

A holistic and critical look at language agents

Yu Su

The Ohio State University

Outline

- Language Agents: Introduction
- The Power of Language: Memory, Planning, Grounding, Tool Use
- Future Directions

Outline

- Language Agents: Introduction
- The Power of Language: Memory, Planning, Tool Use
- Future Directions

The rise, and the divide

Bill Gates

Agents are bringing about the **biggest revolution in computing** since we went from typing commands to tapping on icons.

Andrew Ng

I think AI agentic workflows will drive **massive AI progress** this year.

Sam Altman

2025 is when **agents will work**.



Current agents are just thin wrappers around LLMs.

Autoregressive LLMs can never reason or plan.

Auto-GPT's limitations in ... reveal that it is far from being a practical solution.



Yu Su

@ysu_nlp

...

Quoting [@YiMaTweets](#) "It is industry's job to find how to do better, but academia is to find out *how to do it right.*" While I think there're lots of good industry research doing things right, when it comes to research on agents, I do think academia has unique freedom to explore how to *do it right*. Happy to share some thoughts on that in the CMU Agent workshop next week.



Graham Neubig @gneubig · Apr 25

We're having a big event on agents at CMU on May 2-3 (one week from now), all are welcome! cmu-agent-workshop.github.io

It will feature:

* Invited talks from @alsuhr @ysu_nlp @xinyun_chen_ @MaartenSap and ...

Show more

9:46 PM · Apr 25, 2024 · 9,628 Views

View post engagements

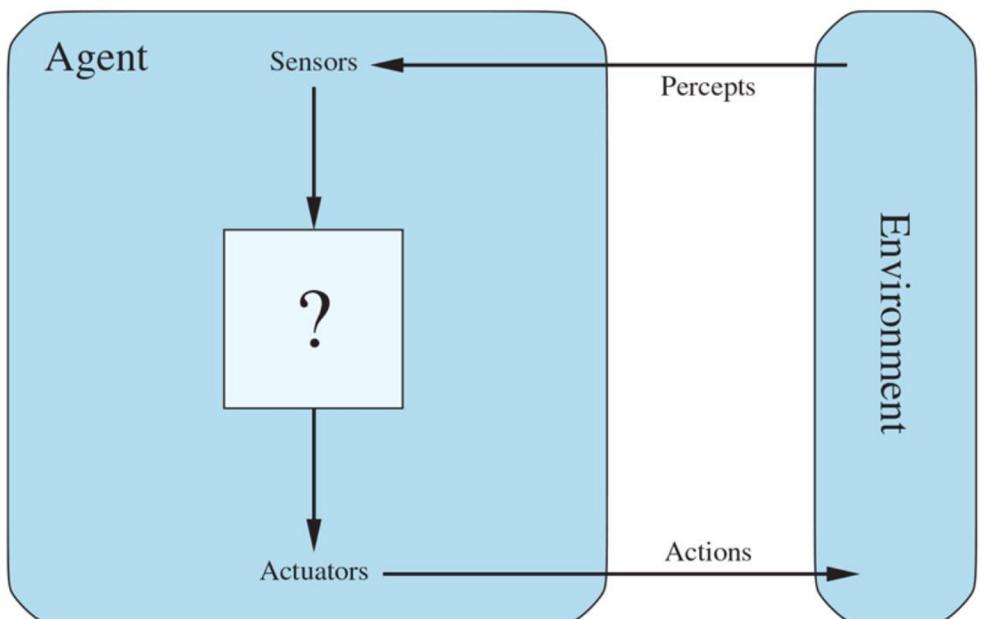
2

5

61

8

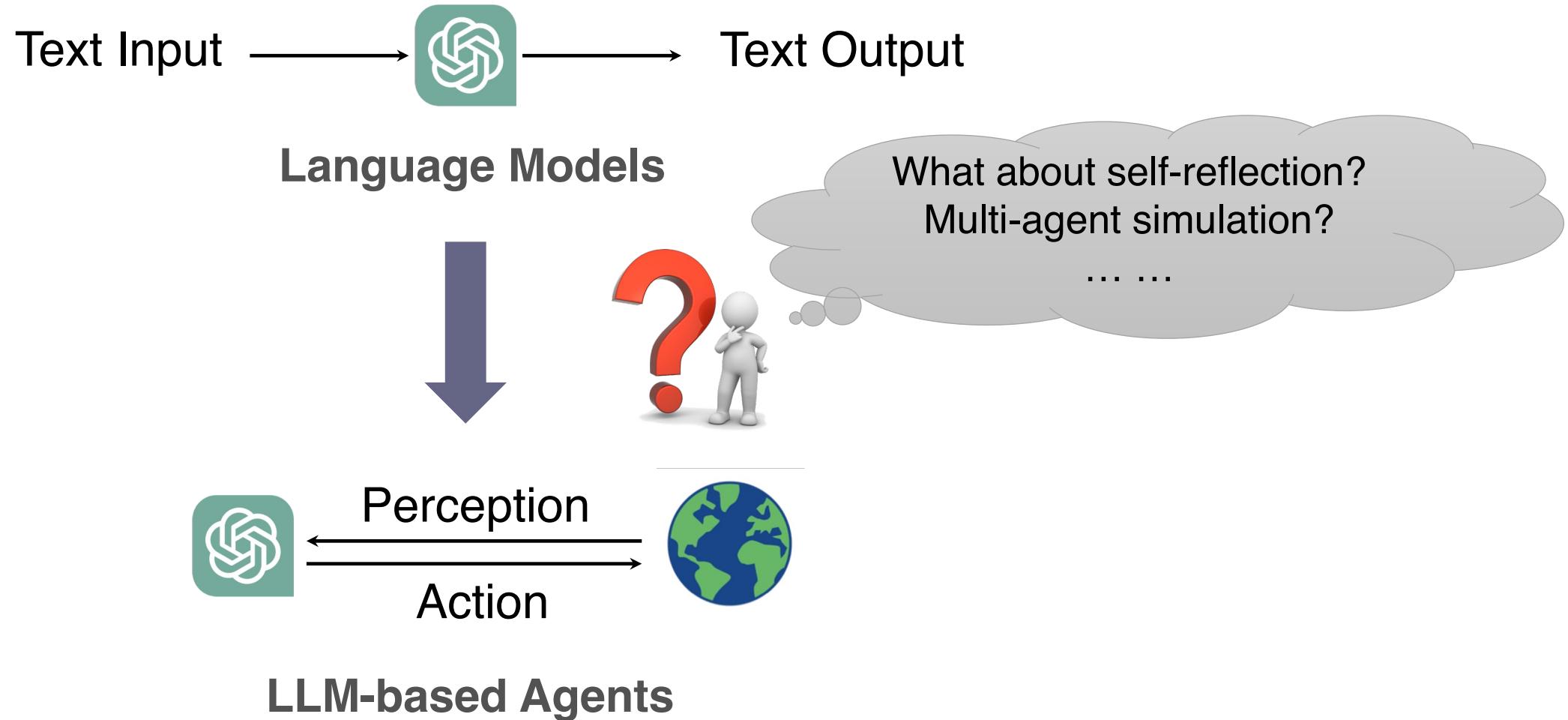
Why agents again?



“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*

'Modern' agent = LLM + external environment?



Two competing views

LLM-first view: We make an LLM into an agent

- Implications: scaffold on top of LLMs, prompting-focused, heavy on engineering

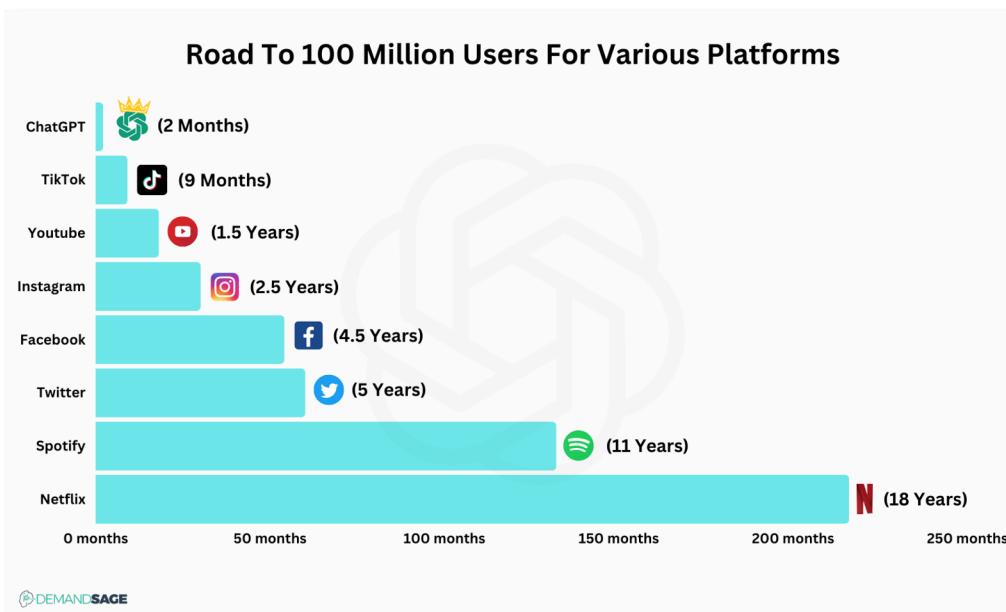
Agent-first view: An AI agent integrates LLMs so it can use language for reasoning and communication

- Implications: All the same challenges faced by previous AI agents (e.g., *perception, reasoning, world models, planning*) still remain, but we need to re-examine them through the new lens of LLMs and tackle the new challenges and opportunities (e.g., *synthetic data, self-reflection, internalized search*)

What's fundamentally different now?

Contemporary AI agents, with integrated LLM(s), can *use language as a vehicle for reasoning and communication*

- ↑ Instruction following, in-context learning, output customization
- ↑ Reasoning (for better acting): state inferences, self-reflection, replanning, etc.

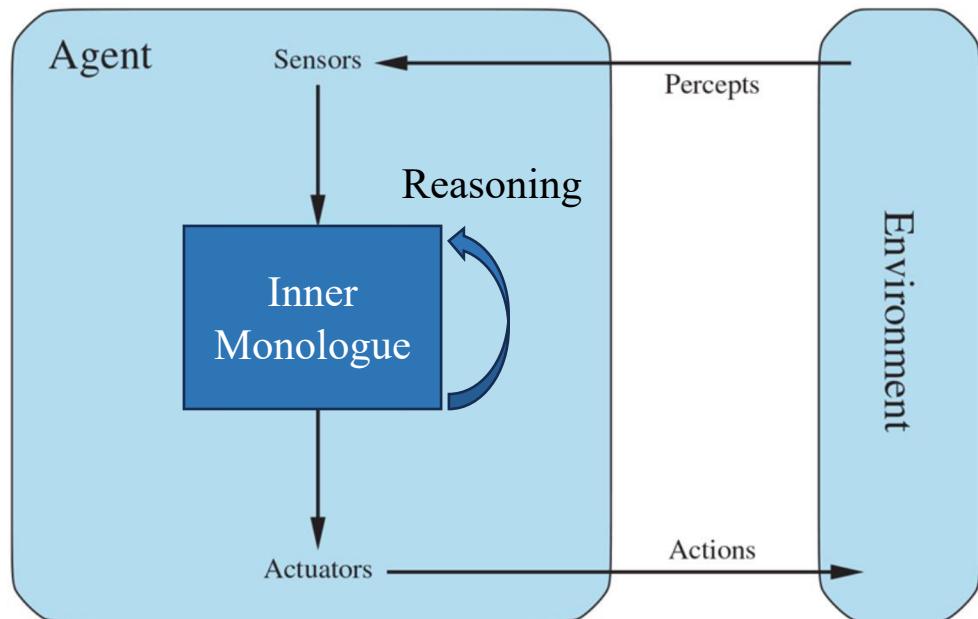


<https://www.demandsage.com/chatgpt-statistics/>



https://www.reddit.com/r/ChatGPT/comments/16jvl4x/wait_actually_yes/

Schematic illustration: updated



- Reasoning by generating tokens is **a new type of action** (vs. actions in external environments)
- **Internal environment**, where reasoning takes place in an inner monologue fashion
- **Self-reflection** is a ‘meta’ reasoning action (i.e., reasoning over the reasoning process), akin to metacognitive functions
- **Reasoning is for better acting**, by inferring environmental states, retrospection, etc.
- **Percept and external action spaces** are substantially expanded, thanks to using language for communication and multimodal perception

Web agent example

American Airlines - Airline ticketing & travel | aa.com/homePage.do

Holiday travel

Arrive at the airport early to allow plenty of time to check in (with or without bags) and clear security before your scheduled departure.

When to arrive at the airport »

English ▾ Search AA.com®

American oneWorld Plan travel Travel information AAdvantage® Log In Join

Earn 50,000 bonus miles with a single pass

Plus, enjoy preferred boarding on American Airlines flights

Have an AAdvantage® account? Log in here.

Book Manage trips / Check-in Flight status

Flight Vacation packages Hotel Car

Round trip One way

From CMH To City or airport

Number of passengers 1

Depart mm/dd/yyyy Return mm/dd/yyyy

Bag and optional fees » Changes to checked bag fees and allowances » ID requirements are changing: Are you REAL ID ready? »

Today's offers

Privacy and cookies

We use cookies on our website. Your interactions and personal data may be collected on our websites by us and our partners in accordance with our privacy policy.

Privacy policy. Dismiss

Task: Search for the flight status of AA 3942 leaving on Dec 29

```
online_demo — node + python demo_text_choice.py — 80x69
geminigby@MacbookAir15 online_demo % ./venv/bin/python demo_text_choice.py
Website: https://www.aa.com/homePage.do
Task: Search for the flight status for the flight AA 3942 leaving on Dec 29
=====
Time step: 0

You are asked to complete the following task: Search for the flight status for the flight AA 3942 leaving on Dec 29
Previous Actions:
None

Action Generation Output (Current Webpage Identification)
The current webpage appears to be the homepage of American Airlines, which offers options for booking flights, vacation packages, hotels, and cars, as well as managing trips/check-ins and checking flight status.

(Previous Action Analysis)
Since there are no previous actions recorded, we are starting fresh with no prior context to consider.

(Screenshot Details Analysis)
The screenshot shows the main booking interface with options for a round trip, one way, and redeeming miles. There is also a navigation menu at the top with options including "Plan travel," "Travel information," "AAdvantage®," "Log in," and "Join." The "Flight status" option is visible, which is relevant to the task at hand. No fields have been filled out yet, and no searches have been initiated.

(Next Action Based on Webpage and Analysis)
To proceed with the task of searching for the flight status of flight AA 3942 leaving on Dec 29, the first next action would be to navigate to the "Flight status" section of the website. This can be done by clicking on the "Flight status" tab located in the top navigation menu.

The target element for the next action is the "Flight status" tab, which is located in the top navigation menu, to the right of the "Manage trips / Check-in" tab.

The corresponding operation is to click on the "Flight status" tab.

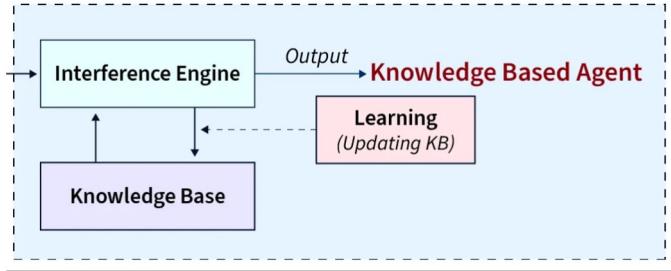
(MultiChoice Question) - Batch 0
If none of these elements match your target element, please select R. None of the other options match the correct element.
A. <button type="button" id="29">Close Global Alert</button>
B. <a id="0">When to arrive at the airport</a>
C. <input type="text" id="45">name=<input type="text" placeholder="Search AA.com®"/>
D. <button type="button" id="33">Submit search</button>
E. <button id="32">English</button>
F. <[type="button"] type="button" id="49">Join</type="button">
G. <[type="button"] type="button" id="48">Log in</type="button">
H. <button type="button" id="34">Close Modal</button>
I. <[tabindex id="51".QSISlider div div { overflow: visible !important;}.QSISlider div span:focus { outline: 2px solid #0078d2;}></[tabindex]>
J. <a role="presentation" id="3">Flight status</a>
K. <a role="presentation" id="2">Manage trips / Check-in</a>
L. <[tabindex role="tab" id="50">Book</tabindex>
M. <input type="radio" id="1">Book</a>
N. <input type="radio" id="35">input value="flight" name="flight" value="flight" />
O. <input type="radio" id="36">input value="roundTripHotel" name="flight" value="roundTripHotel" />
P. <input type="checkbox" id="39">input value="true" name="redeemMiles" value="true" />
Q. <input type="radio" id="38">input value="oneWay" name="tripType" value="oneWay" />
R. None of the other options match the correct element

