

Generalist Web Agents: From Perception to Planning

Boyuan Zheng

04/04/2025



Outline

1 Introduction

2 Environment Perception

3 Planning

4 Self-Improvement

Outline

1

Introduction

2

Environment Perception

3

Planning

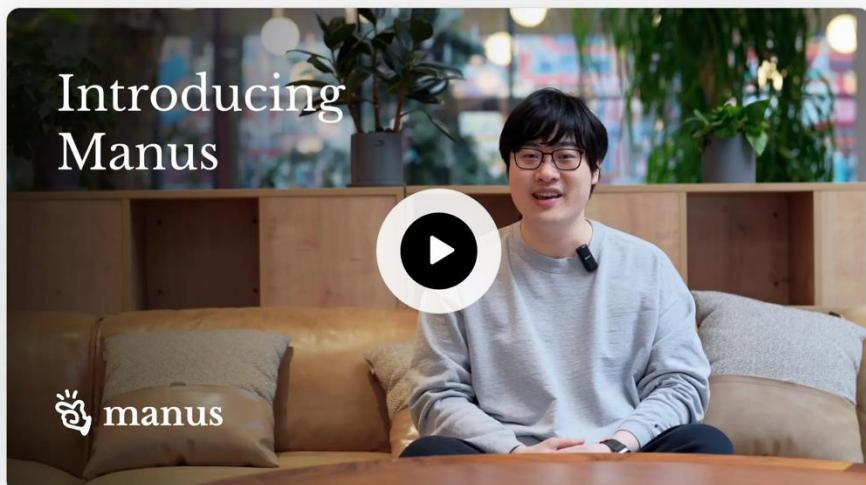
4

Self-Improvement

Coding Agent

Leave it to Manus

Manus is a general AI agent that bridges minds and actions: it doesn't just think, it delivers results. Manus excels at various tasks in work and life, getting everything done while you rest.



Try Manus

manus Comprehensive Tesla Stock Analy... Log in

I'd like a thorough analysis of Tesla stock, including:

Summary: Company overview, key metrics, performance data and investment recommendations

Financial Data: Revenue trends, profit margins, balance sheet and cash flow analysis

Market Sentiment: Analyst ratings, sentiment indicators and news impact

Technical Analysis: Price trends, technical indicators and support/resistance levels

Compare Assets: Market share and financial metrics vs. key competitors

Value Investor: Intrinsic value, growth potential and risk factors

Investment Thesis: SWOT analysis and recommendations for different investor types

manus

I'll help you create a comprehensive analysis of Tesla stock. I'll gather the latest financial data, market sentiment, technical analysis, competitive comparisons, and develop investment recommendations. This will take some time to research thoroughly, but I'll work on it right away and provide you with a detailed report.

Connected to datasource(6)

Manus is replaying the task... Skip to results

Manus's Computer

Manus is using Editor

Creating file tesla_analysis/dashboard/style...

style.css

```
/* Global Styles */
:root {
    --primary-color: #3498db;
    --secondary-color: #2ecc71;
    --accent-color: #e74c3c;
    --neutral-color: #34495e;
    --light-color: #ecf0f1;
    --dark-color: #2c3e50;
    --positive-color: #27ae60;
    --negative-color: #e0392b;
    --neutral-status-color: #f39c12;
    --strength-color: #27ae60;
    --weakness-color: #e74c3c;
    --opportunity-color: #3498db;
    --threat-color: #f39c12;
}

* {
    margin: 0;
    padding: 0;
    box-sizing: border-box;
    font-family: 'Segoe UI', Tahoma, Geneva, Verdana, sans-serif;
}

body {
```

Manus is working: Create interactive ... 11 / 12 ^
0:00 Thinking

Coding Agent

OpenHands

Code Less, Make More

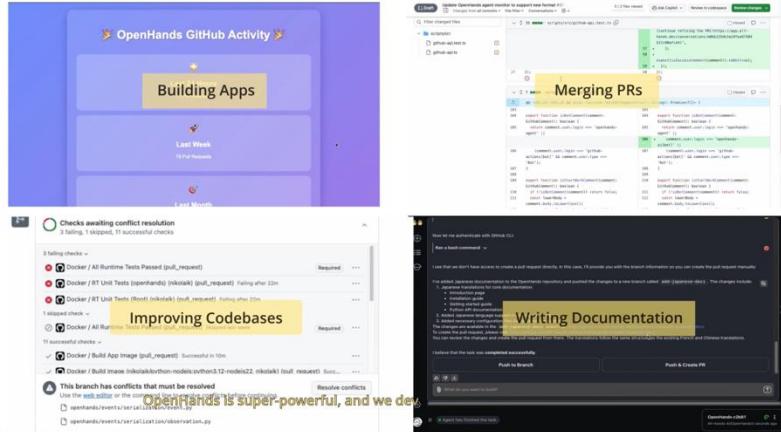
Use AI to tackle the toil in your backlog. Our agents have all the same tools as a human developer: they can modify code, run commands, browse the web, call APIs, and yes—even copy code snippets from StackOverflow.

[Get started with OpenHands.](#)

CONTRIBUTORS 267 STARS 52K COVERAGE 62% LICENSE MIT

SLACK JOIN US DISCORD JOIN US PROJECT CREDITS

PAPER ON ARXIV BENCHMARK SCORE



OpenHands GitHub Activity

Building Apps

Last Week 76 Pull Requests

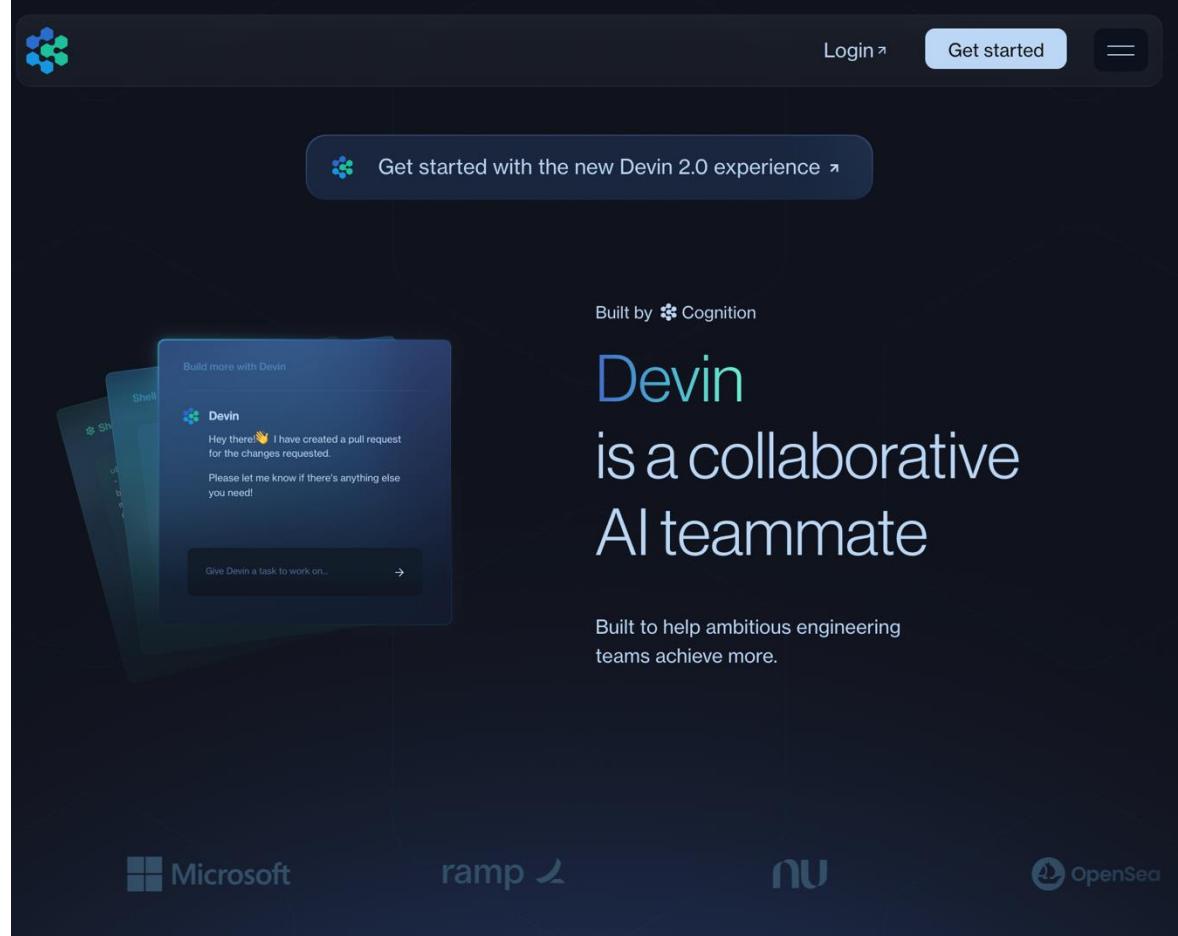
Checks awaiting conflict resolution

Improving Codebases

Merging PRs

Writing Documentation

OpenHands is super-powerful, and we dev...



Login ↗ Get started

Get started with the new Devin 2.0 experience ↗

Built by Cognition

Devin is a collaborative AI teammate

Built to help ambitious engineering teams achieve more.

Microsoft ramp nu OpenSea

Web Agent

Introducing Operator

A research preview of an agent that can use its own browser to perform tasks for you. Available to Pro users in the U.S.

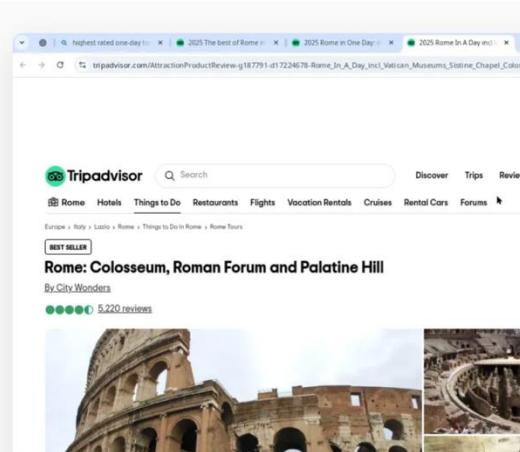
[Go to Operator ↗](#)

Find and book me the highest rated one-day tour of Rome on TripAdvisor.

I'll search for the highest-rated tour of historic Rome on TripAdvisor. Once I find a suitable option, I'll provide you with the details. Let's begin.

Worked for 2 minutes ^

Navigating to TripAdvisor website
Selecting "Things to Do" category
Searching for historic Rome tours
Closing pop-up, continuing tour search
Searching for Historic Rome tours
Exploring all historic Rome tour options
Closing Colosseum tab, resuming tour search
Closing tour pop-up, tab afterward

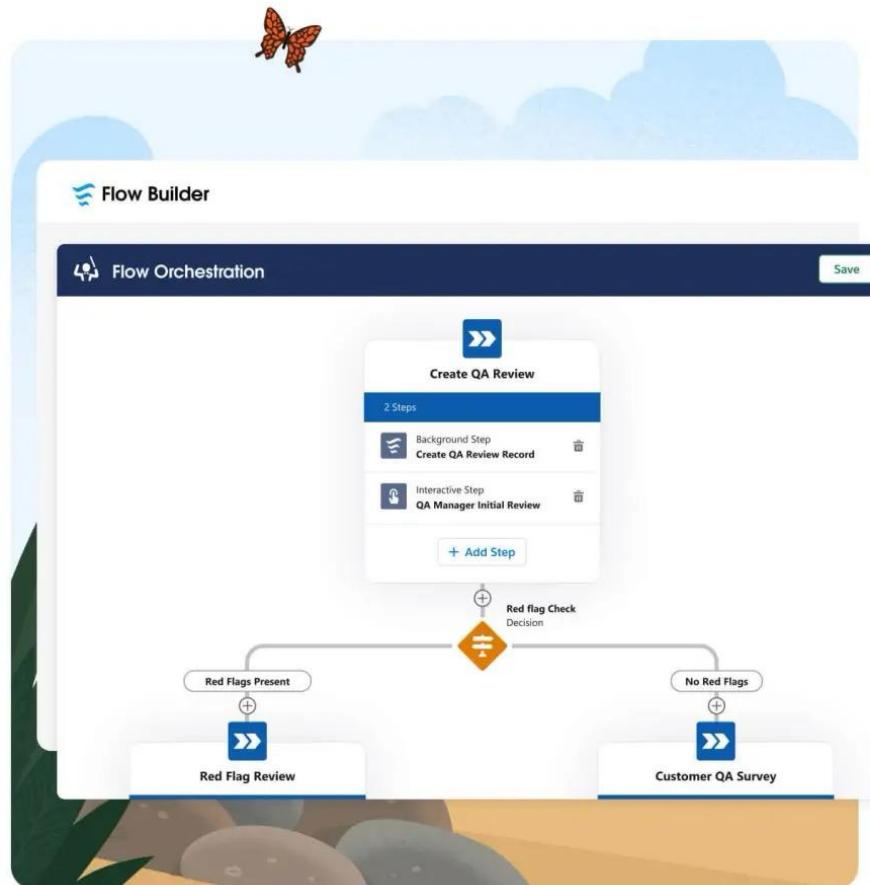


Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku

Oct 22, 2024 • 5 min read



Web Agent Use Case: Process Automation



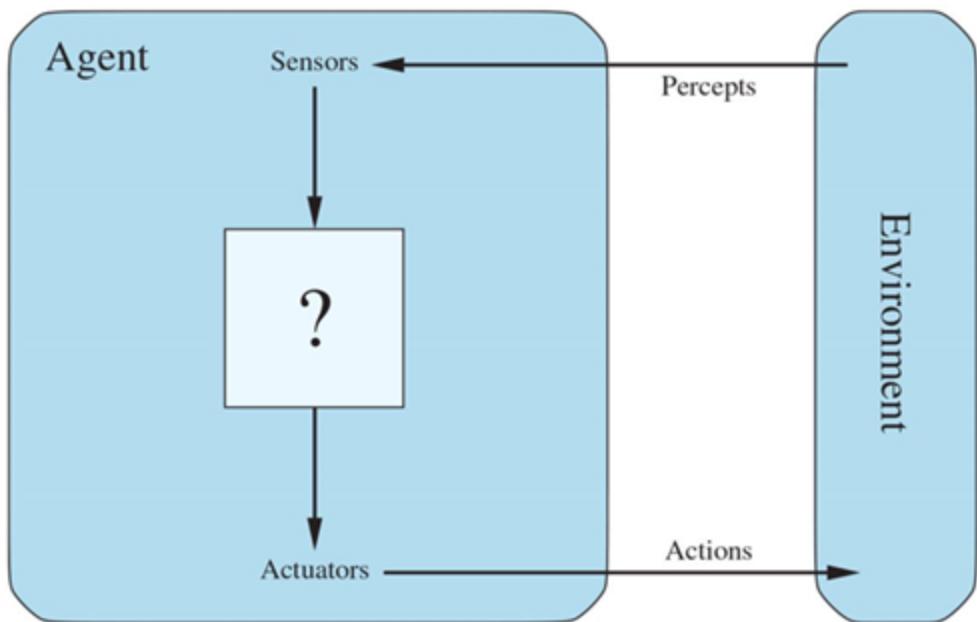
Web Agent Use Case: Auto-Testing

The image shows two screenshots side-by-side. On the left is the homepage of Spur, featuring a 'Launch YC' badge, a counter '171', and the headline 'Your AI QA Engineer!'. Below this is a sub-headline: 'Spur lets you test your website with Natural Language.' At the bottom are 'Book a Demo' and 'Sign Up' buttons, and a 'Backed by Y Combinator' logo. On the right is a screenshot of the Spur test editor. It shows a test titled 'Inviting Co-Founders' from today at 12:17 PM. The expected result is 'Successfully tested inviting co-founders'. The test steps are listed as follows:

- 1 Click on [Finish application]
- 2 Click on [+ Add a co-founder]
- 3 Type [test@yc.com] in the Email field
- 4 Click on [Send Invite]
- 5 Verify that the text on the portal has changed to [Invite sent to test@yc.com]
- 6 Click on [+ Add a co-founder]
- 7 Type [test1@yc.com] in the Email field

The interface includes a preview of the application portal, zoom controls (0.5x, 1x, 2x, 4x, 8x), and a video player.

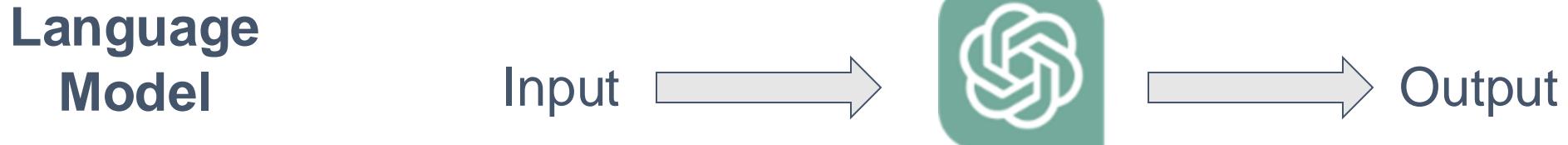
Why Agents Again?



“An **agent** is anything that can be viewed as perceiving its **environment** through **sensors** and acting upon that environment through **actuators**”

— Russel & Norvig, *AI: A Modern Approach*

From Language Model to Language Agent

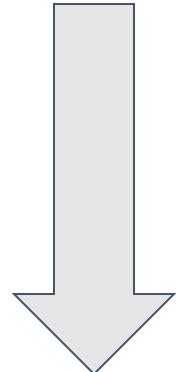


Opportunities of contemporary AI Agent with an integrated LLM:

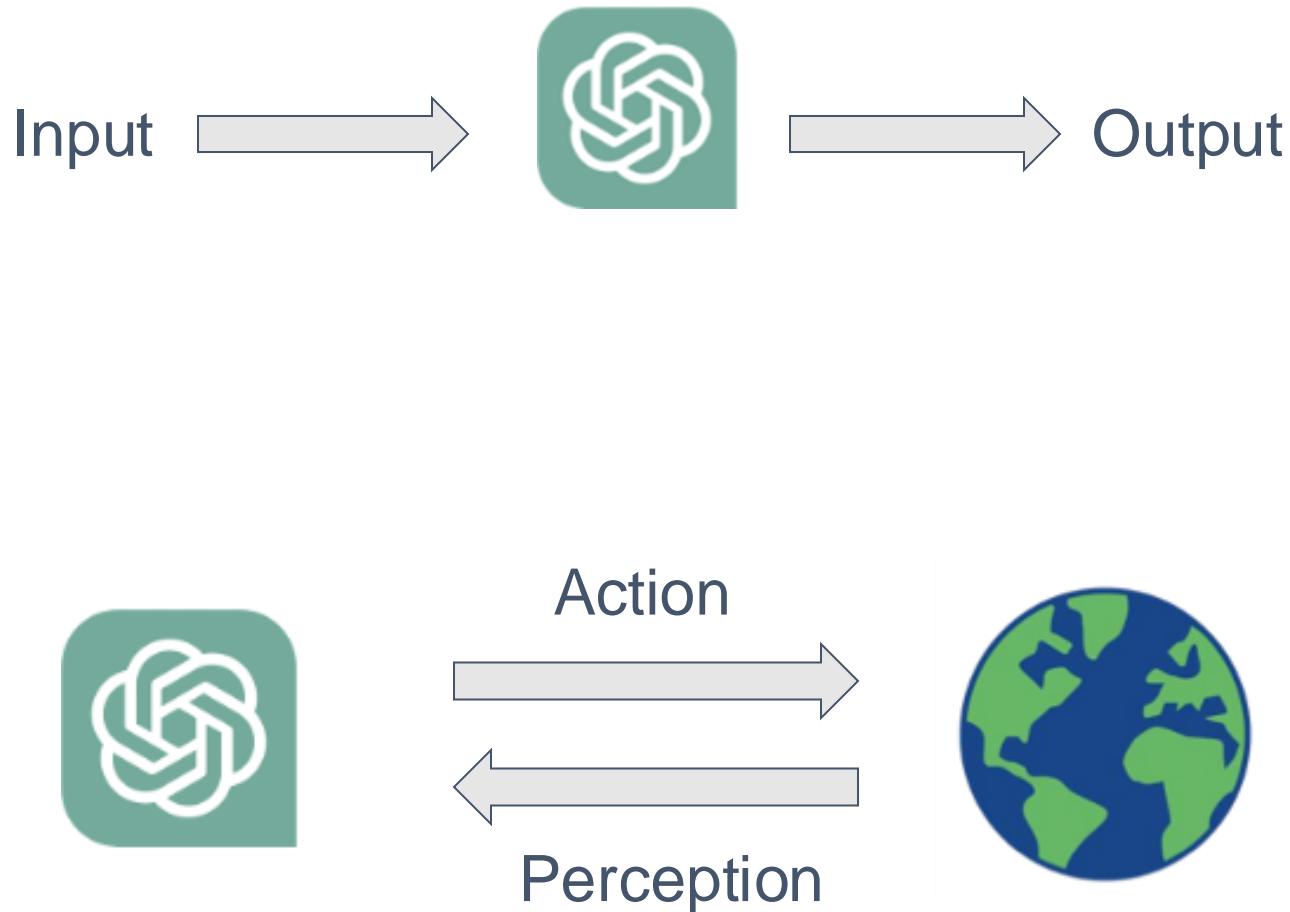
- Environment Perception
- Generalist agent with language
- Reasoning for better acting

From Language Model to Language Agent

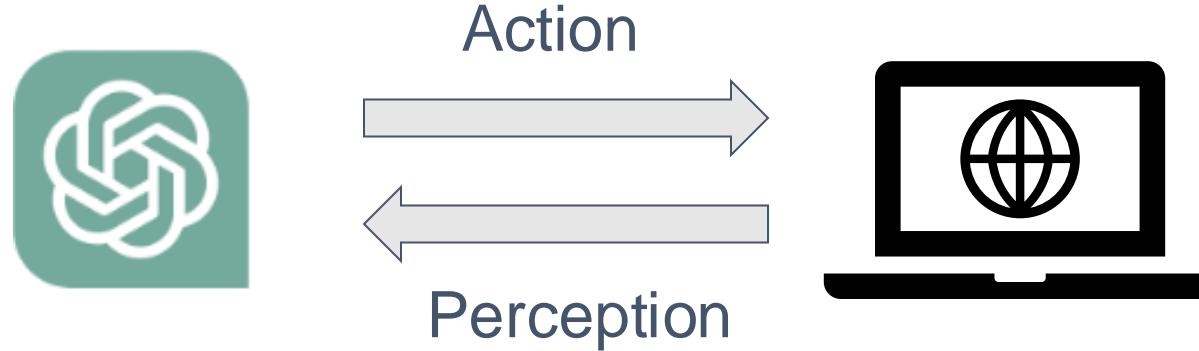
Language
Model



Language
Agent



Web Agent



Web as playgrounds for agent research

- Open, Diverse, Complex Environment
- Highly Structured Environment
- Low Cost
- Lower Safety Concern

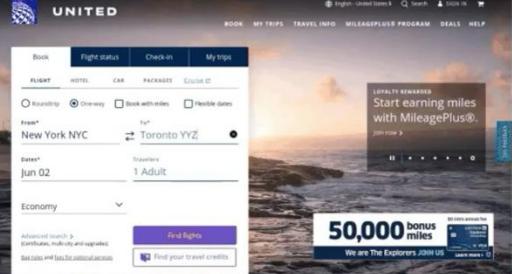
Generalist Web Agent

Cross-Task

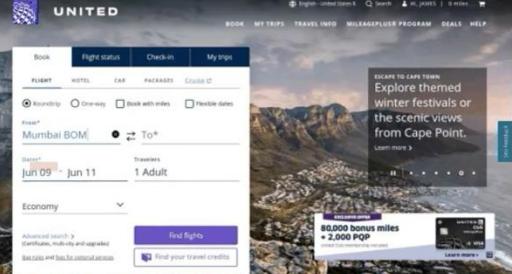
Cross-Website

Cross-Domain

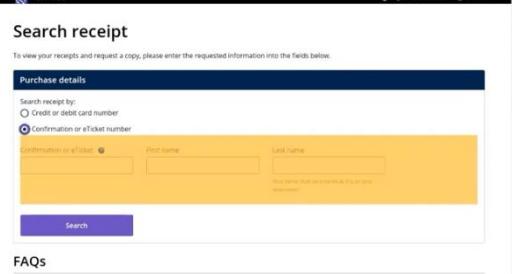
(a) Find one-way flights from New York to Toronto.



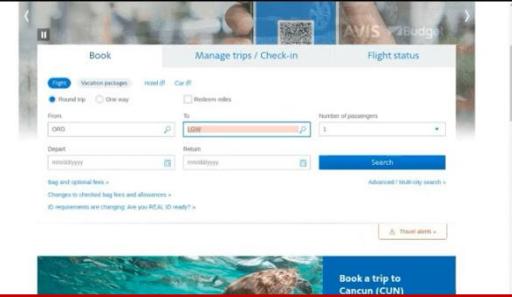
(b) Book a roundtrip on July 1 from Mumbai to London and vice versa on July 5 for two adults...



