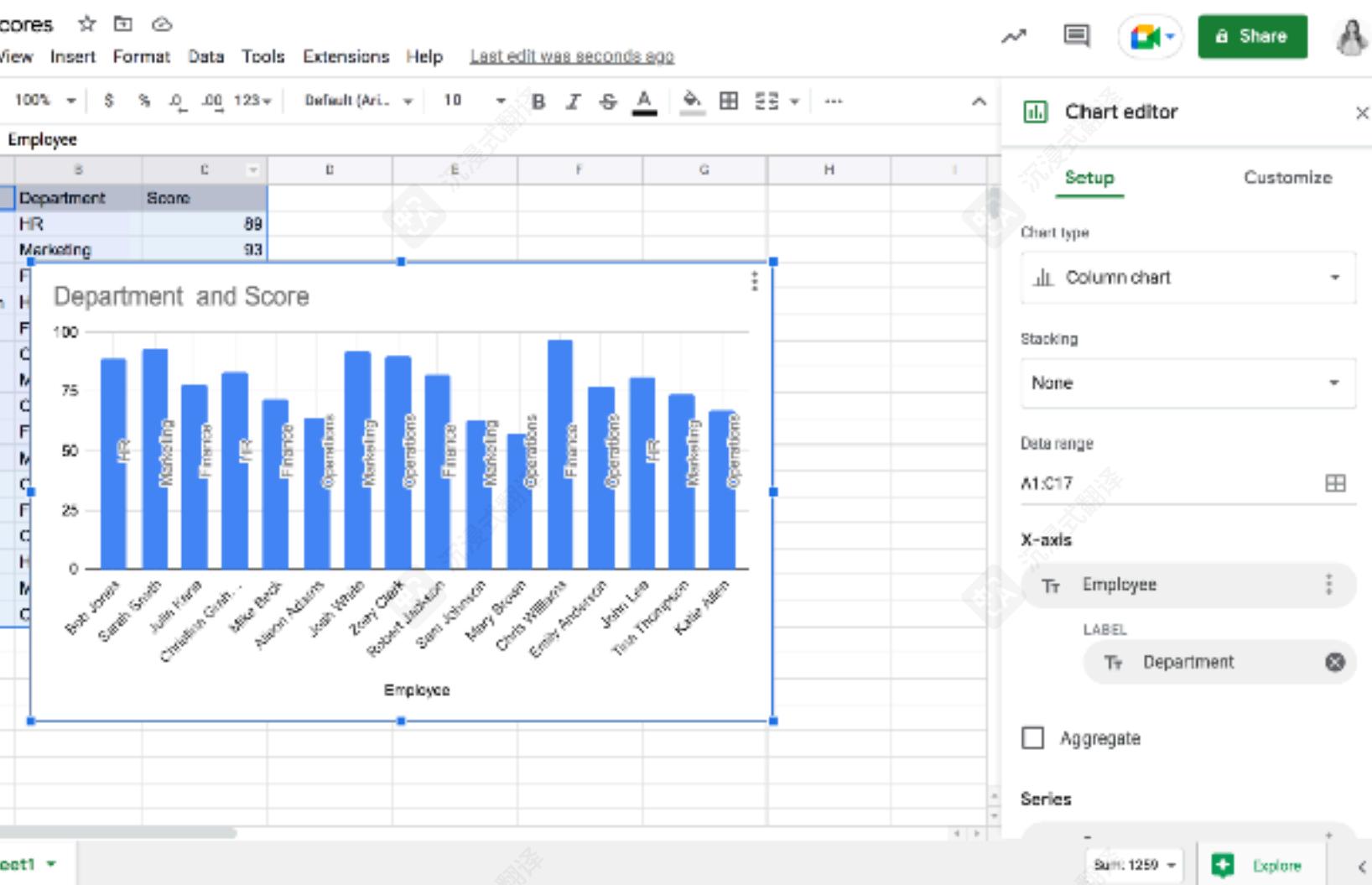
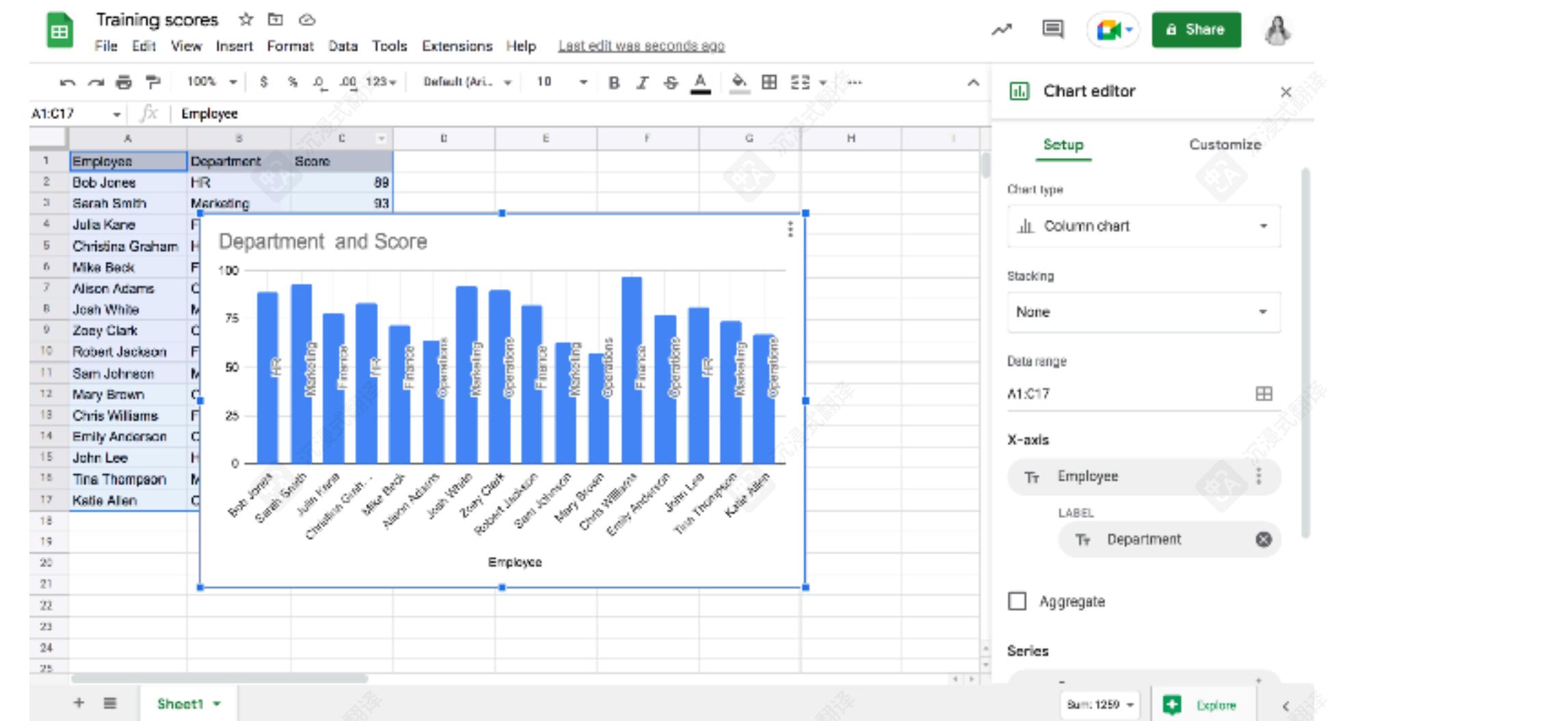
自主代理



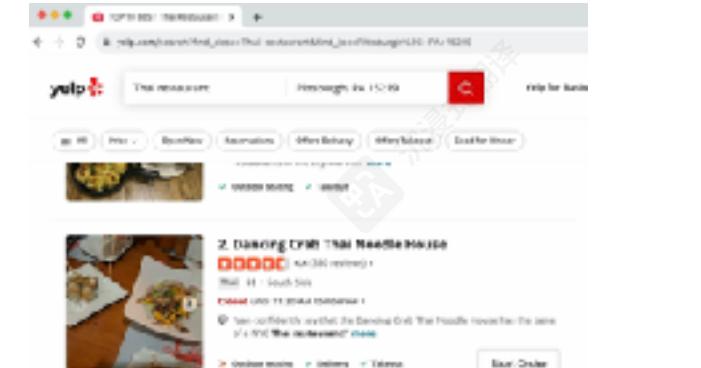
任务：“创建一套 PowerPoint 幻灯片来展示这篇论文的内容。”

Autonomous Agents

自主代理



Web Agents



HTML
Understanding

Web Agents



HTML
Understanding

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

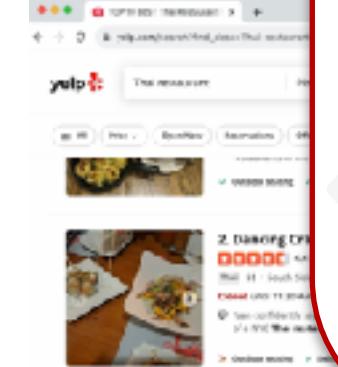
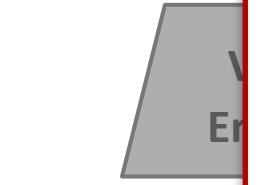
Reiichiro Nakano et al, WebGPT: Browser-assisted Question-Answering with Human Feedback, 2021.

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



Web Agents

Web

Grounding

Shunyu Yao, REACT 在语言模式中协同推理和行动
23s, 20

Jason Wei 等人, 思维链提示激发 La 中的推理
large La
nguage Models, 2022

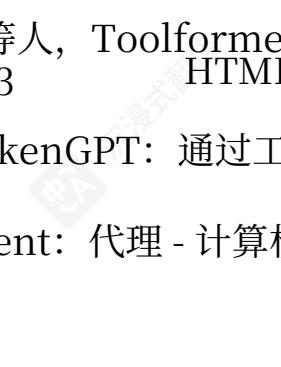
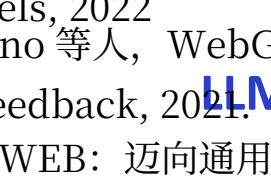
Reiichiro Nakano 等人, WebGPT: 浏览器辅助问答
with Human Feedback, 2021.

邓祥等, MIND2WEB: 迈向通用智能体
2023

Timo Schick 等人, Toolformer: 语言模型可以自我教学
Use Tools, 2023

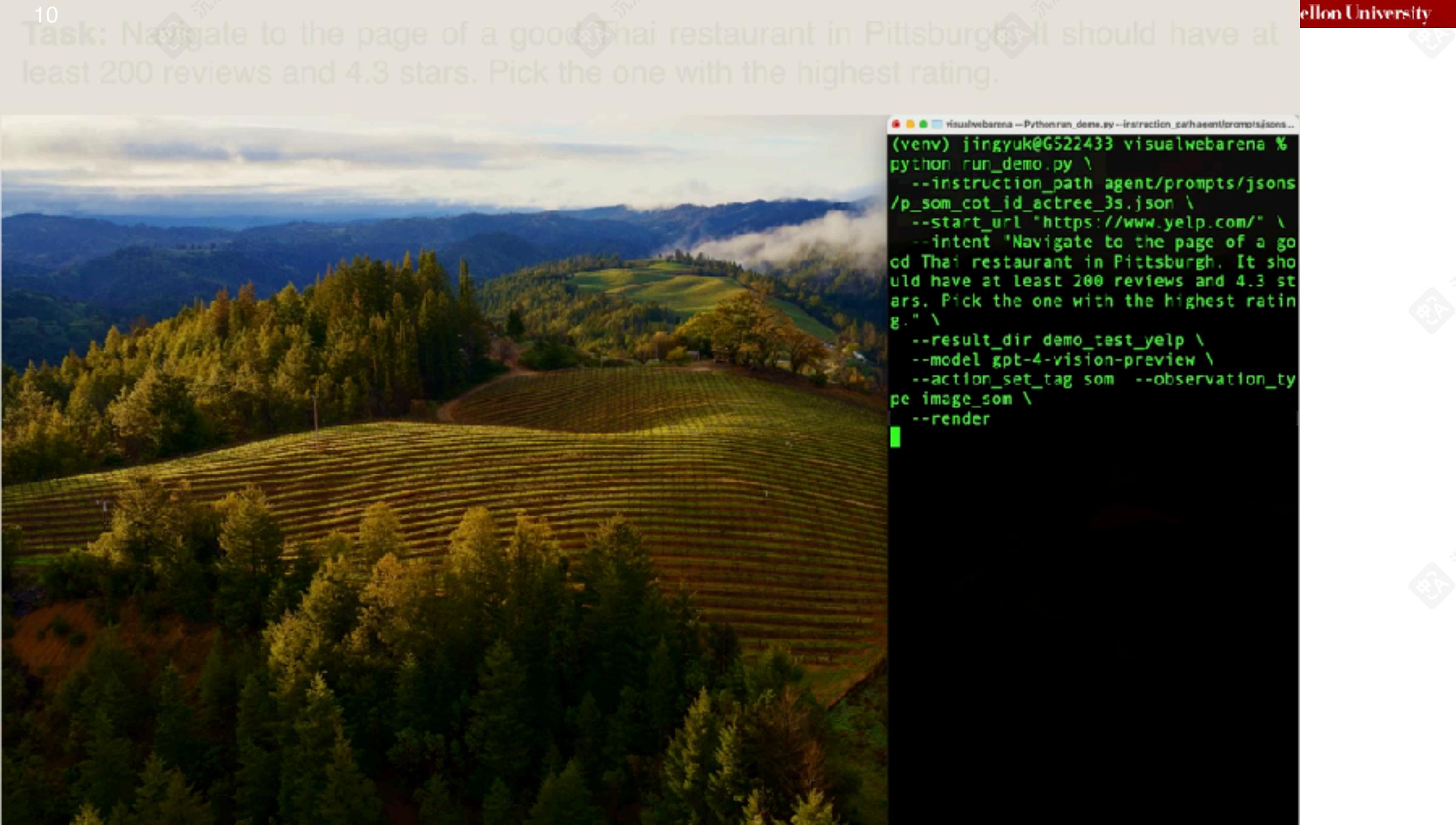
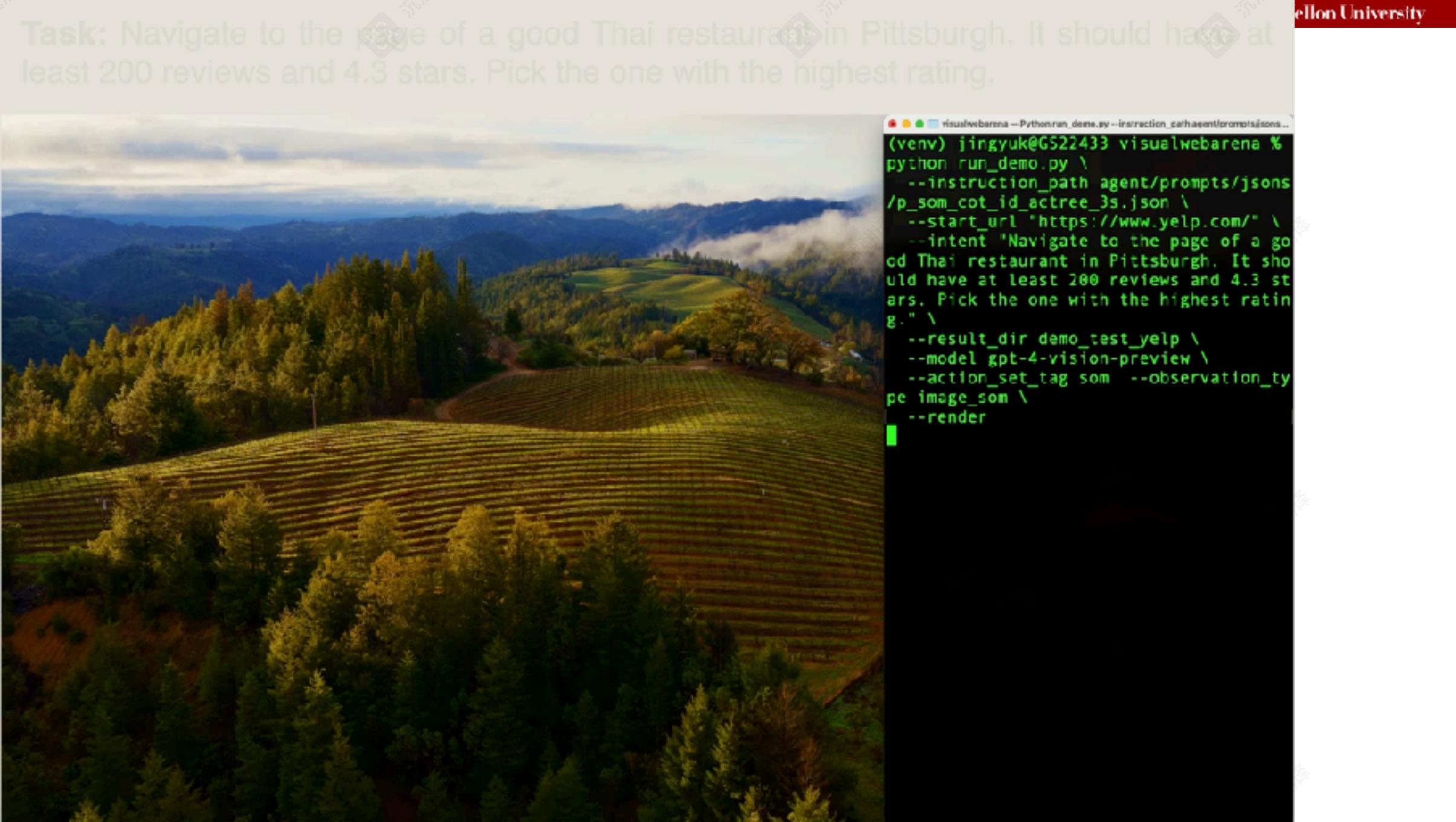
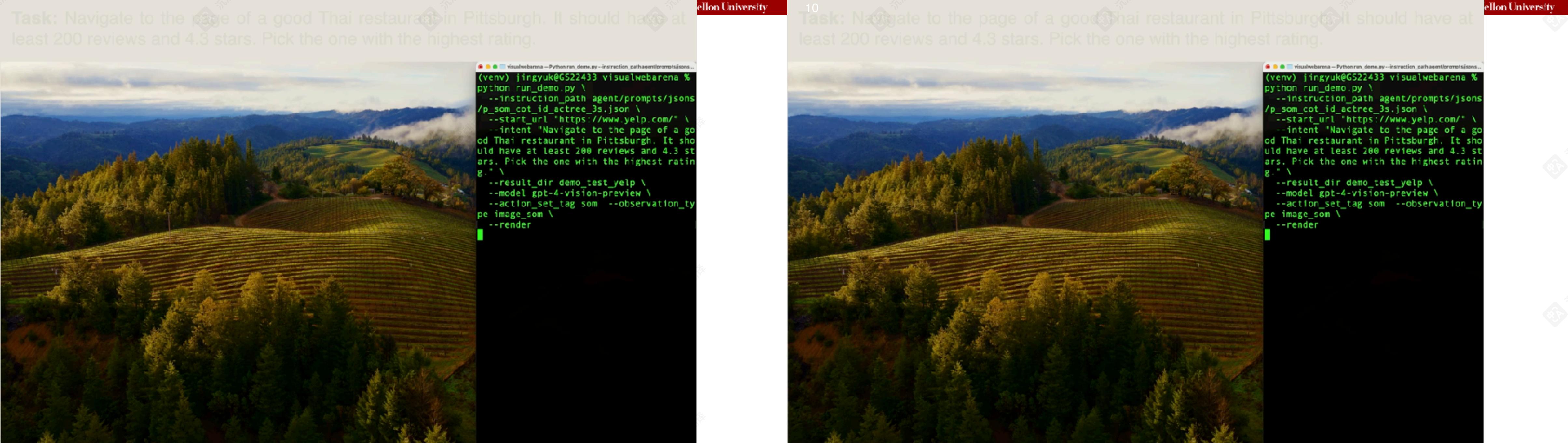
郝世博等, ToolkenGPT: 通过工具嵌入使用海量工具增强冻结语言模型, 2023

杨等 ., SWE-agent: 代理 - 计算机界面实现自动化软件工程 , 2024

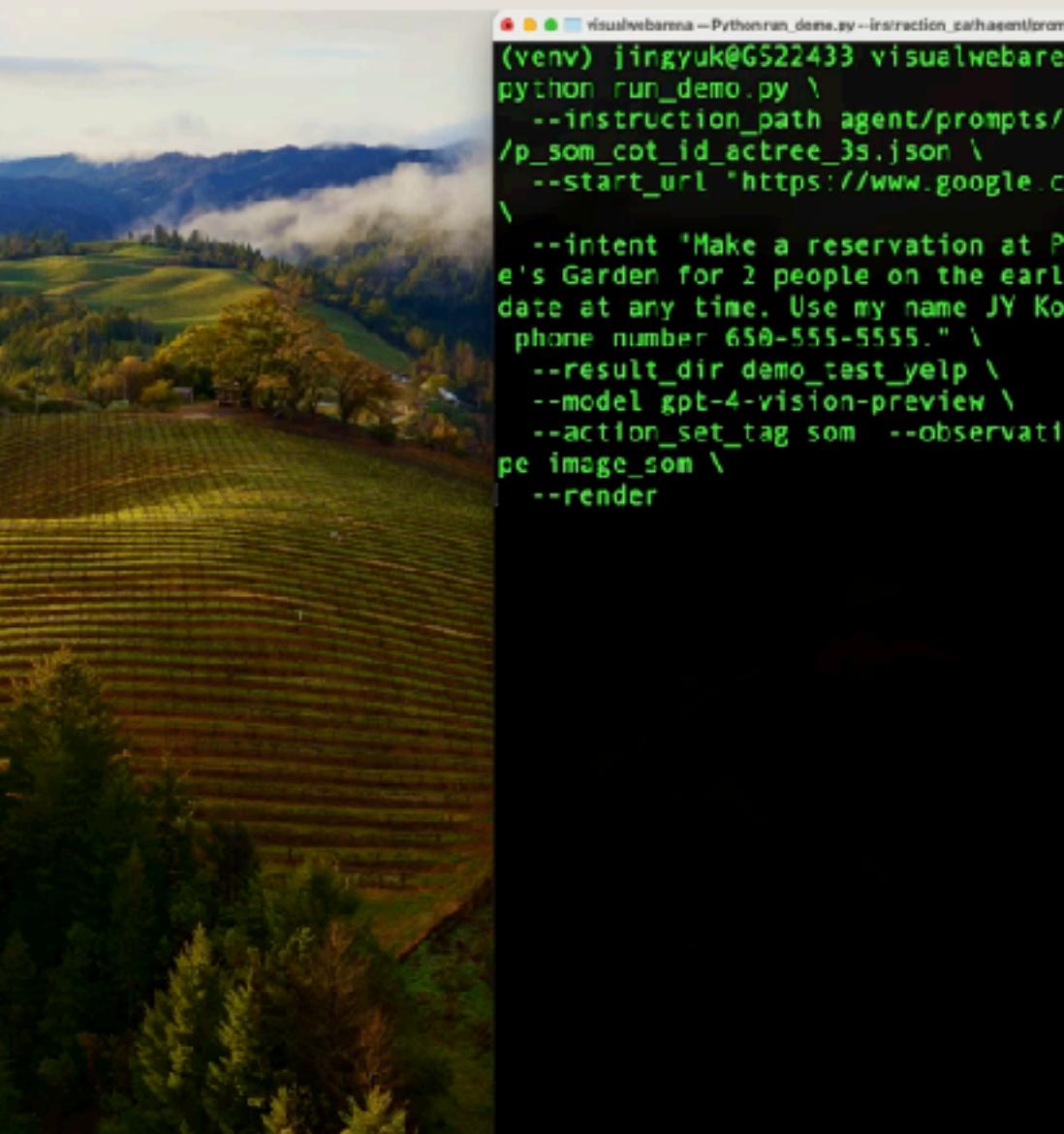


Task: Navigate to a page of a good Thai restaurant in Pittsburgh. It should have at least 200 reviews and 4.3 stars. Pick the one with the highest rating

任务：导航到匹兹堡一家评价良好的泰国餐厅的页面。该餐厅应至少有 200 条评论和 4.3 星。选择评分最高的那家



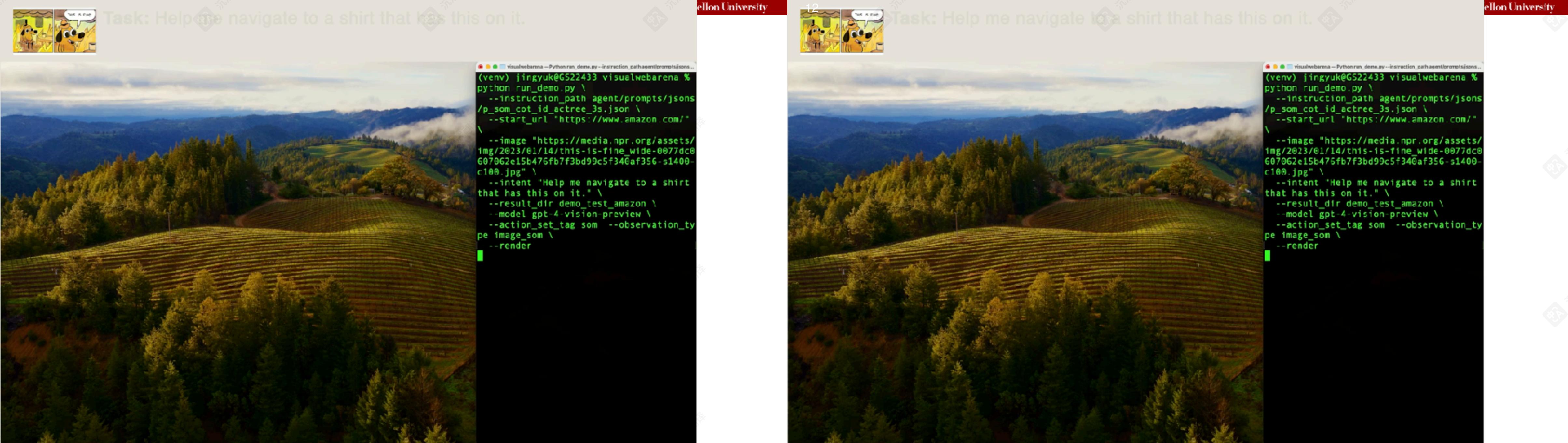
Task: Make a reservation at Pusadee's Garden for 2 people on the earliest date for dinner. Use my name JY Koh and phone number 650-555-5555.



```
visualwebarena -Python run_demo.py --instruction_path agent/prompts/jsons/p_som_cot_id_actree_3s.json \
--start_url "https://www.google.com/" \
--intent "Make a reservation at Pusadee'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_type image_som \
--render
```



```
(venv) jingyuk@G522433 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 Pusadee'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_type image_som \
--render
```



Talk Outline

- VisualWebArena -- Evaluating Multimodal Agents on Realistic Visual Web Tasks (Koh et al., ACL 2024)
- Tree Search for Language Model Agents (Koh, McAleer, Fried, Salakhutdinov, arXiv 2024)
- Towards Internet-Scale Training For Agents (Trabucco, Sigurdsson, Piramuthu, Salakhutdinov, arXiv 2025)

Talk Outline

- VisualWebArena -- 在真实视觉网络任务上评估多模态智能体 (Koh 等人 , ACL 2024)
- 语言模型智能体的树搜索 (Koh, McAleer, Fried, Salakhutdinov, arXiv 2024)
- 面向互联网规模智能体训练 (Trabucco, Sigurdsson, Piramuthu, Salakhutdinov, arXiv 2025)

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

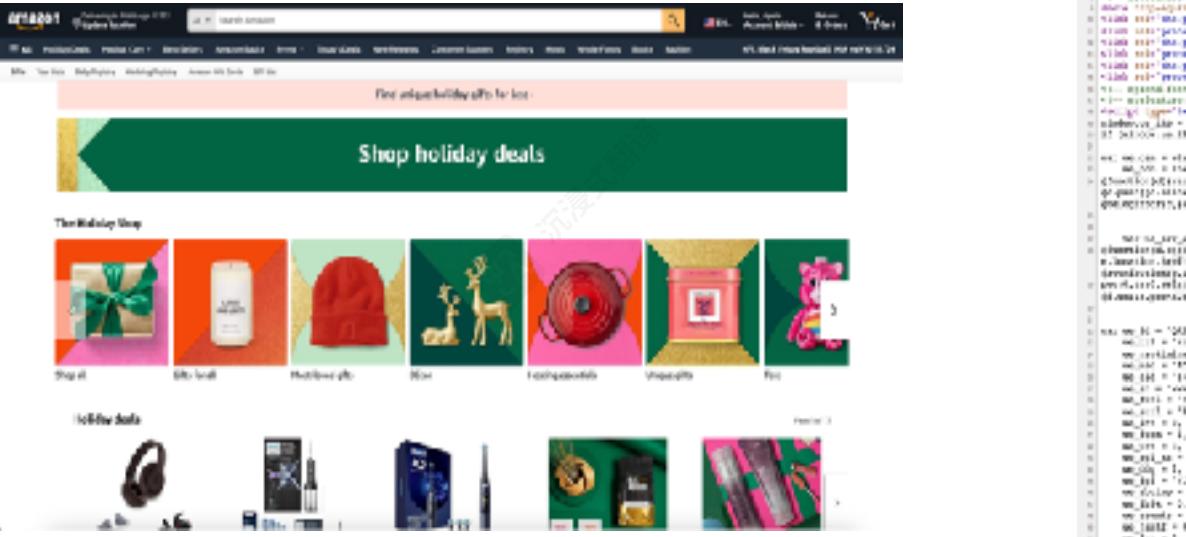
WebArena



Shuyan Zhou FrankXu

- 目前最逼真的网络环境
- 来自热门类别的网站（购物、Reddit、GitLab） - 自托管开源再实现 - 来自真实网站的数据（Amazon、Reddit、GitHub）
- 对人类来说任务很简单（成功率 78%），但对语言模型代理来说很困难（成功率 14%）
- 但：任务设计为仅使用文本和 HTML 源代码
- 混乱的 HTML, JavaScript: 通常为了效率而最小化或压缩
- 交互元素在 HTML 中无法正确显示
 - 例如，页面加载后移动对象的 JavaScript/CSS 代码
- Context length: HTML pages are complex, easily filling up > 100k tokens

HTML is insufficient



```

<div class="grid grid--4x3">
  <a href="#">Shop it</a>
  <a href="#">Gifts</a>
  <a href="#">Multi-gifts</a>
  <a href="#">Box</a>
  <a href="#">Footwear deals</a>
  <a href="#">Polo</a>

```

HTML is insufficient



```

<div class="grid grid--4x3">
  <a href="#">Shop it</a>
  <a href="#">Gifts</a>
  <a href="#">Multi-gifts</a>
  <a href="#">Box</a>
  <a href="#">Footwear deals</a>
  <a href="#">Polo</a>

```

- 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

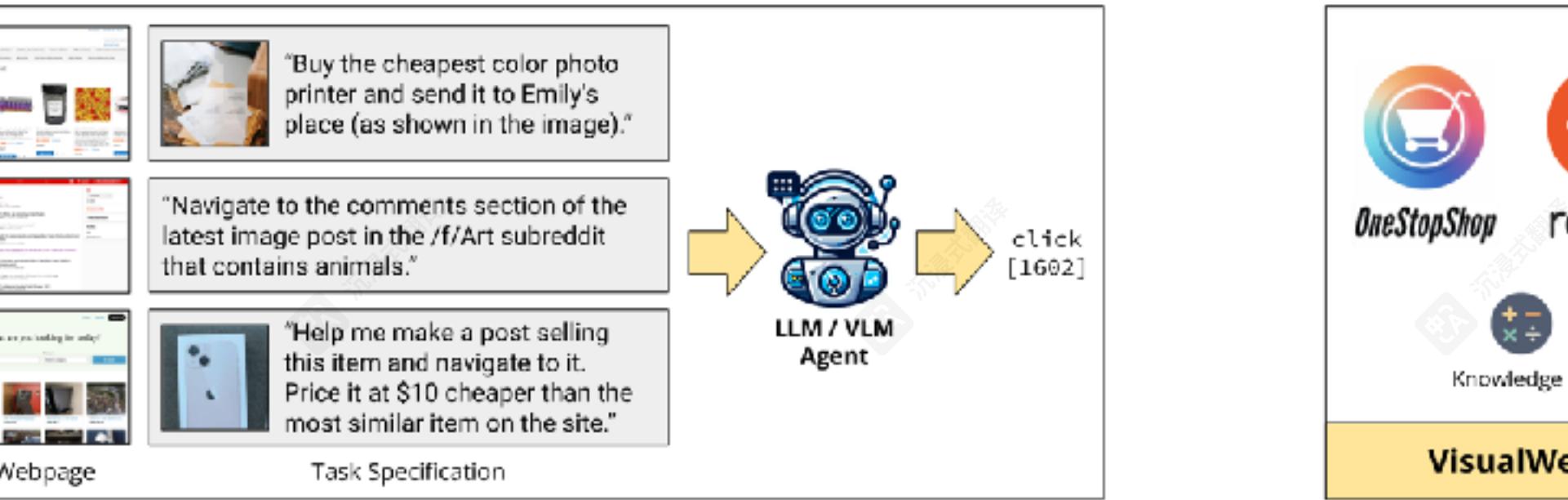
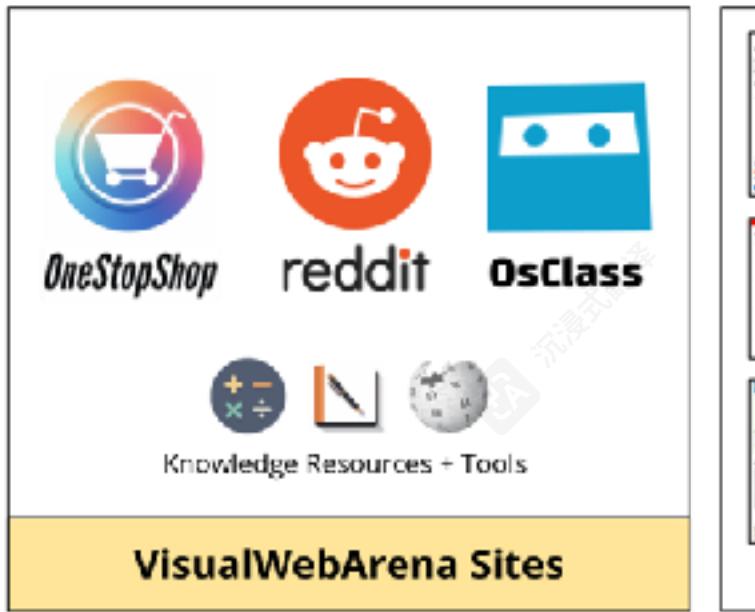
- 乱码 HTML, JavaScript: 通常为了效率会被最小化或压缩
- 交互元素在 HTML 中无法正确显示
 - 例如，页面加载后移动对象的 JavaScript/CSS 代码
 - 空间布局通常也无法很好地传达
- 上下文长度：HTML 页面复杂，容易填满 > 100k 个 token

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



VisualWebArena



JingYu Koh
Shuyan Zhou
Frank Xu

- 构建和跟踪**多模态智能体**
- 我们设计视觉相关的任务来测试这些能力
- 视觉输入（和输出）允许创建独特、有趣且真实的任务



VisualWebArena: Classfields



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

The screenshot shows the OsClass website interface. At the top, there is a navigation bar with the logo, 'My account', 'Logout', and a 'Publish Ad' button. Below the navigation bar, a large search bar asks 'What are you looking for today?'. It contains two input fields: 'Keyword' (with placeholder 'e.g., a blue used car') and 'Category' (with placeholder 'Select a category'). A 'Search' button is located to the right of the category field. Below the search bar, the text 'Latest Listings' is displayed, followed by a grid of four items: 'Nintendo Switch' (\$270.00), 'EII Powered PA Speaker...' (\$150.00), 'xbox series x/ with extras' (\$300.00), and 'Canon EF 100-400mm f/4...' (\$1645.00). Below this grid, there are two more rows of four items each, showing various products like a van, a Marshall speaker, and a shirt.

VisualWebArena: Classfields



任务：找到标价为300-500 美元的这辆自行车，并发表一条评论，报价低于他们的标价 10 美元。

The screenshot shows the OsClass website interface, identical to the one on the left. At the top, there is a navigation bar with the logo, 'My account', 'Logout', and a 'Publish Ad' button. Below the navigation bar, a large search bar asks 'What are you looking for today?'. It contains two input fields: 'Keyword' (with placeholder 'e.g., a blue used car') and 'Category' (with placeholder 'Select a category'). A 'Search' button is located to the right of the category field. Below the search bar, the text 'Latest Listings' is displayed, followed by a grid of four items: 'Nintendo Switch' (\$270.00), 'EII Powered PA Speaker...' (\$150.00), 'xbox series x/ with extras' (\$300.00), and 'Canon EF 100-400mm f/4...' (\$1645.00). Below this grid, there are two more rows of four items each, showing various products like a van, a Marshall speaker, and a shirt.

VisualWebArena: Shopping



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

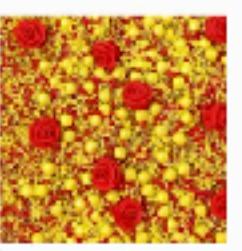
One Stop Market

Product Showcases

 Pro-baked Gingersnap Hobble Kit Value Pouch, 12 oz, Pack of 2, Total 34 oz. ★★★★★ 5 reviews \$19.98 Add to Cart	 Y5 Energy Healthy Energy Drink, 500mg Caffeine from Black and Green Tea, Pomegranate, Blueberry, & Orange, 12 oz, Pack of 24 ★★★★★ 12 reviews \$54.47 Add to Cart	 Unsweetened Iced Tea, Orange, Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch ★★★★★ 4 reviews \$19.96 Add to Cart	 So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project Verified, 9 oz, Tub ★★★★★ 12 reviews \$19.82 Add to Cart	 Unsweetened Iced Tea, Orange, Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch ★★★★★ 4 reviews \$23.99 Add to Cart
---	---	--	---	--

One Stop Market

Product Showcases

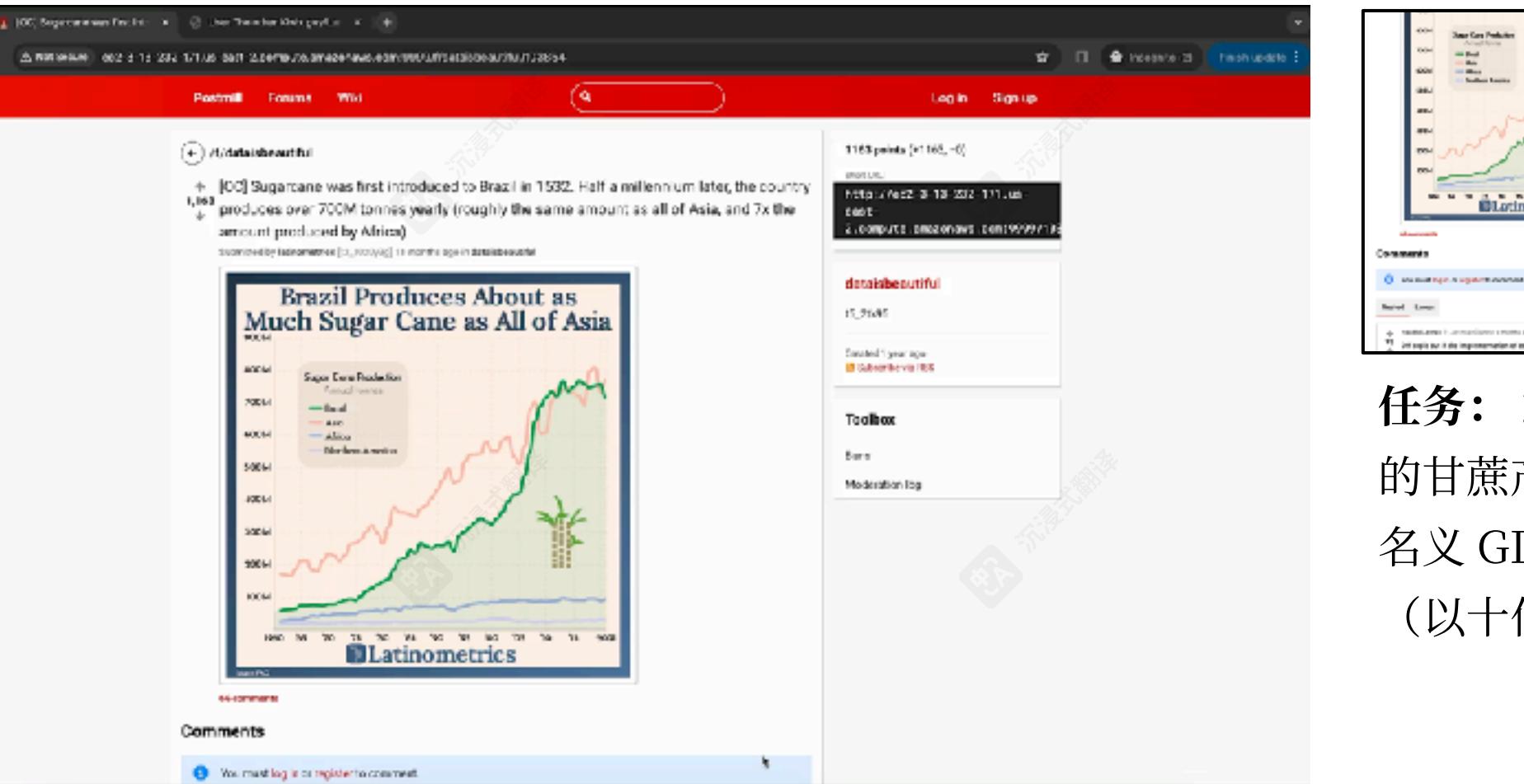
 Pro-baked Gingersnap Hobble Kit Value Pouch, 12 oz, Pack of 2, Total 34 oz. ★★★★★ 5 reviews \$19.98 Add to Cart	 Y5 Energy Healthy Energy Drink, 500mg Caffeine from Black and Green Tea, Pomegranate, Blueberry, & Orange, 12 oz, Pack of 24 ★★★★★ 12 reviews \$54.47 Add to Cart	 Unsweetened Iced Tea, Orange, Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch ★★★★★ 4 reviews \$19.96 Add to Cart	 So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project Verified, 9 oz, Tub ★★★★★ 12 reviews \$19.82 Add to Cart	 Unsweetened Iced Tea, Orange, Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch ★★★★★ 4 reviews \$23.99 Add to Cart
---	---	--	---	--

任务：购买最便宜的有色照片打印机，并将其送到Emily的地方（如图片所示）。

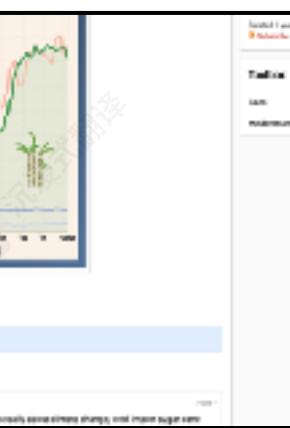
VisualWebArena: Reddit



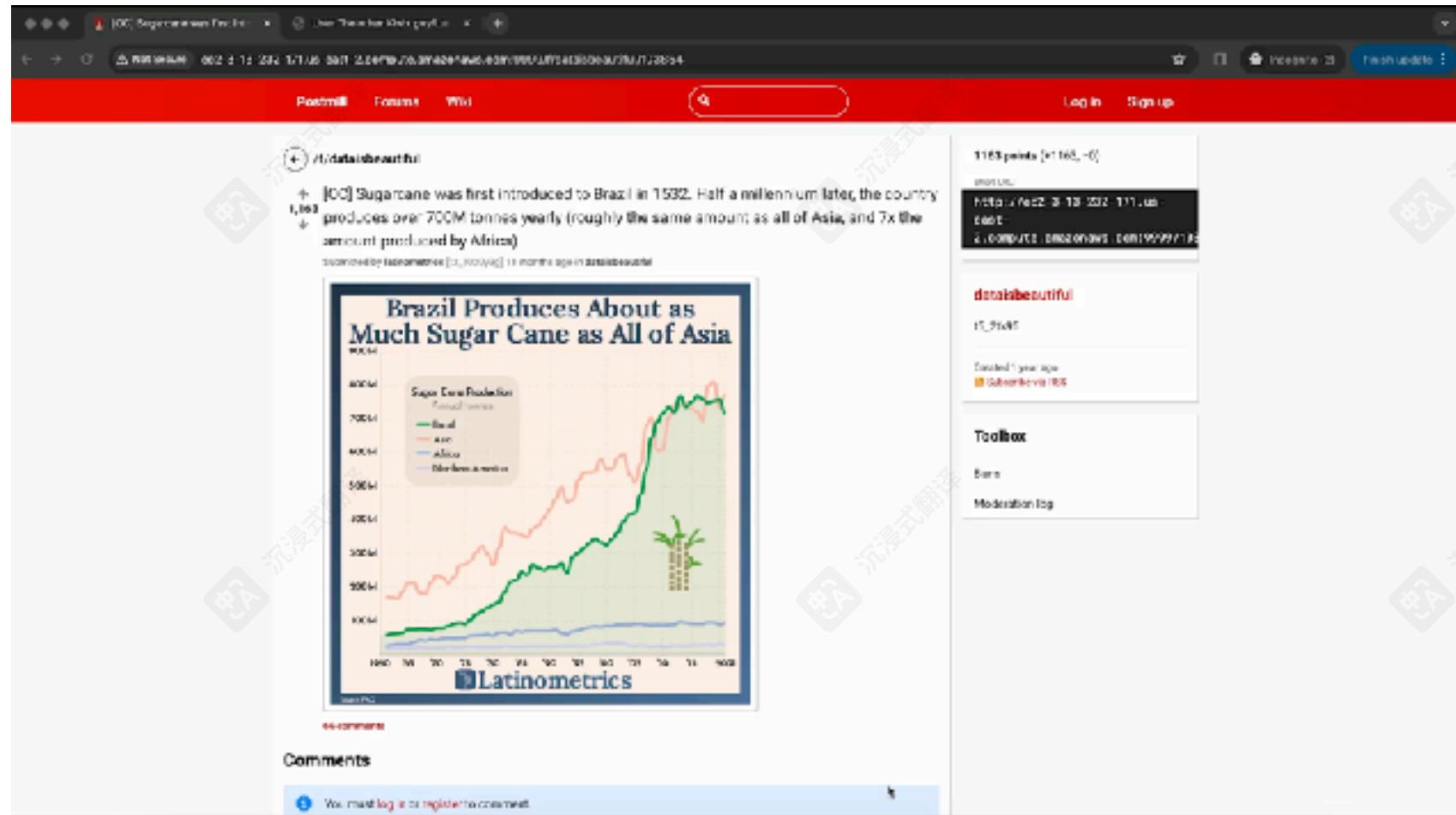
Task: What is the 2022 total nominal GDP of the area that produces most sugarcane in the year of 2021? (in billion)?



VisualWebArena: Reddit



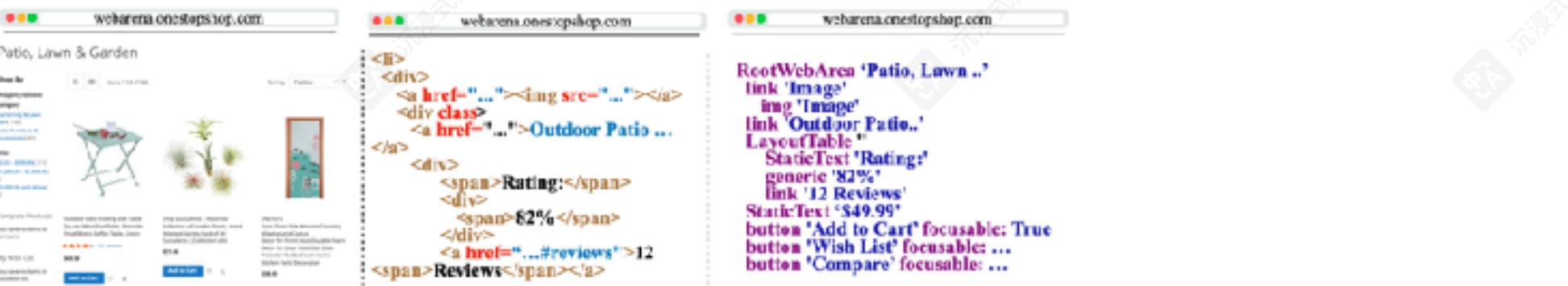
任务：2021 年产量最高的甘蔗产区在 2022 年的名义 GDP 总额是多少？(以十亿为单位)



VisualWebArena

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

- Observations \mathcal{O}



- 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

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

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

VisualWebArena

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

- Observations \mathcal{O}



- 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.

- 确定性转换函数

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

- 奖励函数: $r(\mathbf{a}, \mathbf{s})$

Image Inputs:



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

Shopping

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 > Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier

Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier

INSTOCK SKU: B03STUQ6C
Be the first to review this product

\$2.56

Qty: 1 Add to Cart

Add to Wish List Add to Compare

A screenshot of a shopping website for "One Stop Market". The main focus is a product listing for the "Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier". The product is shown in a large image at the top, followed by its name and a brief description. Below this, there is information about availability ("INSTOCK"), the product code ("SKU: B03STUQ6C"), and a call-to-action button ("Add to Cart"). There are also links for "Add to Wish List" and "Add to Compare". The page includes a search bar and navigation links for various categories like Beauty & Personal Care, Sports & Outdoors, and Home & Kitchen.

图像输入:



任务: “购买最便宜的有色照片打印机并将其送到 Emily 的地方（如图片所示）。

购物

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 > Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier

Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier

INSTOCK SKU: B03STUQ6C
Be the first to review this product

\$2.56

Qty: 1 Add to Cart

Add to Wish List Add to Compare

A screenshot of the same shopping website for "One Stop Market" as the first image. It shows the same product listing for the "Canon PIXMA MG2120 Color Photo Printer with Scanner and Copier". The product image, name, and price are identical. The page layout, including the header, navigation links, and footer, is also consistent with the first screenshot.



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

One Stop Market

Product Showcases

- My-Baked Underwear Housewife Pant Pack, 12x6, Pack of 2, Total 36pc. ★★★★☆ 12 Reviews \$10.98
- VS Energy Infused Tea, Orange, Green Tea, Pomegranate, Decaf, 6 Cans, Case, Pack of 24. ★★★★☆ 12 Reviews \$14.47
- ELIMINATE THE BAIT PREDATOR! Non-Alcohol Colorful Sprinkles | Cafe Decafake Cookie Sprinkles | for Green Gummy Sprinkles | Yellow Gold Red Royal Red Rose Icing Flower Decoupage Sprinkles, 8oz. ★★★★☆ 12 Reviews \$10.42
- DISGRIMOUS DAIRY FREE LOKAWHO Light, Vegan, Free-GMO Project Verified, 9.33, Taro

Printers & Accessories

Shop By

Shopping Systems

Price

Compare Products

Recently Ordered

- My-Baked Underwear Housewife Pant Pack, 12x6, Pack of 2, Total 36pc. ★★★★☆ 12 Reviews \$10.98
- VS Energy Infused Tea, Orange, Green Tea, Pomegranate, Decaf, 6 Cans, Case, Pack of 24. ★★★★☆ 12 Reviews \$14.47
- ELIMINATE THE BAIT PREDATOR! Non-Alcohol Colorful Sprinkles | Cafe Decafake Cookie Sprinkles | for Green Gummy Sprinkles | Yellow Gold Red Royal Red Rose Icing Flower Decoupage Sprinkles, 8oz. ★★★★☆ 12 Reviews \$10.42
- DISGRIMOUS DAIRY FREE LOKAWHO Light, Vegan, Free-GMO Project Verified, 9.33, Taro



任务：“购买最便宜的有色照片打印机并将其送到 Emily 的地方（如图片所示）。

One Stop Market

Product Showcases

- My-Baked Underwear Housewife Pant Pack, 12x6, Pack of 2, Total 36pc. ★★★★☆ 12 Reviews \$10.98
- VS Energy Infused Tea, Orange, Green Tea, Pomegranate, Decaf, 6 Cans, Case, Pack of 24. ★★★★☆ 12 Reviews \$14.47
- ELIMINATE THE BAIT PREDATOR! Non-Alcohol Colorful Sprinkles | Cafe Decafake Cookie Sprinkles | for Green Gummy Sprinkles | Yellow Gold Red Royal Red Rose Icing Flower Decoupage Sprinkles, 8oz. ★★★★☆ 12 Reviews \$10.42
- DISGRIMOUS DAIRY FREE LOKAWHO Light, Vegan, Free-GMO Project Verified, 9.33, Taro

Printers & Accessories

Shop By

Shipping Systems

Price

Compare Products

Recently Ordered

- My-Baked Underwear Housewife Pant Pack, 12x6, Pack of 2, Total 36pc. ★★★★☆ 12 Reviews \$10.98
- VS Energy Infused Tea, Orange, Green Tea, Pomegranate, Decaf, 6 Cans, Case, Pack of 24. ★★★★☆ 12 Reviews \$14.47
- ELIMINATE THE BAIT PREDATOR! Non-Alcohol Colorful Sprinkles | Cafe Decafake Cookie Sprinkles | for Green Gummy Sprinkles | Yellow Gold Red Royal Red Rose Icing Flower Decoupage Sprinkles, 8oz. ★★★★☆ 12 Reviews \$10.42
- DISGRIMOUS DAIRY FREE LOKAWHO Light, Vegan, Free-GMO Project Verified, 9.33, Taro

Step 0: Start on the homepage of OneStopMarket.