```

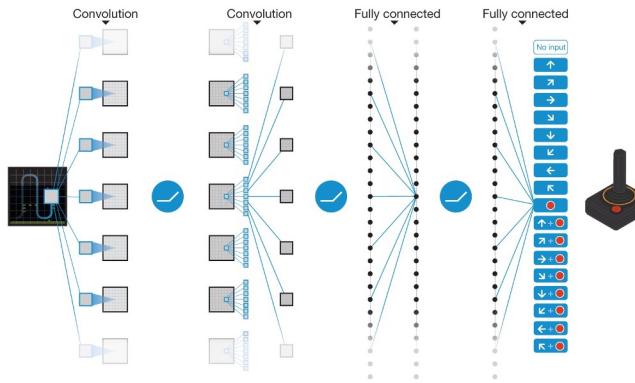
Let's call them language agents

- These contemporary AI agents capable of using language for reasoning and communication are best called “**language agents**,” for language being their most salient trait.
- What about *multimodal agents*?
 - While there’s perception of other modalities, language is still doing the heavy lifting (reasoning and communication)
- What about simply *LLM agents*?
 - The key is using language for reasoning and communication, but that doesn’t have to come from an LLM; that may turn out to be a means to an end
 - Maybe in a few years, we will move beyond LLMs, but the need for universal language understanding and production in agents will remain

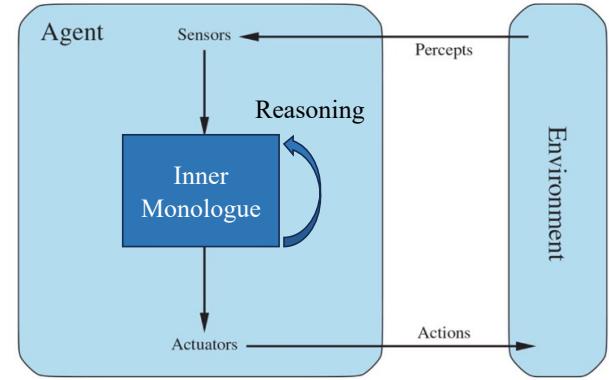
Evolution of AI agents



Logical Agent



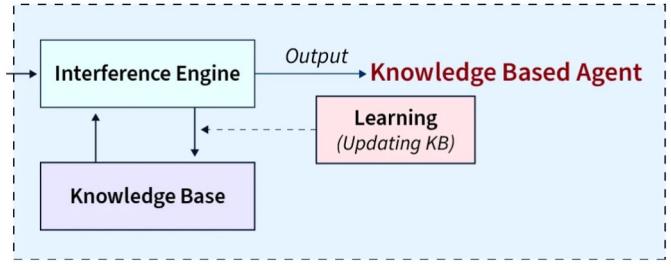
Neural Agent



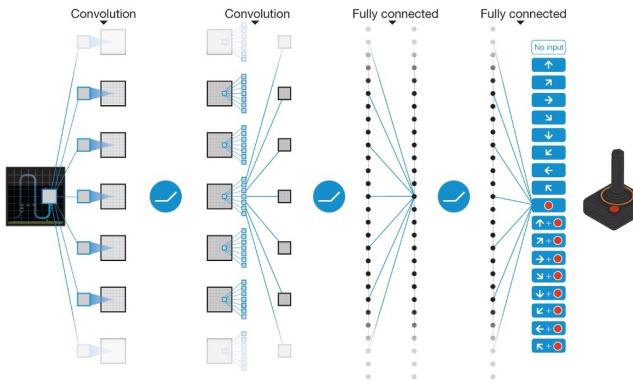
Language Agent

Expressiveness			
Reasoning			
Adaptivity			

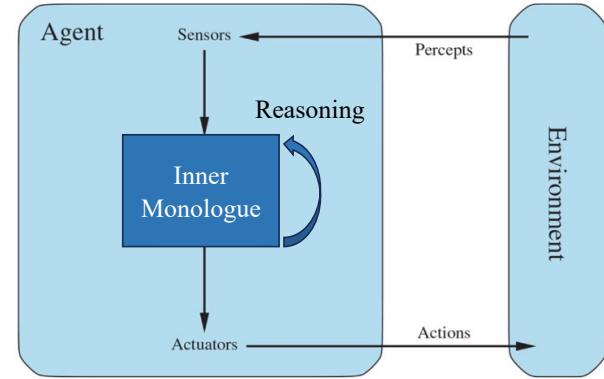
Evolution of AI agents



Logical Agent



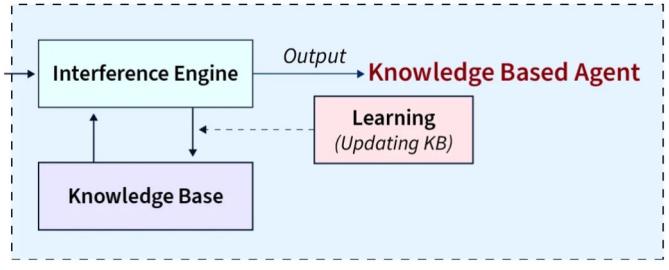
Neural Agent



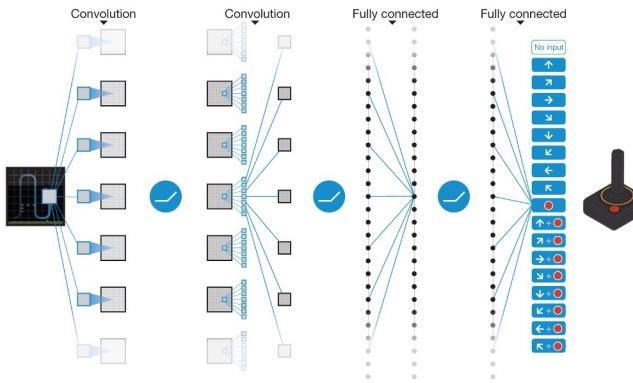
Language Agent

Expressiveness	Low bounded by the logical language		
Reasoning	Logical inferences sound, explicit, rigid		
Adaptivity	Low bounded by knowledge curation		

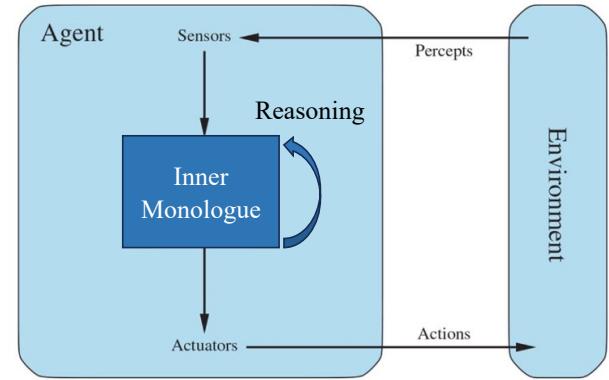
Evolution of AI agents



Logical Agent



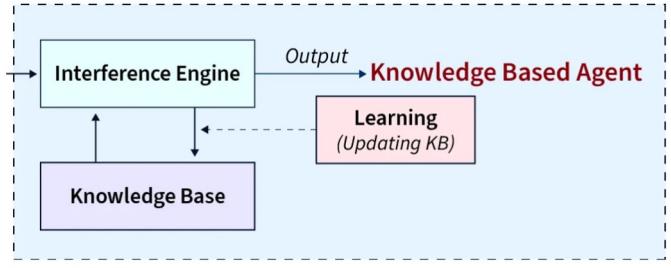
Neural Agent



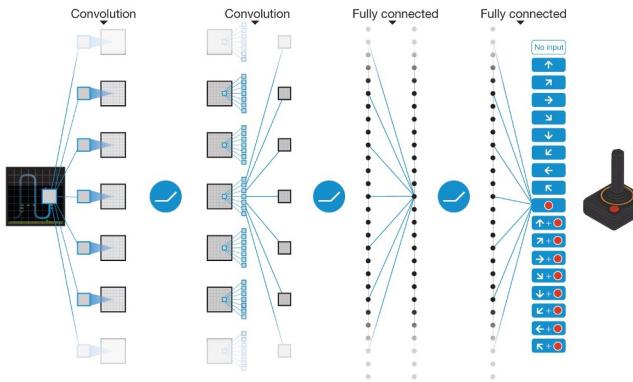
Language Agent

Expressiveness	Low bounded by the logical language	Medium anything a (small-ish) NN can encode	
Reasoning	Logical inferences sound, explicit, rigid	Parametric inferences stochastic, implicit, rigid	
Adaptivity	Low bounded by knowledge curation	Medium data-driven but sample inefficient	

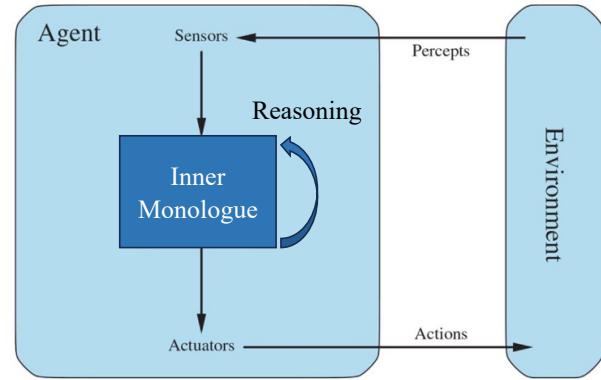
Evolution of AI agents



Logical Agent



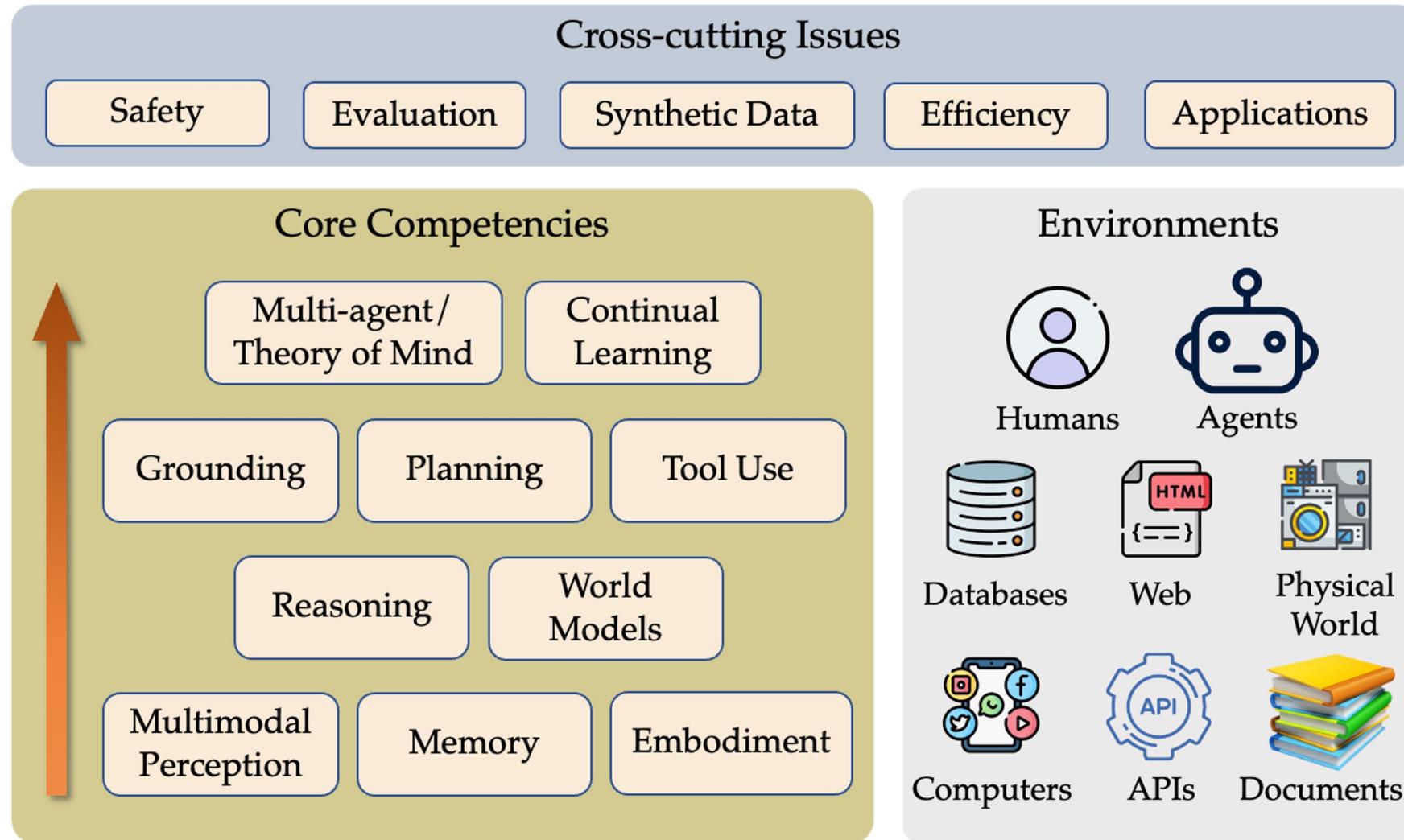
Neural Agent



Language Agent

Expressiveness	Low bounded by the logical language	Medium anything a (small-ish) NN can encode	High almost anything, esp. those verbalizable
Reasoning	Logical inferences sound, explicit, rigid	Parametric inferences stochastic, implicit, rigid	Language-based inferences fuzzy, semi-explicit, flexible
Adaptivity	Low bounded by knowledge curation	Medium data-driven but sample inefficient	High strong prior from LLMs + language use

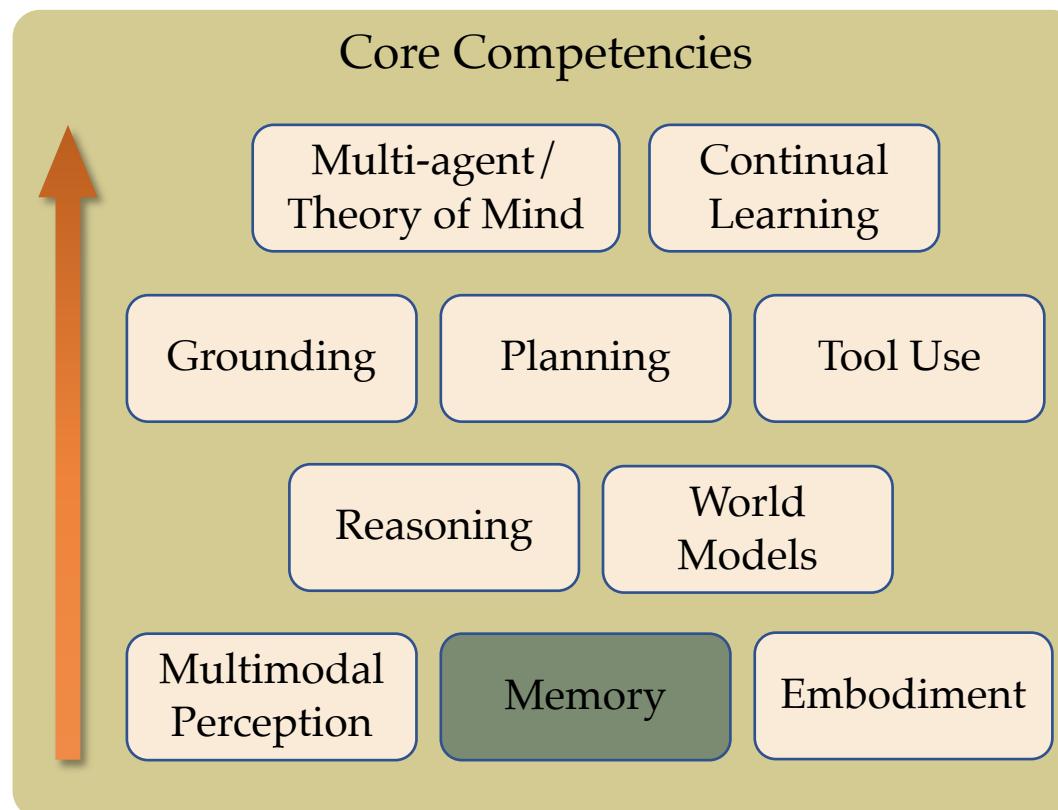
A conceptual framework for language agents



Outline

- Language Agents: Introduction
- The Power of Language: Memory, Planning, Grounding, Tool Use
- Future Directions

Memory



HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models

**Bernal Jiménez Gutiérrez, Yiheng Shu,
Yu Gu, Michihiro Yasunaga, Yu Su**



Long-term memory in LLMs

- LLMs still rely on RAG.

ChatGPT



You

Who won the 2024 Super Bowl?



ChatGPT

The Super Bowl in 2024 was held on February 11, 2024. To find the most current information about the winner, I can browse the internet for you. Let me do that now.

The Kansas City Chiefs won the 2024 Super Bowl against the San Francisco 49ers with a final score of 25-22. The game went into overtime, and the Chiefs secured the win with a touchdown just seconds before the end of the game ([Yahoo Sports](#)) ([KCUR](#)) ([PinkNews](#)) ([Just Jared](#)).