(c) Search receipt with the eTicket 12345678 for the trip reserved by Jason Two



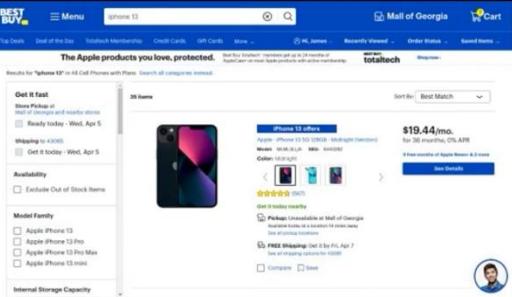
(a) Find a flight from Chicago to London on 20 April and return on 23 April.



(e) Search for the interactions between ibuprofen and aspirin.



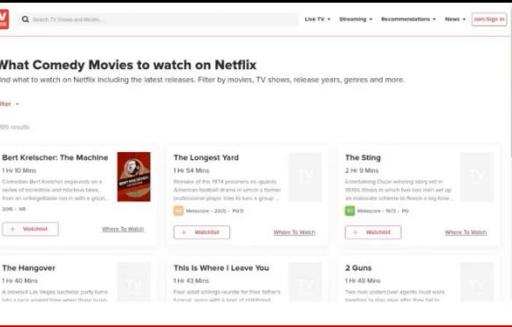
(T) As a verizon user, finance a blue iPhone 13 with 256gb along with monthly apple care.



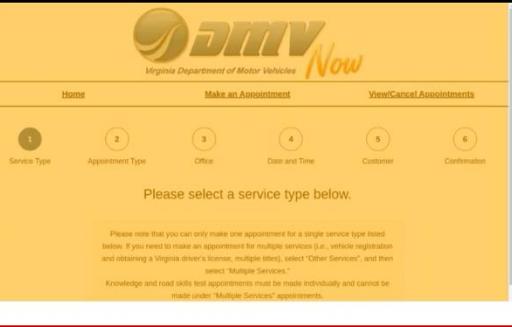
(g) Find Elon Musk's profile and start following, start notifications and like the latest tweet.



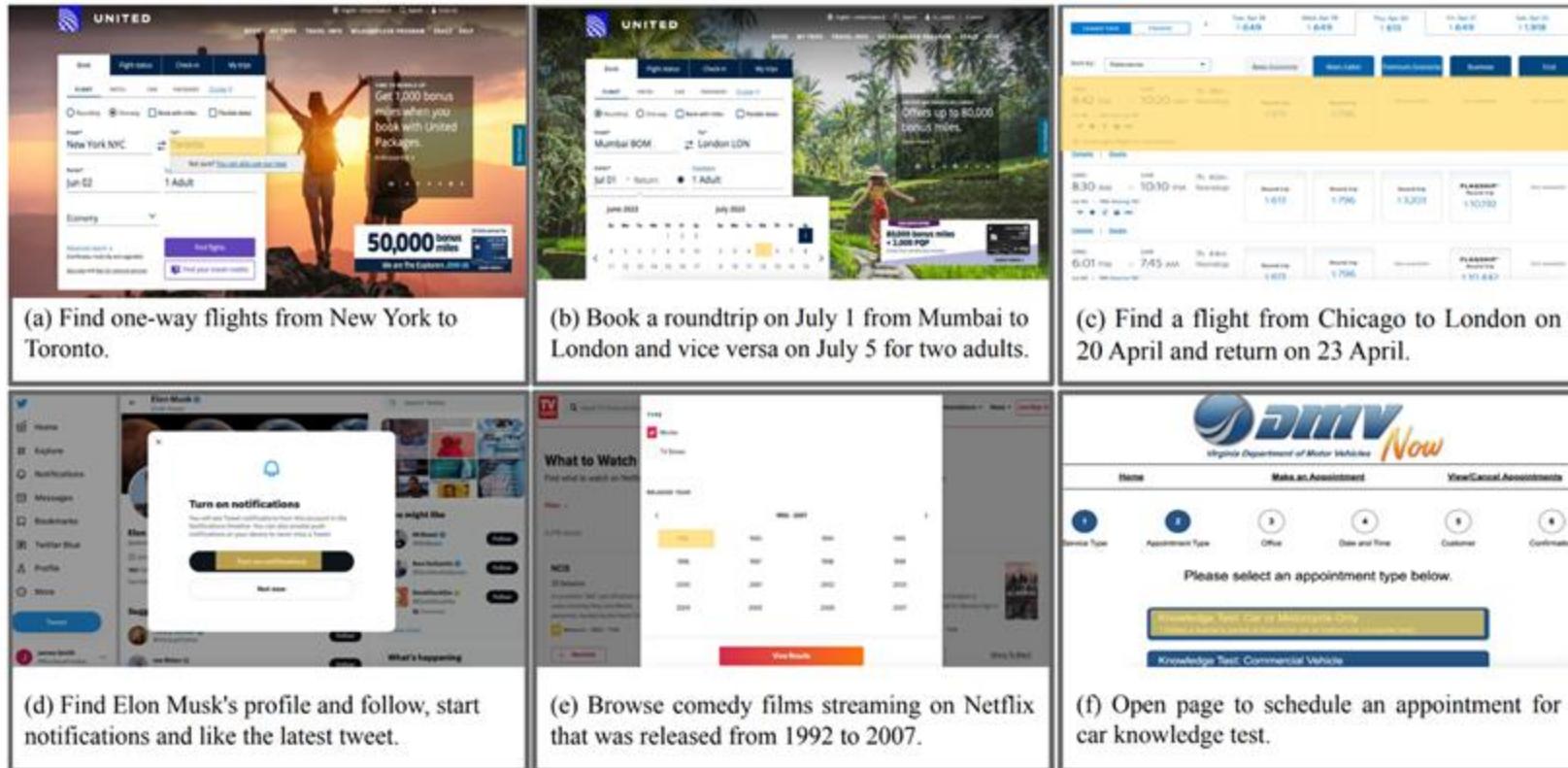
(h) Browse comedy films streaming on Netflix that was released from 1992 to 2007.



(i) Open page to schedule an appointment for car knowledge test.

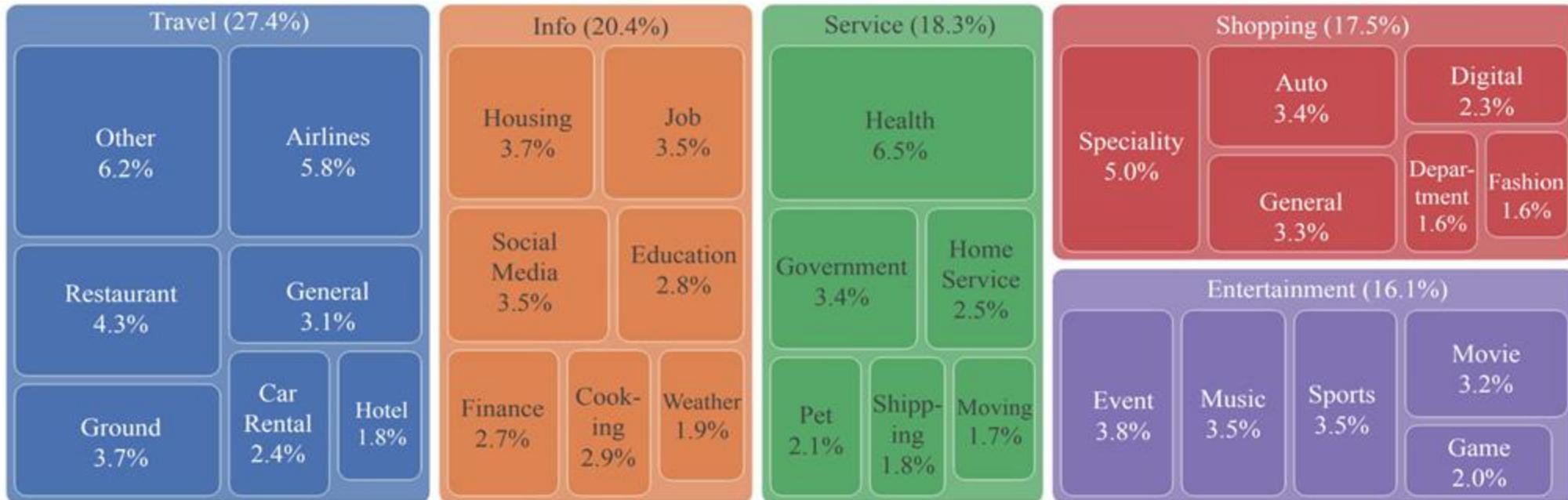


Why a Challenging Mission?



- User tasks: diverse, complex, requiring long-horizon planning
- Real-world websites: complex, dynamic
- Generalize to unseen websites/domains

Mind2Web: A Generalist Web Agent Benchmark



	# Dom.	# Env.	Env. Type	Avg. # Elements	# Tasks	Task Info.	Avg. # Actions
MiniWoB++ [22]	—	100	Simplified mobile websites	28	100	Low-level	3.6
WebShop [40]	1	1	Simplified shopping websites	38	12,000 products	High-level	11.3
RUSS [39]	—	22	Real-world websites	801	80	High & low	5.4
PixelHelp [21]	4	4	Mobile apps	—	187	High & low	-
META-GUI [35]	6	11	Mobile apps	79	1,125 dialogues	High-level	4.3
MoTIF [5]	15	125	Mobile apps	188	756	High & Low	4.4
MIND2WEB	5 / 31	137	Real-world websites	1,135	2,350	High-level	7.3

A task example in Mind2Web

Task Description:

Show me the reviews for the auto repair business closest to 10002.

Action Sequence:

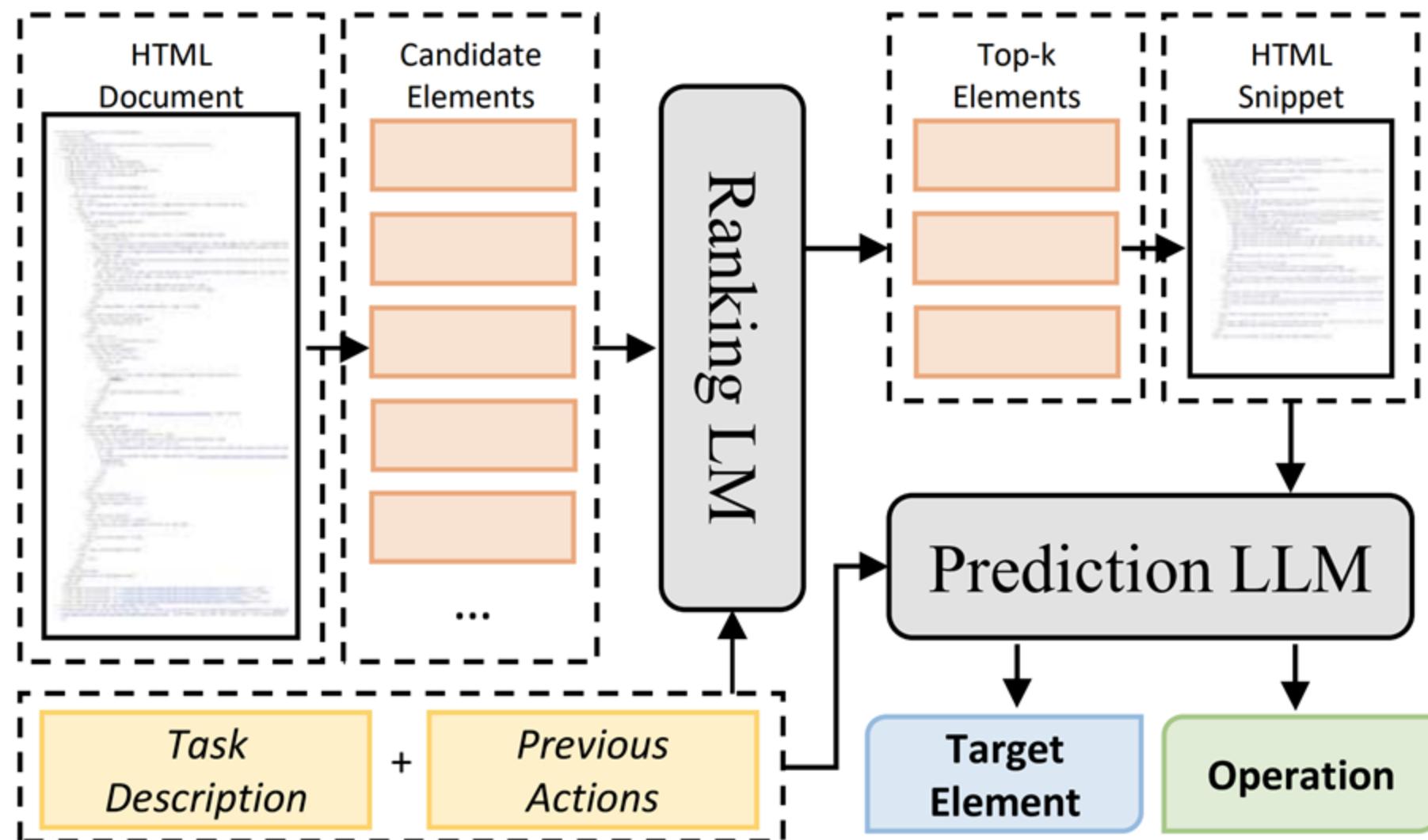
Target Element	Operation
1. [searchbox] Find	TYPE: <i>auto repair</i>
2. [button] Auto Repair	CLICK
3. [textbox] Near	TYPE: 10002
4. [button] 10002	CLICK
5. [button] Search	CLICK
6. [switch] Show BBB Accredited only	CLICK
7. [svg]	CLICK
8. [button] Sort By	CLICK
9. [link] Fast Lane 24 Hour Auto Repair	CLICK
10. [link] Read Reviews	CLICK

Webpage Snapshots:

The figure displays ten sequential screenshots of a web browser window, labeled Action 1 through Action 10, illustrating the process of finding auto repair reviews near the zip code 10002. The browser's address bar shows the URL www.bbb.org. The screenshots show the following sequence of interactions:

- Action 1:** The initial search bar contains the placeholder "Find a Business". The text "" is highlighted in red.
- Action 2:** The search bar now contains the text "Auto Repair". The text "*Auto Repair*" is highlighted in red.
- Action 3:** The search bar now contains the text "Near 10002". The text "**10002**" is highlighted in red.
- Action 4:** The search bar now contains the text "Near 10002". The text "**Search**" is highlighted in red.
- Action 5:** The search results page is shown for "Auto Repair" near "10002". The text "**Search**" is highlighted in red.
- Action 6:** The search results page is shown for "Auto Repair" near "10002". A button labeled "Show BBB Accredited only" is highlighted in red.
- Action 7:** The search results page is shown for "Auto Repair" near "10002". The text "Fast Lane 24 Hour Auto Repair" is highlighted in red.
- Action 8:** The search results page is shown for "Auto Repair" near "10002". The text "Read Reviews" is highlighted in red.
- Action 9:** The search results page is shown for "Auto Repair" near "10002". The text "Read Reviews" is highlighted in red.
- Action 10:** The final result page for "Fast Lane 24 Hour Auto Repair" is shown, displaying the business profile and review section. The text "Read Reviews" is highlighted in red.

The first framework for building a generalist web agent



Our Proposed MindAct

MindAct

• Environment Representation: Simplified HTML

```
<html> <form id=0> <div meta="navigation; sitelinks">
<p> <a> Collect Renaissance </a> <a> Shop Le Meridien
</a> <a> Westin Store </a> <a> Sheraton Store </a>
</p> </div> ... <div> <select id=1 meta="Size; Select a
Size"> <span meta=tablist> <button id=2 meta="button;
tab"> Description </button> ... <a id=3 meta="Shop
Feather & Down Pillow"> <img meta="Product Feather &
Down Pillow"> <p> <a> California Privacy Rights </a>
<a> Privacy Statement </a> <a> Terms of Use </a> <a
id=4> Loyalty Terms </a> ...
```

Based on the HTML webpage above, try to complete the following task:

Task: Search for queen-size pillow protectors from the Marriott shop, and if found, add two pieces to the cart and checkout.

Previous actions:

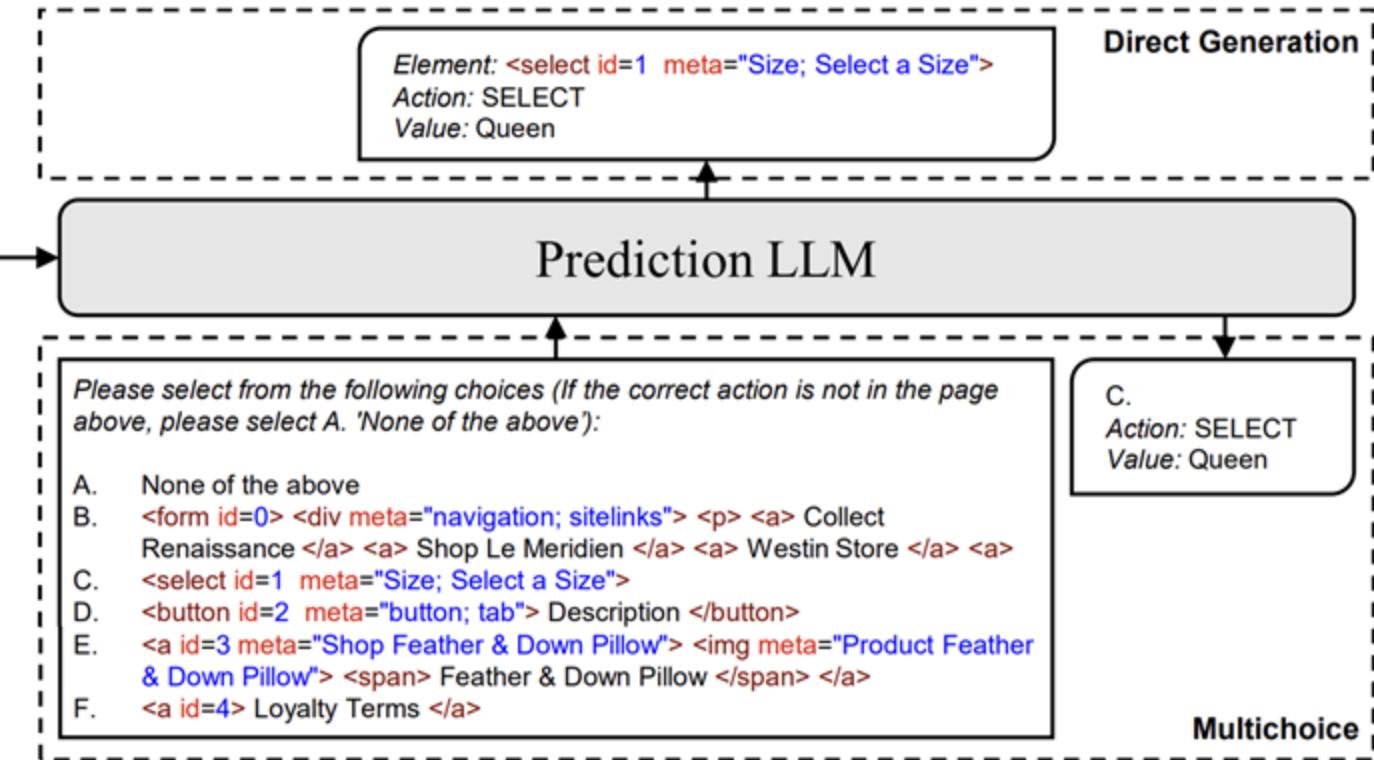
[button] Special Offers -> CLICK

[link] Shop Marriott Opens a new window -> CLICK

[menuitem] category pillows -> CLICK

[span] Pillow Protector -> CLICK

What should be the next action?



Results of MindAct

	Cross-Task				Cross-Website				Cross-Domain			
	Ele. Acc	Op. F1	Step SR	SR	Ele. Acc	Op. F1	Step SR	SR	Ele. Acc	Op. F1	Step SR	SR
Classification	26.8	—	—	—	21.6	—	—	—	24.5	—	—	—
Generation	20.2	52.0	17.5	0.0	13.9	44.7	11.0	0.0	14.2	44.7	11.9	0.4
MINDACT												
w/ Flan-T5 _B	43.6	76.8	41.0	4.0	32.1	67.6	29.5	1.7	33.9	67.3	31.6	1.6
w/ Flan-T5 _L	53.4	75.7	50.3	7.1	39.2	67.1	35.3	1.1	39.7	67.2	37.3	2.7
w/ Flan-T5 _{XL}	55.1	75.7	52.0	5.2	42.0	65.2	38.9	5.1	42.1	66.5	39.6	2.9
w/ GPT-3.5	20.3	56.6	17.4	0.8	19.3	48.8	16.2	0.6	21.6	52.8	18.6	1.0
w/ GPT-4*	41.6	60.6	36.2	2.0	35.8	51.1	30.1	2.0	37.1	46.5	26.4	2.0

Step SR: success rate at each step

SR: Success rate for the whole task

Offline evaluation setting

GPT-4 (3-shot) is close to fine-tuned Flan-T5 models

Results of MindAct

	Cross-Task				Cross-Website				Cross-Domain			
	Ele. Acc	Op. F1	Step SR	SR	Ele. Acc	Op. F1	Step SR	SR	Ele. Acc	Op. F1	Step SR	SR
Classification	26.8	—	—	—	21.6	—	—	—	24.5	—	—	—
Generation	20.2	52.0	17.5	0.0	13.9	44.7	11.0	0.0	14.2	44.7	11.9	0.4
MINDACT												
w/ Flan-T5 _B	43.6	76.8	41.0	4.0	32.1	67.6	29.5	1.7	33.9	67.3	31.6	1.6
w/ Flan-T5 _L	53.4	75.7	50.3	7.1	39.2	67.1	35.3	1.1	39.7	67.2	37.3	2.7
w/ Flan-T5 _{XL}	55.1	75.7	52.0	5.2	42.0	65.2	38.9	5.1	42.1	66.5	39.6	2.9
w/ GPT-3.5	20.3	56.6	17.4	0.8	19.3	48.8	16.2	0.6	21.6	52.8	18.6	1.0
w/ GPT-4*	41.6	60.6	36.2	2.0	35.8	51.1	30.1	2.0	37.1	46.5	26.4	2.0

Step SR: success rate at each step

SR: Success rate for the whole task

Offline evaluation setting

GPT-4 (3-shot) is close to fine-tuned Flan-T5 models, but all models are terrible (2-7% whole task success rate)!

Around six months later ...

Model	Cross-Task			Cross-Website			Cross-Domain					
	Ele.	Acc	Op. F1	Step SR	Ele.	Acc	Op. F1	Step SR	Ele.	Acc	Op. F1	Step SR
Supervised Fine-Tuning												
FLAN-T5-XL	57.1	75.7	53.5	43.8	67.7	41.1	41.4	65.9	38.9			
BLIP-2-T5-XL	50.1	77.0	47.0	39.4	69.3	37.0	41.2	69.3	38.9			
In-Context Learning												
GPT-3.5*	19.4	59.2	16.8	14.9	56.5	14.1	25.2	57.9	24.1			
GPT-4*	40.8	63.1	32.3	30.2	61.0	27.0	35.4	61.9	29.7			
COGAGENT	22.4	53.0	17.6	18.4	42.2	13.4	20.6	42.0	15.5			
?												
	9.7	65.6	8.1	9.1	60.8	7.5	10.9	63.9	8.5			
	21.5	67.7	19.6	17.1	61.3	15.4	20.7	64.3	18.0			
	46.4	73.4	40.2	38.0	67.8	32.4	42.4	69.3	36.8			
	66.4	79.2	61.9	69.5	78.9	65.0	72.8	73.6	62.1			

5~7% gain

?



Around six months later ...

Model	Cross-Task			Cross-Website			Cross-Domain					
	Ele.	Acc	Op. F1	Step SR	Ele.	Acc	Op. F1	Step SR	Ele.	Acc	Op. F1	Step SR
Supervised Fine-Tuning												
FLAN-T5-XL	57.1	75.7	53.5	43.8	67.7	41.1	41.4	65.9	38.9			
BLIP-2-T5-XL	50.1	77.0	47.0	39.4	69.3	37.0	41.2	69.3	38.9			
In-Context Learning												
GPT-3.5*	19.4	59.2	16.8	14.9	56.5	14.1	25.2	57.9	24.1			
GPT-4*	40.8	63.1	32.3	30.2	61.0	27.0	35.4	61.9	29.7			
COGAGENT	22.4	53.0	17.6	18.4	42.2	13.4	20.6	42.0	15.5			
?	9.7	65.6	8.1	9.1	60.8	7.5	10.9	63.9	8.5			
	21.5	67.7	19.6	17.1	61.3	15.4	20.7	64.3	18.0			
	46.4	73.4	40.2	38.0	67.8	32.4	42.4	69.3	36.8			
	66.4	79.2	61.9	69.5	78.9	65.0	72.8	73.6	62.1			

Whole task success rate

	Offline ₀	Offline ₁	Online
FLAN-T5-XL	4.4	24.4	8.9
GPT-4	1.1	12.2	13.3
?	3.3	12.2	37.8
	13.3	27.8	51.1

So, what did we do?

