Multimodal AI

Lecture 10.2 – Interactive Agents

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

Assignments for This Coming Week

Midterm exam grades: mean 84.2, median is 85.7

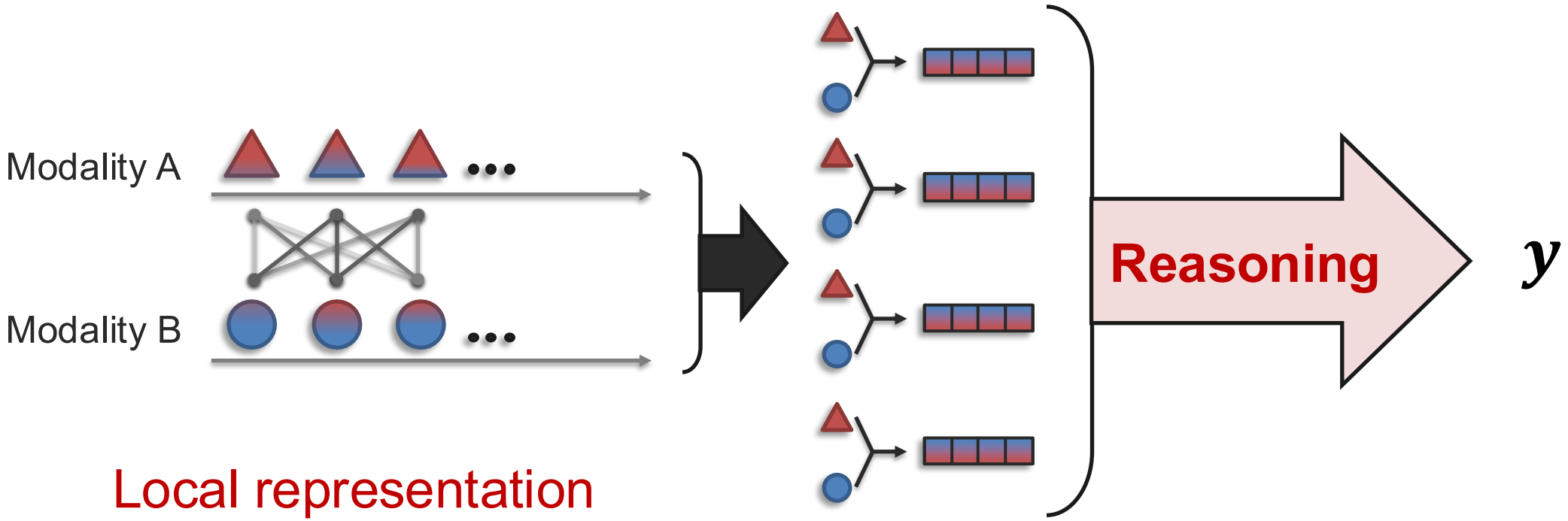
Project midterm due.

- Finalized main ideas and experimental setup, have datasets and baseline models working, detailed error analysis, initial progress towards implementing new ideas.

HW4 out tomorrow, on RL and reasoning.

Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting the structure of the problem.



Local representation

+ Aligned representation

multisensory intelligence

Roadmap

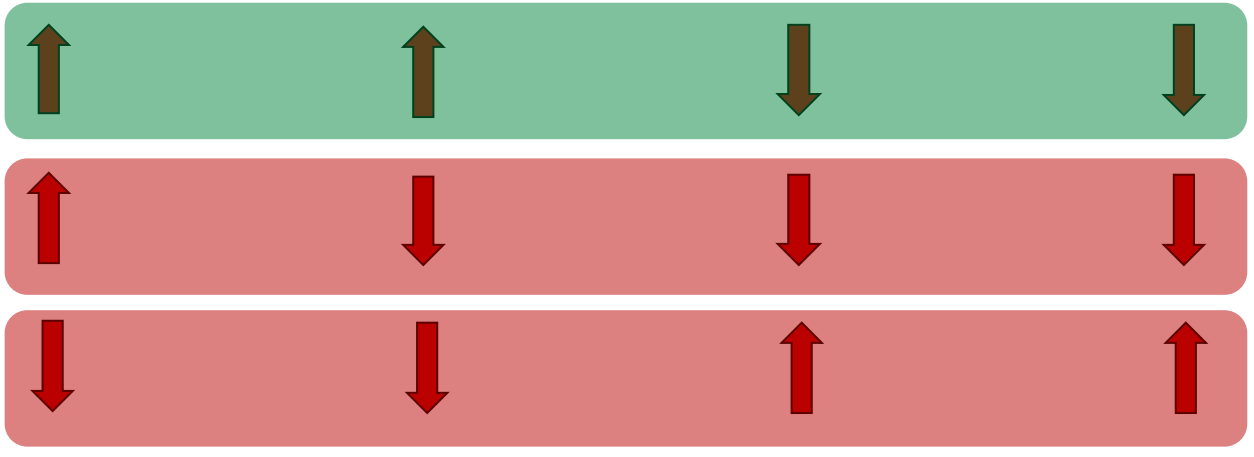
Input, (reasoning step 1, step 2, step 3…), output

Method 1: Direct prompting, no training (e.g., CoT)

Method 2: Supervised fine-tuning
- Assuming you have reasoning traces

Method 3: Reinforcement learning
- Assumes no reasoning traces
- But a reward function scoring reasoning and outputs

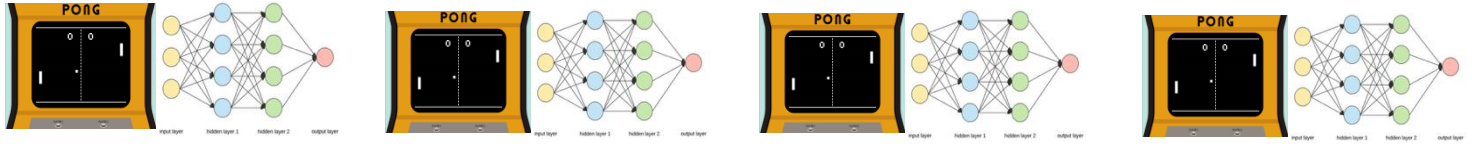
Policy Gradients



Maximize $\sum_i r_i \log p(y_i \mid x_i)$

Maximize $\sum_i r_i \log p(y_i \mid x_i)$

Maximize $\sum_i r_i \log p(y_i \mid x_i)$



multisensory intelligence

Policy Gradients

$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} r(\tau) \nabla_\theta \log \pi_\theta(a_t | s_t)$$

If $r(\tau)$ is positive, increase the probability

If $r(\tau)$ is negative, decrease the probability

But this suffers from high variance

multisensory intelligence

Policy Gradients

The raw reward may not be very meaningful.

What is important then? Whether a reward is higher or lower than what you expect.

-- Compare to a baseline, and use relative improvement

$$\nabla_\theta J(\theta) \approx \sum_{t \geq 0} \left(r(\tau) - b(s_t) \right) \nabla_\theta \log \pi_\theta(a_t | s_t)$$

e.g. exponential moving average of the rewards.

GRPO (Deepseek R1)

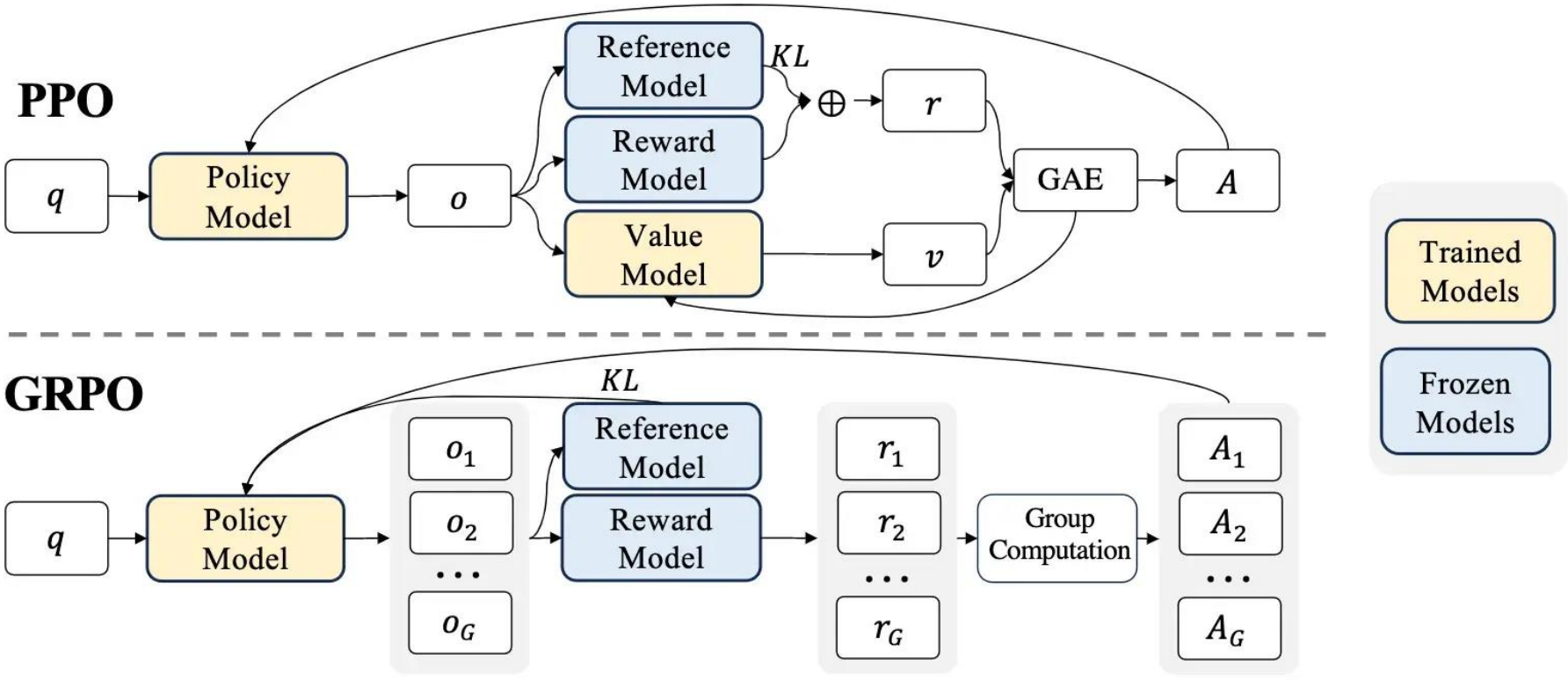


Figure 4 | Demonstration of PPO and our GRPO. GRPO foregoes the value model, instead estimating the baseline from group scores, significantly reducing training resources.

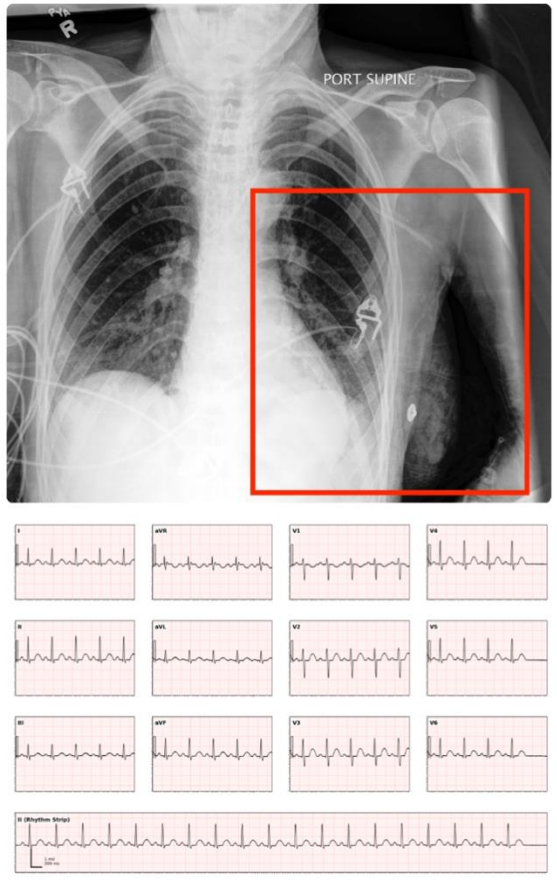
[Guo et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv 2025]

multisensory intelligence

Reasoning Examples

https://github.com/DDVD233/QoQ_Med

Unified reasoning across images and medical sensors



Question: Below is a history of a patient:...How long will the patient stay in the hospital?
A. 0-4 days B. 5-8 days C. 9-12 days D. more than 12 days
Ground truth: D. more than 12 days

Model Reasoning:

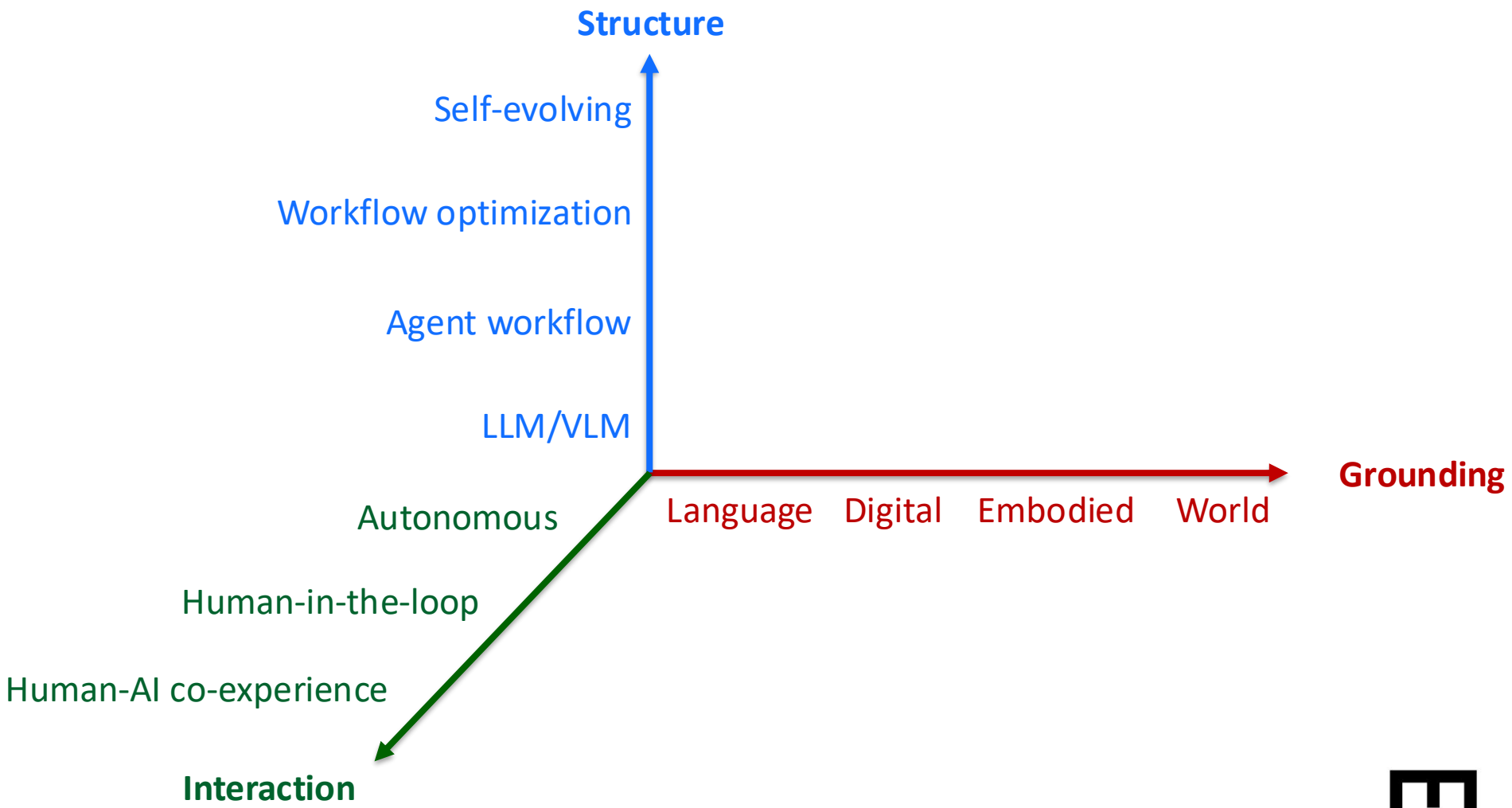
...the patient has been on mechanical ventilation for 96 hours, which is a long duration indicating a complex respiratory condition. The ECG and other lab results do not show any immediate critical issues, but the patient's hematocrit and hemoglobin levels are low, which could be a sign of anemia or other underlying conditions requiring treatment...the high and low heart rate alarms suggest the patient might be at risk for arrhythmias or other issues requiring close monitoring. Given the prolonged ventilation and the need for multiple invasive procedures, the patient is likely to require a significant period in the ICU or a high-dependency unit for close monitoring and management of their respiratory condition.