Current RAG does not always work

Passages



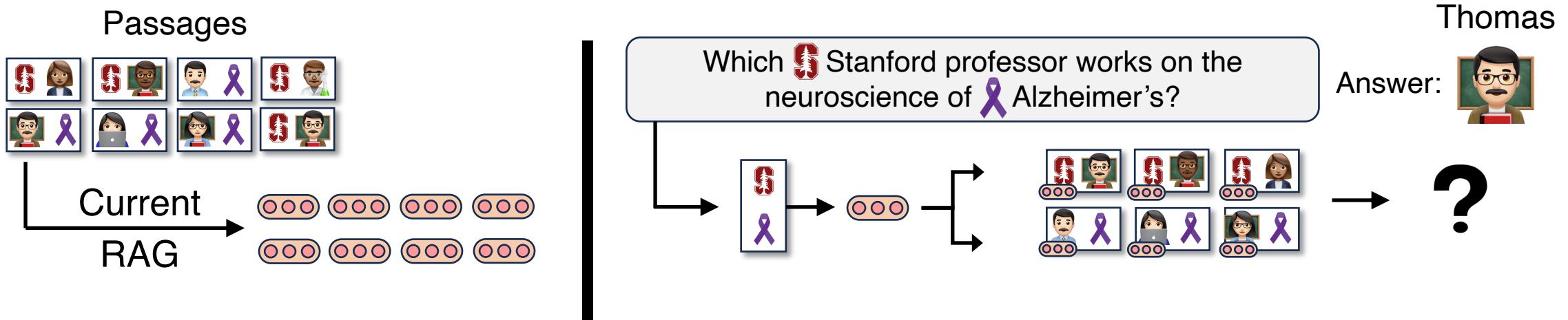
Which Stanford professor works on the neuroscience of Alzheimer's?

Thomas

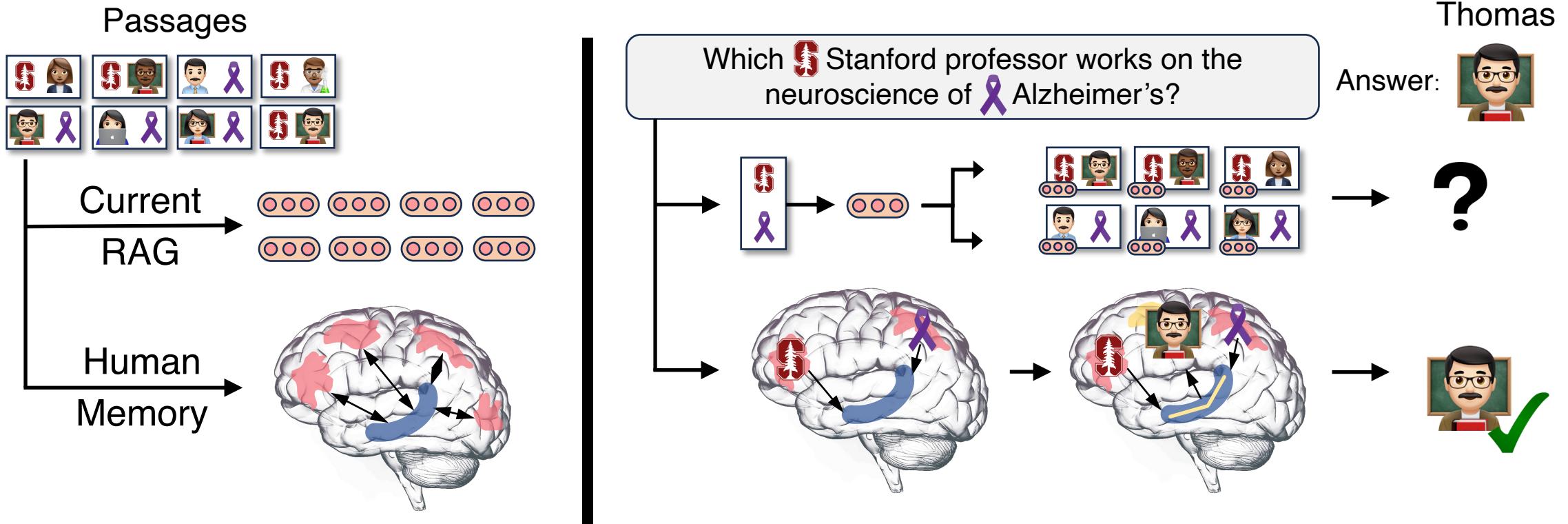
Answer:



Current RAG does not always work

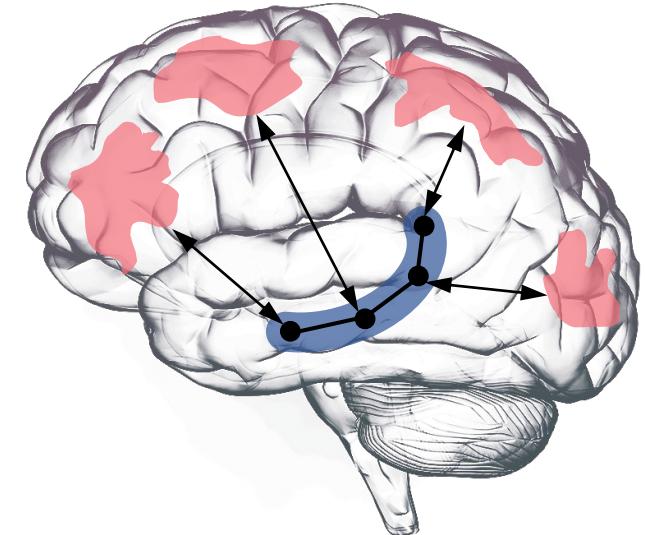


Current RAG does not always work



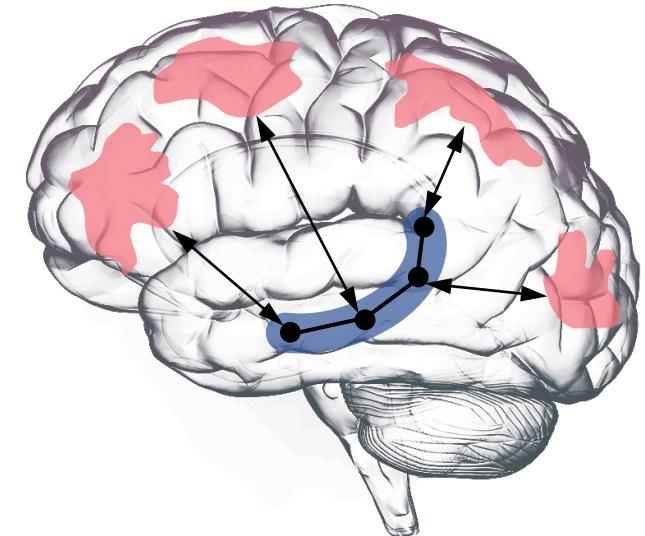
Long-term memory in humans

- One well-established theory of human long-term memory is the **hippocampal indexing theory** (Teyler et al. 1986).
 - The **hippocampus (blue)** is a store for **indices** (which point to memories stored in **neocortex**) and **associations** between them.
 - Memories can be anything (i.e., people, events, places, etc.)

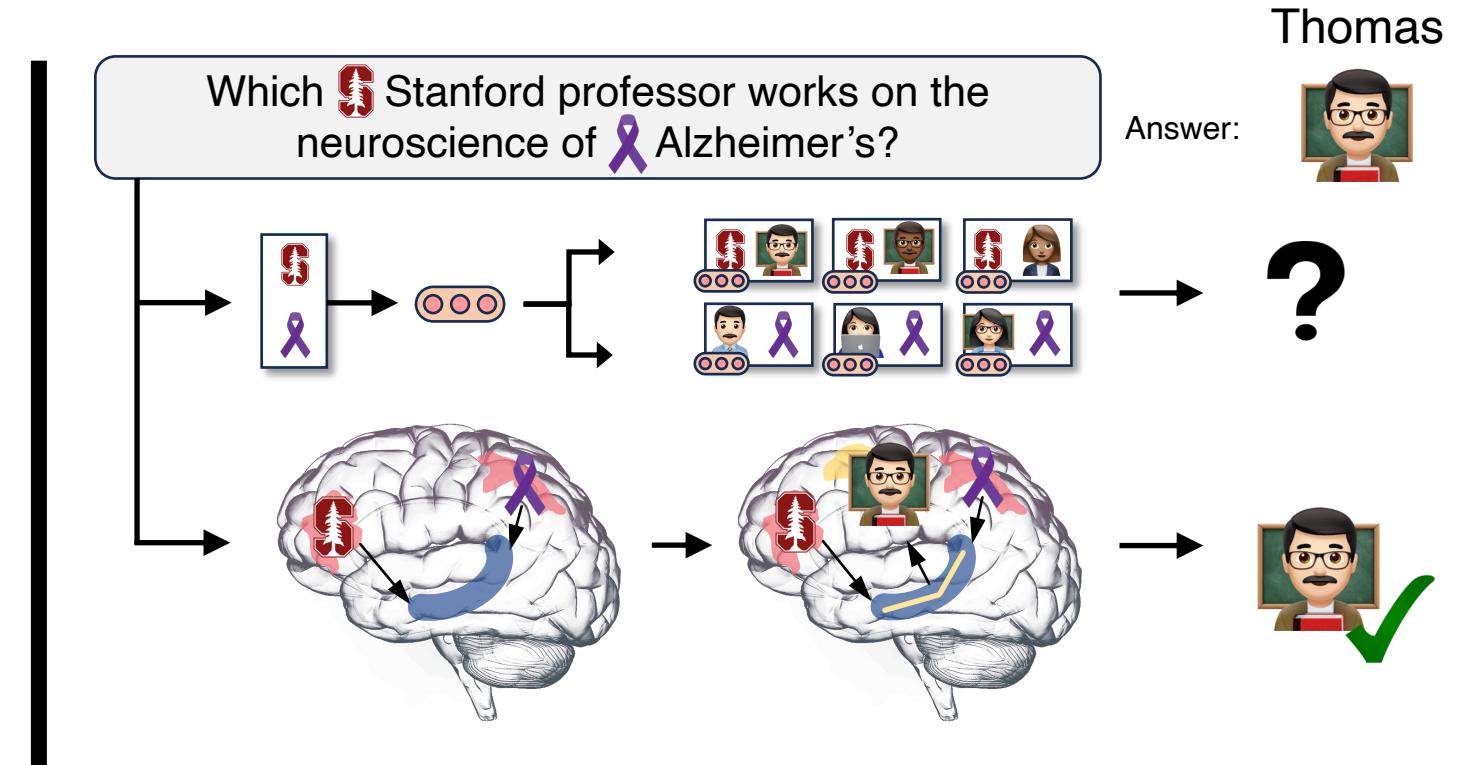
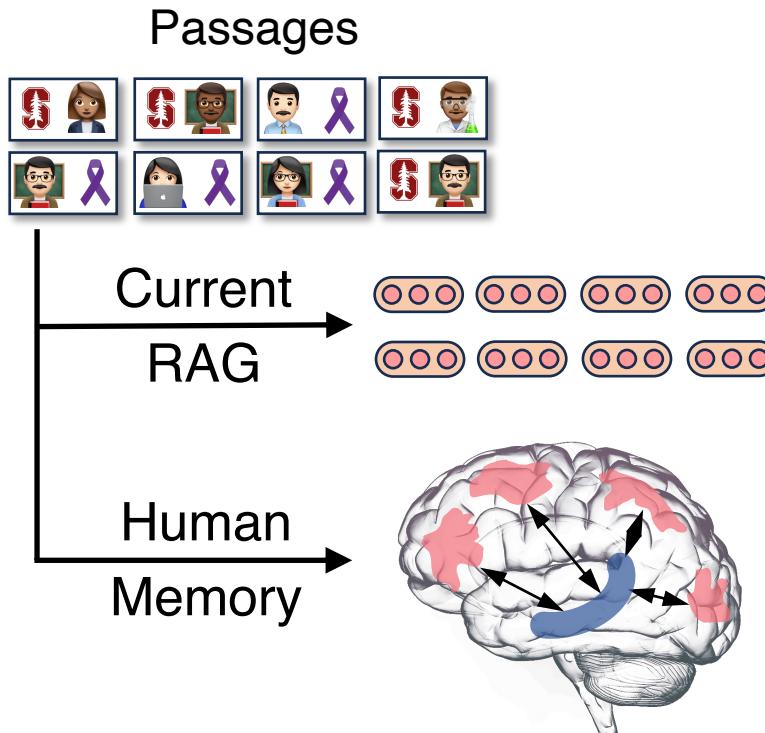


Long-term memory in humans

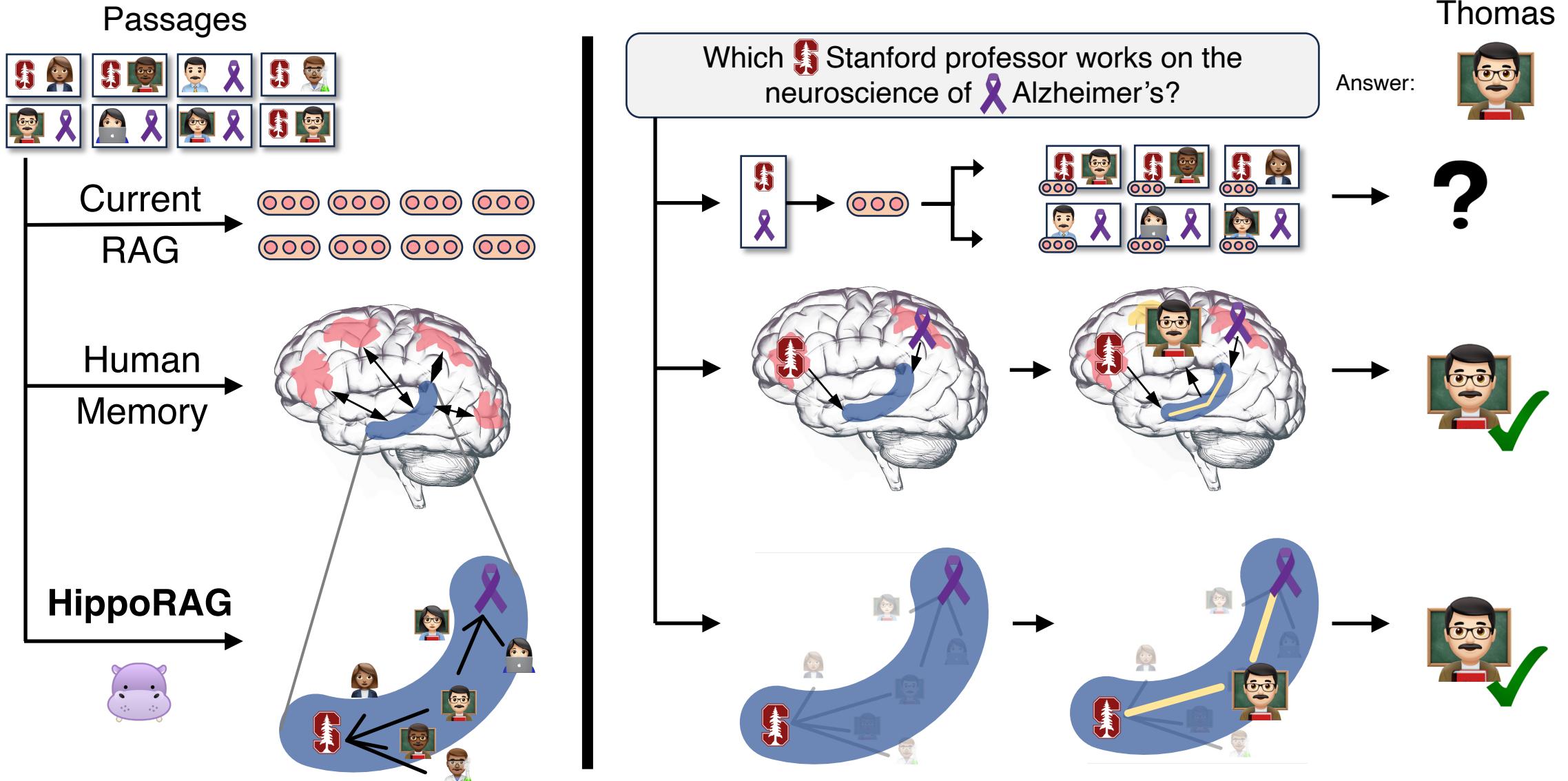
- Indexing procedure enables two fundamental faculties of human memory:
 - **Pattern separation:** process for differentiating memories (**neocortex** and parahippocampus)
 - **Pattern completion:** process for recovering complete memories from relevant associations (mostly **hippocampus**, specifically CA3)