Model	Cross-Task			Cross-Website			Cross-Domain					
	Ele.	Acc	Op. F1	Step SR	Ele.	Acc	Op. F1	Step SR	Ele.	Acc	Op. F1	Step SR
Supervised Fine-Tuning												
FLAN-T5-XL	57.1	75.7	53.5	43.8	67.7	41.1	41.4	65.9	38.9			
BLIP-2-T5-XL	50.1	77.0	47.0	39.4	69.3	37.0	41.2	69.3	38.9			
In-Context Learning												
GPT-3.5*	19.4	59.2	16.8	14.9	56.5	14.1	25.2	57.9	24.1			
GPT-4*	40.8	63.1	32.3	30.2	61.0	27.0	35.4	61.9	29.7			
COGAGENT	22.4	53.0	17.6	18.4	42.2	13.4	20.6	42.0	15.5			
?												
	9.7	65.6	8.1	9.1	60.8	7.5	10.9	63.9	8.5			
	21.5	67.7	19.6	17.1	61.3	15.4	20.7	64.3	18.0			
	46.4	73.4	40.2	38.0	67.8	32.4	42.4	69.3	36.8			
	66.4	79.2	61.9	69.5	78.9	65.0	72.8	73.6	62.1			



	Offline ₀	Offline ₁	Online
FLAN-T5-XL	4.4	24.4	8.9
GPT-4	1.1	12.2	13.3
?	3.3	12.2	37.8
	13.3	27.8	51.1

Whole task success rate

Outline

1

Introduction

2

Environment Perception

3

Planning

4

Self-Improvement

HTML -> Simplified HTML -> Image

**100K
Tokens**

HTML -> Simplified HTML -> Image

```
391 <div backend_node_id="38897">
392   <div backend_node_id="38898">
393     <div backend_node_id="39105">
394       <div backend_node_id="39107">
395         <div backend_node_id="39113" title="AMP Member Discount, Save up to 20%>
396           <img alt="AMP Member Discount logo" backend_node_id="39113" type="image/jpeg"/>
397           <source backend_node_id="39113" type="image/jpeg"/>
398           <img alt="AMP Member Discount, Save up to 20%" backend_node_id="39118" type="image/png"/>
399         </div>
400       </div>
401     </div>
402   </div>
403 </div>
404 <div backend_node_id="39125">
405   <div backend_node_id="39126" title="Save up to 20%>
406     <img alt="Save up to 20% logo" backend_node_id="39126" type="image/png"/>
407   </div>
408 </div>
409 <div backend_node_id="39131" title="Learn More">
410   <div backend_node_id="39132" title="Learn More">
411     <img alt="Learn More logo" backend_node_id="39132" type="image/png"/>
412   </div>
413 </div>
414 <div backend_node_id="39136">
415   <div backend_node_id="39138">
416     <div backend_node_id="39148" title="Don't lift a thing. Get loading help.">
417       <img alt="Don't lift a thing. Get loading help. logo" backend_node_id="39148" type="image/png"/>
418       <source backend_node_id="39148" type="image/webp"/>
419       <source backend_node_id="39148" type="image/jpeg"/>
420       <img alt="Don't lift a thing. Get loading help." backend_node_id="39153" alt="Don't lift a thing. Get loading help." type="image/png"/>
421     </div>
422   </div>
423 </div>
424 <div backend_node_id="39154">
425   <div backend_node_id="39155" title="Compare Prices">
426     <div backend_node_id="39156" title="Compare Prices">
427       <img alt="Compare Prices logo" backend_node_id="39156" type="image/png"/>
428     </div>
429   </div>
430 </div>
431 <div backend_node_id="39163">
432   <div backend_node_id="39167" title="Need a Pick-up Truck?">
433     <div backend_node_id="39171" title="Need a Pick-up Truck?">
434       <img alt="Need a Pick-up Truck? logo" backend_node_id="39171" type="image/png"/>
435       <source backend_node_id="39171" type="image/webp"/>
436       <source backend_node_id="39171" type="image/jpeg"/>
437       <img alt="Need a Pick-up Truck?" backend_node_id="39175" alt="Need a Pick-up Truck?" type="image/png"/>
438     </div>
439   </div>
440 </div>
441 <div backend_node_id="39183">
442   <div backend_node_id="39184" title="In need of a smaller moving truck? Budget.com offers pick-up trucks.">
443     <div backend_node_id="39185" title="Reserve Truck">
444       <div backend_node_id="39186" title="Reserve Truck">
445         <img alt="Reserve Truck logo" backend_node_id="39186" type="image/png"/>
446       </div>
447     </div>
448   </div>
449 </div>
450 <div backend_node_id="39226">
451   <div backend_node_id="39227" name="39226"/>
452   <div backend_node_id="39231" type="button">
453     <div backend_node_id="39232" type="button">
454       <div backend_node_id="39234" type="button">
455         <div backend_node_id="39236" type="button">
456           <div backend_node_id="39238" type="button">
457             <div backend_node_id="39244" type="button">
458               <div backend_node_id="39245" title="Personal Rentals">
459                 <img alt="Personal Rentals logo" backend_node_id="39245" type="image/png"/>
460               </div>
461             </div>
462           </div>
463         </div>
464       </div>
465     </div>
466   </div>
467 </div>
```

```
391 <div backend_node_id="38894" value="18:00 PM">
392   <div backend_node_id="38893" value="18:00 PM">
393     <div backend_node_id="38895" value="18:00 PM">
394       <div backend_node_id="38896" value="18:00 PM">
395         <div backend_node_id="38897" value="18:00 PM">
396           <div backend_node_id="38898" value="18:00 PM">
397             <div backend_node_id="38899" value="18:00 PM">
398               <div backend_node_id="38900" value="18:00 PM">
399                 <div backend_node_id="38901" value="18:00 PM">
400                   <div backend_node_id="38902" value="18:00 PM">
401                     <div backend_node_id="38903" value="18:00 PM">
402                       <div backend_node_id="38904" value="18:00 PM">
403                         <div backend_node_id="38905" value="18:00 PM">
404                           <div backend_node_id="38906" value="18:00 PM">
405                             <div backend_node_id="38907" value="18:00 PM">
406                               <div backend_node_id="38908" value="18:00 PM">
407                                 <div backend_node_id="38909" value="18:00 PM">
408                                   <div backend_node_id="38910" value="18:00 PM">
409                                     <div backend_node_id="38911" value="18:00 PM">
410                                       <div backend_node_id="38912" value="18:00 PM">
411                                         <div backend_node_id="38913" value="18:00 PM">
412                                           <div backend_node_id="38914" value="18:00 PM">
413                                             <div backend_node_id="38915" value="18:00 PM">
414                                               <div backend_node_id="38916" value="18:00 PM">
415                                                 <div backend_node_id="38917" value="18:00 PM">
416                                                   <div backend_node_id="38918" value="18:00 PM">
417                                                     <div backend_node_id="38919" value="18:00 PM">
418                                                       <div backend_node_id="38920" value="18:00 PM">
419                                                         <div backend_node_id="38921" value="18:00 PM">
420               </div>
421             </div>
422           </div>
423         </div>
424       </div>
425     </div>
426   </div>
427 </div>
428 <div backend_node_id="38951">
429   <div backend_node_id="38952">
430     <div backend_node_id="38953" value="Will you return the truck to a different location?">
431       <div backend_node_id="38954" value="Yes">
432         <input backend_node_id="38955" type="radio" name="one-way-radio" value="ON" input_checked="true"/>
433         <span backend_node_id="38956">Yes</span>
434       </div>
435       <div backend_node_id="38957" value="No">
436         <input backend_node_id="38958" type="radio" name="one-way-radio" value="OFF" input_value="OFF"/>
437         <span backend_node_id="38959">No</span>
438       </div>
439     </div>
440   </div>
441 </div>
442 <div backend_node_id="38988" name="divScrl128123ViewSHPPschIsLocal" value="false" input_value="false"/>
443 </div>
444 <div backend_node_id="38989">
445   <div backend_node_id="38990" value="Where do you want to return your truck?">
446     <div backend_node_id="38991" value="US City, State or Zip Code">
447       <input backend_node_id="38991" type="text" placeholder="US City, State or Zip Code"/>
448     </div>
449   </div>
450 </div>
451 <div backend_node_id="39018">
452   <div backend_node_id="39019" value="Flexible Dates">
453     <div backend_node_id="39020" value="Flexible Dates">
454       <div backend_node_id="39021" value="Flexible Dates">
455         <div backend_node_id="39022" value="Flexible Dates">
456           <div backend_node_id="39023" value="Flexible Dates">
457             <div backend_node_id="39024" value="Flexible Dates">
458               <div backend_node_id="39025" value="Flexible Dates">
459                 <div backend_node_id="39026" value="Flexible Dates">
460                   <div backend_node_id="39027" value="Flexible Dates">
461                     <div backend_node_id="39028" value="Flexible Dates">
462                       <div backend_node_id="39029" value="Flexible Dates">
463                         <div backend_node_id="39030" value="Flexible Dates">
464                           <div backend_node_id="39031" value="Flexible Dates">
465                             <div backend_node_id="39032" value="Flexible Dates">
466                               <div backend_node_id="39033" value="Flexible Dates">
467                                 <div backend_node_id="39034" value="Flexible Dates">
468                                   <div backend_node_id="39035" value="Flexible Dates">
469                                     <div backend_node_id="39036" value="Flexible Dates">
470                                       <div backend_node_id="39037" value="Flexible Dates">
471                                         <div backend_node_id="39038" value="Flexible Dates">
472                                           <div backend_node_id="39039" value="Flexible Dates">
473                                             <div backend_node_id="39040" value="Flexible Dates">
474                                               <div backend_node_id="39041" value="Flexible Dates">
475                                                 <div backend_node_id="39042" value="Flexible Dates">
476                                                   <div backend_node_id="39043" value="Flexible Dates">
477                                                     <div backend_node_id="39044" value="Flexible Dates">
478                                                       <div backend_node_id="39045" value="Flexible Dates">
479             </div>
480           </div>
481         </div>
482       </div>
483     </div>
484   </div>
485 </div>
486 <div backend_node_id="39044" value="Add Discount Code">
487   <div backend_node_id="39045" value="Discount Code">
488     <div backend_node_id="39046" value="Discount Code">
489       <div backend_node_id="39047" value="Discount Code">
490         <div backend_node_id="39048" value="Discount Code">
491           <div backend_node_id="39049" value="Discount Code">
492             <div backend_node_id="39050" value="Discount Code">
493               <div backend_node_id="39051" value="Discount Code">
494                 <div backend_node_id="39052" value="Discount Code">
495                   <div backend_node_id="39053" value="Discount Code">
496                     <div backend_node_id="39054" value="Discount Code">
497                       <div backend_node_id="39055" value="Discount Code">
498                         <div backend_node_id="39056" value="Discount Code">
499                           <div backend_node_id="39057" value="Discount Code">
500                             <div backend_node_id="39058" value="Discount Code">
501                               <div backend_node_id="39059" value="Discount Code">
502                                 <div backend_node_id="39060" value="Discount Code">
503                                   <div backend_node_id="39061" value="Discount Code">
504                                     <div backend_node_id="39062" value="Discount Code">
505                                       <div backend_node_id="39063" value="Discount Code">
506                                         <div backend_node_id="39064" value="Discount Code">
507                                           <div backend_node_id="39065" value="Discount Code">
508                                             <div backend_node_id="39066" value="Discount Code">
509                                               <div backend_node_id="39067" value="Discount Code">
510                                                 <div backend_node_id="39068" value="Discount Code">
511                                                   <div backend_node_id="39069" value="Discount Code">
512                                                     <div backend_node_id="39070" value="Discount Code">
513                                                       <div backend_node_id="39071" value="Discount Code">
514             </div>
515           </div>
516         </div>
517       </div>
518     </div>
519   </div>
520 </div>
```

HTML -> Simplified HTML -> Image

Budget®
Truck Rental

Search

Existing Reservations Moving Trucks & Accessories Locations Deals Moving Resources Commercial Truck Rental

Where do you want to pick up your truck?
08817

When do you want to pick up your truck?
mm/dd/yyyy

Select Pick-up Date

APRIL 2023 MAY 2023

S	M	T	W	T	F	S
26	27	28	29	30	31	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	1	2	3	4	5	6

Jump to a new date: April 2023

NOW INCLUDING CARGO VANS!

— A TRUCK LOAD OF SAVINGS —

20% OFF BASE RATES

Budget®

Apply Discount Code 20DIS

Site feedback

Truck rental that fits your budget

Member Benefits AARP

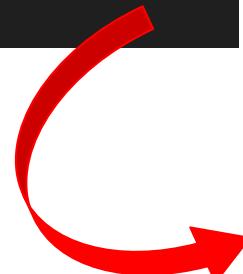
Save up to 20%

HIRE A HELPER

Don't lift a thing. Get loading help.

In need of a smaller moving truck?
Budget.com offers pick-up trucks.

Image: A picture is worth thousands of words



The screenshot shows the homepage of Budget Truck Rental. At the top, there's a search bar with a magnifying glass icon. The main navigation menu includes "Existing Reservations", "Moving Trucks & Accessories", "Locations", "Deals", "Moving Resources", and "Commercial Truck Rental". A banner on the left side allows users to enter pickup and drop-off locations and dates. Another banner in the center promotes "CARGO VANS" and a 20% discount. Below these are two large white vans. A "Member Benefits" section for AARP is on the left, and "HIRE A HELPER" services are in the middle. A pickup truck is shown on the right.

SeeAct Paradigm

Action Generation -> Action Grounding

Rent a truck near zip 08817 on December 10 at 11:30AM returned to the exact location and date.

LMM



Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform a click action.

Grounding

Element: <input id=19 button find your truck />

Operation: CLICK

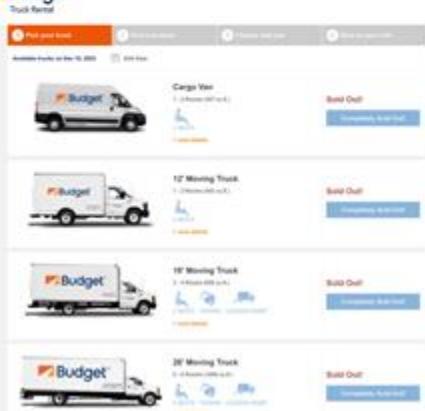


SeeAct



Browser Event

Budget



SeeAct Paradigm

Action Generation -> Action Grounding

Rent a truck near zip 08817 on December 10 at 11:30AM returned to the exact location and date.

LMM



Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform a click action.

Grounding

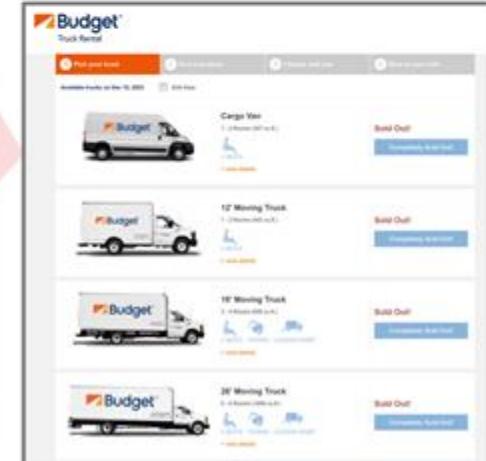
Element: <input id=19 button find your truck />
Operation: CLICK



SeeAct



Browser Event



SeeAct Paradigm

Action Generation -> Action Grounding

Action Description

Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform a click action.



A: <text id=0>Moving Trucks & Accessories</text>
B: <div id=1>Where do you want to pick up your truck?</div>
C: <div id=5>When do you want to pick up your truck?</div>
...
F: <input id=18>No value=true</input>
G: <input id=19 button "find your truck" />
H: None

Element Attributes

TEXT: Find Your Truck
TYPE: BUTTON

A: Where do you want to pick up your truck?
B: 08817
C: 12/10/2023
D: 12/10/2023
E: 11:30 AM
F: Yes
G: Find Your Truck

Image Annotation

CHOICE: G

Textual Choices

CHOICE: G

SeeAct Paradigm

Action Generation -> Action Grounding

Action Description

Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform a click action.



A: <text id=0>Moving Trucks & Accessories</text>
B: <div id=1>Where do you want to pick up your truck?</div>
C: <div id=5>When do you want to pick up your truck?</div>
...
F: <input id=18>No value=true</input>
G: <input id=19 button "find your truck" />
H: None

Element Attributes

TEXT: Find Your Truck
TYPE: BUTTON

A: Where do you want to pick up your truck?
B: 08817
C: 12/10/2023
D: 12/10/2023
E: 11:30 AM
F: Yes
G: Find Your Truck

Image Annotation

CHOICE: G

Textual Choices

CHOICE: G

SeeAct Paradigm

Action Generation -> Action Grounding

Action Description

Move the cursor over the "Find Your Truck" button located in the central portion of the webpage, just below the input fields for rental details, and perform a click action.



A: <text id=0>Moving Trucks & Accessories</text>
B: <div id=1>Where do you want to pick up your truck?</div>
C: <div id=5>When do you want to pick up your truck?</div>
...
F: <input id=18>No value=true</input>
G: <input id=19 button "find your truck" />
H: None

Element Attributes

TEXT: Find Your Truck
TYPE: BUTTON

A: Where do you want to pick up your truck?
B: 08817
C: 12/10/2023
D: 12/10/2023
E: 11:30 AM
F:
G: Find Your Truck

Image Annotation

CHOICE: G

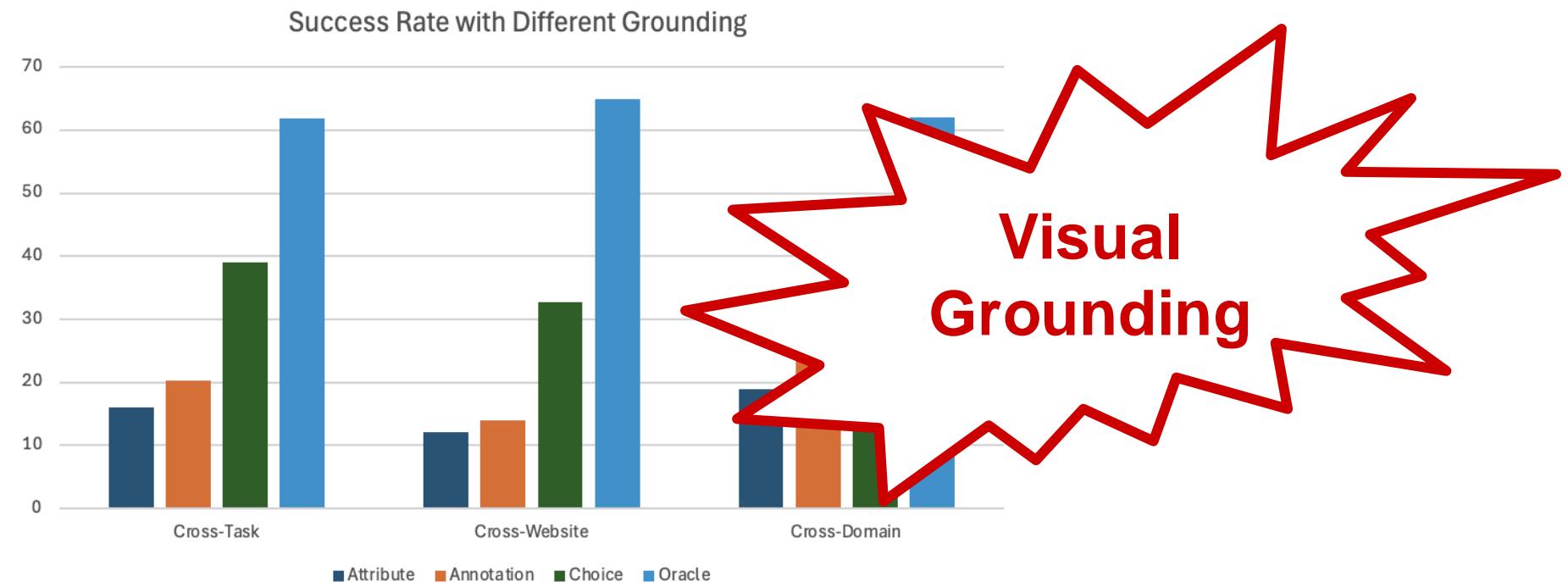
Textual Choices

CHOICE: G

Grounding Strategies

Oracle: Ask humans to identify model's intended actions from Action Description

Large Margin between Oracle and SOTA Grounding method so far



Single Step Success Rate on Mind2Web

Human-like Vision-centric Embodiment is the Future for Web/GUI Agents

- Most comprehensive evaluation of GUI Agents to date
- SeeAct-V + UGround outperforms prior SoTA despite its minimalist design

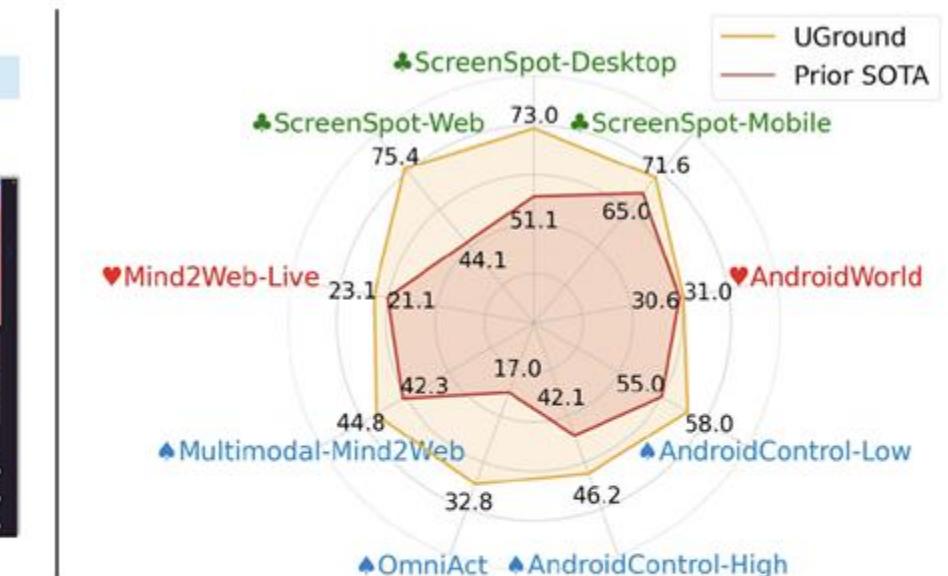
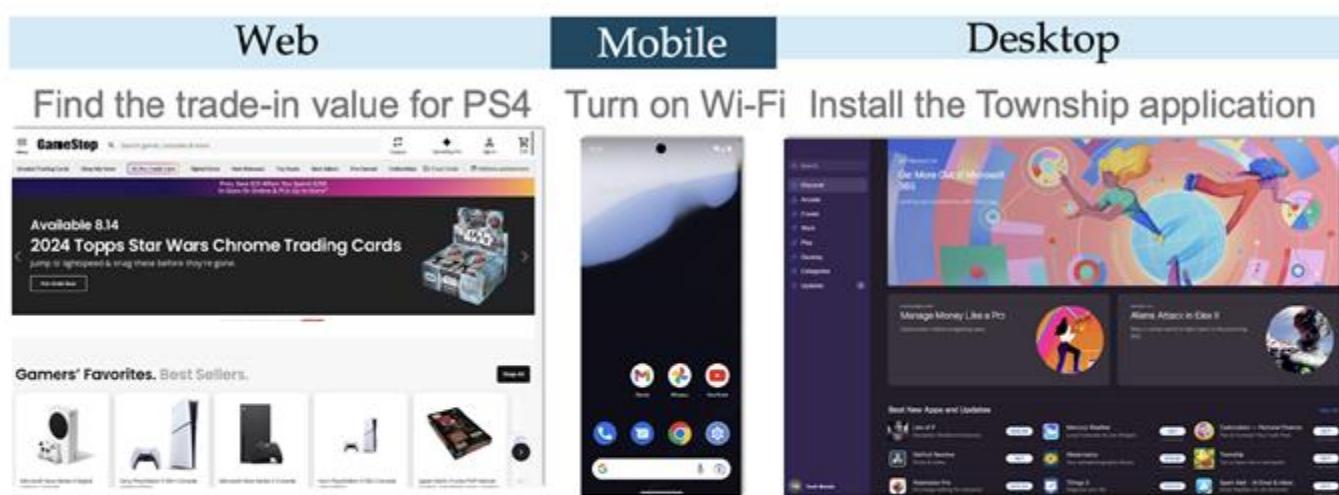


Figure 1: Examples of agent tasks across platforms and performance on **GUI grounding** (♣: ScreenSpot), **offline agent** (♠: Multimodal-Mind2Web, AndroidControl, and OmniAct), and **online agent benchmarks** (♥: Mind2Web-Live and AndroidWorld) when using GPT-4 as the planner.