Answer: D. more than 12 days

[Dai et al., QoQ-Med: Building Multimodal Clinical Foundation Models with Domain-aware GRPO Training. NeurIPS 2025]

multisensory intelligence

My Overview of AI Agents



multisensory intelligence

Language 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 and augment human capabilities

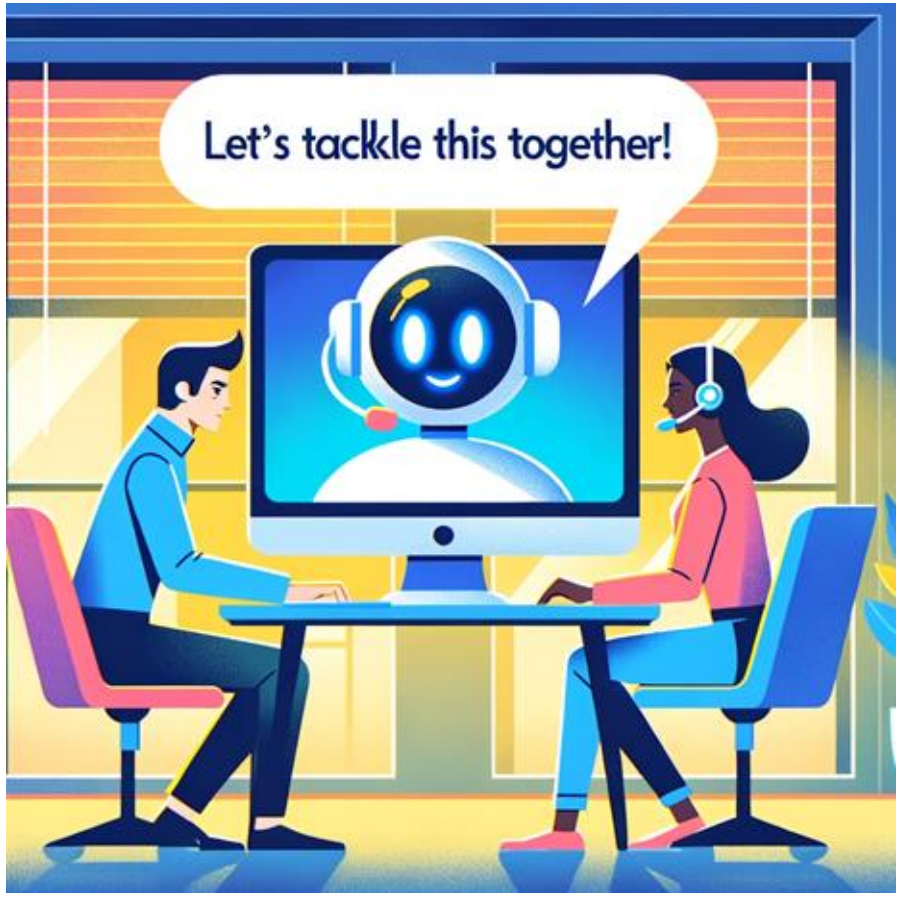
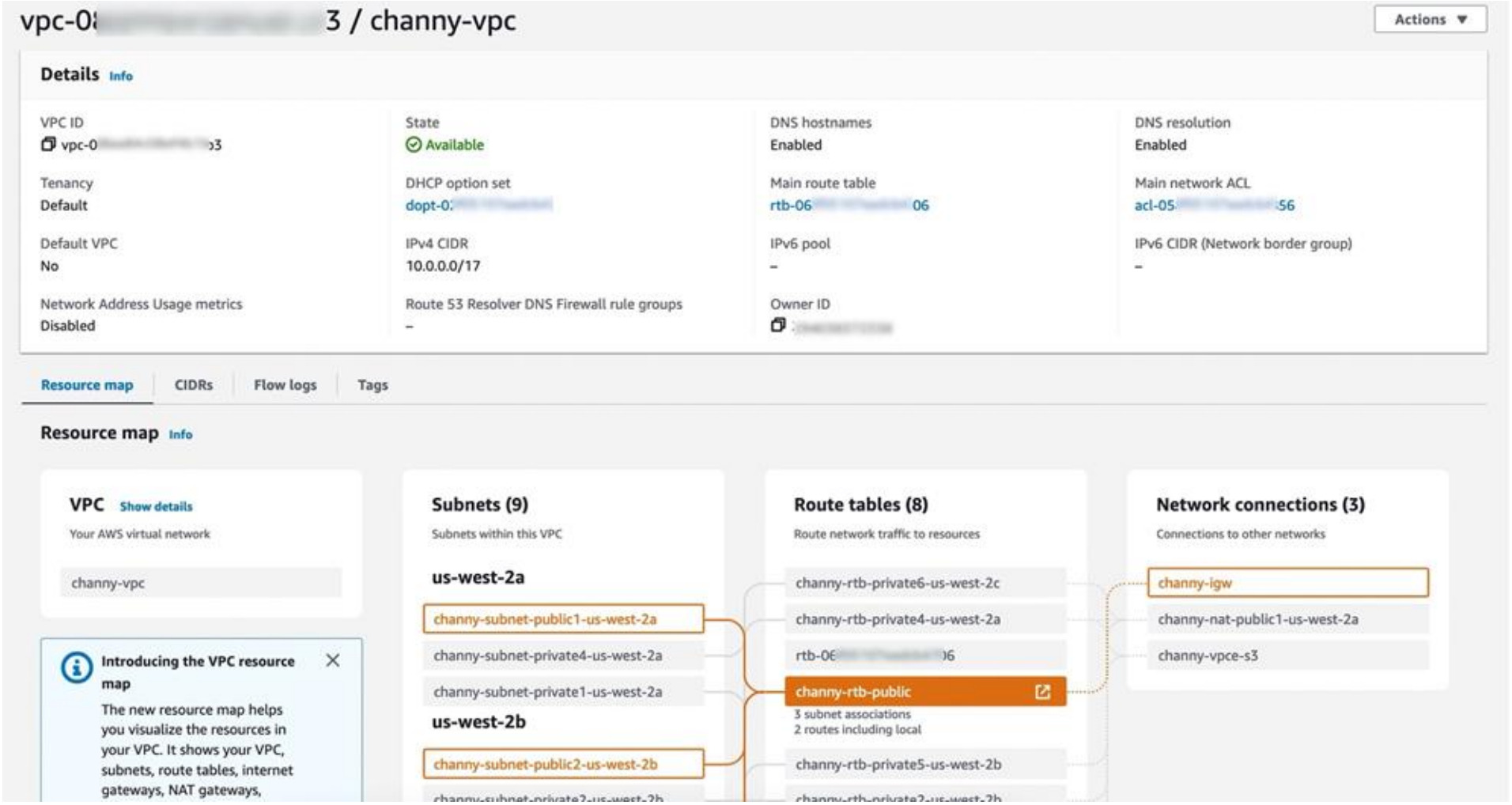
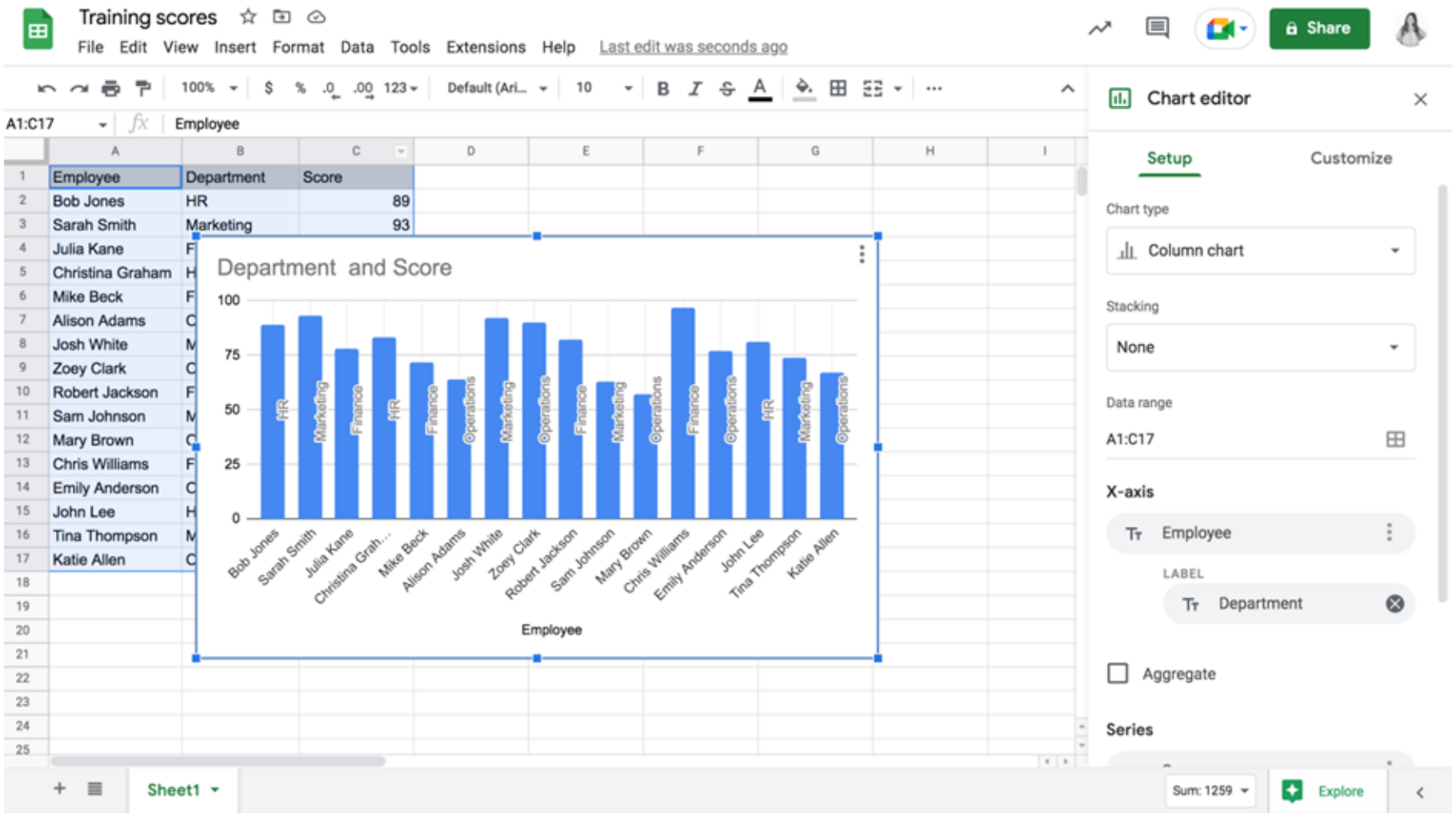


Image generated by DALL·E 3.

Language Agents

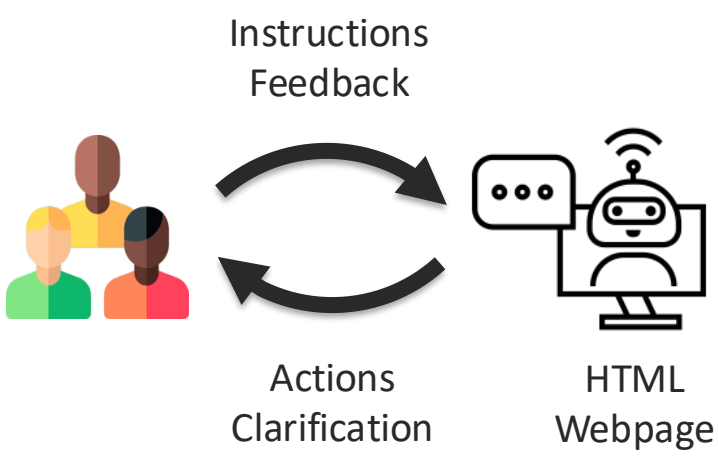


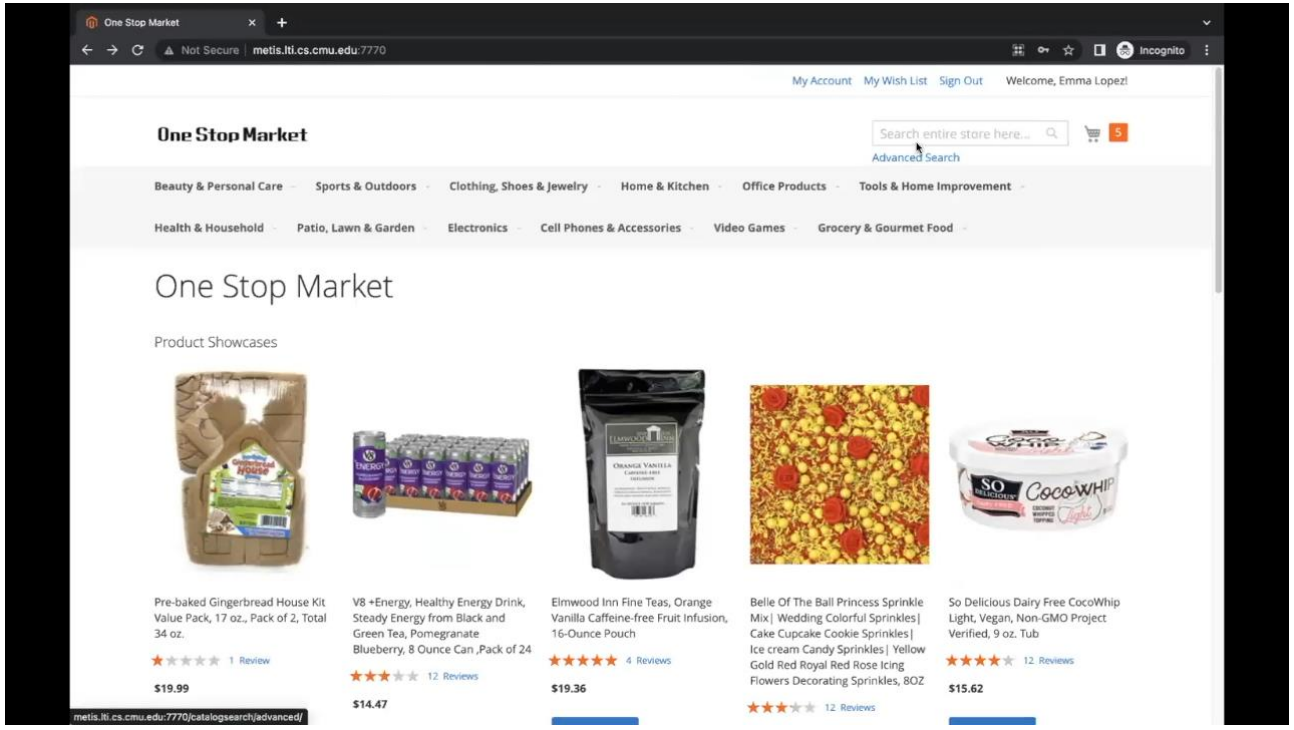
Language Agents



WebArena Environment

Example task: Purchase a set of earphones with at least 4.5 stars in rating and ship it to me.

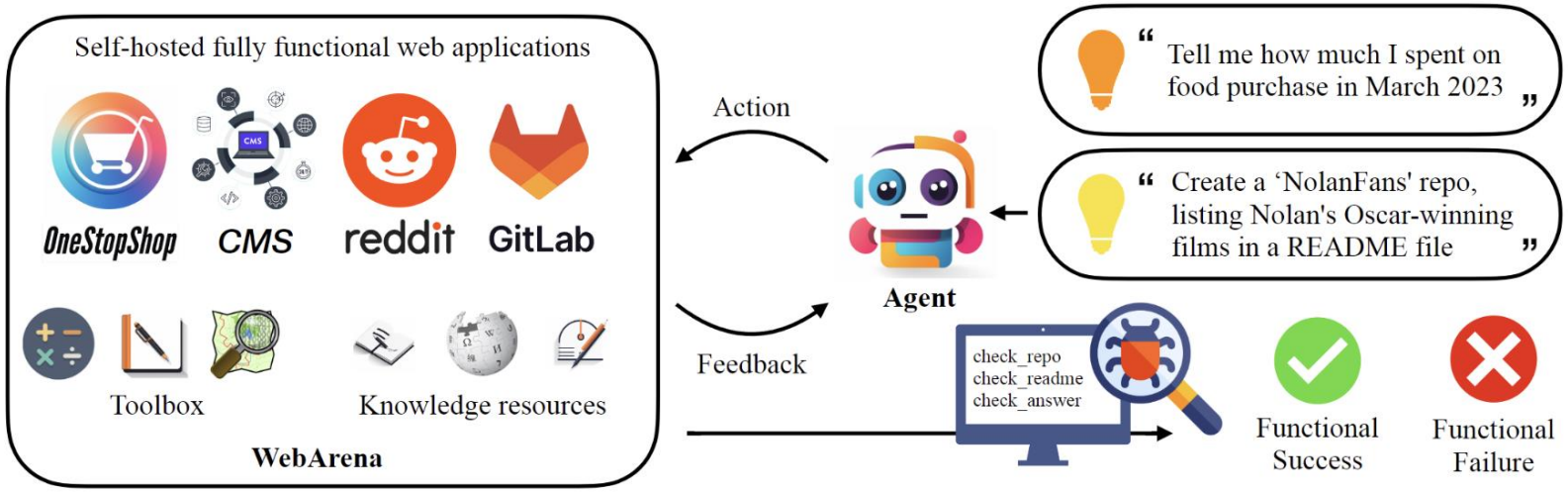




[Zhou et al., WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR 2024]