HippoRAG: Neurobiologically-inspired LTM for LLMs



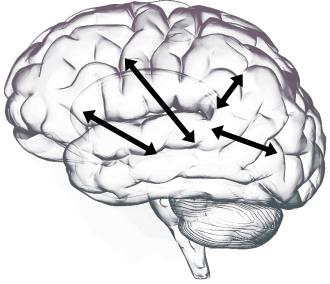
HippoRAG: Neurobiologically-inspired LTM for LLMs



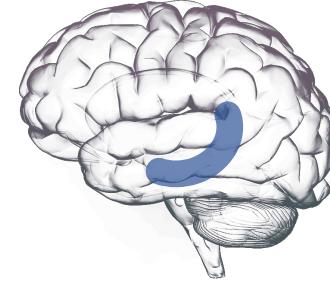
HippoRAG & the three components of LTM



Neocortex
Perception, linguistic
abilities & reasoning

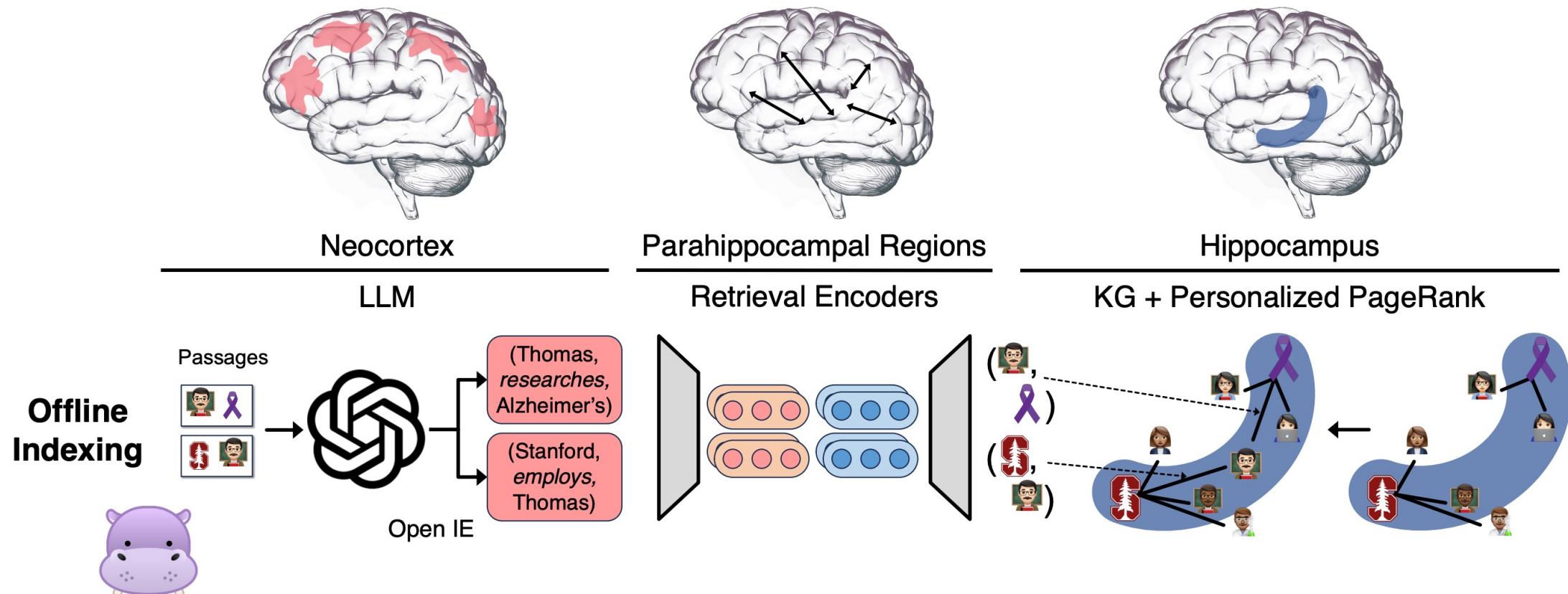


Parahippocampus
Bridge between areas;
working memory

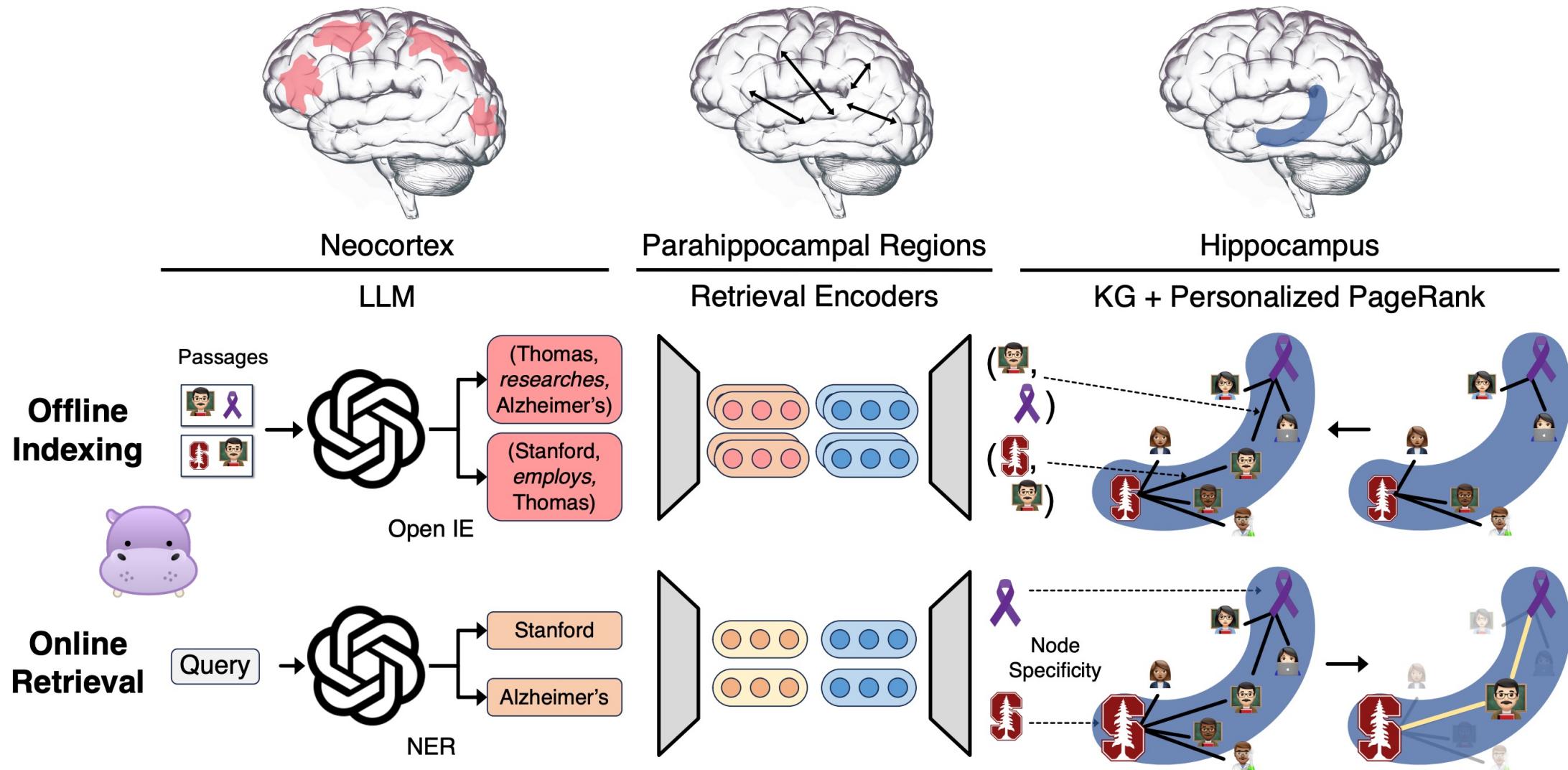


Hippocampus
Indexing & auto-
associative memory

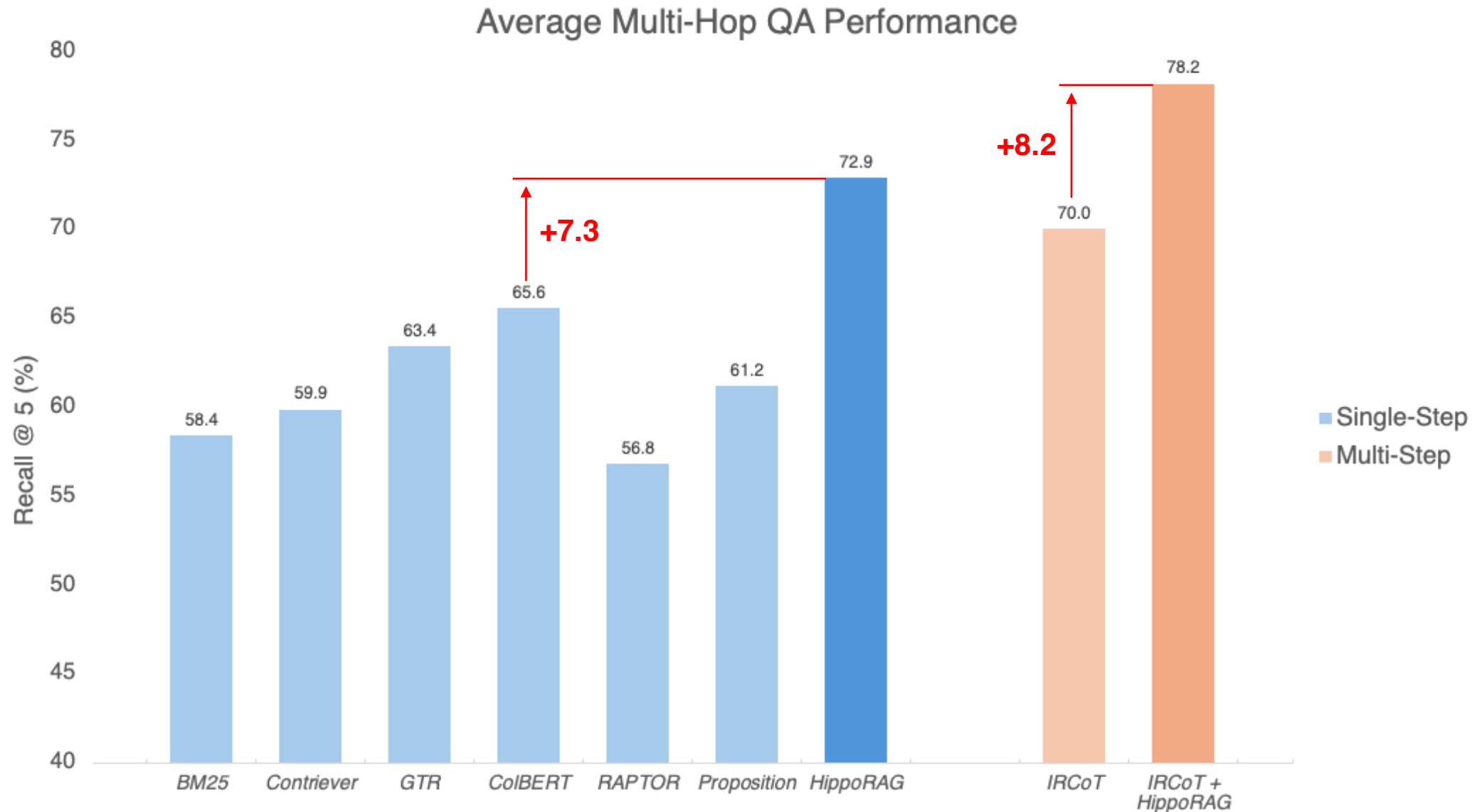
HippoRAG & the three components of LTM



HippoRAG & the three components of LTM



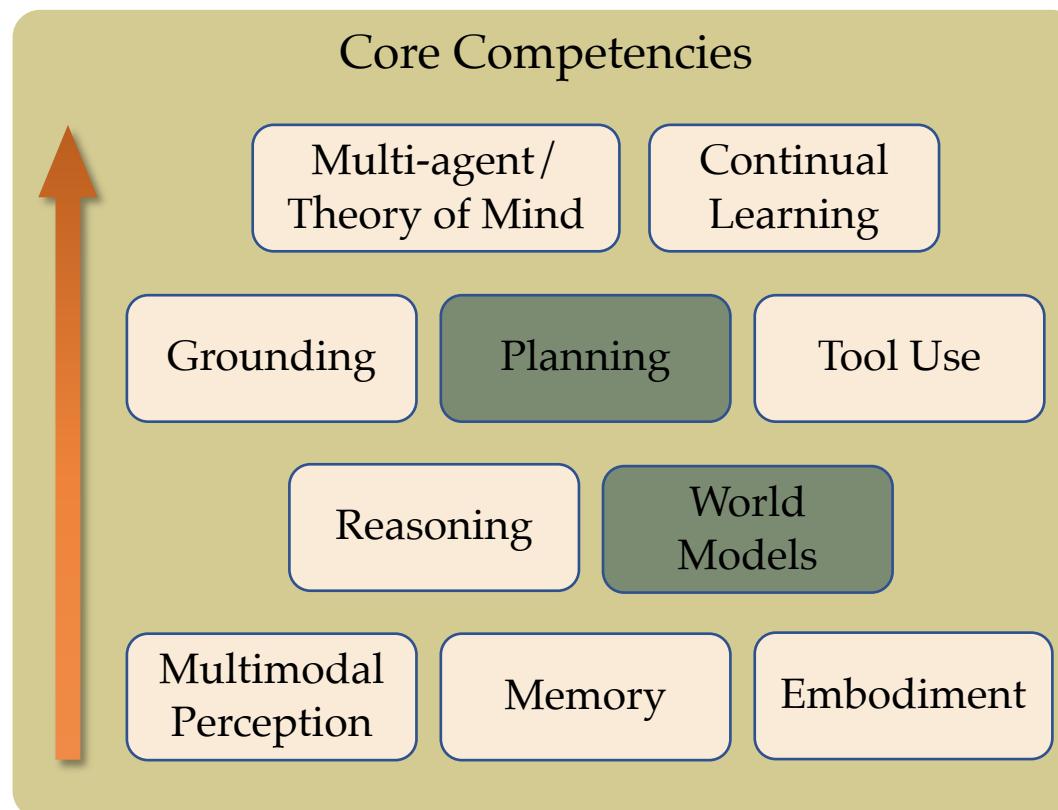
HippoRAG is a SoTA memory retriever



Practical aspects of HippoRAG

- Better retrieval directly translates to better QA (avg. 5% better)
- With *single-step multi-hop retrieval*, it's 10-30 times cheaper and 6-13 times faster than iterative RAG like IRCoT with similar or better recall
- Worried about indexing cost? Llama-3-8B achieves similar performance as GPT-3.5

Planning



Is Your LLM Secretly a World Model of the Internet?