online_demo — node + python demo_text_choice.py — 80x69
geminigby@MacbookAir15 online_demo % ./venv/bin/python demo_text_choice.py
Website: https://apple.com
Task: Compare iPhone 15 Pro Max with iPhone 13 Pro Max

Online evaluation tool that allows SeeAct to interact with live websites

Online Evaluation

	Offline ₀	Offline ₁	Online
FLAN-T5-XL	4.4	24.4	8.9
GPT-4	1.1	12.2	13.3
SEEACT _{Choice}	3.3	12.2	37.8
SEEACT _{Oracle}	13.3	27.8	51.1

Whole task success rate (%)

SeeAct Codebase

An interface between Agent and Website

- Perception
- Action Execution

```
import asyncio
import os
from seeact.agent import SeeActAgent

# Setup your API Key here, or pass through environment
os.environ["OPENAI_API_KEY"] = "Your API KEY Here"

async def run_agent():
    agent = SeeActAgent(model="gpt-4-turbo")
    await agent.start()
    while not agent.complete_flag:
        prediction_dict = await agent.predict()
        await agent.execute(prediction_dict)
    await agent.stop()

if __name__ == "__main__":
    asyncio.run(run_agent())
```

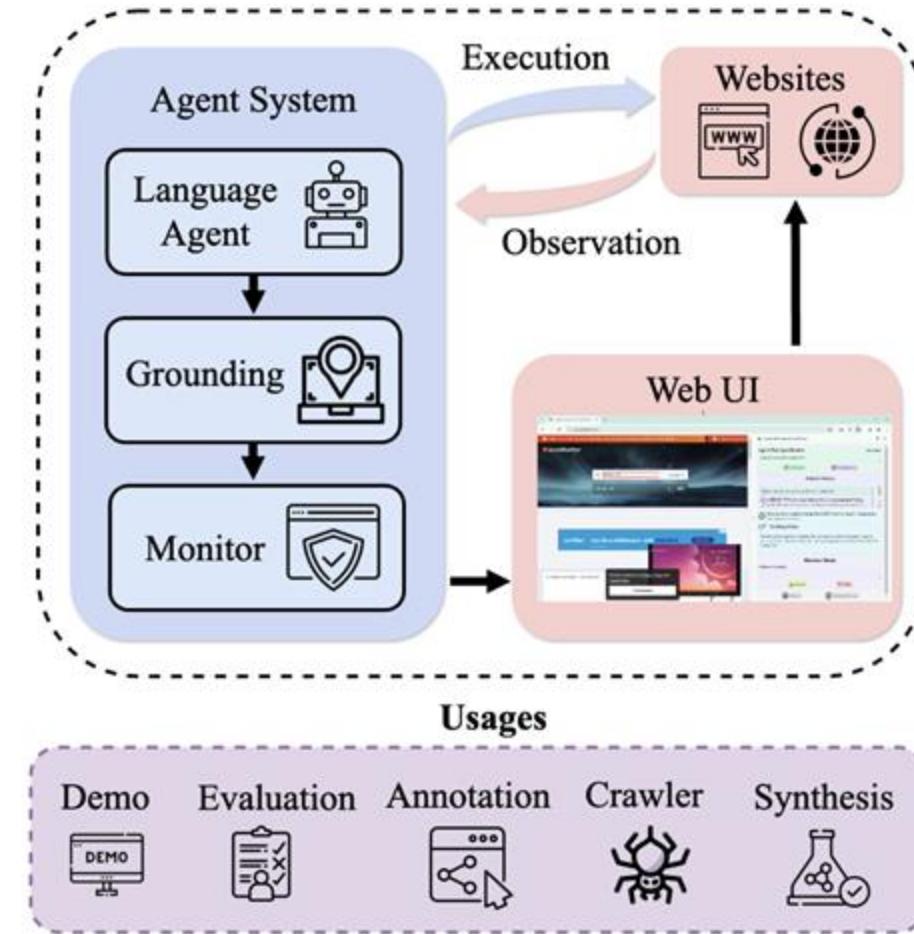
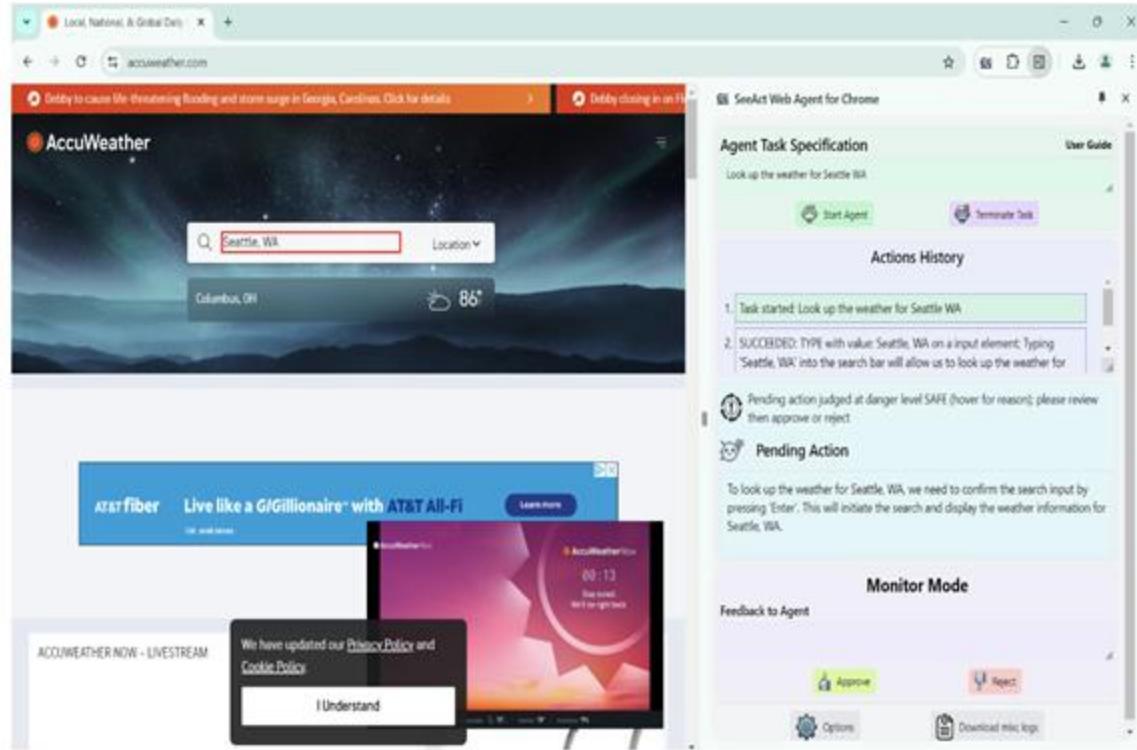
The screenshot shows a web browser window displaying the Apple website. The main content is a product page for the iPhone 15 Pro Max in Titanium. The sidebar contains several panels of text, likely representing the state of the SeeAct agent's perception and action planning process.

- ELEMENTS:** A list of UI elements identified by the agent.
- ACTION: CLICK:** A section showing the current target element for a click action.
- PREVIOUS ACTION ANALYSIS:** A log of previous actions taken by the agent.
- SCREENSHOT DETAILS ANALYSIS:** An analysis of the current screenshot, identifying links under headings like "Shop and Learn," "Accessories," "Apple Store," "For Business," "Apple Values," etc.
- NEXT ACTION BASED ON WEBSITE ANALYSIS:** A suggestion to compare the iPhone 15 Pro Max with the iPhone 13 Pro Max.
- QUESTION:** A multiple-choice question asking which element matches the target element.
- RECOMMENDING OUTPUT:** A section showing the recommended action.
- WEBSITE STATE:** A summary of the current state of the website being analyzed.

At the bottom, there are two buttons: "Action: Navigate to the 'iPhone' category" and "Task: Compare iPhone 15 Pro Max with iPhone 13 Pro Max".

WebOlympus: An Ecosystem for Web Agent

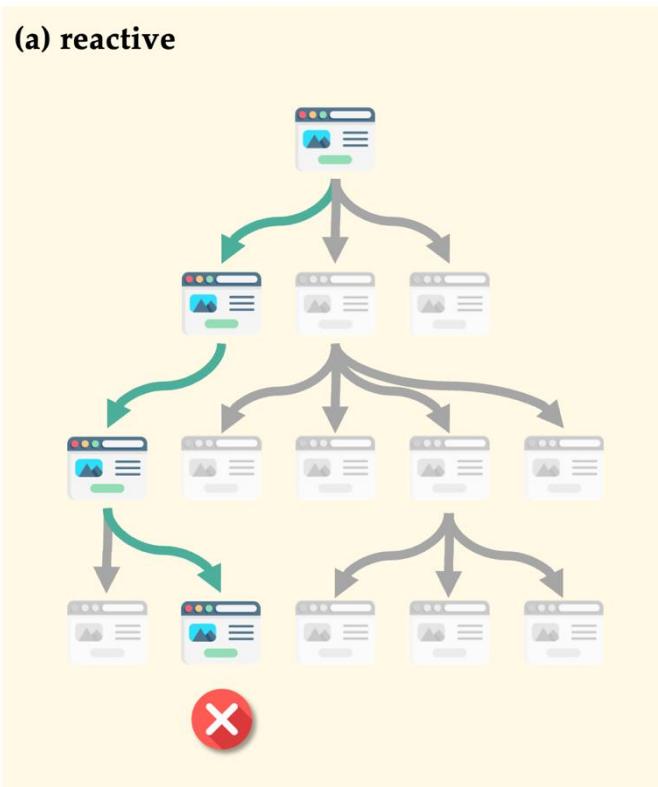
An ecosystem for web agent



Outline

- 1 Introduction
- 2 Environment Perception
- 3 Planning
- 4 WebDreamer: Model-based Planning for Web Agent

Planning Paradigms for Language Agents

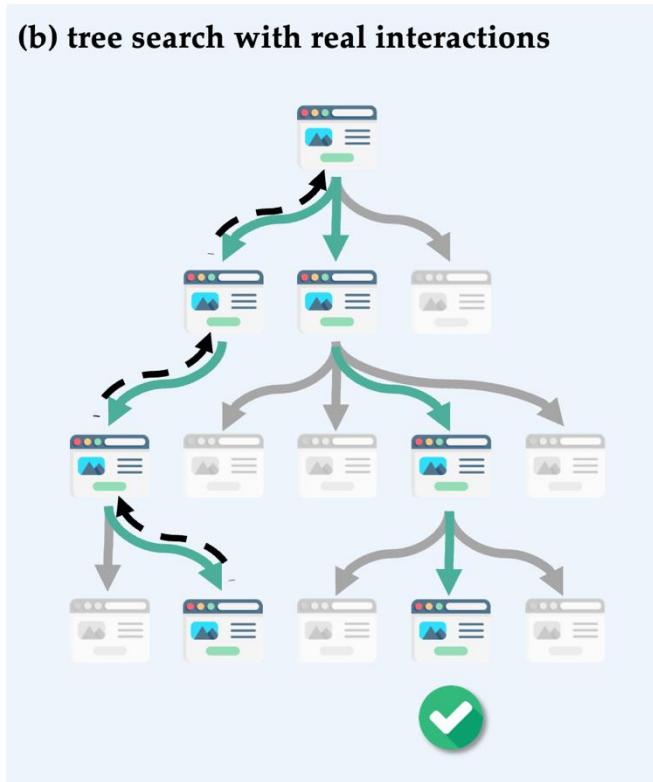
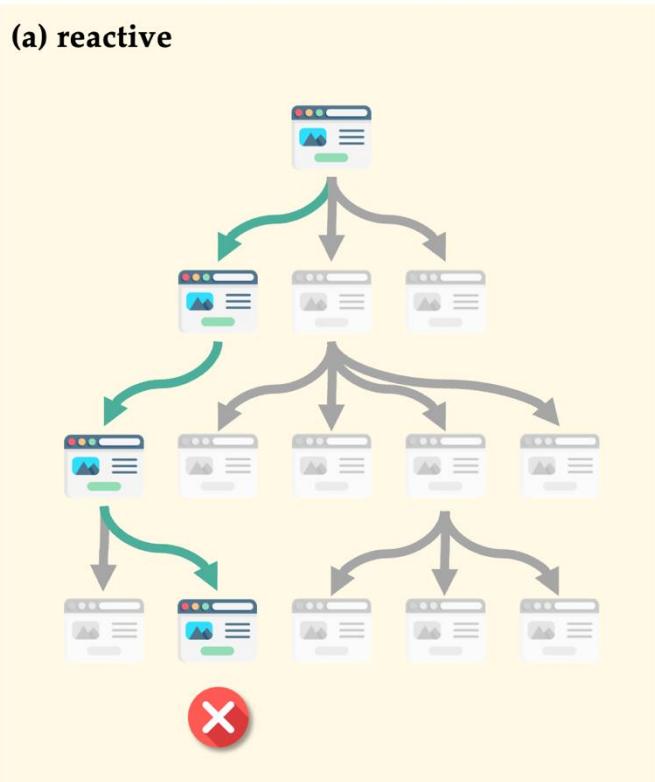


fast, easy to implement



greedy, short-sighted

Planning Paradigms for Language Agents



fast, easy to implement



greedy, short-sighted

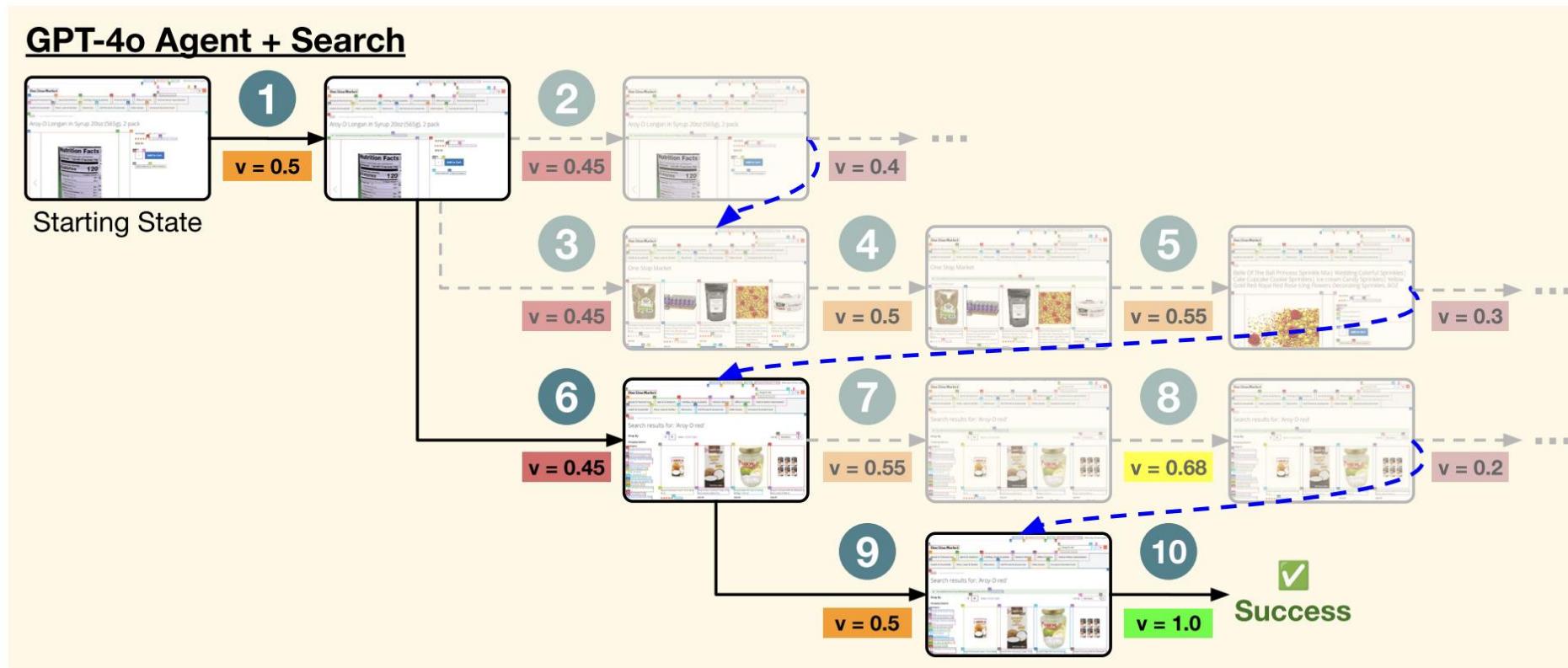


systematic exploration



irreversible actions,
unsafe, slow

Planning Paradigms for Language Agents



- State-Changing Action

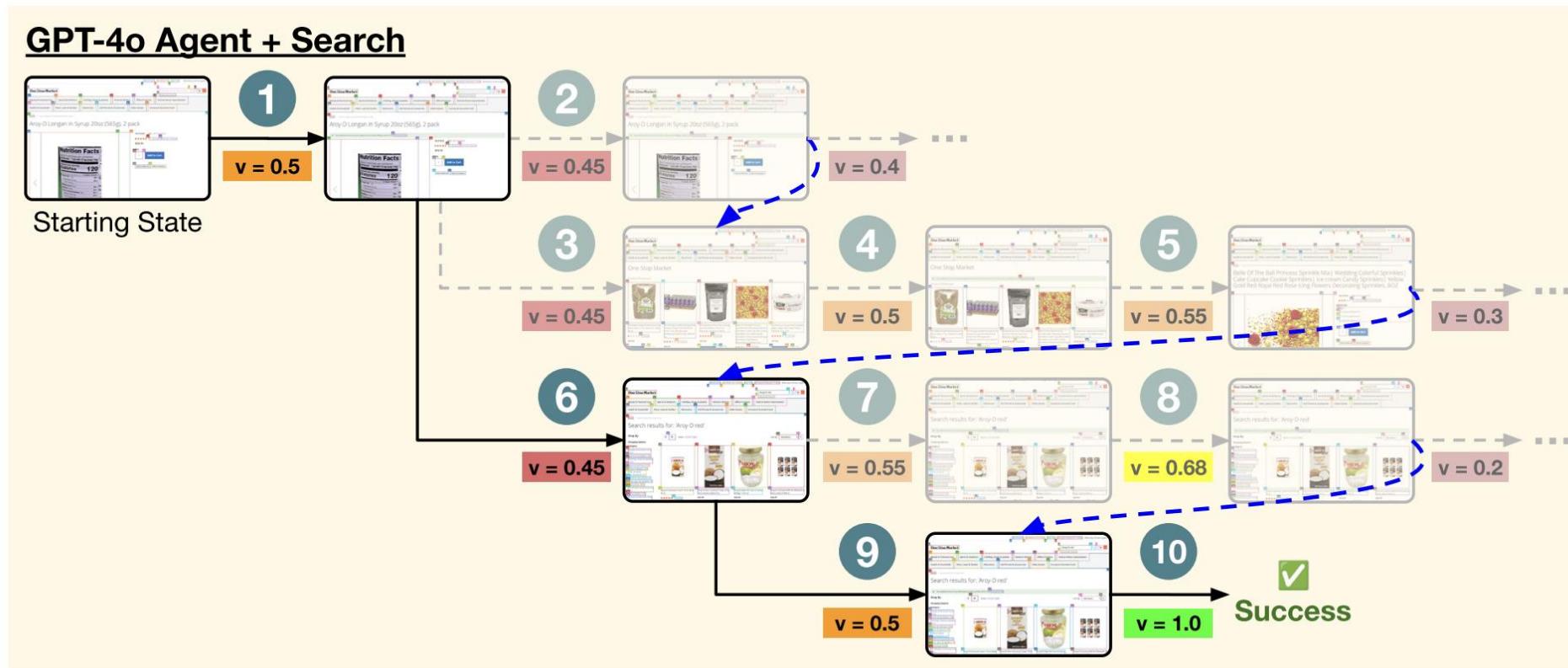
State-Changing Actions

The image displays two web pages side-by-side, each featuring a large green or red text overlay indicating the level of state change required for a specific action.

Tesla Demo Drive Model 3 (Left): This page shows a form for scheduling a demo drive. A green box highlights the "SAFE" state for selecting a location from a dropdown menu. The dropdown includes options like Gilroy-500 Automall Dr, Gilroy CA and Monterey-Seaside Seaside CA, with the latter being highlighted by a green border. Other fields include date (August 4, 2024) and time (10:00 AM). A yellow box highlights the "Schedule Demo Drive" button at the bottom.

Avis Address & Phone Number (Right): This page shows a form for saving addresses and phone numbers. A red box highlights the "HIGH" state for entering the primary address (281 W Lane Ave, Columbus). Below it, a blue box highlights the "LOW" state for accepting privacy choices, with a red border around the "Accept All Cookies" button.

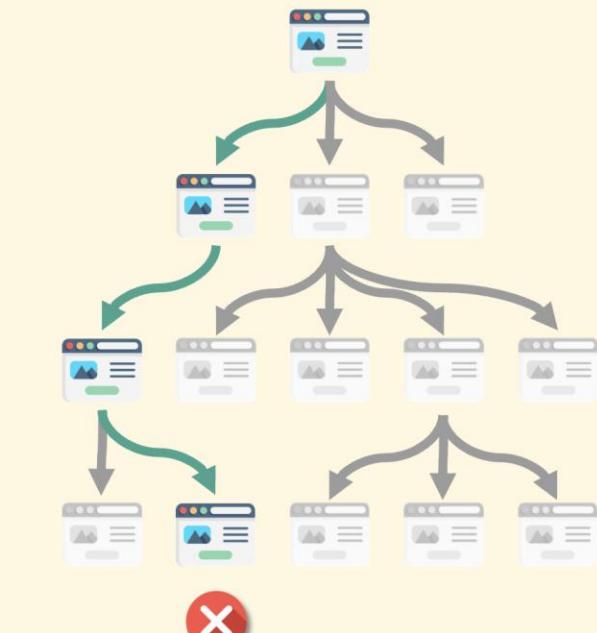
Planning Paradigms for Language Agents



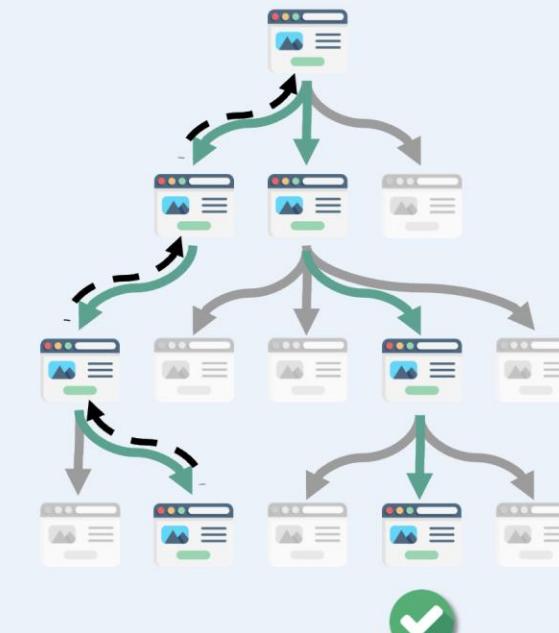
- State-Changing Action
- Tracing Back?
- Limited Efficiency

Planning Paradigms for Language Agents

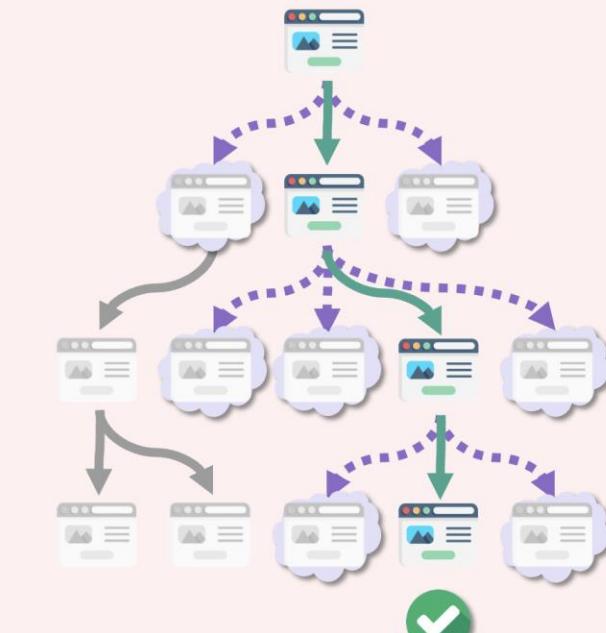
(a) reactive



(b) tree search with real interactions



(c) model-based planning



fast, easy to implement



greedy, short-sighted



systematic exploration



irreversible actions,
unsafe, slow



faster, safer,
systematic exploration



how to get a world model

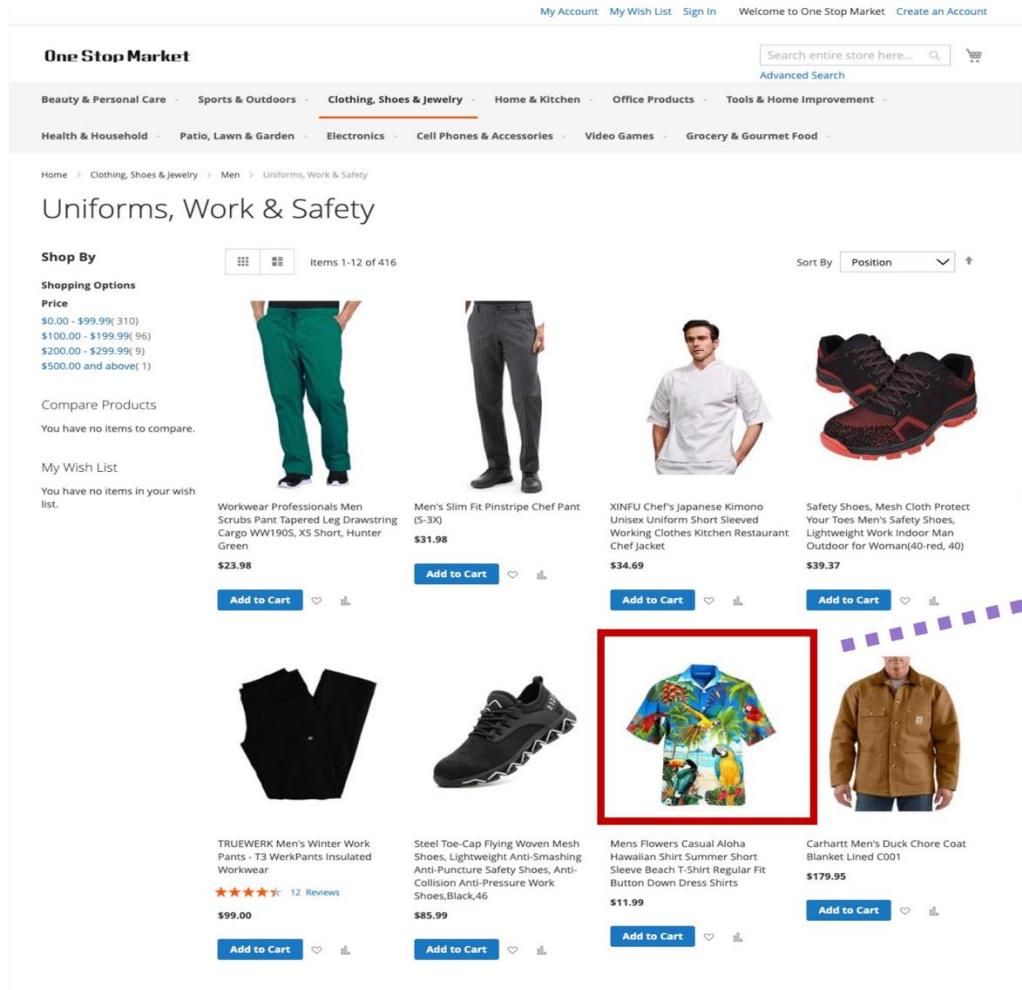
World Model

A model capable of predicting environment transition:

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

If I do this (a_t) on (s_t), what would happen next (s_{t+1})?