multisensory intelligence

WebArena Environment

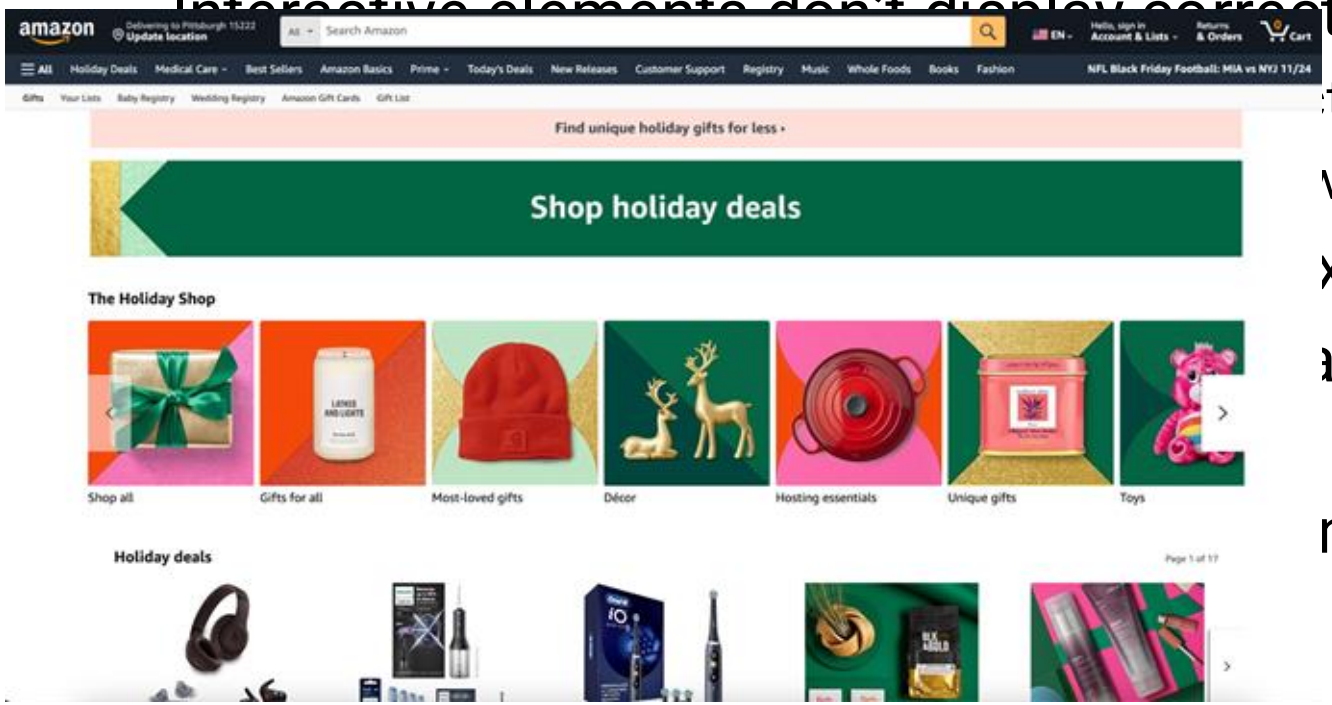


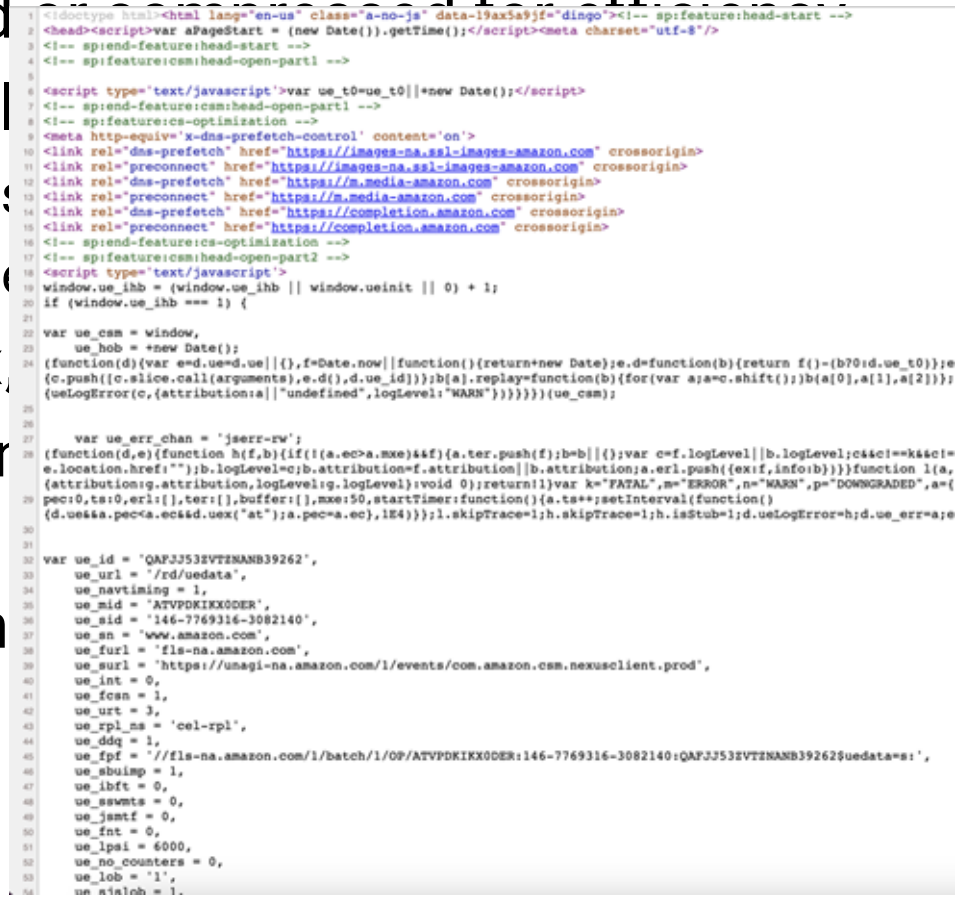
- Websites from four popular categories (shopping, CMS, Reddit, GitLab)
  - Self-hosted open source re-implementations
  - Data from real websites (Amazon, Reddit, GitHub)
- Tasks easy for humans (78% success) but difficult for LLM agents (14%)
- **But:** Tasks are designed to use just text and HTML source code

[Zhou et al., WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR 2024]

HTML is Insufficient

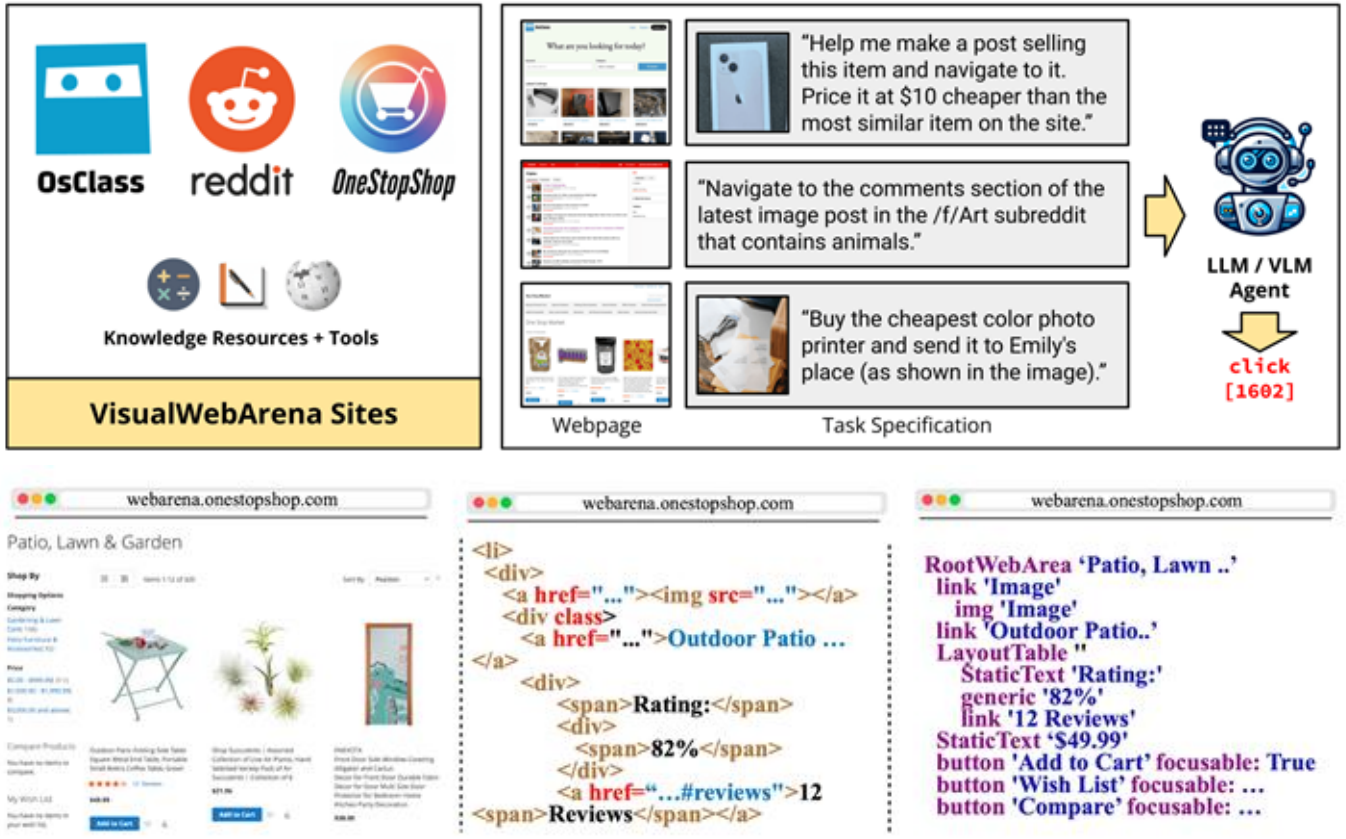
- Messy HTML, JavaScript: usually minified or compressed for efficiency
- Interactive elements don't display correctly





VisualWebArena: A Visually Grounded Benchmark

- Benchmark and track the progress of **multimodal agents**

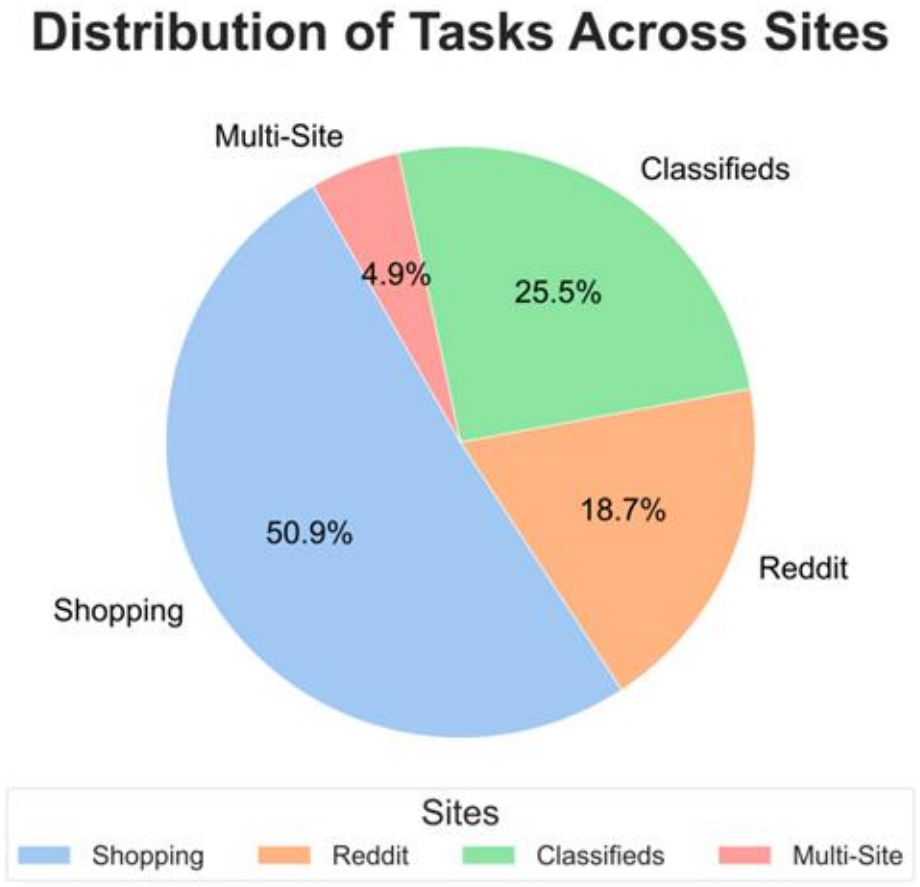


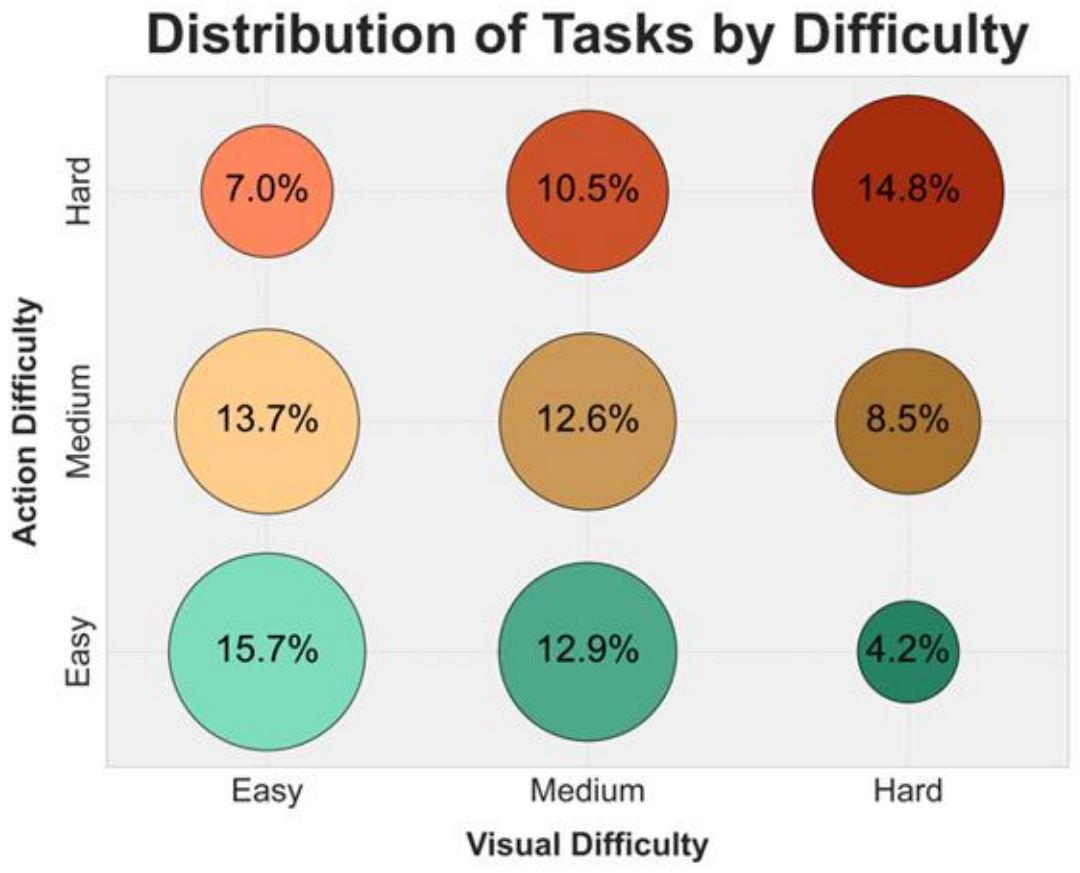
Action Type a	Description
`click [elem]`	Click on element `elem`.
`hover [elem]`	Hover on element `elem`.
`type [elem] [text]`	Type `text` on element `elem`.
`press [key_comb]`	Press a key combination.
`new_tab`	Open a new tab.
`tab_focus [index]`	Focus on the i-th tab.
`tab_close`	Close current tab.
`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.

[Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. ACL 2024]

multisensory intelligence

VisualWebArena: A Visually Grounded Benchmark



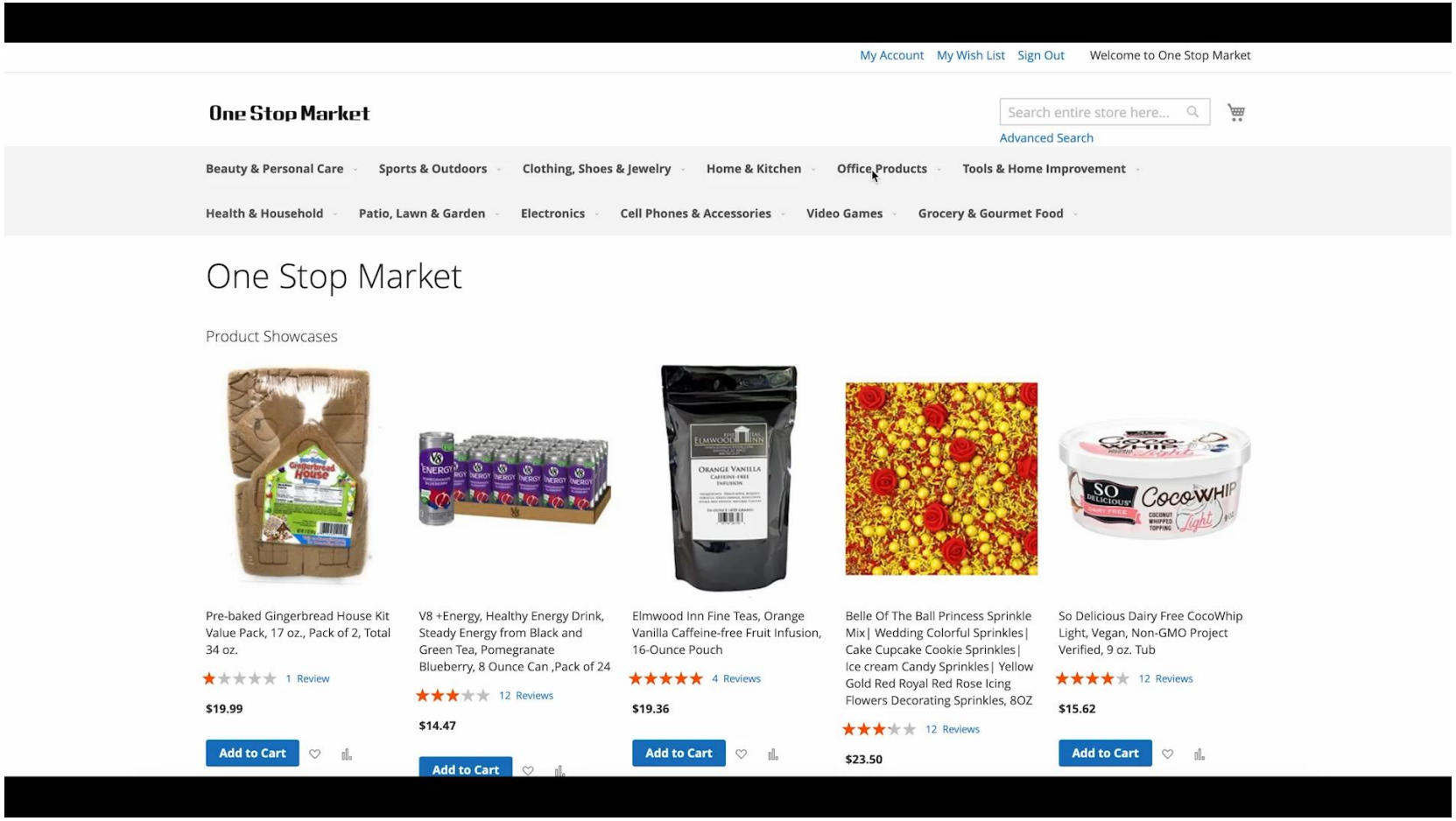


[Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. ACL 2024]

multisensory intelligence

VisualWebArena Shopping Example

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



[Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. ACL 2024]

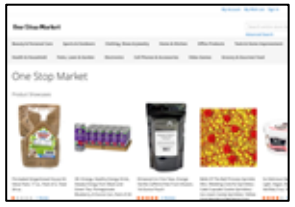
Execution-Based Evaluation

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

→ `exact_match(â, "US0378331005")`


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

→
```
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")
```



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

→ `eval_fuzzy_image_match(s, a*)`

"Navigate to my listing of the white car and change the price to $25,000. Update the price in the description as well."