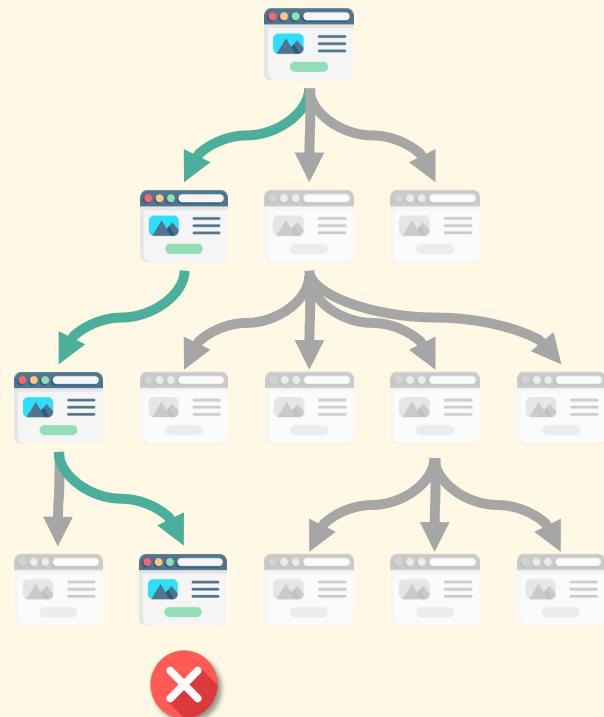
Model-based Planning for Web Agents

Yu Gu*, Boyuan Zheng*, Boyu Gou, Kai Zhang,
Cheng Chang, Sanjari Srivastava, Yanan Xie, Peng Qi,
Huan Sun, Yu Su



Planning paradigms for language agents

(a) reactive

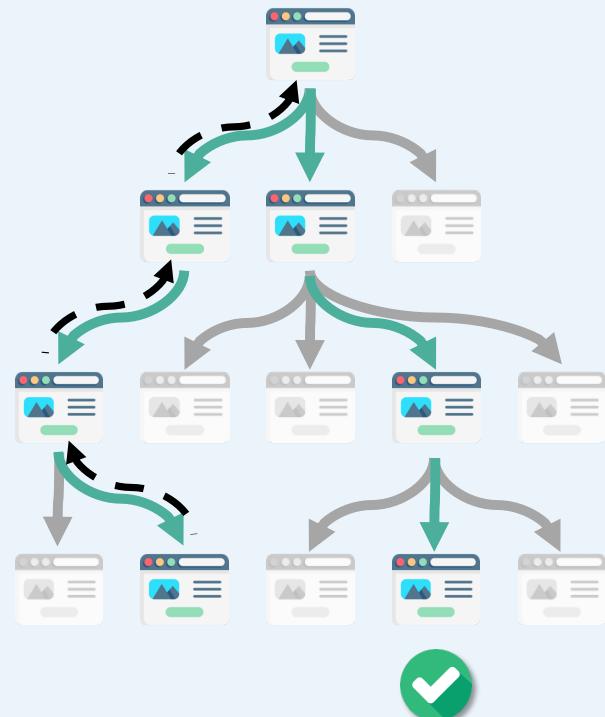


fast, easy to implement



greedy, short-sighted

(b) tree search with real interactions

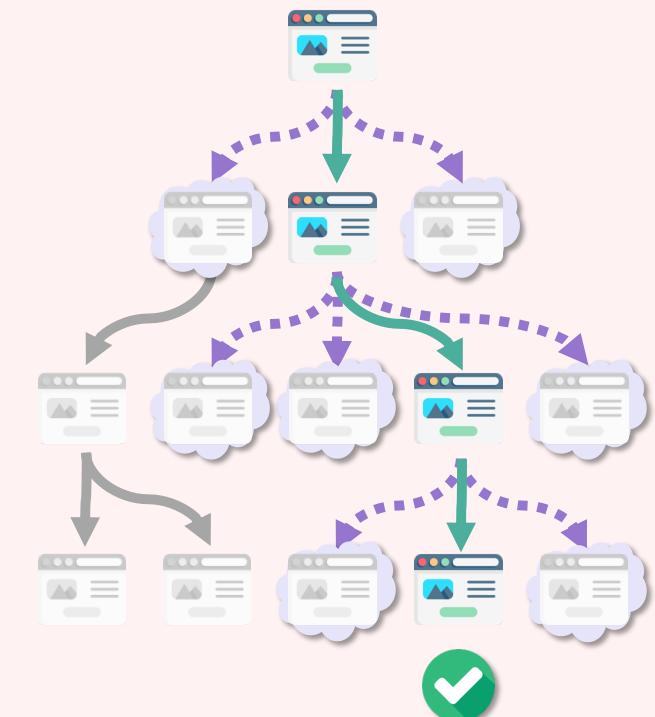


systematic exploration



irreversible actions,
unsafe, slow

(c) model-based planning



faster, safer,
systematic exploration



how to get a world model?

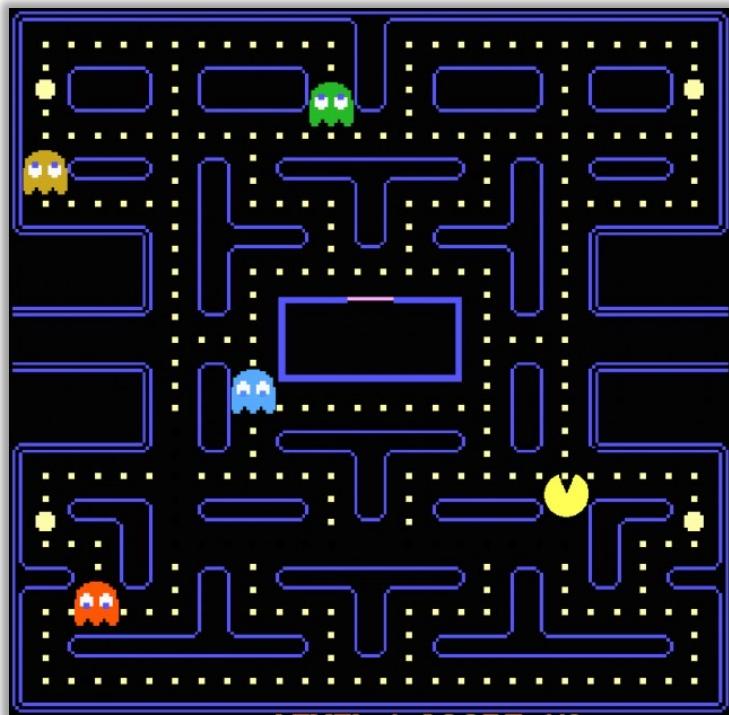
What's ... a world model?

A computational model of environment transition dynamics

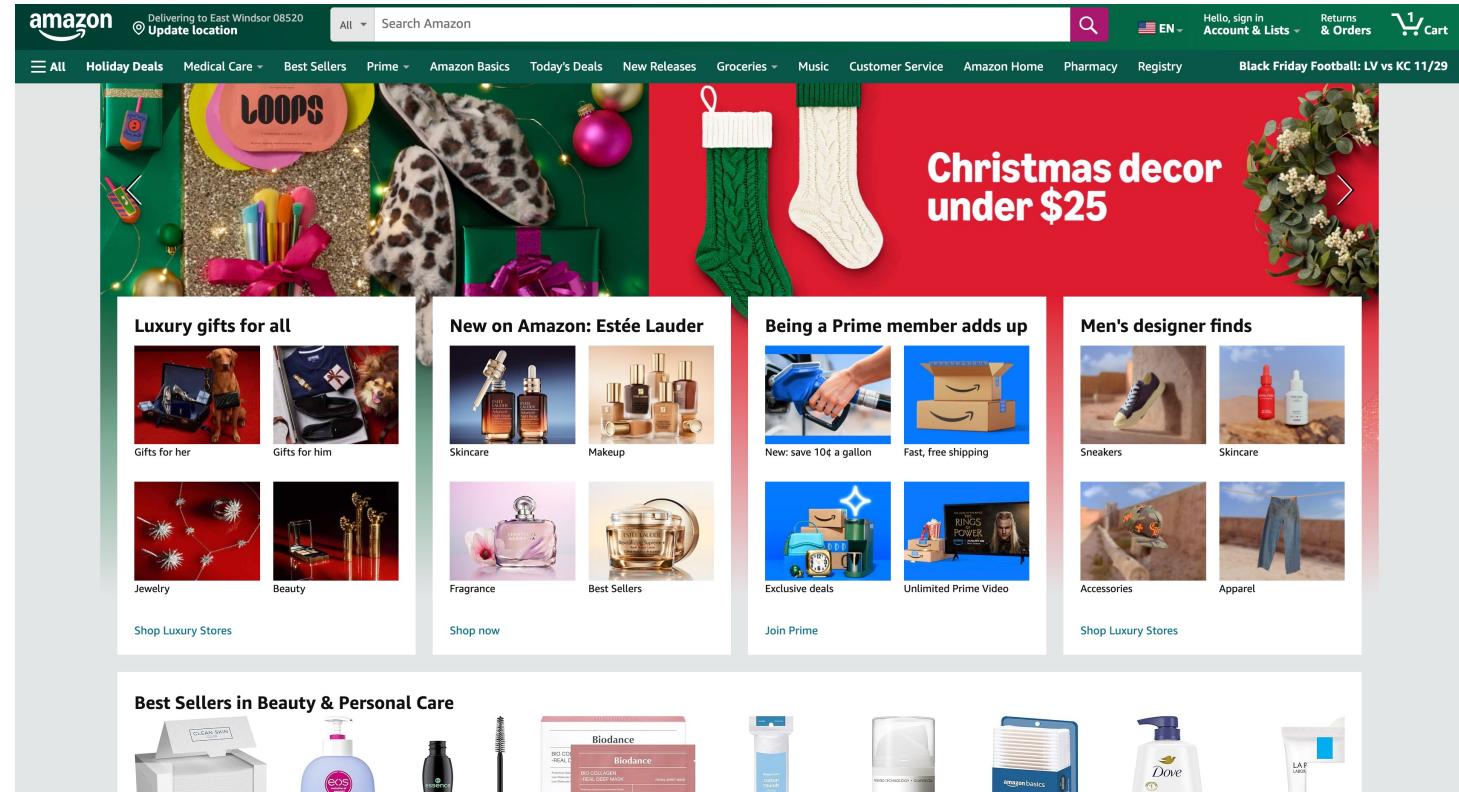
$$\hat{T}: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$$

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

Why hasn't it been done already?



V.S



The image shows the homepage of the Amazon website. At the top, there's a search bar with the placeholder "Search Amazon". Below the search bar is a navigation menu with links to "All", "Holiday Deals", "Medical Care", "Best Sellers", "Prime", "Amazon Basics", "Today's Deals", "New Releases", "Groceries", "Music", "Customer Service", "Amazon Home", "Pharmacy", "Registry", and "Black Friday Football: LV vs KC 11/29". On the right side of the header, there are links for "Hello, sign in", "Account & Lists", ">Returns & Orders", and a shopping cart icon with the number "1". The main content area features a large banner for "Christmas decor under \$25" with images of various festive items. Below the banner are several promotional boxes: "Luxury gifts for all" (with sub-sections for "Gifts for her" and "Gifts for him"), "New on Amazon: Estée Lauder" (showing skincare and makeup products), "Being a Prime member adds up" (with sub-sections for "Exclusive deals" and "Unlimited Prime Video"), and "Men's designer finds" (with sub-sections for "Sneakers", "Skincare", "Accessories", and "Apparel"). At the bottom, there's a section for "Best Sellers in Beauty & Personal Care" featuring various skincare and personal care products.

And **billions of other websites** on the Internet!

LLMs can predict state transitions

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

One Stop Market

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

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

Home > Clothing, Shoes & Jewelry > Men > Uniforms, Work & Safety

Uniforms, Work & Safety

Shop By

Shopping Options

Price

- \$0.00 - \$99.99(310)
- \$100.00 - \$199.99(96)
- \$200.00 - \$299.99(9)
- \$500.00 and above(1)

Compare Products
You have no items to compare.

My Wish List
You have no items in your wish list.

Items 1-12 of 416

Sort By Position ↑

Image	Name	Price	Add to Cart
	Workwear Professionals Men Scrubs Pant Tapered Leg Drawstring Cargo WW1905, XS Short, Hunter Green	\$23.98	Add to Cart
	Men's Slim Fit Pinstripe Chef Pant (S-3X)	\$31.98	Add to Cart
	XINFU Chef's Japanese Kimono Unisex Uniform Short Sleeved Working Clothes Kitchen Restaurant Chef Jacket	\$34.69	Add to Cart
	Safety Shoes, Mesh Cloth Protect Your Toes Men's Safety Shoes, Lightweight Work Indoor Man Outdoor for Woman(40-46, 40)	\$39.37	Add to Cart
	TRUEWERK Men's Winter Work Pants - T3 WerkPants Insulated Workwear	★★★★★ 12 Reviews \$99.00	Add to Cart
	Steel Toe-Cap Flying Woven Mesh Shoes, Lightweight Anti-Smashing Anti-Puncture Safety Shoes, Anti-Collision Anti-Pressure Work Shoes, Black, 46	\$11.99	Add to Cart
	Mens Flowers Casual Aloha Hawaiian Shirt Summer Short Sleeve Beach T-Shirt Regular Fit Button Down Dress Shirts	\$179.95	Add to Cart
	Carhartt Men's Duck Chore Coat Blanket Lined C001	\$85.99	Add to Cart



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 planner for web agents

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

③ Type 'Disk'

① Click 'Office Products'

② Click 'Electronics'

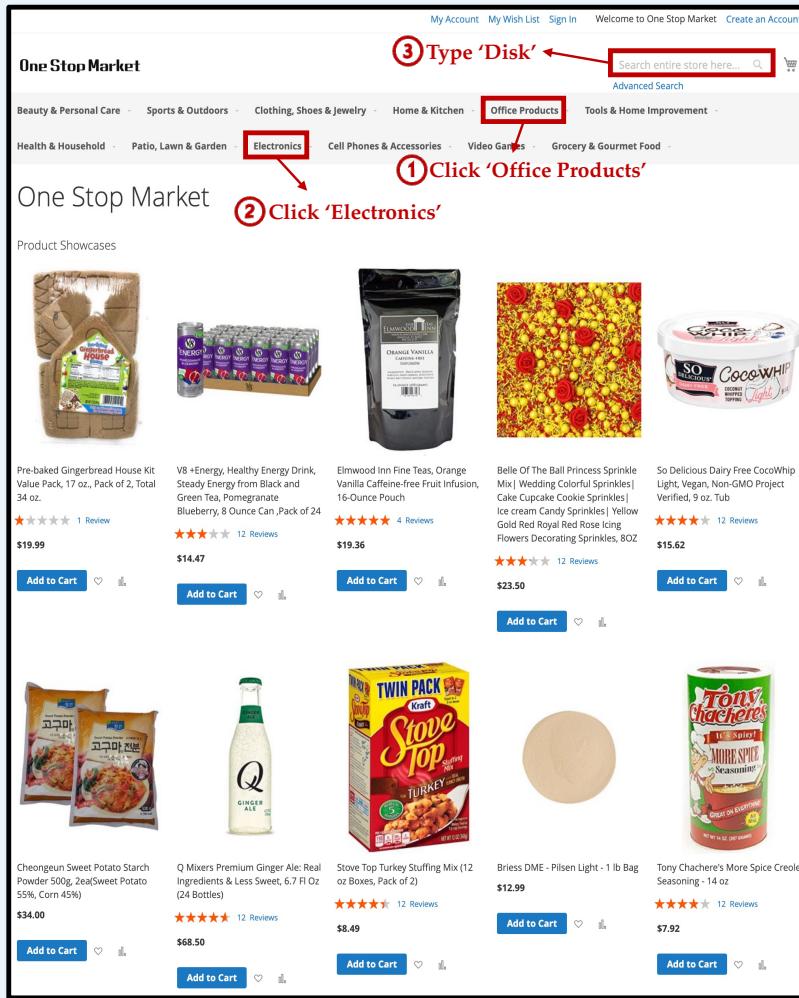
Product Showcases

- Pre-baked Gingerbread House Kit Value Pack, 17 oz., Pack of 2, Total 34 oz.
★ ★ ★ ★ 1 Review
\$19.99
[Add to Cart](#)
- VR +Energy, Healthy Energy Drink, Steady Energy From Black and Green Tea, Pomegranate Blueberry, 8 Ounce Can, Pack of 24
★ ★ ★ ★ 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 Colorful Sprinkles| Cake Cupcake Cookie Sprinkles | Ice cream Candy Sprinkles | Yellow Gold Royal Red Rose King 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
★ ★ ★ ★ 12 Reviews
\$15.62
[Add to Cart](#)

- Cheongeon Sweet Potato Starch Powder 500g, ZeSweet Potato 55%, 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](#)
- Kraft Stove Top Turkey Stuffing Mix - Twin Pack
★ ★ ★ ★ 12 Reviews
\$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](#)