LLM can predict state transitions(reasonably good)



The page will navigate to a detailed product page for the "Mens Flowers Casual Aloha Hawaiian Shirt Summer Short Sleeve Beach T-Shirt Regular Fit Button Down Dress Shirts." This new page will likely contain additional information about the product including more detailed specifications, customer reviews, larger images, sizing options, and possibly a larger "Add to Cart" button. Other elements from the current category view like the grid of products will be replaced with the detailed view of this specific product.

WebDreamer: model-based planning for web agents

Please navigate to the 'Data Storage' category and purchase the least expensive disk with 512GB of storage.

The screenshot shows a web browser displaying the One Stop Market website. The URL is 'http://www.onestopmarket.com'. The page title is 'One Stop Market'. The navigation bar includes links for 'My Account', 'My Wish List', 'Sign in', 'Welcome to One Stop Market', and 'Create an Account'. The main menu categories are 'Beauty & Personal Care', 'Sports & Outdoors', 'Clothing, Shoes & Jewelry', 'Home & Kitchen', 'Office Products' (which is highlighted with a red box), 'Tools & Home Improvement', 'Health & Household', 'Patio, Lawn & Garden', and 'Electronics'. A search bar at the top right contains the placeholder text 'Search entire store here...'. A red arrow points to this search bar with the instruction '③ Type "Disk"'. Below the search bar, a red box highlights the 'Office Products' link with the instruction '① Click "Office Products"'. Another red arrow points to the 'Electronics' link in the main menu with the instruction '② Click "Electronics"'. The main content area displays various product categories like 'Product Showcases' and 'Food & Beverage'. A search results section for 'Disk' is shown, with the first item being 'Western Digital WD Blue 500GB 2.5" SATA III Internal Hard Drive'.

One Stop Market

My Account My Wish List Sign in Welcome to One Stop Market Create an Account

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

③ Type 'Disk'

① Click 'Office Products'

② Click 'Electronics'

Product Showcases

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

V8+Energy Healthy Energy Drink, Steady Energy from Black and Green Tea, Pomegranate
16-Ounce Pouch
★ ★ ★ ★ 12 Reviews
\$14.47 Add to Cart

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

Belle Of The Ball Princess Sprinkle Mix [Wedding Sprinkles] Cupcake Sprinkles
Ice cream Candy Sprinkles! Yellow Gold Red Royal Rose King Flowers Decorating Sprinkles, BOZ
★ ★ ★ ★ 12 Reviews
\$23.50 Add to Cart

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

Chengjung Sweet Potato Starch Powder 500g, 2ea(Sweet Potato 65%, Corn 45%)
\$34.00 Add to Cart

Q Mixers Premium Ginger Ale: Real Ingredients & Less Sweet, 6.7 fl Oz (24 Bottles)
★ ★ ★ ★ 12 Reviews
\$68.50 Add to Cart

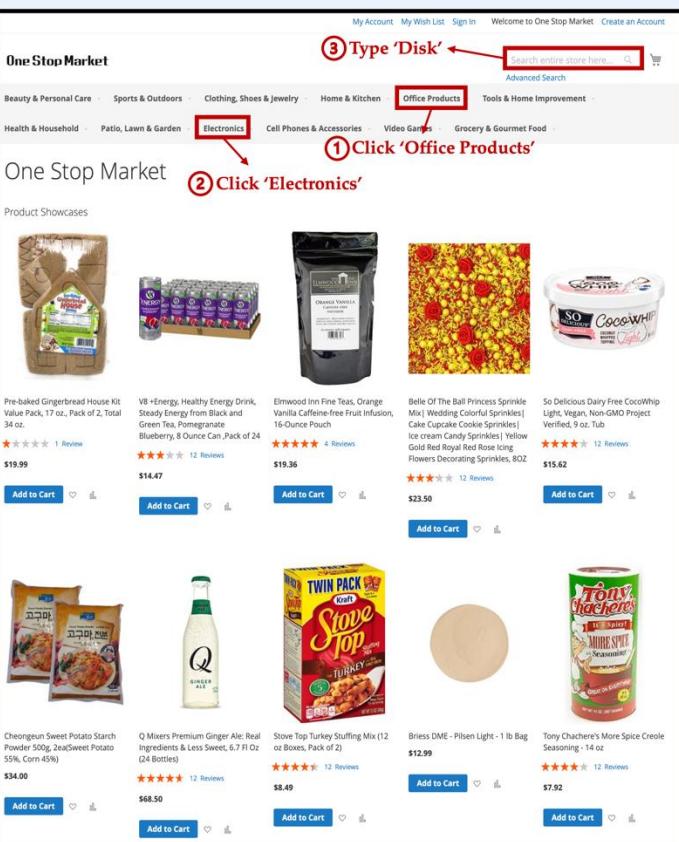
Stove Top Turkey Stuffing Mix (12 oz Boxes, Pack of 2)
\$8.49 Add to Cart

Briess DME - Pilsen Light - 1 lb Bag
\$12.99 Add to Cart

Tony Chachere's More Spice Creole Seasoning - 14 oz
★ ★ ★ ★ 12 Reviews
\$7.52 Add to Cart

WebDreamer: model-based planning for web agents

Please navigate to the 'Data Storage' category and purchase the least expensive disk with 512GB of storage.



The screenshot shows a search interface with a search bar containing 'Search entire store here...'. A red box highlights the search bar, and a red arrow points to it with the instruction '③ Type 'Disk''. Above the search bar, there is a navigation bar with links: My Account, My Wish List, Sign In, Welcome to One Stop Market, and Create an Account. Below the search bar, there is a main menu with categories: Beauty & Personal Care, Sports & Outdoors, Clothing, Shoes & Jewelry, Home & Kitchen, Office Products (highlighted with a red box), Tools & Home Improvement, Health & Household, Patio, Lawn & Garden, Electronics (highlighted with a red box), Cell Phones & Accessories, Video Games, and Grocery & Gourmet Food. Below the menu, there is a section titled 'One Stop Market' with a heading 'Product Showcases'. It displays several products with their names, prices, and 'Add to Cart' buttons. Red boxes and arrows indicate steps: '① Click 'Office Products'' points to the 'Office Products' link in the menu; '② Click 'Electronics'' points to the 'Electronics' link in the menu; and '③ Type 'Disk'' points to the search bar.

Stage I: Simulation

1 The 'Office Products' category will display three sub-categories: 'Office Electronics', 'Office & School Supplies', and 'Office Furniture & Lighting'.

WebDreamer: model-based planning for web agents

Please navigate to the 'Data Storage' category and purchase the least expensive disk with 512GB of storage.

One Stop Market

Product Showcases

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

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

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

Belle Of The Ball Princess Sprinkle
Mix [With Edible Cupcake Sprinkles]
Ice cream Candy Sprinkles | Yellow
Ice Gold Red Royal Rose Icing
Flowers Decorating Sprinkles, 8OZ
★★★★★ 12 Reviews
\$15.62

So Delicious Dairy Free CocoWhip
Light, Vegan, Non-GMO Project
Verified, 9 oz. Tub

Cheongeon Sweet Potato Starch
Flour 500g, 2ea(Sweet Potato
55%, Corn 45%)
\$34.00

Q Mixers Premium Ginger Ale: Real
Ingredients & Less Sweet, 6.7 Fl Oz
(24 Bottles)
★★★★★ 12 Reviews
\$68.50

Kraft Stove Top Turkey Stuffing Mix (12
Box, Pack of 2)
★★★★★ 12 Reviews
\$8.49

Briss DME - Pilsen Light - 1 lb Bag
\$12.99

Tony Chachere's More Spice Creole
Seasoning - 14 oz
★★★★★ 12 Reviews
\$7.92

Stage I: Simulation

1 Click 'Office Products'

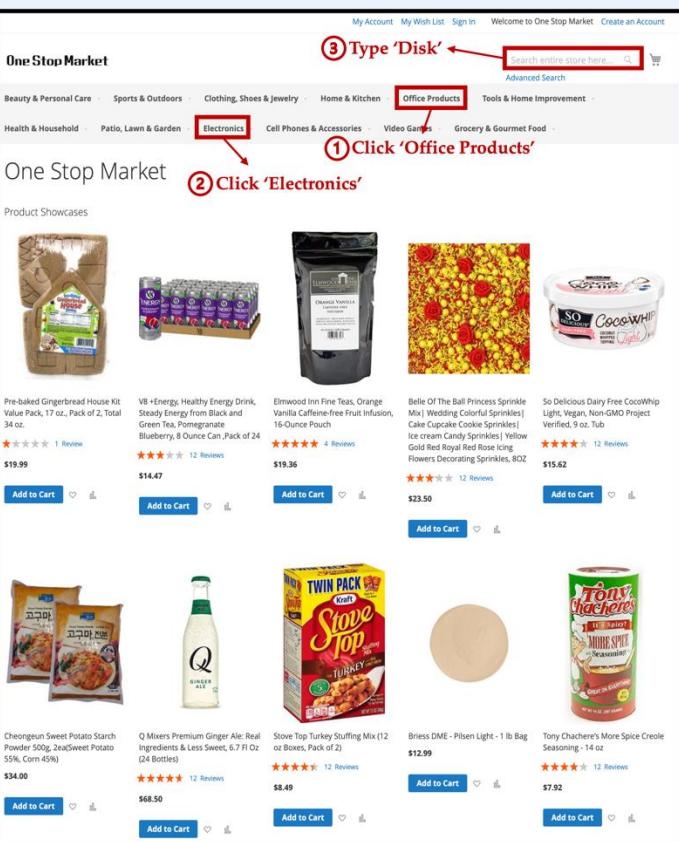
The 'Office Products' category will display three sub-categories: 'Office Electronics', 'Office & School Supplies', and 'Office Furniture & Lighting'.

Click 'Office Electronics'

The webpage will display 'Office Electronics' sub-category results with products, and the sub-menu will show 'Printers&Accessories' and other categories.

WebDreamer: model-based planning for web agents

Please navigate to the 'Data Storage' category and purchase the least expensive disk with 512GB of storage.



One Stop Market

My Account My Wish List Sign In Welcome to One Stop Market Create an Account

One Stop Market

Product Showcases

1 Click 'Office Products'

2 Click 'Electronics'

3 Type 'Disk'

Office Products

Search results 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

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

V8 +Energy Healthy Energy Drink, Steady Energy from Black and Green Tea, Pomegranate ★★★★☆ 12 Reviews \$14.47 Add to Cart

Elmwood Inn Fine Teas, Orange Vanilla Caffeine-free Fruit Infusion, 5 Ounce Can 16-Ounce Pouch \$19.36 Add to Cart

Bell-O Of The Ball Princess Sprinkle Mix | Wafer & Colorful Sprinkles | Ice cream Candy Sprinkles | Yellow Gold Red Royal Rose Icing, Flowers Decorating Sprinkles, 8oz ★★★★☆ 12 Reviews \$23.50 Add to Cart

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

Chongju Sweet Potato Starch Powder 500g, 2ea(Sweet Potato 5%, Corn 45%) \$34.00 Add to Cart

Q Mixers Premium Ginger Ale, Real Ingredients & Less Sweet, 6.7 fl oz (24 Bottles) ★★★★☆ 12 Reviews \$68.50 Add to Cart

Stove Top Turkey Stuffing Mix (12 oz Boxes, Pack of 2) \$8.49 Add to Cart

Briess DME - Pilsen Light - 1 lb Bag \$12.99 Add to Cart

Tony Chachere's More Spice Creole Seasoning - 14 oz ★★★★☆ 12 Reviews \$7.92 Add to Cart

Stage I: Simulation

1 The 'Office Products' category will display three sub-categories: 'Office Electronics', 'Office & School Supplies', and 'Office Furniture & Lighting'.

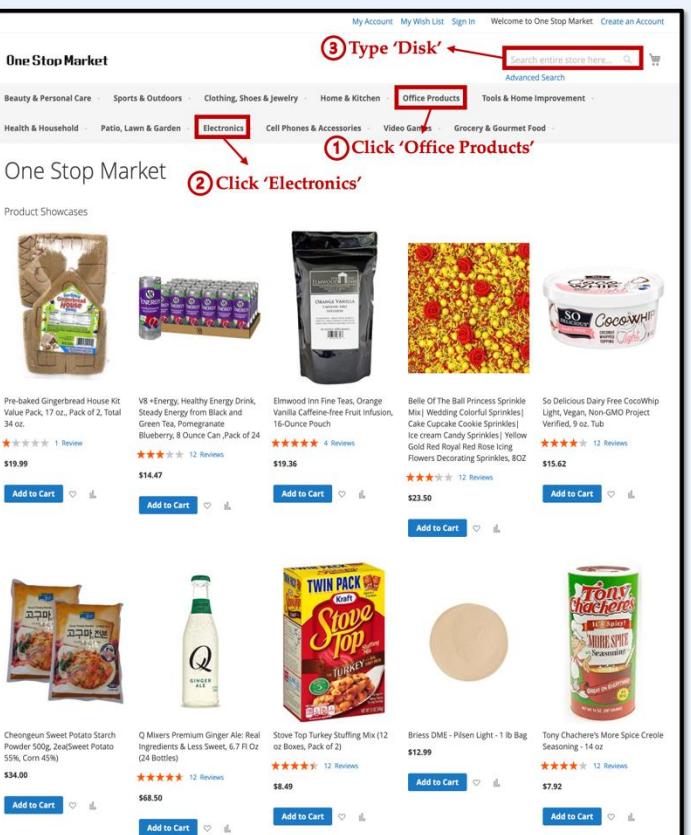
Click 'Office Electronics'

The webpage will display 'Office Electronics' sub-category results with products, and the sub-menu will show 'Printers&Accessories' and other categories.

v = 0.4

WebDreamer: model-based planning for web agents

Please navigate to the 'Data Storage' category and purchase the least expensive disk with 512GB of storage.

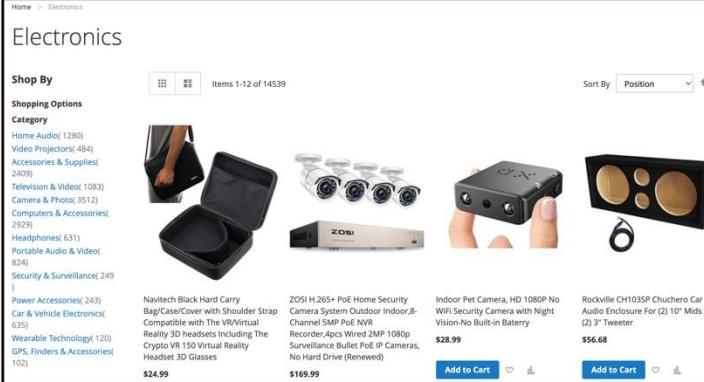


The screenshot shows a shopping website with various product categories like Beauty & Personal Care, Sports & Outdoors, Clothing, Shoes & Jewelry, Home & Kitchen, Office Products, Tools & Home Improvement, Health & Household, and more. A search bar at the top has the placeholder 'Search products here...'. A red box highlights the 'Office Products' link, labeled ① Click 'Office Products'. Another red box highlights the search bar, labeled ③ Type 'Disk'. Below the search bar, a red box highlights the 'Add to Cart' button for a product.

Stage I: Simulation

1. Click 'Office Electronics'
The 'Office Products' category will display three sub-categories: 'Office Electronics', 'Office & School Supplies', and 'Office Furniture & Lighting'. v = 0.4
2. Click 'Computer & Accessories'
The 'Electronics' category will display three sub-categories: 'Computers & Accessories', 'Accessories & Supplies', and 'Car & Vehicle Electronics'. v = 0.8
3. Click 'Electronics'
The webpage will display search results, including a list of products, each of which includes the product title, price, and an 'Add to Cart' button. v = 0.1

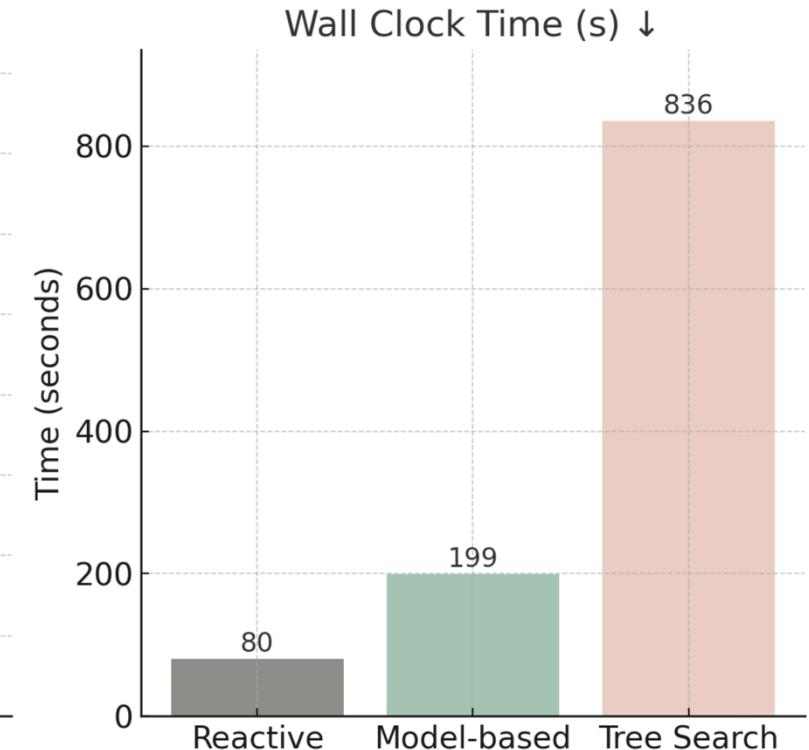
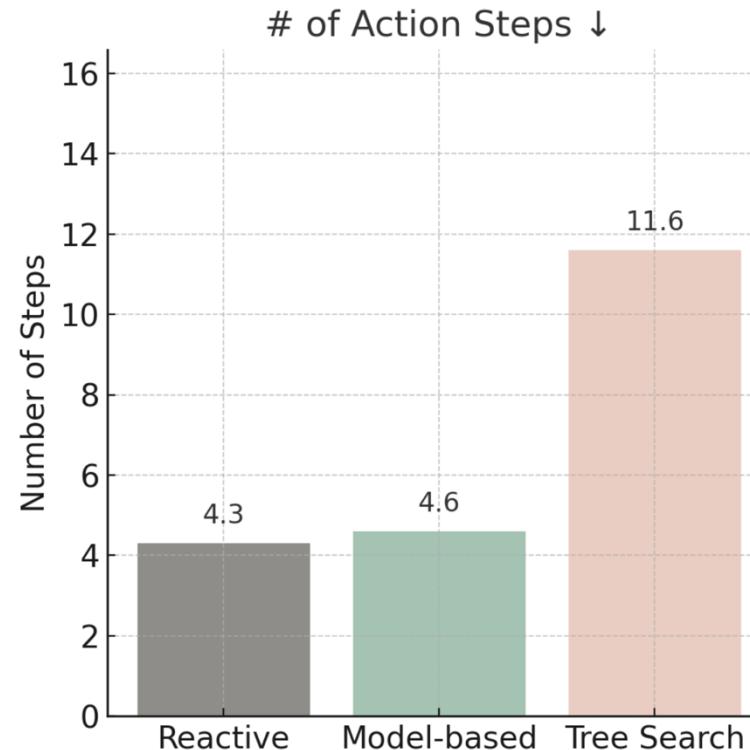
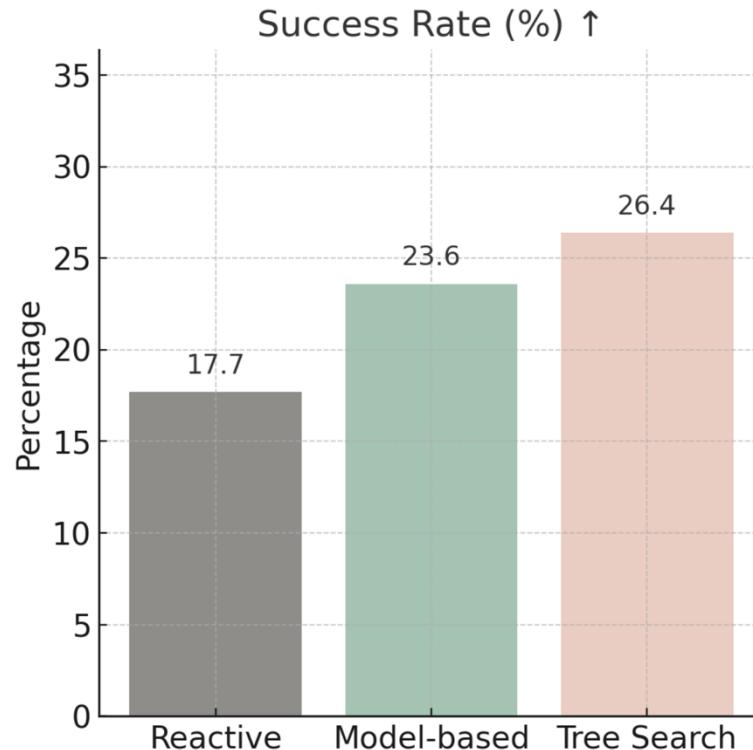
Stage II: Execution



The screenshot shows the 'Electronics' category page with various sub-categories listed on the left. A green arrow points from Stage II to the page, and a green box highlights the 'Add to Cart' button for a product.

Results on VisualWebArena

Model-based planning is more accurate than reactive planning and more efficient than tree search



Outline

1

Introduction

2

Environment Perception

3

Planning

4

Self-Improvement

What's next step?