Step 1: Navigate to the printers category.

步骤 0：从 OneStopMarket 的主页开始。

步骤 1：导航到打印机类别。



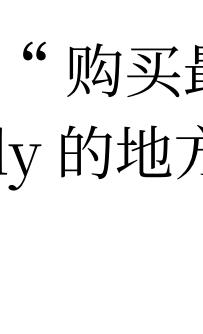
Task: “Buy the cheapest color photo printer and send it to Emily's place (as shown in the image).”

This screenshot shows a list of printers on a website. The top navigation bar includes categories like BEAUTY & PERSONAL CARE, SPORTS & OUTDOOR, WILDLIFE, TOYS & JEWELRY, HOME & KITCHEN, OFFICE PRODUCTS, and HOME & ROOM IMPROVEMENT. Below the navigation, there are filters for Shipping Options (Free, \$0.00 - \$1,000.00, \$1,000.00 and above) and a price range (\$0.00 - \$1,000.00, \$1,000.00 and above). The main content area displays four printer models: a white printer with a blue top, a black printer with a blue top, a black printer with a blue top, and a white printer with a blue top. Each item has a price, a 'Compare' link, and 'Add to Cart' buttons.

Step 2: Sort by descending price.

This screenshot shows a product page for the Canon PIXMA MG2120 Color Photo Printer. The page includes a large image of the printer, its name, a price of \$2.55, and a 'Buy It Now' button. Below the main image are two smaller images of the printer. The page also features a 'Compare' link and 'Add to Cart' buttons.

Step 3: Click on the cheapest color photo printer.



任务：“购买最便宜的有色照片打印机并将其送到 Emily 的地方（如图片所示）。

This screenshot shows a list of printers on a website, identical to the one in the previous step. It includes the same navigation bar, filters, and printer models. The prices are listed as \$2.55 for the Canon PIXMA MG2120.

步骤 2：按降序价格排序。

This screenshot shows a product page for the Canon PIXMA MG2120 Color Photo Printer, identical to the one in the previous step. It includes the same large image, price of \$2.55, and 'Buy It Now' button.

步骤 3：点击最便宜的有色照片打印机。



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 Pet Supplies Electronics Dell Monitors & Accessories Video Games Grocery & Gourmet Food

Shopping Cart

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

Move to Wishlist Edit Removal

Customer Shopping Update Shopping Cart Proceed to Checkout Check Out with Multiple Addresses

Privacy and Cookie Policy Search Terms

One Stop Market

Shipping Review & Payments

Shipping Address

Emily Lopez
7713 5th MILES DR
San Mateo, California 94401
United States
BOSTON

Order Summary

1 Item in Cart

Subtotal \$2.56
Shipping (Flat Rate - Fixed) \$0.00
Order Total \$2.56
Apply Discount/Coupons

Proceed to Checkout Next

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 Pet Supplies Electronics Dell Monitors & Accessories Video Games Grocery & Gourmet Food

Shopping Cart

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

Move to Wishlist Edit Removal

Customer Shopping Update Shopping Cart Proceed to Checkout Check Out with Multiple Addresses

Privacy and Cookie Policy Search Terms

One Stop Market

Shipping Review & Payments

Shipping Address

Emily Lopez
7713 5th MILES DR
San Mateo, California 94401
United States
BOSTON

Order Summary

1 Item in Cart

Subtotal \$2.56
Shipping (Flat Rate - Fixed) \$0.00
Order Total \$2.56
Apply Discount/Coupons

Proceed to Checkout Next

Step 4: Add it to the shopping cart.

Step 5: Proceed to checkout



任务：“购买最便宜的有色照片打印机并将其送到 Emily 的地方（如图片所示）。

步骤 4：添加到购物车。

步骤 5：继续结账



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

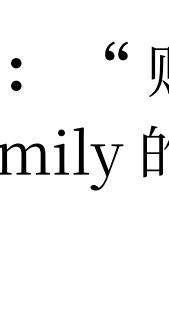
One Stop Market
Shipping
Shipping Address
To Name * Emily
List Name * Casper
Company
Direct Address * USWBS, UHtown
Country * United States
State/Province * New York
City * Marmesse
Zip/Postal Code * 12945
House Number * USWBS123-2
 Save in address book
Cancel **SHIP HERE**

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

One Stop Market
Shipping Address
To Name * Emily
List Name * Casper
Company
Direct Address * USWBS, UHtown
Country * United States
State/Province * New York
City * Marmesse
Zip/Postal Code * 12945
House Number * USWBS123-2
 Save in address book
Cancel **SHIP HERE**

Thank you for your purchase!
Your order number is: 040000198.
We'll email you an order confirmation with details and tracking info.
Continue Shopping

Step 7: Place the order



任务：“购买最便宜的有色照片打印机并将其送到 Emily 的地方（如图片所示）。

One Stop Market
Shipping
Shipping Address
To Name * Emily
List Name * Casper
Company
Direct Address * USWBS, UHtown
Country * United States
State/Province * New York
City * Marmesse
Zip/Postal Code * 12945
House Number * USWBS123-2
 Save in address book
Cancel **SHIP HERE**

One Stop Market
Shipping Address
To Name * Emily
List Name * Casper
Company
Direct Address * USWBS, UHtown
Country * United States
State/Province * New York
City * Marmesse
Zip/Postal Code * 12945
House Number * USWBS123-2
 Save in address book
Cancel **SHIP HERE**

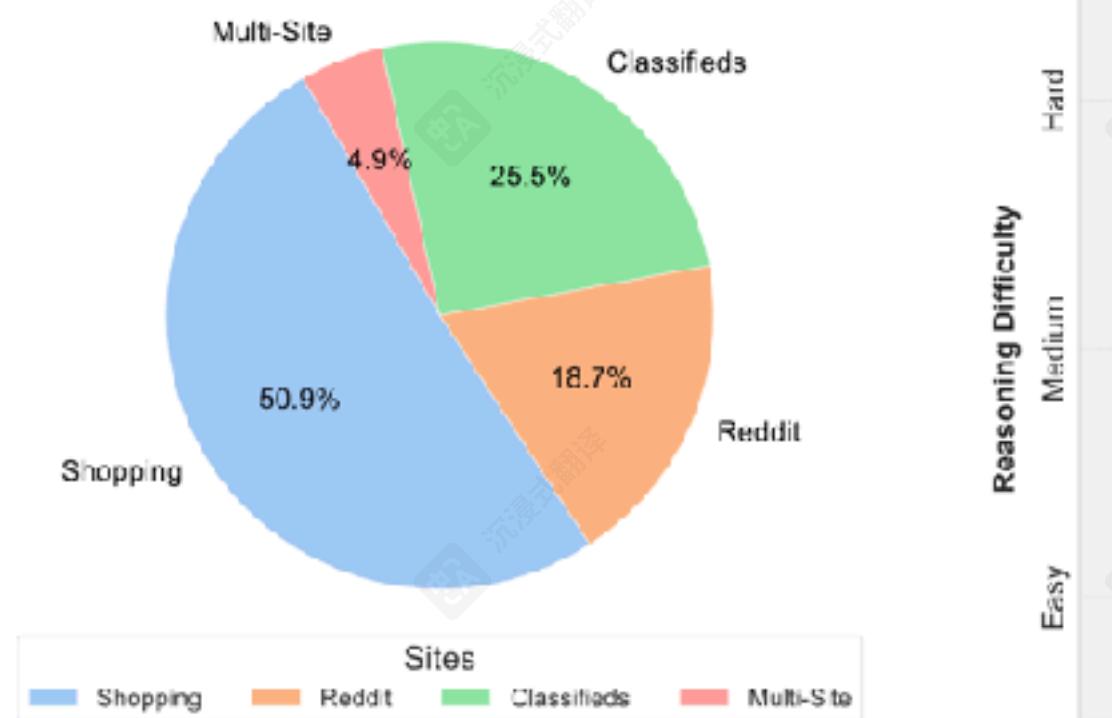
Thank you for your purchase!
Your order number is: 040000198.
We'll email you an order confirmation with details and tracking info.
Continue Shopping

步骤 6：编辑地址为 Emily 的地方。

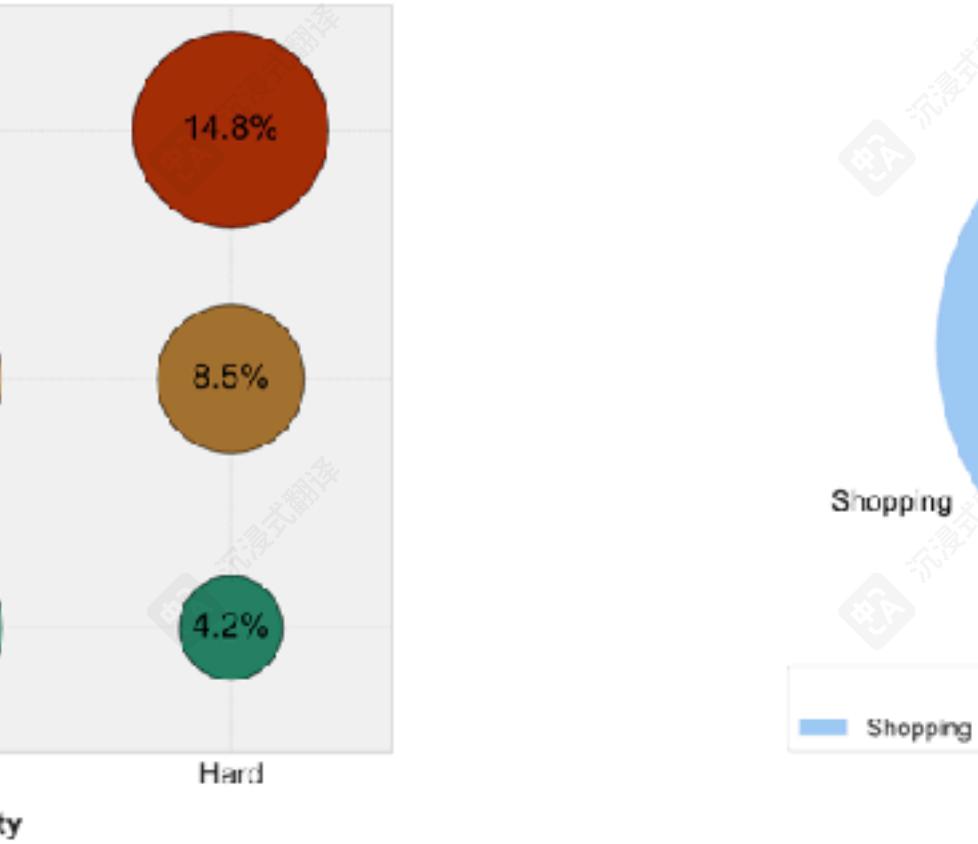
步骤 7：下单

VisualWebArena

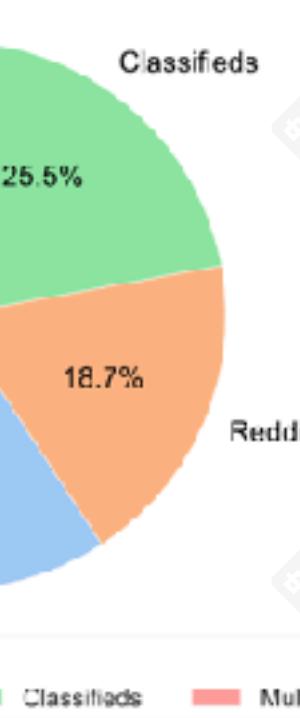
Distribution of Tasks Across Sites



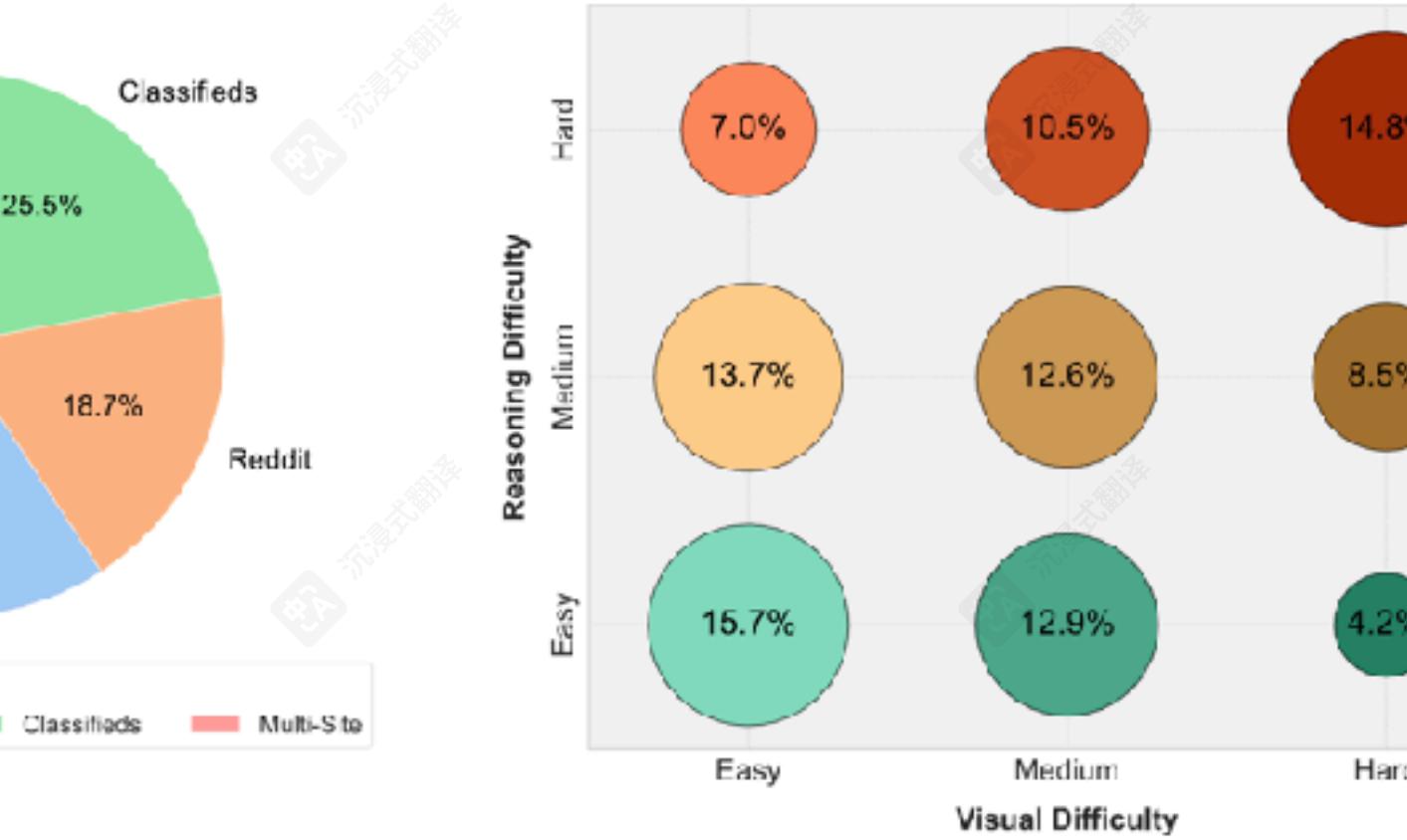
Distribution of Tasks by Difficulty



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.	<code>exact_match(\$a, "US0378331005")</code>
	Add something like what the man is wearing to my wish list.	<code>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")</code>
	Create a post for each of the following images in the most related forums.	<code>eval_fuzzy_image_match(\$s, \$a*)</code>
	Navigate to my listing of the white car and change the price to \$25000. Update the price in the description as well.	<code>url="/index.php?page=item&id=84144" must_include(\$a, "\$25000 OR \$25,000") must_exclude(\$a, "\$30000 OR \$30,000")</code>

基于执行的评估

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.	<code>exact_match(\$a, "US0378331005")</code>
	Add something like what the man is wearing to my wish list.	<code>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")</code>
	Create a post for each of the following images in the most related forums.	<code>eval_fuzzy_image_match(\$s, \$a*)</code>
	Navigate to my listing of the white car and change the price to \$25000. Update the price in the description as well.	<code>url="/index.php?page=item&id=84144" must_include(\$a, "\$25000 OR \$25,000") must_exclude(\$a, "\$30000 OR \$30,000")</code>

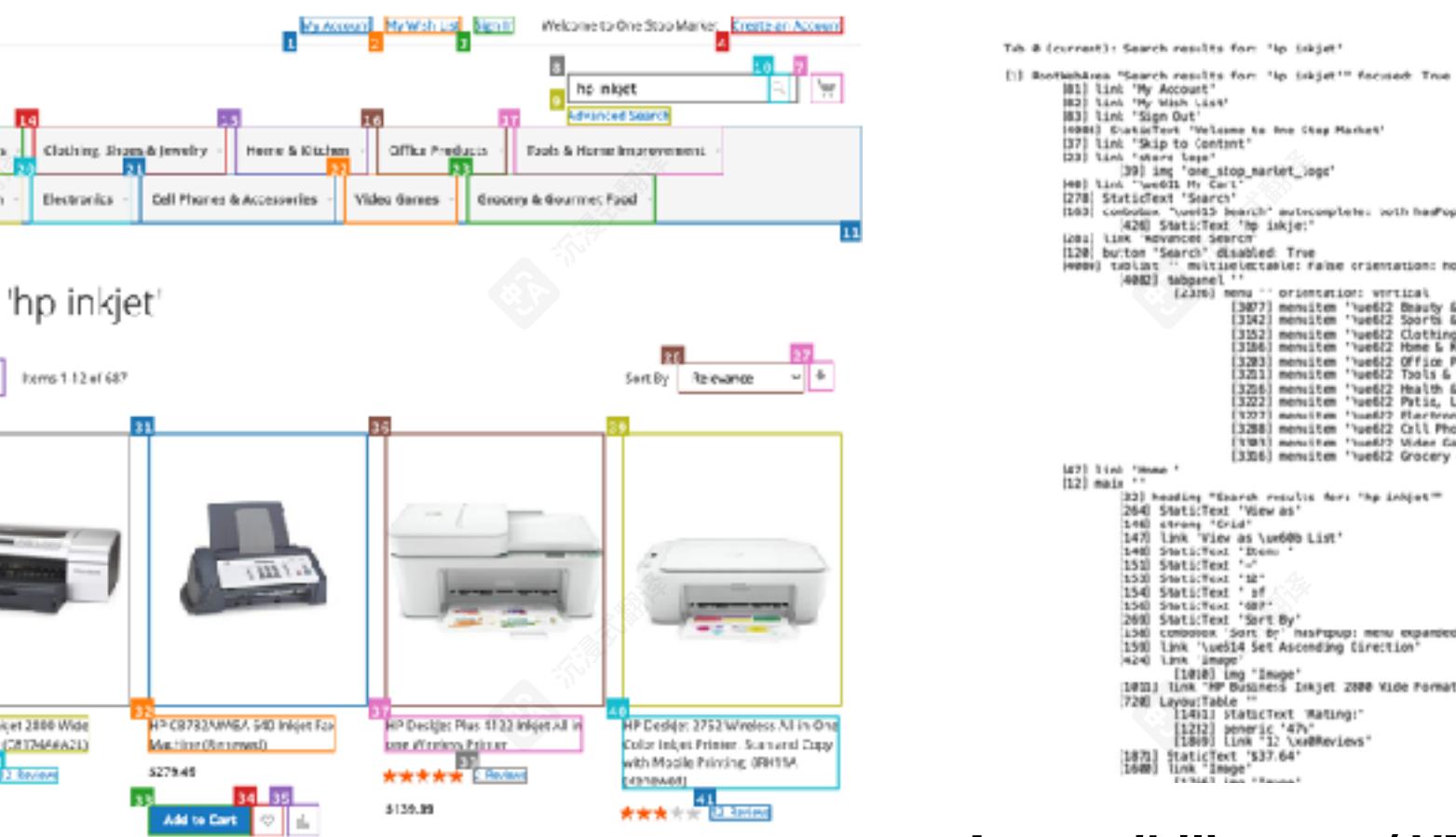
LLM and VLM Agents

LLM 和 VLM 代理

Visual Language Models as Agents

Accessibility tree / HTML representations: Cluttered with unnecessary information, long confusing context.

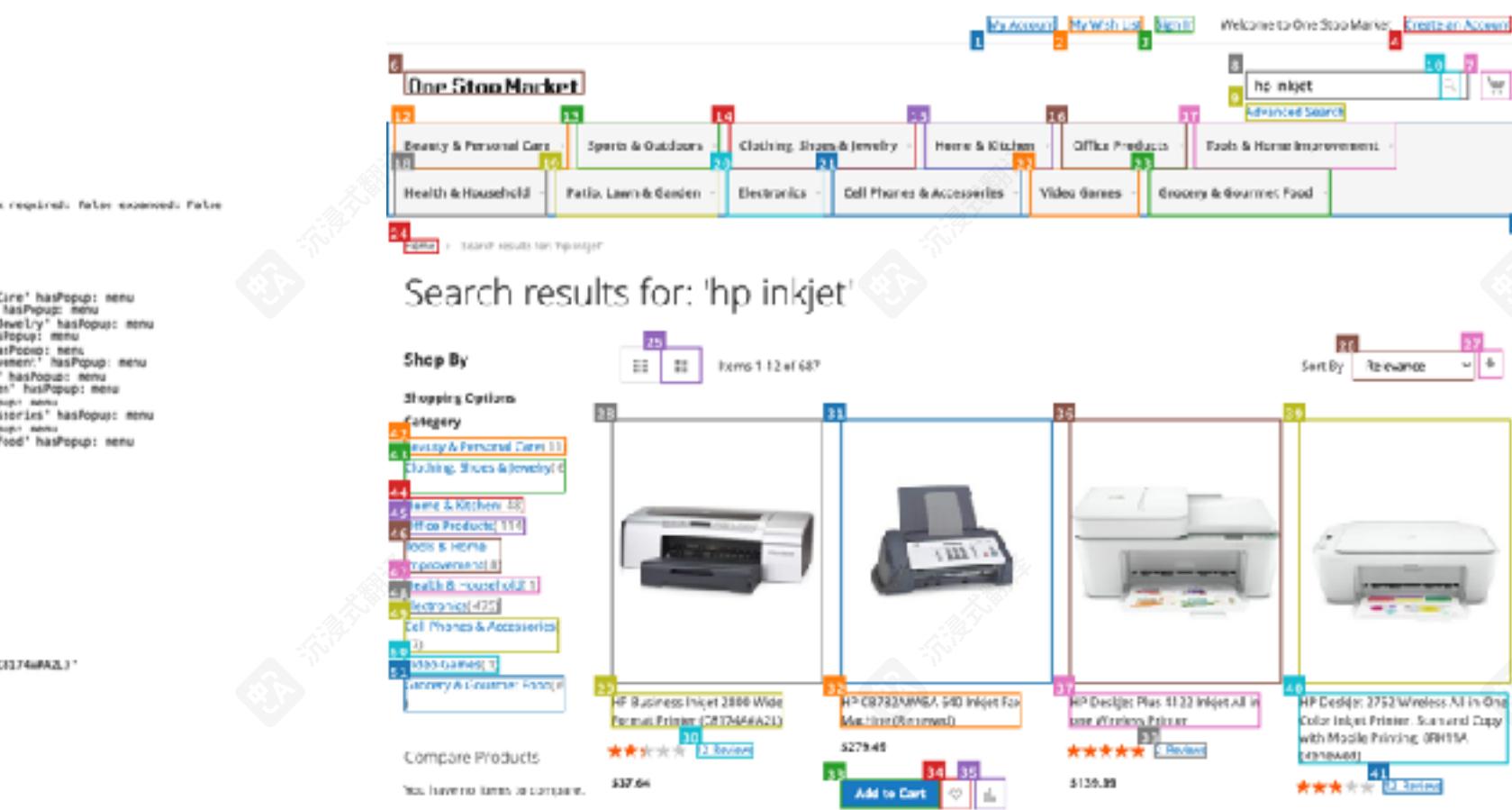
M: Simplified representation with Set-of-
(M) prompting over interactable elements.



卷之三

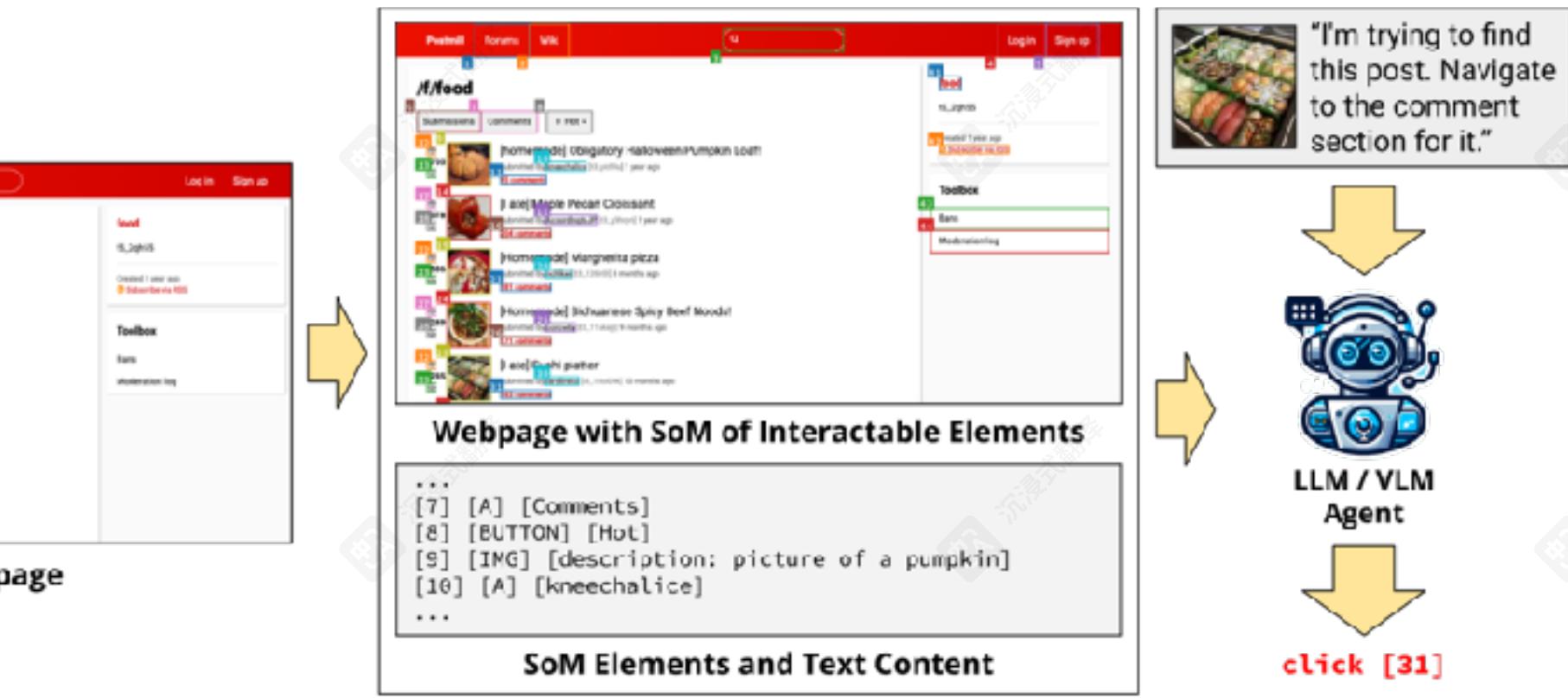
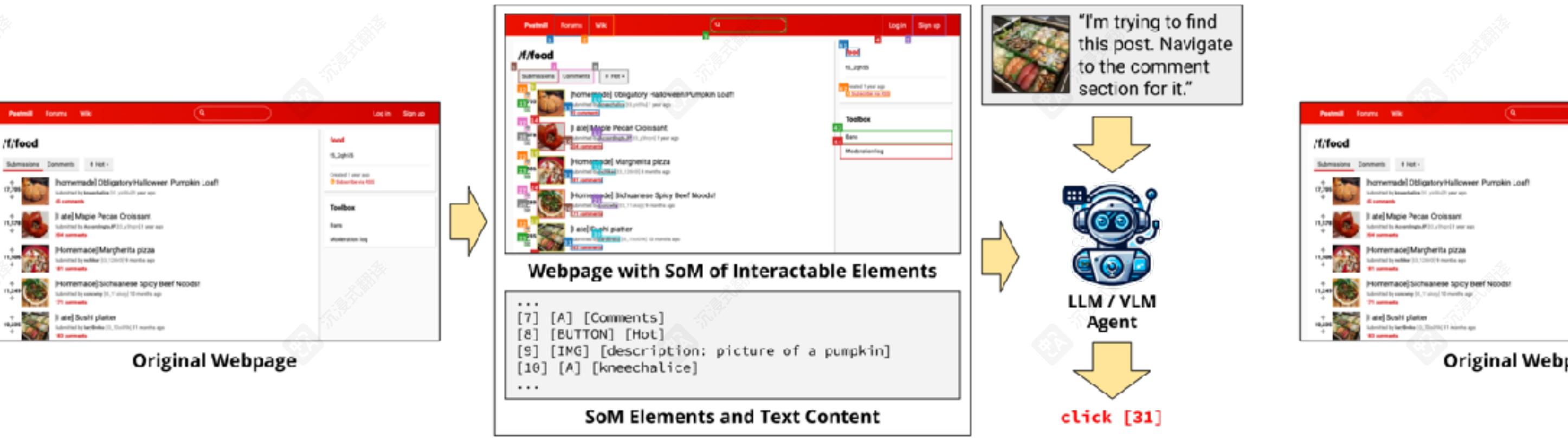
-
d with
long and

VLM + SoM: 简化表示，包含集合 - 标记（SoM）
通过可交互元素进行提示。

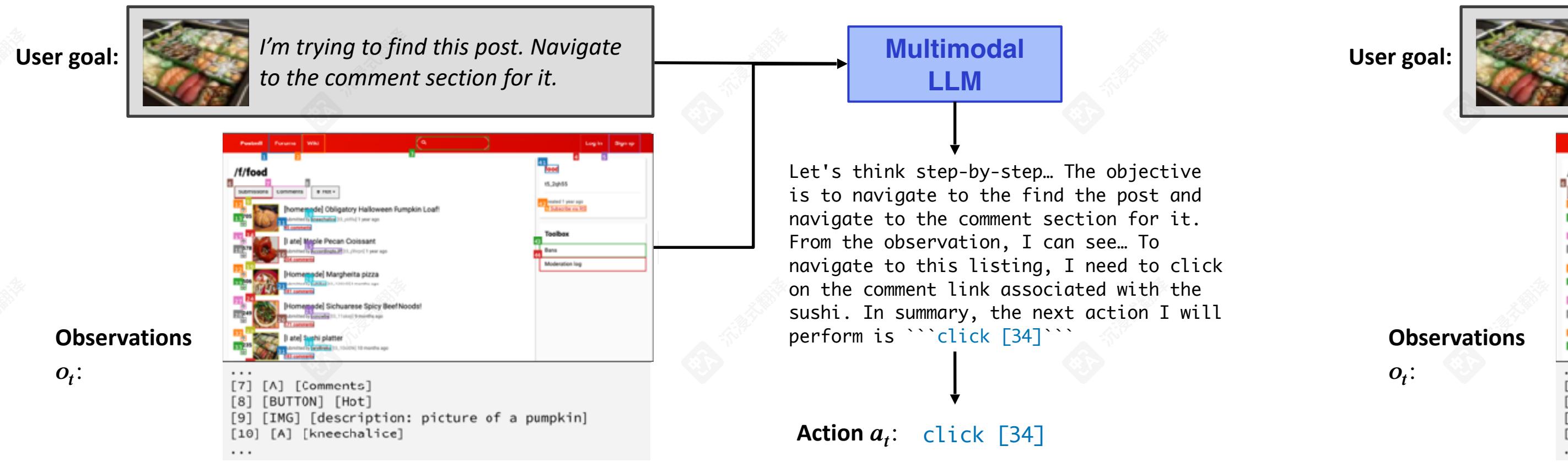


Visual Language Models as Agents

视觉语言模型作为智能体

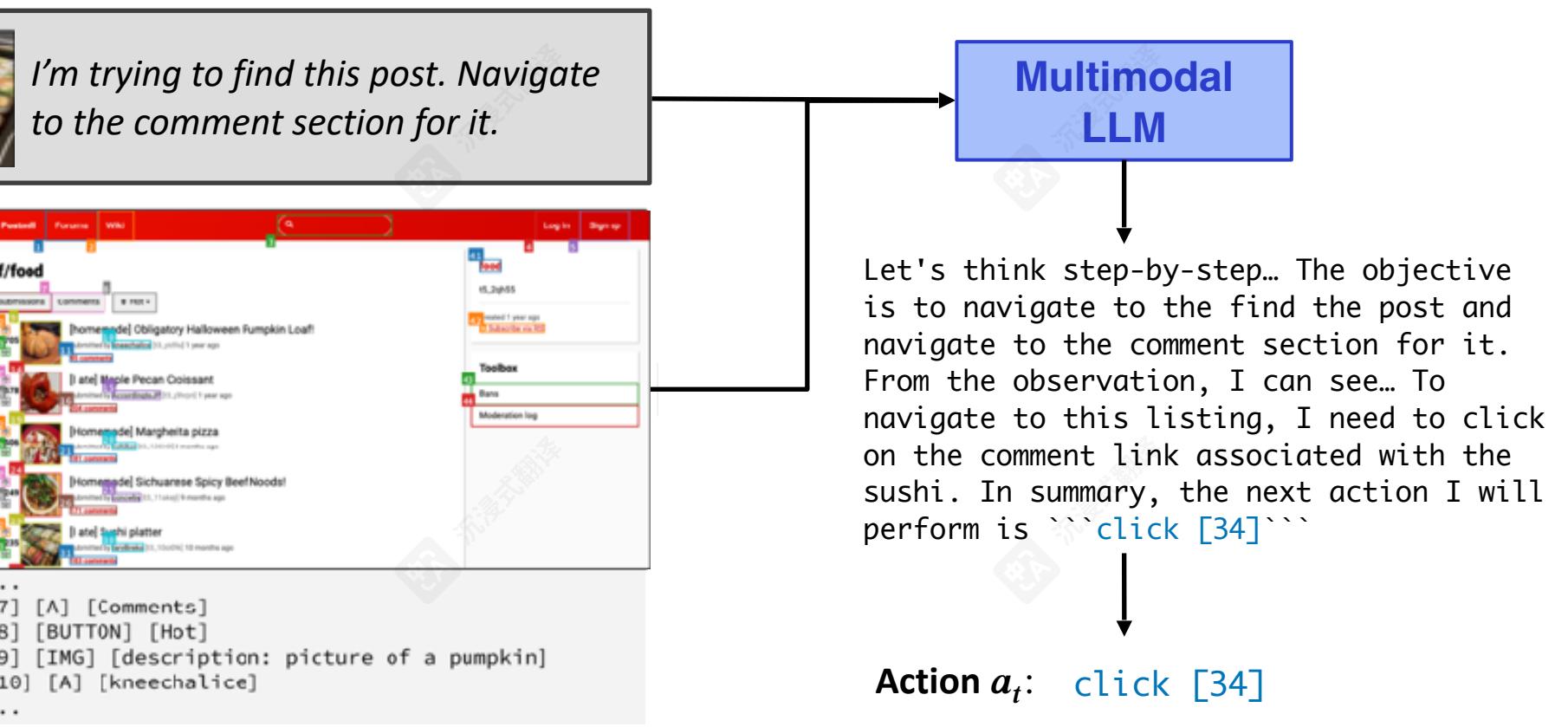


Visual Language Models as Agents



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

视觉语言模型作为智能体

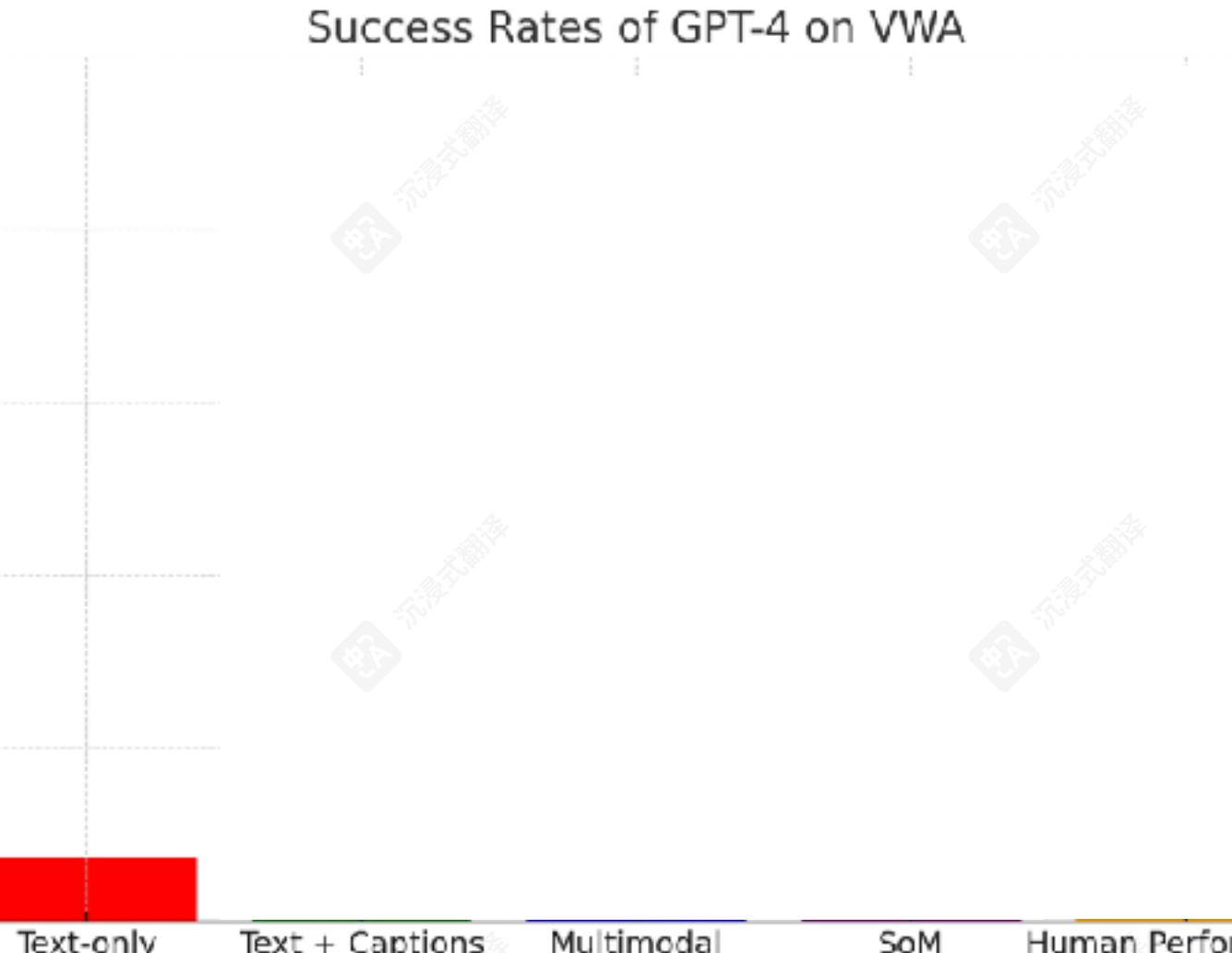


VLM + SoM: 简化表示，使用标记集（SoM）通过可交互元素进行提示。

Baseline Agents



基准代理



Baseline Agents: Text-based LLMs

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

Model Type	LLM Backbone	Visual Backbone	Inputs	Success Rate (\uparrow)
Text-only	LLaMA-2-70B	-	Accessibility Tree	1.10%
	Mixtral-8x7B			1.76%
	Gemin-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			1.87%
	GPT-3.5			2.75%
	GPT-3.5			2.97%
	Gemin-Pro			3.85%
	GPT-4			12.75%

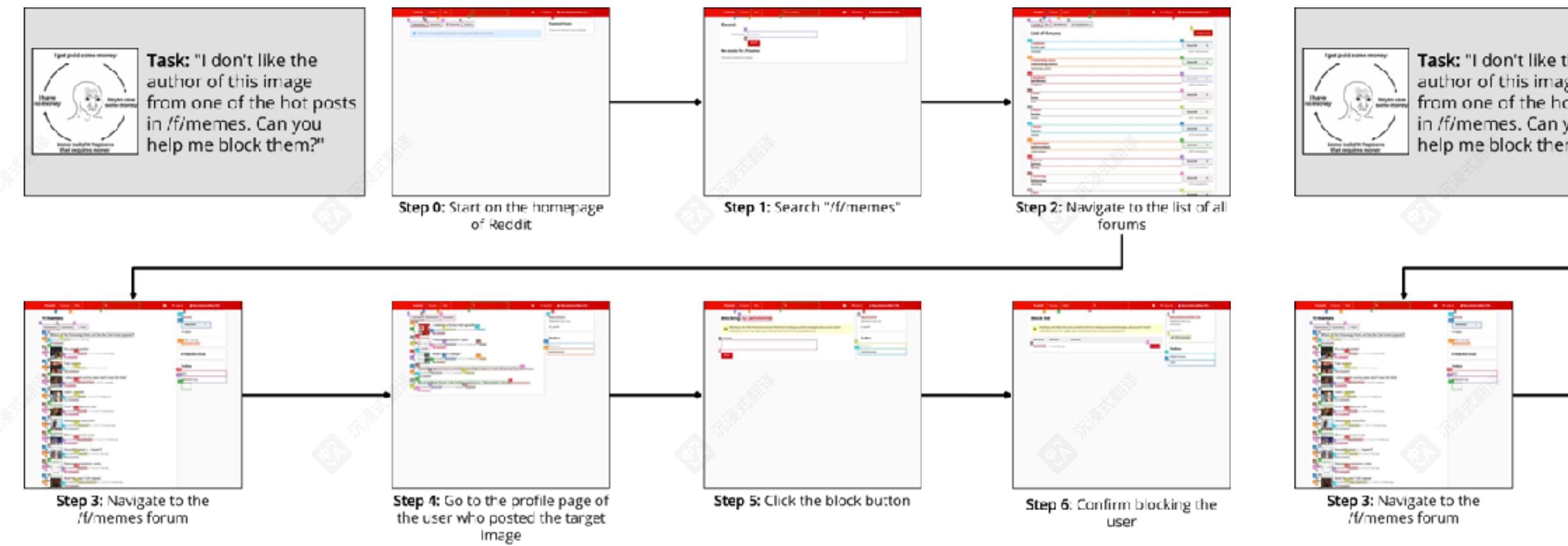
Baseline Agents: 基于文本的 LLMs

Baseline Agents: Multimodal LLMs

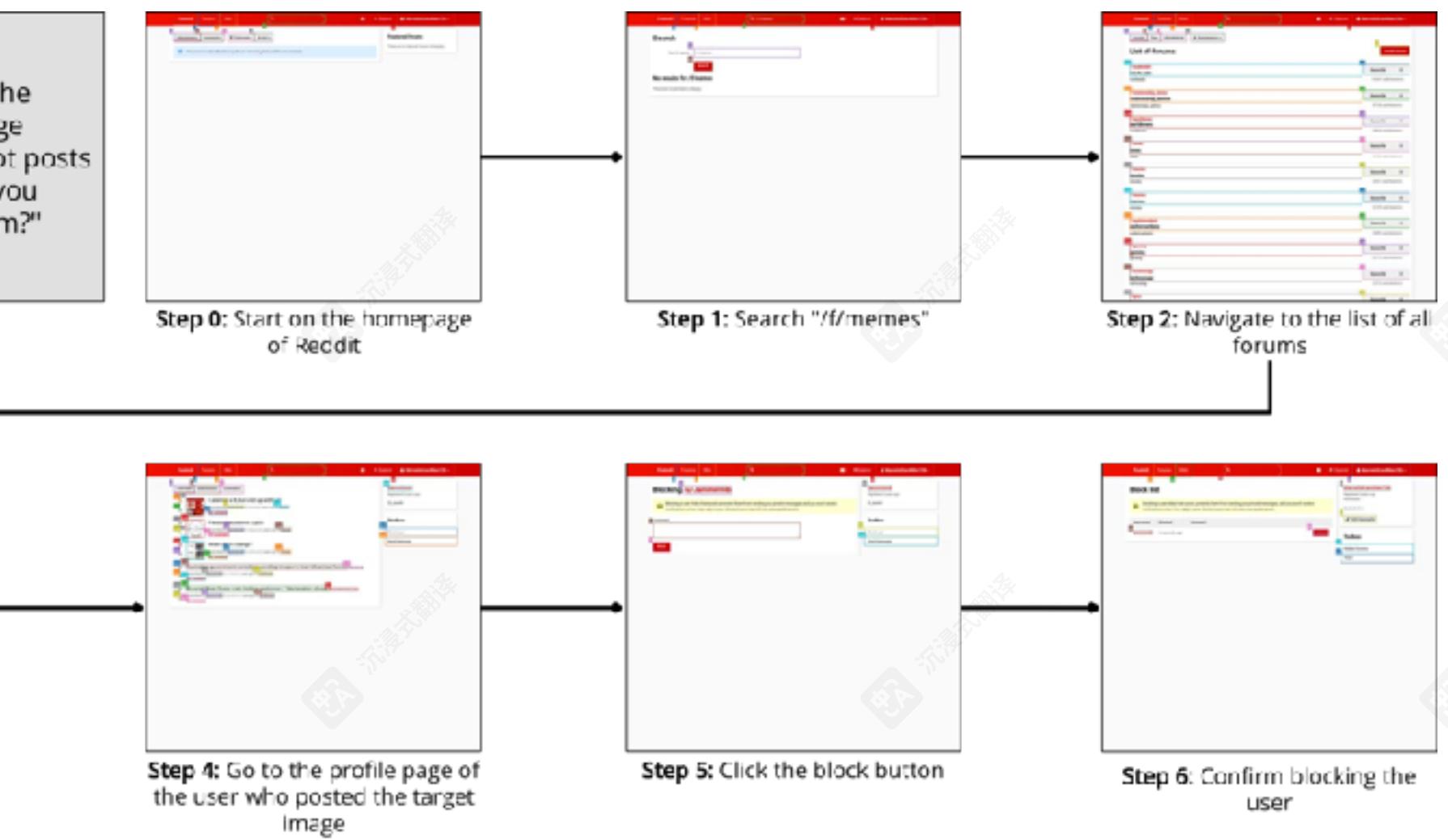
Model Type	Multimodal Model	Inputs	Success Rate (↑)
Multimodal	IDEFICS-80B-Instruct	Image + Captions + Accessibility Tree	0.77%
	CogVLM		0.33%
	Gemini-Pro		6.04%
	GPT-4V		15.05%
Multimodal (SoM)	IDEFICS-80B-Instruct	Image + Captions + SoM	0.99%
	CogVLM		0.33%
	Gemini-Pro		5.71%
	GPT-4V		16.37%
Human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

基准代理：多模态 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%



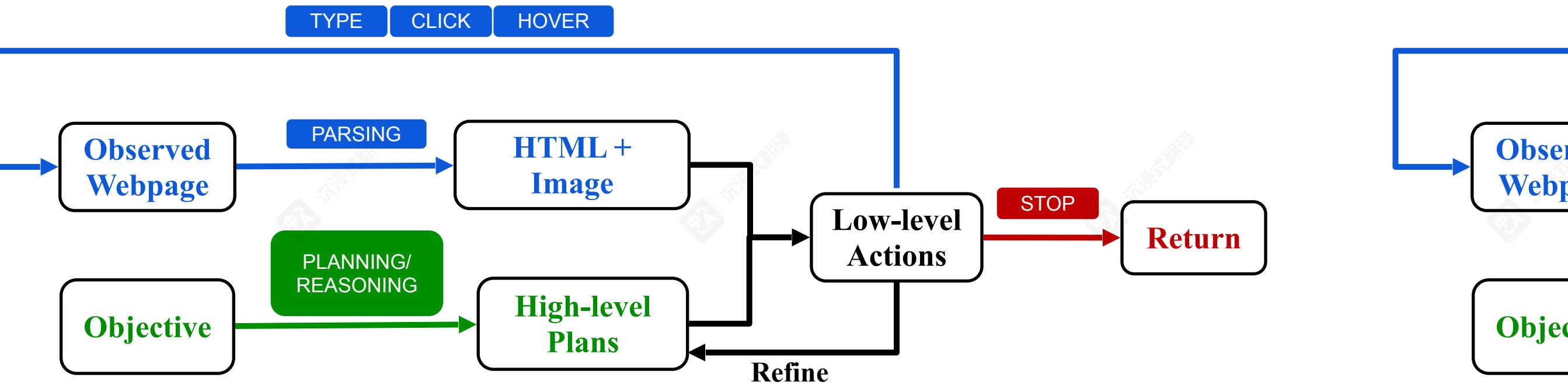
Successful execution trajectory of the GPT-4V + SoM agent on the task for blocking a user that posted a certain picture



GPT-4V + SoM 代理在阻止发布特定图片的用户任务上的成功执行轨迹

Web Agent Architecture

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

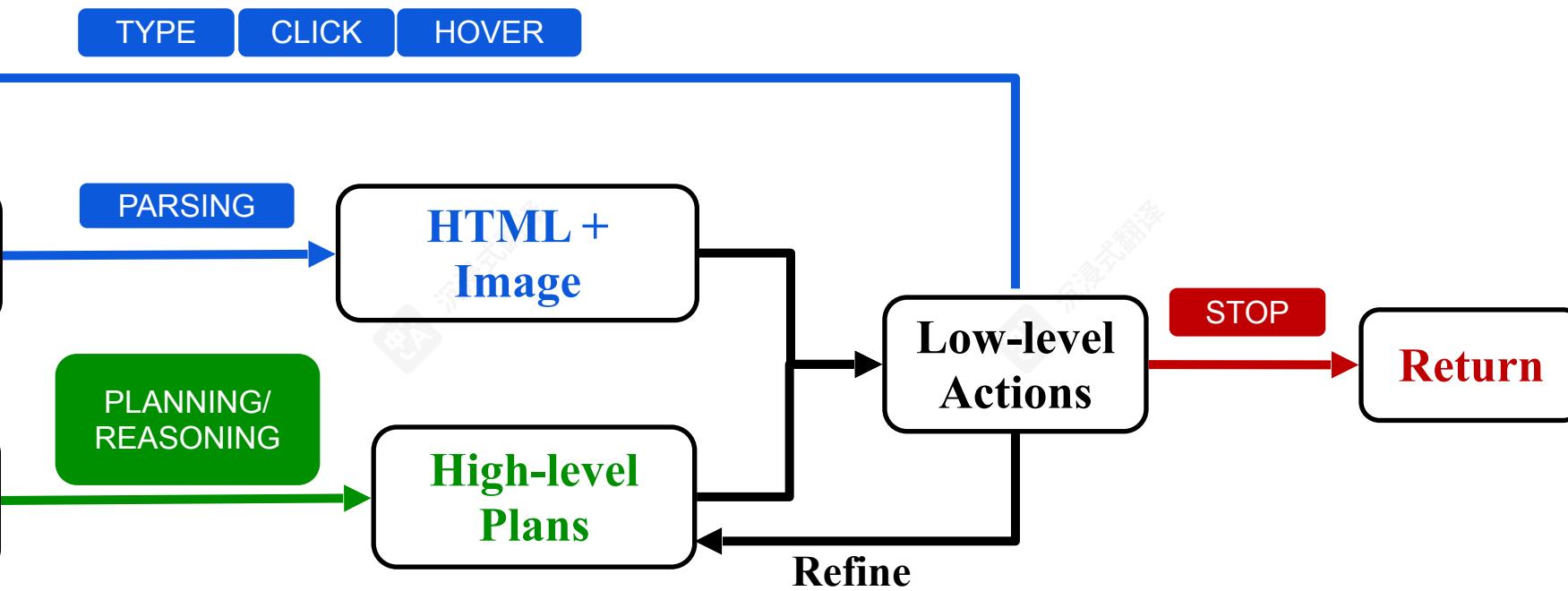


Web Agent Architecture

- 我们交互式代理的模型架构:

- 高级规划与推理 - 观察解析 - 低级动

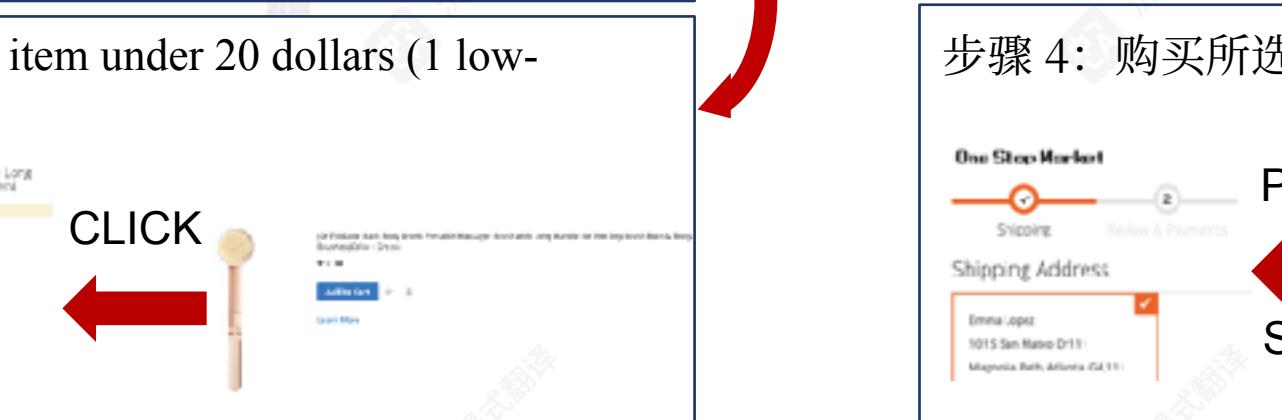
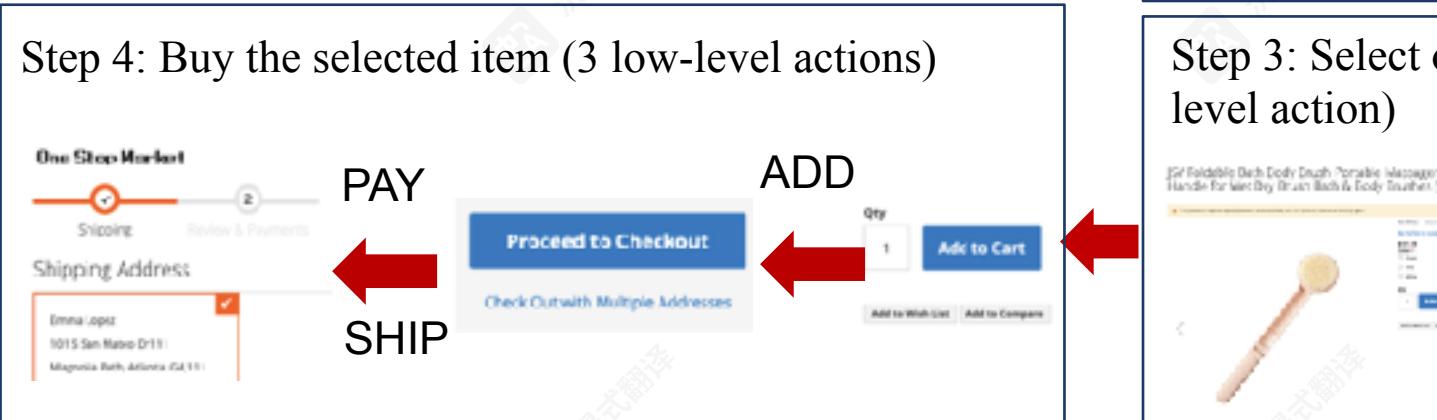
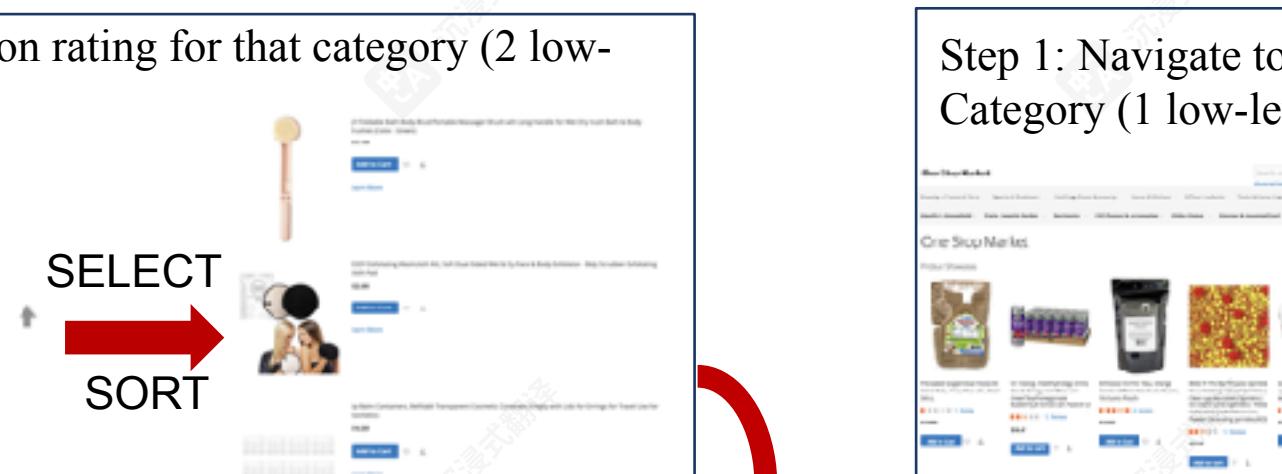
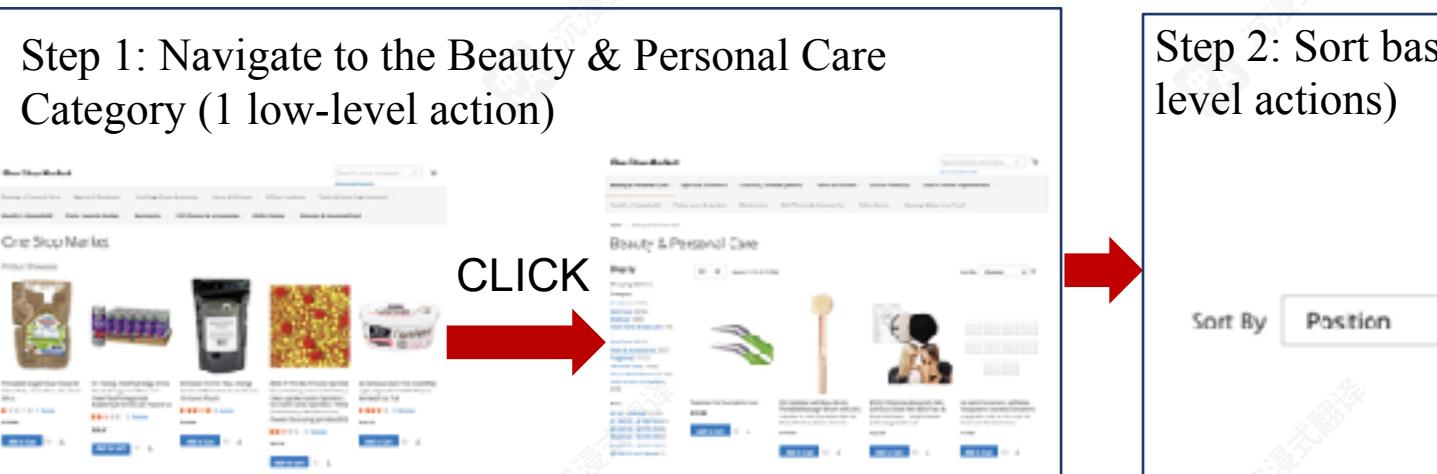
- 作生成



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.