WebDreamer: model-based planner for web agents

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



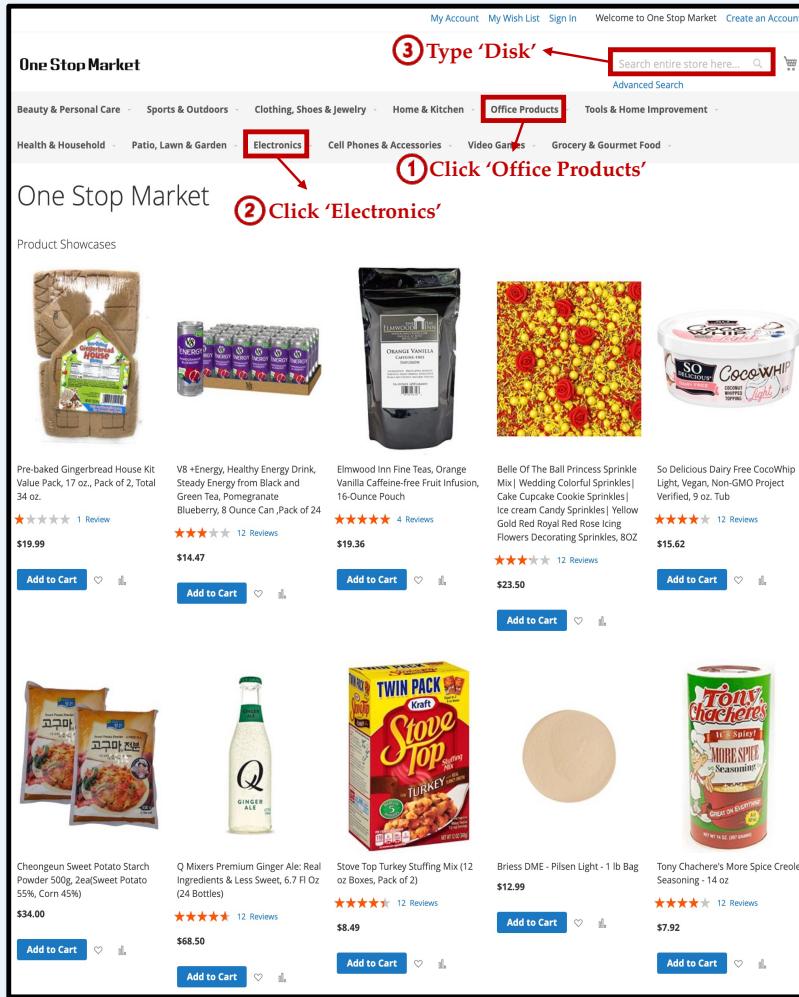
Stage I: Simulation



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

WebDreamer: model-based planner for web agents

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



Stage I: Simulation



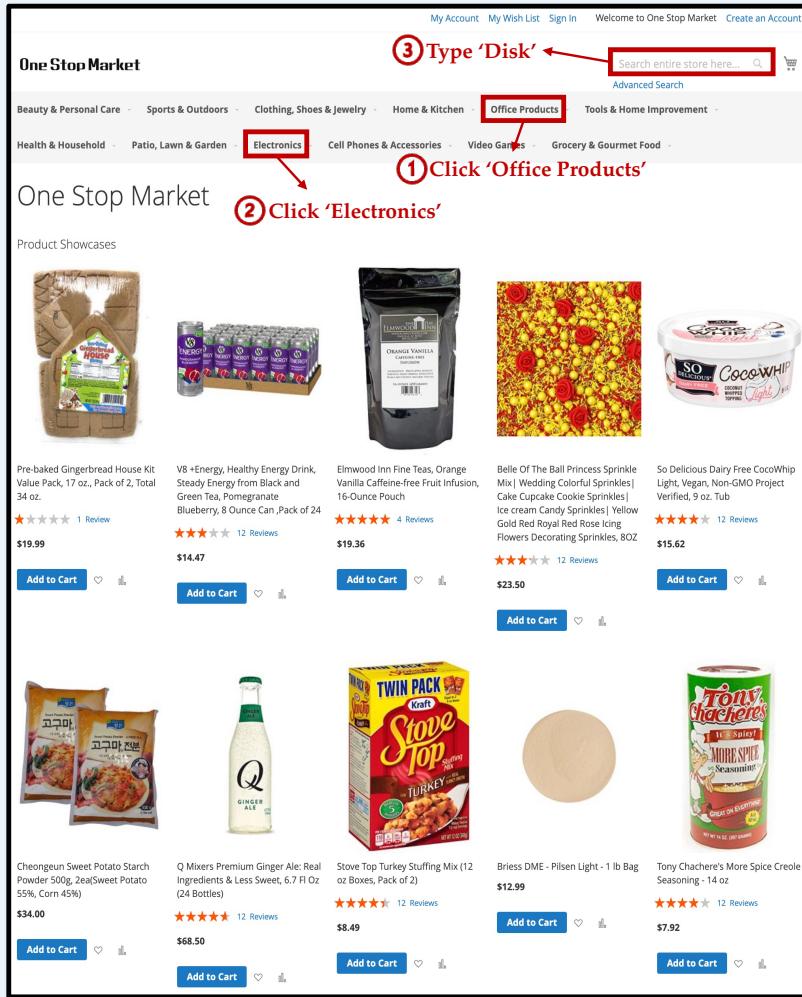
The 'Office Products' category will display three sub-categories: 'Office Electronics', 'Office & School Supplies', and 'Office Furniture & Lighting'.

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 planner for web agents

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



Stage I: Simulation



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

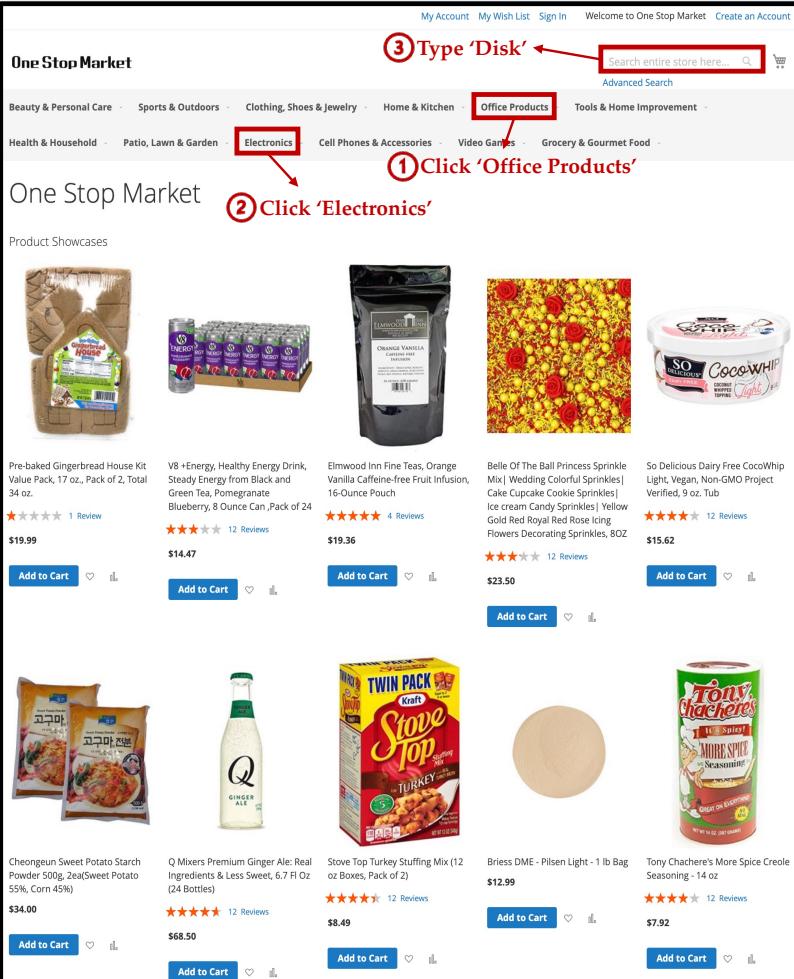
Click 'Office Electronics'

v = 0.4

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 planner for web agents

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



The screenshot shows a navigation bar with links for My Account, My Wish List, Sign In, Welcome to One Stop Market, and Create an Account. Below the navigation bar, there are several categories: 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, and Grocery & Gourmet Food. A search bar at the top right has the placeholder "Search entire store here...". Below the categories, there are product cards for items like a gingerbread house kit, energy drinks, tea, and various food products. Some products have "Add to Cart" buttons.

Stage I: Simulation

① Click 'Electronics'

② Type 'Disk'

③ Click 'Office Products'

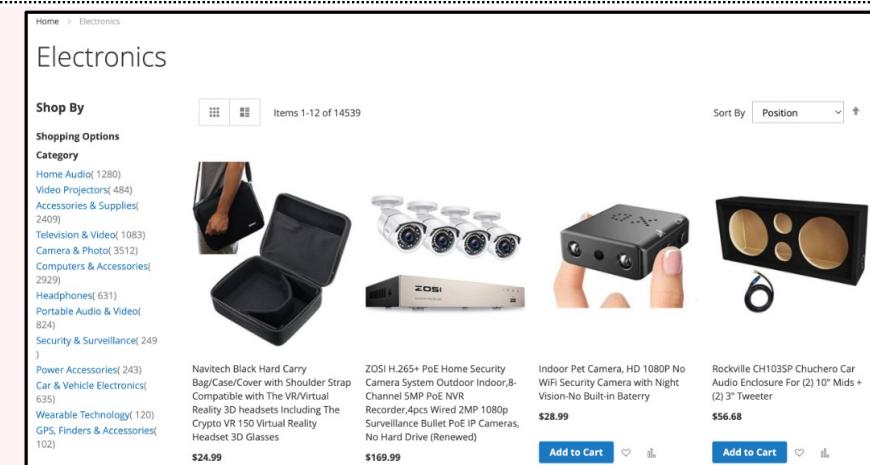
1 The 'Office Products' category will display three sub-categories: 'Office Electronics', 'Office & School Supplies', and 'Office Furniture & Lighting'. v = 0.4

2 The 'Electronics' category will display three sub-categories: 'Computers & Accessories', 'Accessories & Supplies', and 'Car & Vehicle Electronics'. v = 0.8

3 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

2 Click 'Electronics'

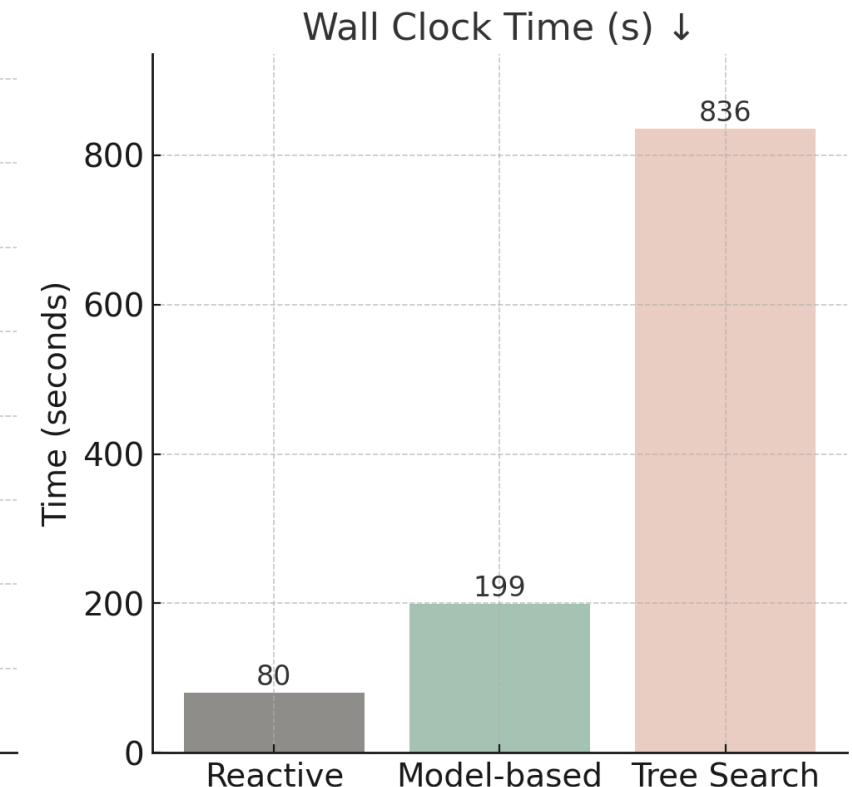
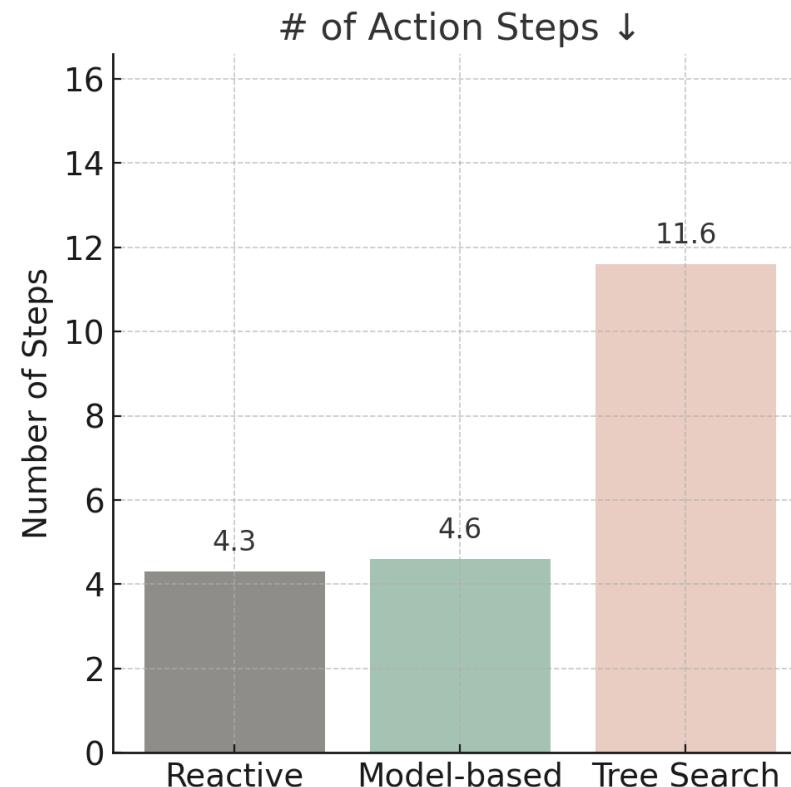
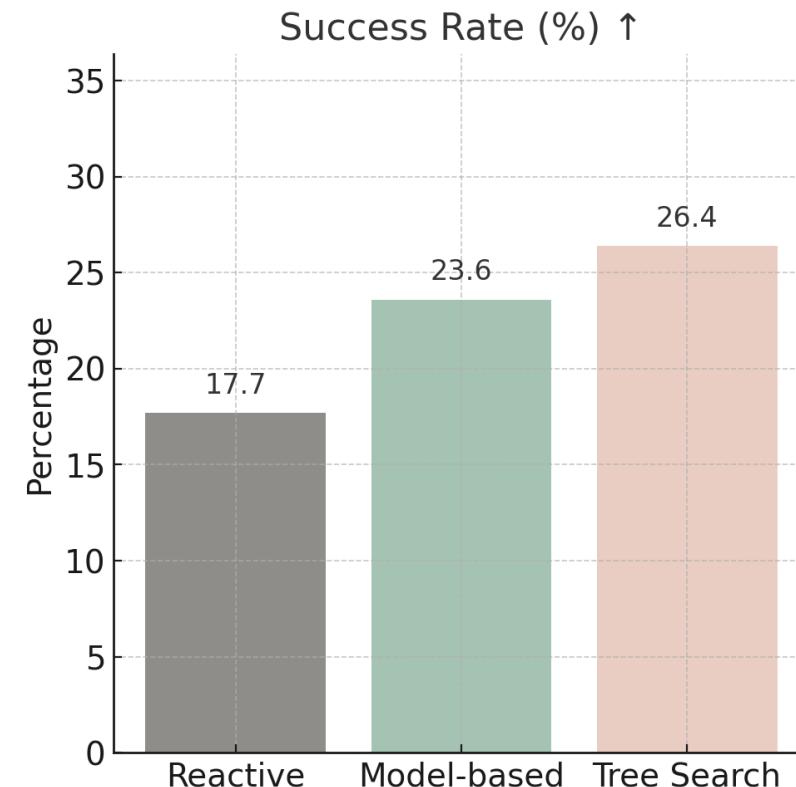


The screenshot shows the 'Electronics' category page with a sidebar for 'Shop By' and a main area displaying 14539 items. Sub-categories listed include Home Audio, Video Projectors, Accessories & Supplies, Television & Video, Camera & Photo, Computers & Accessories, Headphones, Portable Audio, Security & Surveillance, Power Accessories, Car & Vehicle Electronics, Wearable Technology, GPS, Finders & Accessories, Navitech Black Hard Carry Bag/Cover with Shoulder Strap, ZOSI H.265+ PoE Home Security Camera System Outdoor Indoor 8-Channel 5MP PoE NVR, Indoor Pet Camera, and Rockville CH103SP Chuchero Car Audio Enclosure For (2) 10" Mids + (2) 3" Tweeter.