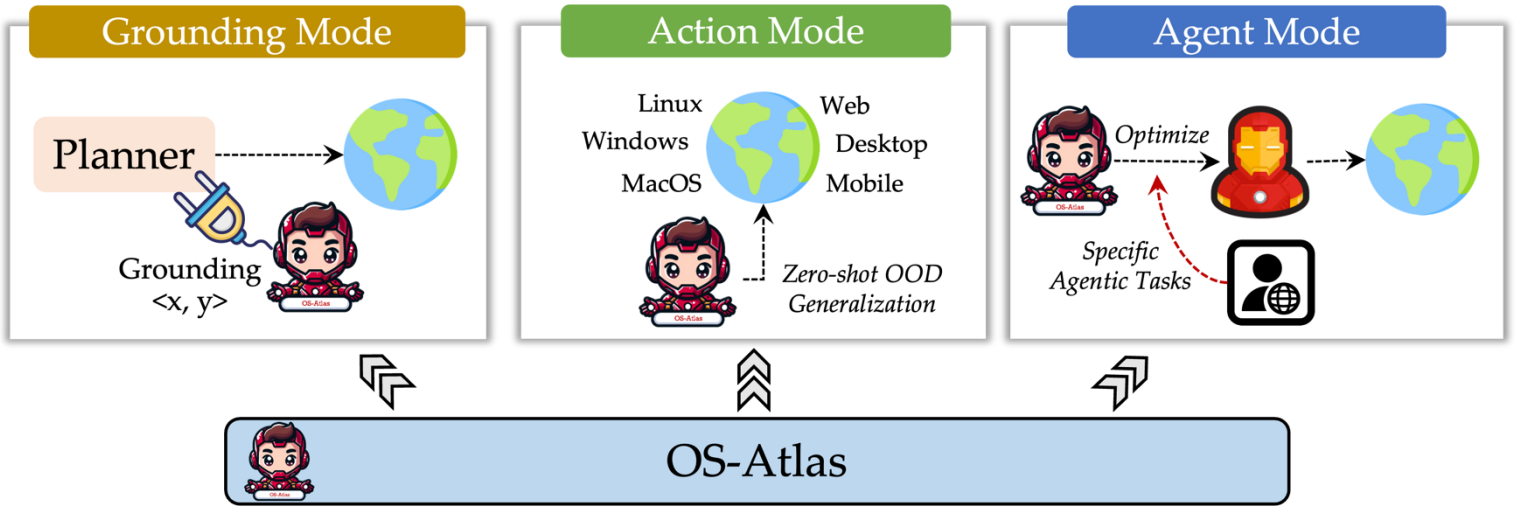
→
```
url="/index.php?page=item&id=84144"
must_include(â, "$25000 |OR| $25,000")
must_exclude(â, "$30000 |OR| $30,000")
```

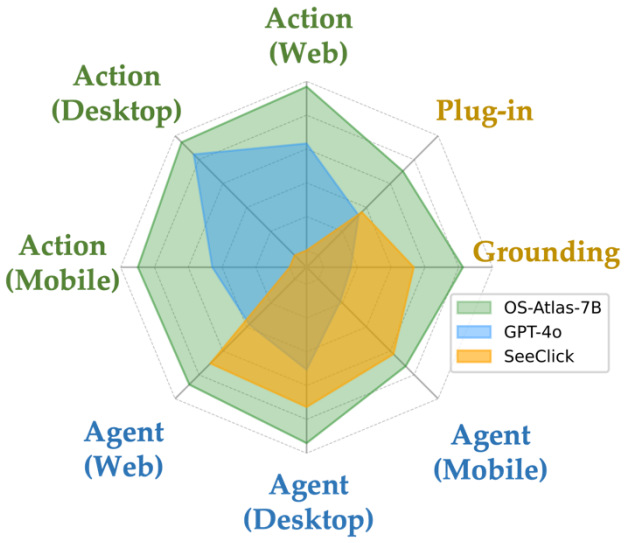
Webpage(s) **Task Specification** **Automatic Functional Evaluation**

[Koh et al., VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. ACL 2024]

OS-Atlas

Web + desktop + mobile GUI grounding



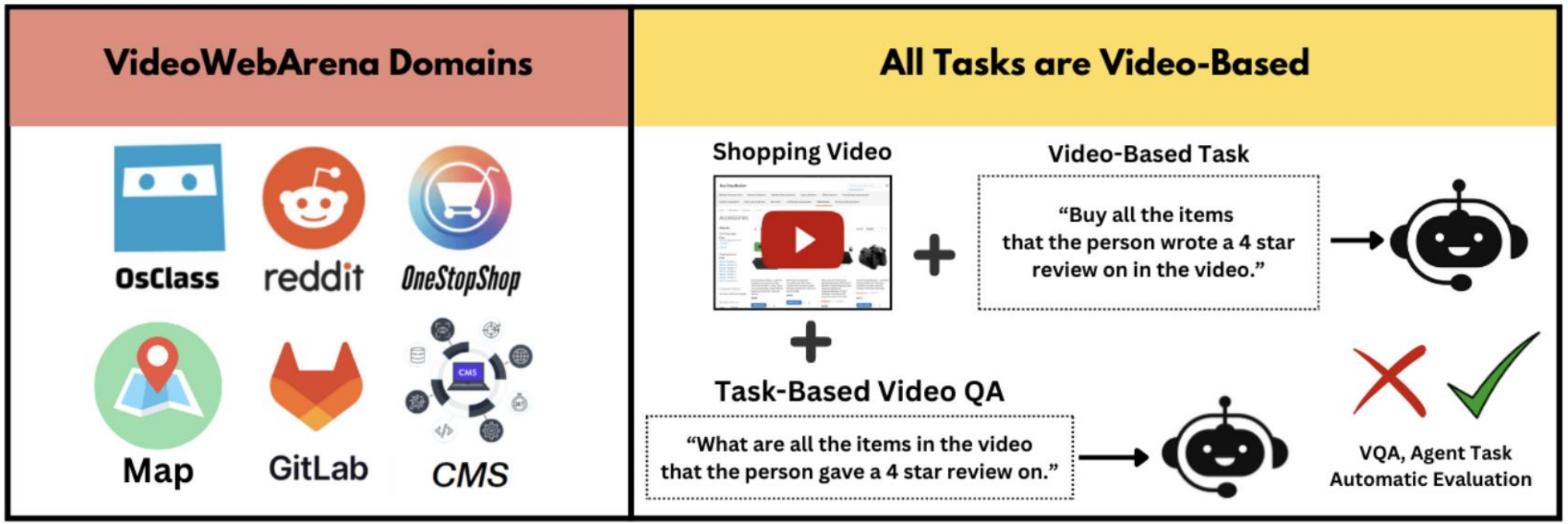


[Wu et al., OS-Atlas: A Foundation Action Model for Generalist GUI Agents. ICLR 2025]

multisensory intelligence

VideoWebArena

Extension to video understanding web tasks

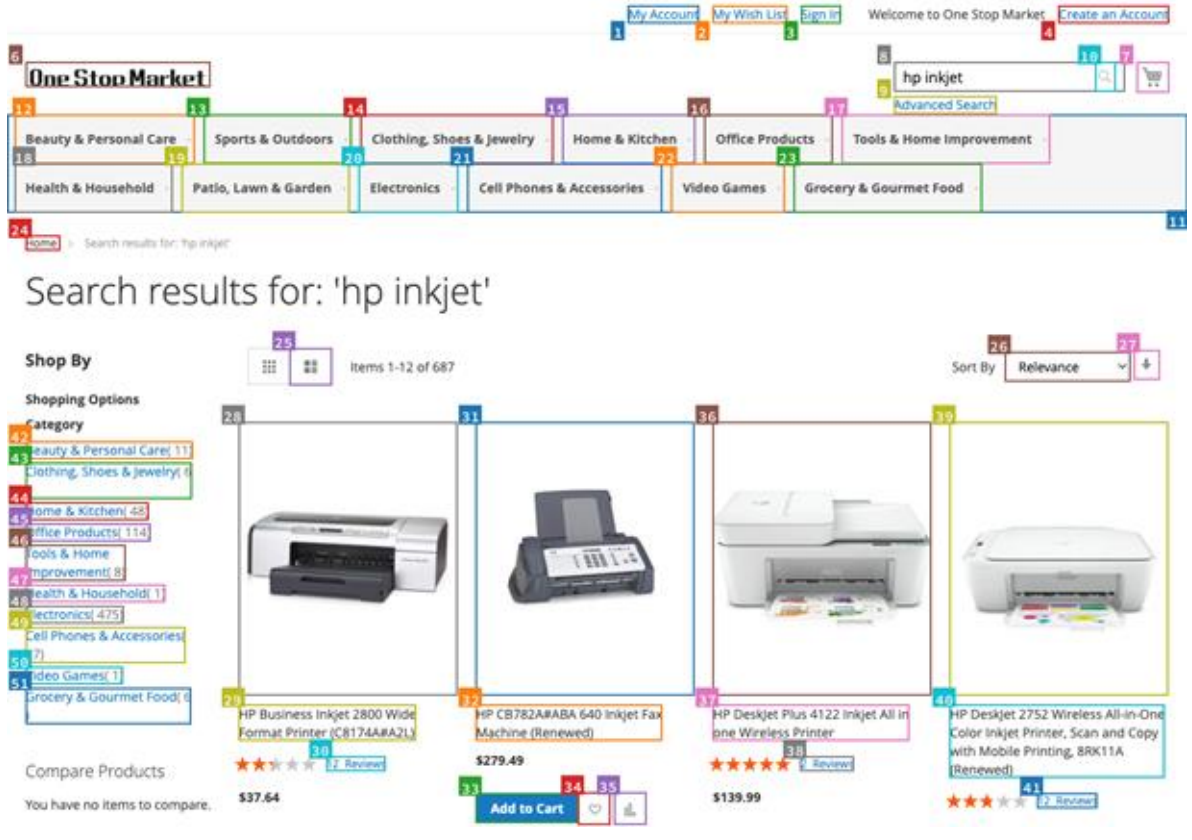


[Jang et al., VideoWebArena: Evaluating Multimodal Agents on Video Understanding Web Tasks. ICLR 2025]

multisensory intelligence

VLMs as Agents

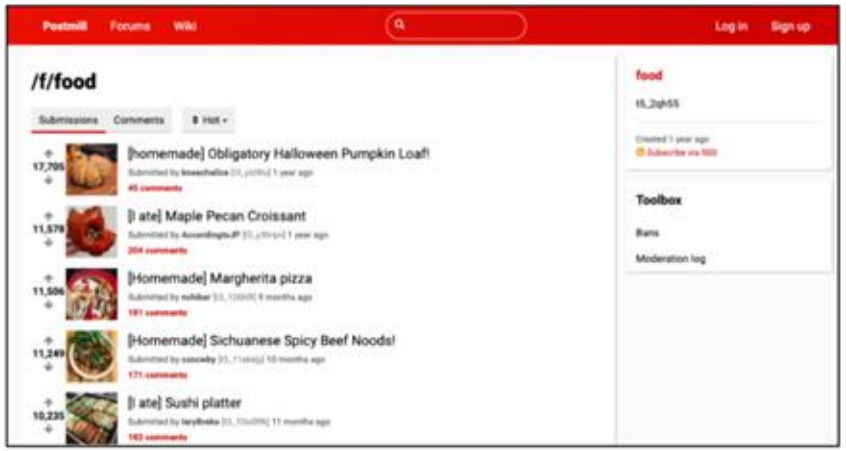
```
Tab 0 (current): Search results for: 'hp inkjet'

[1] RootWebArea "Search results for: 'hp inkjet'" focused: True
        [81] link 'My Account'
        [82] link 'My Wish List'
        [83] link 'Sign Out'
        [4086] StaticText 'Welcome to One Stop Market'
        [37] link 'Skip to Content'
        [23] link 'store logo'
                [39] img 'one_stop_market_logo'
        [40] link '\ue611 My Cart'
        [278] StaticText 'Search'
        [163] combobox '\ue615 Search' autocomplete: both hasPopup: listbox required: False expanded: False
                [426] StaticText 'hp inkjet'
        [281] link 'Advanced Search'
        [120] button 'Search' disabled: True
        [4080] tablist '' multiselectable: False orientation: horizontal
                [4082] tabpanel ''
                        [2326] menu '' orientation: vertical
                                [3077] menuitem '\ue622 Beauty & Personal Care' hasPopup: menu
                                [3142] menuitem '\ue622 Sports & Outdoors' hasPopup: menu
                                [3152] menuitem '\ue622 Clothing, Shoes & Jewelry' hasPopup: menu
                                [3166] menuitem '\ue622 Home & Kitchen' hasPopup: menu
                                [3203] menuitem '\ue622 Office Products' hasPopup: menu
                                [3211] menuitem '\ue622 Tools & Home Improvement' hasPopup: menu
                                [3216] menuitem '\ue622 Health & Household' hasPopup: menu
                                [3222] menuitem '\ue622 Patio, Lawn & Garden' hasPopup: menu
                                [3227] menuitem '\ue622 Electronics' hasPopup: menu
                                [3288] menuitem '\ue622 Cell Phones & Accessories' hasPopup: menu
                                [3303] menuitem '\ue622 Video Games' hasPopup: menu
                                [3316] menuitem '\ue622 Grocery & Gourmet Food' hasPopup: menu
        [47] link 'Home '
        [12] main ''
                [32] heading "Search results for: 'hp inkjet'"
                [264] StaticText 'View as'
                [146] strong 'Grid'
                [147] link 'View as \ue60b List'
                [148] StaticText 'Items '
                [151] StaticText '-'
                [153] StaticText '12'
                [154] StaticText ' of '
                [156] StaticText '687'
                [269] StaticText 'Sort By'
                [158] combobox 'Sort By' hasPopup: menu expanded: False
                [159] link '\ue614 Set Ascending Direction'
                [424] link 'Image'
                        [1010] img 'Image'
                [1011] link 'HP Business Inkjet 2800 Wide Format Printer (CB174A#A2L)'
                [720] LayoutTable ''
                        [1451] StaticText 'Rating:'
                        [1232] generic '47%'
                        [1869] link '12 \xa0Reviews'
                [1871] StaticText '$37.64'
                [1600] link 'Image'
```



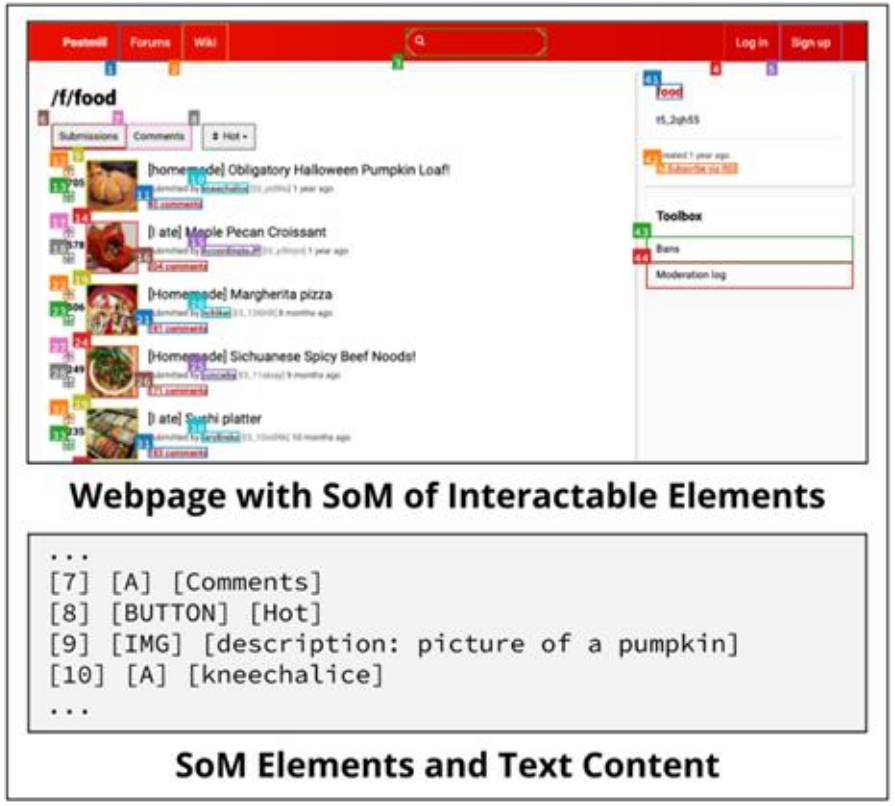
Accessibility tree / HTML: Cluttered with unnecessary information, long and confusing context.