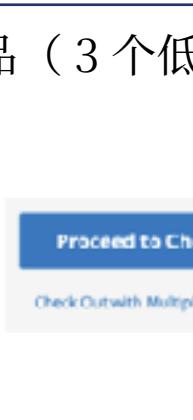
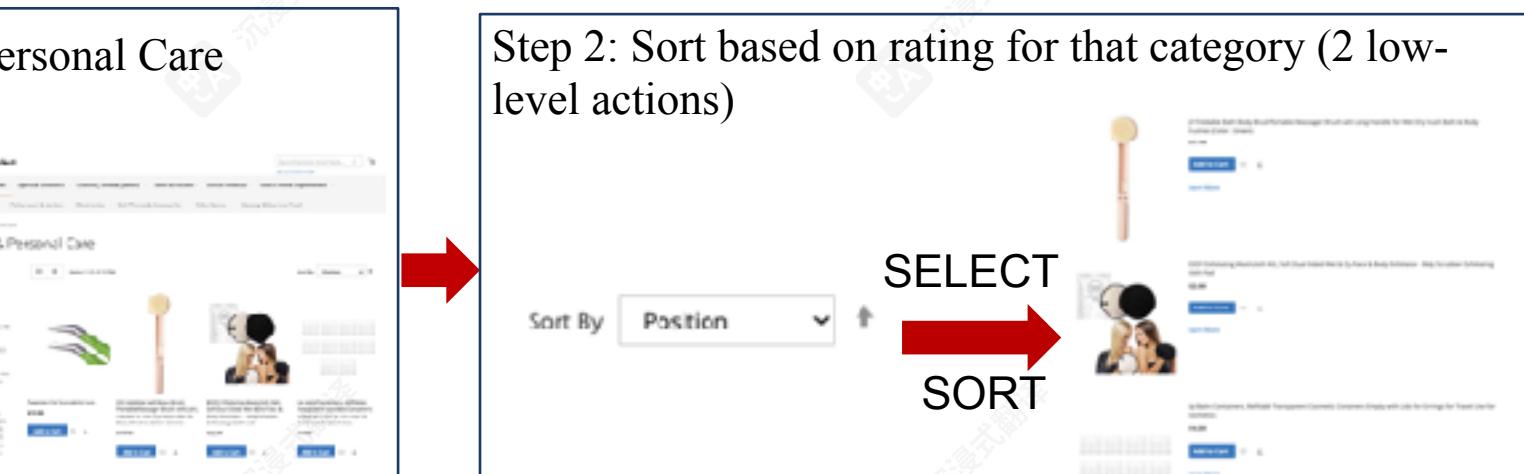
规划

高级计划

对于长序列和复杂目标非常

重要。

任务：在预算低于 20 美元的情况下，从美容及个人护理类别中选择评分最高的产品。

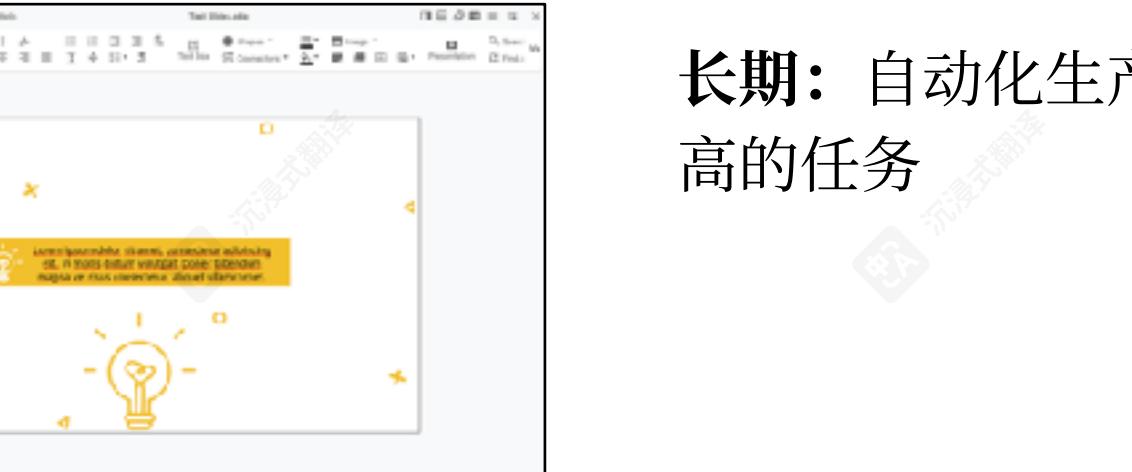


Measuring Productive Tasks

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.

衡量生产性任务

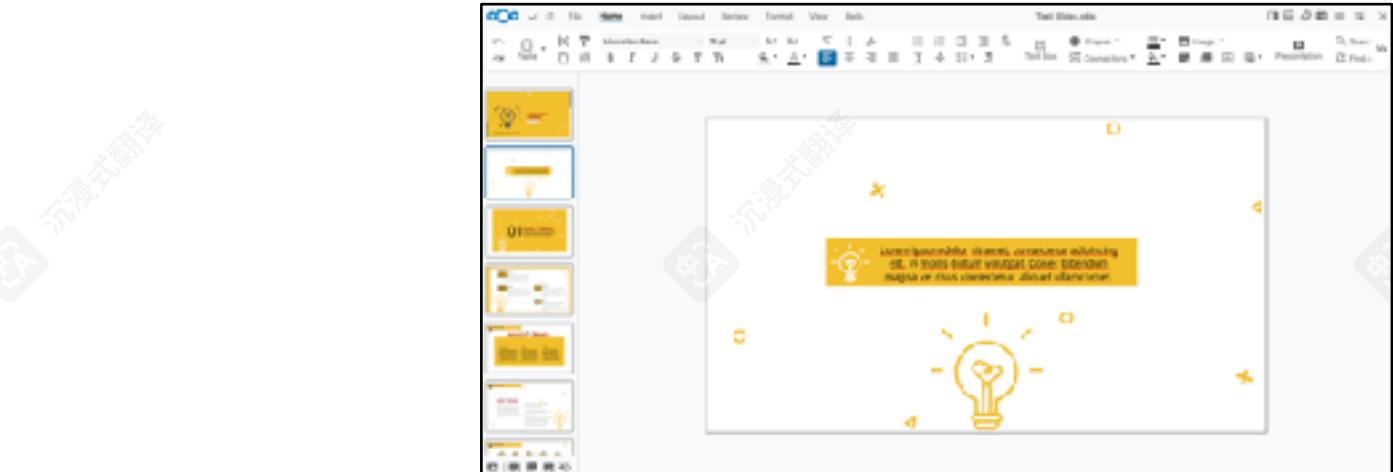
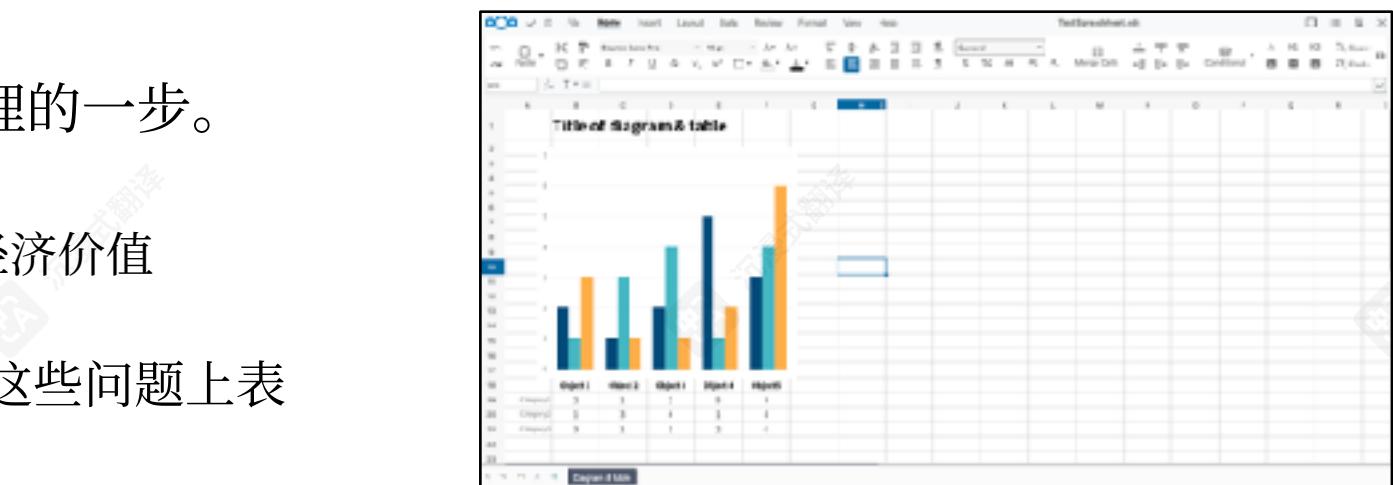
VisualWebArena 是构建通用代理的第一步。

但：

- 任务不是很 有影响力的：不代表显著的经济价值

- 任务更简单，因为当前的 LLM 代理在这些问题上表现不佳

长期：自动化生产性、经济价值高的任务



来自 [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

常见故障模式

- 长时程推理和规划:
 - 模型在两个网页之间振荡，或陷入死循环
 - 正确执行任务但随后撤销
 - 智能体倾向于过早停止探索 / 执行
- Failures in visual processing
 - Clicking the wrong item
 - Identifying specific items in complex webpages
 - Spatial reasoning (“what are the prices of products in the first row?”)

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.

什么是缺失的？

- 我们需要做更多来缩小差距：
 - **推理和规划** 跨越长期目标
 - 允许代理**搜索**，并行执行和协调多个实例并请求澄清 / 确认 - 强大的视觉 - 语言 - 代码模型 - 确定代理的适当抽象级别（HTML/ 屏幕截图 /API）
- **多模态模型：**许多现实世界的任务需要视觉基础才能有效解决（例如，涉及 PowerPoint、Excel、Photoshop 的每个任务）。为了开发强大的通用代理，我们需要训练和构建强大的视觉 - 语言模型。

Talk Outline

- VisualWebArena -- Evaluating Multimodal Agents on Realistic Visual Web Tasks (Koh et al., ACL 2024)
- Tree Search for Language Model Agents (Koh, McAleer, Fried, Salakhutdinov, arXiv 2024)
- Towards Internet-Scale Training For Agents (Trabucco, Sigurdsson, Piramuthu, Salakhutdinov, arXiv 2025)

Talk Outline

- VisualWebArena -- 在真实视觉网络任务上评估多模态智能体 (Koh 等人 , ACL 2024)
- 语言模型代理的树搜索 (Koh, McAleer, Fried, Salakhutdinov, arXiv 2024)
- 迈向互联网规模智能体训练 (Trabucco, Sigurdsson, Piramuthu, Salakhutdinov, arXiv 2025)

Exponential Error Compounding in Agents

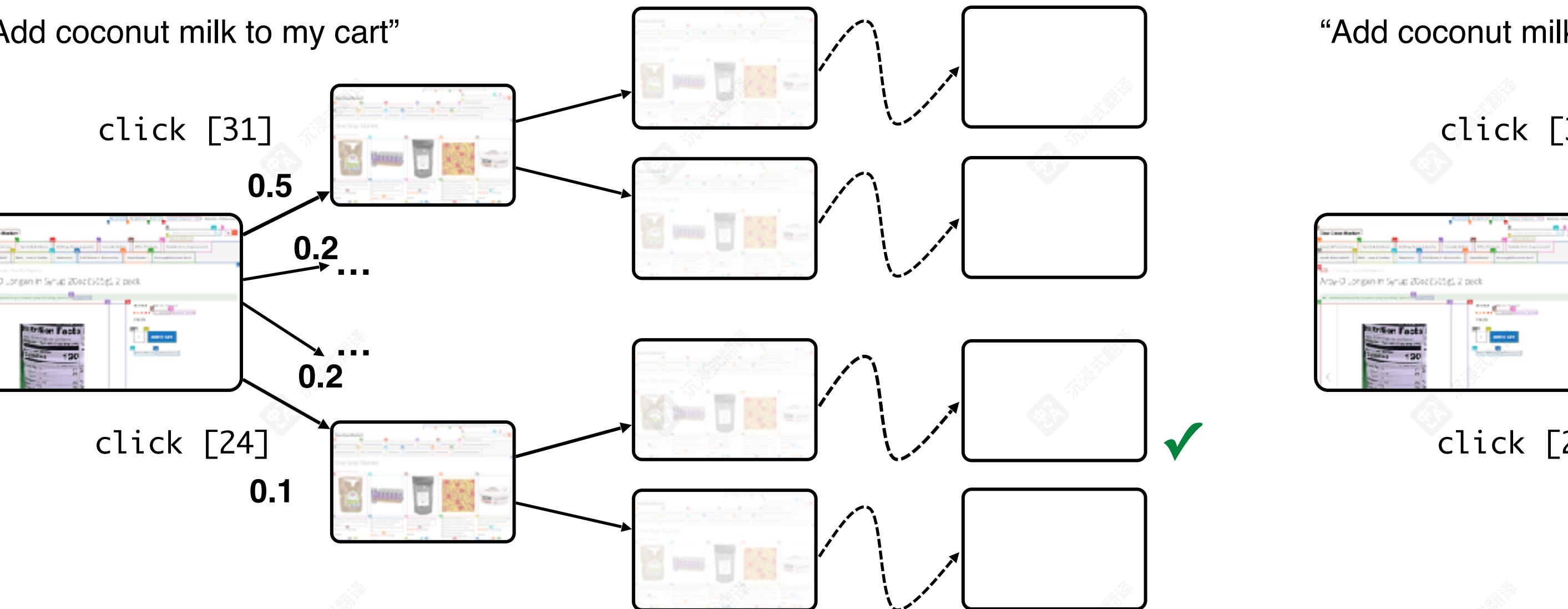
Accuracy @ k steps:					
1 (single step)	5	10	30	50	
90%	59.05%	34.87%	4.24%	0.52%	
95%	77.38%	59.87%	21.46%	7.69%	
99%	95.10%	90.44%	73.97%	60.50%	
99.9%	99.50%	99.00%	97.04%	95.12%	
99.99%	99.95%	99.90%	99.70%	99.50%	

智能体中的指数误差累积

Accuracy @ k steps:					
1 (单步)	5	10	30	50	
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%	

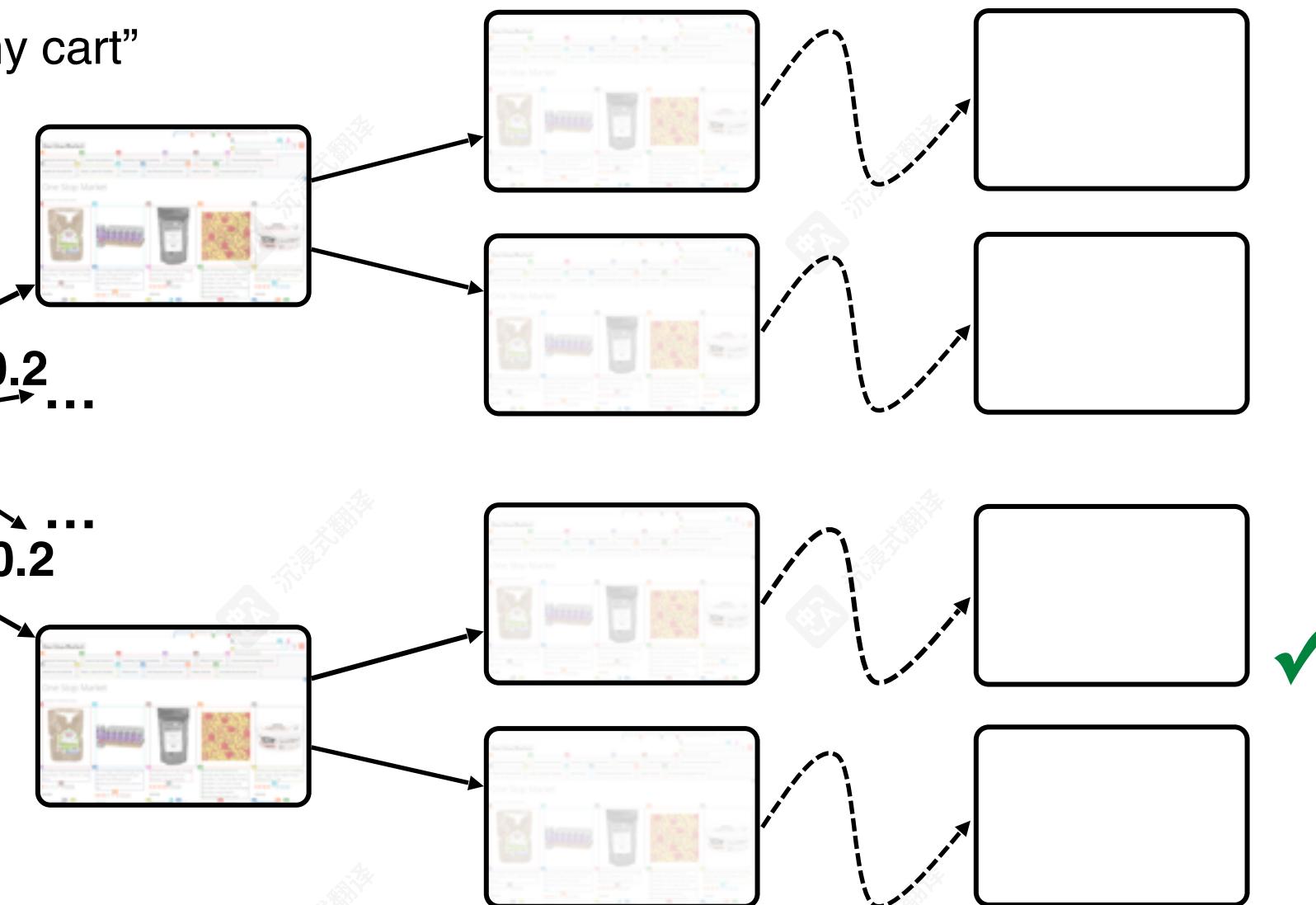
Local Decisions; Global Consequences

“Add coconut milk to my cart”



Local Decisions; Global Consequences

“Add coconut milk to my cart”



Local Decisions; Global Consequences

“Add coconut milk to my cart”

click [31]

0.5

0.2
...

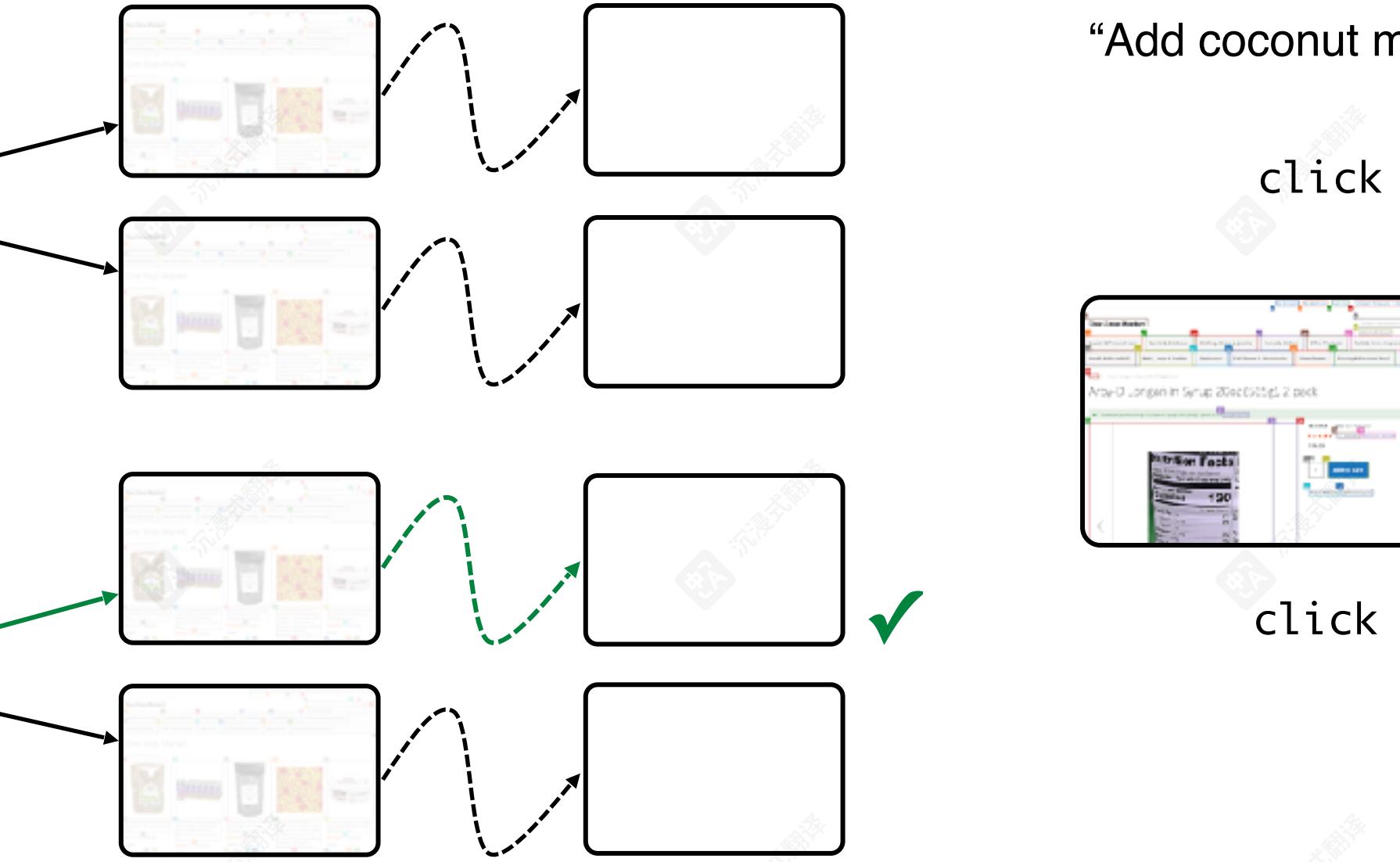
0.2
...

click [24]

0.1

0.2
...

0.2
...



Local Decisions; Global Consequences

“Add coconut milk to my cart”

click [31]

0.5

0.2
...

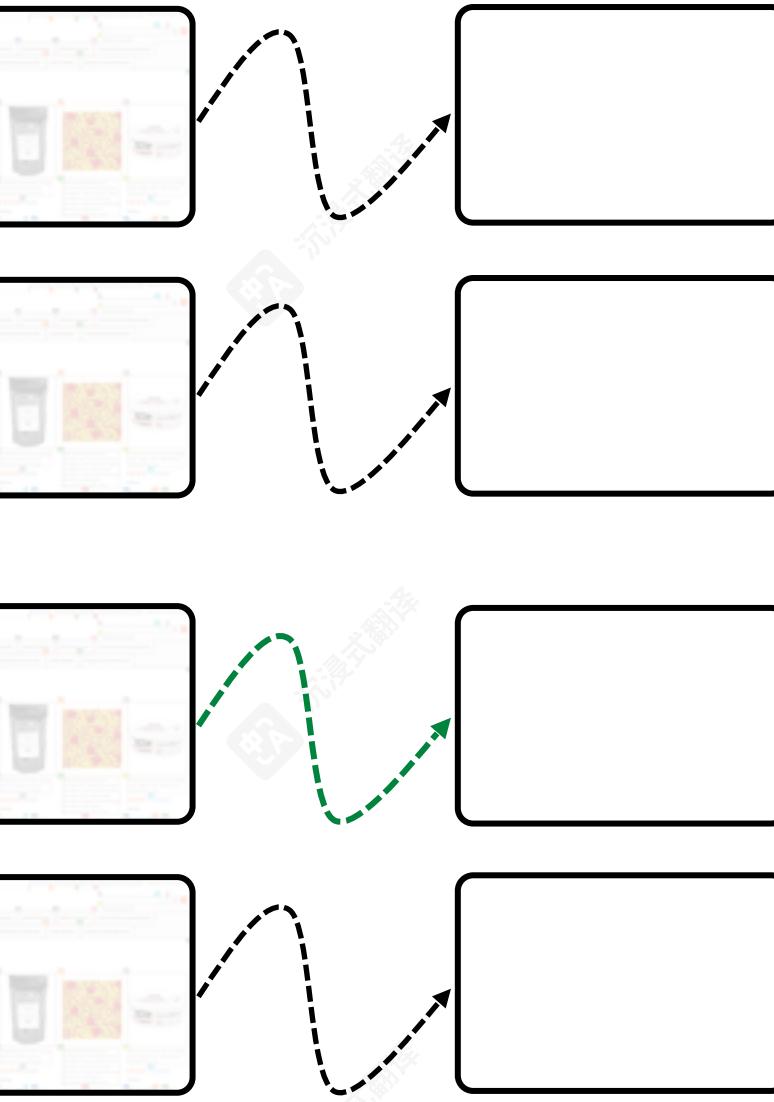
0.2
...

click [24]

0.1

0.2
...

0.2
...



Search By Repeated Sampling

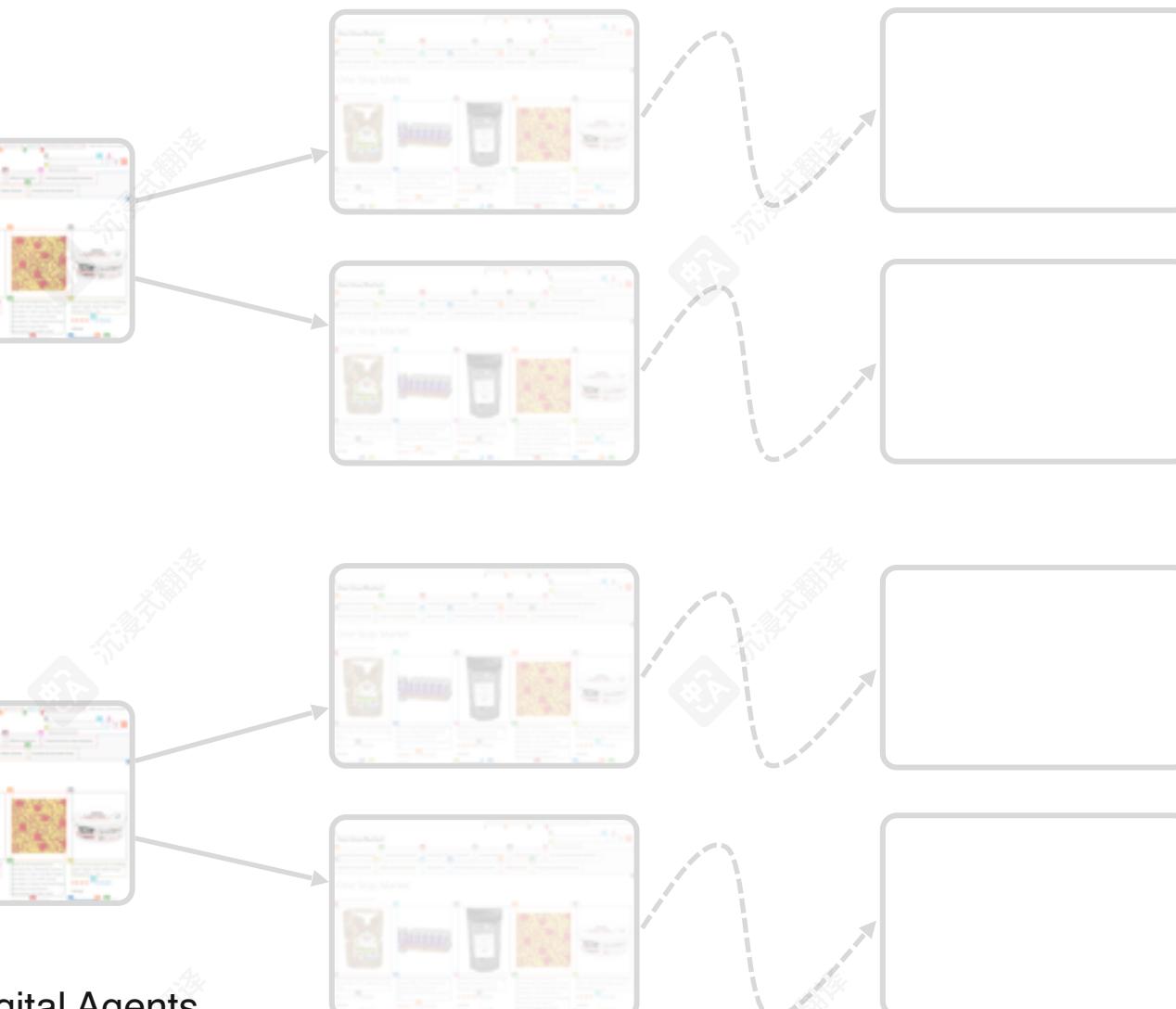
1. Sample actions from the language model until [STOP].



Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

通过重复采样搜索

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



Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

Search By Repeated Sampling

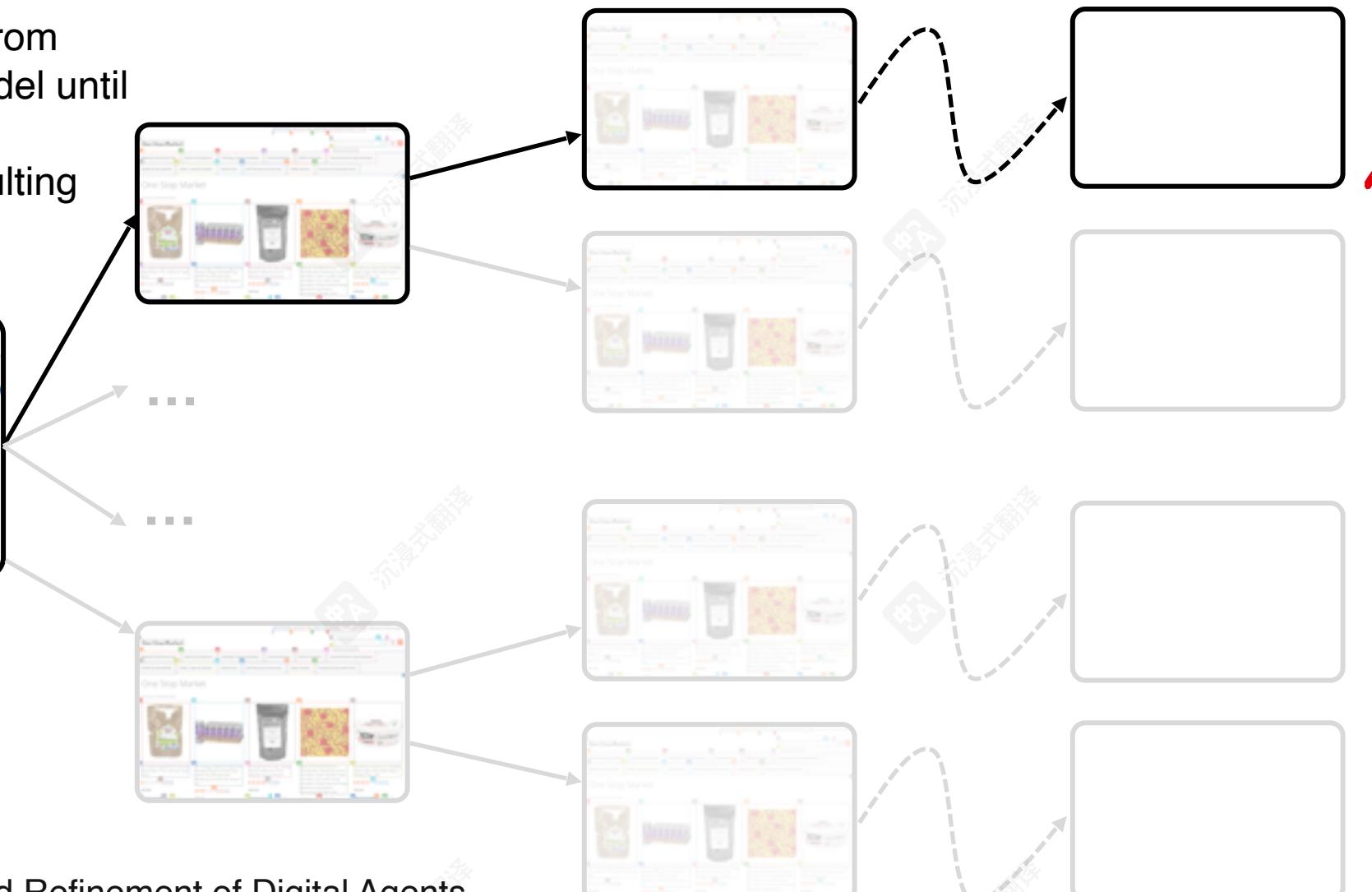
1. Sample actions from the language model until [STOP].
2. Evaluate the resulting trajectory
3. Repeat?



Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

通过重复采样搜索

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



Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

Search By Repeated Sampling

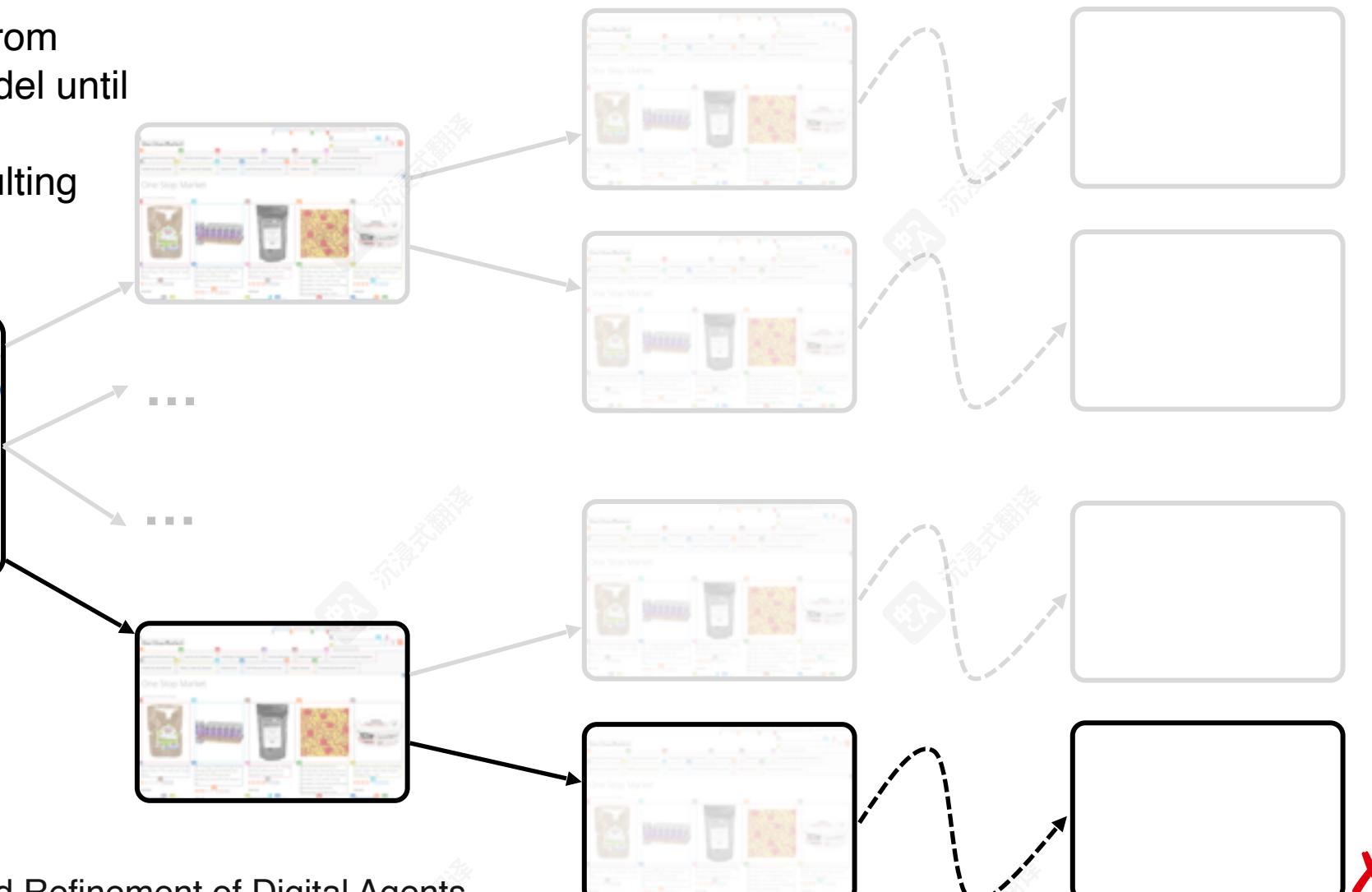
1. Sample actions from the language model until [STOP].
2. Evaluate the resulting trajectory
3. Repeat?



Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

通过重复采样搜索

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



Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

Search By Repeated Sampling

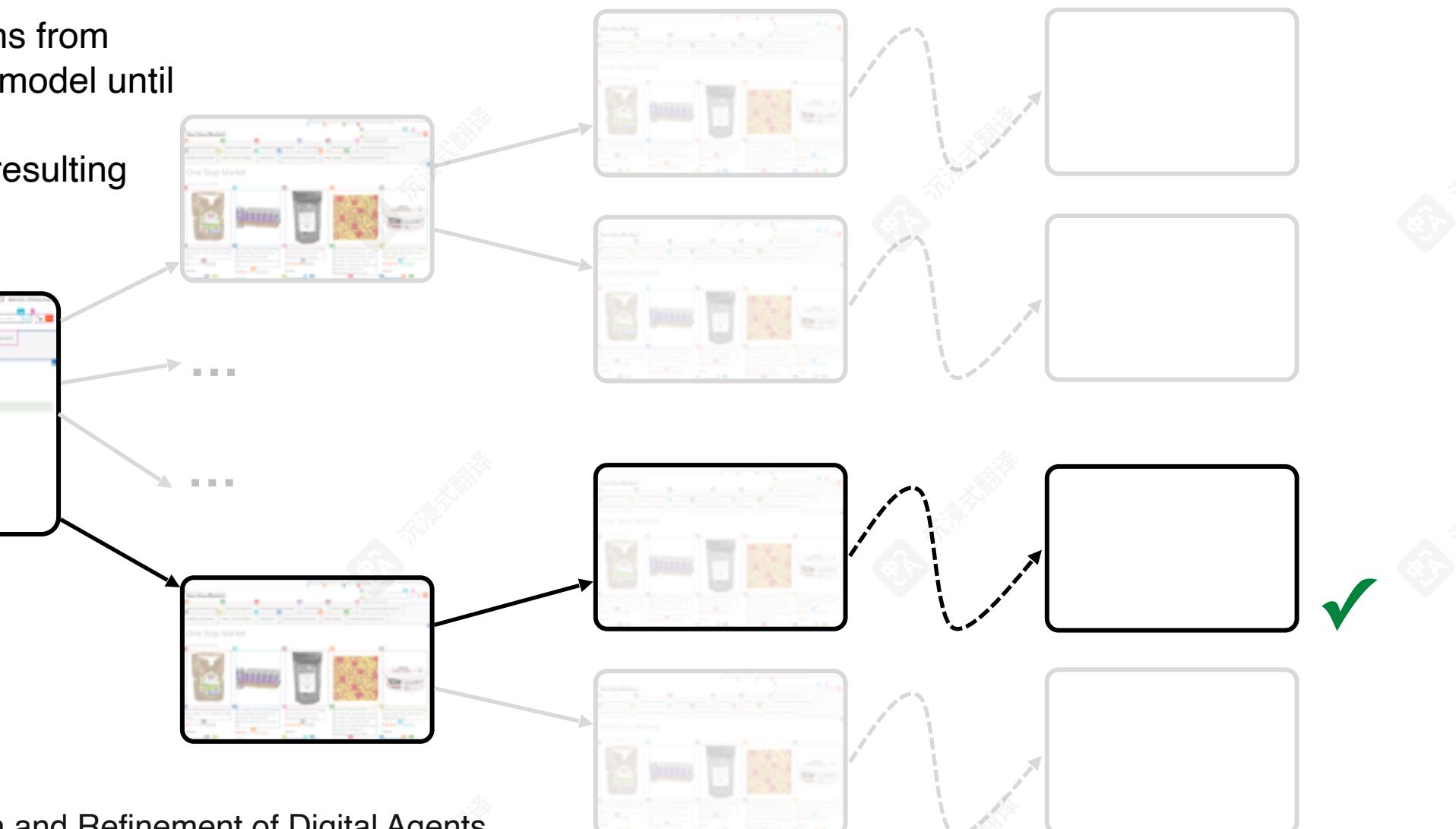
1. Sample actions from the language model until [STOP].
2. Evaluate the resulting trajectory
3. Repeat?



Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

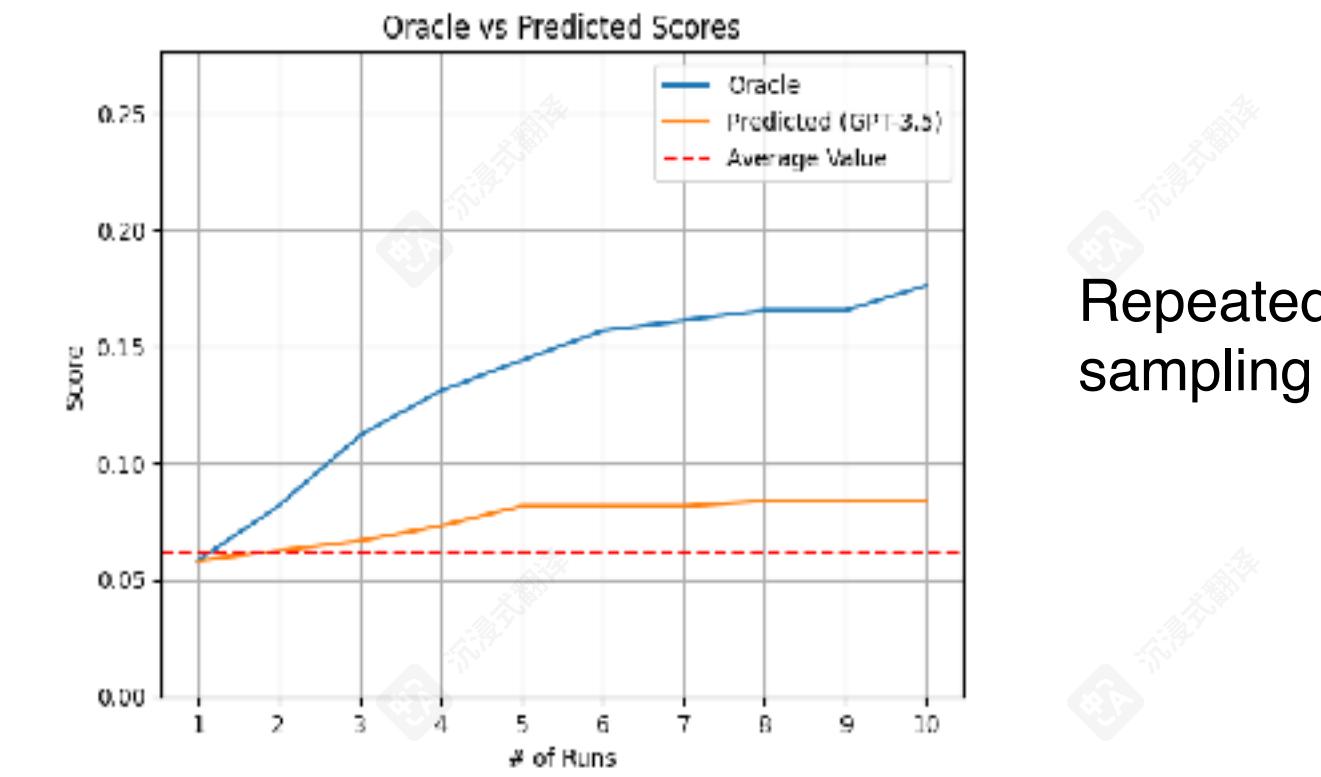
通过重复采样搜索

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



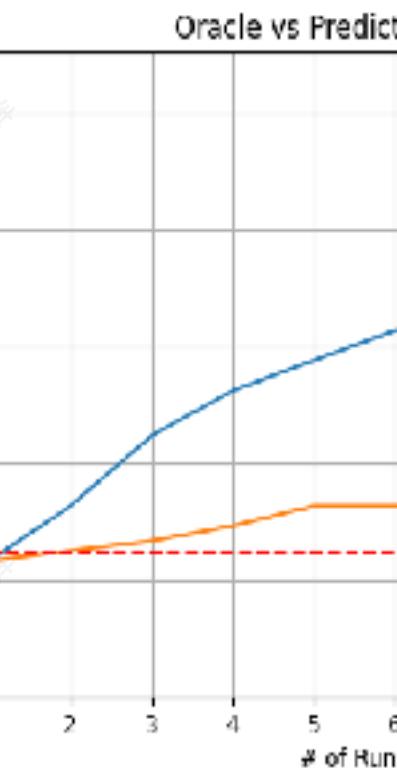
Pan et al. 2024,
Autonomous Evaluation and Refinement of Digital Agents

Search By Repeated Sampling



Repeated sampling helps!

搜索重复采样



重复采样有帮助!

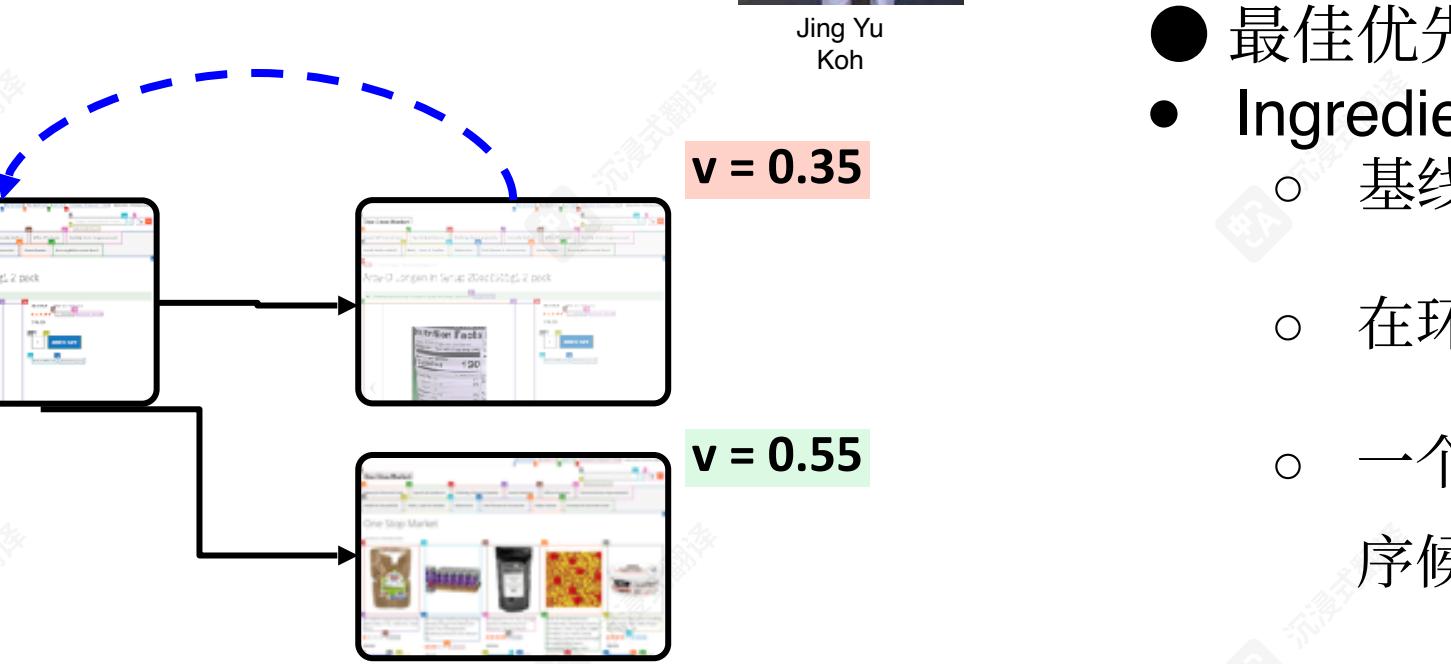
- But the space is exponentially large. Can we guide exploration?
- Key idea: apply value function to intermediate nodes.