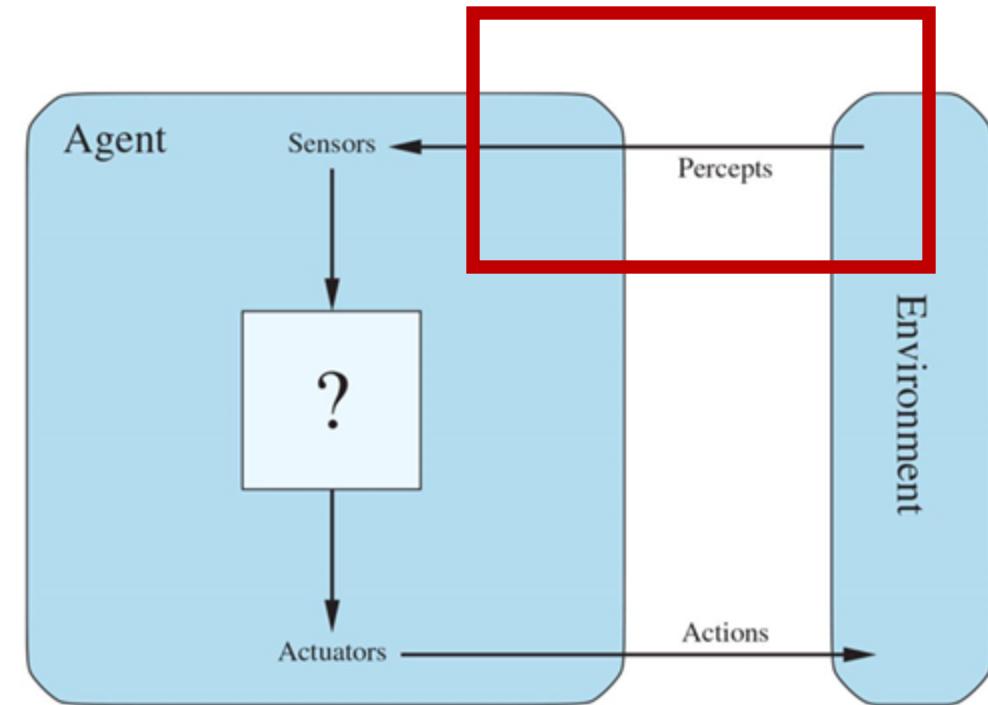
Where have we arrived at?

Perception

Perception:

HTML/DOM: MindAct

Screenshot: SeeAct, WebGUM, CogAgent

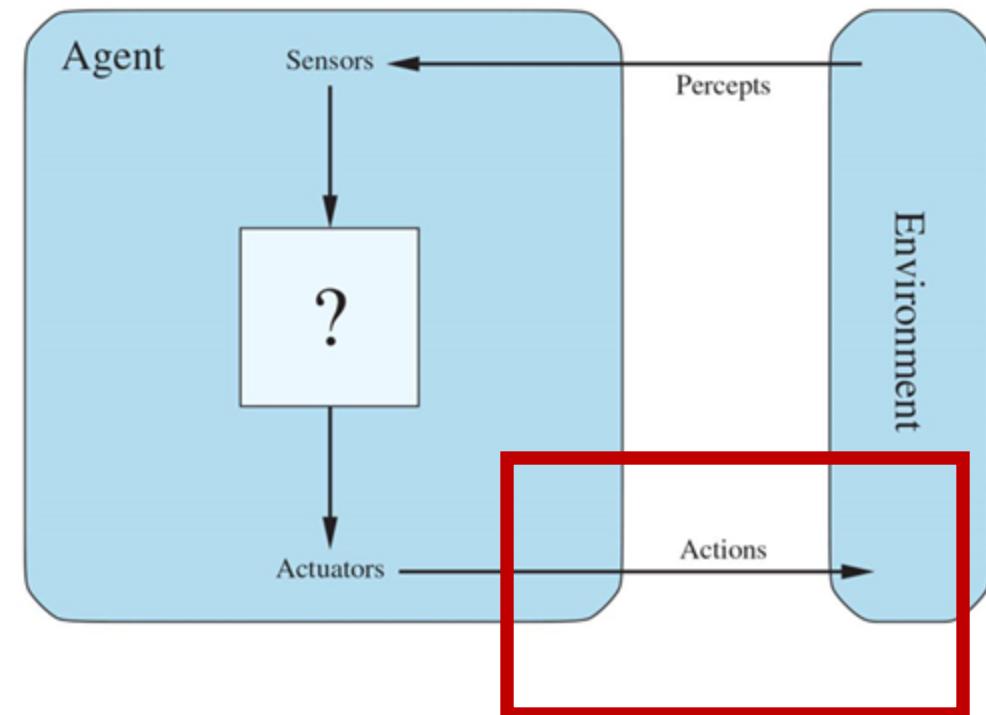
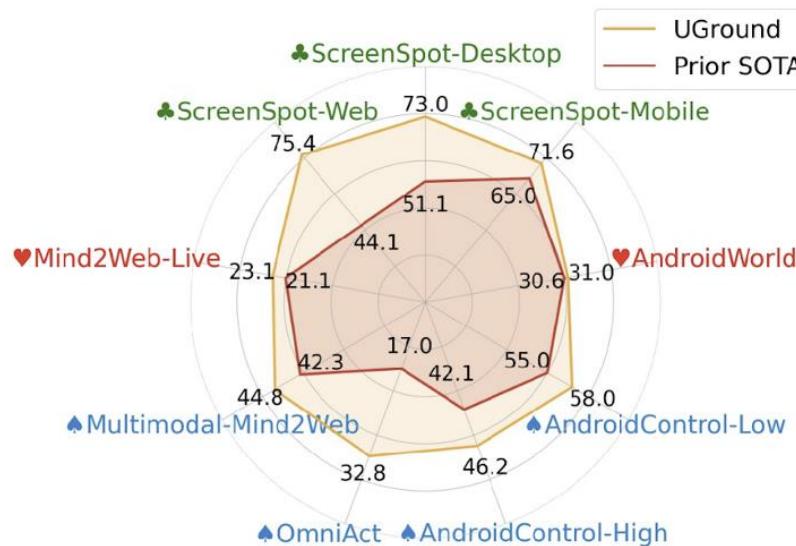


Action

Action:

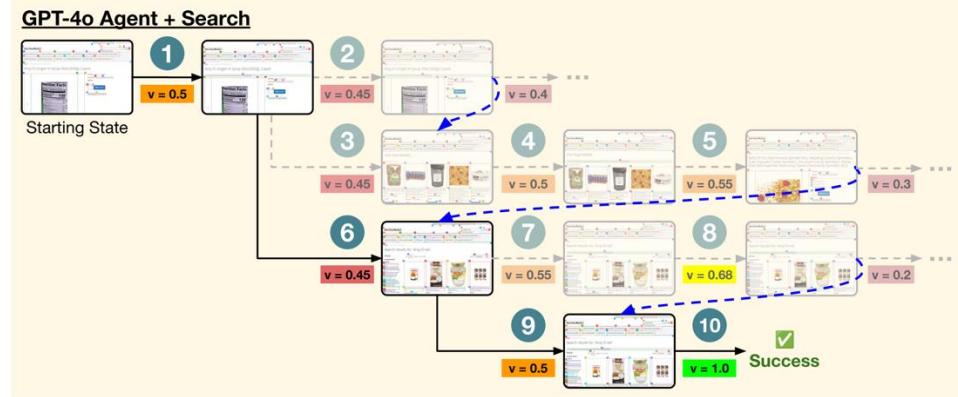
Weak Grounding: SOM, SeeAct

Pixel-Level: UGround, SeeClick, etc.

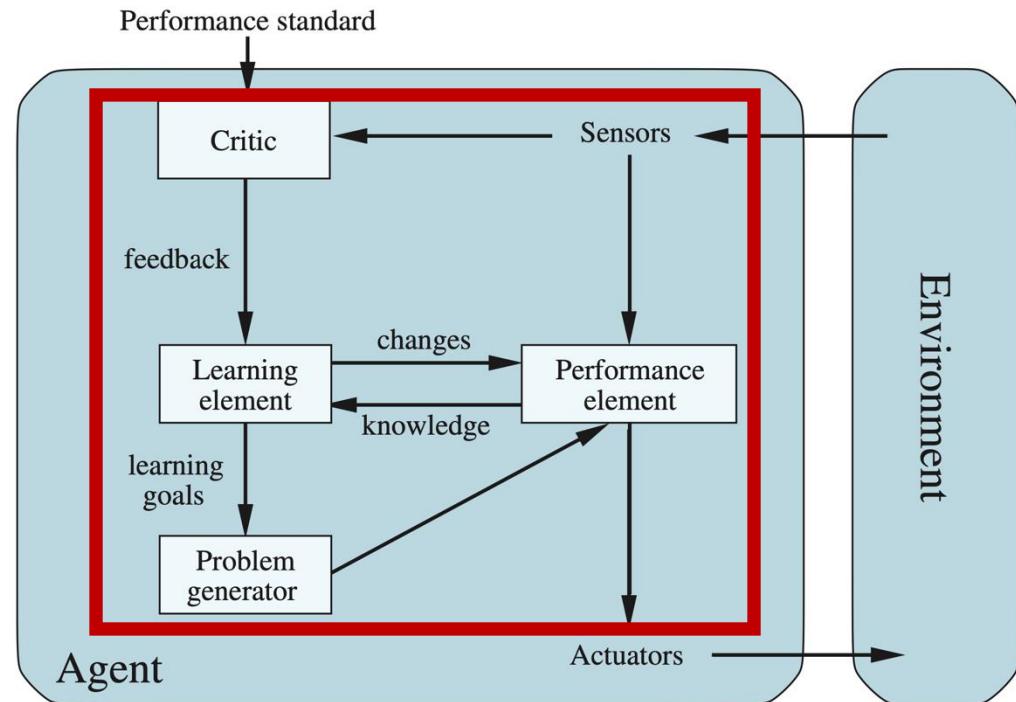
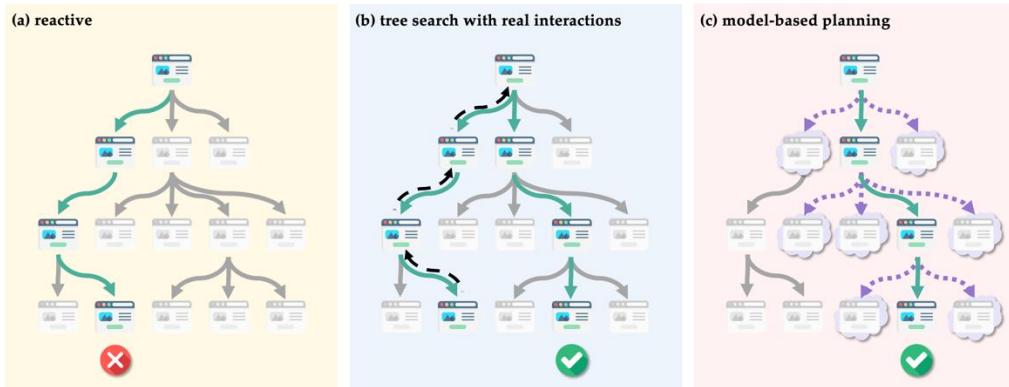


Planning

Search over the environment: Search-Agent



Planning with World Model: WebDreamer



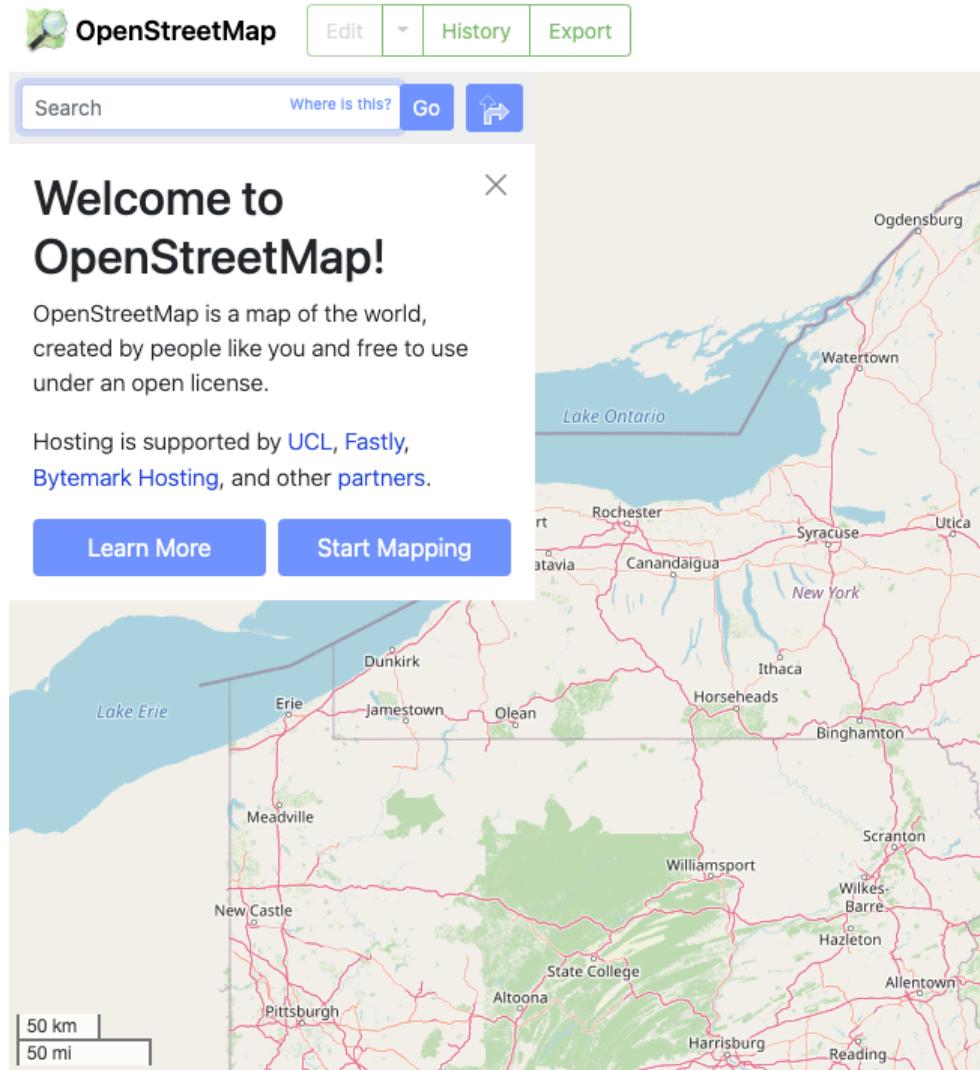
What's next step?

Agent has reached a decent performance of 30~50%.

How reach closer to 100% to be actually useful and robust?

What current agents are capable of?

Task Proposal



1. **Get Directions Between Two Points:**
 - Usefulness: 5 (Finding directions is a frequent task for users of a map service.)
 - Steps: Click on 'Find directions between two points' link, enter starting point, enter destination, click 'Go'. Total: 4 actions.
 - Total Rating: 9

2. **Export Map Data:**
 - Usefulness: 4 (Exporting map data is useful for offline usage or data analysis.)
 - Steps: Click on 'Export' link, select export options, click 'Export'. Total: 3 actions.
 - Total Rating: 7

3. **Search for Location:**
 - Usefulness: 5 (Common and frequent usage of map services.)
 - Steps: Enter text into 'Search' textbox, click 'Go'. Total: 2 actions.
 - Total Rating: 7

4. **Show My Location:**
 - Usefulness: 4 (Useful for quickly finding your current location on the map.)
 - Steps: Click 'Show My Location' button. Total: 1 action.
 - Total Rating: 5

The skill 'Get Directions Between Two Points' has the highest total rating and usefulness, making it the most beneficial skill to propose for this task.

Execution

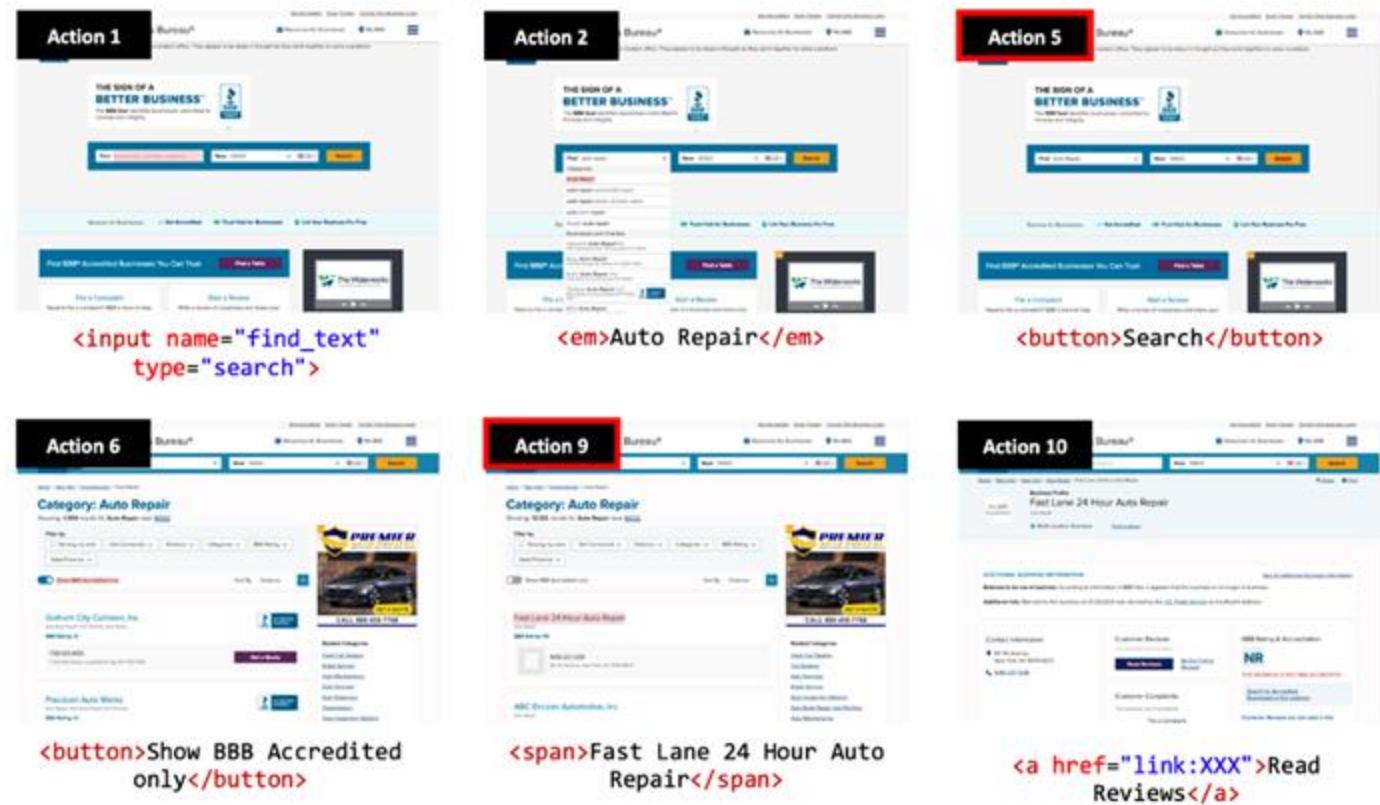
Task Description:

Show me the reviews for the auto repair business closest to 10002.

Action Sequence:

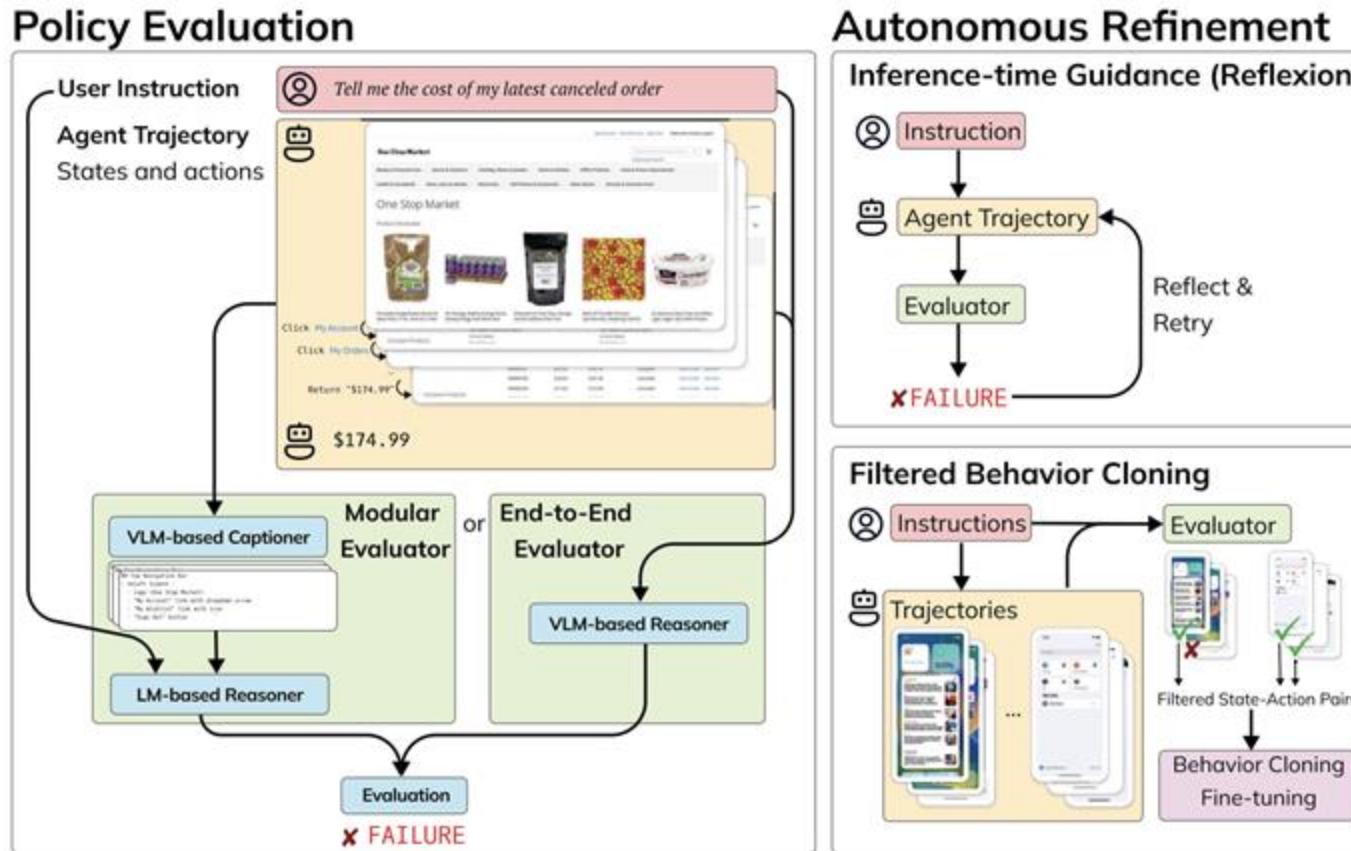
Target Element	Operation
1. [searchbox] Find	TYPE: <i>auto repair</i>
2. [button] Auto Repair	CLICK
3. [textbox] Near	TYPE: <i>10002</i>
4. [button] 10002	CLICK
5. [button] Search	CLICK
6. [switch] Show BBB Accredited only	CLICK
7. [svg]	CLICK
8. [button] Sort By	CLICK
9. [link] <i>Fast Lane 24 Hour Auto Repair</i>	CLICK
10. [link] <i>Read Reviews</i>	CLICK

Webpage Snapshots:



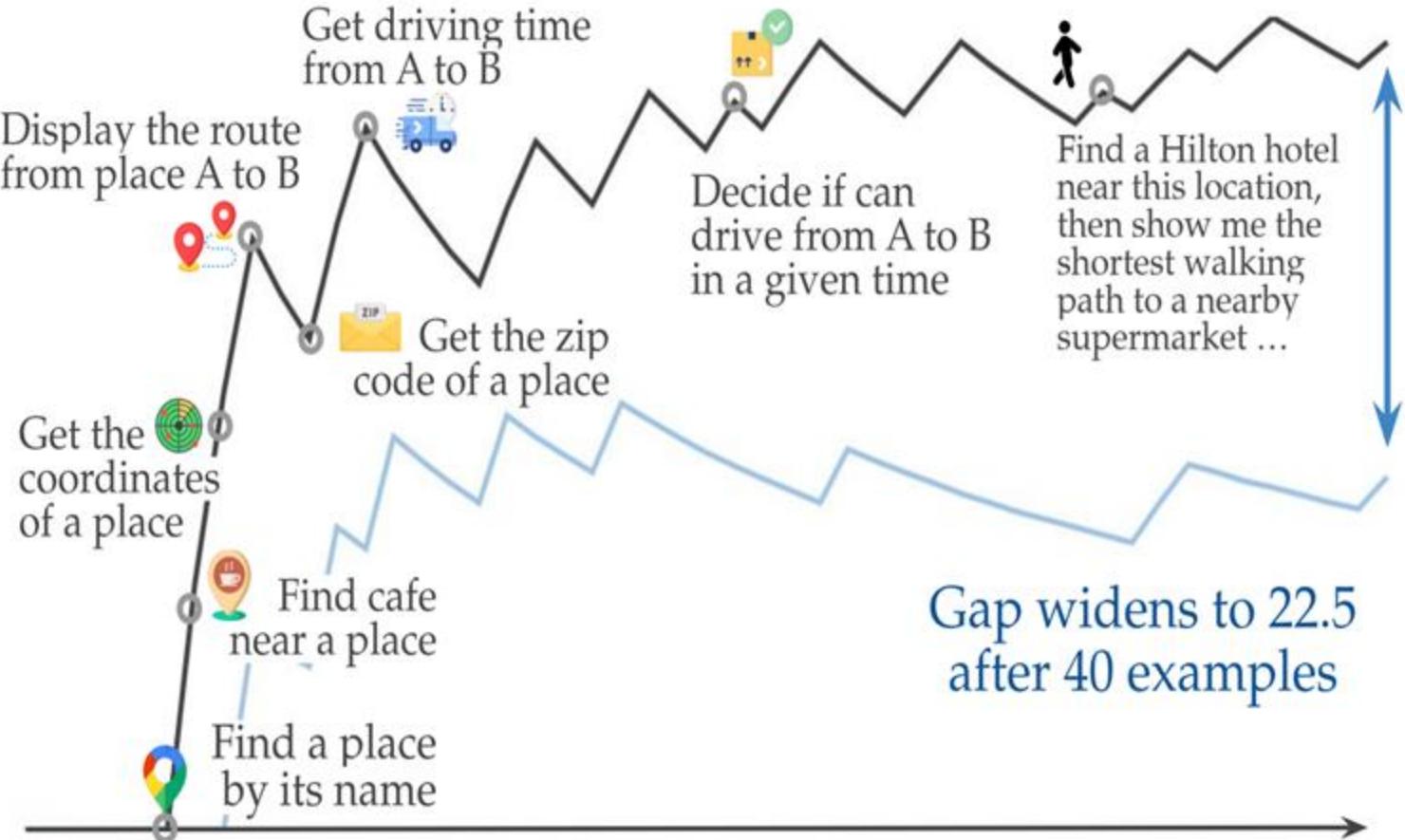
Critic

Given action sequence and instruction, label if the task is successfully completed



Self-Improvement through exploration

1. Propose tasks
2. Explore websites
3. Accumulate skills
4. Leverage knowledge



What's the Abstraction Level?