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

VLMs as Agents



Original Webpage



Webpage with SoM of Interactable Elements

SoM Elements and Text Content

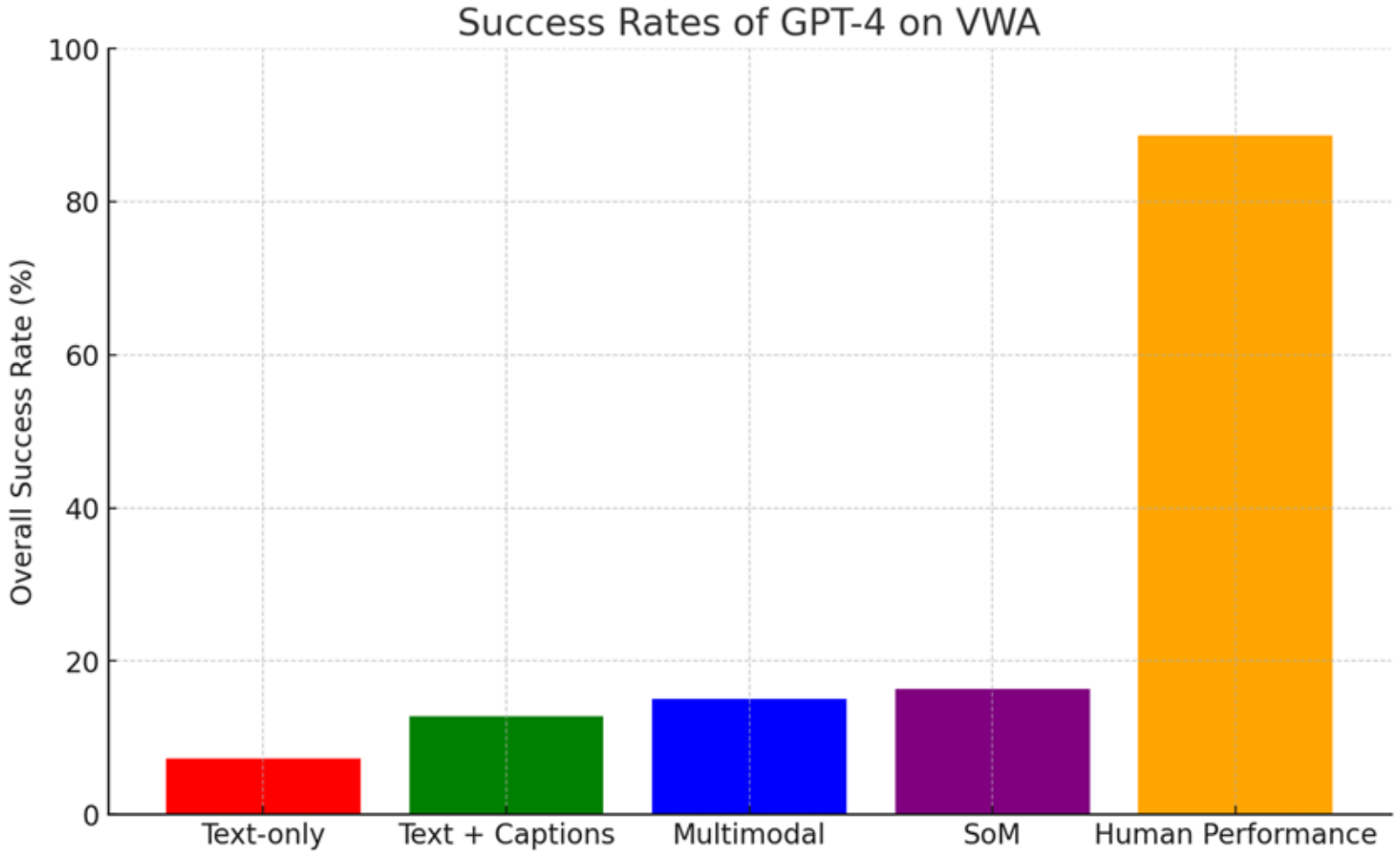
"I'm trying to find this post. Navigate to the comment section for it."

LLM / VLM Agent

click [31]

multisensory intelligence

Baseline Agents



multisensory intelligence

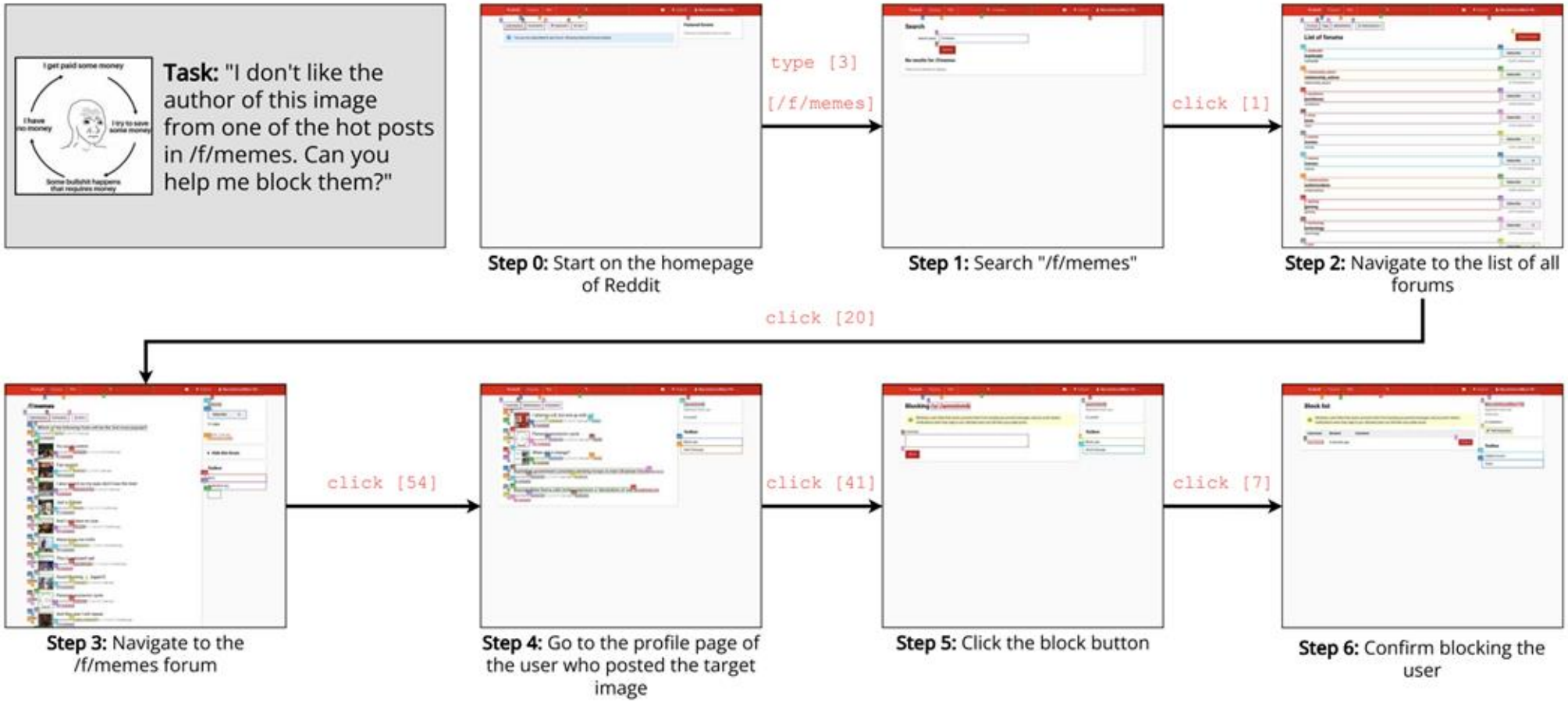
Text-only LLM Agents

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

Multimodal Agents

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%

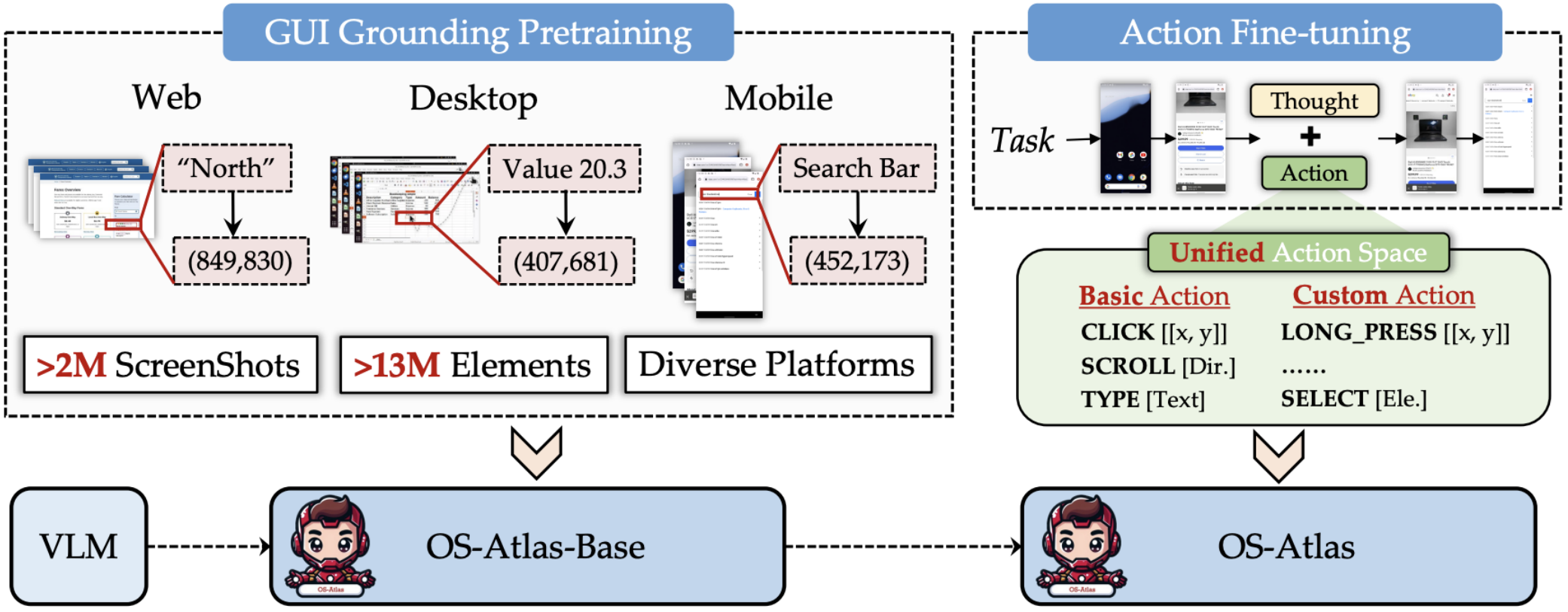
Example Success Trajectory



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

OS-Atlas

Web + desktop + mobile GUI grounding



[Wu et al., OS-Atlas: A Foundation Action Model for Generalist GUI Agents. ICLR 2025]

multisensory intelligence

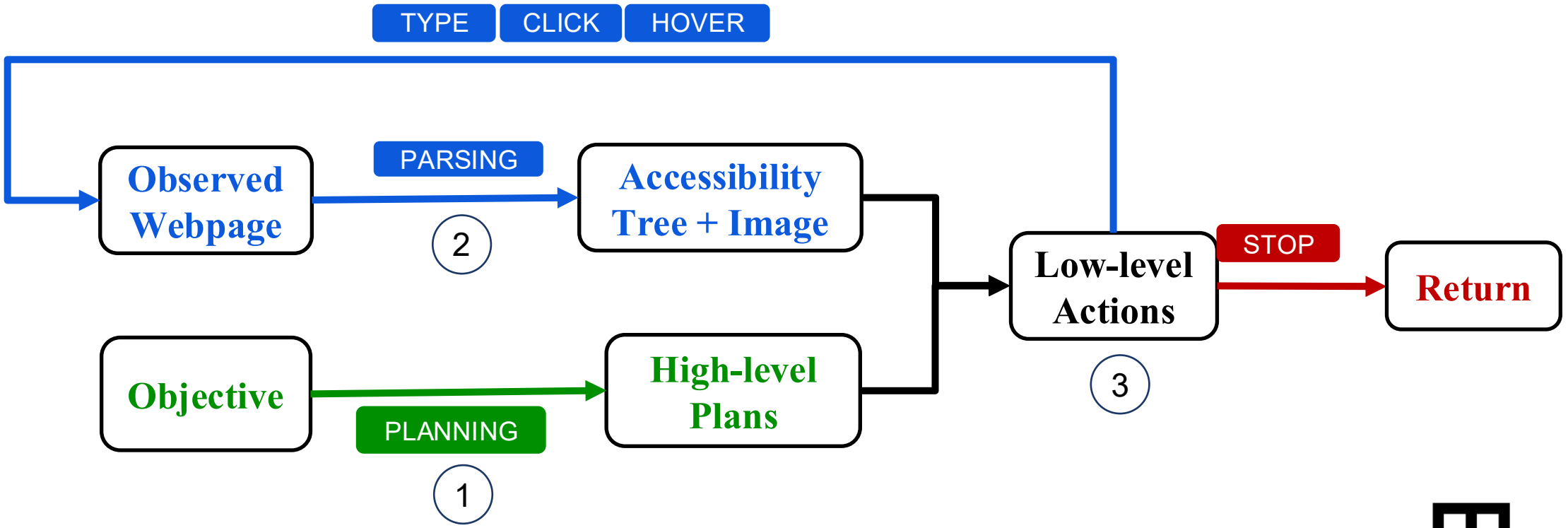
Common Failure Modes

- **Long horizon reasoning and planning**
  - Getting stuck in loops
  - Correctly performing tasks but undoing them

- **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?”)

Agents + Reasoning

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

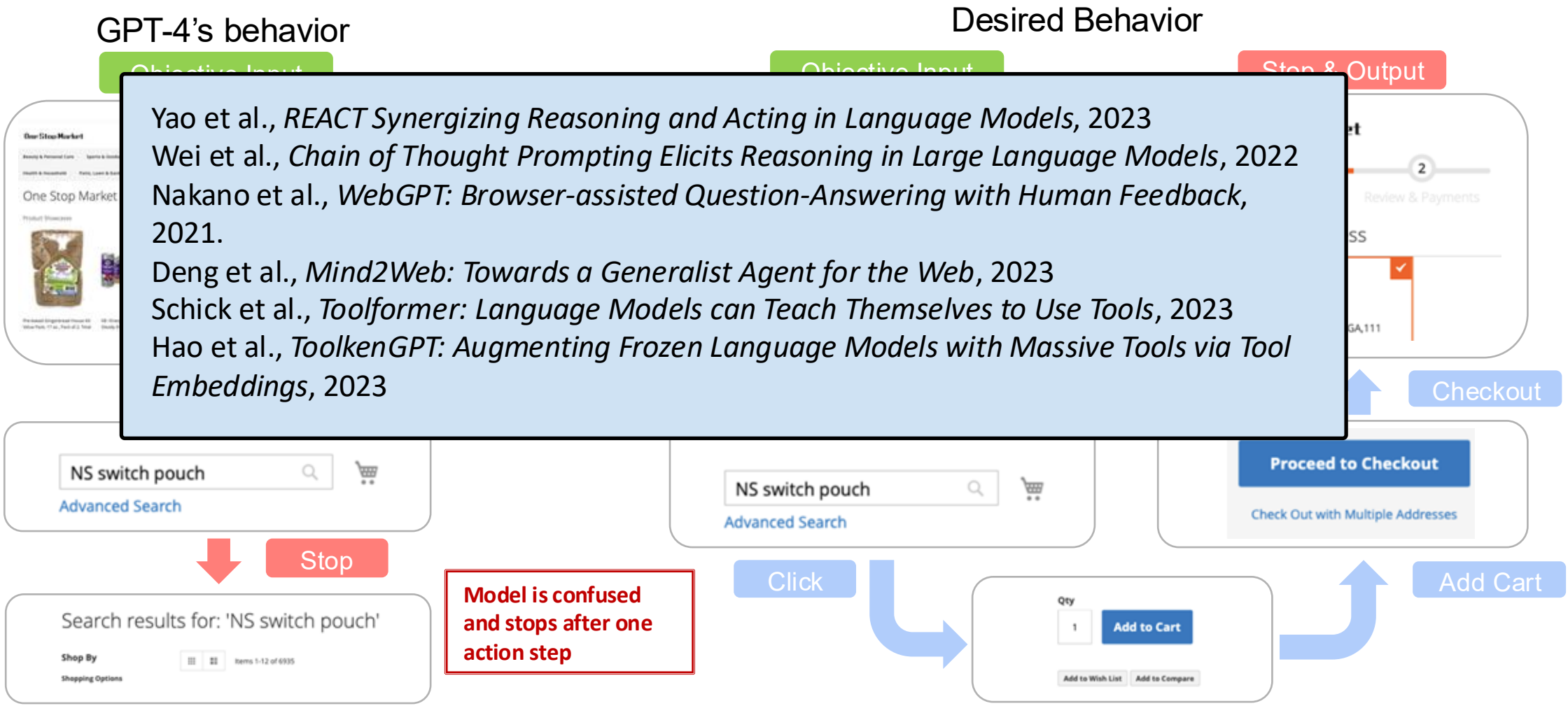


[Zhou et al., WebArena: A Realistic Web Environment for Building Autonomous Agents. ICLR 2024]
[Jang et al., VideoWebArena: Evaluating Multimodal Agents on Video Understanding Web Tasks. ICLR 2025]

multisensory intelligence

Agents + Reasoning

Task: Buy the highest rated product from the NS switch pouch category within a budget under 60 dollars.