- But the 空间是指数级的。我们能引导探索 ?
- 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.



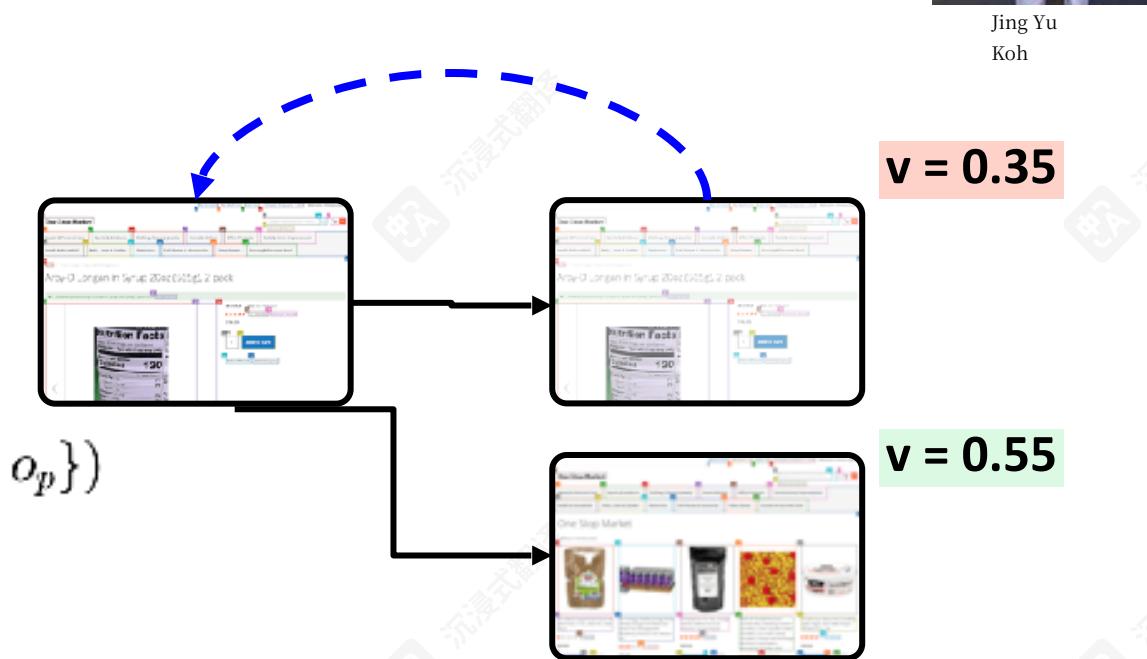
Jing Yu Koh



Jing Yu Koh

我们的方法：树搜索

- 最佳优先搜索算法
- Ingredients:
 - 基线智能体来提出动作。
 - 在环境中回溯的方法。
 - 一个价值函数来评分和重新排序候选状态。



Jing Yu Koh



Jing Yu Koh

在这项工作中，我们提示一个多模态 LLM (GPT-4o) 来充当评估者。



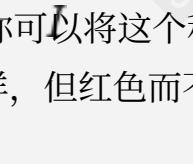
Task Instruction (): “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

Legend

1 Step sequence

$v=1.0$ State values

→ Backtracking



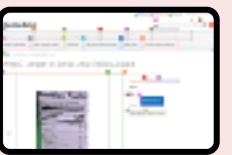
任务指令 (): “你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到比较页面吗？”

Legend

1 步骤序列

$v=1.0$ $v=0.4$ State values

→ Backtracking

GPT-4o Agent**GPT-4o Agent + Search**

Starting State

GPT-4o Agent**GPT-4o Agent + Search**

起始状态



Task Instruction (): “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
→ Backtracking

v = 1.0 State values

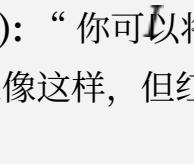
GPT-4o Agent



GPT-4o Agent + Search



Starting State



任务指令 ():

“你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到对比页面吗？”

Legend

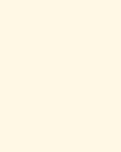
1 步骤顺序
→ Backtracking

v = 1.0 State values

GPT-4o Agent



GPT-4o Agent + 搜索



Starting State

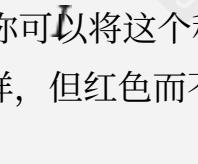


Task Instruction (): “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

Legend

1 Step sequence
→ Backtracking

v = 1.0 State values



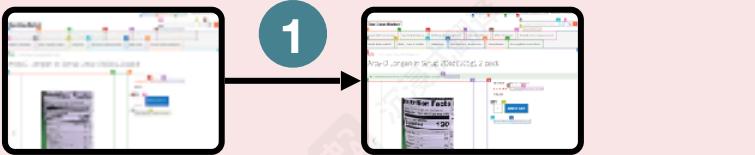
任务指令 (): “你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到对比页面吗？”

Legend

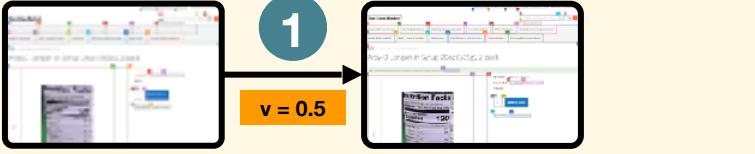
1 步骤顺序
→ Backtracking

v = 1.0 State values

GPT-4o Agent



GPT-4o Agent + Search



Starting State

GPT-4o Agent



GPT-4o Agent + 搜索



Starting State



Task Instruction (): “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

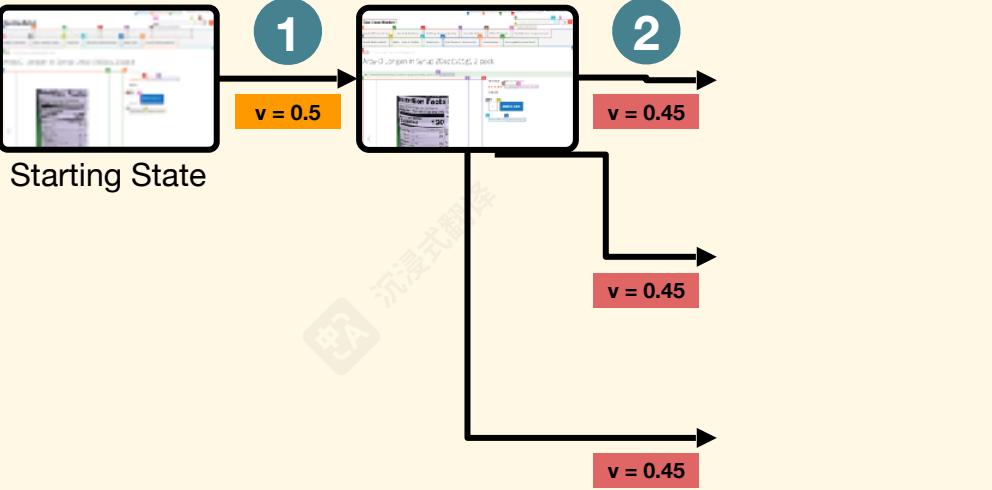
- 1 Step sequence
- Backtracking

v = 1.0 State values

GPT-4o Agent



GPT-4o Agent + Search



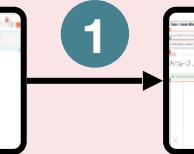
任务指令 (): “你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到对比页面吗？”

Legend

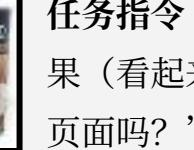
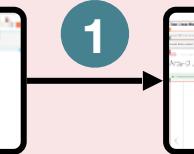
- 1 步骤顺序
- Backtracking

v = 1.0 State values

GPT-4o Agent



GPT-4o Agent + 搜索



任务指令 (): “你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到对比页面吗？”

Legend

- 1 步骤顺序
- Backtracking

v = 1.0 State values



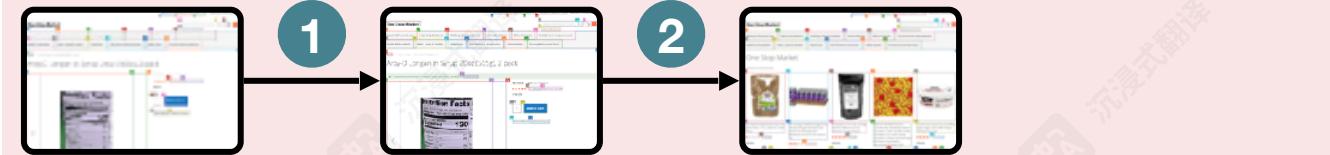
Task Instruction (): “Can you add this and the other canned fruit (of the same brand) that looks like this, but red instead of brown to the comparison page?”

Legend

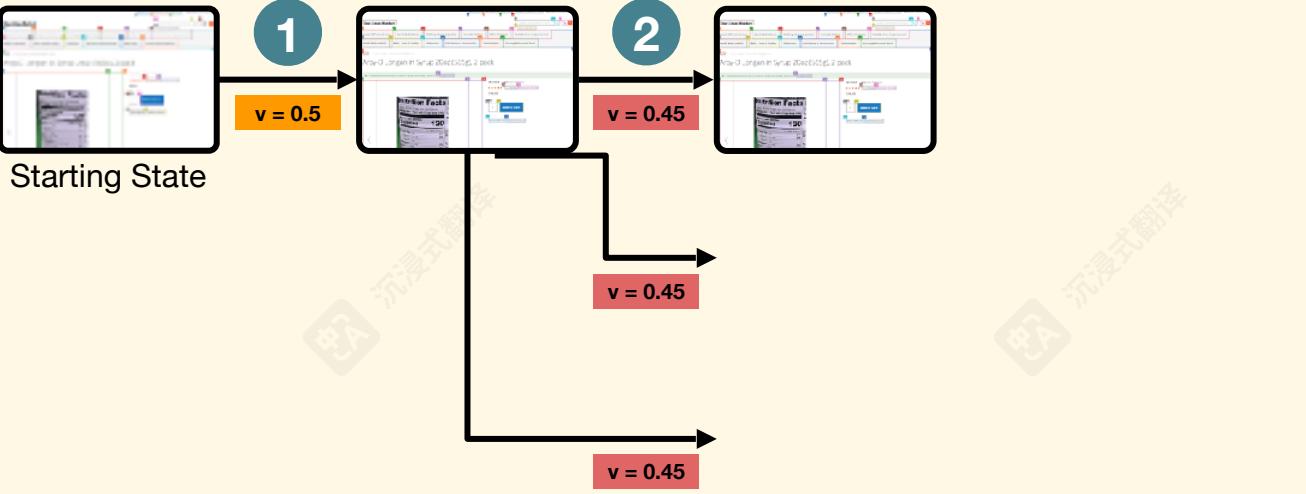
- 1 Step sequence
→ Backtracking

v = 1.0 State values

GPT-4o Agent



GPT-4o Agent + Search



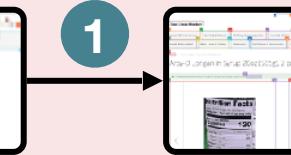
任务指令 (): “你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到对比页面吗？”

Legend

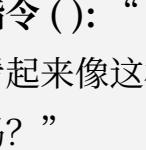
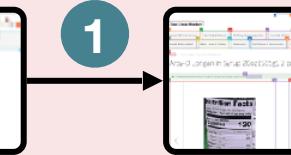
- 1 步骤顺序
→ Backtracking

v = 1.0 State values

GPT-4o Agent

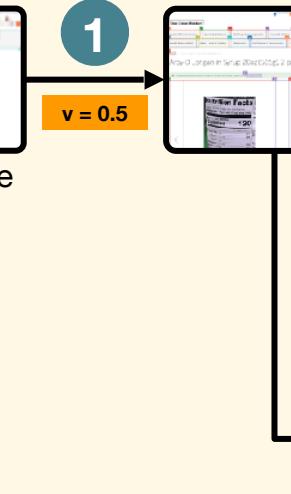


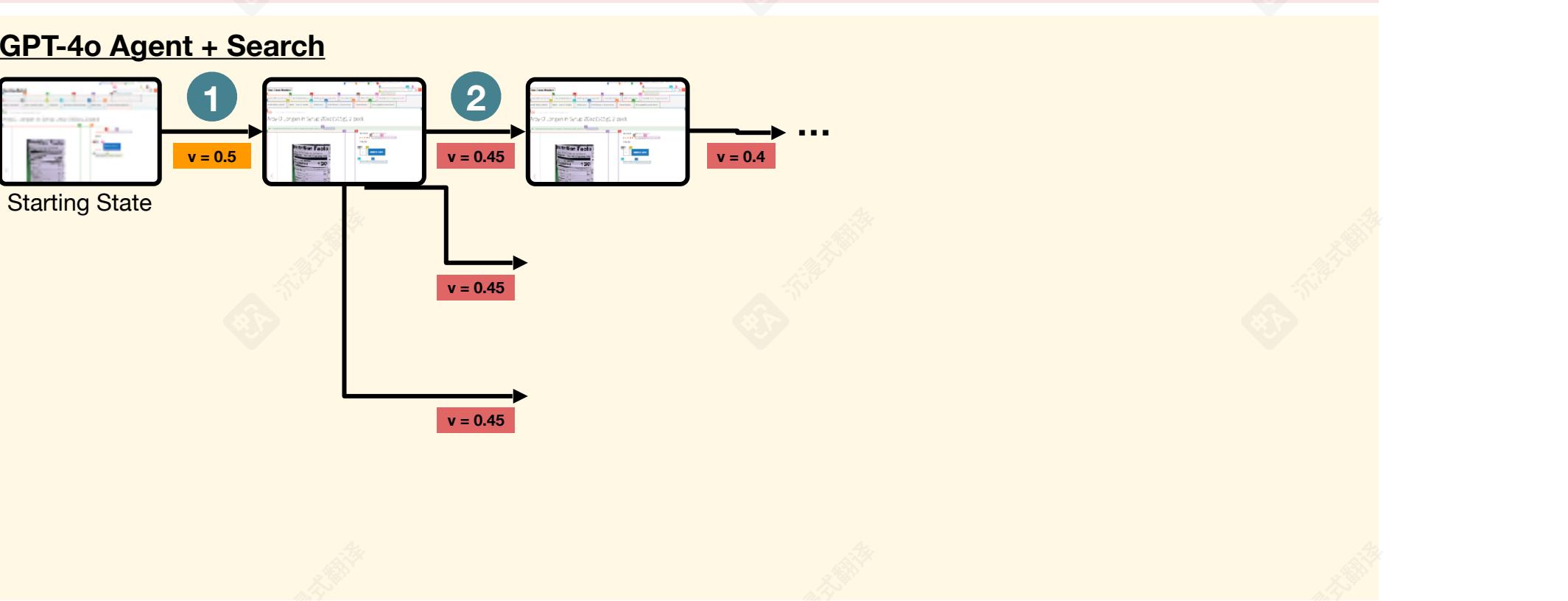
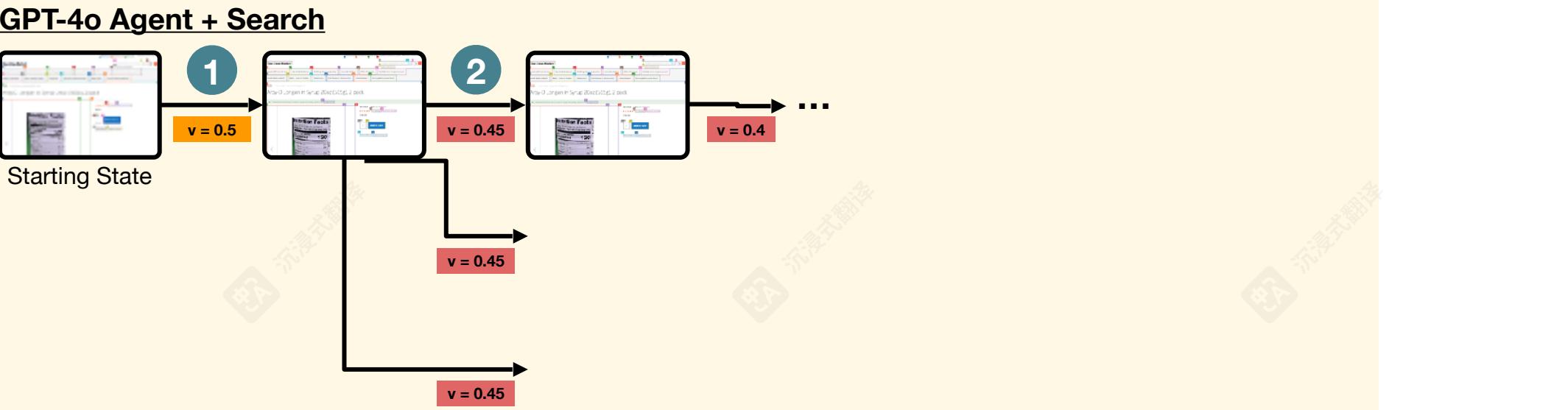
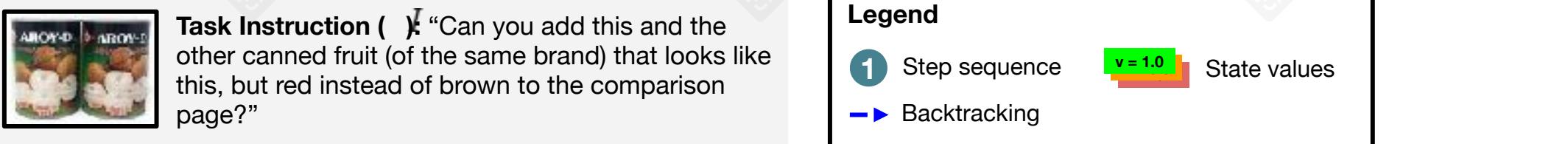
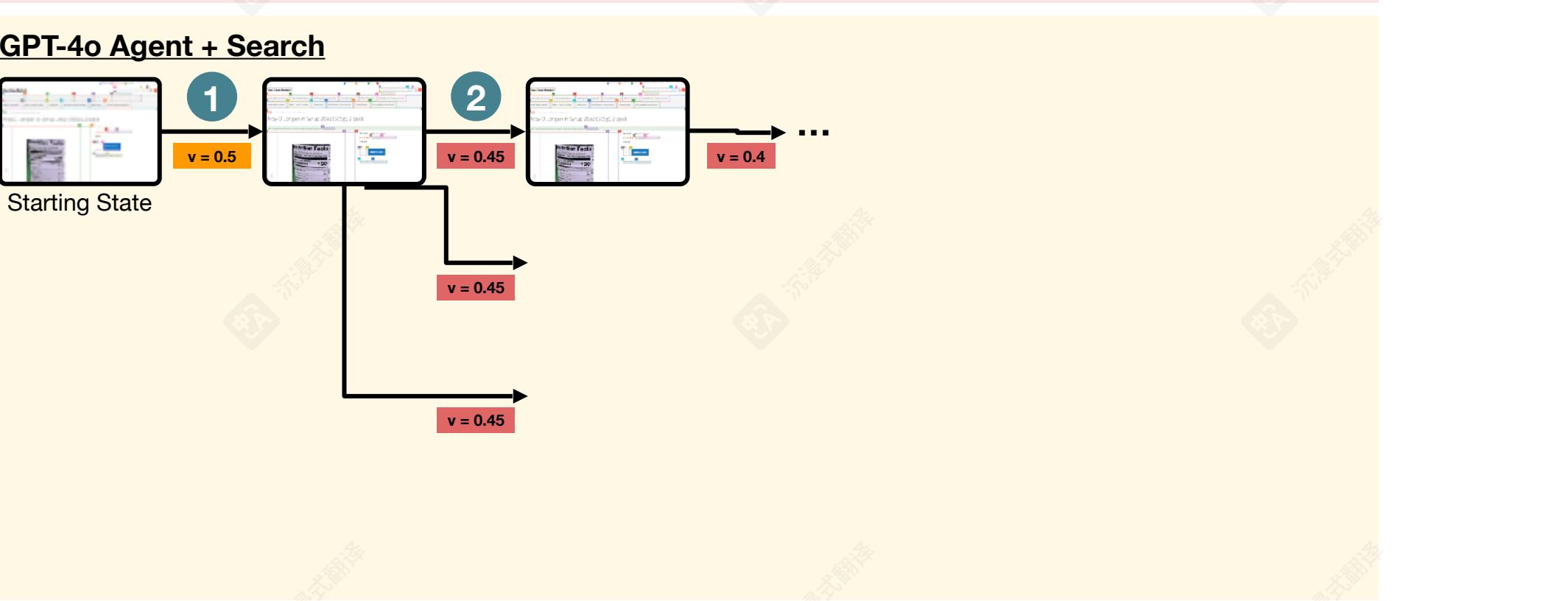
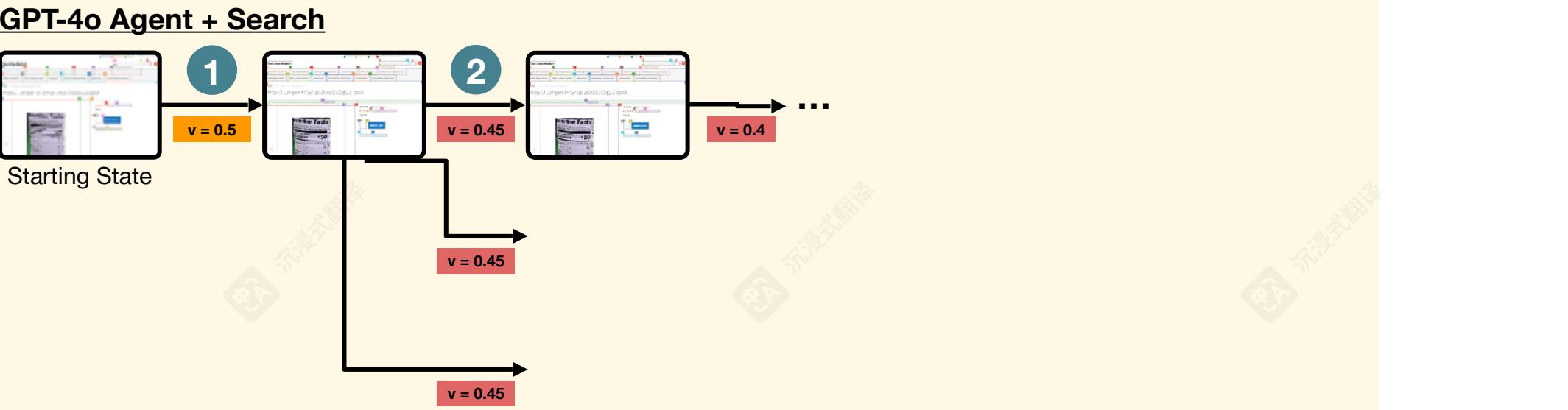
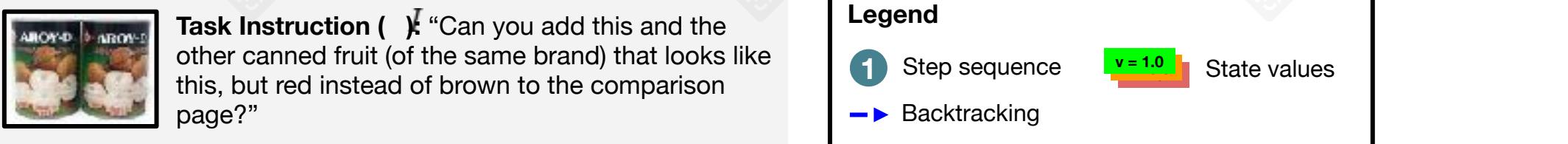
GPT-4o Agent + 搜索



任务指令 (): “你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到对比页面吗？”

GPT-4o Agent + 搜索







Task Instruction (): “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
- Backtracking

v = 1.0 State values

GPT-4o Agent



GPT-4o Agent + Search



Starting State



任务指令 (): “你可以将这个和其它同品牌的罐头水果（看起来像这样，但红色而不是棕色）添加到对比页面吗？”

Legend

- 1 步骤顺序
- Backtracking

v = 1.0 State values

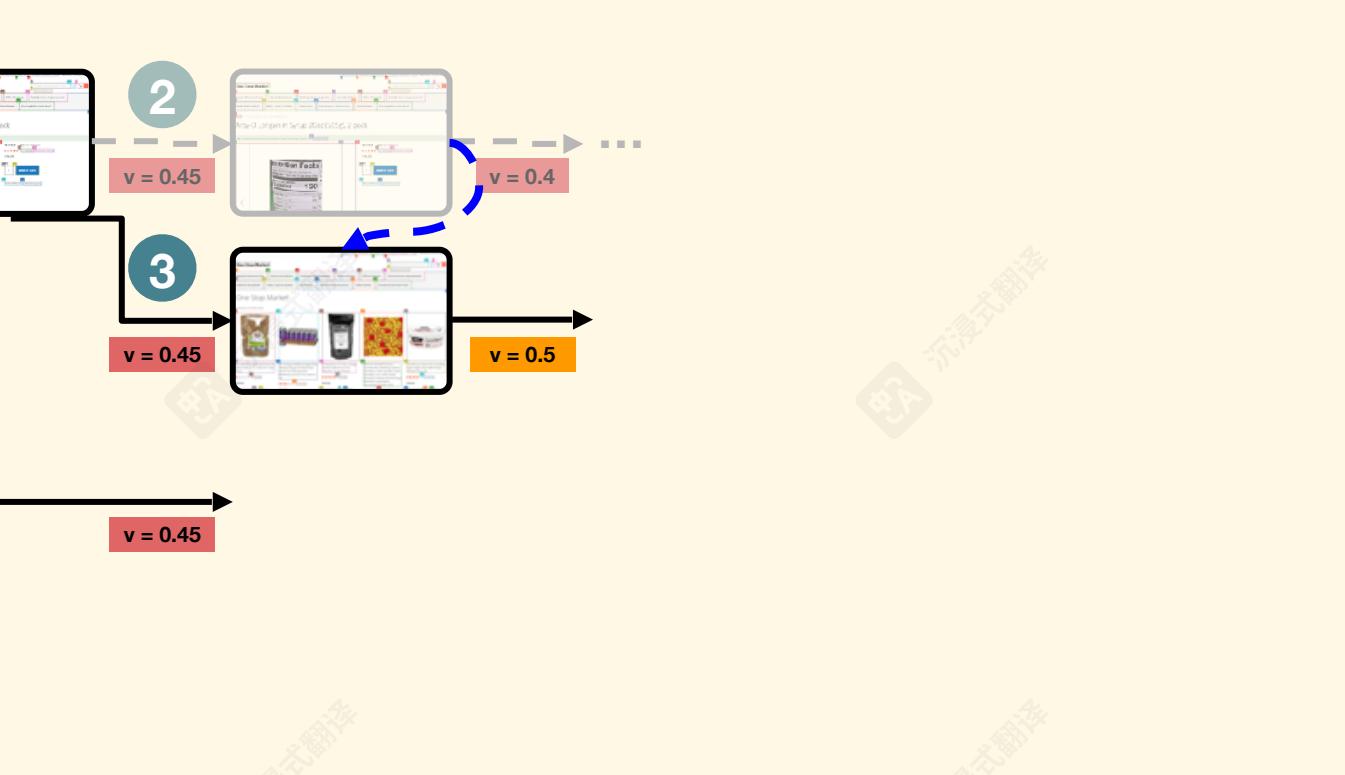
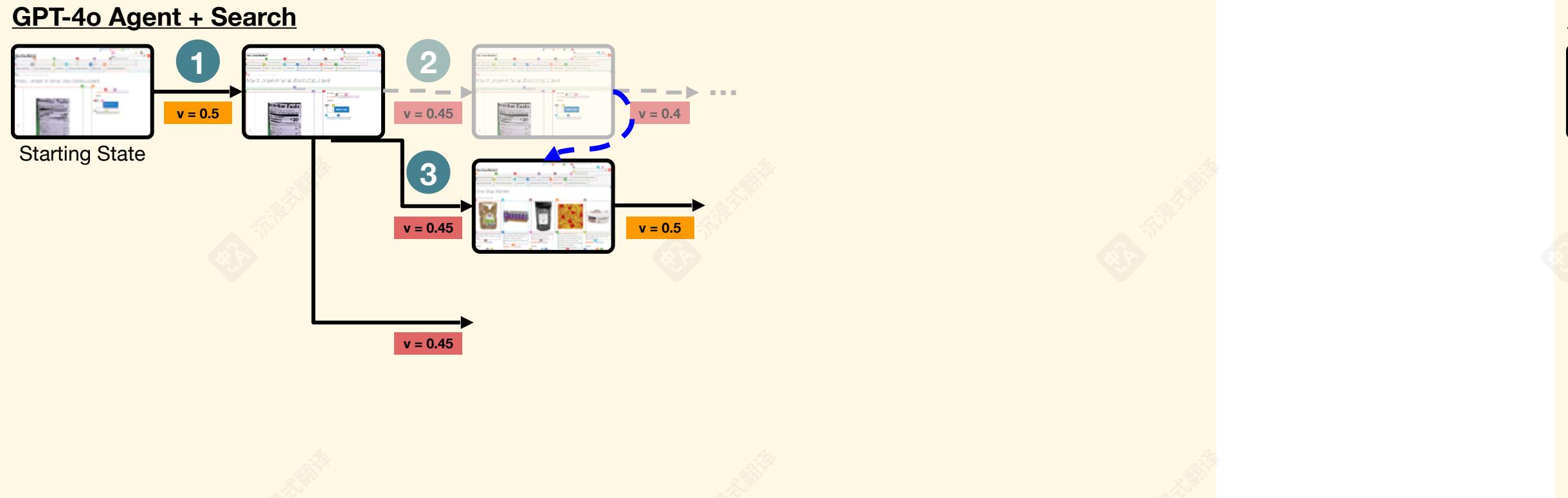
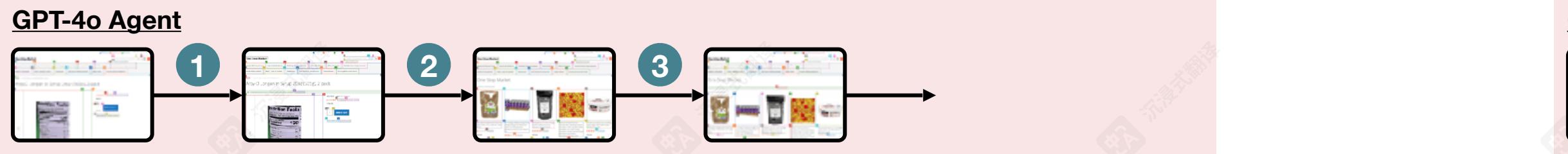
GPT-4o Agent

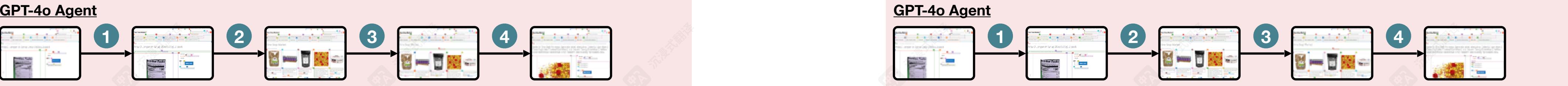
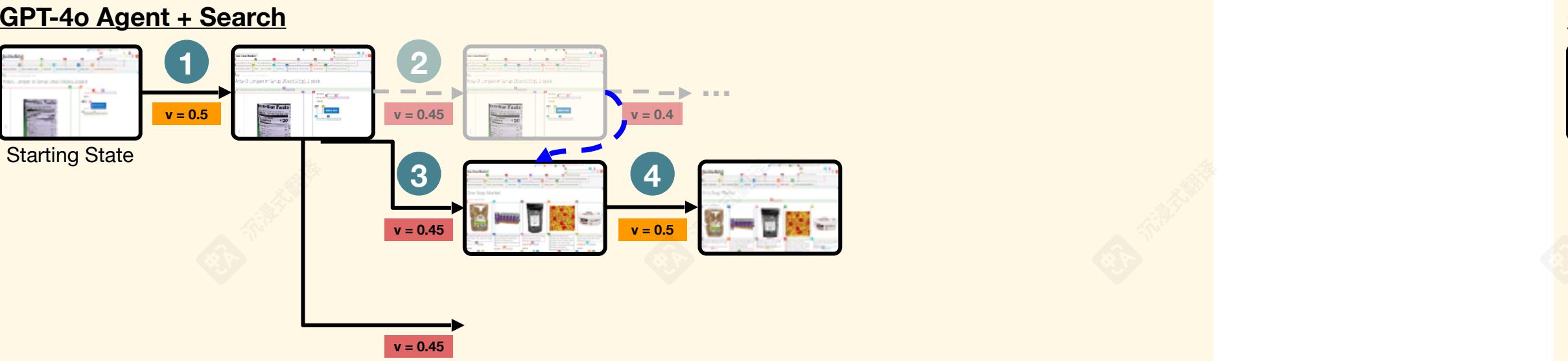
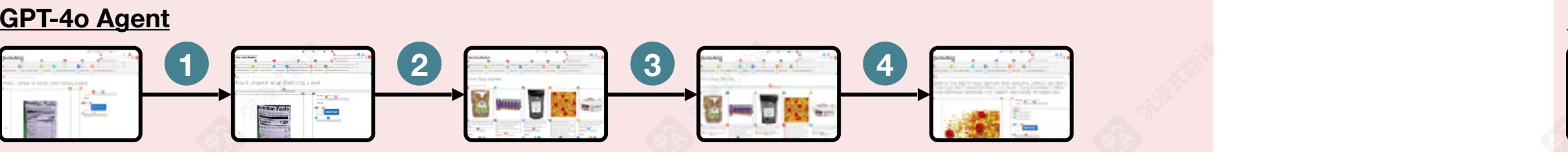


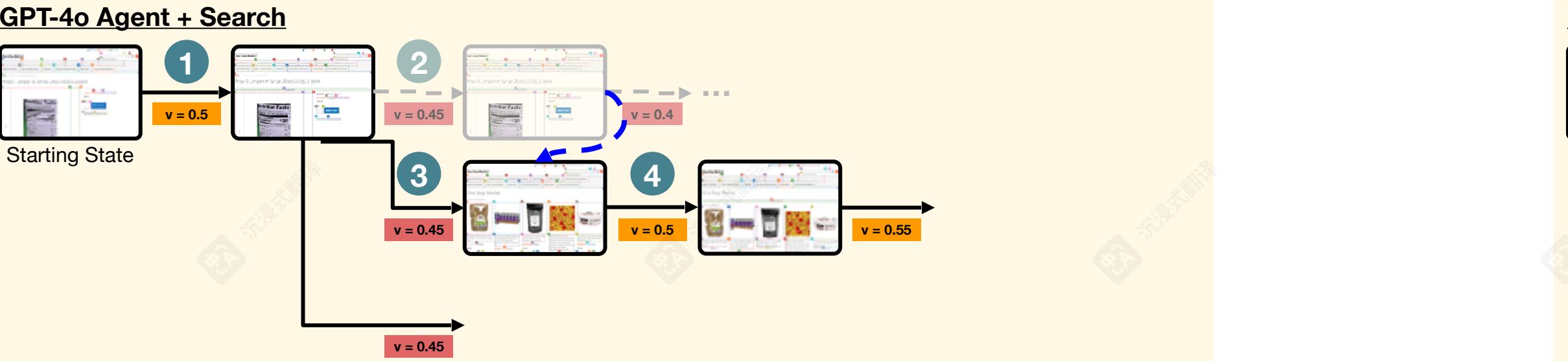
GPT-4o Agent + 搜索

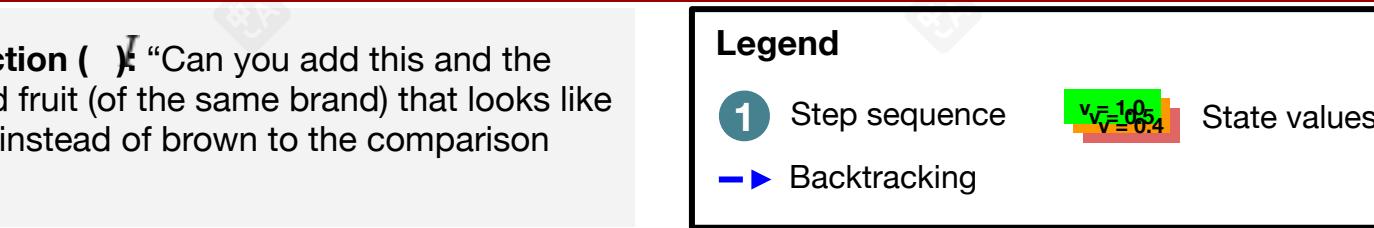
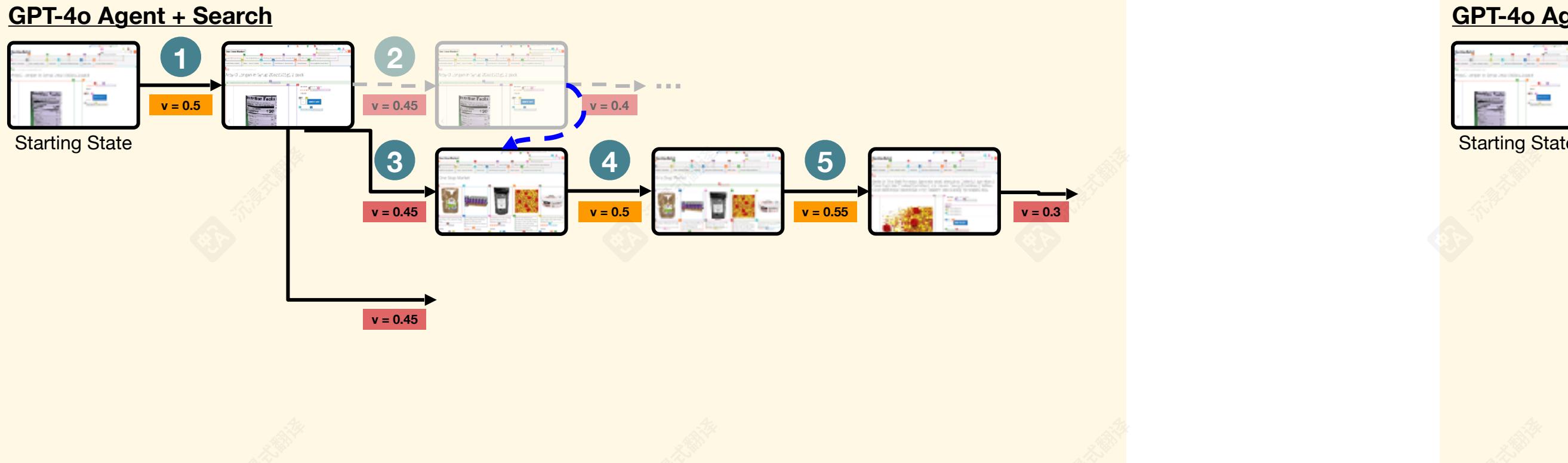
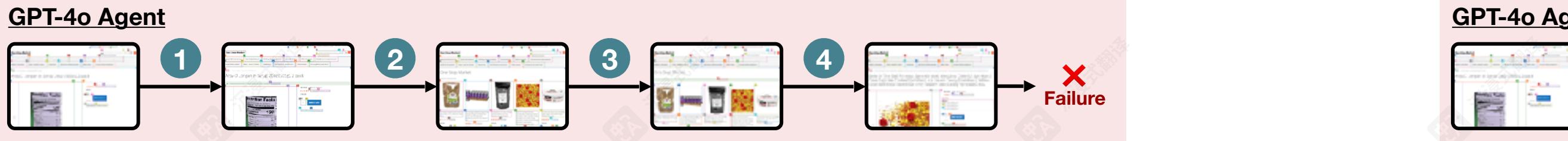
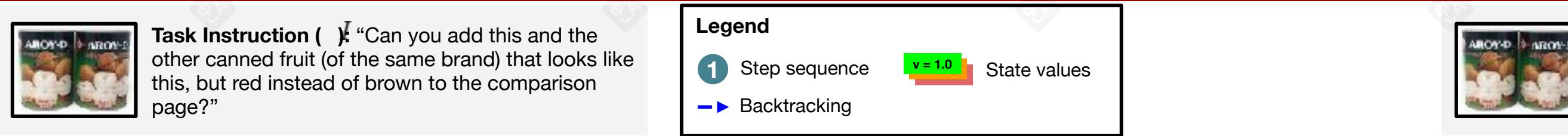


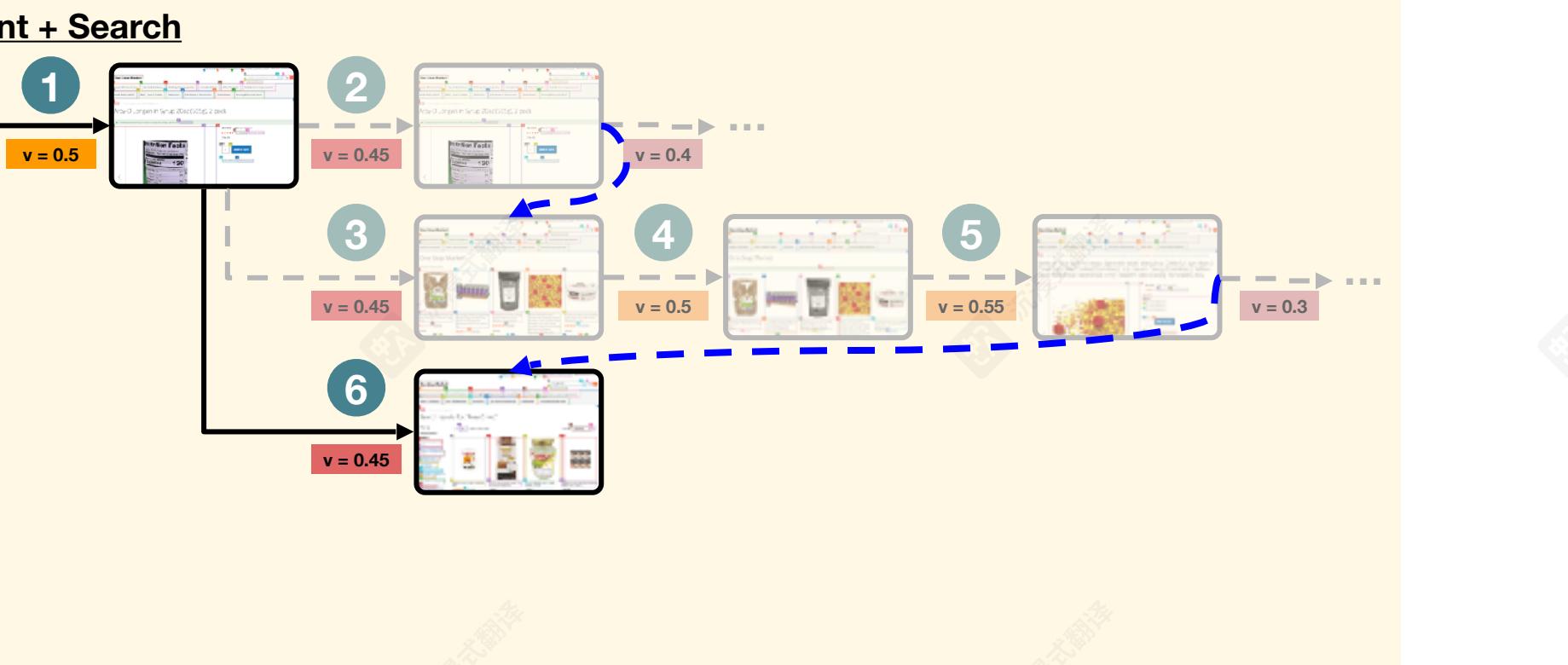
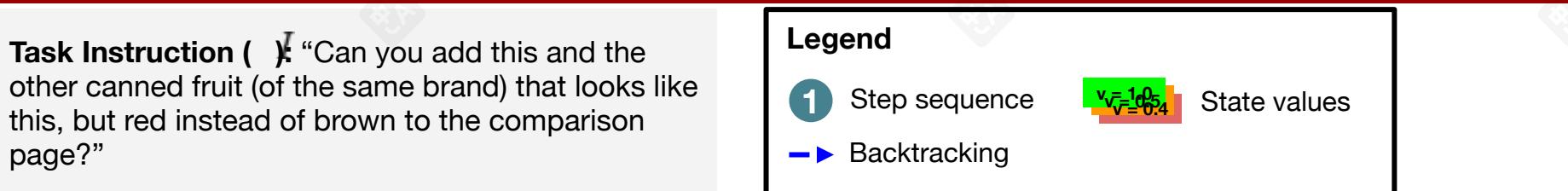
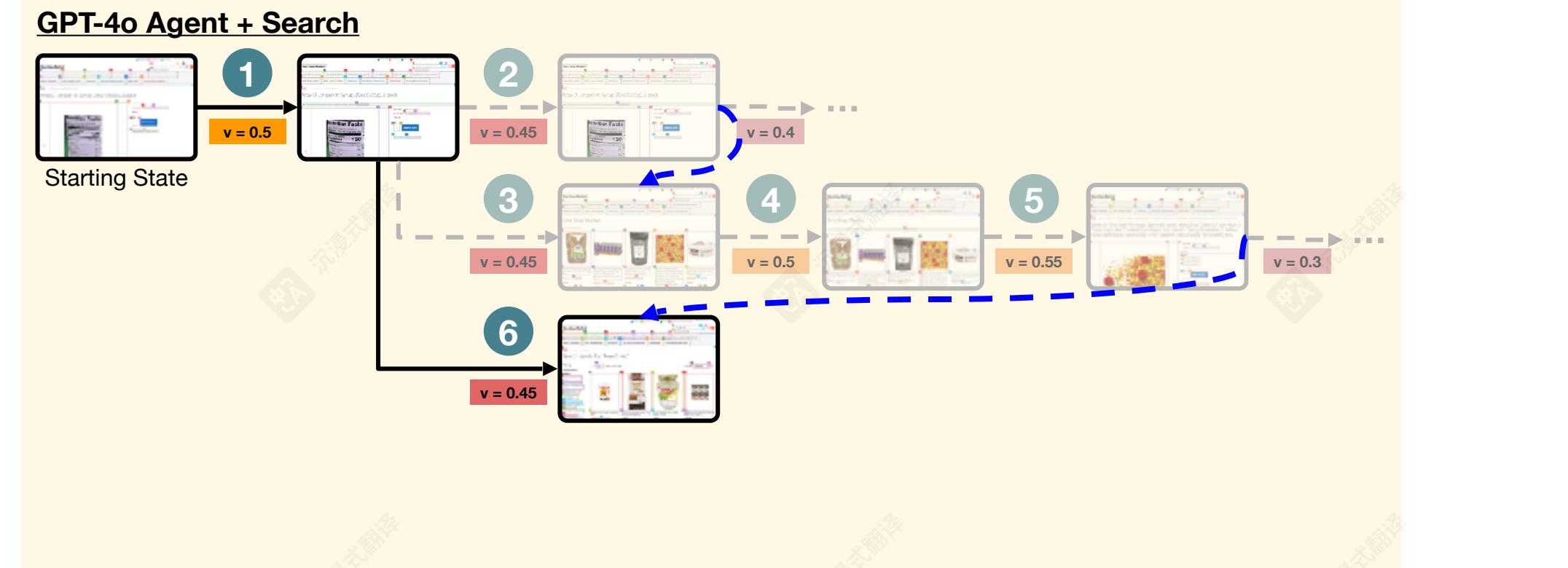
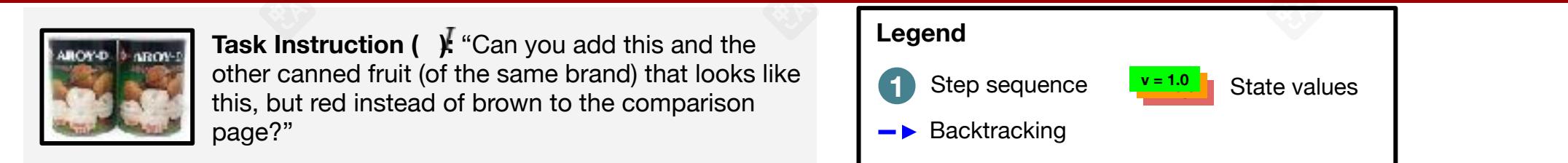
Starting State

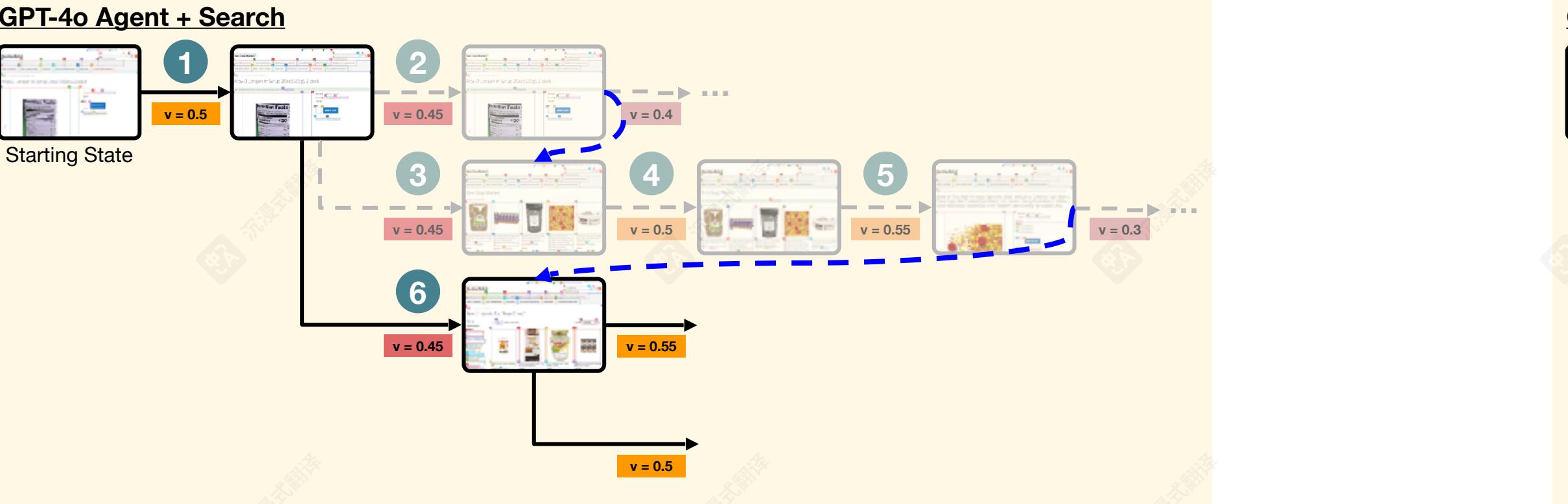
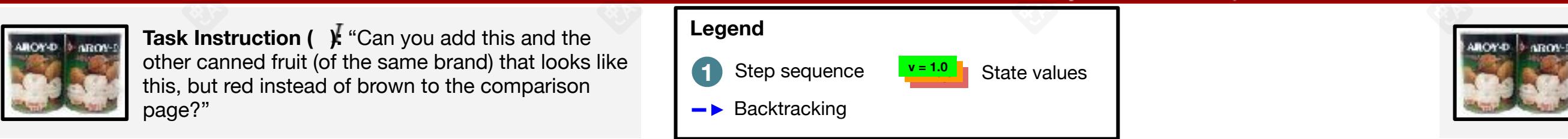