Trajectories: Accumulate knowledge in an inexplicit way

- Heavy: Training Policy
- Update & Adapt to new websites? Catastrophic forgetting
- Website changes?

Workflow: Textual Description of Procedure

- Lengthy
- How to update workflow?
- How to leverage multiple attempts into a workflow?

What's the Abstraction Level?

APIs:

- Light-Weighted
- Easy to Debug
- Easy for trial-and-error

```
async def identify_pill(page, imprint, color=None, shape=None):  
    """  
        Automates the process of identifying a pill using the Pill Identifier  
        feature on Drugs.com.  
  
        Parameters:  
        - page: The Playwright page object.  
        - imprint: The imprint on the pill to be identified.  
        - color: (Optional) The color of the pill.  
        - shape: (Optional) The shape of the pill.  
  
        This function navigates to the Pill Identifier page, agrees to the terms,  
        inputs the pill's characteristics,  
        and submits the information for identification.  
  
        Usage Log:  
        - Successfully navigated to the Pill Identifier page and submitted pill  
        information for identification.  
        - Inputted imprint '93 5510', color 'White', and shape 'Oval' and  
        successfully submitted for identification.  
        - Encountered issues with strict mode violations when attempting to click  
        the 'Search' button due to multiple matches.  
        - Updated to use a more specific selector for the 'Search' button to avoid  
        strict mode violations.  
    """  
  
    import re  
  
    await page.goto("https://www.drugs.com/pill_identification.html")  
    await page.get_by_role("link", name="Agree and Continue").click()  
    await page.get_by_role("textbox", name="Pill Imprint").fill(imprint)  
    if color:  
        await page.get_by_role("group", name="Color and shape (optional)").  
            get_by_role("combobox", name="Color  
                (optional)").select_option(color)  
    if shape:  
        await page.get_by_role("group", name="Color and shape (optional)").  
            get_by_role("combobox", name="Shape  
                (optional)").select_option(shape)  
  
    search_button =  
        page.locator("button.ddc-btn.ddc-btn-block[data-submit-loading]")  
    await search_button.click()
```

What if?

We have an algorithm:

Explore environment and accumulate skill&knowledge into APIs

A plug-and-play module for any agent capable of function calling

Overview

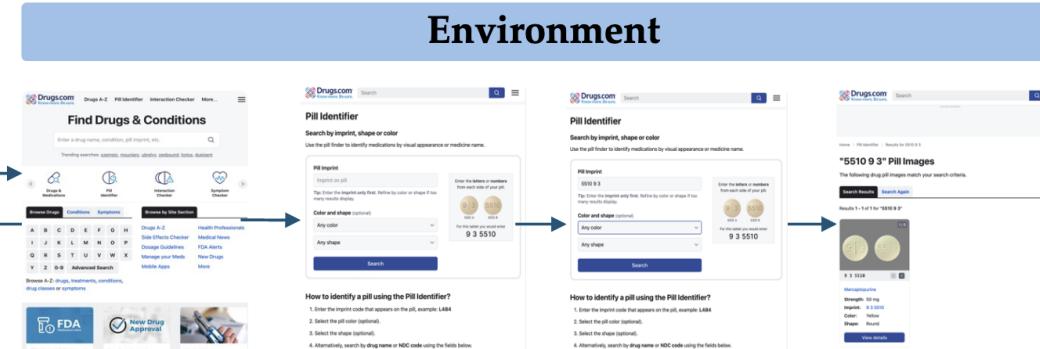
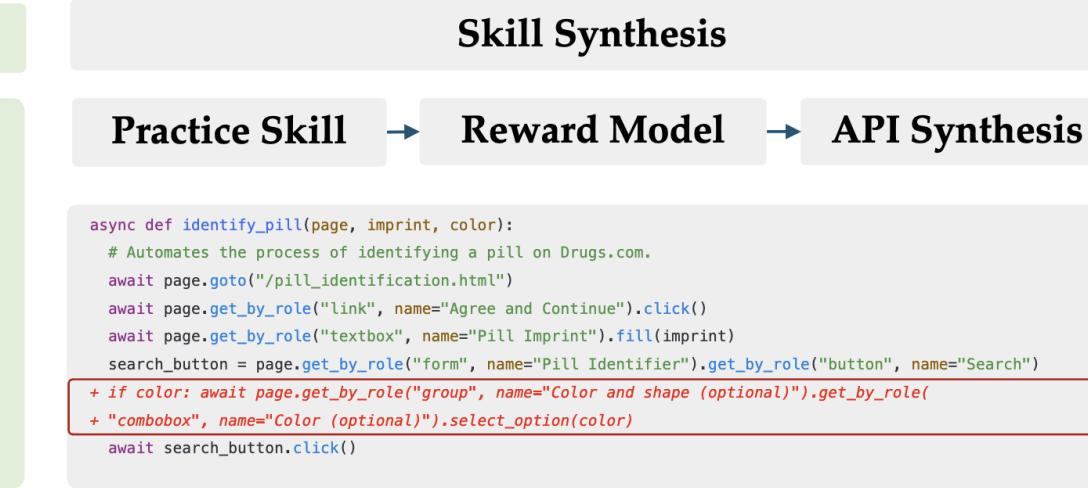
Skill Library

- check_drug_interaction(drug_name)
- subscribe_to_newsletters(email)
- search_FDA_alert()
- check_side_effects(drug_name)
- identify_pill(imprint, color)

Skill Proposal

time useful

Identify Pill Using Pill Identifier	3	5	
Check Side Effects for a Drug	3	4	
Search a Condition's Treatment	3	4	
Browse Drugs by Letter	2	4	



Verification

Test Cases → Execution

Error:
Warning: Unused parameter 'color'
→ identify_pill(page, imprint, color)
The parameter 'color' is defined but never used in the function body



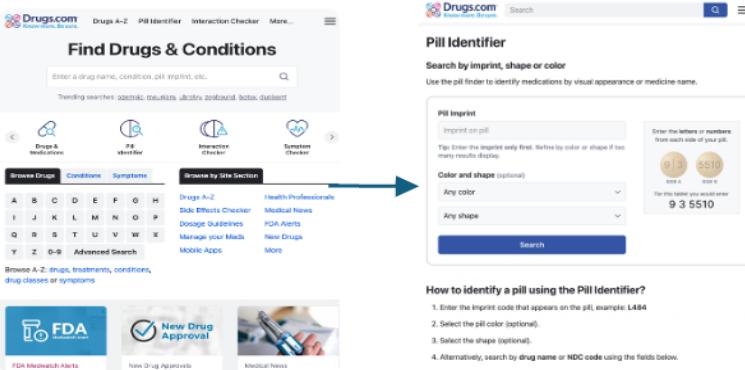
Error: Search results validation failed.
API returned pills matching imprint '5510' but the wrong color (expected: yellow).

Diagnosis:
Added color selection using the dropdown menu
Made color filtering optional

Uses 'if color': to make color filtering optional
Properly implemented color parameter in the search function, locating the "Color and shape" group element

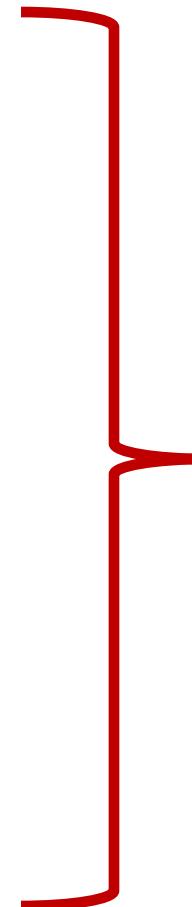
Skill Proposal

Environment



Skill Library

-  check_drug_interaction(drug_name)
-  subscribe_to_newsletters(email)
-  search_FDA_alert()
-  check_side_effects(drug_name)
-  identify_pill(imprint, color)



Skill Proposal

time useful

Identify Pill Using Pill Identifier 3 5 ✓

Check Side Effects for a Drug 3 4 ✗

Search a Condition's Treatment 3 4 ✗

Browse Drugs by Letter 2 4 ✗

Skill Synthesis

Skill Library

-  check_drug_interaction(drug_name)
-  subscribe_to_newsletters(email)
-  search_FDA_alert()
-  check_side_effects(drug_name)
-  identify_pill(imprint, color)

Skill Synthesis

Practice Skill

Reward Model

API Synthesis

```
async def identify_pill(page, imprint, color):
    # Automates the process of identifying a pill on Drugs.com.
    await page.goto("/pill_identification.html")
    await page.get_by_role("link", name="Agree and Continue").click()
    await page.get_by_role("textbox", name="Pill Imprint").fill(imprint)
    search_button = page.get_by_role("form", name="Pill Identifier").get_by_role("button", name="Search")

    await search_button.click()
```

Execution

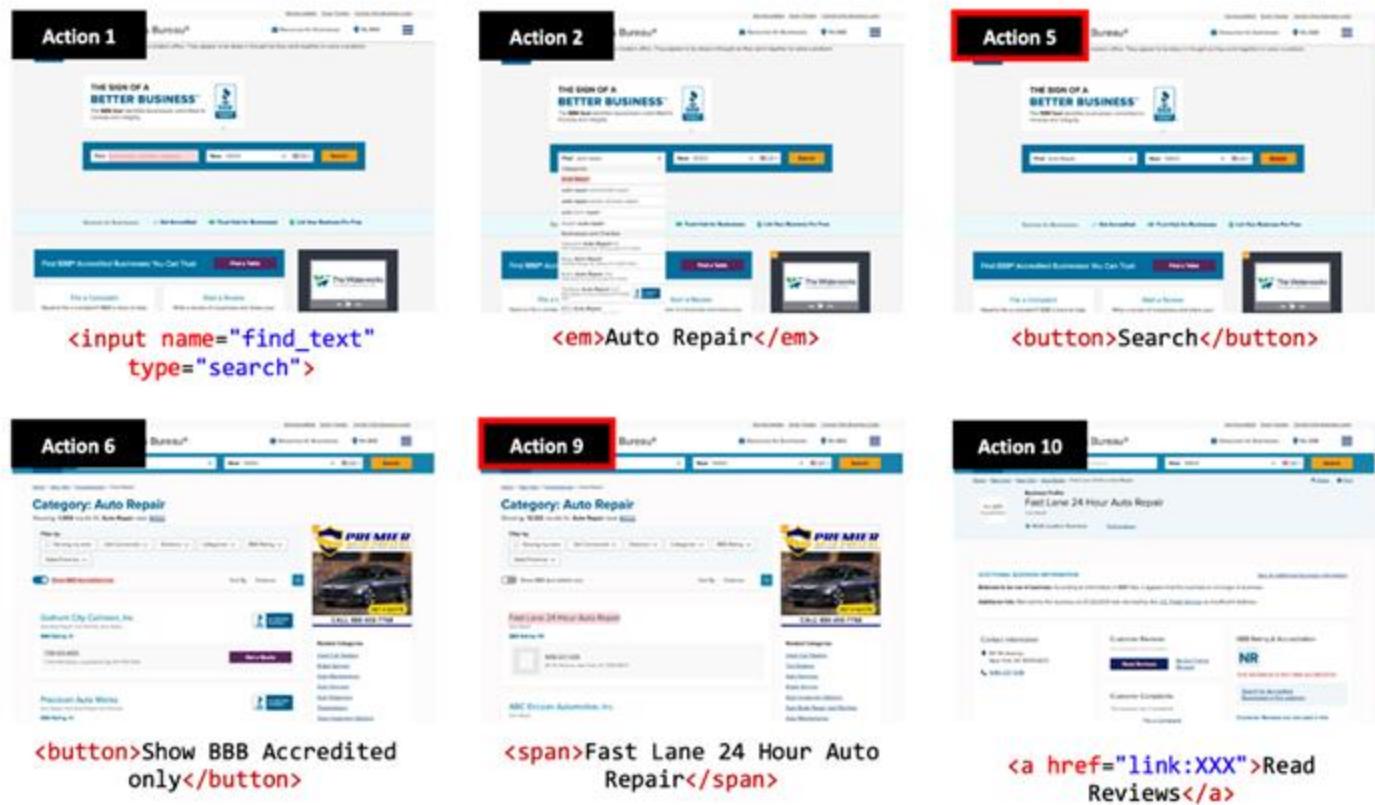
Task Description:

Show me the reviews for the auto repair business closest to 10002.

Action Sequence:

Target Element	Operation
1. [searchbox] Find	TYPE: <i>auto repair</i>
2. [button] Auto Repair	CLICK
3. [textbox] Near	TYPE: <i>10002</i>
4. [button] 10002	CLICK
5. [button] Search	CLICK
6. [switch] Show BBB Accredited only	CLICK
7. [svg]	CLICK
8. [button] Sort By	CLICK
9. [link] <i>Fast Lane 24 Hour Auto Repair</i>	CLICK
10. [link] <i>Read Reviews</i>	CLICK

Webpage Snapshots:



Skill Synthesis

Skill Library

-  check_drug_interaction(drug_name)
-  subscribe_to_newsletters(email)
-  search_FDA_alert()
-  check_side_effects(drug_name)
-  identify_pill(imprint, color)

Skill Synthesis

Practice Skill

Reward Model

API Synthesis

```
async def identify_pill(page, imprint, color):
    # Automates the process of identifying a pill on Drugs.com.
    await page.goto("/pill_identification.html")
    await page.get_by_role("link", name="Agree and Continue").click()
    await page.get_by_role("textbox", name="Pill Imprint").fill(imprint)
    search_button = page.get_by_role("form", name="Pill Identifier").get_by_role("button", name="Search")

    await search_button.click()
```

Skill Synthesis

Skill Library

-  check_drug_interaction(drug_name)
-  subscribe_to_newsletters(email)
-  search_FDA_alert()
-  check_side_effects(drug_name)
-  identify_pill(imprint, color)

Skill Synthesis

Practice Skill → Reward Model → API Synthesis

```
async def identify_pill(page, imprint, color):
    # Automates the process of identifying a pill on Drugs.com.
    await page.goto("/pill_identification.html")
    await page.get_by_role("link", name="Agree and Continue").click()
    await page.get_by_role("textbox", name="Pill Imprint").fill(imprint)
    search_button = page.get_by_role("form", name="Pill Identifier").get_by_role("button", name="Search")

    await search_button.click()
```

Verification

```
async def identify_pill(page, imprint, color):
    # Automates the process of identifying a pill on Drugs.com.
    await page.goto("/pill_identification.html")
    await page.get_by_role("link", name="Agree and Continue").click()
    await page.get_by_role("textbox", name="Pill Imprint").fill(imprint)
    search_button = page.get_by_role("form", name="Pill Identifier").get_by_role("button", name="Search")

    await search_button.click()
```

The screenshot shows the Drugs.com Pill Identifier search interface. At the top, there is a navigation bar with the Drugs.com logo, a search bar, and a magnifying glass icon. Below the header, the title "Pill Identifier" is displayed, followed by the sub-instruction "Search by imprint, shape or color". A tip below the input field says, "Use the pill finder to identify medications by visual appearance or medicine name." The main search area has a "Pill Imprint" input field containing "5510 9 3". To the right of this field is a note: "Enter the letters or numbers from each side of your pill." Below the imprint field are dropdown menus for "Color and shape (optional)" set to "Any color" and "Any shape". To the right of these dropdowns is a diagram of a tablet with two circular sides labeled "SIDE A" and "SIDE B". The left side shows "9 | 3" and the right side shows "5510". Below the diagram, the text "For this tablet you would enter 9 3 5510" is displayed. At the bottom of the search area is a large blue "Search" button.

How to identify a pill using the Pill Identifier?

1. Enter the imprint code that appears on the pill, example: L484
2. Select the pill color (optional).
3. Select the shape (optional).
4. Alternatively, search by drug name or NDC code using the fields below.

Verification

```
async def identify_pill(page, imprint, color):
    # Automates the process of identifying a pill on Drugs.com.
    await page.goto("/pill_identification.html")
    await page.get_by_role("link", name="Agree and Continue").click()
    await page.get_by_role("textbox", name="Pill Imprint").fill(imprint)
    search_button = page.get_by_role("form", name="Pill Identifier").get_by_role("button", name="Search")

    await search_button.click()
```

Pill Imprint
5510 9 3

Tip: Enter the imprint only first. Refine by color or shape if too many results display.

Color and shape (optional)

Any color

Any shape

Search

Enter the letters or numbers from each side of your pill.

9 | 3 5510

SIDE A SIDE B

For this tablet you would enter
9 3 5510

Verification

Test Cases

Execution

Error:



Warning: Unused parameter 'color'
-> identify_pill(page, imprint, color)

The parameter 'color' is defined but never used in the function body

Error: Search results validation failed.

API returned pills matching imprint '5510' but the wrong color (expected: yellow).

Diagnosis:

Added color selection using the dropdown menu
Made color filtering optional

Uses 'if color': to make color filtering optional
Properly implemented color parameter in the search function, locating the "Color and shape" group element

Verification

```
async def identify_pill(page, imprint, color):
    # Automates the process of identifying a pill on Drugs.com.
    await page.goto("/pill_identification.html")
    await page.get_by_role("link", name="Agree and Continue").click()
    await page.get_by_role("textbox", name="Pill Imprint").fill(imprint)
    search_button = page.get_by_role("form", name="Pill Identifier").get_by_role("button", name="Search")
    + if color: await page.get_by_role("group", name="Color and shape (optional)").get_by_role(
    + "combobox", name="Color (optional)").select_option(color)
    await search_button.click()
```

Pill Imprint

Tip: Enter the imprint only first. Refine by color or shape if too many results display.

Color and shape (optional)

Any color

Any shape

Search

Enter the letters or numbers from each side of your pill.

9 | 3 5510

SIDE A SIDE B

For this tablet you would enter

9 3 5510

Verification

Test Cases

Execution



Error:

Warning: Unused parameter 'color'
-> identify_pill(page, imprint, color)
The parameter 'color' is defined but never used in the function body

Error: Search results validation failed.

API returned pills matching imprint '5510' but the wrong color (expected: yellow).

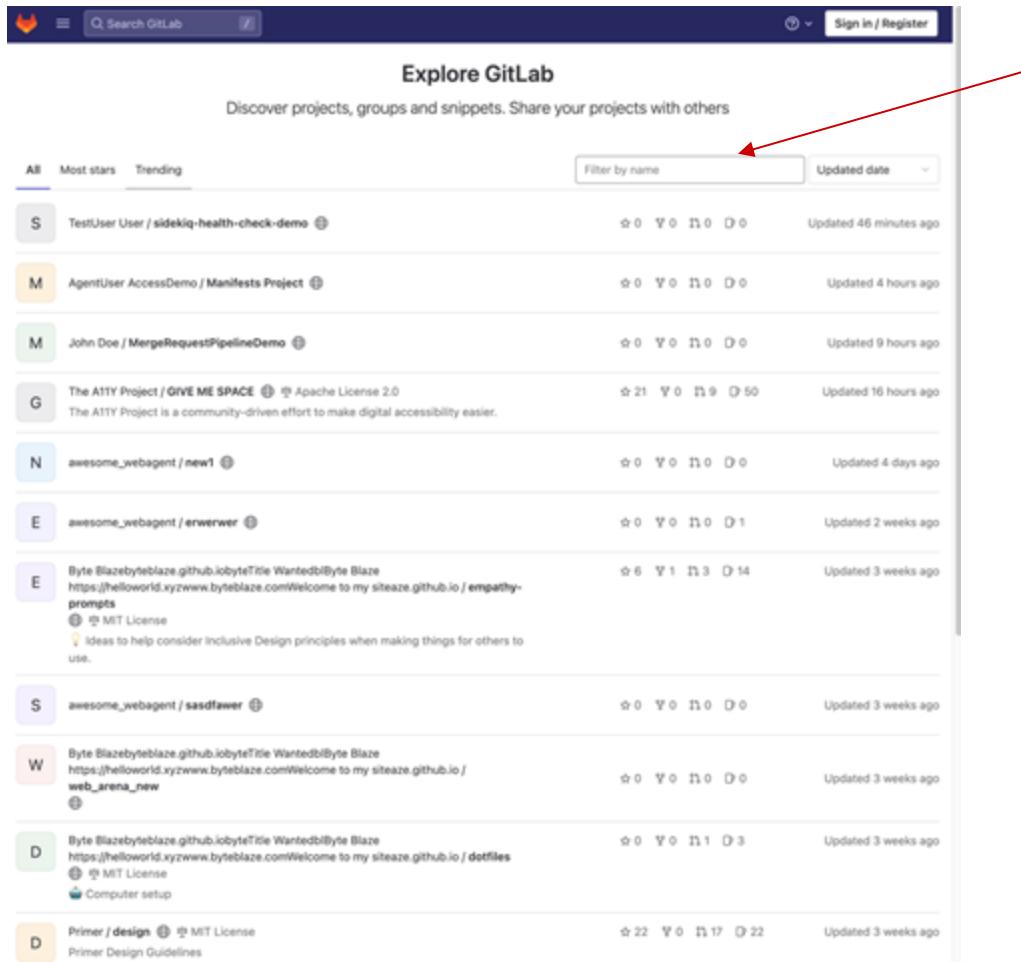
Diagnosis:

Added color selection using the dropdown menu
Made color filtering optional

Uses 'if color': to make color filtering optional
Properly implemented color parameter in the search function, locating the "Color and shape" group element

More Examples of Synthesized APIs

APIs in GitLab



The screenshot shows the GitLab Explore page with a list of projects. At the top right, there is a 'Sign in / Register' button. Below it, a search bar contains the placeholder 'Search GitLab'. To the right of the search bar is a 'Filter by name' input field, which is highlighted with a red arrow pointing to the corresponding documentation. The main area displays a table of projects with columns for name, stars, forks, issues, and last updated.

Project	Stars	Forks	Issues	Last Updated
S TestUser User / sidekiq-health-check-demo	0	0	110	0 days ago
M AgentUser AccessDemo / Manifests Project	0	0	110	0 days ago
M John Doe / MergeRequestPipelineDemo	0	0	110	0 days ago
G The ATTY Project / GIVE ME SPACE	21	0	119	0 days ago
N awesome_webagent / new1	0	0	110	0 days ago
E awesome_webagent / erwerwer	0	0	110	0 days ago
E Byte Blazebyteblaze.github.io	6	1	113	0 days ago
S awesome_webagent / sasdfawer	0	0	110	0 days ago
W Byte Blazebyteblaze.github.io	0	0	110	0 days ago
D Byte Blazebyteblaze.github.io	0	0	111	0 days ago
D Primer / design	22	0	117	0 days ago

async def search_projects_by_keyword(keyword: str):
 """
 Search for projects by a specific keyword.

 This function locates the 'Filter by name' search box, types the given keyword, and presses 'Enter' to execute the search operation.

 Usage:
 Call this function with the desired keyword to perform a search operation within the projects.

 :param keyword: The keyword to search for in projects.
 """

 searchbox_locator = page.get_by_role('main', name='').get_by_role('link', name='Filter by name')
 await searchbox_locator.click()
 await searchbox_locator.fill(keyword)
 await searchbox_locator.type(keyword)
 await page.keyboard.press('Enter')

async def explore_and_clone_repository(project_name: str) -> str:
 """
 Explore and clone a project repository by extracting its clone URL.