Yao et al., *REACT Synergizing Reasoning and Acting in Language Models*, 2023
Wei et al., *Chain of Thought Prompting Elicits Reasoning in Large Language Models*, 2022
Nakano et al., *WebGPT: Browser-assisted Question-Answering with Human Feedback*, 2021.
Deng et al., *Mind2Web: Towards a Generalist Agent for the Web*, 2023
Schick et al., *Toolformer: Language Models can Teach Themselves to Use Tools*, 2023
Hao et al., *ToolkenGPT: Augmenting Frozen Language Models with Massive Tools via Tool Embeddings*, 2023

Generating High-level Plans

What is the price range of wireless earphone in the One Stop Market?

Generated High-level Plan

[Search] wireless earphone in the One Stop Market

[Find] the price range of wireless earphone in the One Stop Market

TOO AMBIGUOUS
HARD TO UNDERSTAND

Human Annotated High-level Plan

[Search] wireless earphones in the One Stop Market.

[Sort] the price of the wireless earphones from low to high

[Find] the first item

[Sort] the price of the wireless earphones from high to low

[Find] the first item

[Identify] the price range of wireless earphones in the One Stop Market

multisensory intelligence

Agents + Human-in-the-loop

Asking humans for clarification when it is uncertain about generated plans

What is the price range of wireless earphone in the One Stop Market?

Zero-shot high-level plan

- [Search] wireless earphone in the One Stop Market
- [Find] the price range of wireless earphone in the One Stop Market

Too ambiguous

Few-shot high-level plan

- [Search] wireless earphone in the One Stop Market
- [Find] the price of the first wireless earphone
- [Read] the price range of wireless earphones in the One Stop Market

Didn't sort

Few-shot high-level plan with human clarification

- [Search] wireless earphone in the One Stop Market
- [Sort] the price of the wireless earphones from low to high
- [find] the lowest priced wireless earphone
- [sort] the price of wireless earphones from high to low
- [find] the highest priced wireless earphone

Subsequent model generations become correct after clarification

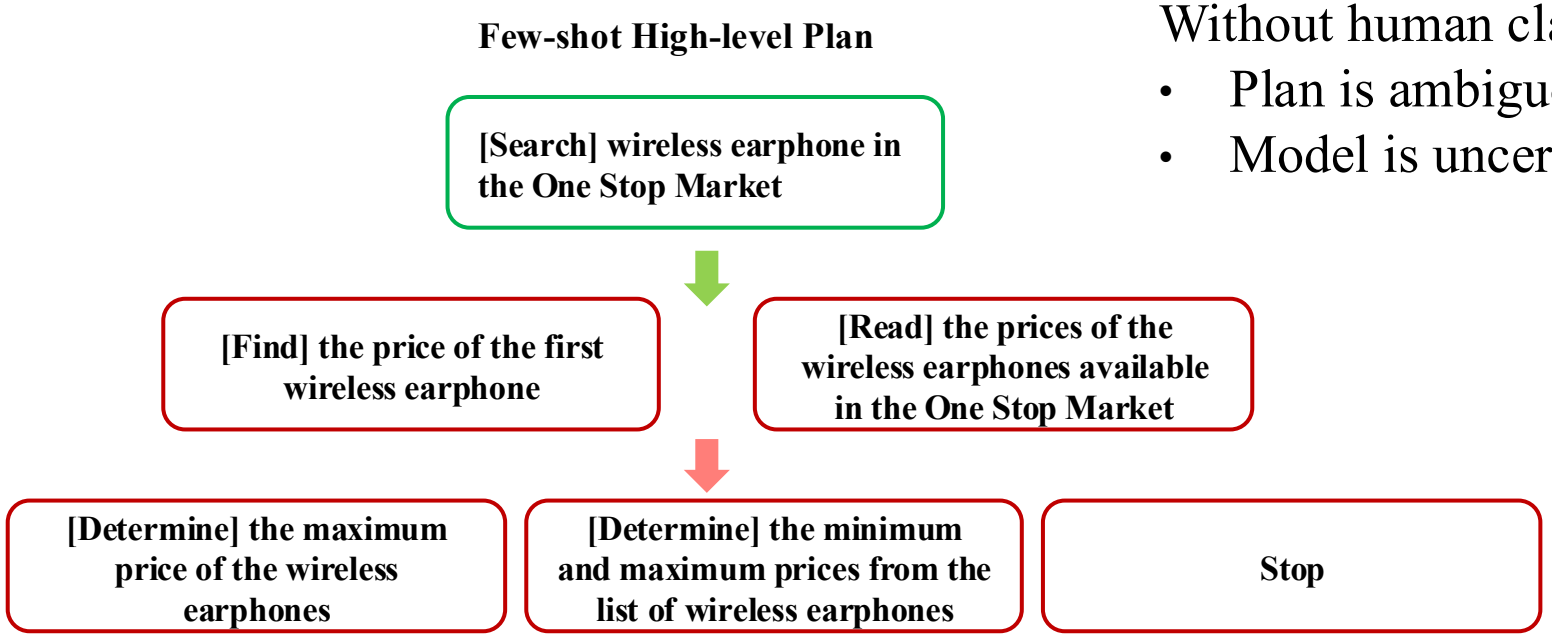
Ground truth high-level plan

- [Search] wireless earphones in the One Stop Market.
- [Sort] the price of the wireless earphones from low to high
- [Find] the first item
- [Sort] the price of the wireless earphones from high to low
- [Find] the first item
- [Identify] the price range of wireless earphones in the One Stop Market

[Kaufmann et al., A Survey of Reinfocement Learning from Human Feedback. arXiv 2024]
[Wang et al., Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. CHI 2024]

Estimating Uncertainty

Sampling high-level plans generated by the model multiple times to estimate uncertainty



Without human clarification:

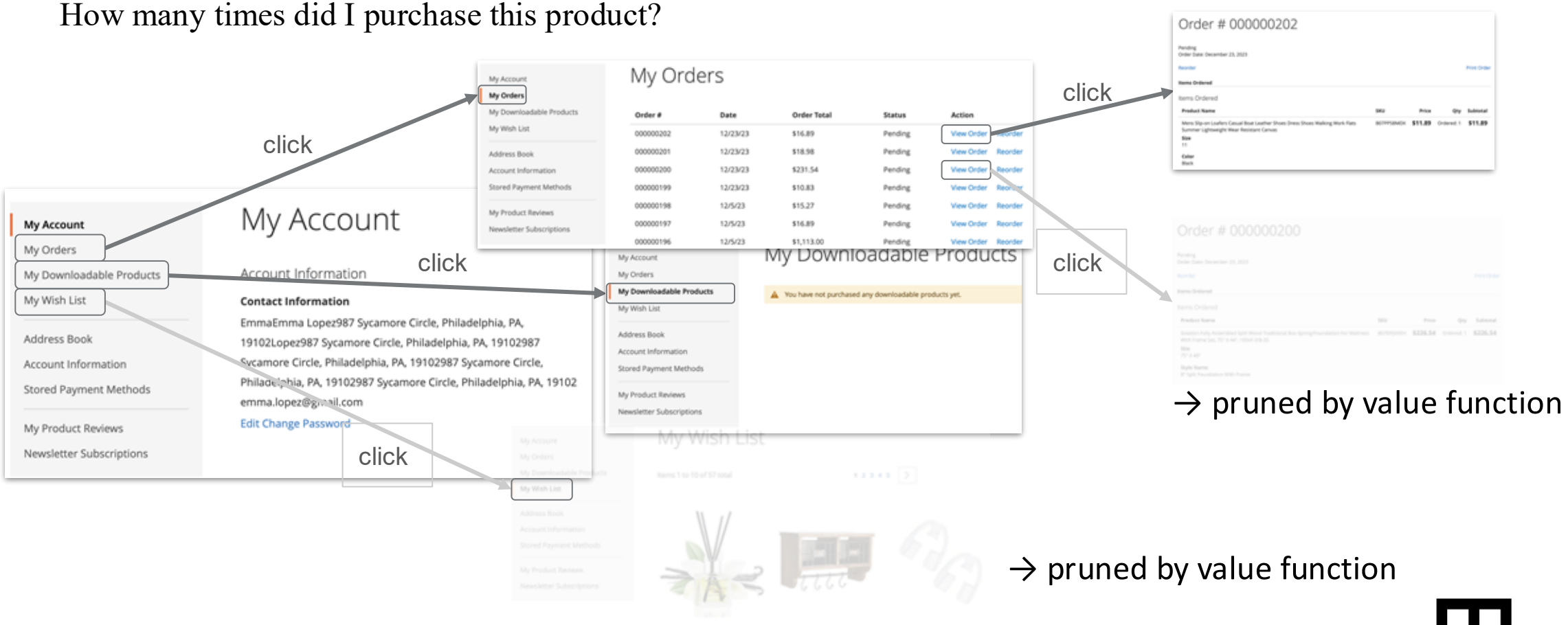
- Plan is ambiguous
- Model is uncertain about the overall plan

multisensory intelligence

Agents + Search

Searching over low-level actions – recall reinforcement learning

How many times did I purchase this product?



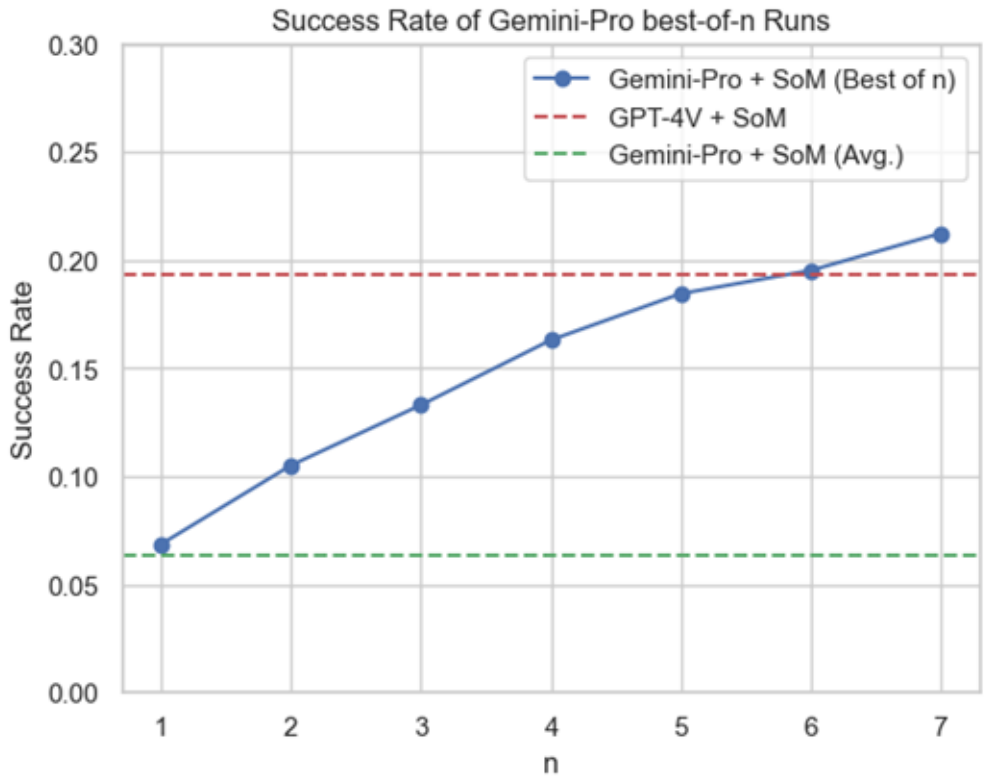
[Yao et al., Tree of Thoughts: Deliberate Problem Solving with Large Language Models. NeurIPS 2023]
[Koh et al., Tree Search for Language Model Agents. TMLR 2025]

multisensory intelligence

Agents + Search

Idea: Allow the agent to execute and coordinate multiple instances in parallel for better exploration.

Research Problem: Learning a good value function for search and re-ranking of trajectories.

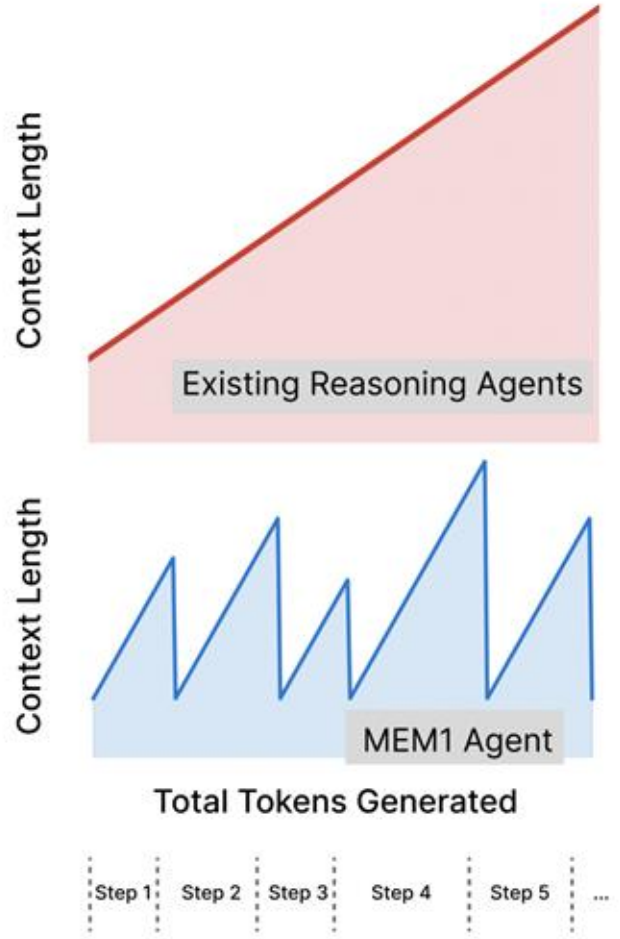


Gemini-Pro performance on VisualWebArena. With an accurate value function, exploration and search can substantially improve performance.

[Yao et al., Tree of Thoughts: Deliberate Problem Solving with Large Language Models. NeurIPS 2023]
[Koh et al., Tree Search for Language Model Agents. TMLR 2025]

Agents + Memory

https://github.com/MIT-MI/MEM1



Key problem: As reasoning context grows longer (more modalities and longer context), new complexity and forgetting challenges.

Proposal: Introduce an internal state that integrates reasoning and memory. At each step, the model **updates this internal state and discards all previous information**.