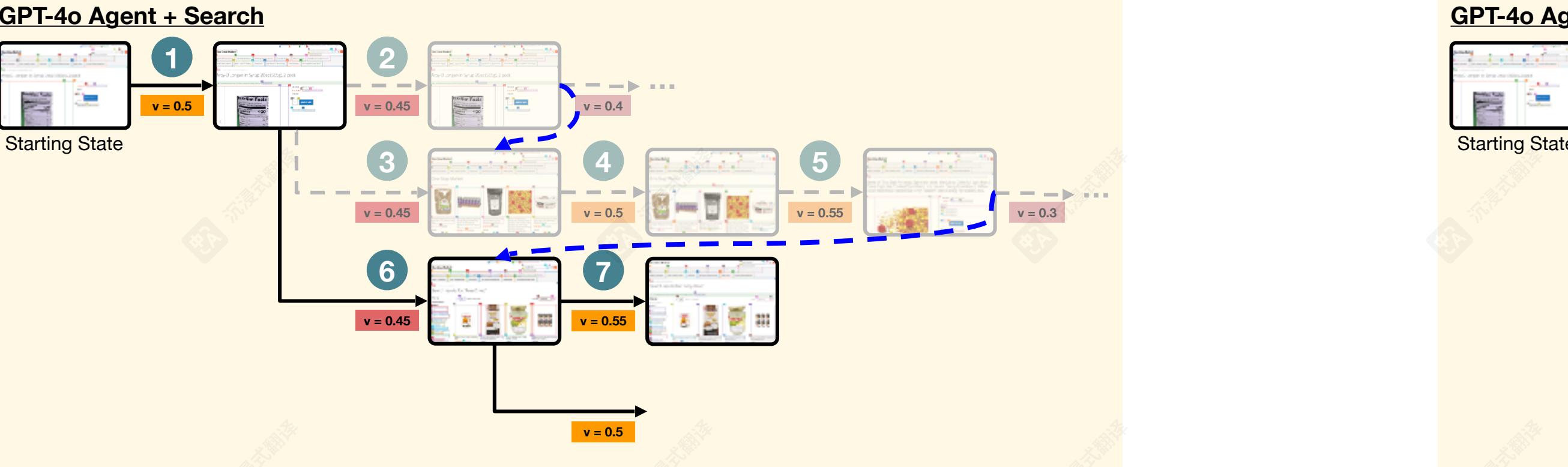
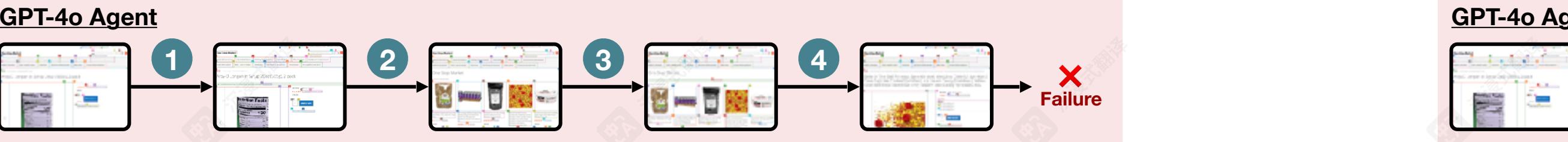
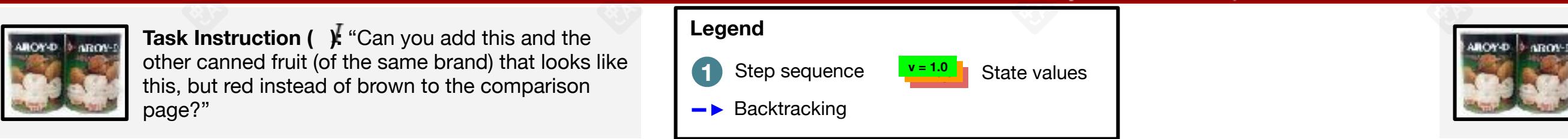


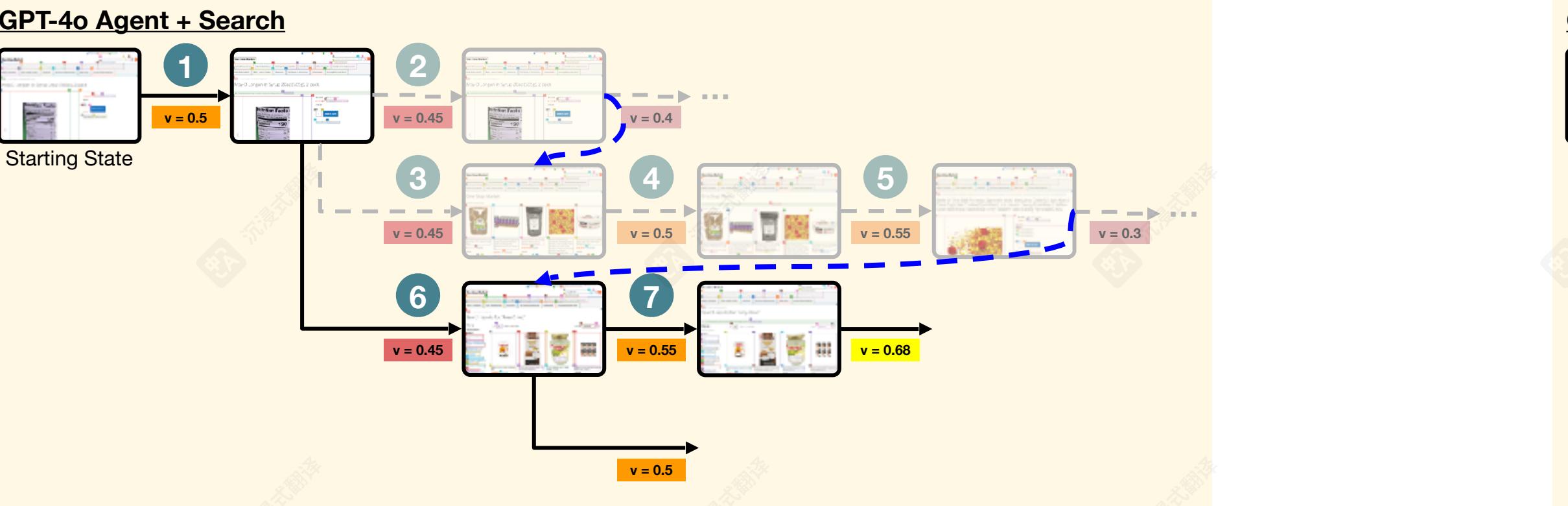
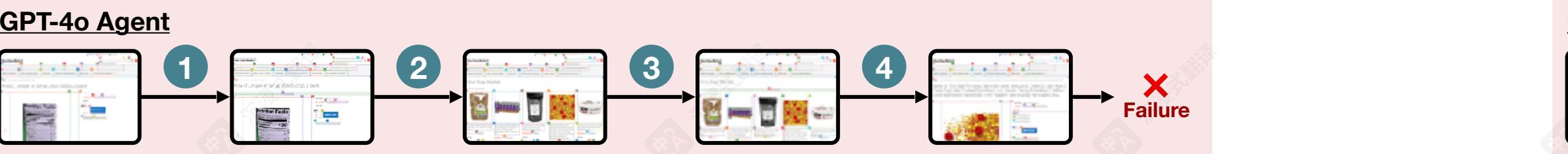
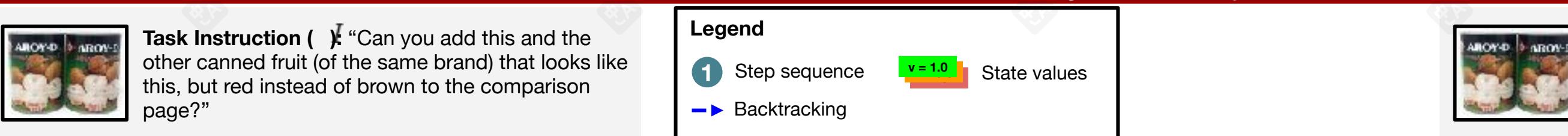


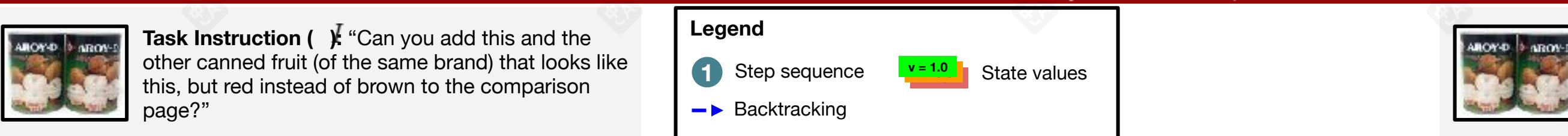


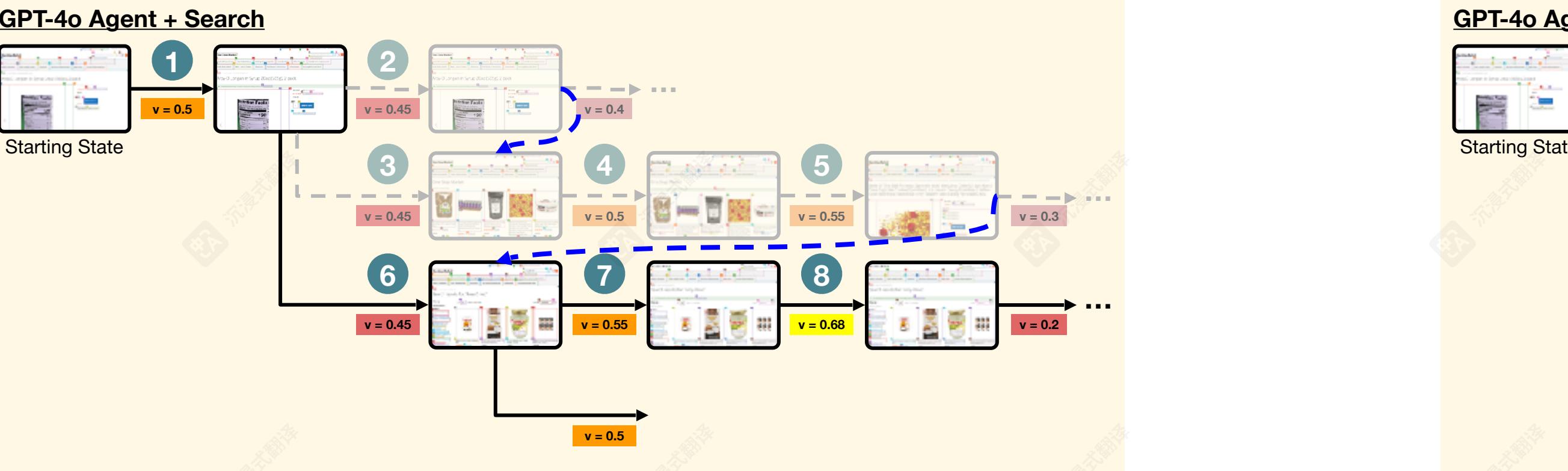
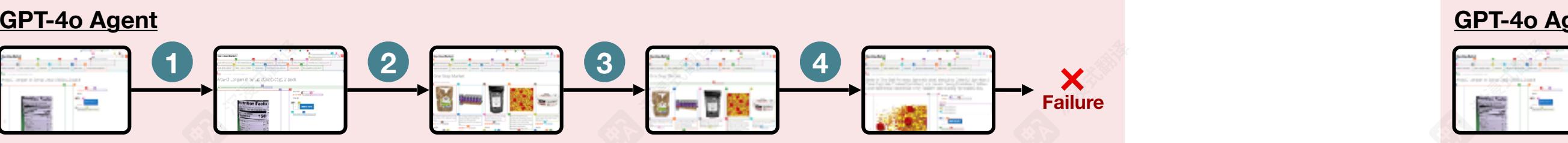
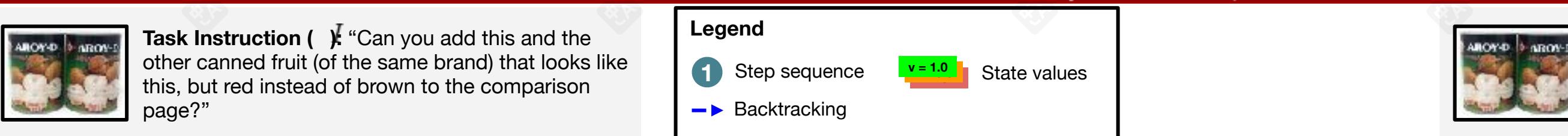


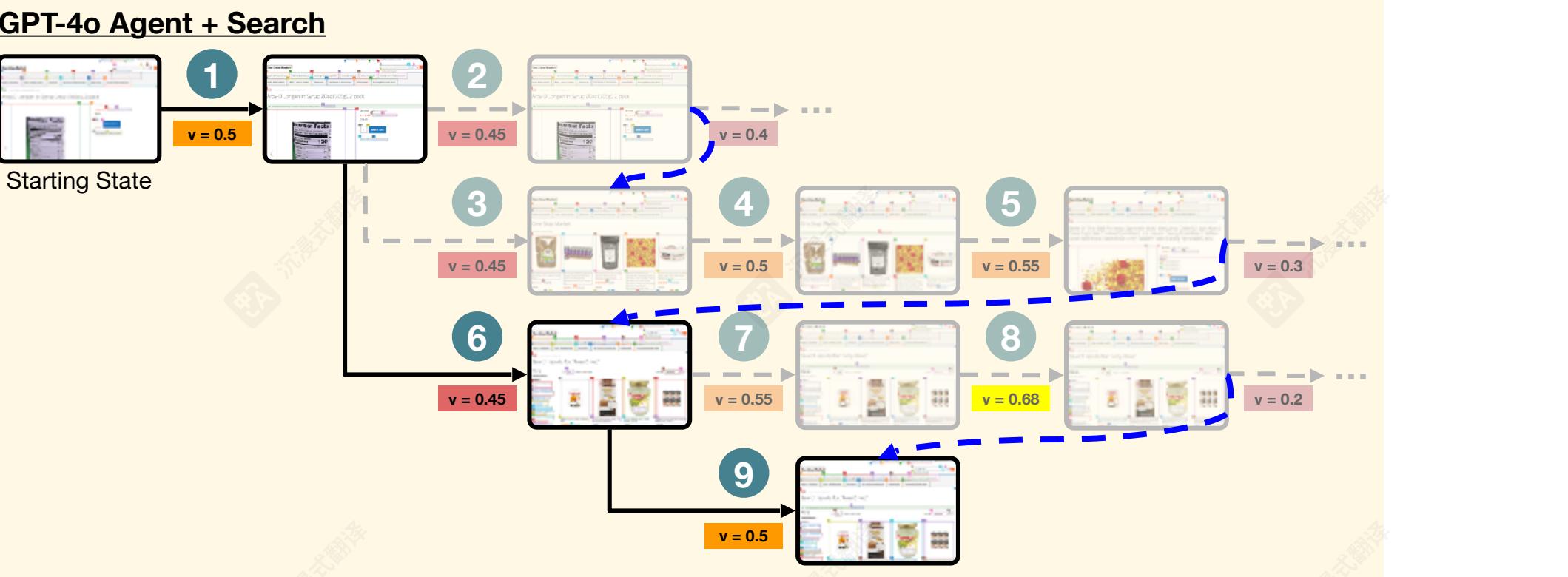
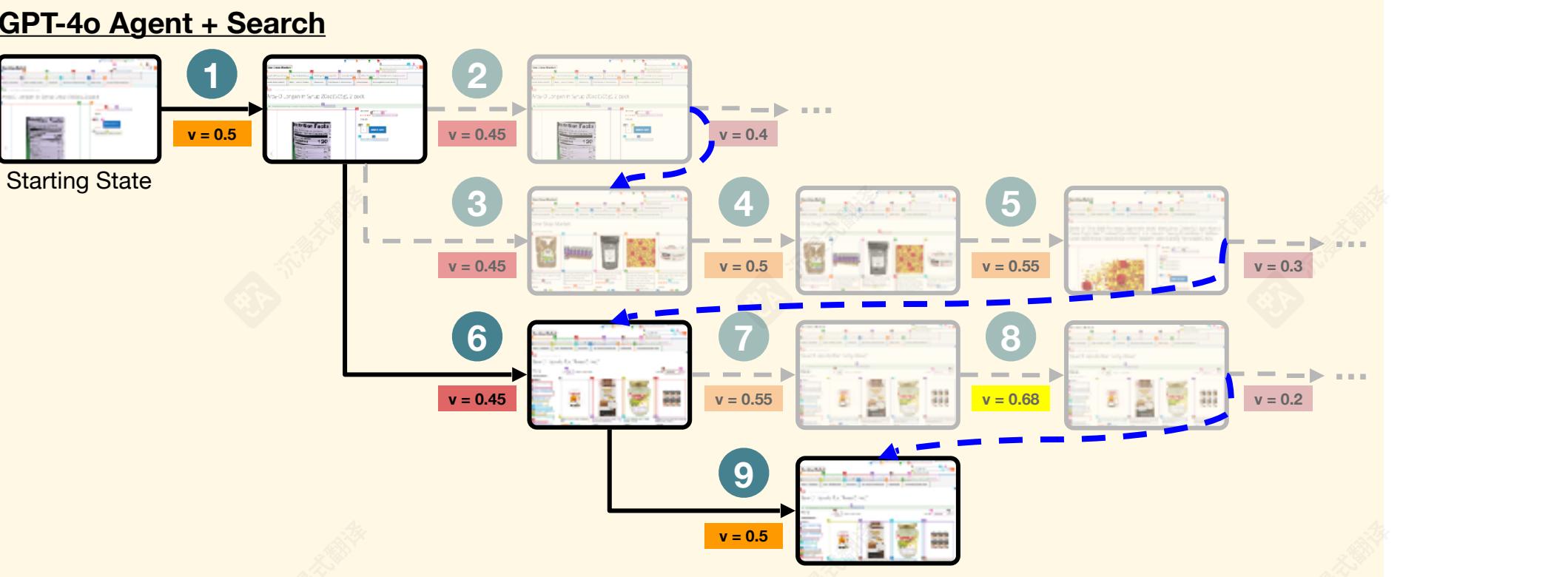
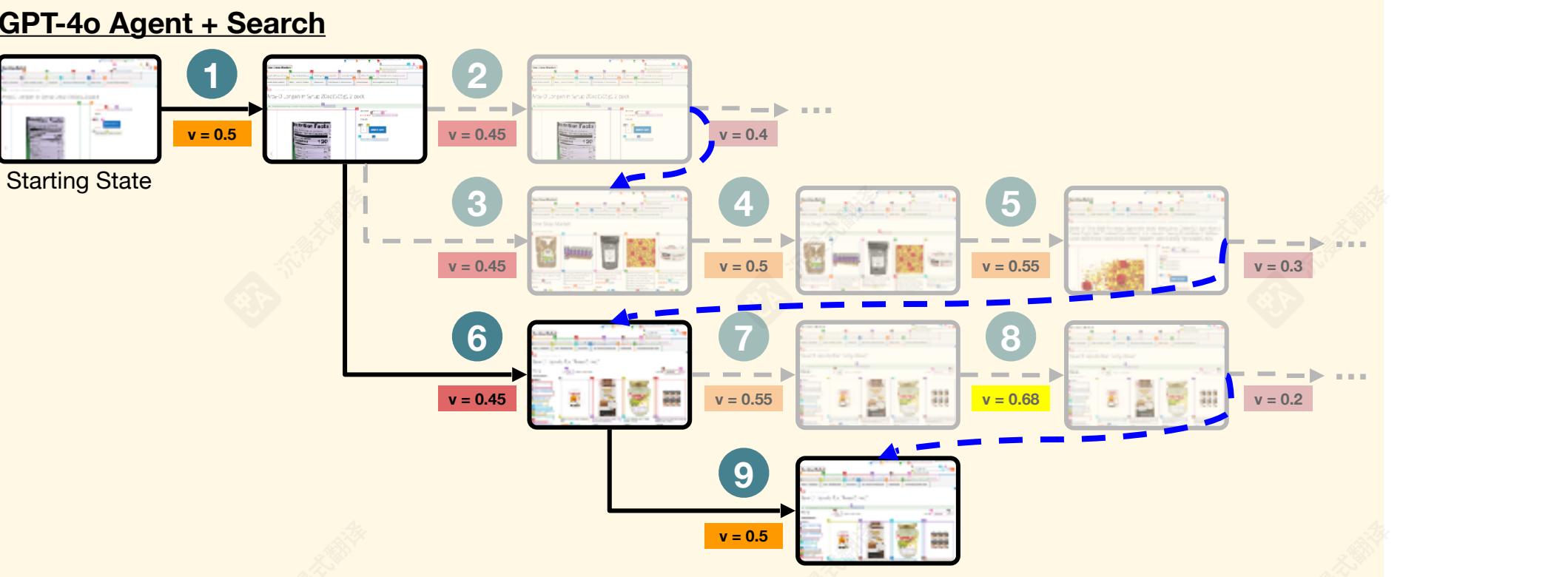
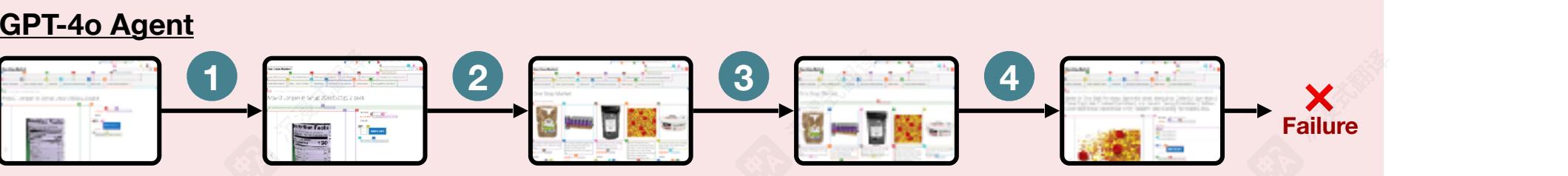


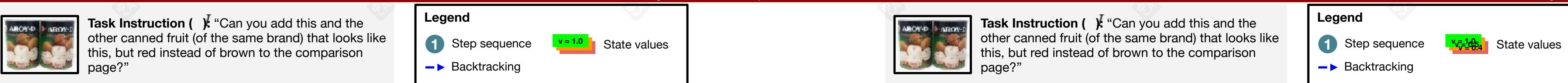
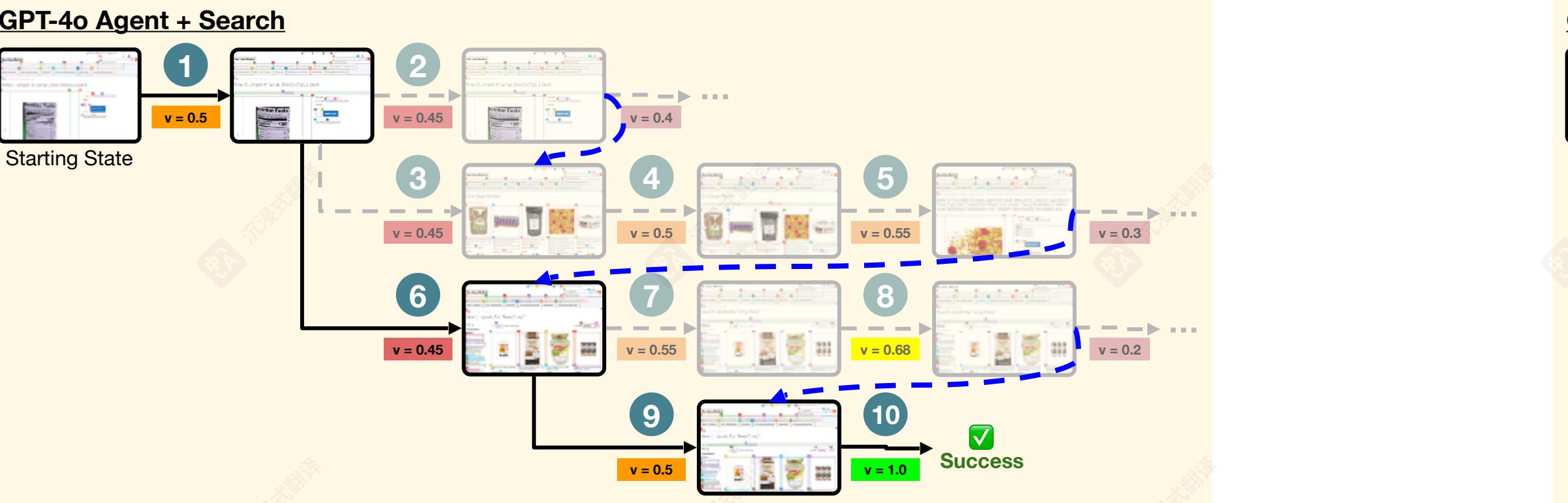
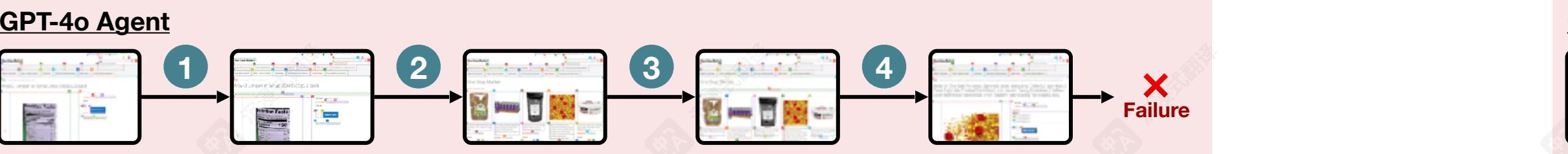
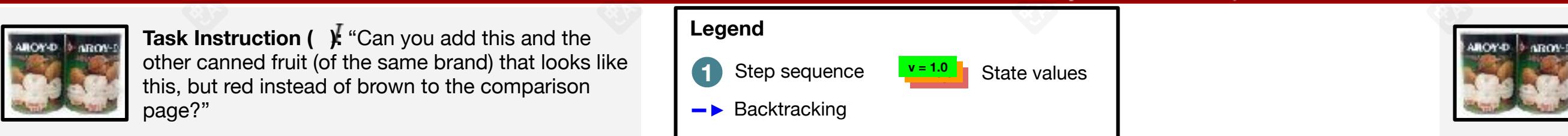






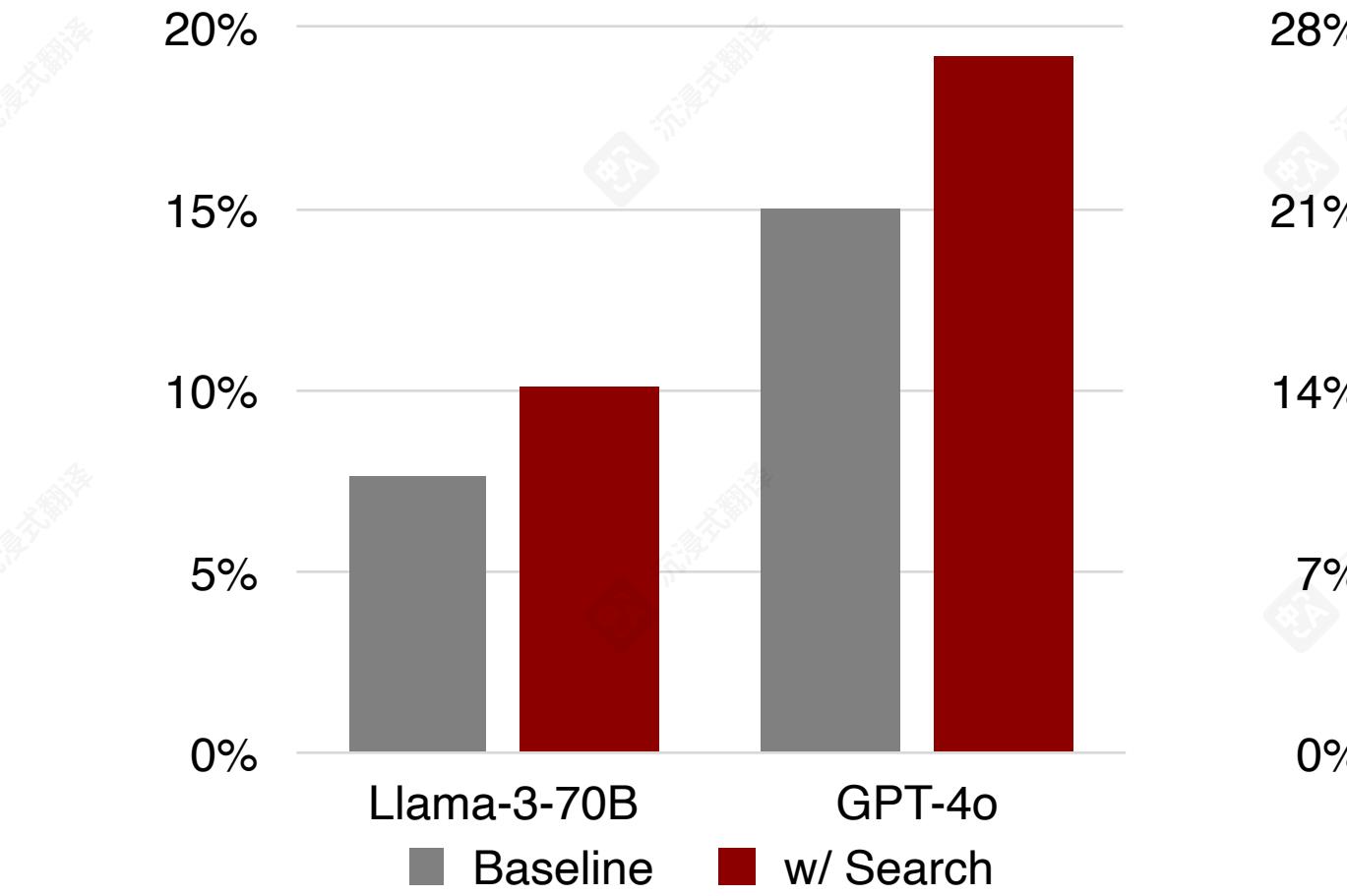




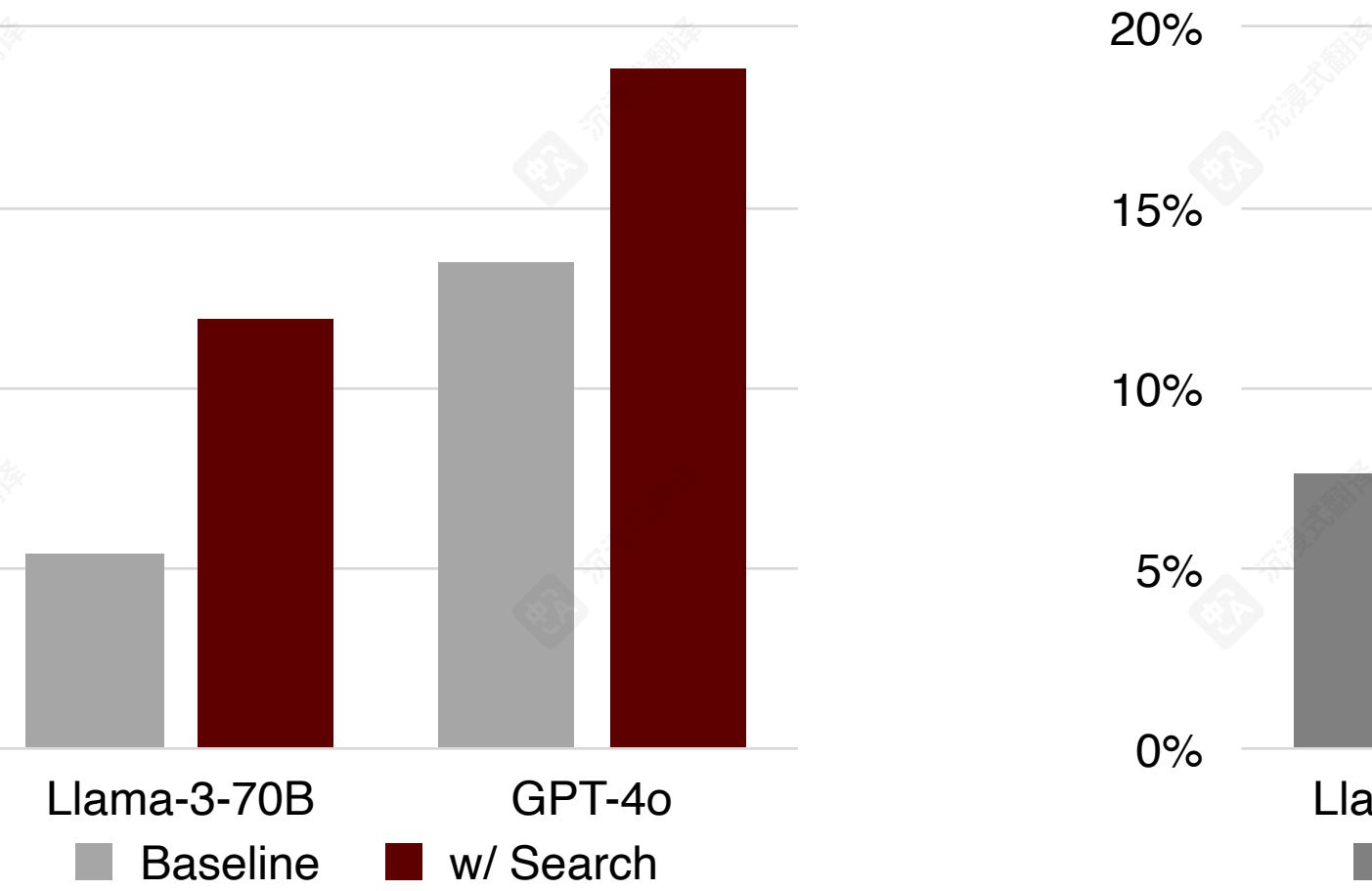


Results

WebArena

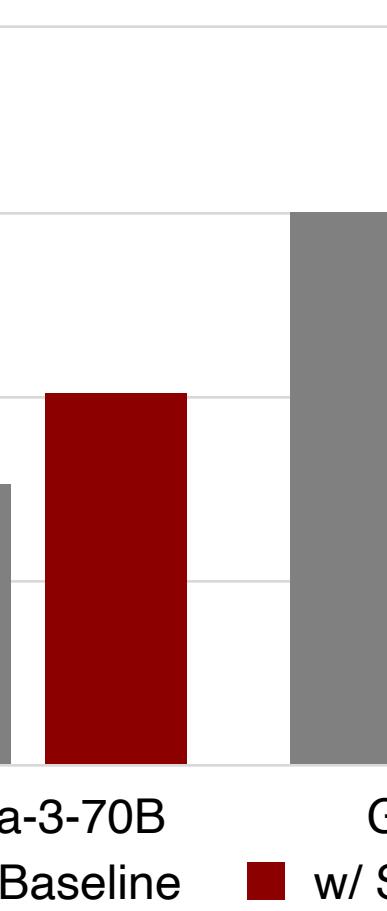


VisualWebArena

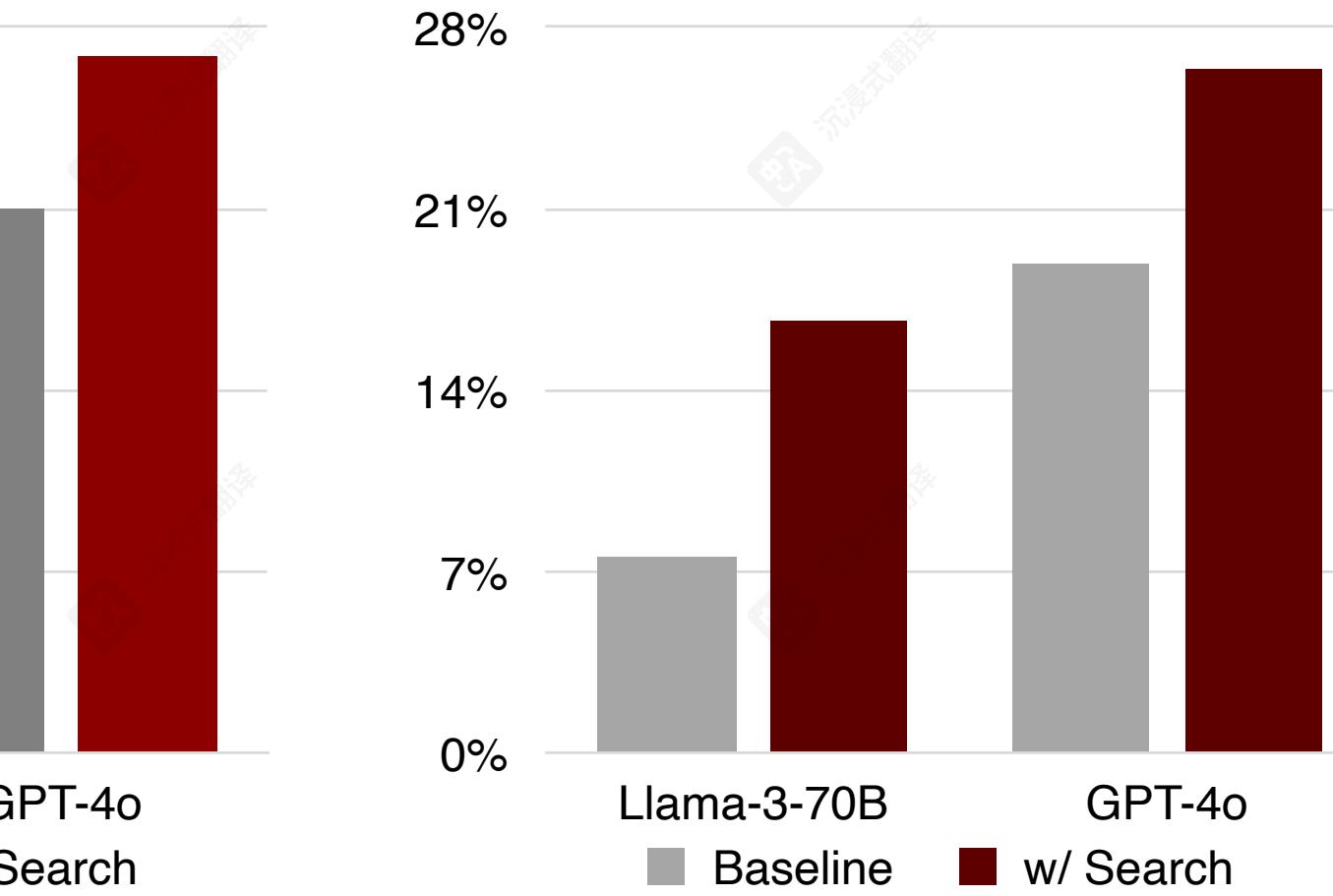


结果

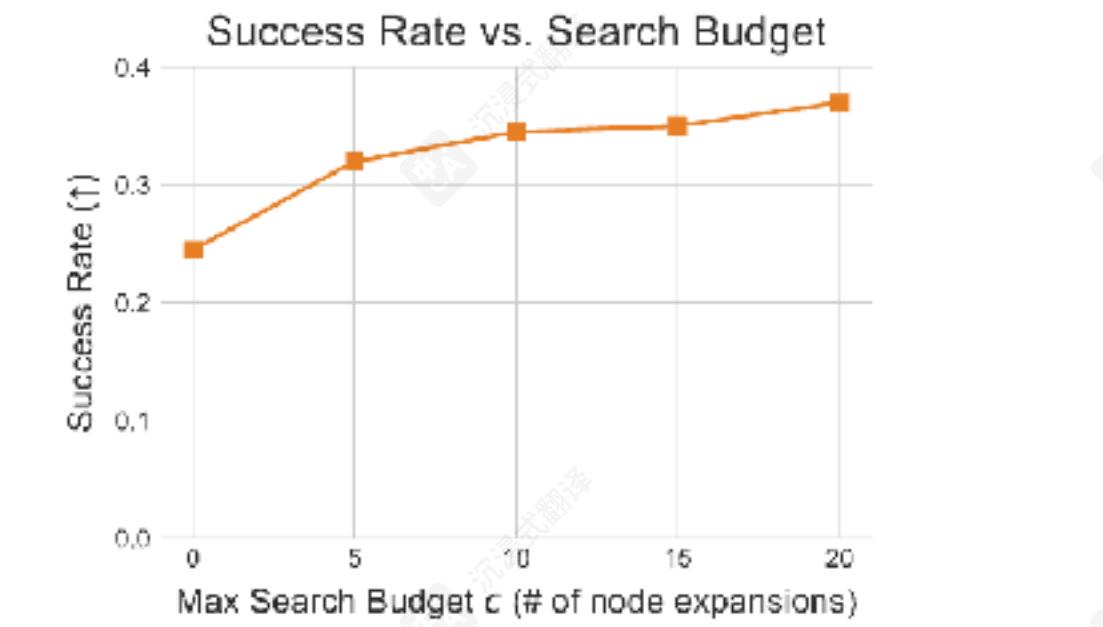
WebArena



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 (↑)	Δ
0	1	24.5%	0%
1	3	26.0%	+6%
	5	32.0%	+31%
2	3	31.5%	+29%
	5	35.0%	+43%
3	5	35.5%	+45%
	5	37.0%	+51%

Success rate (SR) and relative change over the baseline (Δ) 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 $c = 20$.

消融实验



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 (↑)	Δ
0	1	24.5%	0%
1	3	26.0%	+6%
	5	32.0%	+31%
2	3	31.5%	+29%
	5	35.0%	+43%
3	5	35.5%	+45%
	5	37.0%	+51%

Success rate (SR) and relative change over the baseline (Δ) 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.

Value Function	SR (\uparrow)
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.

Ablations

- 一个好的价值函数至关重要。
- 对于基础智能体策略和价值函数的改进，仍然有很大的提升空间。

Value Function	SR (\uparrow)
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%

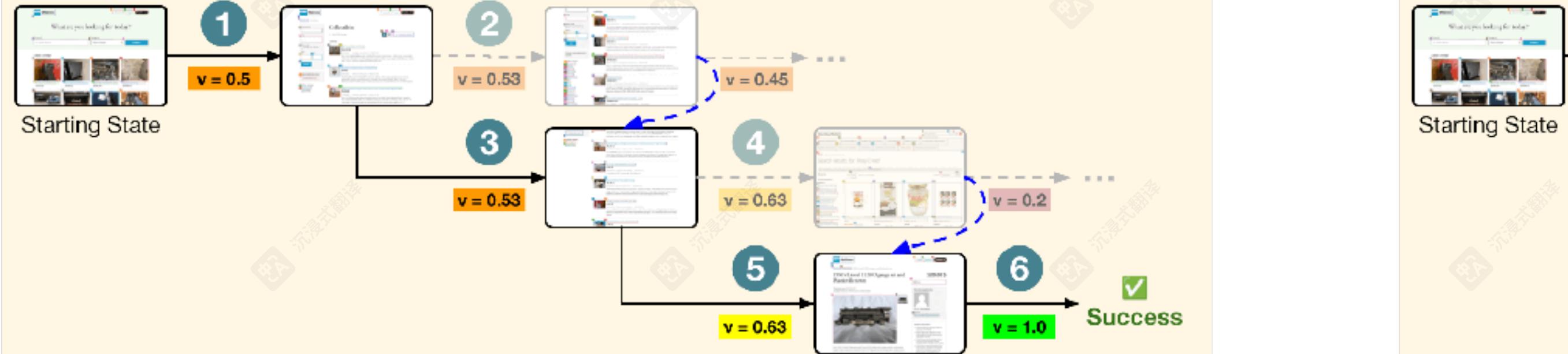
Table 3: Success rate of the GPT-4o agent with different value functions.

Qualitative Results



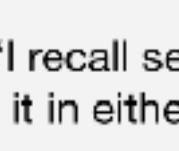
Task Instruction (I): “I recall seeing this exact item on the site, help me find the most recent post of it. I recall seeing it in either the Collectibles or Antiques section.”

GPT-4o Agent + Search



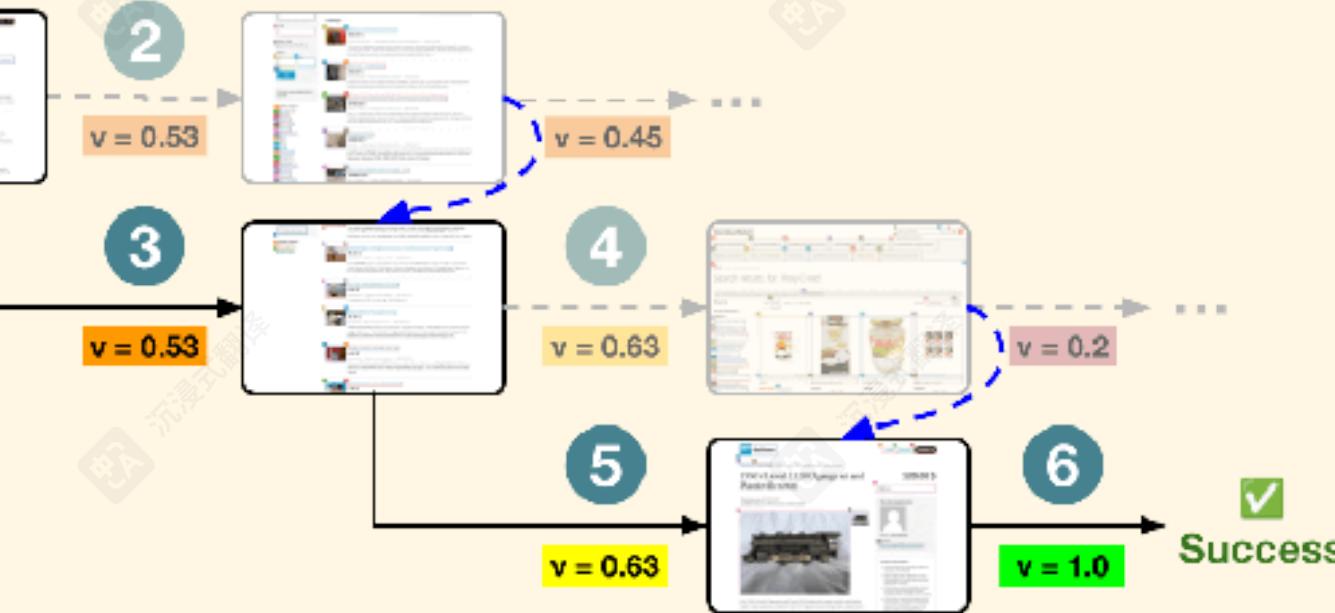
Legend: ① Search sequence -> Backtracking v = 1.0 State values

定性结果



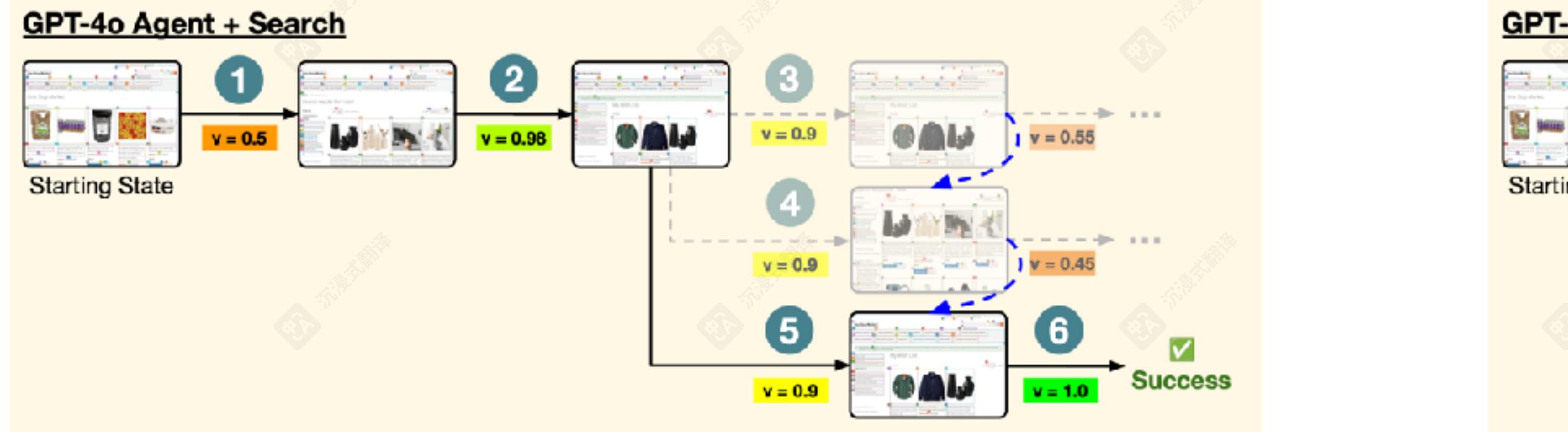
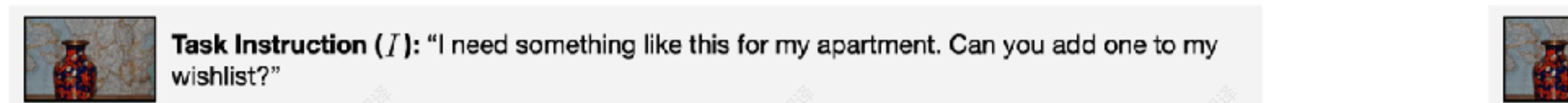
Task Instruction (I): “I recall seeing this exact item on the site, help me find the most recent post of it. I recall seeing it in either the Collectibles or Antiques section.”