44

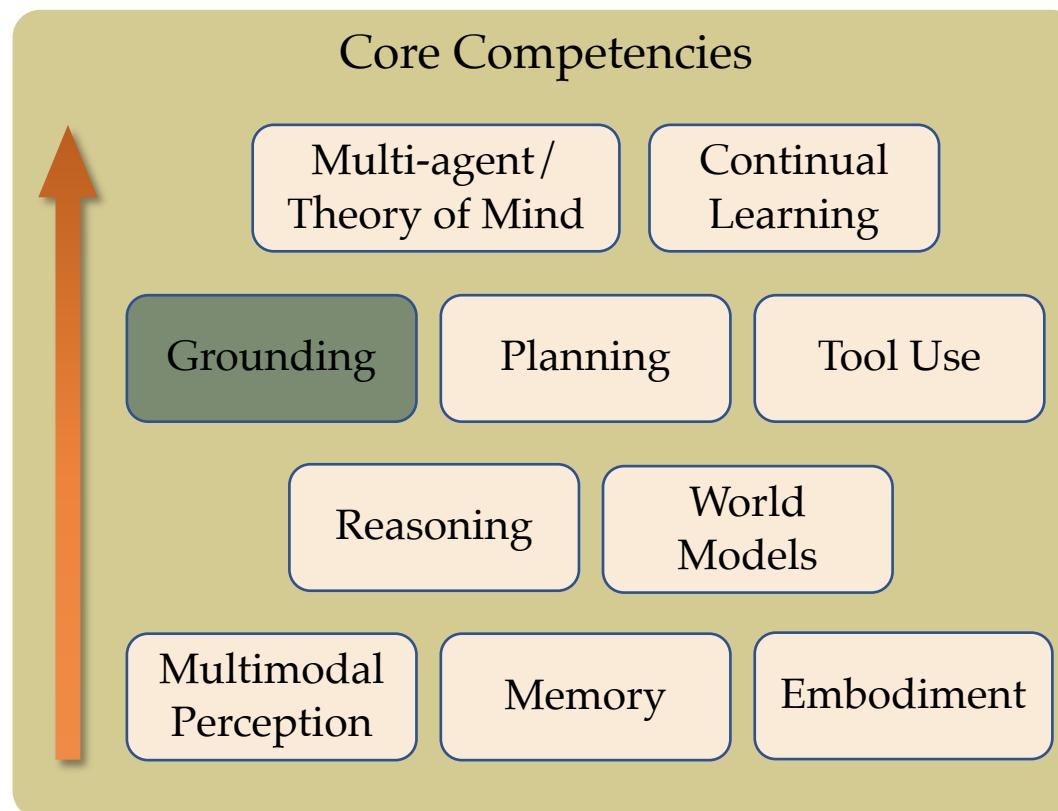
Results on VisualWebArena

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

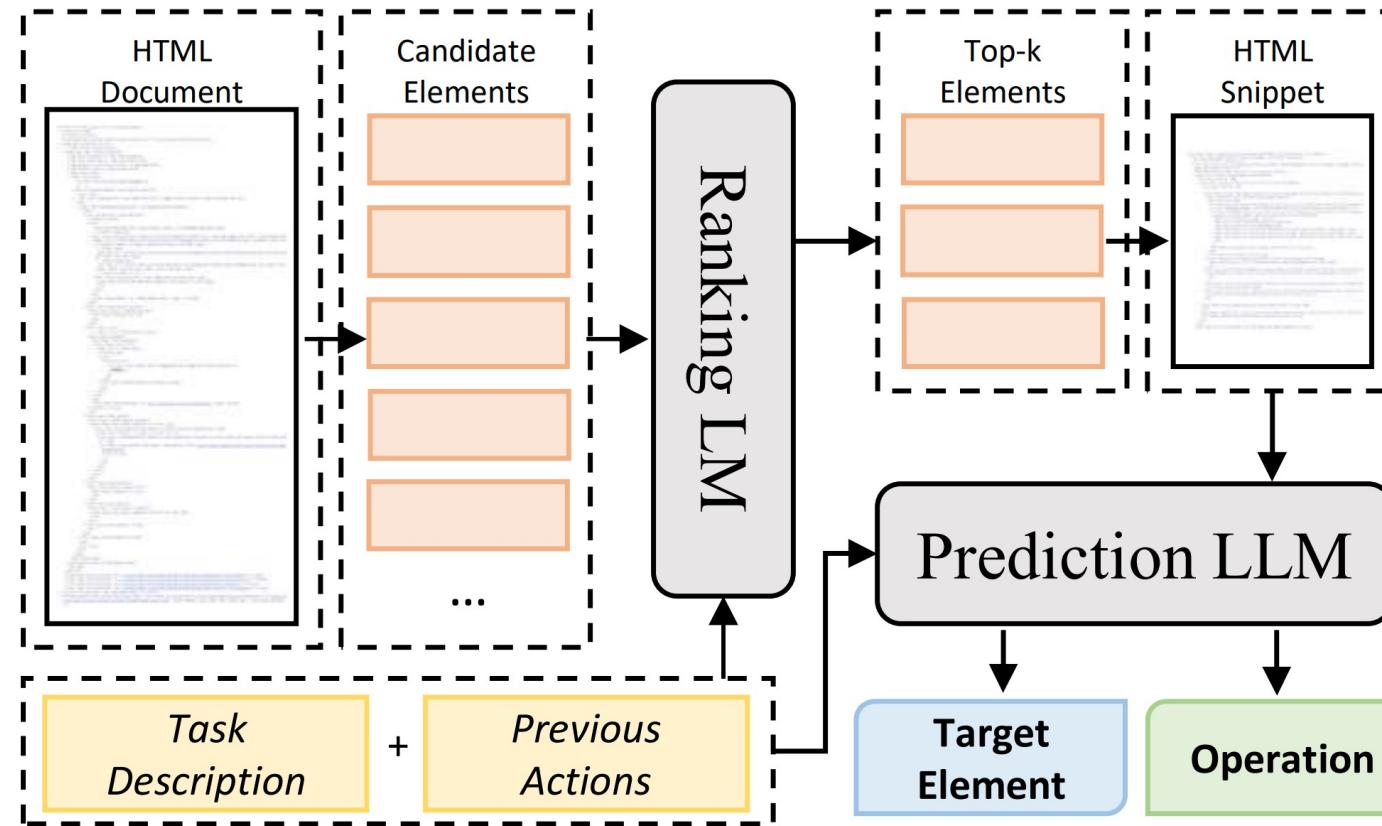


Baselines are from Koh et al., (2024)

Memory



Embodiment of Web/GUI Agents: Evolution



Mind2Web (23' May)

Sensory Inputs	HTML/DOM	
Effectors	Multi-choice Selection	

Embodiment of Web/GUI Agents: Evolution

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: [Moving Trucks & Accessories](#)
B:
...
F: No name="one-way-radio"
G:
H: None

Element Attributes

TEXT: Find Your Truck
TYPE: BUTTON



Boyuan Zheng

A screenshot of a form titled "Moving Trucks & Accessories". The form has several input fields: 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: "Will you return the truck to a different location?", and G: "Find Your Truck". The "Find Your Truck" button is highlighted with a red border. There are also radio buttons for "Yes" and "No" under the "Will you return the truck to a different location?" section.

Image Annotation

CHOICE: G



Boyu Gou

Textual Choices

CHOICE: G

Mind2Web (23' May)

Sensory Inputs

HTML/DOM

Effectors

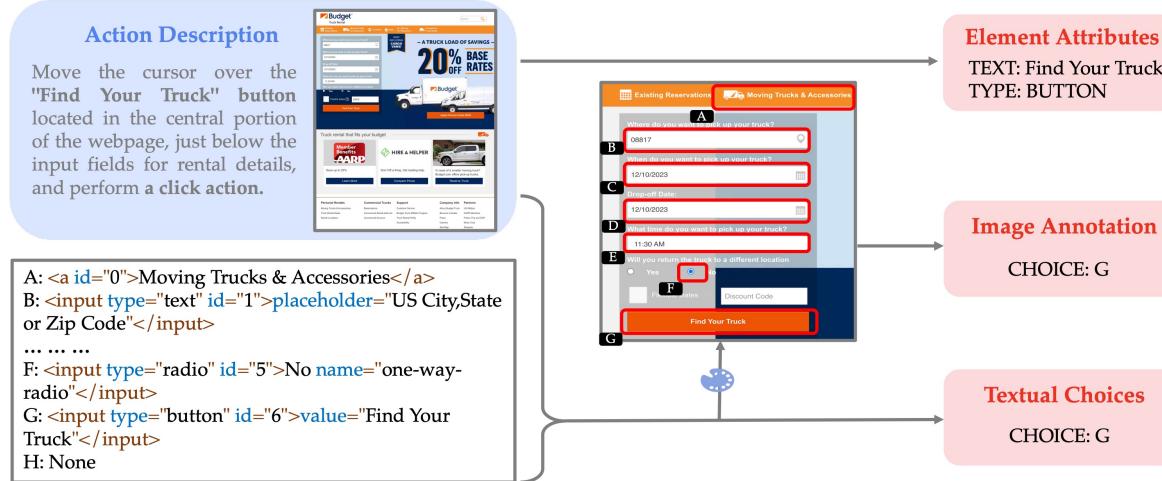
Multi-choice Selection

SeeAct (24' Jan)

Screenshot + DOM

Multi-choice Selection

Embodiment of Web/GUI Agents: Evolution



Text-based Representations: Limitations

- *Noisy and incomplete*
 - 95.9% of home pages have accessibility conformance errors
 - Avg. 56.8 errors per page¹
- *Additional input increases latency and inference costs*
 - Consuming more tokens
 - Difficult and time-consuming to get
 - Compounding over long horizon

Mind2Web (23' May)	SeeAct (24' Jan)		
Sensory Inputs	HTML/DOM	Sensory Inputs	Screenshot + DOM
Effectors	Multi-choice Selection	Effectors	Multi-choice Selection

¹ <https://webaim.org/projects/million/>

Embodiment of Web/GUI Agents: Evolution

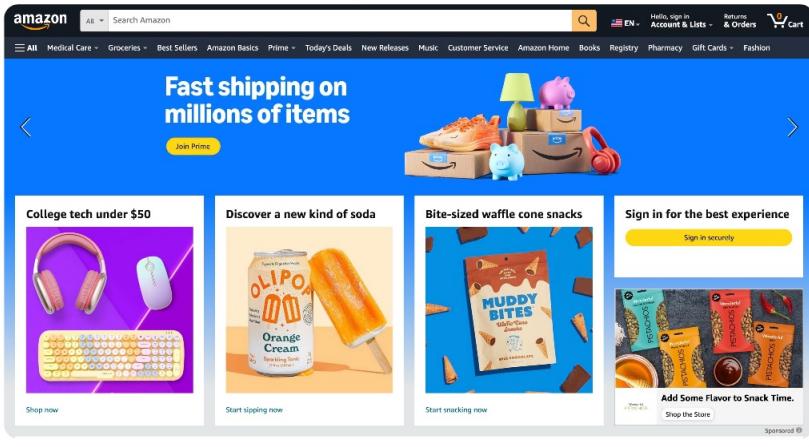


SeeAct-V: Human-like, Vision-centric GUI Agent

Boyu Gou

Vision-Only Observation

TASK: Find the cheapest 4k monitor

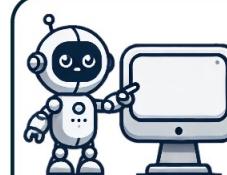


Planning



Element Description:
The search bar at the top of the page
Action: Type
Value: 4k monitor

Grounding

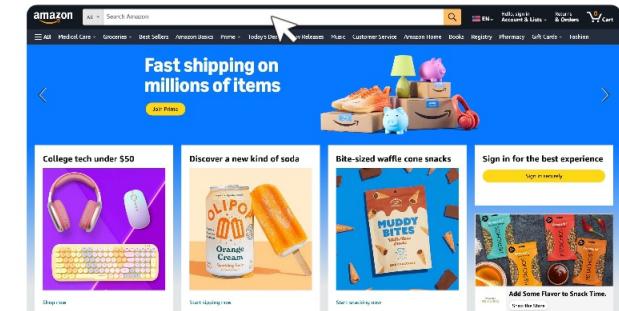


What are the pixel coordinates of
the element corresponding to "..."?
(556, 26)

Human-like Operation



Click(556, 26)
Type("4k monitor")



Mind2Web (23' May)

Sensory Inputs

HTML/DOM

SeeAct (24' Jan)

Screenshot + DOM

SeeAct-V (24' Aug)

Screenshot Only

Effectors

Multi-choice Selection

Multi-choice Selection

Pixel-level Operations

Also see CogAgent and SeeClick for prior work along this line

Visual Grounding is the Bottleneck

GPT-4V(ision) is a Generalist Web Agent, if Grounded

Boyuan Zheng¹ Boyu Gou¹ Jihyung Kil¹ Huan Sun¹ Yu Su¹

<https://osu-nlp-group.github.io/SeeAct>

Abstract

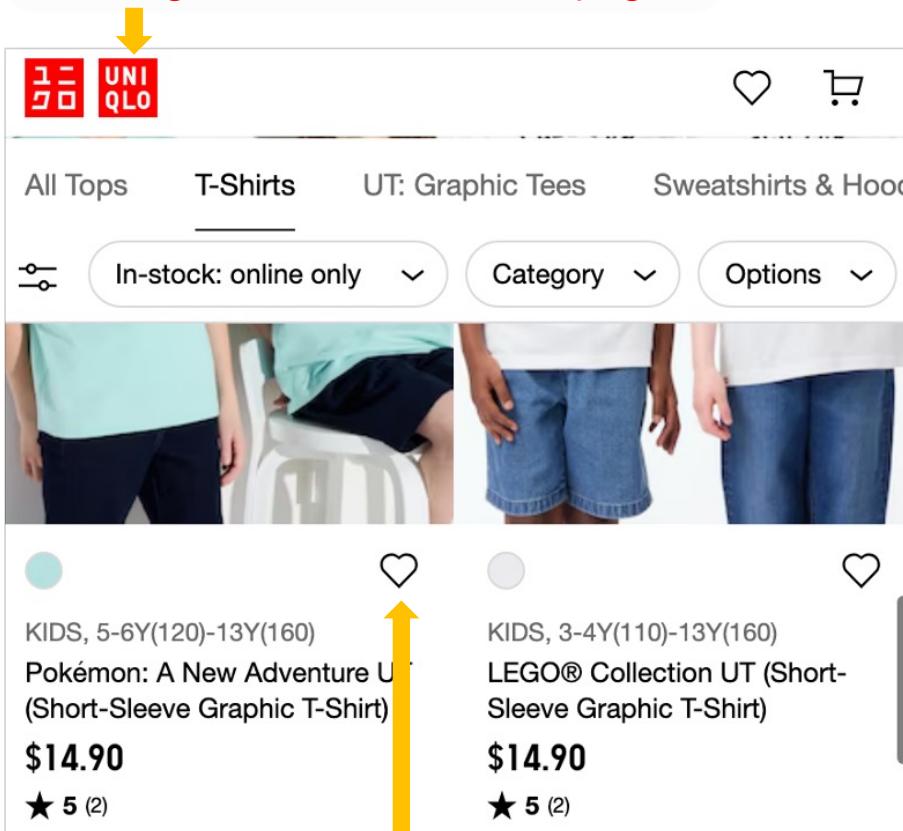
The recent development on large multimodal models (LMMs), especially GPT-4V(ision) and Gemini, has been quickly expanding the capability boundaries of multimodal models beyond traditional tasks like image captioning and visual question answering. In this work, we explore the potential of LMMs like GPT-4V as a generalist web agent that can follow natural language instructions to complete tasks on any given website. We propose SEEACT, a generalist web agent that harnesses the power of LMMs for integrated visual understanding and acting on the web. We evaluate on the recent MIND2WEB benchmark. In addition to standard offline evaluation on cached websites, we enable a new online evaluation setting by developing a tool that allows running web agents on live websites. We show that GPT-4V presents a great potential for web agents—it can successfully complete 51.1% of the tasks on live websites if we manually ground its textual plans into



Figure 1: SEEACT leverages an LMM like GPT-4V to visually perceive websites and generate plans in textual forms. The textual plans are then grounded onto the HTML elements and operations to act on the website.

Referring Expressions for GUIs are Diverse

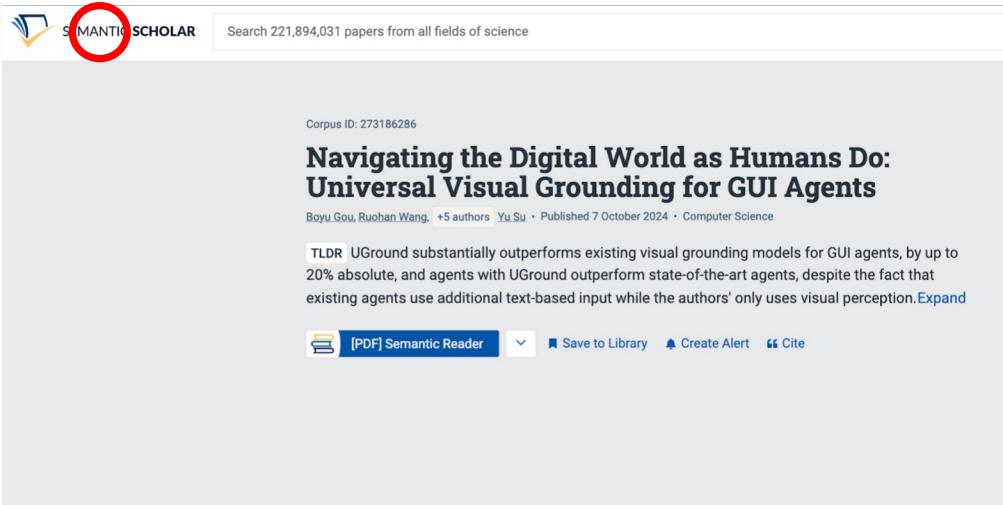
1. Red icon labeled “UNIQLO”
2. Button at the top left corner
3. Navigate back to the homepage



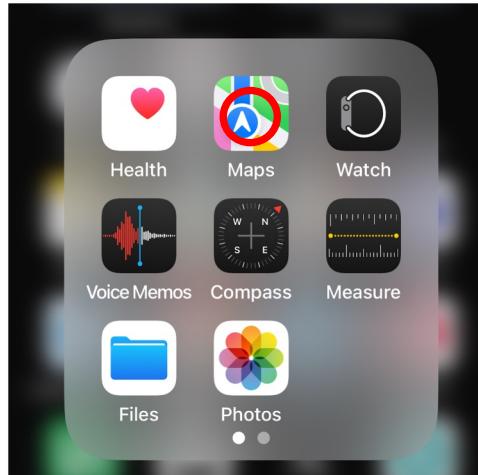
1. Hollow heart button
2. Button below the Pokémon shirt
3. Favor the Pokémon shirt

- **Visual Referring Expressions**
 - Salient visual features like textual content, element type (button, input field, checkbox, etc.), shape, color, ...
- **Positional Referring Expressions**
 - including **absolute** (e.g., “*at the top left of the page*”) and **relative** positions (e.g., “*to the right of element X*”)
- **Functional Referring Expressions**
 - Referring to elements by their functions
- **Hybrid**
 - “*click the heart button under the Pokemon shirt to add it to favorite.*”

Shared Designs across GUIs



“Go to homepage”



“Open Maps”

The screenshot shows the "arXiv info" page. At the top, there's a red circle highlighting the arXiv logo. The page features a sidebar with navigation links: Home, About, Help, Brand, Labs, Policies (with a dropdown menu for arXiv Policies, Code of conduct, Privacy policy, Identity, affiliation, and registration, Submission terms and agreement, Paper ownership, General submission policies, Submission schedule and cutoff time, Content Moderation, and License and copyright). The main content area is titled "Availability of submissions" and contains text about submission timelines and identifiers. It also includes a note about moderation and a section about arXiv-id assignments.