Advantages:

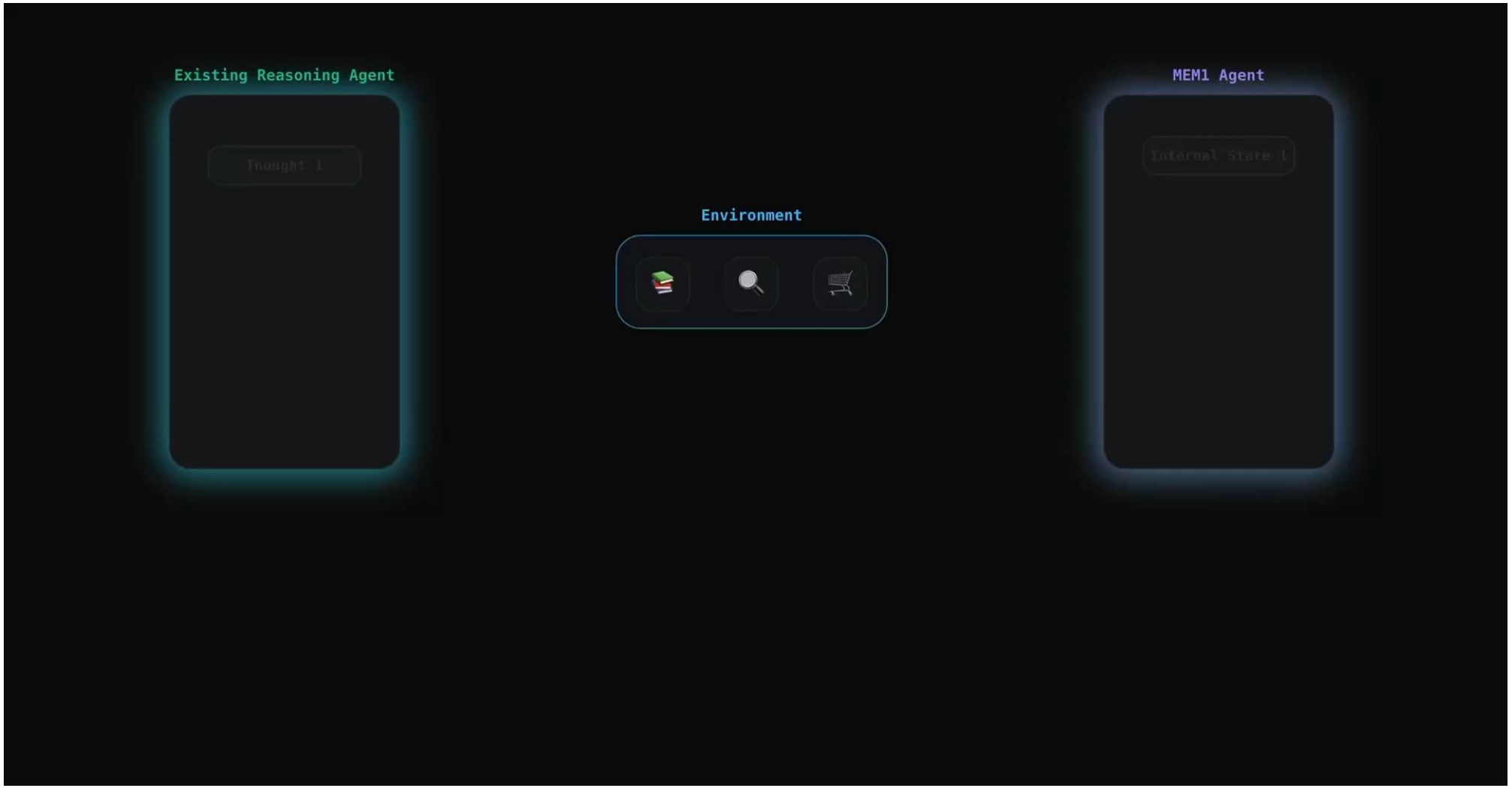
1. Enable near-constant memory ✅
2. Support end-to-end optimization ✅

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

MEM1: Memory Efficient Reasoning

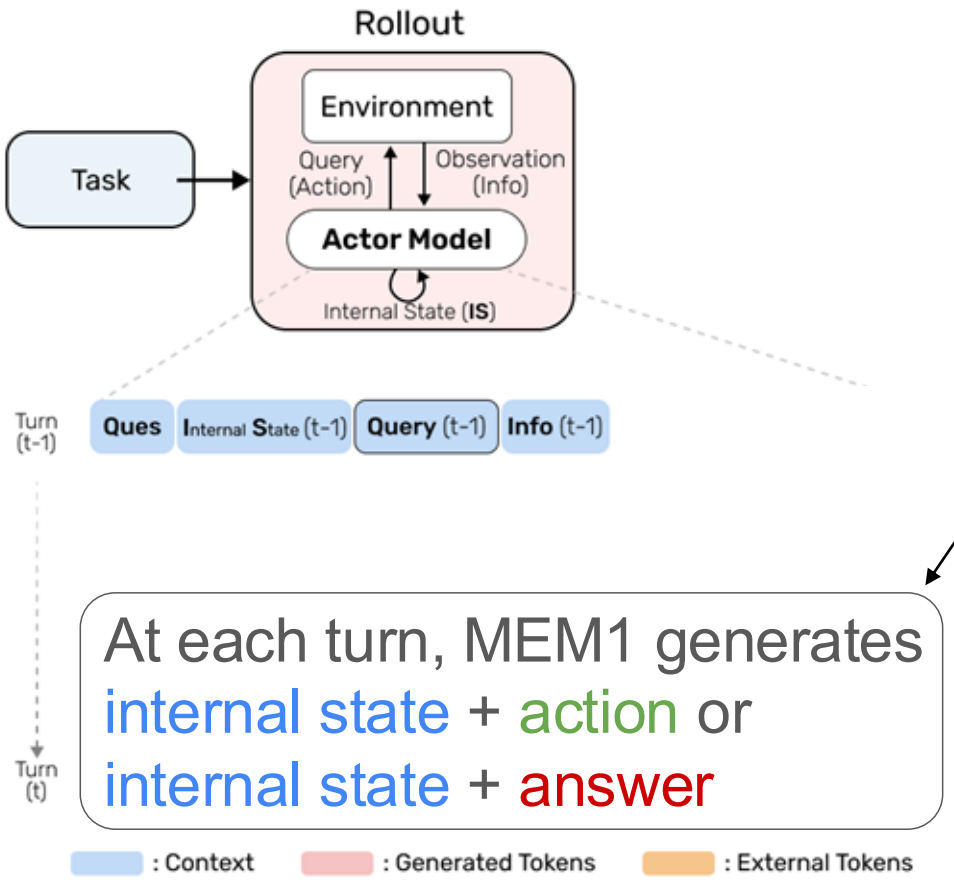
https://github.com/MIT-MI/MEM1



[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

MEM1: Rollout with Internal State

https://github.com/MIT-MI/MEM1



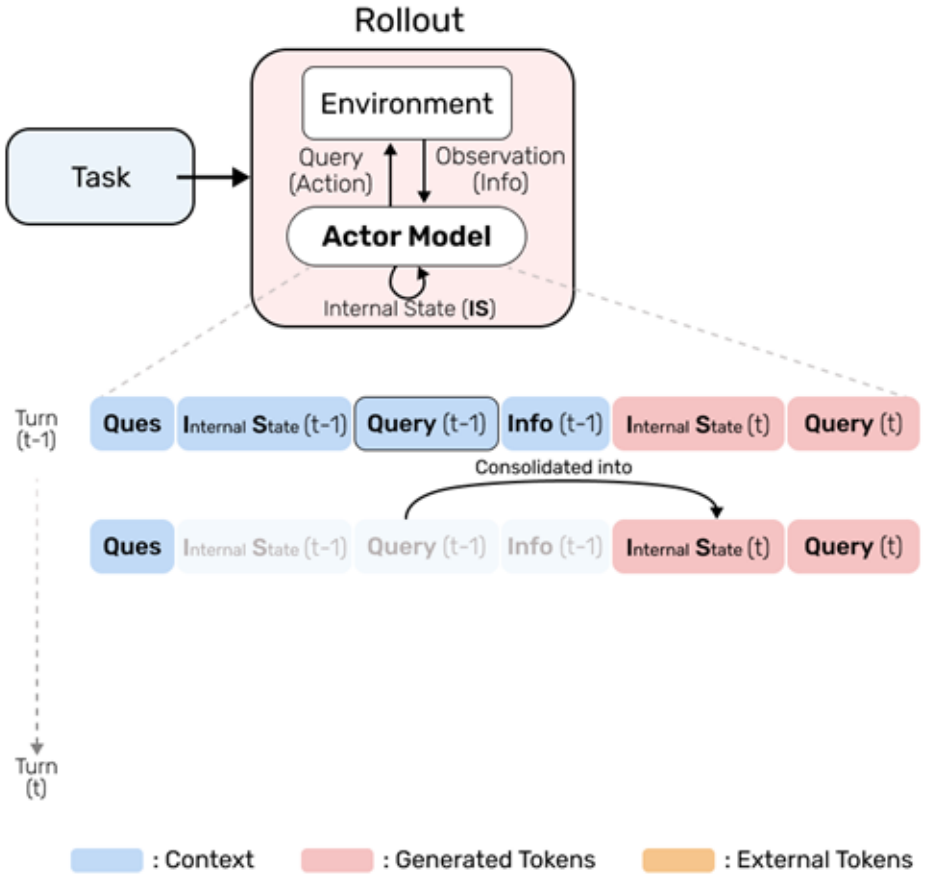
Prompt 1: Single-Objective Task (QA)

```
You will answer a complex question through iterative
reasoning, summarization, and web searches.

At each step, you can see the question, previous reasoning
and summary in <internal_state> ... </internal_state>,
search query in <search> ... </search>, and the returned
information in <information> ... </information> (except the
first step where you will be given only the question). Then,
you should:

1. Update a concise, cumulative summary with essential
information and conduct reasoning inside <internal_state>
</internal_state>. This is your persistent memory and
should include all important information from previous
<internal_state> </internal_state> and <information>
</information> (i.e. information and answers already found
for questions).

2. Then choose one:
- Issue a query (i.e., key words / phrases for search)
inside <search> </search> (you may search repeatedly until
the answer is clear). This query will be used to conduct
search and return the results in <information> results
</information>
- Provide the final concise answer (no explanations) if no
additional information is needed inside <answer> </answer>.
The answer should be concise and only contain the words
necessary to answer the question.

Question: [QUESTION]
```

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

MEM1: Rollout with Internal State

https://github.com/MIT-MI/MEM1



Key idea: Memory Consolidation

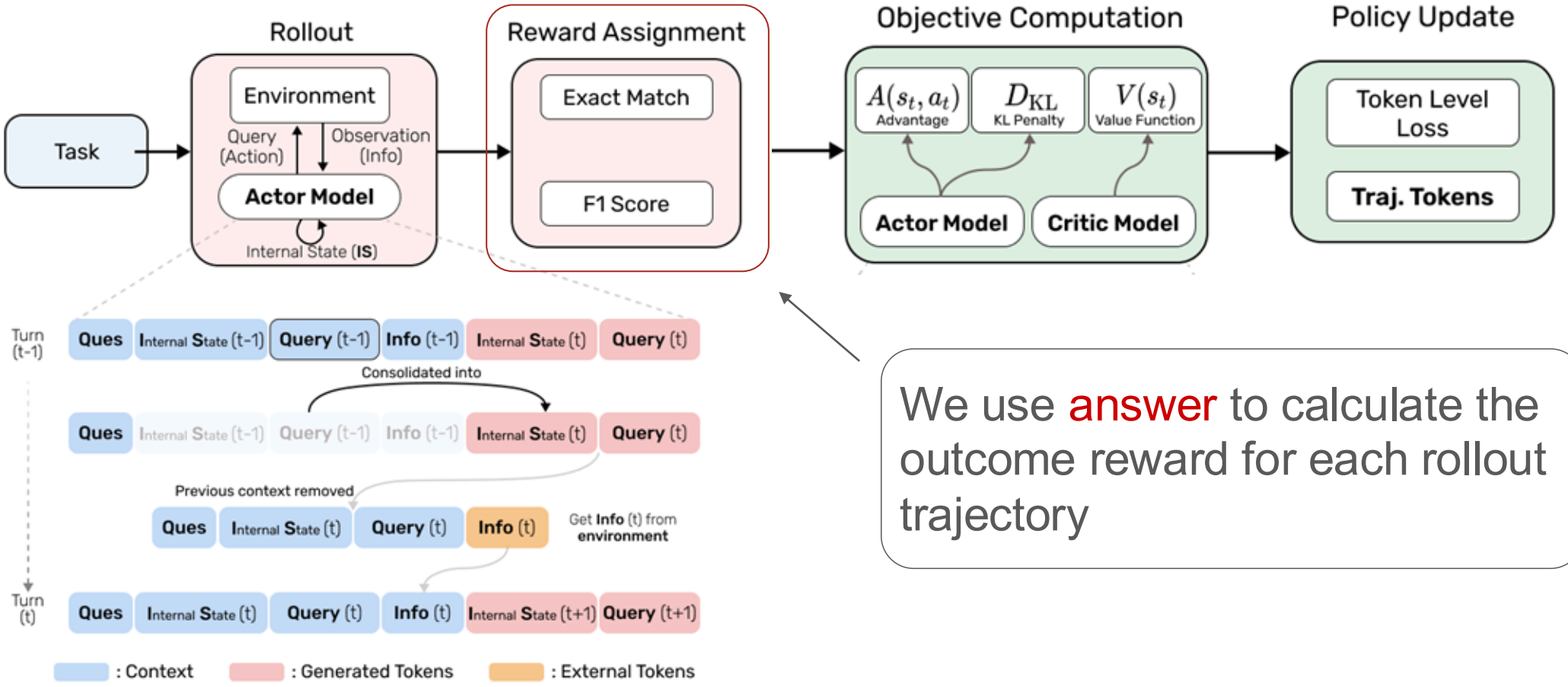
- MEM1 keeps context size constant by pruning old states after each turn.
- All prior knowledge is compressed into the latest *Internal State*, so only the latest internal state, query, and info are kept in memory.
- Forces the agent to continuously integrate, update, and compress knowledge at every step.

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

MEM1: Outcome Reward

https://github.com/MIT-MI/MEM1



We use answer to calculate the outcome reward for each rollout trajectory

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

MEM1 Results: WebAgents

https://github.com/MIT-MI/MEM1

Model	Avg Final Reward ↑	Peak Token ($\times 10^3$) ↓	Dependency ($\times 10^6$) ↓	Inference Time Per Traj (s) ↓
GPT-4o	25.48	5.30 ± 1.23	3.99 ± 1.16	*N/A*
GPT-4o (truncate)	13.82	0.99 ± 0.99	0.81 ± 0.23	*N/A*
GPT-4o (A-MEM)	24.50	1.84 ± 0.06	0.31 ± 0.11	*N/A*
Qwen2.5-7B-Instruct	18.42	5.64 ± 1.34	3.38 ± 0.89	12.31 ± 1.82
Qwen2.5-14B-Instruct	12.34	5.44 ± 0.92	3.30 ± 0.61	18.17 ± 2.32
Agent-FLAN-7B	40.35	3.37 ± 1.12	2.18 ± 1.62	9.95 ± 6.19
Agent-R-8B	63.91	*N/A*	*N/A*	*N/A*
AgentLM-7B	63.60	2.24 ± 0.40	0.28 ± 0.07	3.91 ± 1.07
AgentLM-13B	70.80	2.36 ± 0.46	0.30 ± 0.08	5.23 ± 1.59
MEM1-WebShop	**70.87**	$\mathbf{0.81 \pm 0.10}$	$\mathbf{0.15 \pm 0.16}$	$\mathbf{2.61 \pm 0.48}$

MEM1 trained on WebShop achieves **state-of-the-art performance**, with **the fewest peak tokens** and **the shortest inference time.**

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

MultihopQA Task Construction

https://github.com/MIT-MI/MEM1

Existing datasets often focus on **single-objective tasks**

An example question from HotpotQA:

- *Which US state is nicknamed The Equality State?*

We scale up existing datasets by **composing multiple tasks** into one and training the model on compositional tasks

An example task of our 2-objective HotpotQA dataset:

- *Which US state is nicknamed The Equality State?; Which US State is nicknamed The Beaver State?*

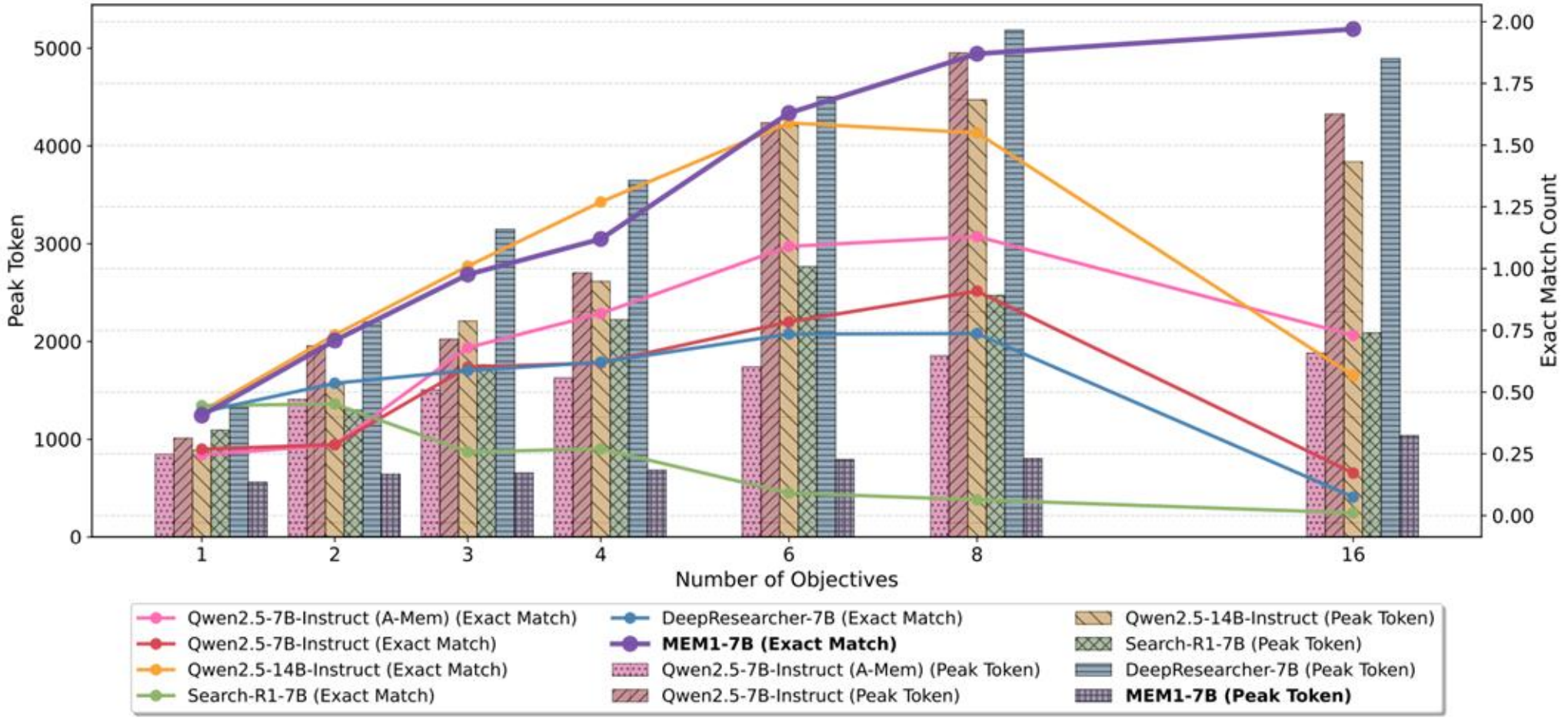
~32.2% of compositional tasks involve two closely related questions

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

MEM1 Results: MultihopQA

https://github.com/MIT-MI/MEM1



When # Questions ↑ -> Performance : Baseline Methods ↓; MEM1 ↑

Context : Baseline Methods ↑; MEM1 –

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

Emergent Memory Behaviors

https://github.com/MIT-MI/MEM1

🔍 MEM1 learns to store key extracted info about each question separately

MEM1: ...
According to the provided information:
- They had 9 World Series appearances in Brooklyn.
- They have had 11 World Series appearances in Los Angeles.
Therefore, ... Now, to answer the second question, ...