GPT-4o Agent + Search

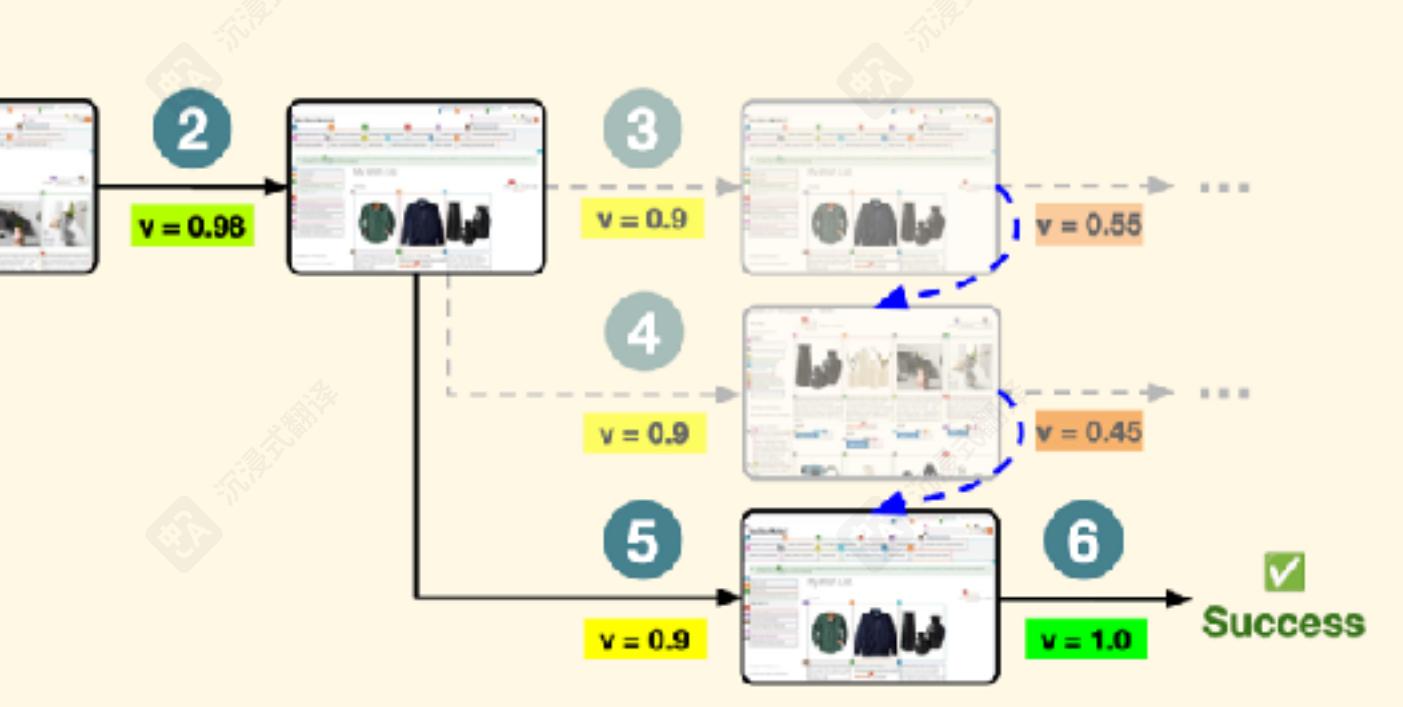
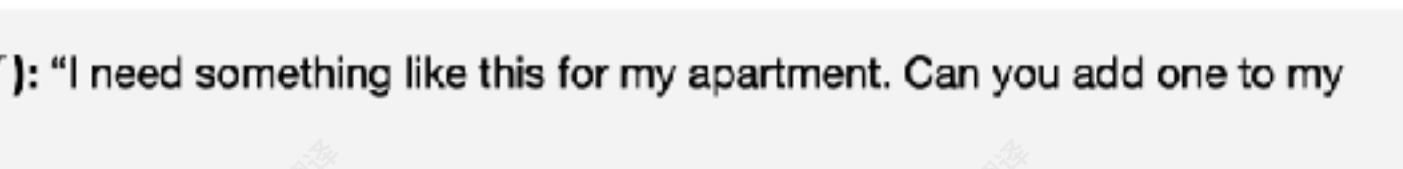


Legend: ① Search sequence -> Backtracking v = 1.0 State values

Qualitative Results



定性结果



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

局限性

- 搜索缓慢 - 我们以一种相对原始的方式实现了回溯（将操作存储在队列中，再次获取它们以回到原始状态）
- 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?**

当前工作

- 将搜索作为策略改进函数
- 通过微调而不是提示来改进值函数
- 探索改进基线代理与在推理时 `<code>` 进行更多搜索 `</code>` 之间的计算权衡
- 如果我们没有完美的模拟器 ——`<code>` 我们如何才能大规模收集数据 `</code>?`

Talk Outline

- VisualWebArena -- Evaluating Multimodal Agents on Realistic Visual Web Tasks (Koh et al., ACL 2024)
- Tree Search for Language Model Agents (Koh, McAleer, Fried, Salakhutdinov, arXiv 2024)
- Towards Internet-Scale Training For Agents (Trabucco, Sigurdsson, Piramuthu, Salakhutdinov, arXiv 2025)

Talk Outline

- VisualWebArena -- 在真实视觉网络任务上评估多模态智能体 (Koh 等人 , ACL 2024)
- 语言模型智能体的树搜索 (Koh, McAleer, Fried, Salakhutdinov, arXiv 2024)
- 面向互联网规模的智能体训练 (Trabucco, Sigurdsson, Piramuthu, Salakhutdinov, arXiv 2025)

Agents Suffer From A Data Problem

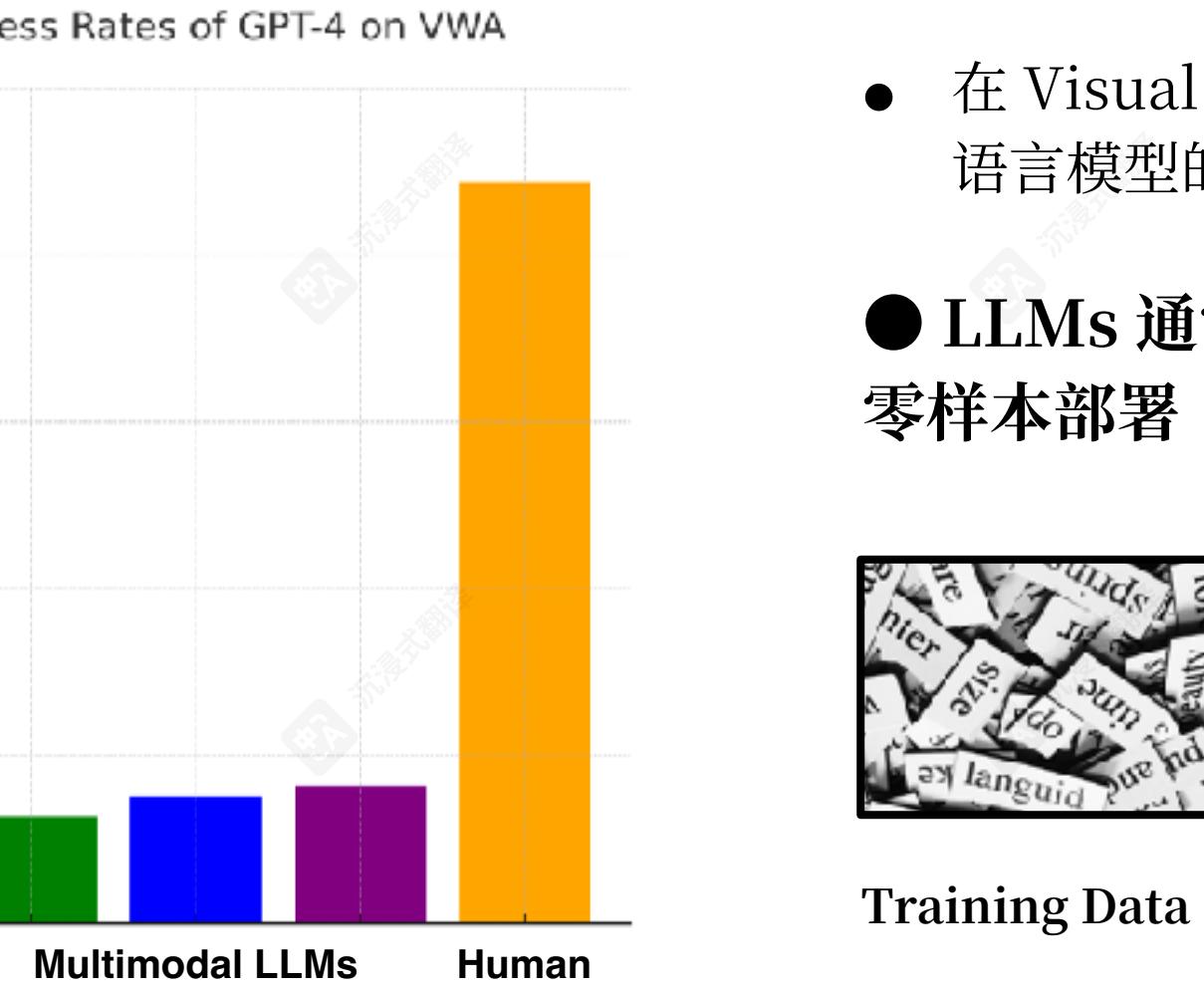
- Top LLMs fall short of humans by 68.92% on Visual Web Arena
- LLMs are often **trained offline**, then **deployed zero-shot** as agents



Training Data



Test Environment



Agents Suffer From A Data Problem

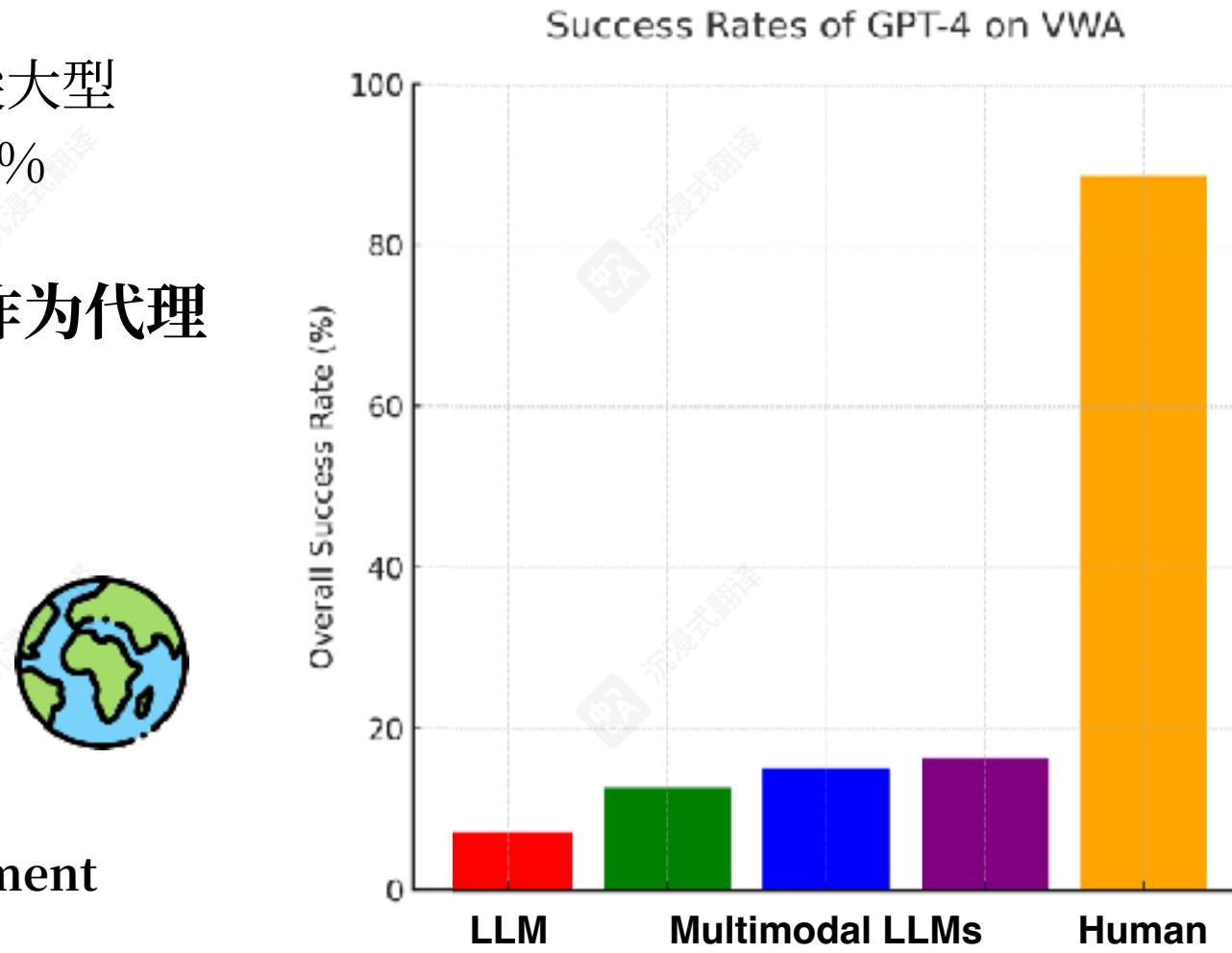
- 在 Visual WebArena 上，顶尖大型语言模型的性能比人类低 68.92%
- LLMs 通常离线训练，然后作为代理零样本部署



Training Data

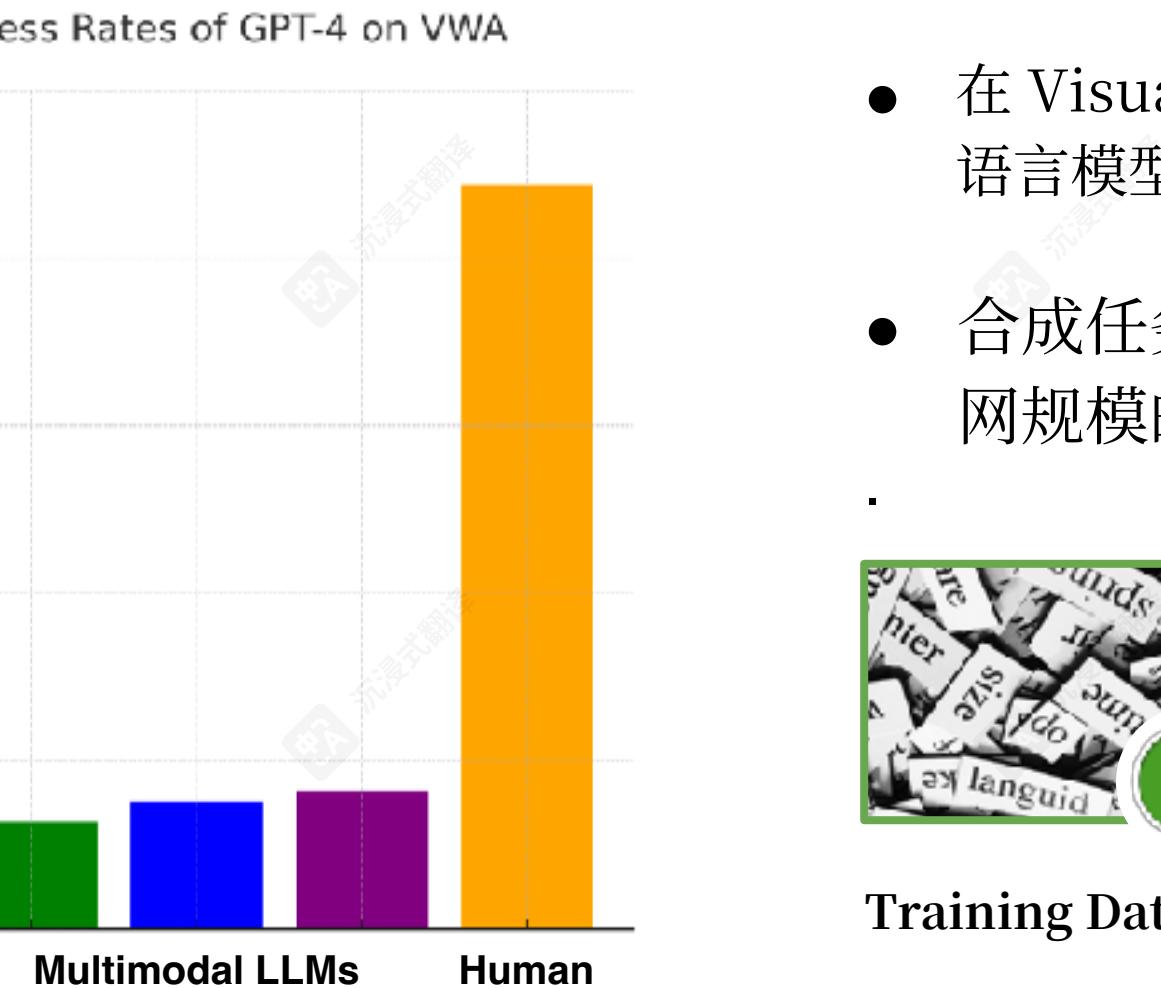
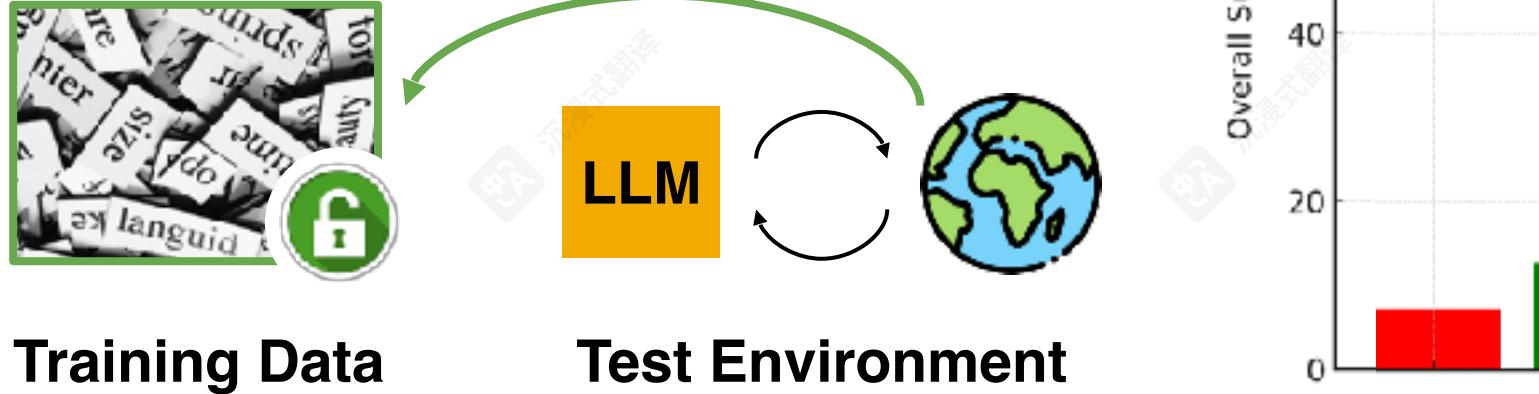


Test Environment



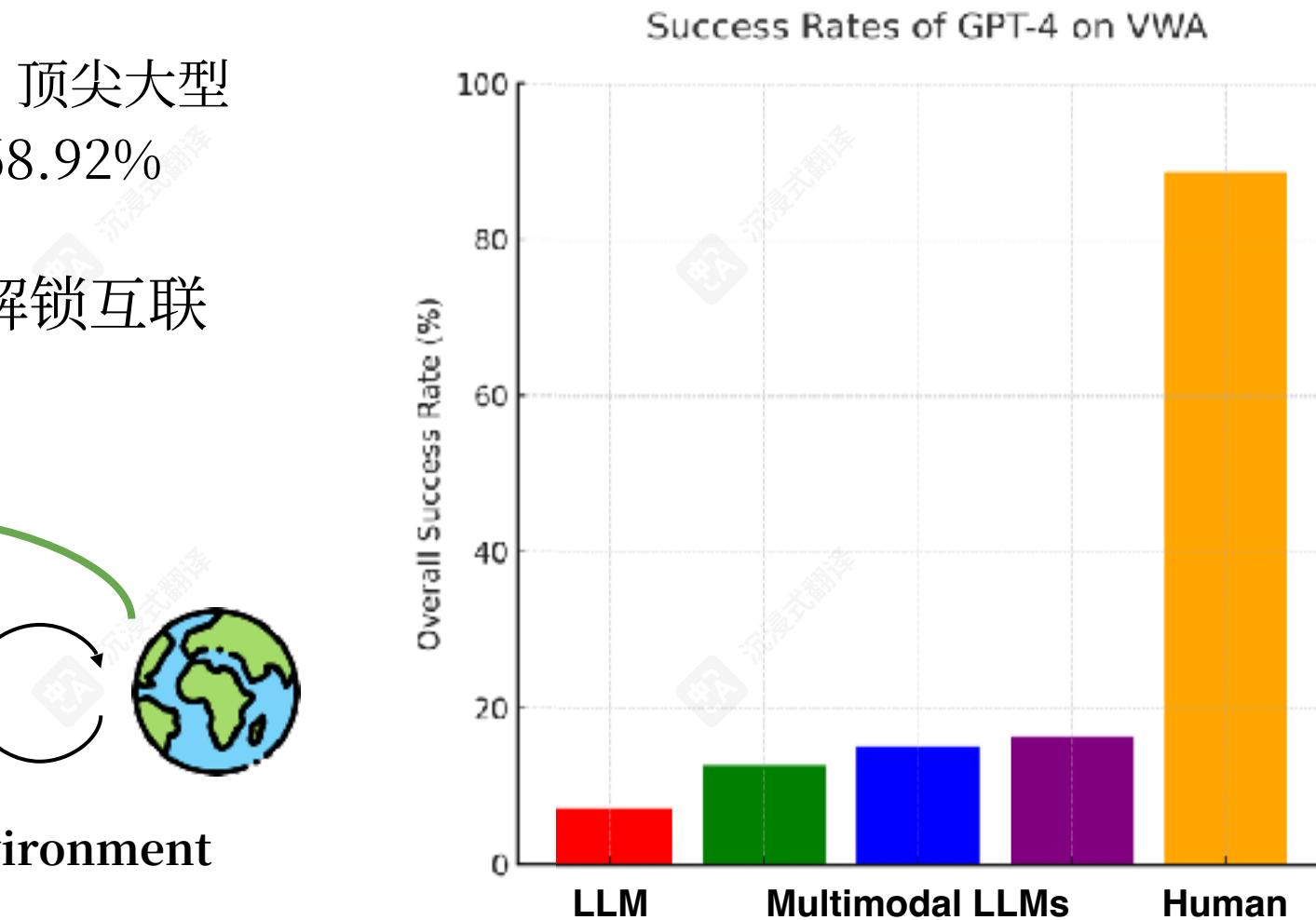
Agents Suffer From A Data Problem

- Top LLMs fall short of humans by 68.92% on Visual Web Arena
- Can **synthetic tasks** unlock internet-scale training for agents?



Agents Suffer From A Data Problem

- 在 Visual WebArena 上，顶尖大型语言模型的性能比人类低 68.92%
- 合成任务能否为智能体解锁互联网规模的训练？



Towards Internet-Scale Training For Agents (InSTA)



- Can synthetic tasks unlock internet-scale training for agents?
- **Key Idea:** use Llama to **generate and verify** synthetic agentic tasks

Stage 1: Task Generation

www.github.com

LLM Find a codebase for generating images with Flux.1 [dev].

Stage 2: Task Evaluation



Codebase found:
Flux supported:
Task solved:

Stage 3: Data Collection

www.github.com
www.stackoverflow.com
www.uefi.org
www.javatpoint.com
manuals.playstation.net
calculator.bcis.co.uk
research.vu.nl
...
(150k sites)

面向代理的互联网规模训练 (InSTA)

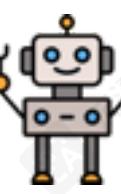
- 合成任务能否解锁代理的互联网规模训练?
- **Key Idea:** use Llama to **generate and verify** synthetic agentic tasks

阶段 1: 任务生成

www.github.com

LLM 寻找一个使用 Flux.1 [dev] 生成图像的代码库。

阶段 2: 任务评估



Codebase found:
Flux supported:
Task solved:

阶段 3: 数据收集

www.github.com
www.stackoverflow.com
www.uefi.org
www.javatpoint.com
manuals.playstation.net
calculator.bcis.co.uk
research.vu.nl
...
(150k sites)



Use Llama To Generate Agentic Tasks

- Given a web domain as text (i.e. merseyferries.co.uk)
- Propose a realistic task that an average user could complete in one session.

Use Llama 生成代理任务

- 给定一个网站域名作为文本（例如 .merseyferries.co.uk ）
- 提出一个普通用户可以在一次会话中完成的现实任务。

Use Llama To Generate Agentic Tasks

- Given a web domain as text (i.e. merseyferries.co.uk)
- Propose a realistic task that an average user could complete in one session.

Llama-3.1-70B-Instruct: describe the “merseyferries.co.uk” site

Mersey Ferries (merseyferries.co.uk) is the **official website for the famous Mersey Ferry service** in Liverpool, England. This website serves as a one-stop platform for users to access various services and information related to the ferry operation. Here are some things a user can do on this site:

1. **Plan a Journey**: Users can plan their journey by checking the ferry timetable, which includes departure and arrival times for both commuter and cruise services.
2. **Book Tickets**: Users can book tickets online for both commuter and cruise services, including the popular Manchester Ship Canal Cruise and the Liverpool Bay Cruise.

Use Llama To Generate Agentic Tasks

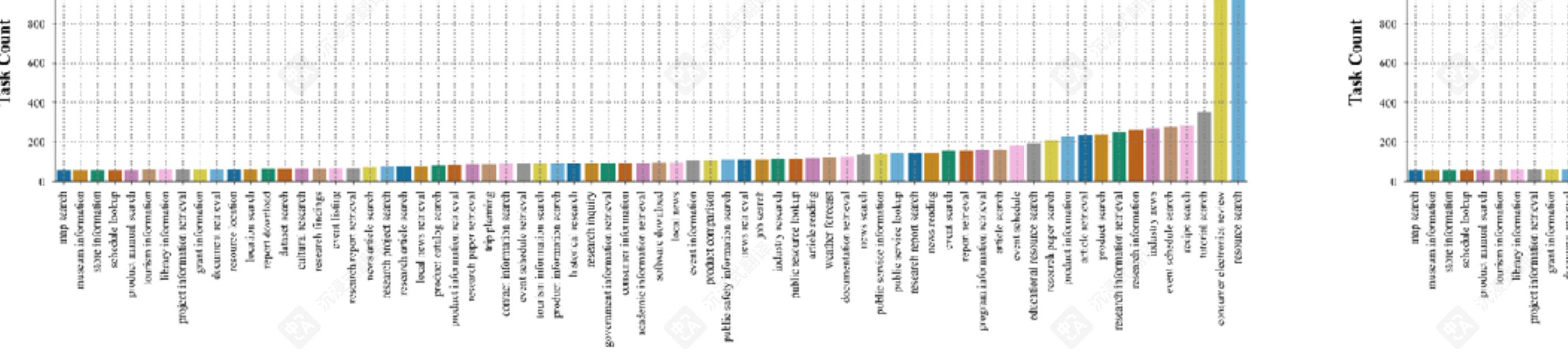
- 给定一个文本形式的网站域名（例如 merseyferries.co.uk）
- 提出一个普通用户可以在一次会话中完成的现实任务。

Llama-3.1-70B-Instruct: 描述“merseyferries.co.uk”网站

Mersey Ferries (merseyferries.co.uk) 是 **利物浦, 英格兰著名渡轮服务的官方网站**。该网站为用户提供了一个一站式平台，以获取与渡轮运营相关的各种服务和信息。以下是用户可以在此网站上进行的一些操作：

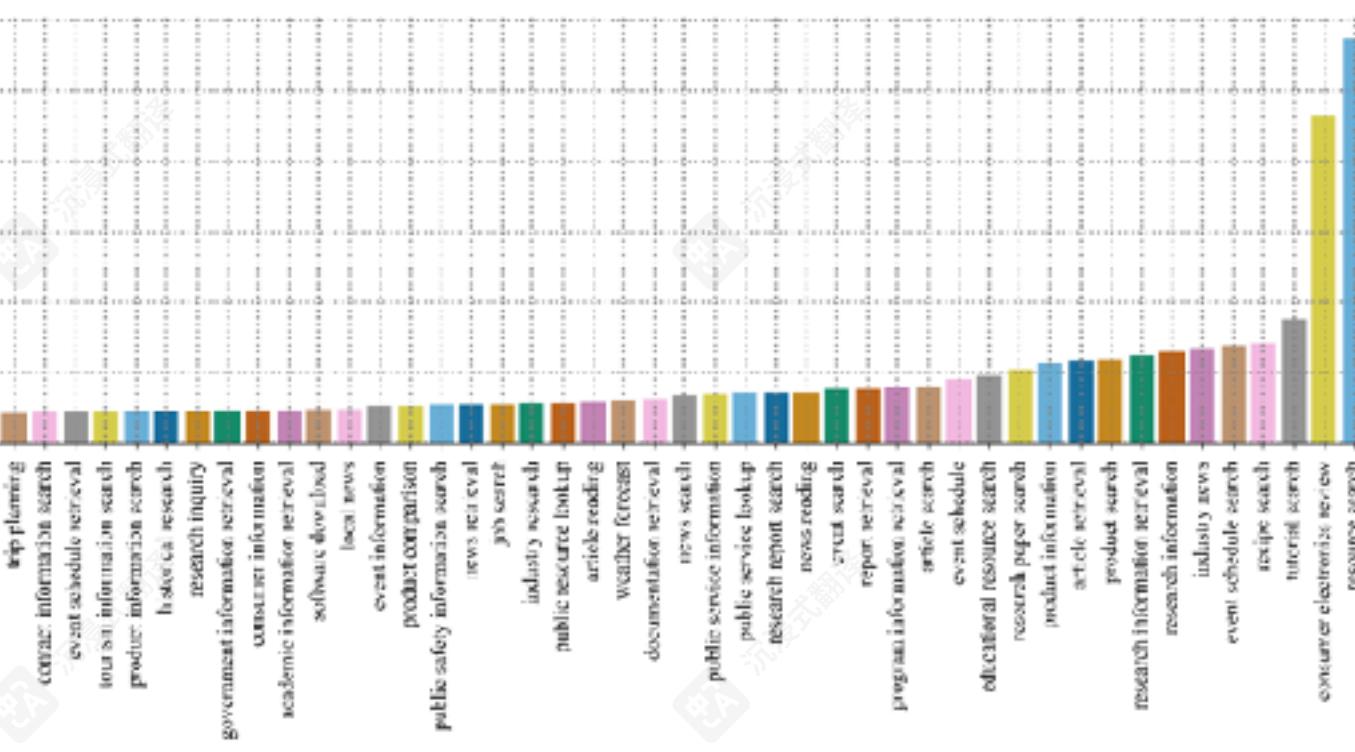
1. **规划旅程**：用户可以通过查看渡轮时刻表来规划他们的旅程，该时刻表包括通勤和游轮服务的出发和到达时间。
2. **预订门票**：用户可以在线预订通勤和游船服务，包括热门的曼彻斯特运河游船和利物浦湾游船。

Use Llama To Generate Agentic Tasks



- Heavy-tailed distribution, most tasks involve information retrieval
- Crucially, **tasks do not modify the state of the internet**

Use Llama 生成代理任务



- 重尾分布，大多数任务涉及信息检索

- 关键地，任务不会修改互联网的状态

Generated Tasks

- Tasks are **diverse**, and many require multiple steps of reasoning (i.e., what makes a font suitable for a children's book)

Web Domain	Generated Task
wordpress.org	Find a free and popular theme for a personal blog.
policies.google.com	Read Google's terms of service for using YouTube.
ec.europa.eu	Retrieve a report on the EU's climate change policy.
vimeo.com	Find a short film on environmental conservation.
fonts.adobe.com	Browse fonts suitable for a children's book.
apps.apple.com	Find the top-rated free productivity app for iPhone.

Generated Tasks

- 任务**是**多样的，并且许多任务需要多步骤的推理（例如，什么使一种字体适合儿童读物）

Web Domain	Generated Task
wordpress.org	Find a free and popular theme for a personal blog.
policies.google.com	Read Google's terms of service for using YouTube.
ec.europa.eu	Retrieve a report on the EU's climate change policy.
vimeo.com	Find a short film on environmental conservation.
fonts.adobe.com	Browse fonts suitable for a children's book.
apps.apple.com	Find the top-rated free productivity app for iPhone.

Generated Tasks

- Llama can **identify facts** that a site is likely to contain, such as the meaning of the Om symbol

Web Domain	Generated Task
ancient-symbols.com	Look up the meaning of the Om symbol in ancient cultures.
petsforhomes.com.au	Find a list of available dogs for adoption in New South Wales.
timorousbeasties.com	View the latest fabric designs by the Timorous Beasties studio.
shop.nikon-image.com	Compare prices of the Nikon D850 and D500 cameras.
blueridgecountry.com	Find a scenic hiking trail in the Blue Ridge Mountains.
awg-fittings.com	Find the dimensions of a 1/2\" NPT fitting.

Generated Tasks

- Llama 可以识别一个网站可能包含的事实，例如 Om 符号的含义

Web Domain	Generated Task
ancient-symbols.com	查询古奥姆符号的含义
petsforhomes.com.au	在 New Sou 找到可领养的狗的列表
timorousbeasties.com	查看 Timorous Beas 最新的织物设计
shop.nikon-image.com	比较 Nikon D850 和 D500 相机的价格
blueridgecountry.com	在蓝岭山脉找到一条风景优美的徒步小径
awg-fittings.com	找到 1/2\" NPT 管件的尺寸。

Generated Tasks

- Llama has broad knowledge of sites, such as for timorousbeasties.com, an independent Scottish design studio (fairly obscure)

Web Domain	Generated Task
ancient-symbols.com	Look up the meaning of the Om symbol in ancient cultures.
petsforhomes.com.au	Find a list of available dogs for adoption in New South Wales.
timorousbeasties.com	View the latest fabric designs by the Timorous Beasties studio.
shop.nikon-image.com	Compare prices of the Nikon D850 and D500 cameras.
blueridgecountry.com	Find a scenic hiking trail in the Blue Ridge Mountains.
awg-fittings.com	Find the dimensions of a 1/2\" NPT fitting.

Generated Tasks

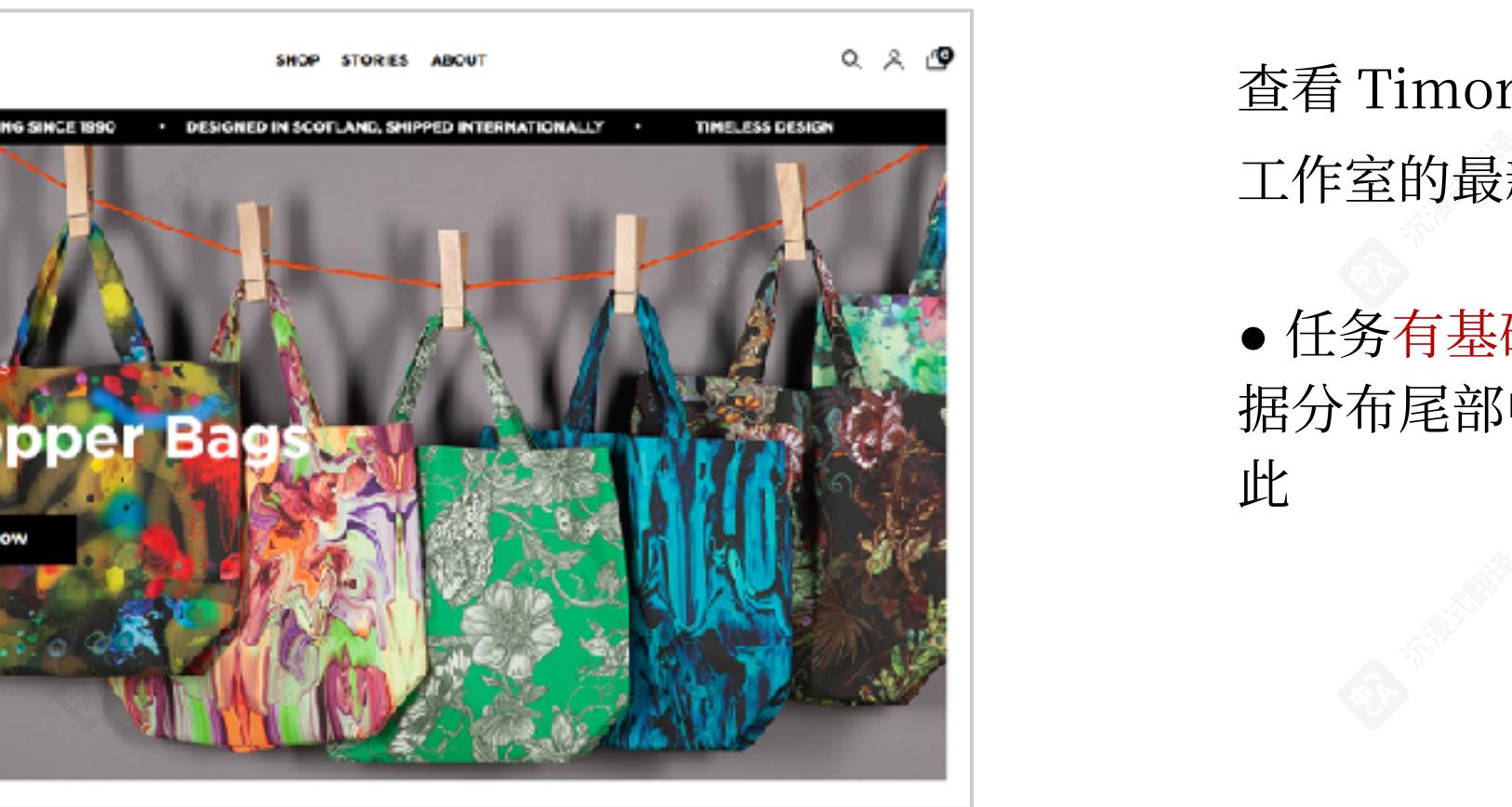
- Llama has broad knowledge of sites, such as for timorousbeasties.com, an independent Scottish design studio (fairly obscure)

Web Domain	Generated Task
ancient-symbols.com	查询古奥姆符号的含义
petsforhomes.com.au	在 New Sou 找到可领养的狗的列表
timorousbeasties.com	查看 Timorous Beas 最新的织物设计
shop.nikon-image.com	比较 Nikon D850 和 D500 相机的价格
blueridgecountry.com	在蓝岭山脉找到一条风景优美的徒步小径
awg-fittings.com	找到 1/2\" NPT 管件的尺寸。

Generated Tasks

View the latest fabric designs by the Timorous Beasties studio

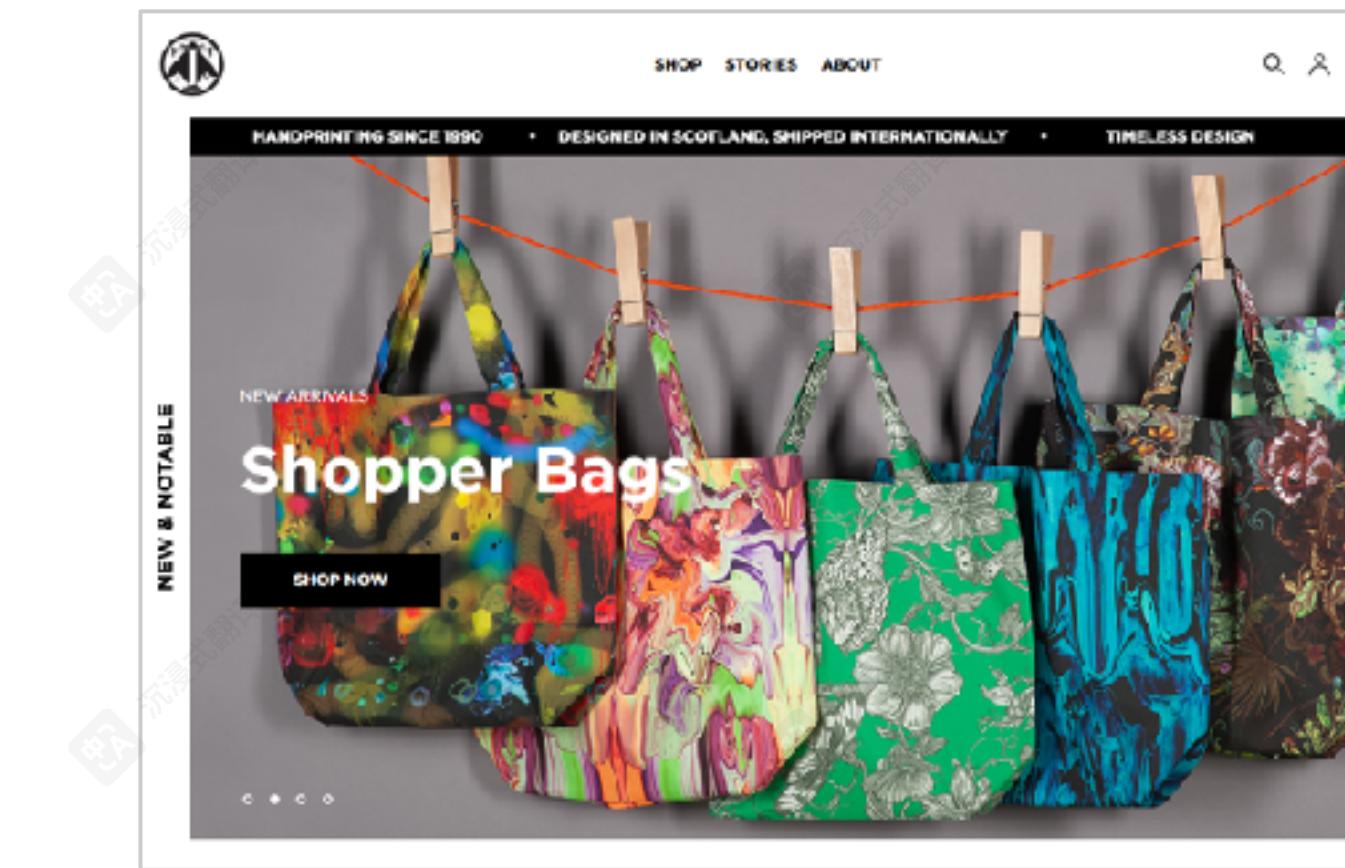
- Tasks are **grounded**, even for sites in the tail of the data distribution



Generated Tasks

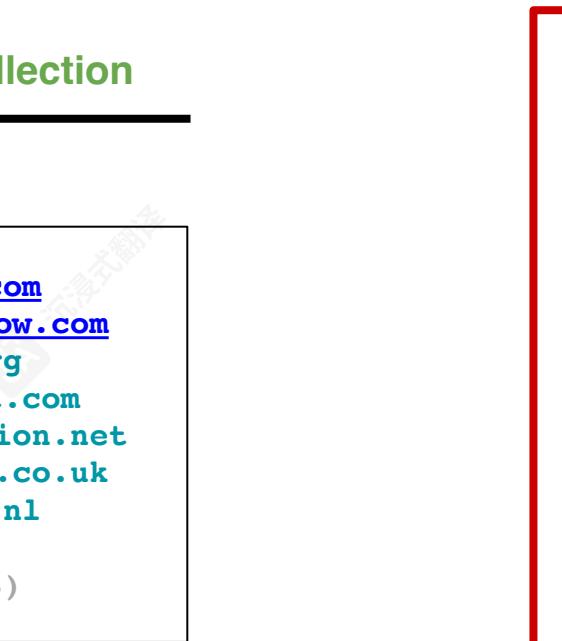
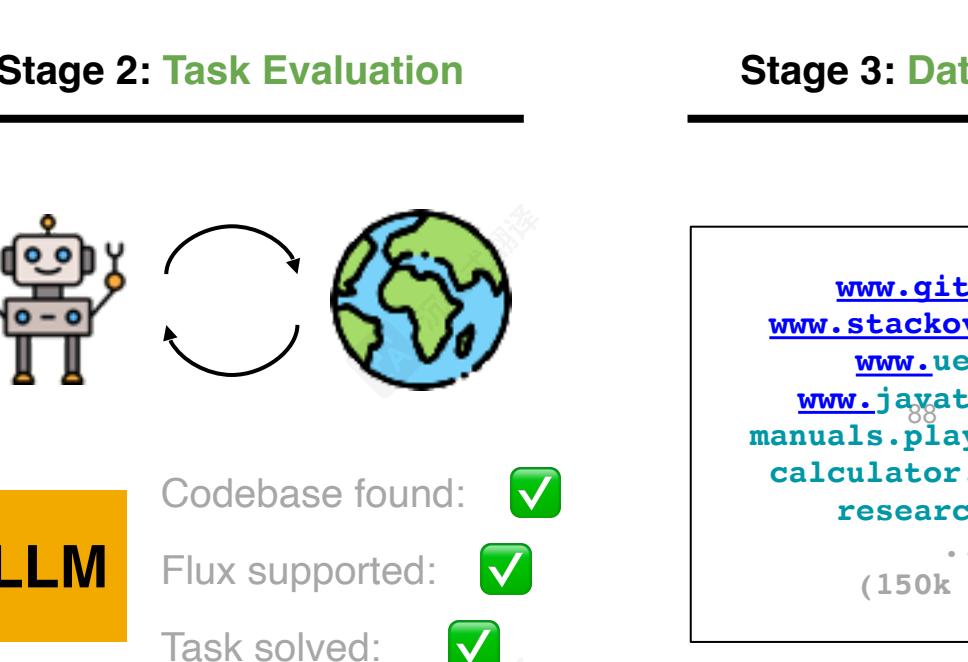
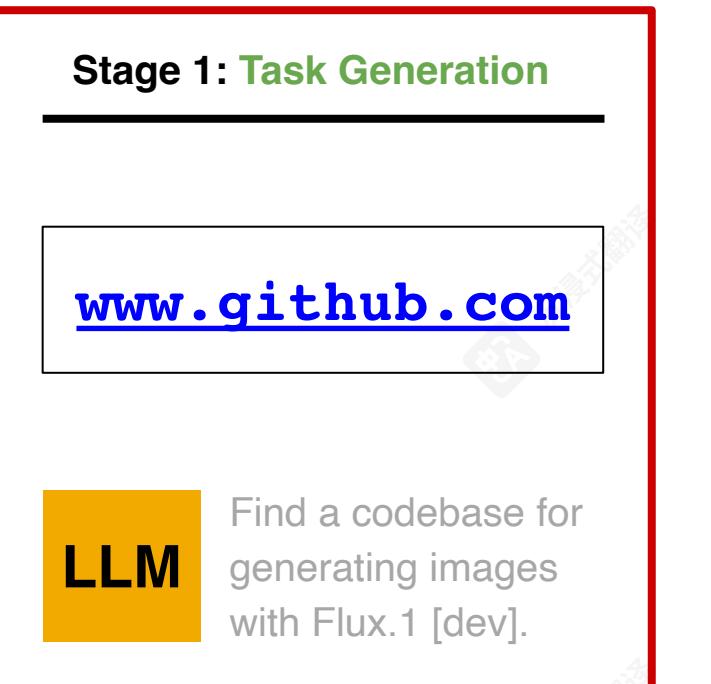
查看 Timorous Beasties 工作室的最新面料设计

- 任务**有基础**, 即使对于数据分布尾部中的站点也是如此



The Data Pipeline

- **Key Idea:** use Llama to **generate and verify** synthetic agentic tasks.



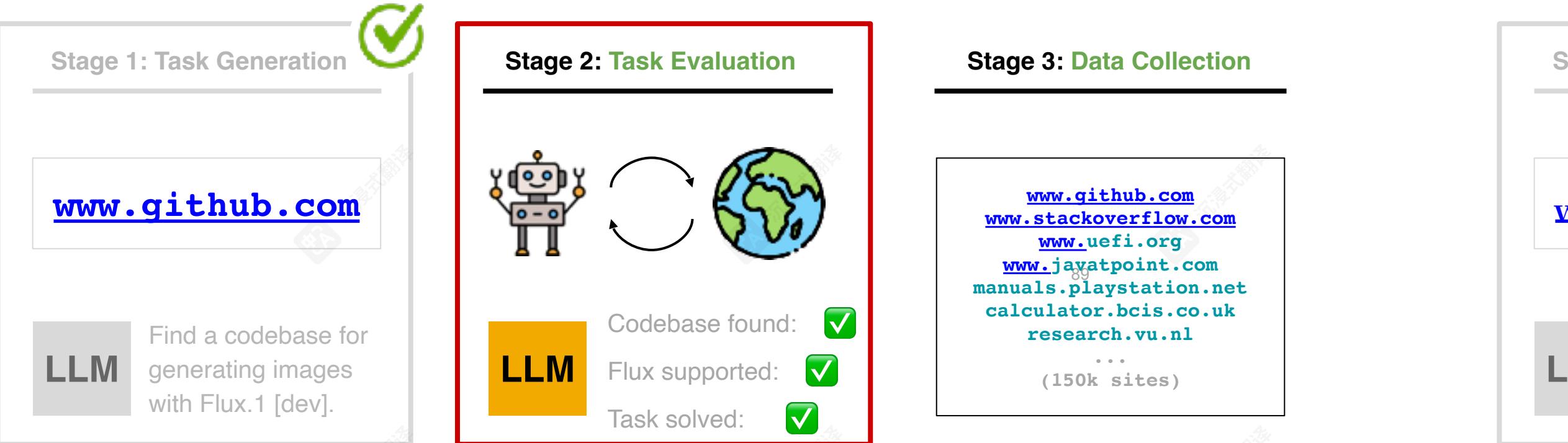
数据管道

- **关键点:** 使用 Llama 来生成和验证合成代理任务。



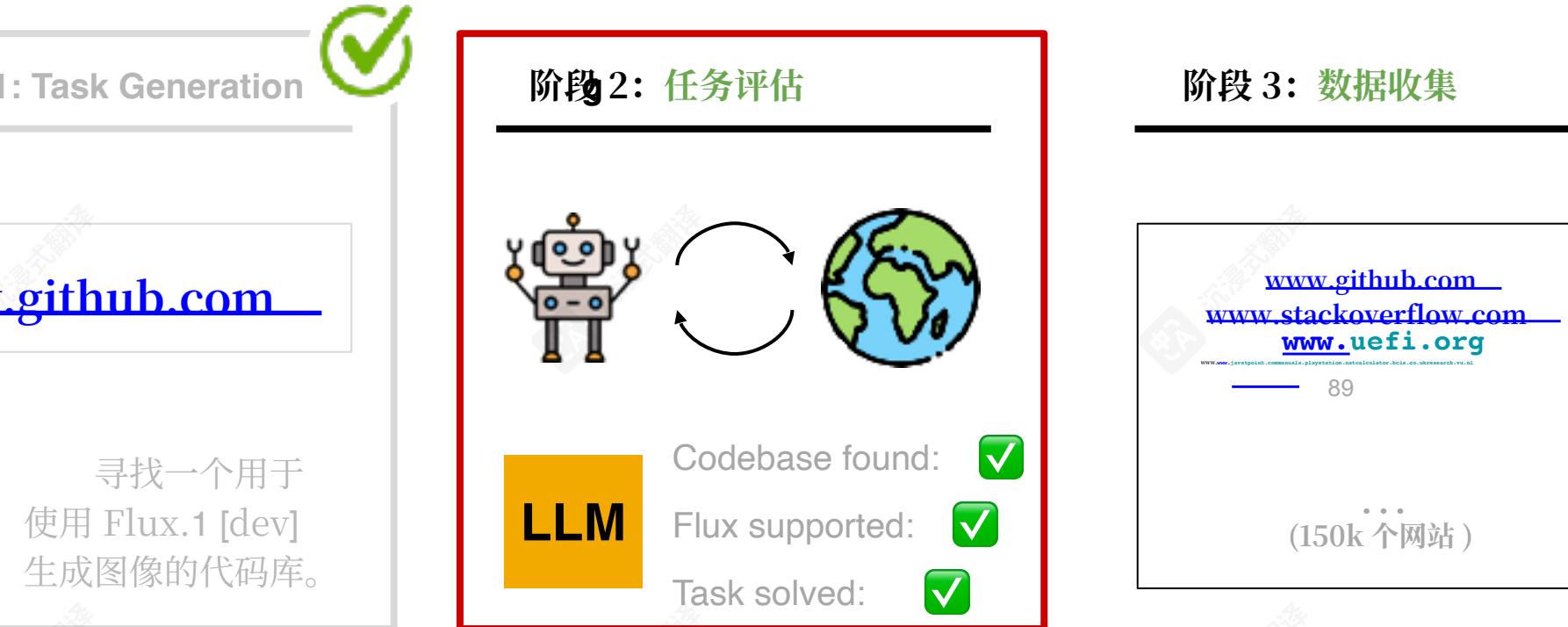
The Data Pipeline

- **Key Idea:** use Llama to **generate and verify** synthetic agentic tasks
- How do we know **when tasks are solved?** Build on Llama models



数据管道

- **核心思想：** 使用 Llama 来 **生成和验证** 合成代理任务
- 我们如何知道 **任务何时被解决？** 基于 Llama 模型



Automatic Task Verification

- How do we know when tasks are solved?
 - Observe a sequence of actions, and the last observation
 - Estimate the **probability the task is a success** at the final step

$$V_{\text{LLM}}(s_T, a_{1:T}) = P(\text{success} | s_T, a_{1:T})$$

自动任务验证

- 我们如何知道任务何时被解决?
 - 观察一系列动作，以及最后的观察
 - Estimate 匹配 **任务成功的概率** 在最终状态

$$V_{\text{LLM}}(s_T, a_{1:T}) = P(\text{success} | s_T, a_{1:T})$$

Automatic Task Verification

- How do we know when tasks are solved?
 - Observe a sequence of actions, and the last observation
 - Estimate the **probability the task is a success** at the final step

$$V_{\text{LLM}}(s_T, a_{1:T}) = P(\text{success} | s_T, a_{1:T})$$

- Prompt Llama to rate its confidence the task is solved on a scale from **0.0 (not possible)** → **1.0 (certain)**
- Of 150,000 tasks, 14.6% are rated as a success with a **confidence of 1.0** by the Llama verifier, roughly 22,000 tasks

自动任务验证

- 我们如何知道任务何时被解决?
 - 观察一系列动作，以及最后的观察
 - Estimate the **probability the task is a success** at the final step

$$V_{\text{LLM}}(s_T, a_{1:T}) = P(\text{success} | s_T, a_{1:T})$$

- 提示 Llama 对其任务是否解决的信心进行评分，从 0.0 (不可能) → 1.0 (确定)
- 在 150,000 个任务中，14.6% 被 Llama 验证器以 1.0 的信心评为成功，大约有 22,000 个任务

Exploring The Data: InvasivePlantAtlas.org

Proposed Task: find invasive plant species native to North America

Agent Response:

To find invasive plant species native to North America, I will first click on the link "All Species" to view the list of all species.

Here is the action in JSON format:

```
```json
{
 "action_key": "click",
 "action_kwargs": {},
 "target_element_id": 12
}
```

```



探索数据：InvasivePlantAtlas.org

建议任务：寻找原产于北美的入侵植物物种

Agent Response:

要寻找原产于北美的入侵植物物种，我将首先点击链接“所有物种”以查看所有物种的列表。

以下是 JSON 格式的操作：

```
```json { "action_key": "点击", "action_kwargs": {}, "target_element_id": 12 }```

```



# Exploring The Data: InvasivePlantAtlas.org

## Agent Response:

To find invasive plant species native to North America, I will first click on the link "Home" to go back to the main page and then click on the link "Sources" to view the list of sources that provide information on invasive plant species native to North America.

However, since the link "Sources" is not on the current page, I will first click on the link "Home" to go back to the main page.

Here is the action in JSON format:

```
```
json
{
    "action_key": "click",
    "action_kwargs": {},
    "target_element_id": 1
}
```

```

Misread  
Page

The screenshot shows the 'Sources' page of the Invasive Plant Atlas. At the top, there's a navigation bar with links for Home, About, ENAVIATE BY Google, and a search bar. Below the navigation is a green header bar with links for Home, Species, Glossary, Herbs/Forbs, Shrubs/Subshrubs, Trees, Video, All Species, Images, Parks, and Sources. The main content area has a title 'Plant Species Reported to be Invasive in Natural Areas' and a subtitle explaining the data. It includes a table titled '1301 Records' listing various invasive plant species with their scientific names, common names, number of countries, and habit. A red box highlights the 'Sources' link in the green header bar.

| Subject Name            | Scientific Name                               | No. of Countries | Habit             |
|-------------------------|-----------------------------------------------|------------------|-------------------|
| catsearpea              | Abrus precatorius L.                          | 54               | Vines             |
| hairy Indian mallow     | Abutilon grandifolium (Willd.) Sweet          | 9                | Shrub or Subshrub |
| velvetleaf              | Abutilon theophrasti Medic                    | 1342             | Forbs/Herbs       |
| earleaf acacia          | Acacia auriculiformis R. Cunningham ex Benth. | 21               | Shrub or Subshrub |
| invadentra wattle       | Acacia kalkora F. Muell.                      | 14               | Hardwood Tree     |
| small Philippine acacia | Acacia melanoxylon R. Br. ex Willd.           | 4                | Hardwood Tree     |
| cyclopia acacia         | Acacia cyclops A. Cunn. ex G. Don             | 3                | Shrub or Subshrub |
| silver wattle           | Acacia dealbata Link                          | 32               | Shrub or Subshrub |
| Sydney golden wattle    | Acacia longifolia (Astr.) Willd.              | 16               | Hardwood Tree     |
| black wattle            | Acacia mearnsii de Wildeman                   | 11               | Hardwood Tree     |
| black acacia            | Acacia melanoxylon R. Br. ex Willd. f.        | 36               | Hardwood Tree     |
| Kerguelan thorn         | Acacia paradoxa DC.                           | 9                | Shrub or Subshrub |
| golden wattle           | Acacia pycnantha Benth.                       | 6                | Hardwood Tree     |
| overkilling wattle      | Acacia retinodes Edmest.                      | 12               | Hardwood Tree     |

## Agent Response:

为查找原产于北美的入侵植物物种，我将首先点击“主页”链接返回主页面，然后点击“来源”链接查看提供有关原产于北美入侵植物物种信息的来源列表。

然而，由于“来源”链接不在当前页面，我将首先点击“主页”链接返回主页面。

以下是 JSON 格式的操作  
````json {"action_key": "点击", "action_kwargs": {}, "target_element_id": 1}````

误读
页面

The screenshot shows the 'Sources' page of the Invasive Plant Atlas. At the top, there's a navigation bar with links for Home, About, ENAVIATE BY Google, and a search bar. Below the navigation is a green header bar with links for Home, Species, Glossary, Herbs/Forbs, Shrubs/Subshrubs, Trees, Video, All Species, Images, Parks, and Sources. The main content area has a title 'Plant Species Reported to be Invasive in Natural Areas' and a subtitle explaining the data. It includes a table titled '1301 Records' listing various invasive plant species with their scientific names, common names, number of countries, and habit. A red box highlights the 'Sources' link in the green header bar.

| Subject Name | Scientific Name | No. of Countries | Habit |
|-------------------------|---|------------------|-------------------|
| catsearpea | Abrus precatorius L. | 54 | Vines |
| hairy Indian mallow | Abutilon grandifolium (Willd.) Sweet | 9 | Shrub or Subshrub |
| velvetleaf | Abutilon theophrasti Medic | 1342 | Forbs/Herbs |
| earleaf acacia | Acacia auriculiformis R. Cunningham ex Benth. | 21 | Shrub or Subshrub |
| invadentra wattle | Acacia kalkora F. Muell. | 14 | Hardwood Tree |
| small Philippine acacia | Acacia melanoxylon R. Br. ex Willd. | 4 | Hardwood Tree |
| cyclopia acacia | Acacia cyclops A. Cunn. ex G. Don | 3 | Shrub or Subshrub |
| silver wattle | Acacia dealbata Link | 32 | Shrub or Subshrub |
| Sydney golden wattle | Acacia longifolia (Astr.) Willd. | 16 | Hardwood Tree |
| black wattle | Acacia mearnsii de Wildeman | 11 | Hardwood Tree |
| black acacia | Acacia melanoxylon R. Br. ex Willd. f. | 36 | Hardwood Tree |
| Kerguelan thorn | Acacia paradoxa DC. | 9 | Shrub or Subshrub |
| golden wattle | Acacia pycnantha Benth. | 6 | Hardwood Tree |
| overkilling wattle | Acacia retinodes Edmest. | 12 | Hardwood Tree |

Exploring The Data: InvasivePlantAtlas.org

Agent Response:

To find invasive plant species native to North America, I will first click on the link "Sources" to view the list of sources that provide information on invasive plant species native to North America.