 This function automates the process of navigating to the 'Explore' section of the website, selecting a project from the list, accessing the project page, and retrieving the clone URL by interacting with the 'Clone' system.

 Usage:
 Call this function with the 'project_name' of the project you want to clone. It will return the clone URL of the repository as a string.

 :param project_name: The name of the project to explore and clone.
 :return: A string containing the clone URL to clone the project repository.
 """
 await page.get_by_role('main', name='').get_by_role('link', name='Explore')
 await page.get_by_role('main', name='').click()
 await page.get_by_role('main', name='').get_by_role('link', name='Clone')
 await page.get_by_role('main', name='').click()
 await page.get_by_role('main', name='').get_by_role('button', name='Copy URL')
 await page.get_by_role('main', name='').click()
 return f'{url}/git@192.168.112.29:2222/{project_name}.git'.format(url=project_name,
 project_name=project_name.replace(' ', '_').lower())

async def merge_project_branches(project_name: str, merge_request_title: str):
 """
 Explore and merge branches for a project that has open merge requests.

 This function navigates to the list of open merge requests for a given project, selects a specific merge request by its title, explores the merging options, and performs the merge if it is deemed acceptable (even if there are unverified changes).

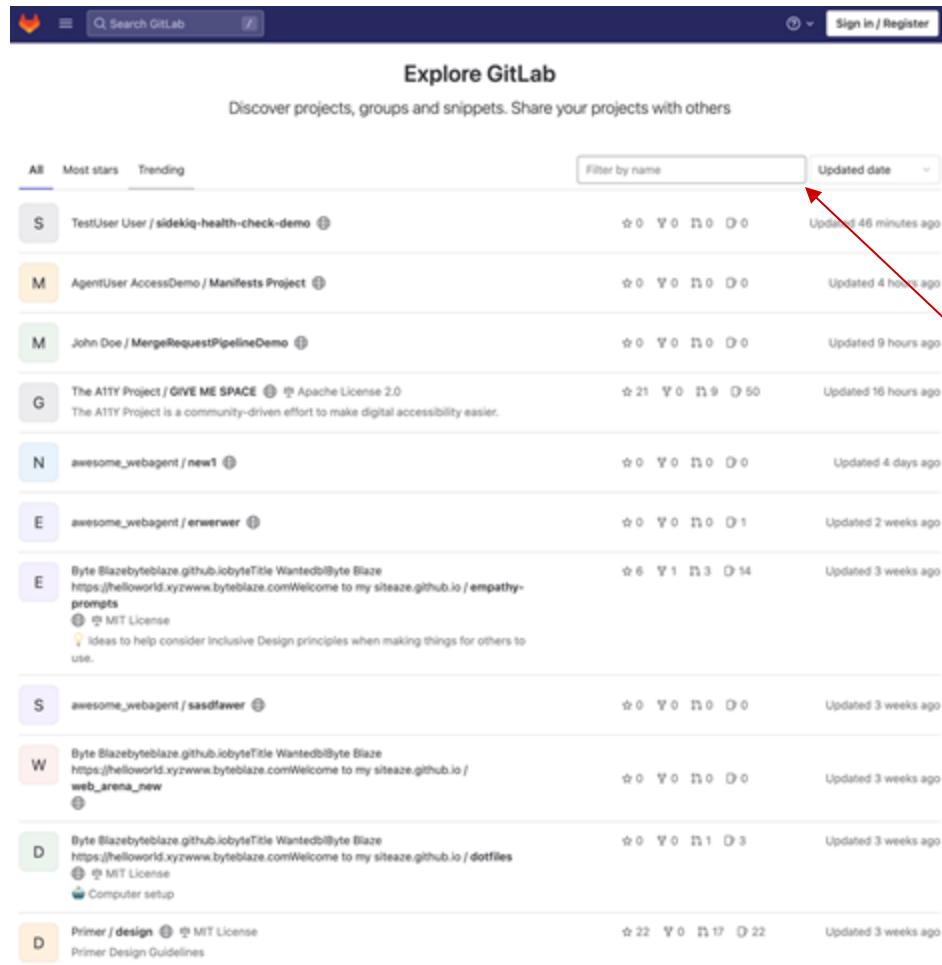
 Usage:
 Call this function with the 'project_name' to open the merge request list, and 'merge_request_title' to select and merge the branches.

 :param project_name: The name of the project which contains the merge requests.
 :param merge_request_title: The title of the merge request to explore and potentially merge.
 """

 await page.get_by_role('main', name='').get_by_role('link', name='2').click()
 await page.get_by_role('main', name='').get_by_role('link', name=merge_request_title).click()
 await page.get_by_role('main', name='').get_by_role('group', name='').get_by_role('button', name='Merge...').click()
 await page.get_by_role('dialog', name='Merge unverified changes?').get_by_role('content', name='Merge unverified changes').click()
 return 'The merge has been completed successfully.'

More Examples of Synthesized APIs

APIs in GitLab



The screenshot shows the GitLab Explore page with a list of projects. At the top, there is a search bar labeled 'Search GitLab' and a 'Sign in / Register' button. Below the search bar, the title 'Explore GitLab' is displayed, followed by the subtitle 'Discover projects, groups and snippets. Share your projects with others'. There are three tabs at the top: 'All', 'Most stars', and 'Trending'. The 'All' tab is selected. The main area lists projects with columns for the project name, star count, fork count, issue count, merge request count, and last update time. A red arrow points from the 'search_projects_by_keyword' function's 'Filter by name' description to the actual 'Filter by name' search box on the GitLab page.

Project	Stars	Forks	Issues	Merge Requests	Last Updated
TestUser User / sidekiq-health-check-demo	0	0	0	0	Updated 46 minutes ago
AgentUser AccessDemo / Manifests Project	0	0	0	0	Updated 4 hours ago
John Doe / MergeRequestPipelineDemo	0	0	0	0	Updated 9 hours ago
The ATTY Project / GIVE ME SPACE	21	0	19	50	Updated 16 hours ago
The ATTY Project is a community-driven effort to make digital accessibility easier.					
awesome_webagent / new1	0	0	0	0	Updated 4 days ago
awesome_webagent / erwerwer	0	0	0	1	Updated 2 weeks ago
Byte BlazebyteBlaze.github.iobyteTitle Wanteddb!Byte Blaze	6	1	3	14	Updated 3 weeks ago
https://helloworld.xyzwww.byteblaze.comWelcome to my siteaze.github.io / empathy-prompts					
Byte BlazebyteBlaze.github.iobyteTitle Wanteddb!Byte Blaze	0	0	0	0	Updated 3 weeks ago
https://helloworld.xyzwww.byteblaze.comWelcome to my siteaze.github.io / web_arena_new					
Byte BlazebyteBlaze.github.iobyteTitle Wanteddb!Byte Blaze	0	1	1	3	Updated 3 weeks ago
https://helloworld.xyzwww.byteblaze.comWelcome to my siteaze.github.io / dotfiles					
Byte BlazebyteBlaze.github.iobyteTitle Wanteddb!Byte Blaze	22	0	17	22	Updated 3 weeks ago
Primer / design					
Primer Design Guidelines					

```
async def search_projects_by_keyword(keyword: str):
```

```
    """
```

Search for projects by a specific keyword.

This function locates the 'Filter by name' search box, types the given keyword, and presses 'Enter' to execute the search operation.

Usage:

Call this function with the desired keyword to perform a search operation within the projects.

```
:param keyword: The keyword to search for in projects.
```

```
"""
```

```
searchbox_locator = page.get_by_role('main', name='').get_by_role(
    'searchbox', name='Filter by name')
await searchbox_locator.click()
await searchbox_locator.fill('')
await searchbox_locator.type(keyword)
await page.keyboard.press('Enter')
```

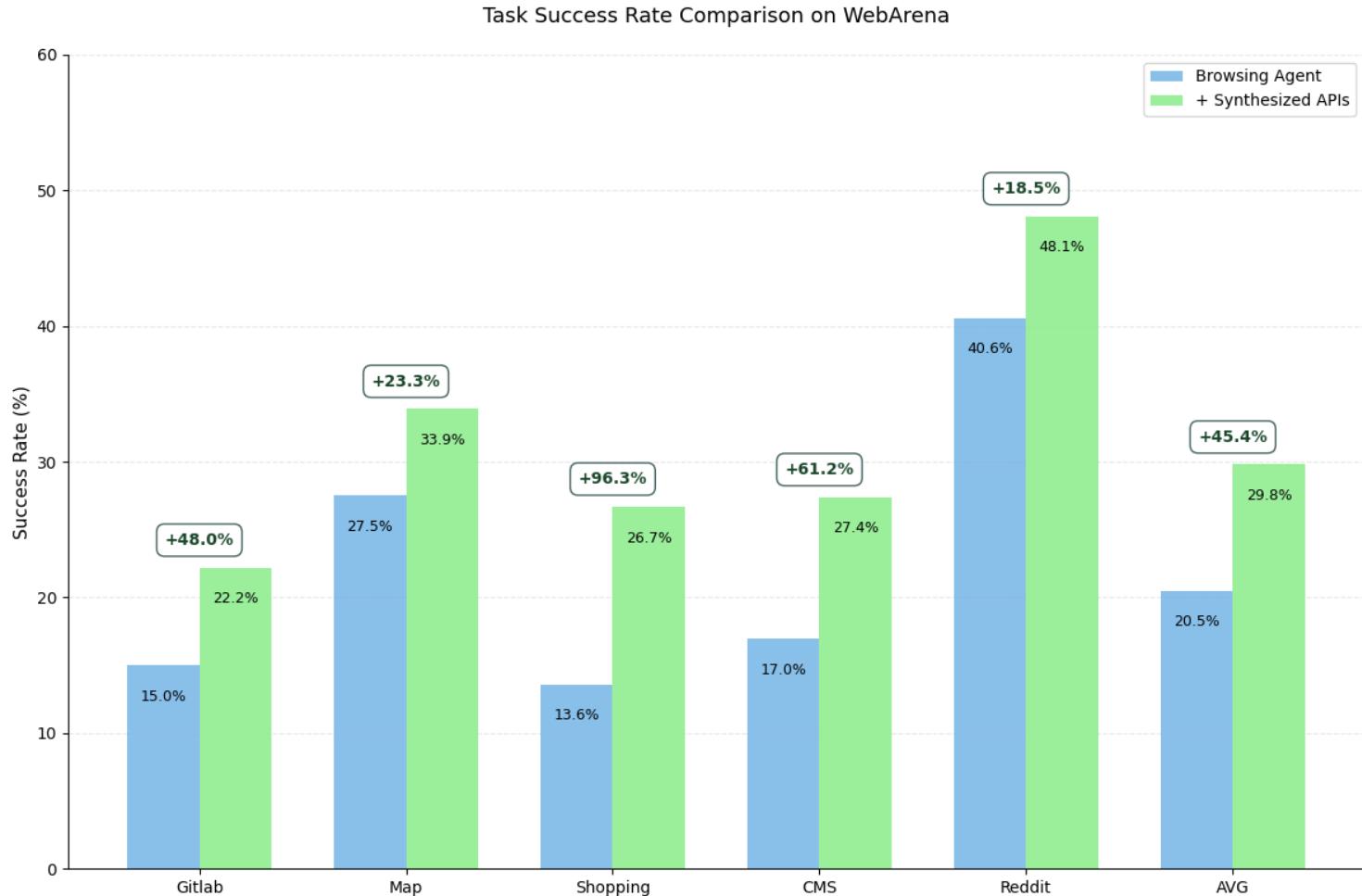
Evaluation on Sandbox

Exploration:

Synthesize APIs with 160 Iterations

Agents:

- Browsing Agent:
 - Browser Action
- +API Synthesis
 - Extend action space with APIs



How far are we from human-crafted APIs?

- Human Crafted APIs:
 - Mining APIs from sandbox source code

Not realistic, but can serve for a case study

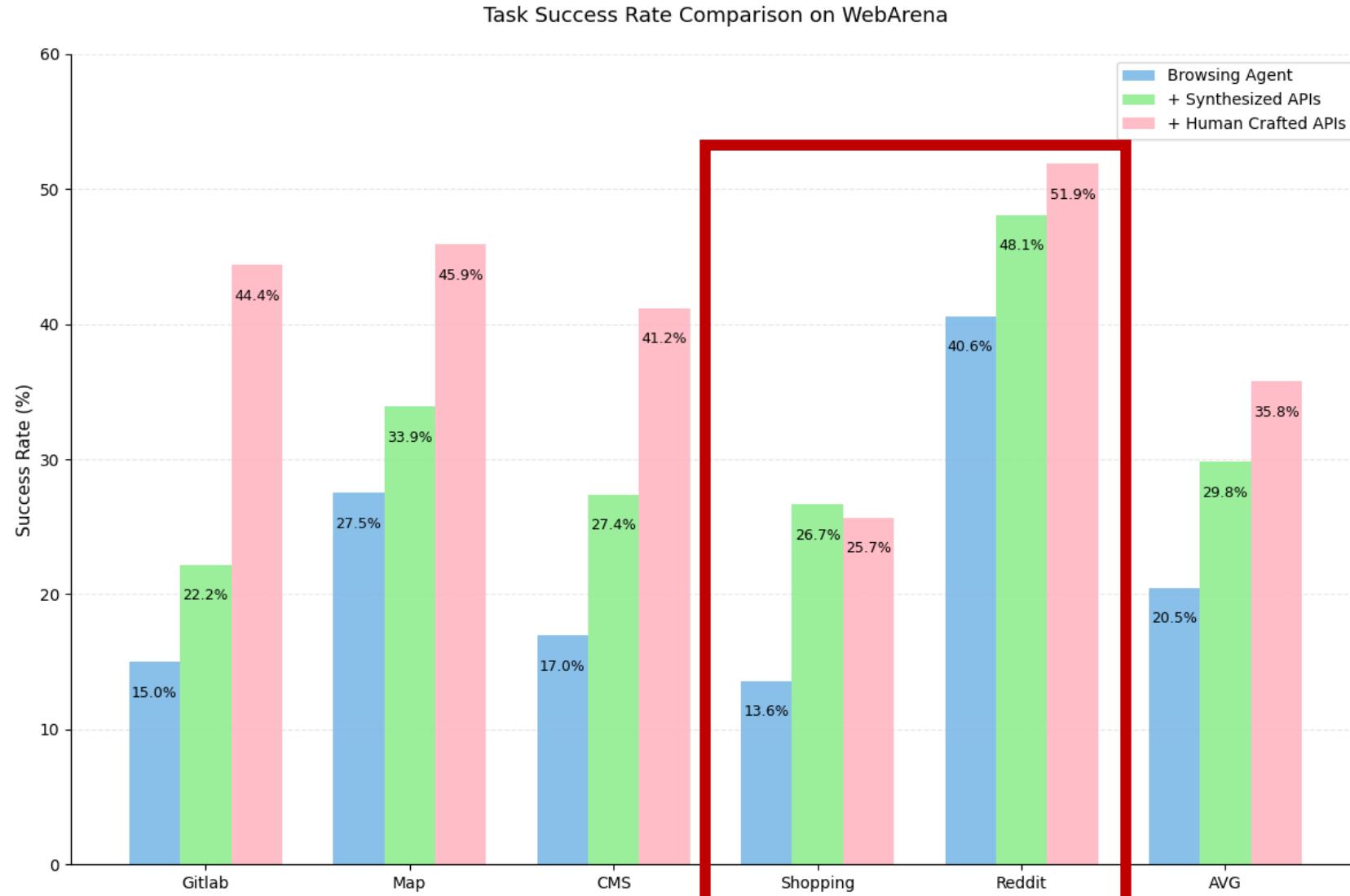
API Documentation	<pre># Commits ## GET /api/{id}/commits: Get a list of commits in a project. Attribute Type Description `id` integer/string The ID or path of the project. `since` string Only commits after or on this date. `until` string Only commits before or on this date. Output: JSON containing all commits that meet the given criteria.</pre>
API Calling	<pre><execute_ipython> requests.get('gitlab.com/api/allmyproject/commits') </execute_ipython></pre>
JSON Output	<pre>[.....{ "id": "ed37a2f2", "created_at": "2023-03-13T21:04:49.000-04:00", "title": "Update README.md", "message": "Update README.md", "author": "SaptakS", }]</pre>

Comparing with human crafted APIs

API Support Level

MEDIUM:
Admin Access
Limited APIs

LOW:
Manually written
automation Code

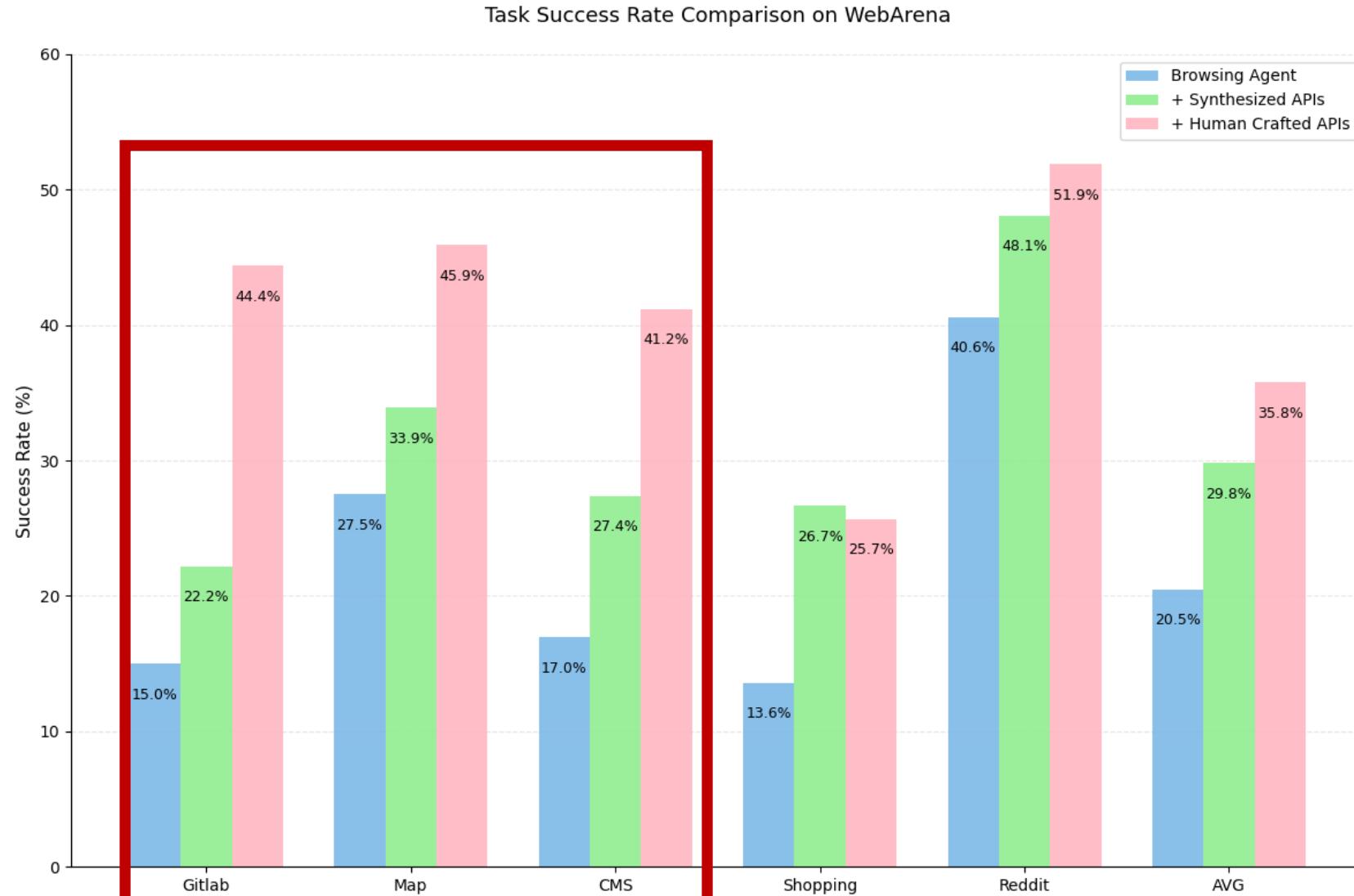


Comparing with human crafted APIs

HIGH API support:

Admin Access

High Quality API



Generalization to weaker Agent

- Empowering smaller models with APIs synthesized by larger models
- Exploration: GPT-4o
- Inference: GPT-4o-mini

	SKILLWEAVER					
GPT-4o	17.8	27.5	19.8	18.7	37.7	22.6
+ Skills	22.2	33.9	27.2	25.8	50.0	29.8
Δ	↑ 25%	↑ 23%	↑ 38%	↑ 38%	↑ 33%	↑ 32%
GPT-4o-mini	6.1	10.3	11.8	3.3	18.9	9.2
+ Skills	8.9	16.7	17.1	7.7	26.4	14.1
Δ	↑ 46%	↑ 62%	↑ 46%	↑ 133%	↑ 40%	↑ 45%

Scaling to Live Websites

Exploration:

80 Iterations

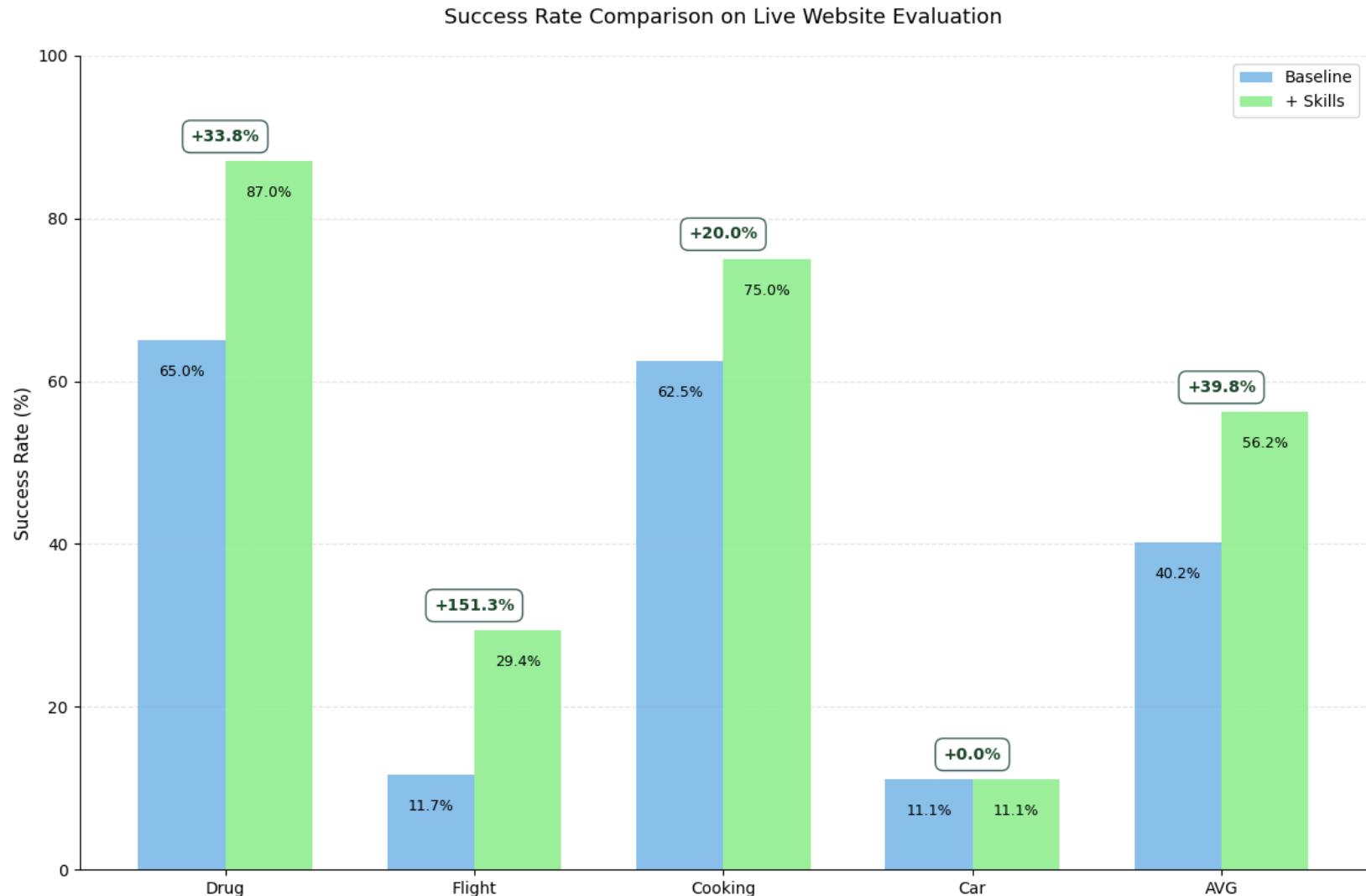
Evaluation:

Drug: 23

Flight: 17

Cooking: 8

Car: 9



Thank you!
&
Questions?