From the provided information:
- The show originally ran on ABC from 2011 to 2017 for six seasons...

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

Emergent Memory Behaviors

https://github.com/MIT-MI/MEM1

🔍 MEM1 learns to explicitly update its memory before moving on to the next question

[Omitted History]
Env: [Search Results]

MEM1: : After searching for the information, we found out that Beniamino Di Giacomo was born in Porto Recanati. **Now we can add this information to our cumulative summary: Beniamino Di Giacomo was born in Porto Recanati.** To answer the second question, 'In what city was Tim Alexander born?', we need to perform a web search...

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

Emergent Memory Behaviors

https://github.com/MIT-MI/MEM1

🔍 MEM1 learns to skip one question after several searches without finding useful information.

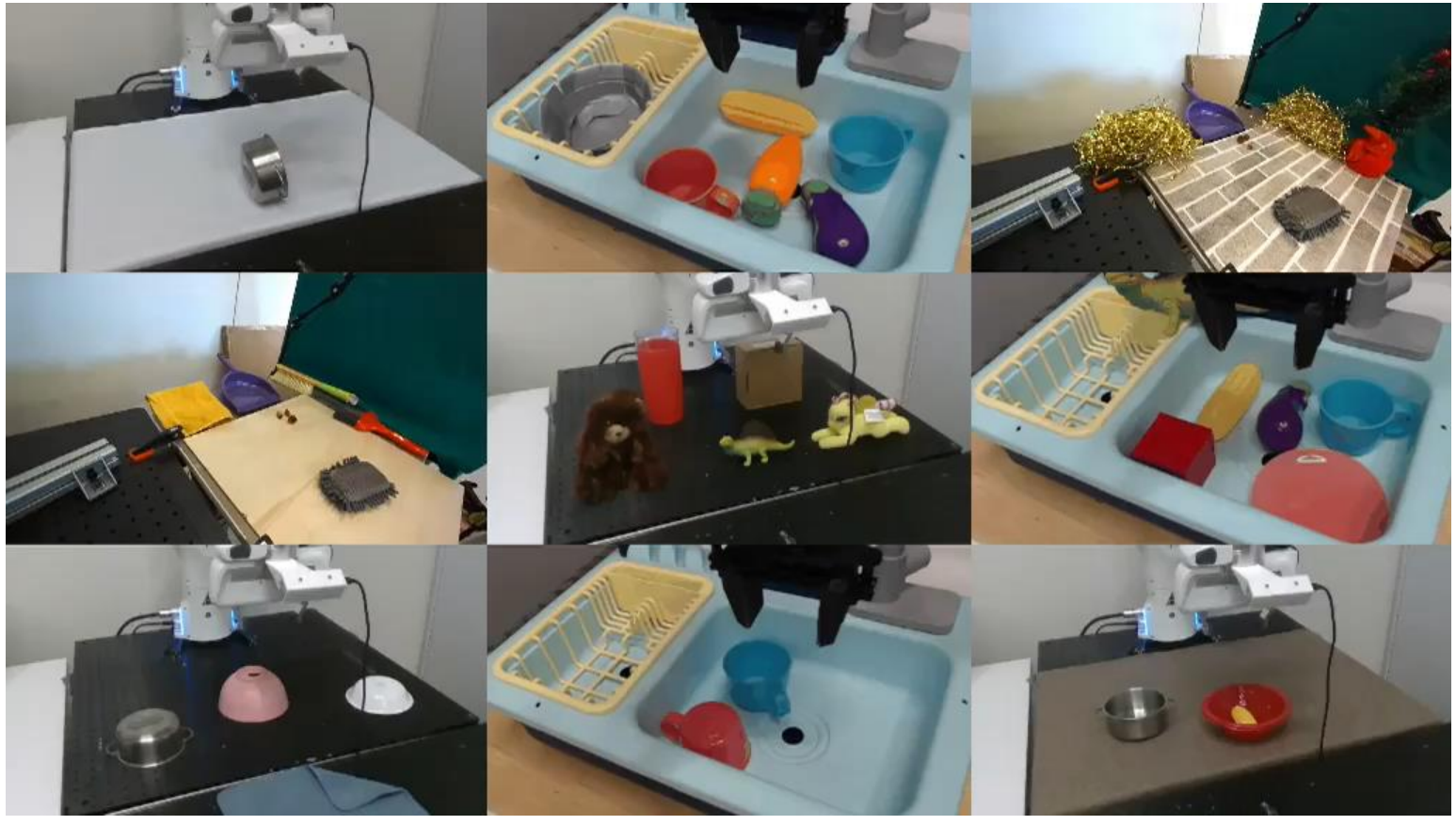
MEM1: Unfortunately, based on the information provided in the documents, I am unable to determine the name of the hotel designed by Veldon Simpson that was the largest in the world upon opening. **Moving on to the second question, ...**

[Zhou et al., Learning to Synergize Memory and Reasoning for Efficient Long-Horizon Agents. ICLR 2026]

multisensory intelligence

Embodied Agents

Generate precise robotics control directly via trained vision language models.

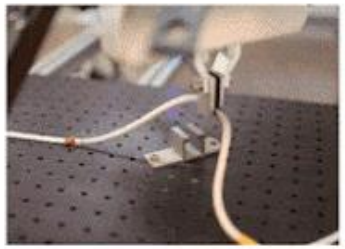


[Kim et al., OpenVLA: An Open-Source Vision-Language-Action Model. 2024]

multisensory intelligence

Real Example: Google RT-X

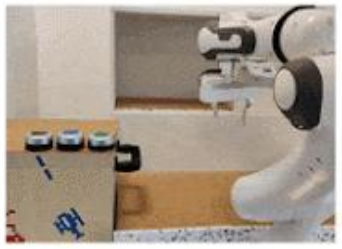
CLVR, USC



RAIL, UC Berkeley

CILVR, NYU

AUTOLab, UC Berkeley

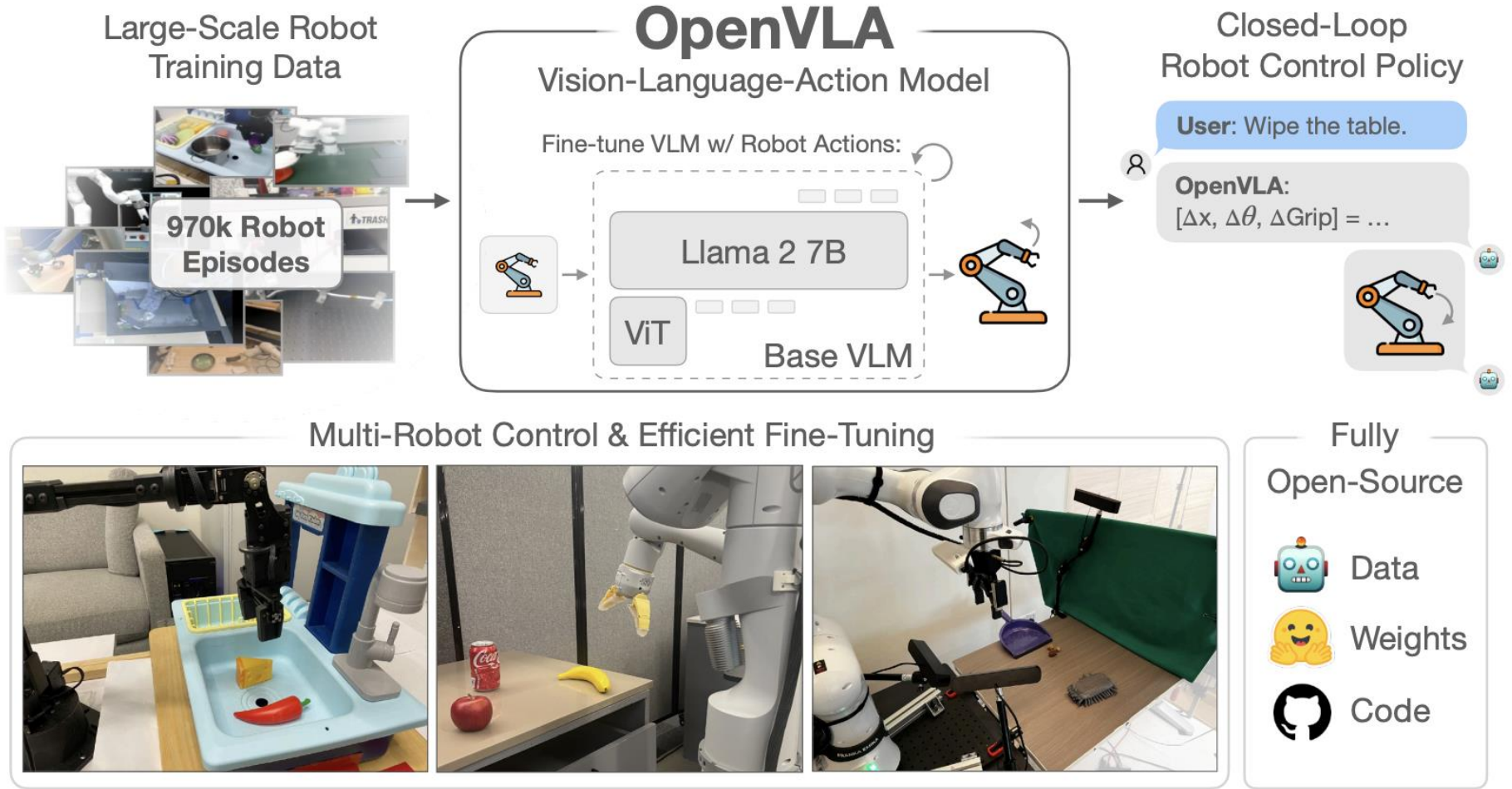


AiS, University of Freiburg

https://deepmind.google/discover/blog/scaling-up-learning-across-many-different-robot-types/

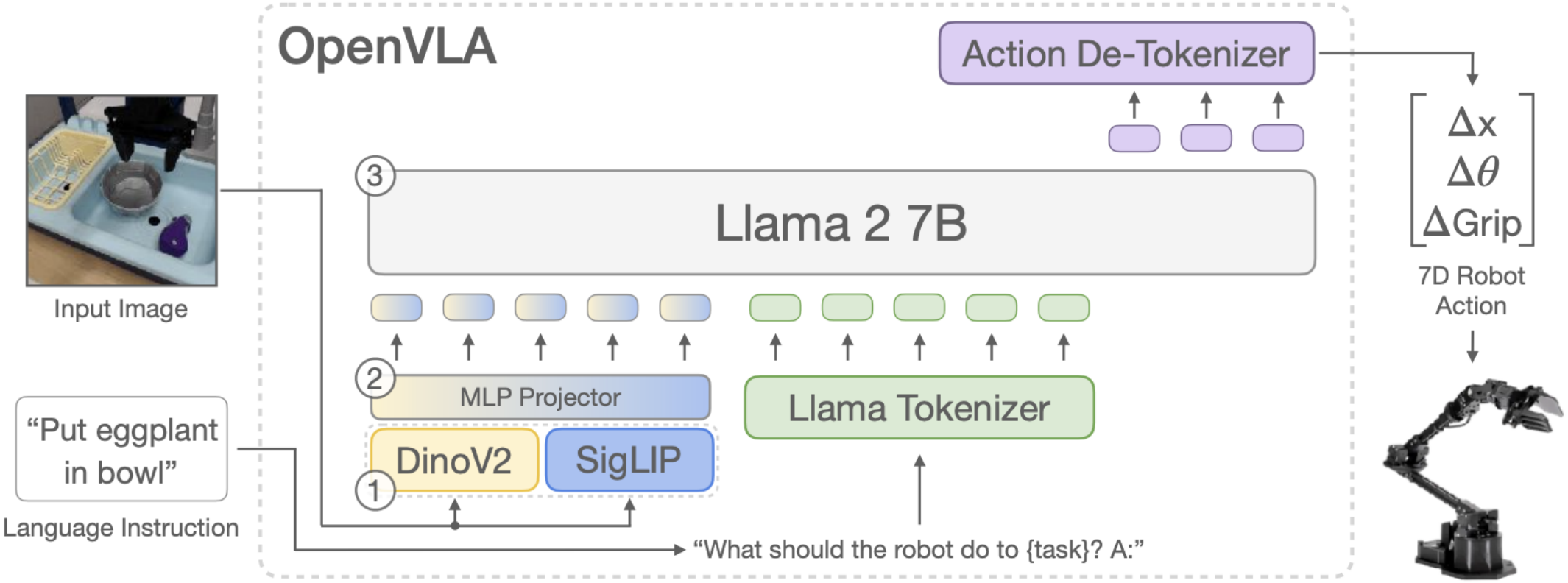
multisensory intelligence

Vision Language Action Models



OpenVLA: https://openvla.github.io/

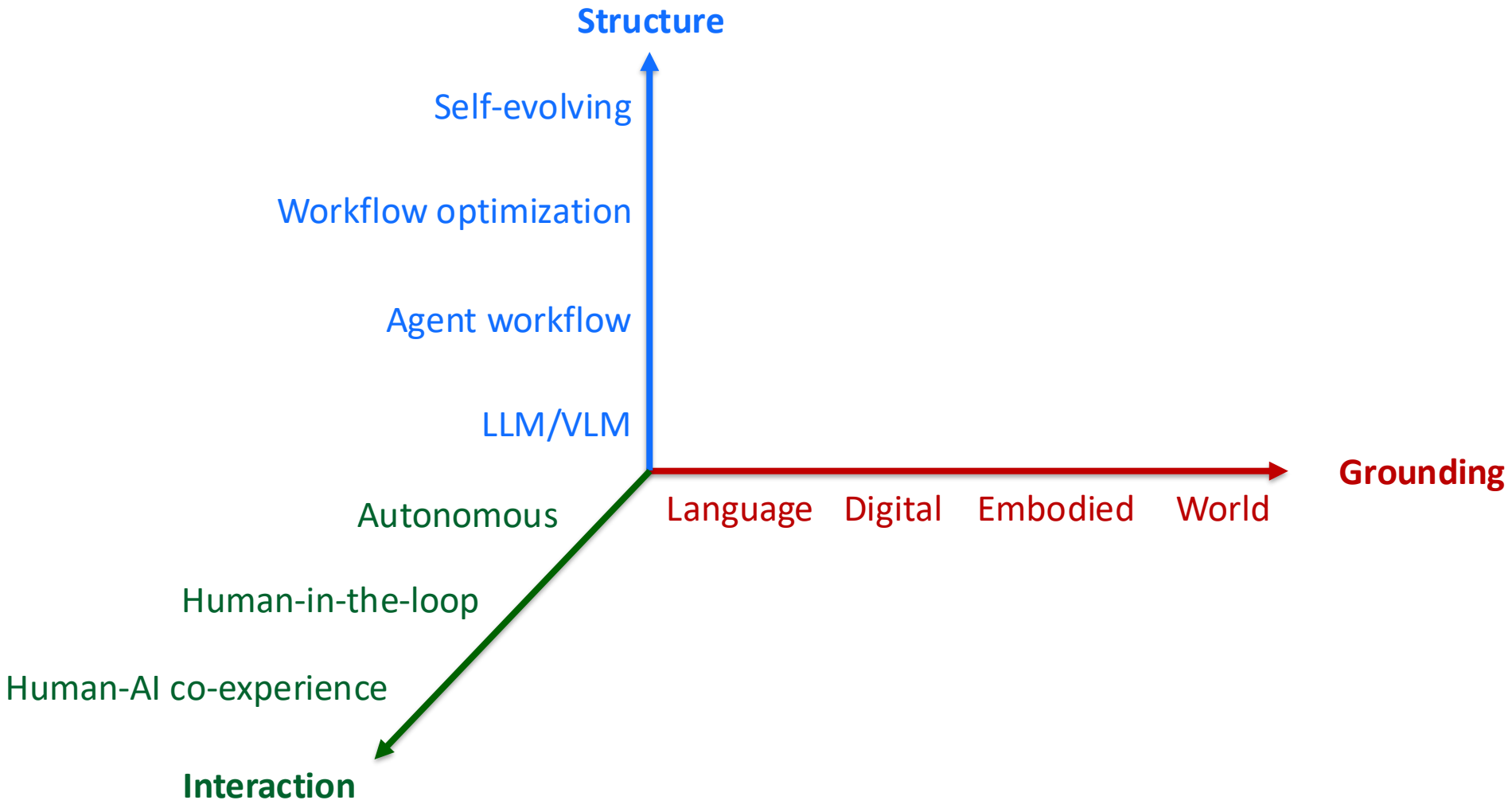
Vision Language Action Models



OpenVLA: https://openvla.github.io/

multisensory intelligence

Summary: My Overview of AI Agents



Assignments for This Coming Week

Project midterm due.

HW4 out, on RL and reasoning.