Here is the action in JSON format:

```
```json
{
 "action_key": "click",
 "action_kwargs": {},
 "target_element_id": 15
}
```

```



The screenshot shows the header of the Invasive Plant Atlas of the United States website. The navigation bar includes links for Home, Aquatics, grasses, Herbs/Folks, Shrubs/Subshrubs, trees, vines, All Species, Images, Rank, and Sources. The 'Sources' link is highlighted with a red box. Below the header, there's a large green banner with the text 'Invasive Plant Atlas of the United States'. To the right of the banner, there's a sidebar titled 'EDDMaps' with a sub-section for 'invasive species news' containing several news items.

Agent Response:

为查找原产于北美的入侵植物物种，我将首先点击“来源”链接，查看提供有关原产于北美入侵植物物种信息的来源列表。

以下是 JSON 格式的操作：

```
.. .json { "action_key": "点击",
action_kwargs": {}, "target_
element_id": 15 } ..
```



The screenshot shows the same website interface as the previous one, but the 'Sources' link in the header is now highlighted with a red box. The rest of the page content, including the banner and sidebar, remains the same.

Exploring The Data: InvasivePlantAtlas.org

Agent Response:

To find invasive plant species native to North America, I will first click on the link "FONA" under the "Native Range and Taxonomic References" section, which stands for "Flora of North America". This will take me to the Flora of North America website, where I can search for invasive plant species native to North America.

Here is the action in JSON format:

```
```json
{
 "action_key": "click",
 "action_kwargs": {},
 "target_element_id": 25
}
```

```



The screenshot shows the 'Native Range and Taxonomic References' section of the website. A red box highlights the 'FONA' link, which is described in the JSON action as element 25.

探索数据：InvasivePlantAtlas.org

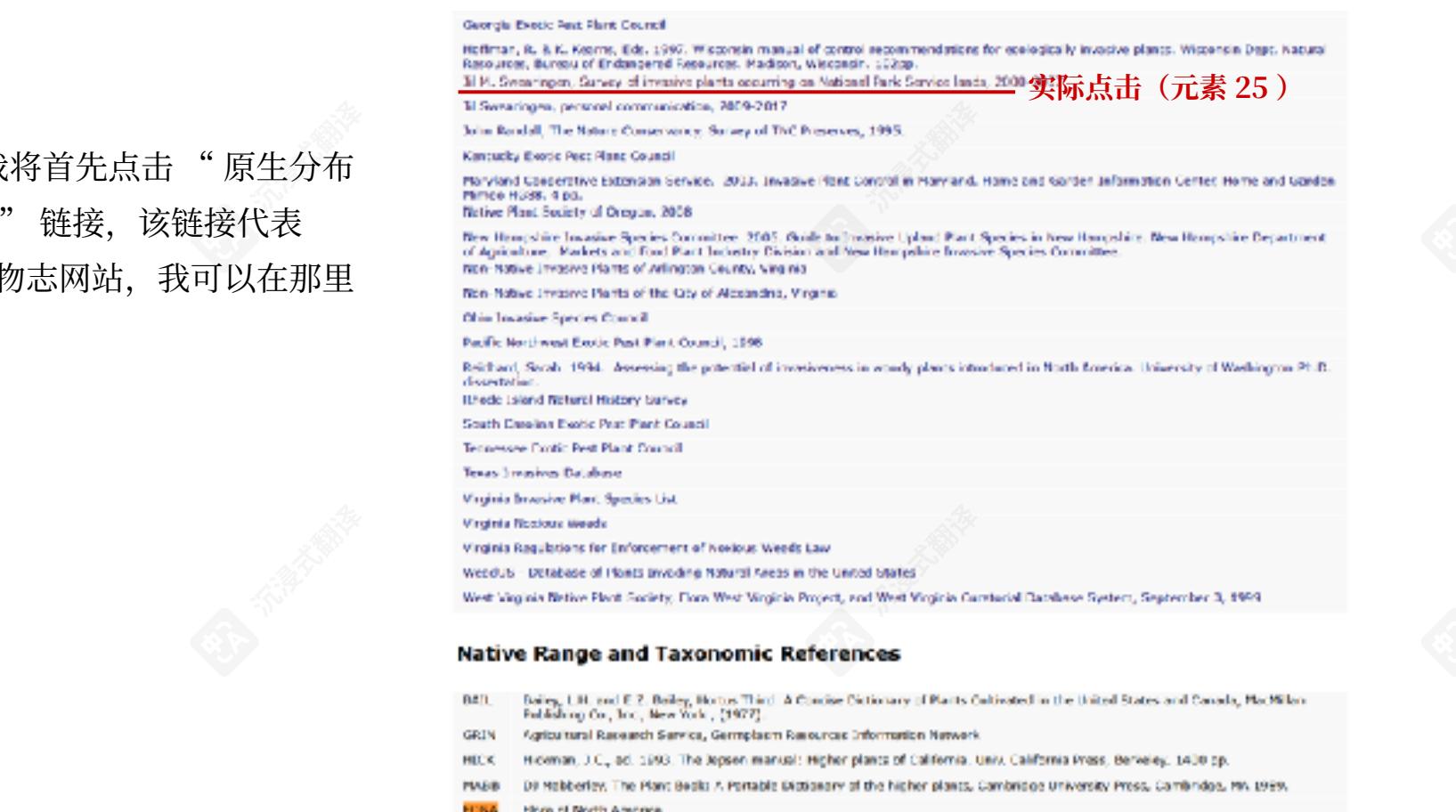
Agent Response:

为查找原产于北美的入侵植物物种，我将首先点击“原生分布和分类参考文献”部分下的“FONA”链接，该链接代表“北美植物志”。这将带我到北美植物志网站，我可以在那里搜索原产于北美的入侵植物物种。

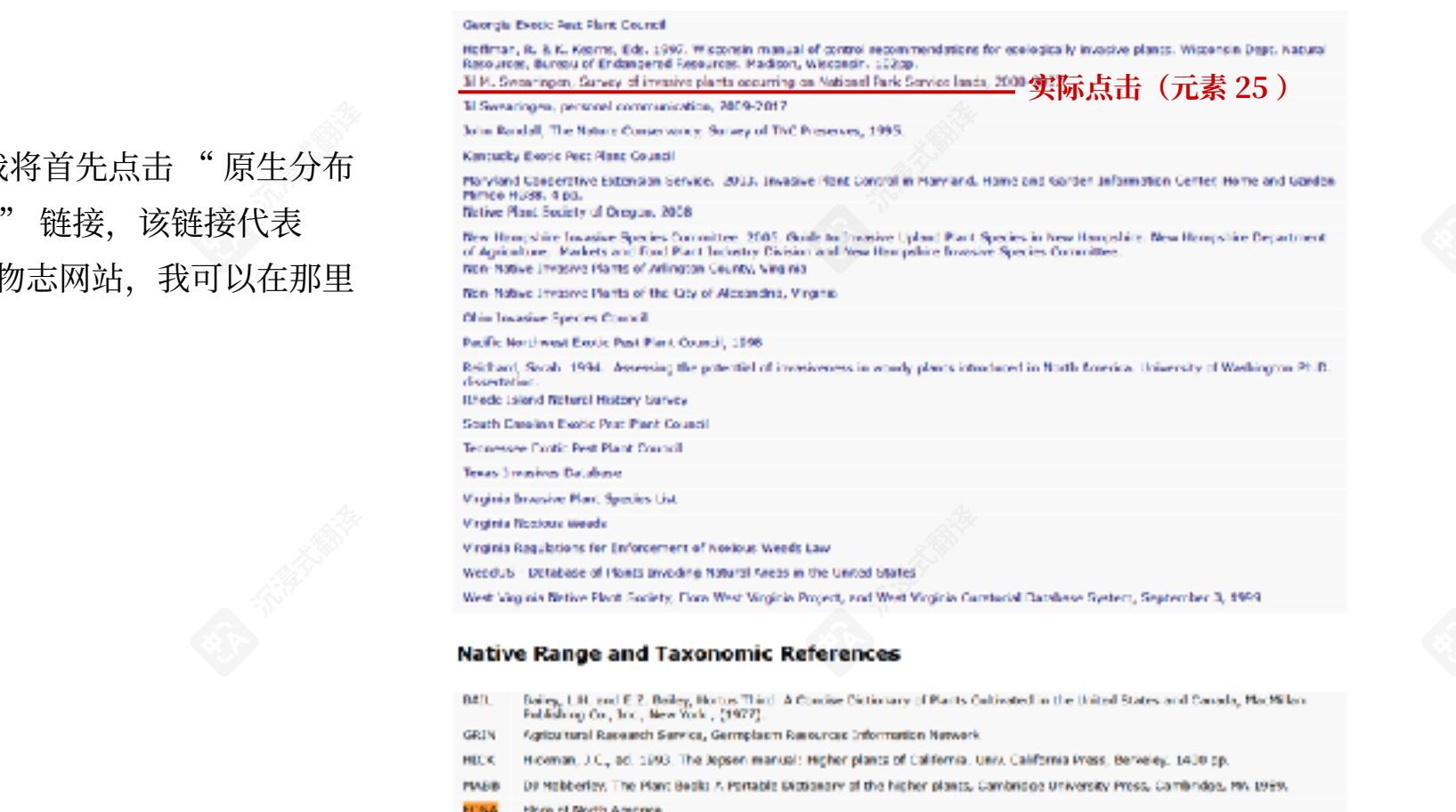
以下是 JSON 格式的操作：

```
```json { "action_key": "点击", "action_kwargs": {}, "target_element_id": 25 }```

```



The screenshot shows the 'Native Range and Taxonomic References' section of the website. A red box highlights the 'FONA' link, which is described in the JSON action as element 25.



The screenshot shows the 'Native Range and Taxonomic References' section of the website. A red box highlights the 'FONA' link, which is described in the JSON action as element 25.

# Exploring The Data: InvasivePlantAtlas.org

## Agent Response:

To find invasive plant species native to North America, I will first click on the link "Sort on "U.S. Nativity"" to sort the list by the nativity of the plant species in the United States.

Here is the action in JSON format:

```
```json
{
    "action_key": "click",
    "action_kwargs": {},
    "target_element_id": 20
}
```

```



| Subject Name          | Scientific Name                                     | Family | U.S. Nativity |
|-----------------------|-----------------------------------------------------|--------|---------------|
| hairy jimson weed     | Atropa belladonna (W. B. S.) Sweet                  | 茄科     | Exotic        |
| black nettle          | Abutilon theophrasti Medicus                        | 锦葵科    | Exotic        |
| black locust          | Acacia melanoxylon R. Br. ex Willd.                 | 豆科     | Exotic        |
| Japanese maple        | Acer palmatum Thunb.                                | 槭树科    | Exotic        |
| Norway maple          | Acer platanoides L.                                 | 槭树科    | Exotic        |
| rough maidenhair fern | Adiantum hispidulum L.                              | 水龙骨科   | Exotic        |
| debt maidenhair       | Adiantum nidus-sacrum H. Pol.                       | 水龙骨科   | Exotic        |
| joined goutgrass      | Agropyron cylindricum Host                          | 禾本科    | Exotic        |
| teal                  | Agave stricta Perrine                               | 天门冬科   | Exotic        |
| creeping zinnia       | Ageratina adenophora (Strong.) King & H.B. Robinson | 菊科     | Exotic        |
| creeping crotonweed   | Ageratum riparia (Regel) King & H.E. Robins.        | 菊科     | Native        |
| tropical elephantfoot | Agenetum comosum L.                                 | 禾本科    | Native        |
| colonial buttergrass  | Agrostis capillaris L.                              | 禾本科    | Exotic        |
| redtop                | Agrostis gigantea Roth                              | 禾本科    | Exotic        |
| creeping bentgrass    | Agrostis stolonifera L.                             | 禾本科    | Exotic        |

## Agent Response:

为查找原产于北美的入侵植物物种，我将首先点击链接“按“美国原产”排序”以按美国植物物种的原产地对列表进行排序。

以下是 JSON 格式的操作：

```
.. json { "action_key": "点击",
 "action_kwargs": {}, "target_
 element_id": 20 } ..
```



| Subject Name          | Scientific Name                                     | Family | U.S. Nativity |
|-----------------------|-----------------------------------------------------|--------|---------------|
| hairy jimson weed     | Atropa belladonna (W. B. S.) Sweet                  | 茄科     | Exotic        |
| black nettle          | Abutilon theophrasti Medicus                        | 锦葵科    | Exotic        |
| black locust          | Acacia melanoxylon R. Br. ex Willd.                 | 豆科     | Exotic        |
| Japanese maple        | Acer palmatum Thunb.                                | 槭树科    | Exotic        |
| Norway maple          | Acer platanoides L.                                 | 槭树科    | Exotic        |
| rough maidenhair fern | Adiantum hispidulum L.                              | 水龙骨科   | Exotic        |
| debt maidenhair       | Adiantum nidus-sacrum H. Pol.                       | 水龙骨科   | Exotic        |
| joined goutgrass      | Agropyron cylindricum Host                          | 禾本科    | Exotic        |
| teal                  | Agave stricta Perrine                               | 天门冬科   | Exotic        |
| creeping zinnia       | Ageratina adenophora (Strong.) King & H.B. Robinson | 菊科     | Exotic        |
| creeping crotonweed   | Ageratum riparia (Regel) King & H.E. Robins.        | 菊科     | Native        |
| tropical elephantfoot | Agenetum comosum L.                                 | 禾本科    | Native        |
| colonial buttergrass  | Agrostis capillaris L.                              | 禾本科    | Exotic        |
| redtop                | Agrostis gigantea Roth                              | 禾本科    | Exotic        |
| creeping bentgrass    | Agrostis stolonifera L.                             | 禾本科    | Exotic        |

# Exploring The Data: InvasivePlantAtlas.org

## Agent Response:

To find invasive plant species native to North America, I will first scroll down the page to view the list of invasive plant species that are native to North America.

However, since the list is already sorted by "U.S. Nativity" in ascending order, I can see that the first 65 plant species are native to North America.

Here is the action in JSON format:

```
 ``json
 "action_key": "stop",
 "action_kwargs": {
 "answer": "The first 165 p
list are invasive plant species na
merica."
 },
 "target_element_id": null
```



# InvasivePlantA

植物物种，我将

美国原产性”

停止",

5 种植物是原

null };



# Exploring The Data: InvasivePlantAtlas.org

**Proposed Task:** find invasive plant species native to North America

- Despite reasoning failures, the agent **self-corrected**, and found the target information
- The Llama verifier evaluates this trajectory as **successful** with confidence = 1.0



The screenshot shows the homepage of the Invasive Plant Atlas of the United States. The main title is "Invasive Plant Atlas of the United States". Below it is a navigation bar with links: Home, About, Enhanced in Google, Home, Aquatics, Grasses, Herbs/Ferbs, Shrubs/Solidagines, Trees, Vines, All species, Images, Plans, Sources. A green banner below the navigation bar says "Plant Species Reported to be Invasive in Natural Areas List". Below the banner is a table titled "SPP Species" with columns: Subject Name, Scientific Name, Family, U.S. Native. The table lists various plant species, all of which are marked as "Native".

| Subject Name              | Scientific Name                                        | Family         | U.S. Native |
|---------------------------|--------------------------------------------------------|----------------|-------------|
| Water speedwell           | Veronica persicae (L.)                                 | Plantaginaceae | Native      |
| cronwheat                 | Verbesina encelioides (Cav.) Benth. & Hook. f. ex Gray | Asteraceae     | Native      |
| Snowberry-like bush       | Triadica sebifera (L.)                                 | Tiliaceae      | Native      |
| M. ALNUSMUS               | Betula populifolia (Muhl.) Kuntze                      | Betulaceae     | Native      |
| Yucca whipplei            | Yucca whipplei (T. & G.) Mill.                         | Agavaceae      | Native      |
| STUDIES                   | Geocryptis insulae (L.) R. Br.                         | Pandanaceae    | Native      |
| cutleaf nightshade        | Solanum elaeagnifolium                                 | Solanaceae     | Native      |
| American black nightshade | Solanum americanum P. Mill.                            | Solanaceae     | Native      |
| knobcone, fossil          | Sequoia sempervirens (Pursh) Reichenb.                 | Fagaceae       | Native      |
| highbush blueberry        | Rubus leucodermis Link                                 | Rosaceae       | Native      |
| anglestem blackberry      | Rubus cuneifolius Ell.                                 | Rosaceae       | Native      |
| Douglas fir               | Pseudotsuga menziesii (Mirbel) Franco                  | Pinaceae       | Native      |
| balsam poplar             | Populus tremuloides Michx.                             | Salicaceae     | Native      |
| curl for all              | Pithecellobium dulce (Jacq.) Bonpl.                    | Astereaceae    | Native      |
| Hudson pine               | Pinus strobus L.                                       | Pinaceae       | Native      |
| eastern white pine        | Pinus strobus L.                                       | Pinaceae       | Native      |
| Meaterry pine             | Pinus contorta D. Don                                  | Pinaceae       | Native      |
| ponderosa pine            | Pinus ponderosa P. C. Lawson                           | Pinaceae       | Native      |
| yellow pine               | Pinus Jeffreyi (Carr.) Saff.                           | Pinaceae       | Native      |

# 探索数据：InvasivePlantAtlas.org

**建议任务：**寻找原产于北美的  
入侵植物物种

- 尽管推理失败，代理自我纠正，并找到了目标信息

- Llama 验证器将此轨迹评估为成  
功，置信度 = 1.0



The screenshot shows the homepage of the Invasive Plant Atlas of the United States. The main title is "Invasive Plant Atlas of the United States". Below it is a navigation bar with links: Home, About, Enhanced in Google, Home, Aquatics, Grasses, Herbs/Ferbs, Shrubs/Solidagines, Trees, Vines, All species, Images, Plans, Sources. A green banner below the navigation bar says "Plant Species Reported to be Invasive in Natural Areas List". Below the banner is a table titled "SPP Species" with columns: Subject Name, Scientific Name, Family, U.S. Native. The table lists various plant species, all of which are marked as "Native".

| Subject Name              | Scientific Name                                        | Family         | U.S. Native |
|---------------------------|--------------------------------------------------------|----------------|-------------|
| Water speedwell           | Veronica persicae (L.)                                 | Plantaginaceae | Native      |
| cronwheat                 | Verbesina encelioides (Cav.) Benth. & Hook. f. ex Gray | Asteraceae     | Native      |
| Snowberry-like bush       | Triadica sebifera (L.)                                 | Tiliaceae      | Native      |
| M. ALNUSMUS               | Betula populifolia (Muhl.) Kuntze                      | Betulaceae     | Native      |
| Yucca whipplei            | Yucca whipplei (T. & G.) Mill.                         | Agavaceae      | Native      |
| STUDIES                   | Geocryptis insulae (L.) R. Br.                         | Pandanaceae    | Native      |
| cutleaf nightshade        | Solanum elaeagnifolium                                 | Solanaceae     | Native      |
| American black nightshade | Solanum americanum P. Mill.                            | Solanaceae     | Native      |
| knobcone, fossil          | Sequoia sempervirens (Pursh) Reichenb.                 | Fagaceae       | Native      |
| highbush blueberry        | Rubus leucodermis Link                                 | Rosaceae       | Native      |
| anglestem blackberry      | Rubus cuneifolius Ell.                                 | Rosaceae       | Native      |
| Douglas fir               | Pseudotsuga menziesii (Mirbel) Franco                  | Pinaceae       | Native      |
| balsam poplar             | Populus tremuloides Michx.                             | Salicaceae     | Native      |
| curl for all              | Pithecellobium dulce (Jacq.) Bonpl.                    | Astereaceae    | Native      |
| Hudson pine               | Pinus strobus L.                                       | Pinaceae       | Native      |
| eastern white pine        | Pinus strobus L.                                       | Pinaceae       | Native      |
| Meaterry pine             | Pinus contorta D. Don                                  | Pinaceae       | Native      |
| ponderosa pine            | Pinus ponderosa P. C. Lawson                           | Pinaceae       | Native      |
| yellow pine               | Pinus Jeffreyi (Carr.) Saff.                           | Pinaceae       | Native      |

Find the opening hours  
for La Sagrada Familia.

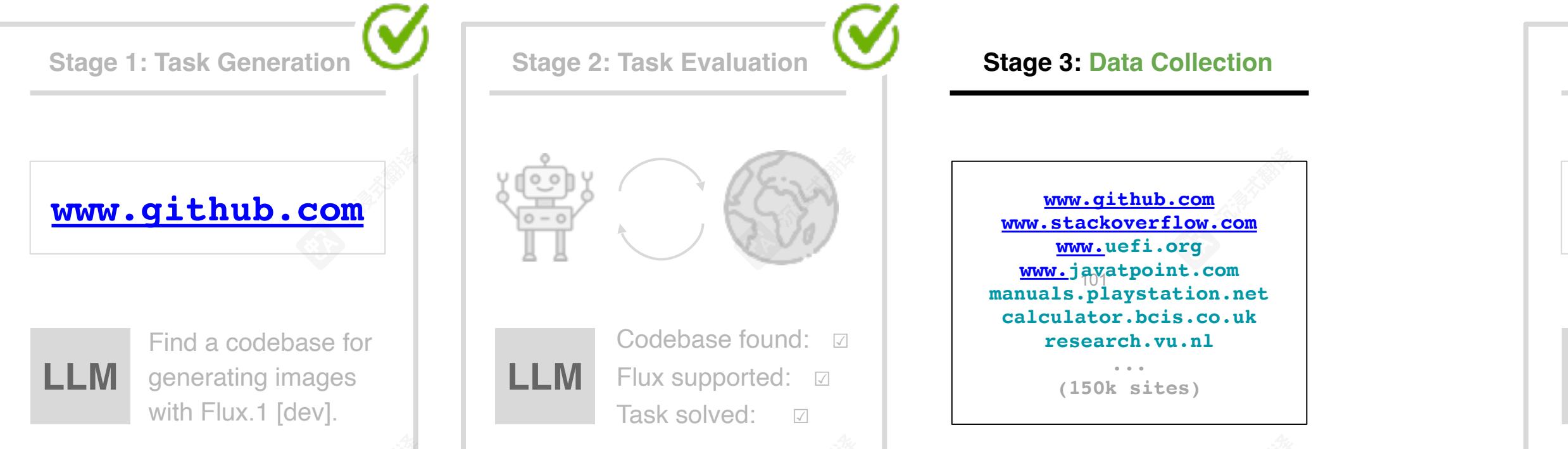
Find the opening hours  
for La Sagrada Familia.

Find information on the European Union's climate action policies.

Find information on the European Union's climate action policies.

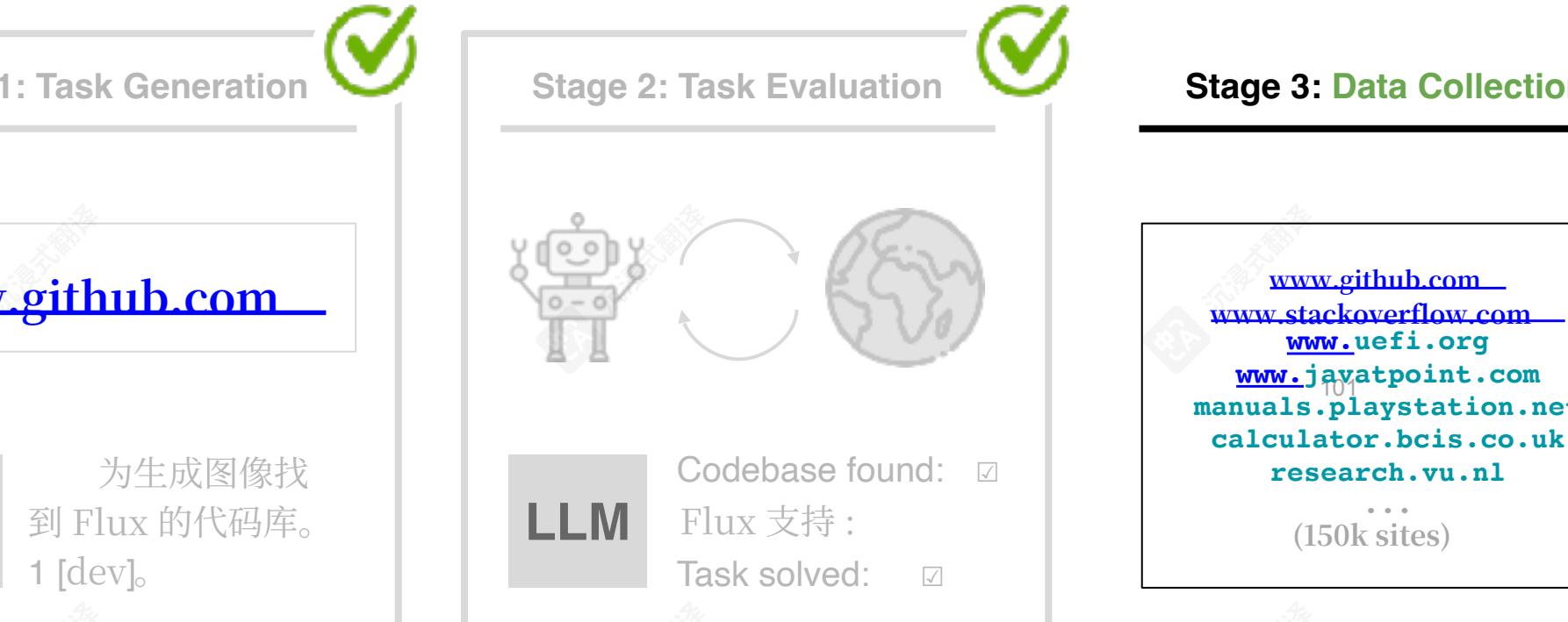
# The Data Pipeline

- We've covered **generation and verification** of synthetic agentic tasks



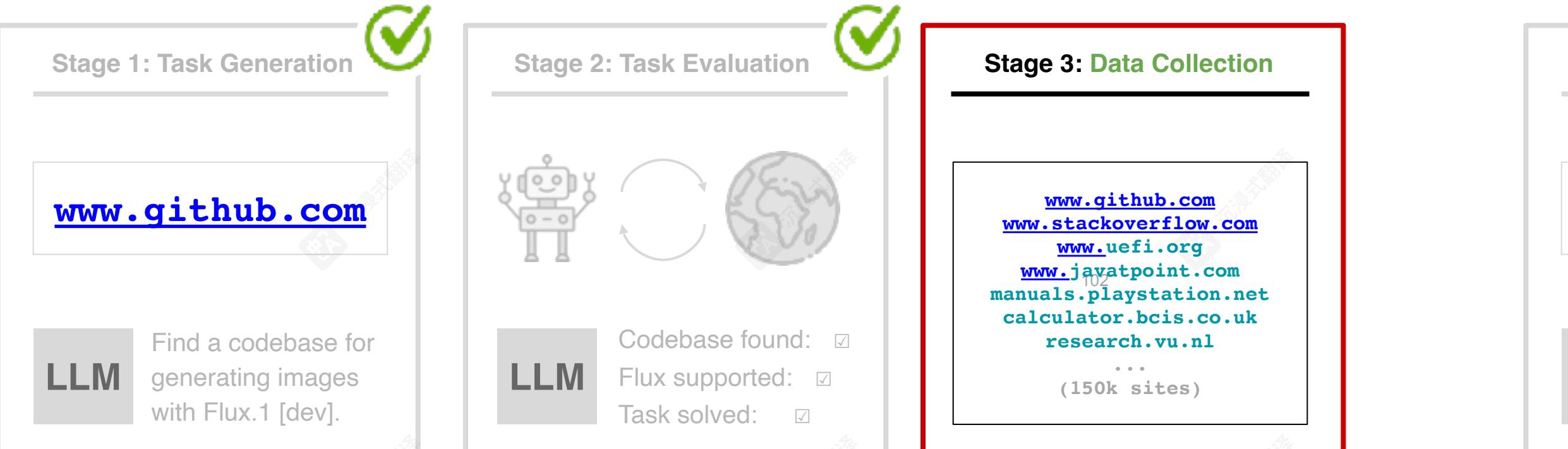
# 数据管道

- 我们已经涵盖了 生成和验证 合成代理任务的 {v1}



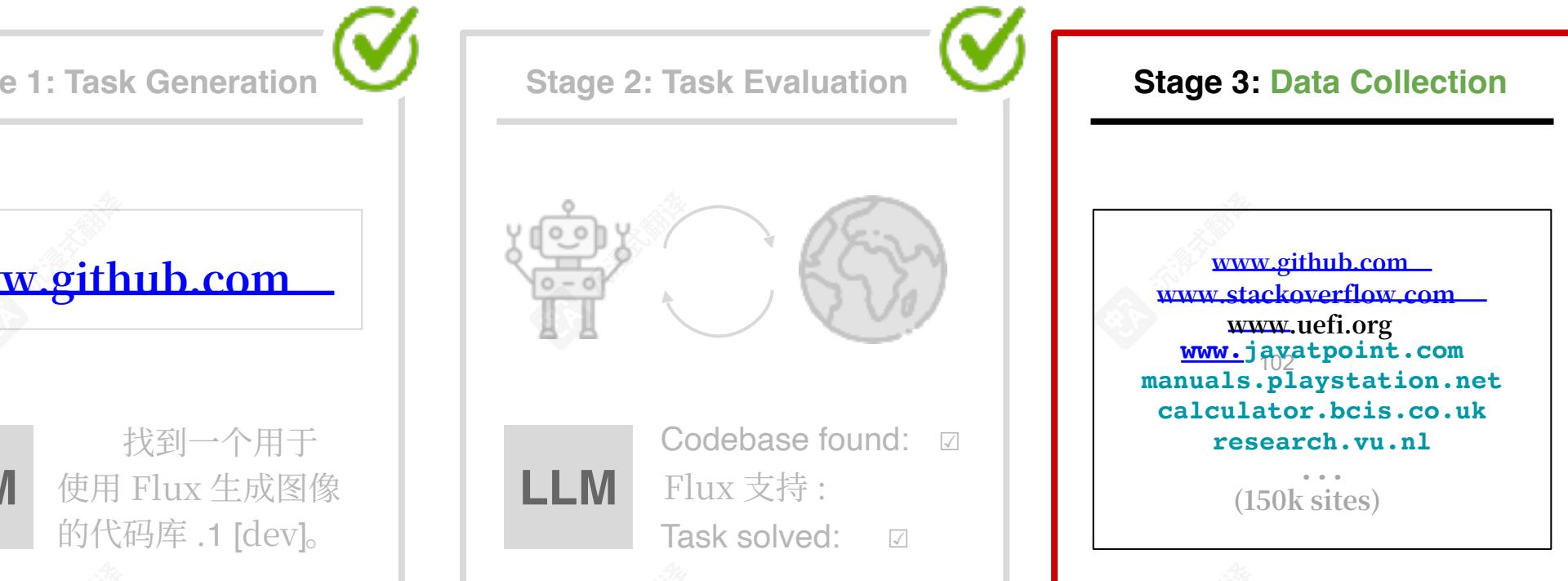
# The Data Pipeline

- We've covered **generation and verification** of synthetic agentic tasks
- Now we can **scale up** data collection



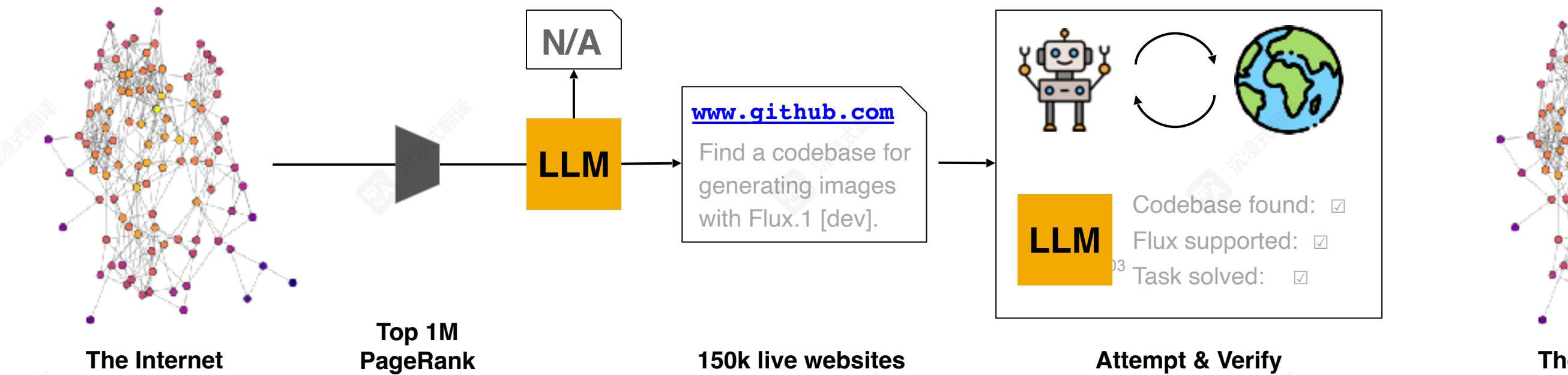
# 数据管道

- 我们已经涵盖了 生成和验证合成代理任务的方面
- 现在我们可以 扩大数据收集



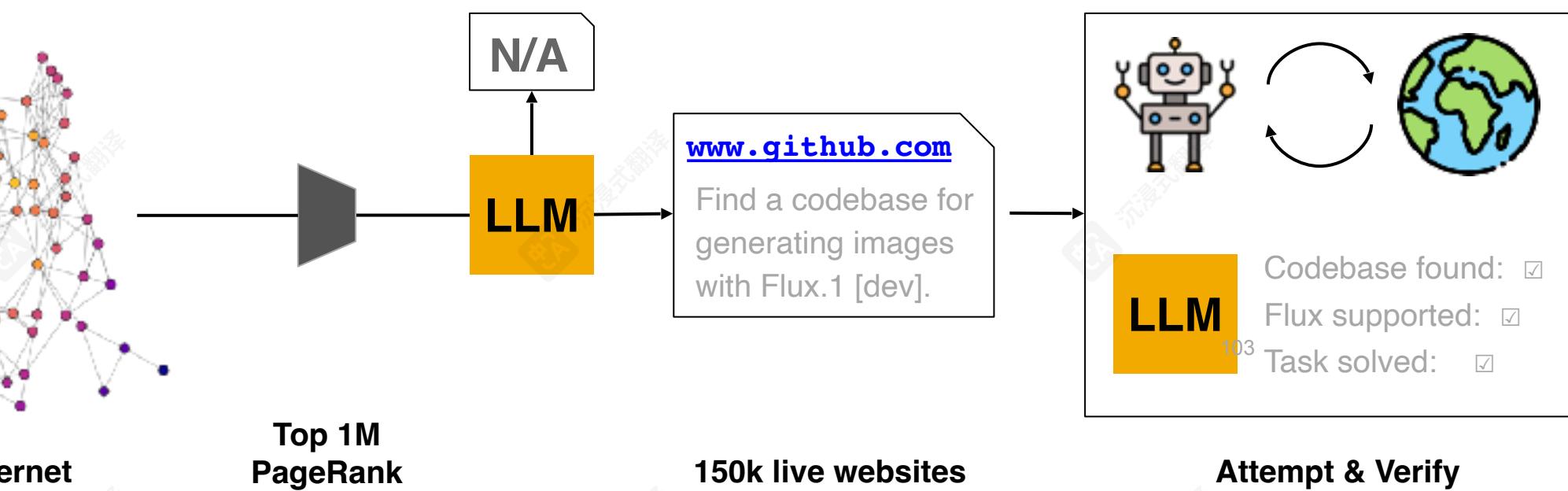
# Scaling Up To 150k Live Websites

- We can use the **Common Crawl PageRank** to find important sites
  - 97% **accuracy** in detecting and filtering harmful content
  - 89% **success rate** in generating feasible tasks
  - 82% **accuracy** in judging successful task completions

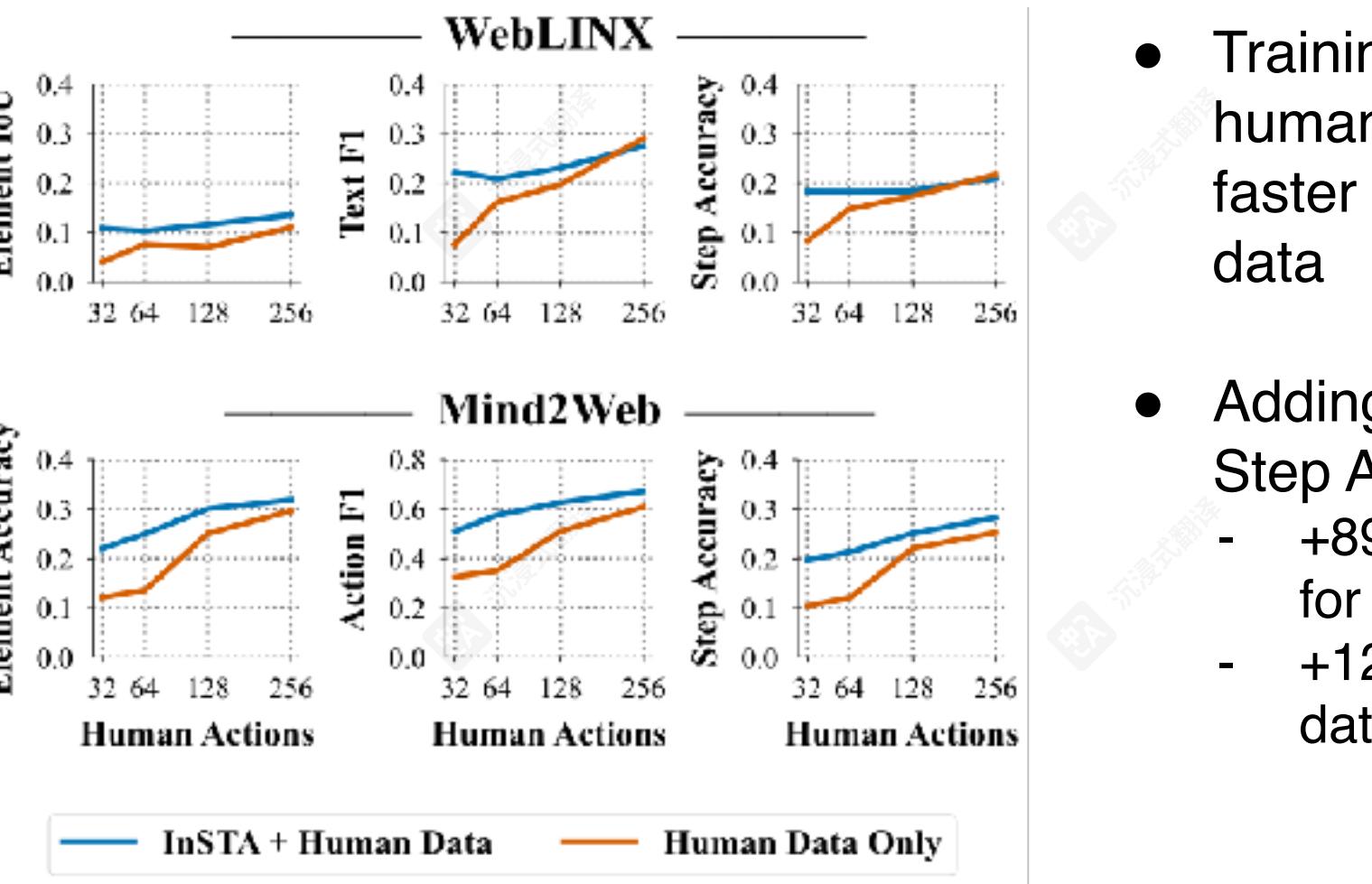


# 扩展至 15 万活跃网站

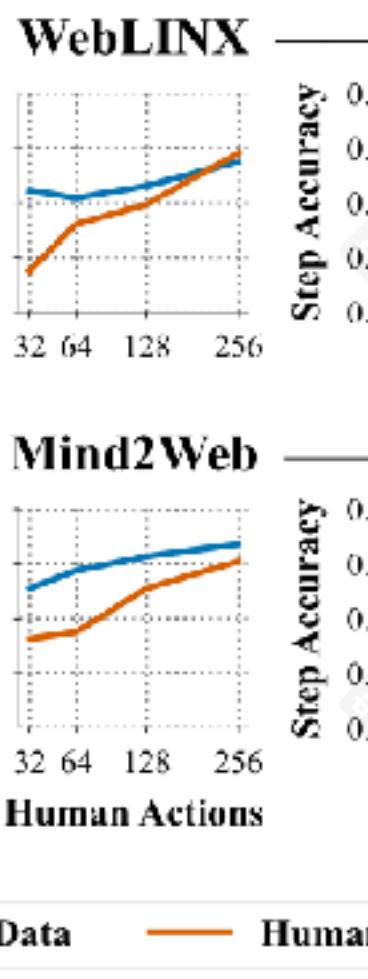
- 我们可以使用 **Common CrawlPageRank** 来查找重要网站
  - 97% 准确率在检测和过滤有害内容 89% 成功率在生成
  - 可行任务 82% 准确率在判断任务成功完成



# Results: Improving Efficiency

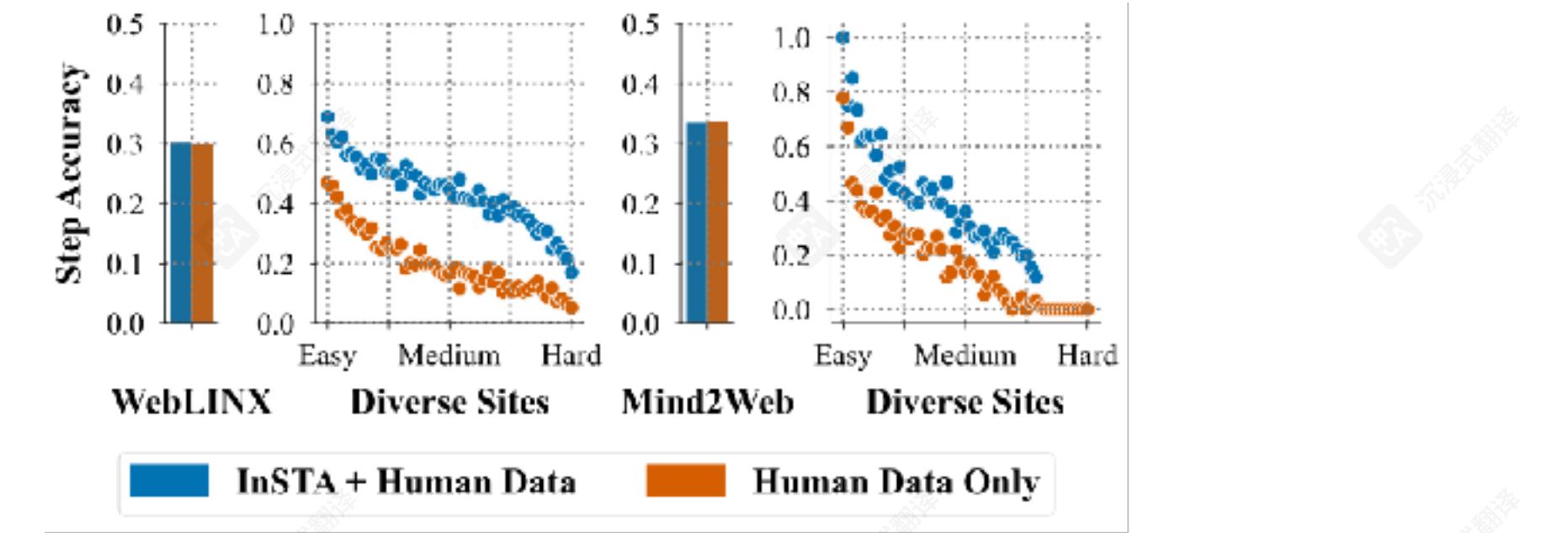


- Training on synthetic and human demonstrations scale faster than training on human data
- Adding synthetic data improves Step Accuracy by
  - +89.5% relative to human data for Mind2Web
  - +122.1% relative to human data for WebLINX



- 在合成数据和人演示规模上训练比在人类数据上训练更快
- 添加合成数据提高了 StepAccuracy，提高了 +89.5% 相对于人类数据（ Mind2Web ） - +122.1% 相对于 人类数据（ WebLINX ）

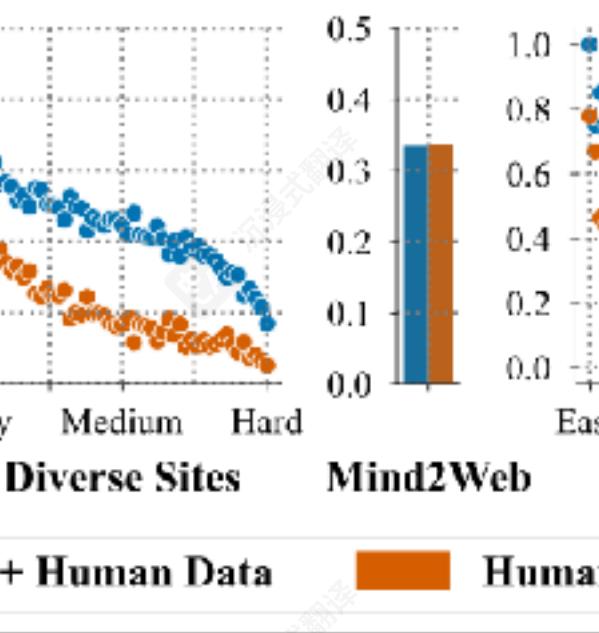
# Results: Improving Generalization



仅使用人类演示进行训练难以泛化

- Training with only human demonstrations struggle with generalization
- Adding synthetic data improves generalization by
  - +149.0% for WebLINX
  - +156.3% for Mind2Web

# 结果：提高泛化能力



仅使用人类演示进行训练难以泛化

- 仅使用人类演示进行训练难以泛化
- 添加合成数据通过 +149.0% 提高了 WebLINX 的泛化能力 - +156.3% 提高了 Mind2Web 的泛化能力

## Next Steps

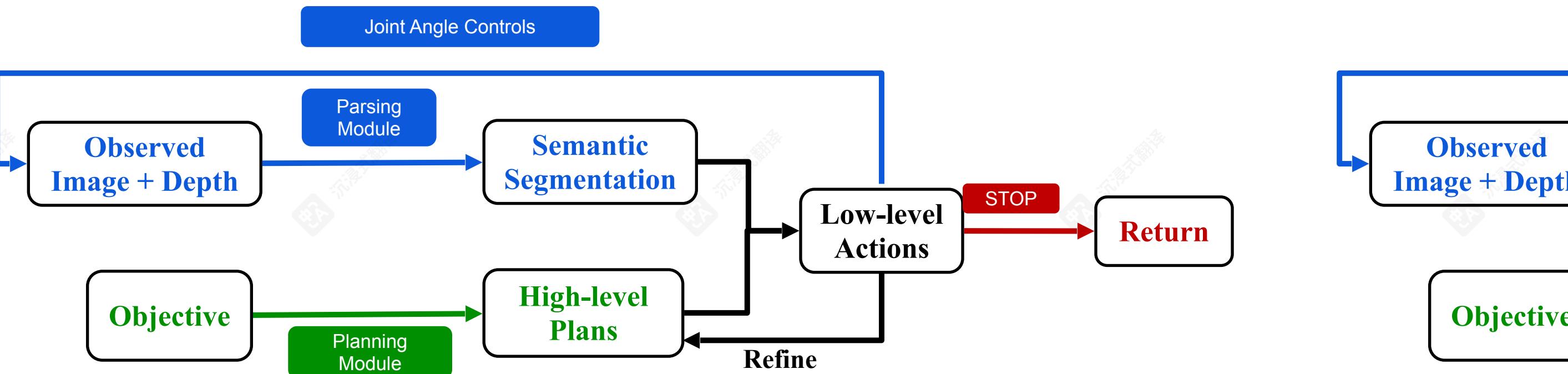
- There are 385M unique domains in the Common Crawl PageRank, suggesting another 1000x more data could be available by scaling further
- Moving towards **online RL**

## 下一步

- Common Crawl PageRank 中有 3.85 亿个唯一域名，表明通过进一步扩展，还有 1000 倍更多的数据可能可用
- 向在线强化学习过渡

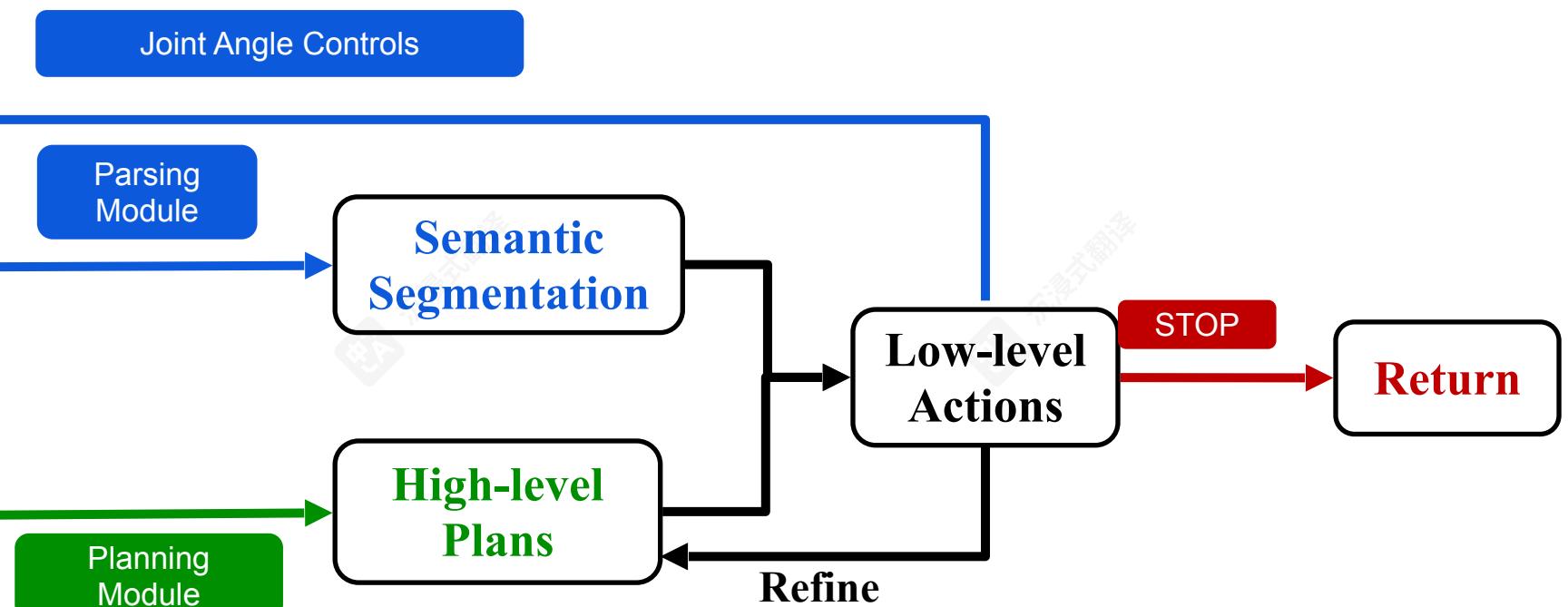
# Physical Agent: Long-horizon Robotic Manipulation Task

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

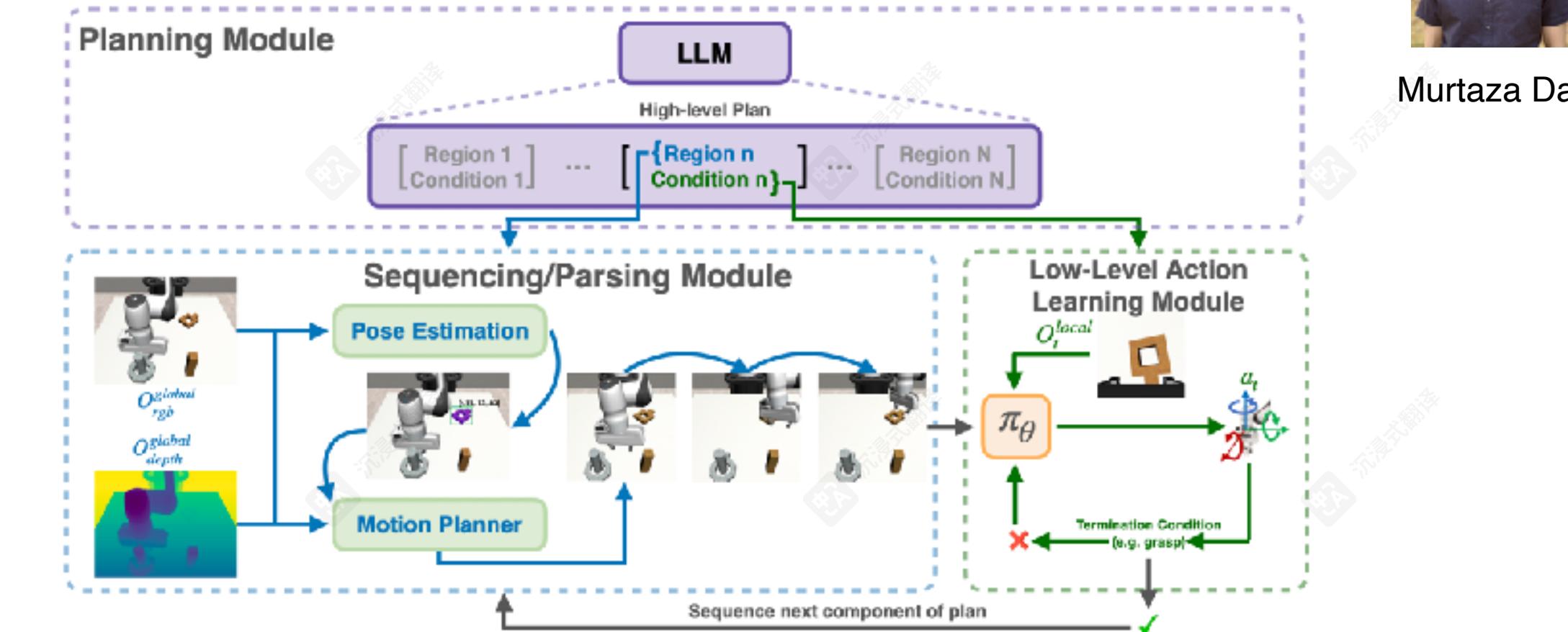


# 物理代理：长时程机器人操作任务

- 我们交互式代理的模型架构：
  - 高级规划
  - 观测解析
  - 低级动作生成

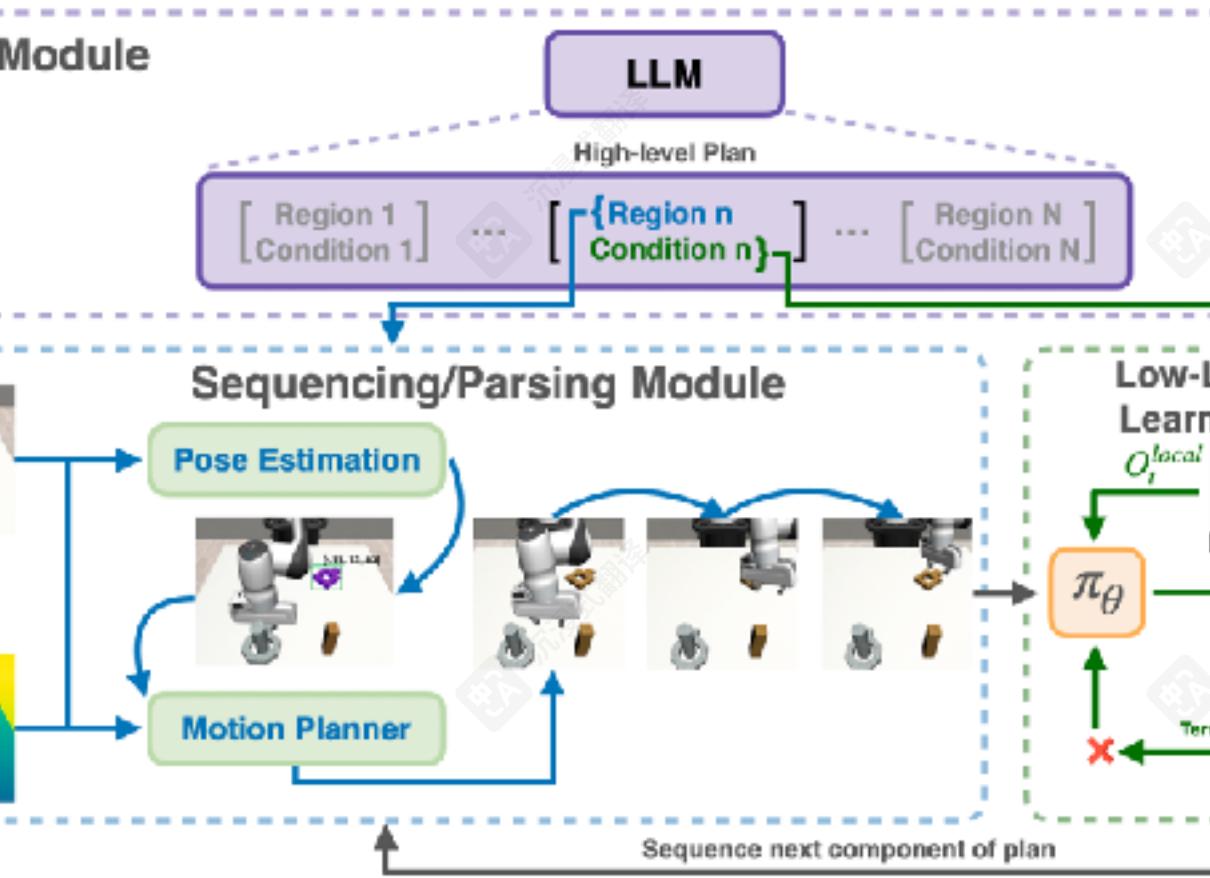


# Plan-Sequence-Learn



Murtaza Dalal

# Plan-Sequence-Learn



Murtaza Dalal

# 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.

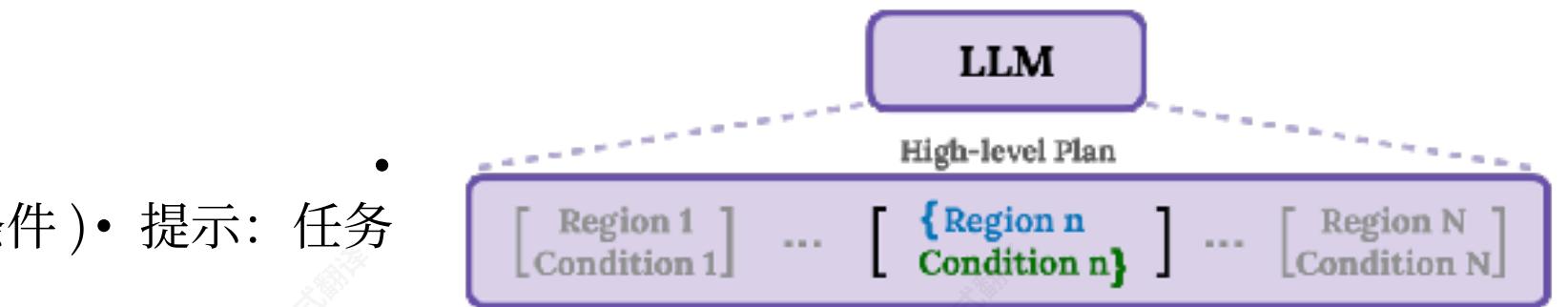
## Output:

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

# 规划模块

- 结构化语言计划：( 对象 , 条件 )

- 提示：任务描述、条件、对象、格式



**阶段终止条件:** ( 抓取 , 放置 ).

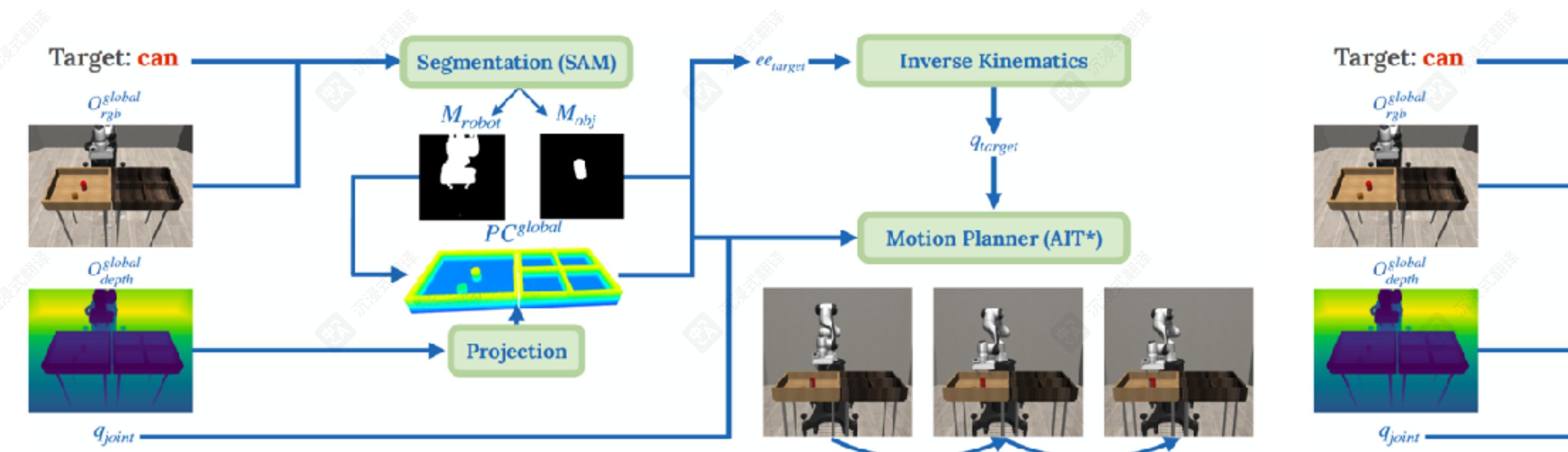
**任务描述:** 银色螺母放在银色插销上，金色螺母放在金色插销上。给我一个简单的计划来解决这个问题，只能使用阶段终止条件。确保计划遵循以下格式，并注意考虑对象几何形状。

**输出格式:** 一个列表，其中每个元素看起来像：(< 对象 / 区域 >,< 阶段终止条件 >)。不要输出其他任何内容。

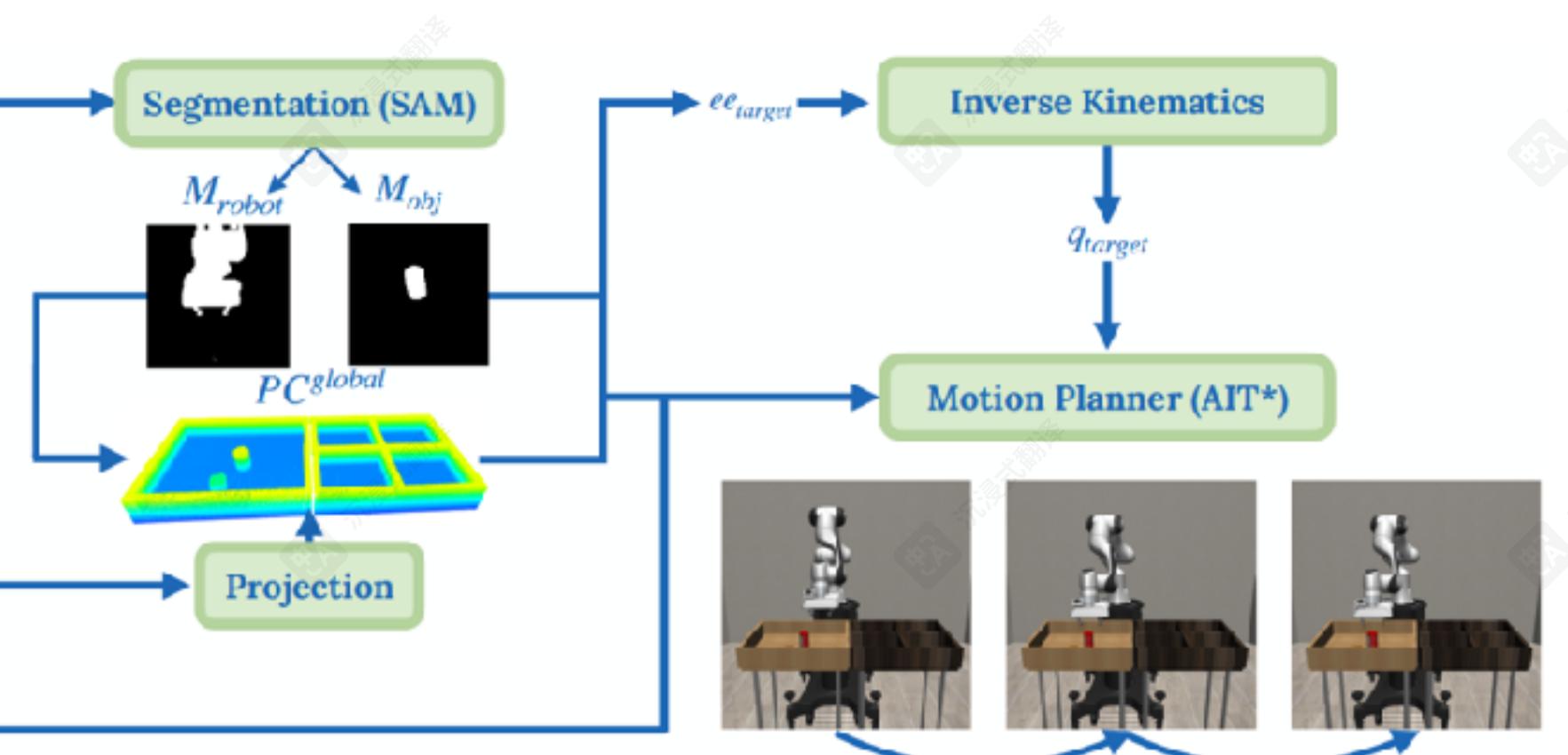
## 输出:

```
[("银螺母", "抓取"), ("银销钉", "放置"), ("金螺母", "抓取"), ("金销钉", "放置")]
```

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

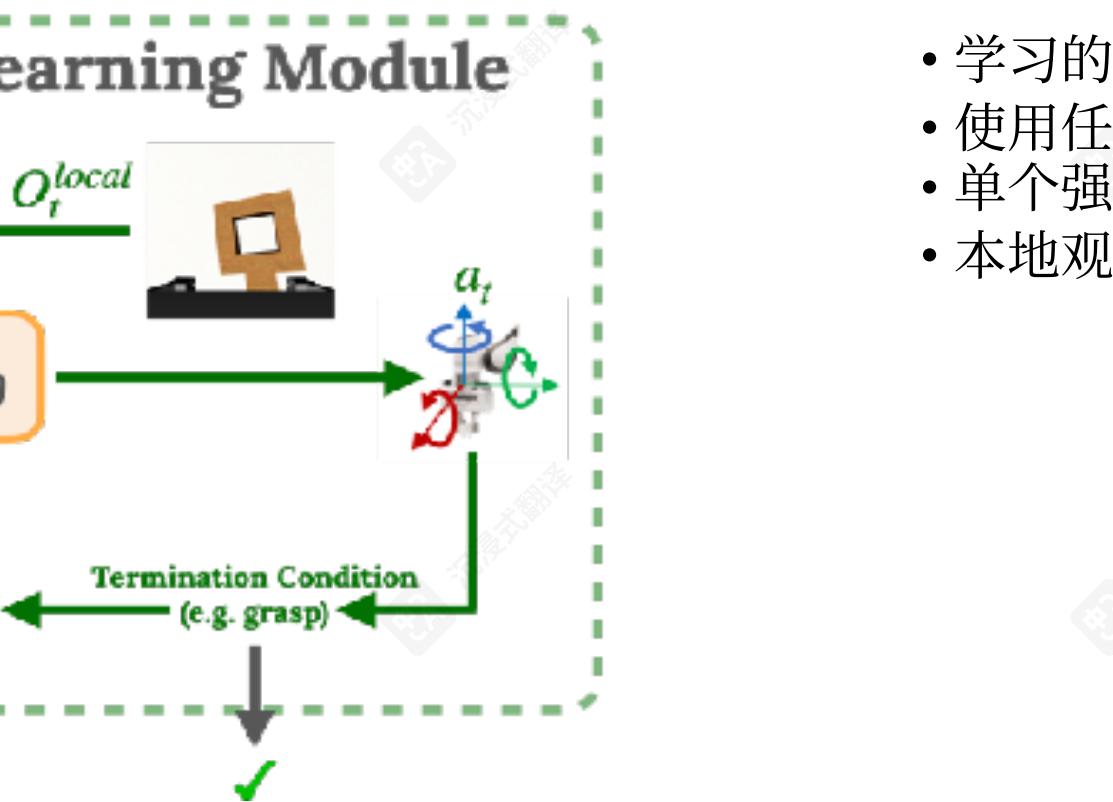


Sequencing/Parsing Module: 在场景  
中将语言计划进行基础定位



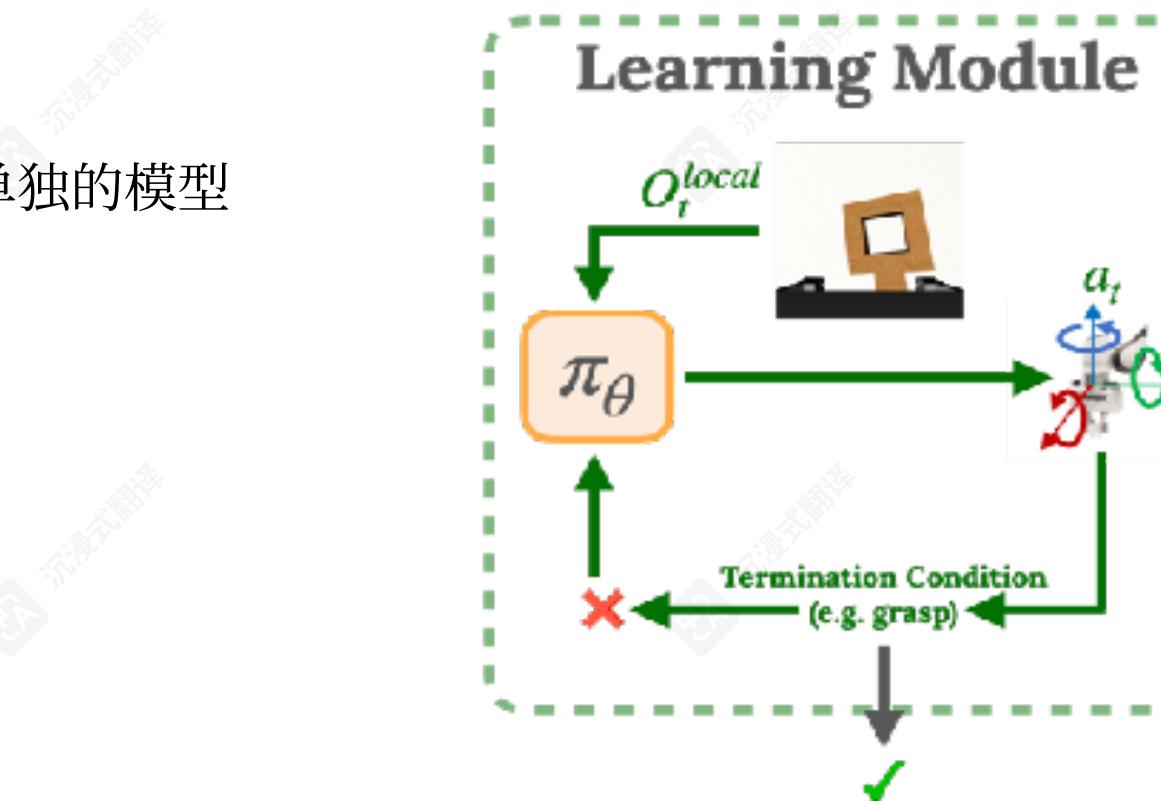
## Learning Low-level Actions Module: Learning Local Control

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



## 学习低级动作模块：学习本地控制

- 学习的交互强化学习策略
- 使用任务奖励进行训练
- 单个强化学习模型而不是每个阶段单独的模型
- 本地观察而不是全局观察



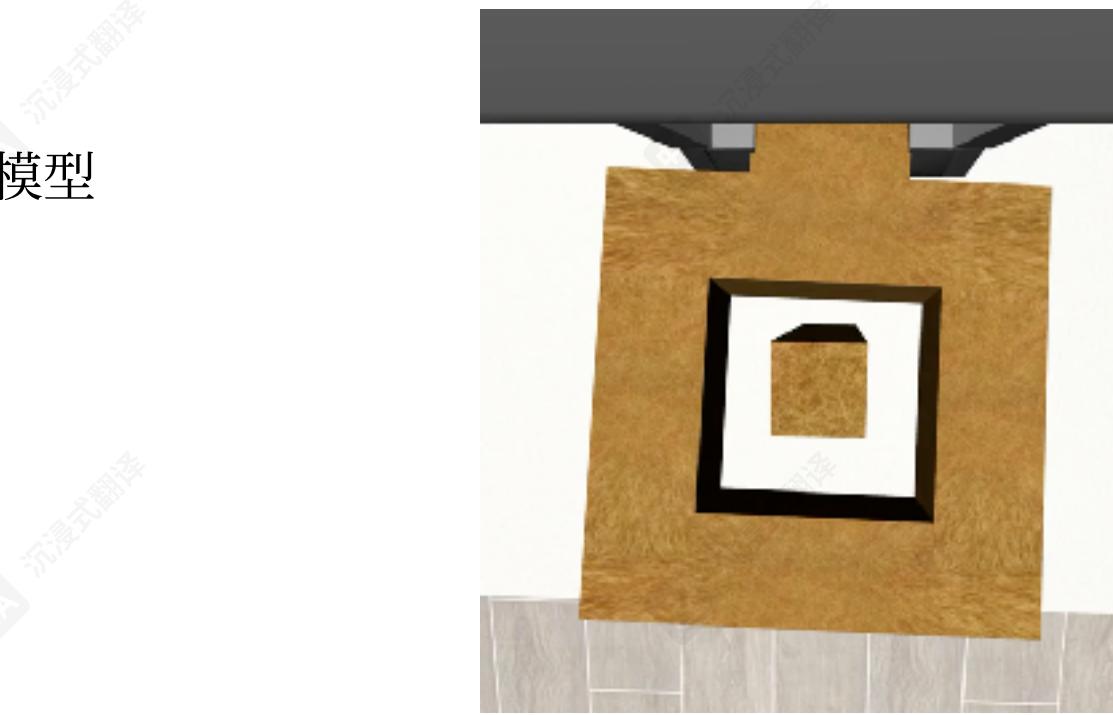
## Learning Low-level Actions Module: Learning Local Control

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

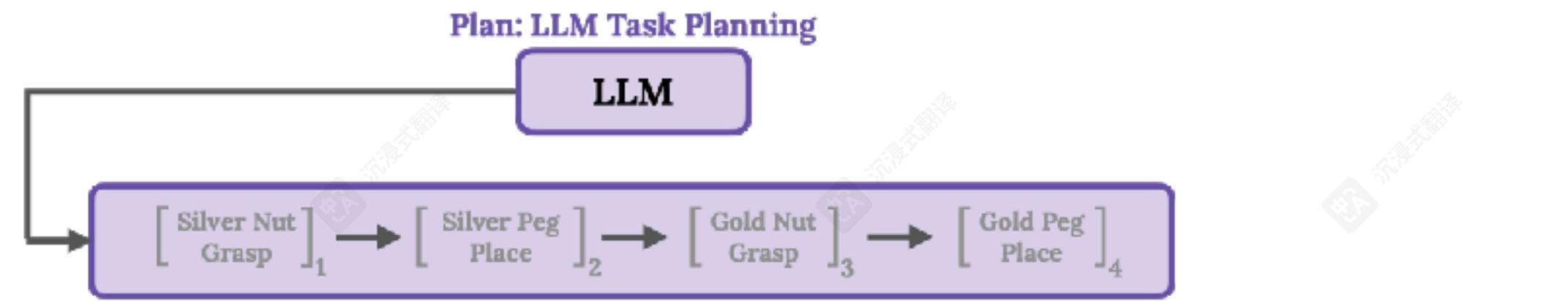


## 学习低级动作模块：学习本地控制

- 学习的交互 RL 策略
- 使用任务奖励进行训练
- 单个 RL 模型而不是每个阶段单独的模型
- 本地观察而不是全局观察



# Full Pipeline Example



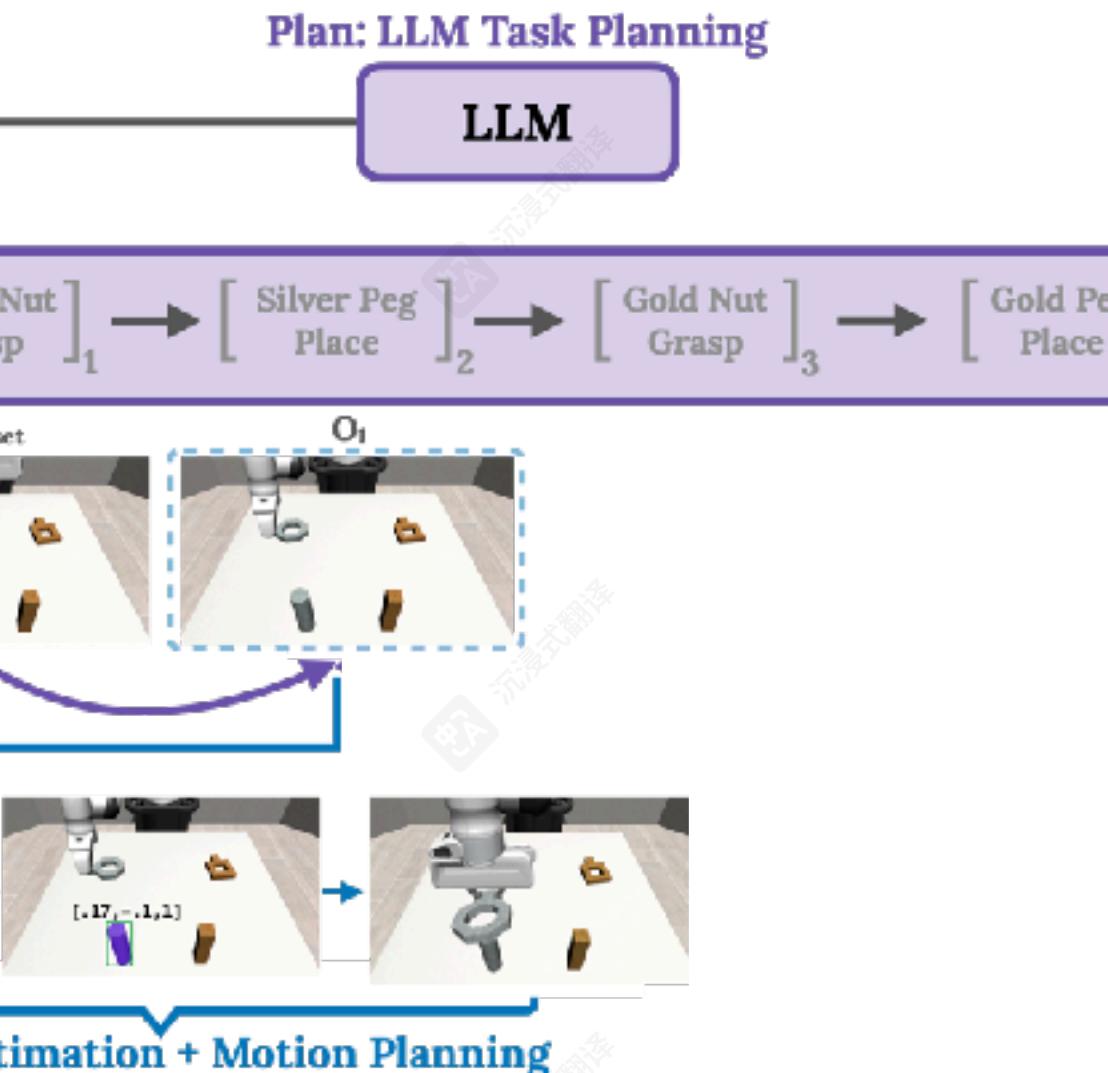
# Full Pipeline Example



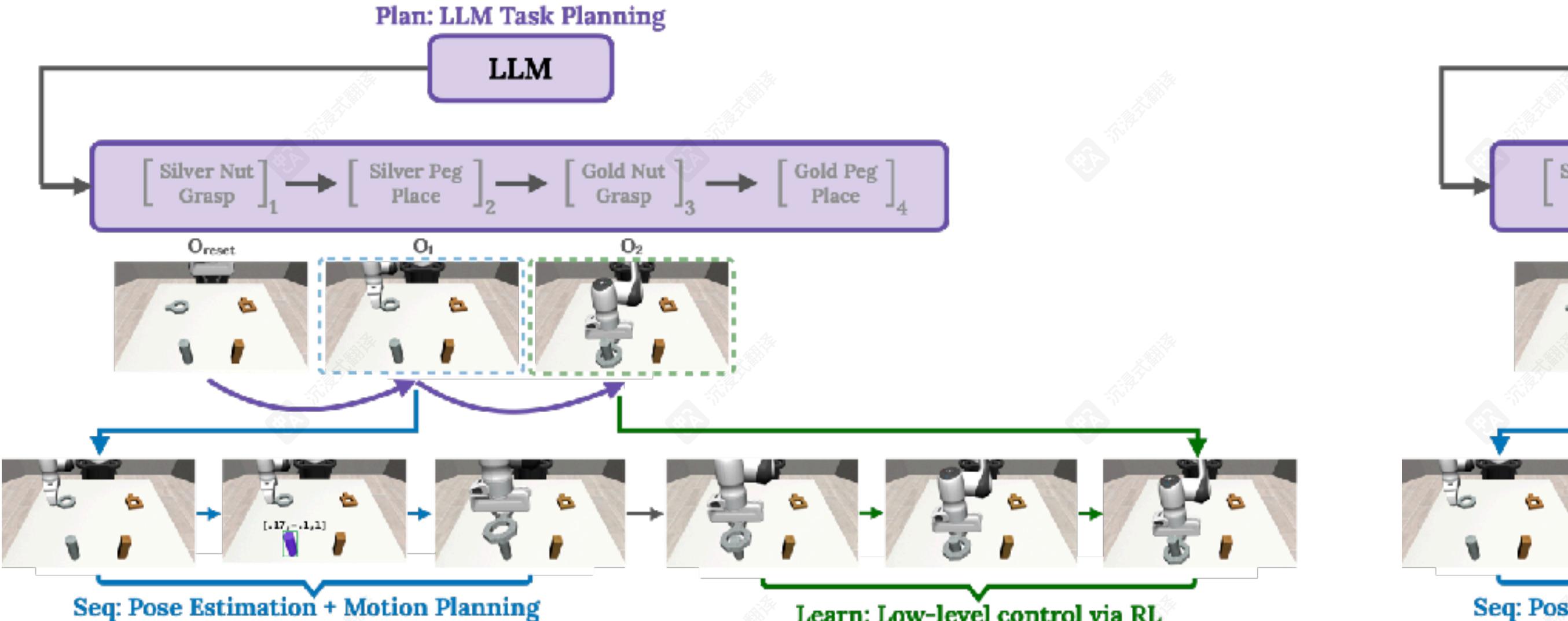
# Full Pipeline Example



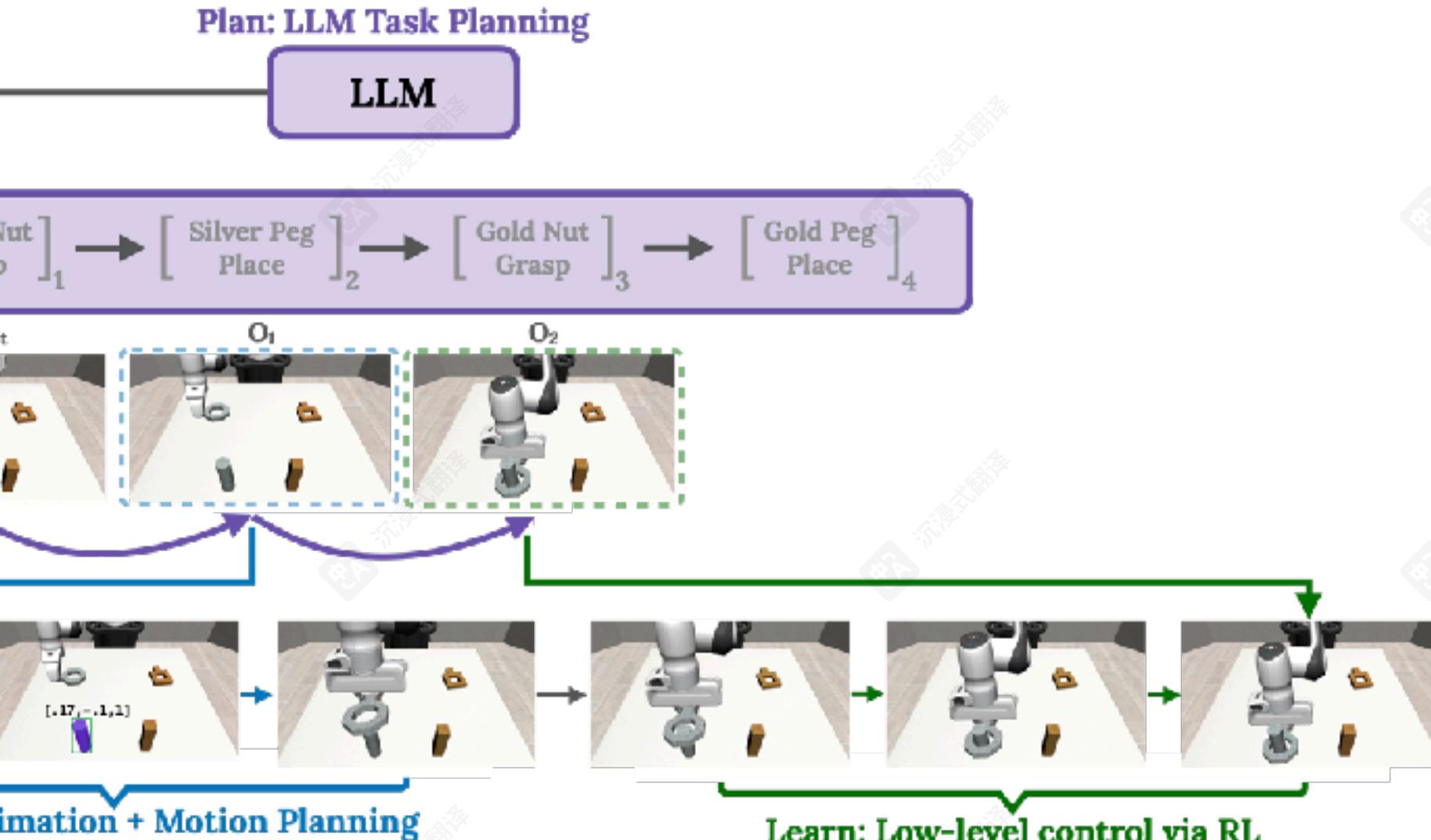
# Full Pipeline Example



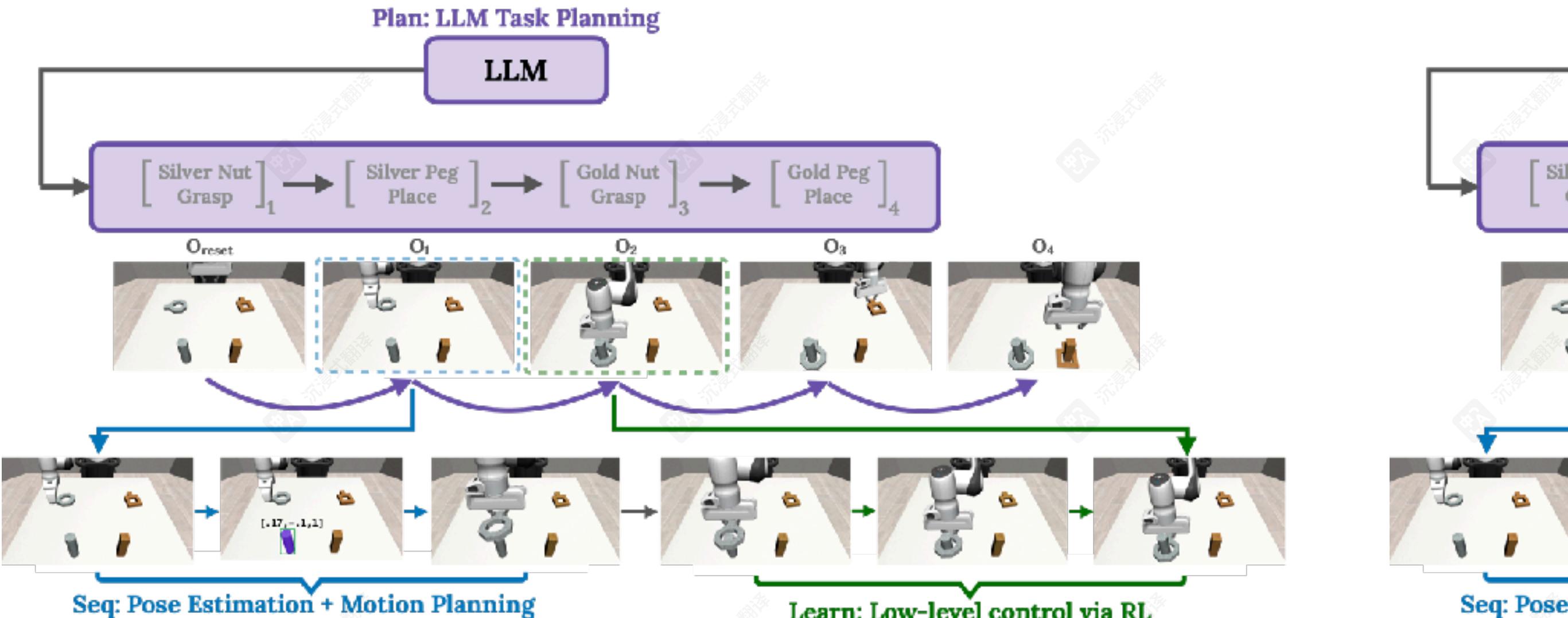
# Full Pipeline Example



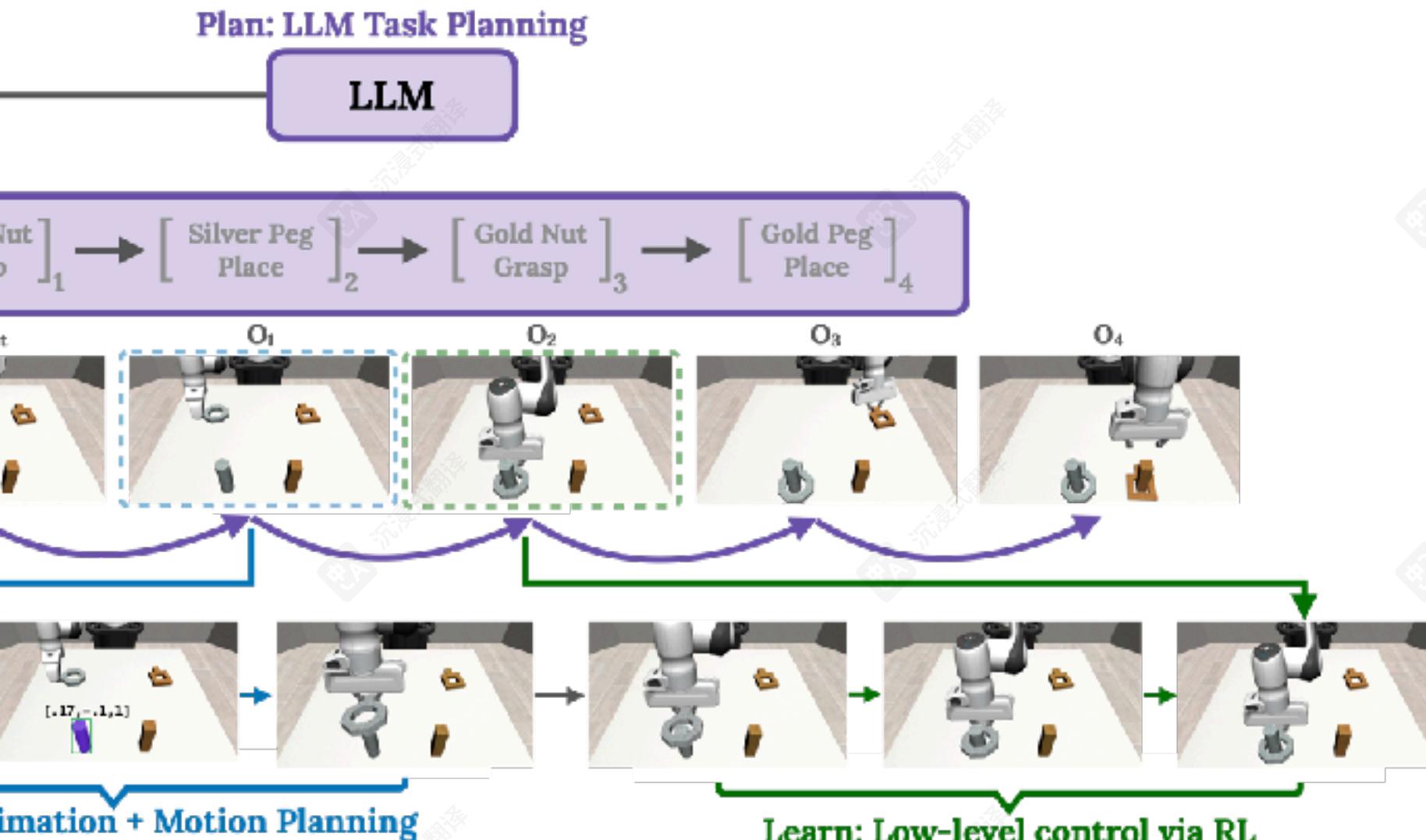
# Full Pipeline Example

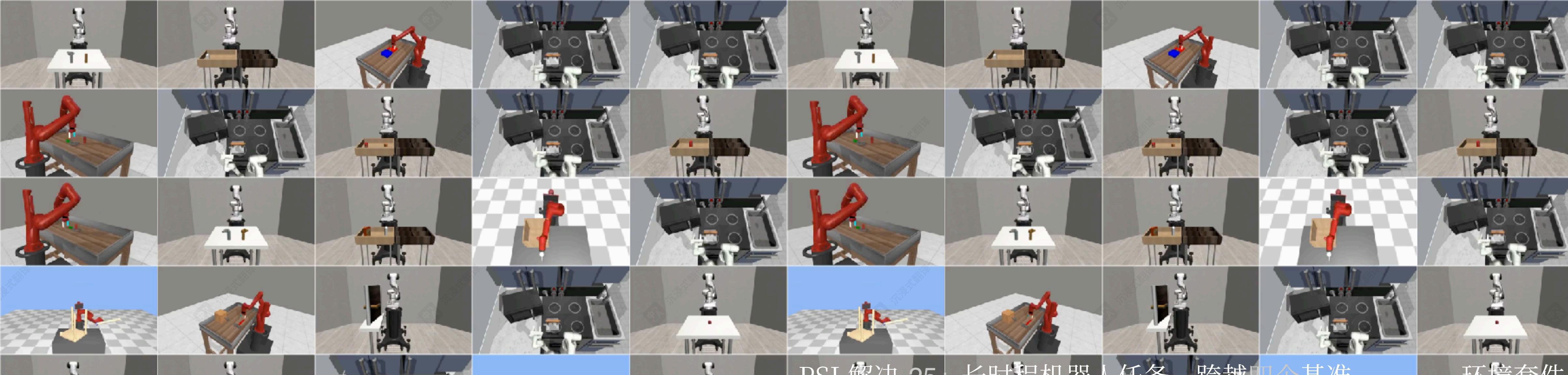


# Full Pipeline Example



# Full Pipeline Example





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

PSL解决 25+ 长时程机器人任务，跨越四个基准

成功率超过85%

NVIDIA CONFIDENTIAL。不要分发。

环境套件，

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



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



Generalizes to Novel Object Geometries/Categories



Generalizes to Novel Object Geometries/Categories



Manipulate novel objects with unseen receptacles



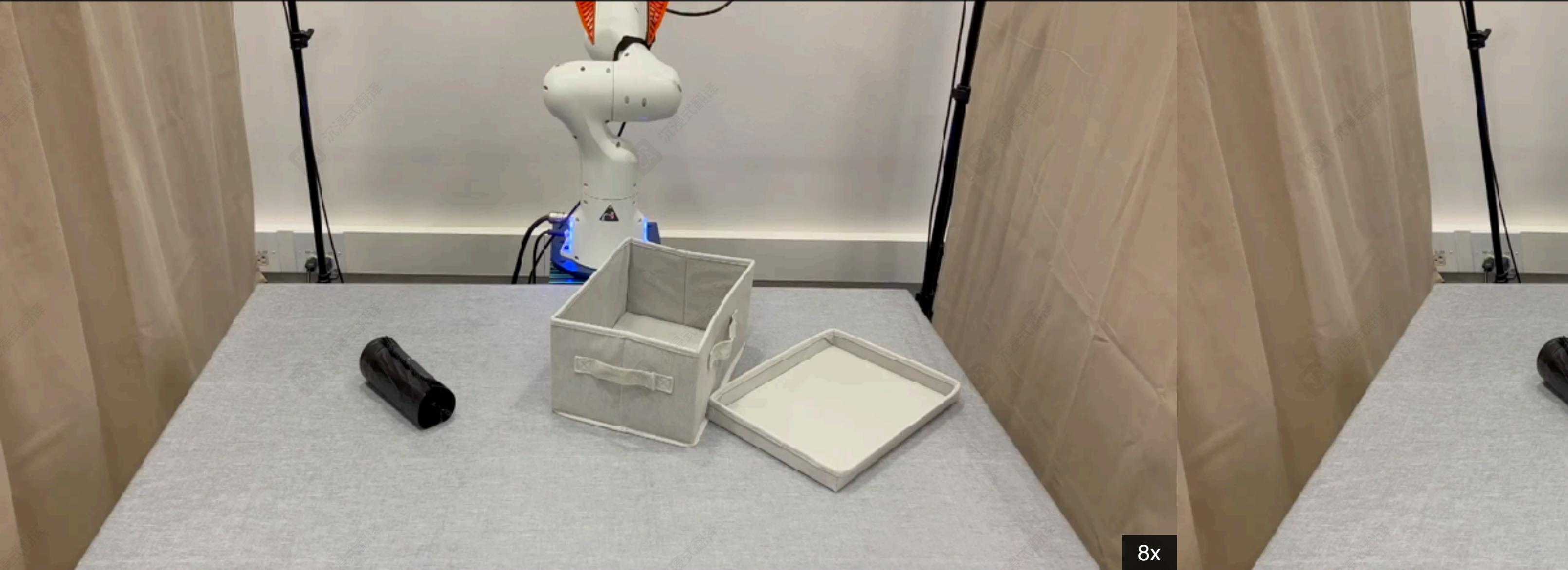
Manipulate novel objects with unseen receptacles



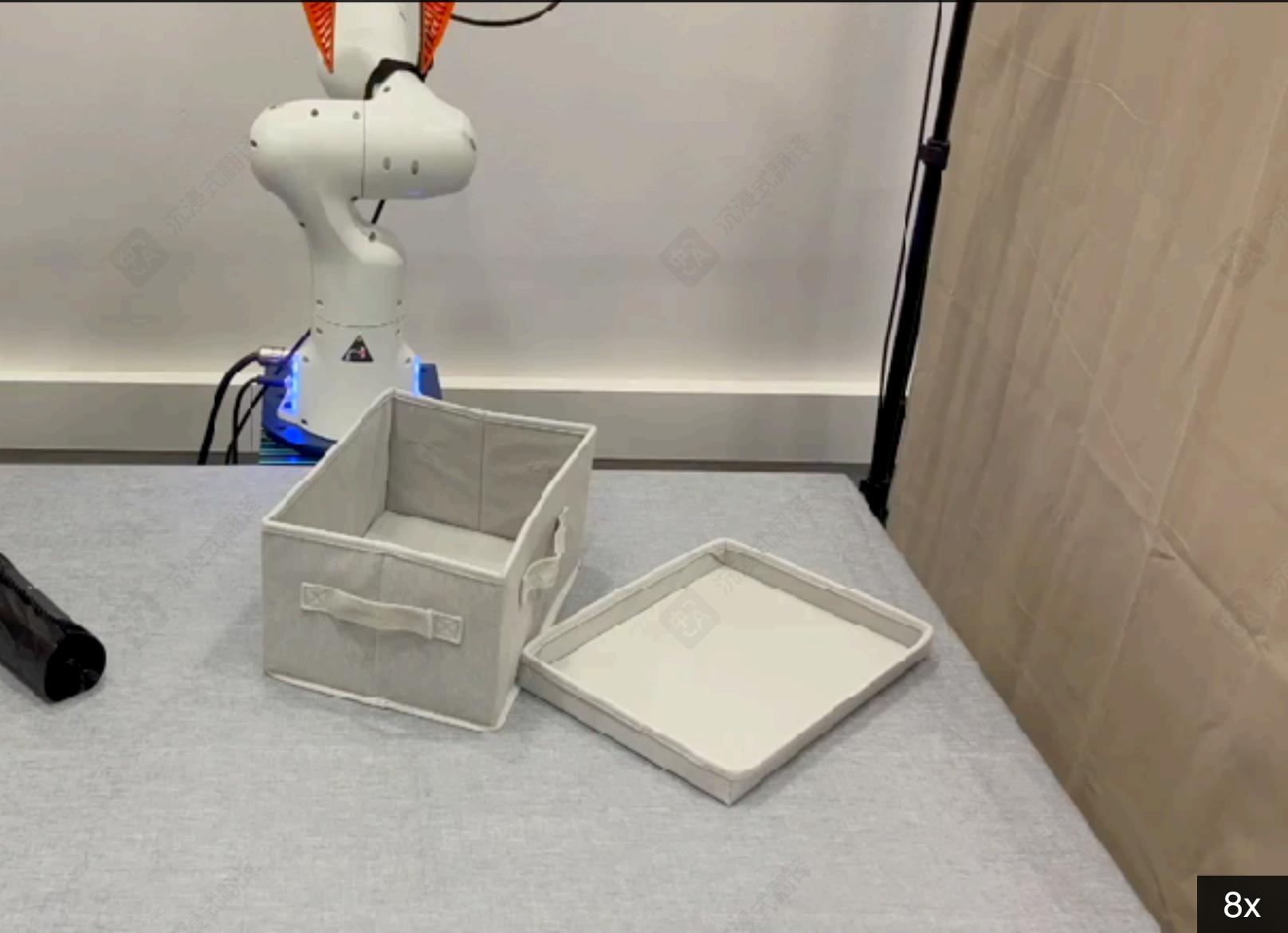
8x

8x

Manipulate Deformable Objects (not observed in sim!)



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
- Data pipeline for large-scale generation and verification of synthetic web tasks, powered by Llama models

## Summary

- VisualWebArena: 一个用于严格评估和提升自主多模态网络代理能力的真实任务基准
- 推理时搜索算法，旨在提升语言模型代理在真实网络任务中的能力
- 用于大规模生成和验证合成网络任务的数据管道，由 Llama 模型提供支持

# 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
- Data pipeline for large-scale generation and verification of synthetic web tasks, powered by Llama models
- **AI Safety and robustness, especially in the age of autonomous systems.**

# Summary

- VisualWebArena: 一个设计用于严格评估和提升自主多模态网络代理能力的真实任务基准
- Inference-time search algorithm designed to enhance the capabilities of language model agents on realistic web tasks
- 用于大规模生成和验证合成网络任务的管道，由 Llama 模型驱动
- 人工智能安全性和鲁棒性，尤其是在自主系统时代。

# 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



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

Carnegie Mellon University

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



Even while we were recording demonstrations of computer use for today's launch, we encountered some amusing errors. In one, Claude accidentally clicked to stop a long-running screen recording, causing all footage to be lost. In another, Claude suddenly took a break from our coding demo and began to peruse photos of Yellowstone National Park.

We expect that computer use will rapidly improve to become faster, more reliable, and more useful for the tasks our users want to complete. It'll also become much easier to implement for those with less software-development experience. At every stage, our researchers will be working closely with our safety teams to ensure that Claude's new capabilities are accompanied by the appropriate safety measures.

Even while we were recording demonstrations of computer use for today's launch, we encountered some amusing errors. In one, Claude accidentally clicked to stop a long-running screen recording, causing all footage to be lost. In another, Claude suddenly took a break from our coding demo and began to peruse photos of Yellowstone National Park.

We expect that computer use will rapidly improve to become faster, more reliable, and more useful for the tasks our users want to complete. It'll also become much easier to implement for those with less software-development experience. At every stage, our researchers will be working closely with our safety teams to ensure that Claude's new capabilities are accompanied by the appropriate safety measures.

Thank you

感谢