“Availability of submissions”

Submissions to arXiv are typically made public Sunday through Thursday, with no announcements Friday or Saturday. Check the [current time](#) at the arXiv main site, with the amount of time remaining before the deadline. Submissions are made public as part of the scheduled announcement process. This includes new submissions as well as [replacements](#), [withdrawal notices](#), [cross listings](#) and [journal reference](#). Journal reference submissions do not appear in the mailings, but are still processed at that time.

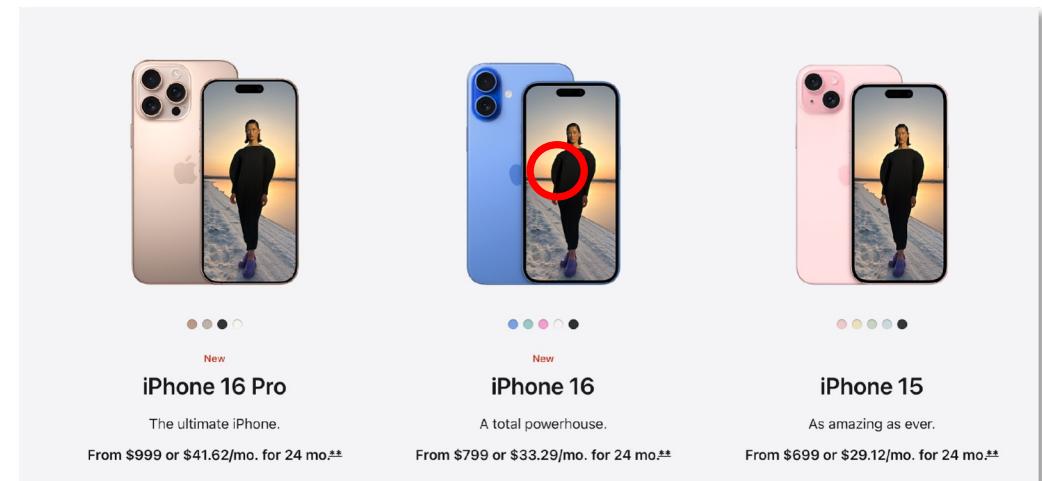
Note that all submissions are subject to [moderation](#) which may result in delays.

A note about arXiv-id assignments

The final [arXiv identifier](#) cannot be assigned until the paper is announced, so please do not ask for it ahead of announcement. It is not possible to make exceptions, nor accelerate the announcement schedule regardless of your circumstance.

Note: The arXiv identifier cannot be back-dated, so identifiers will be assigned in the month of first

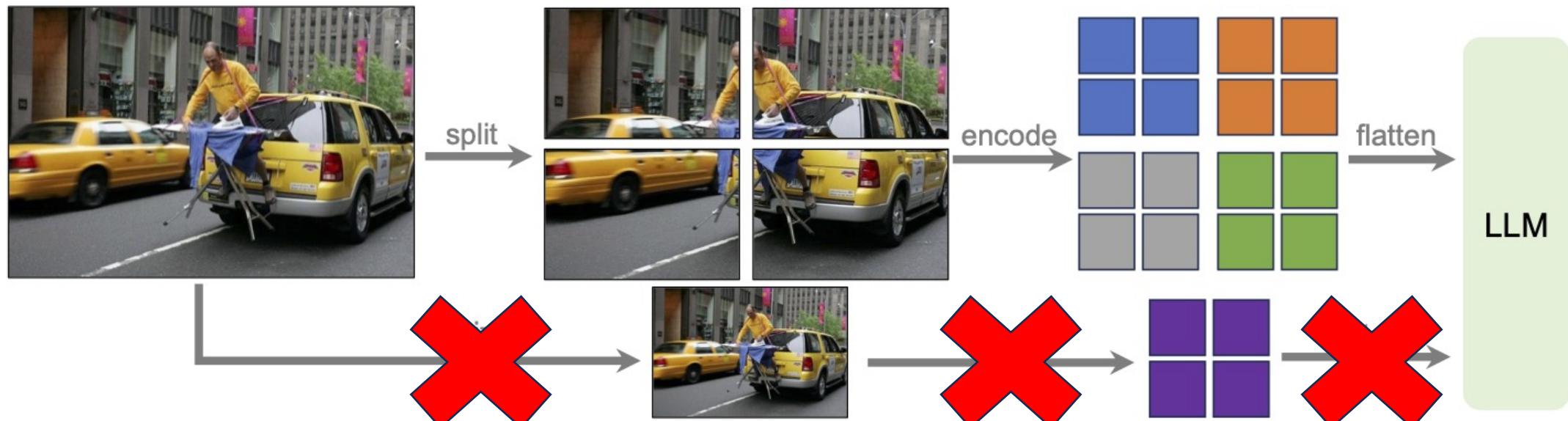
“Go to homepage”



“iPhone 16”

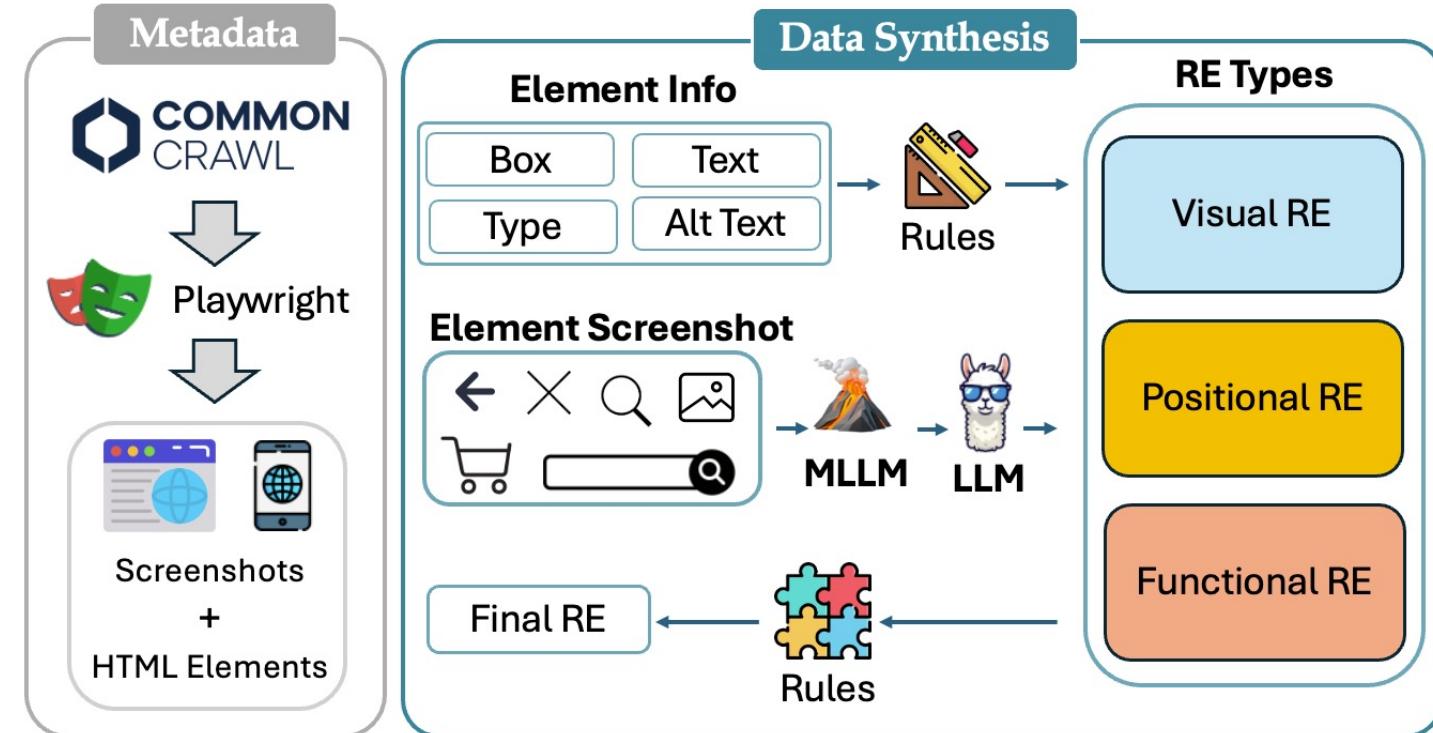
UGround: Universal Visual Grounding for GUI Agents

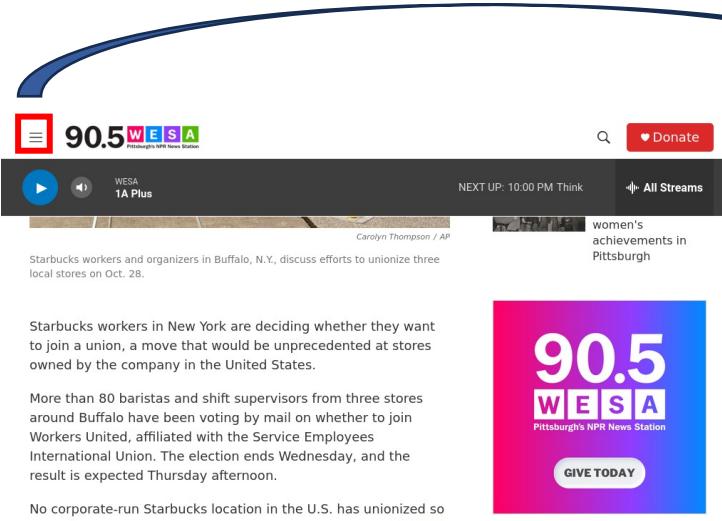
- What do we mean by universal?
 - Generalize across different web, desktop, and mobile environments
 - Plug and play with different planners / LLMs
 - Handle input images of different resolutions
- Modeling
 - Input: Screenshot + RE; Output: Pixel Coordinates
 - LLaVA backbone: CLIP@224px, Vicuna-1.5-7B-16K, AnyRes



Synthetic Data is Key

- Collect all relevant information about an element
 - Visual (i.e., screenshot)
 - HTML attributes like inner-text, alt-text as additional visual cues
 - Accessibility labels like alt, aria-label as functionality cues
- Generate referring expressions using (multimodal) LLMs and rules
- + Positional information





Screenshot



HTML

```
type: button  
aria-label:  
menu
```



MLLM Interpretation

The image shows a familiar "hamburger menu" icon, which consists of three horizontal, evenly spaced lines stacked vertically. This icon is often used in web and mobile interfaces to indicate a collapsible or expandable menu, commonly referred to as the "menu" or "navigation" icon. Its minimalistic design helps users quickly recognize it as a toggle for hidden menu items or additional navigation options. Given the context from the HTML information, it appears this icon is intended to open a menu when clicked.



Rules

Absolute Position: Top Left Corner

Relative Position: to the left of "Donate"

...



LLM Rewrite

three-line menu icon



Final Referring Expression

three-line menu icon, at the top left corner of the page

Training Data Statistics

Dataset	Annotation	# of Elements	# of Screenshots	Platform
Web-Hybrid (Ours)	Rule + LLM	9M	773K	Web
Web-Direct (Ours)	GPT	408K	408K	Web
GUIAct (Chen et al., 2024)	GPT + Human	140K	13K	Web
AndroidControl (Li et al., 2024b)	Human	47K	47K	Android
Widget Caption (Li et al., 2020b)	Human	41K	15K	Android
UIBert (Bai et al., 2021)	Human	16K	5K	Android
AITZ (Zhang et al., 2024b)	GPT + Human	8K	8K	Android
Total		10M	1.3M	Web + Android

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

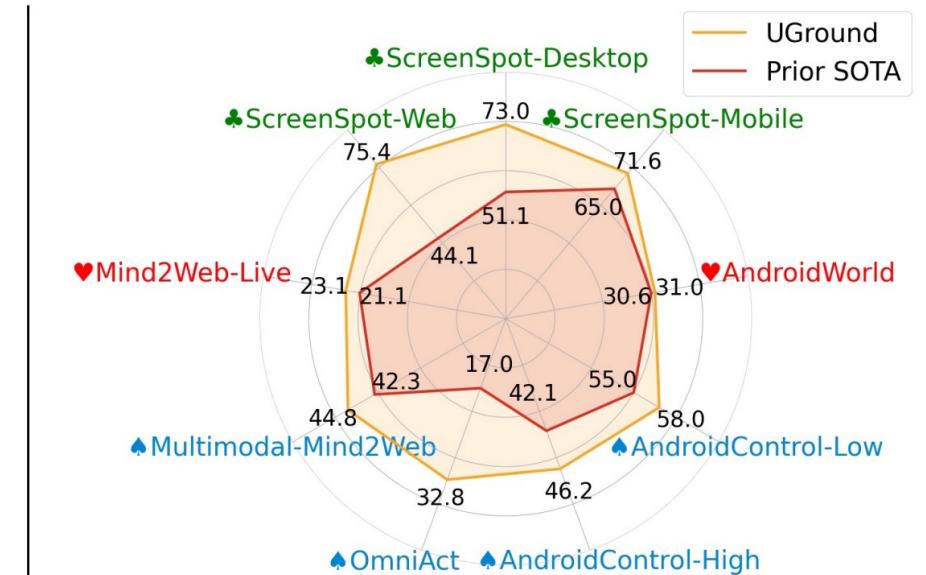
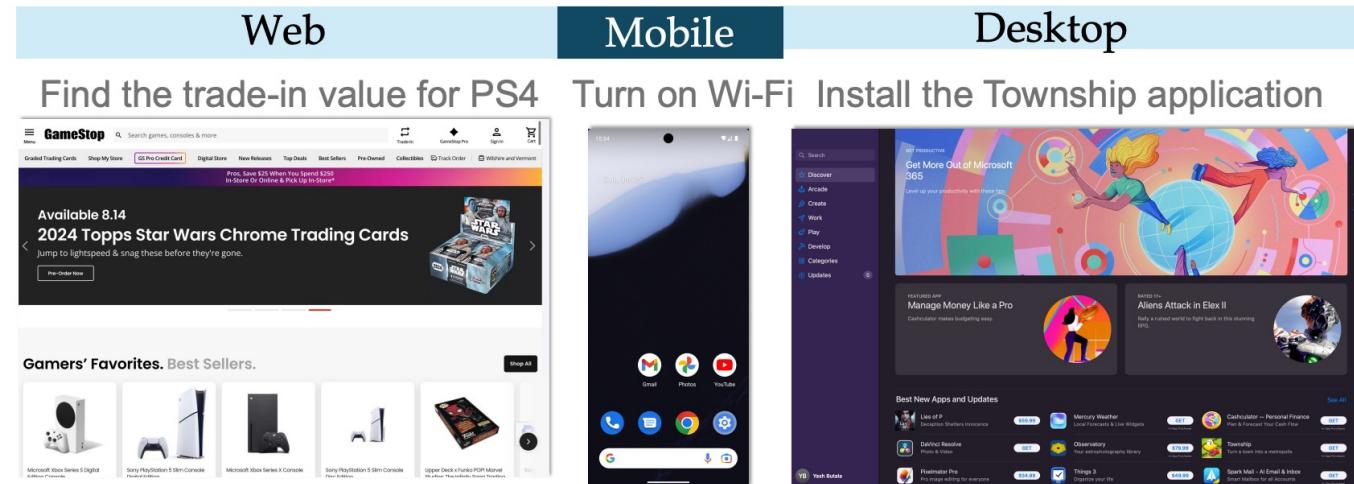
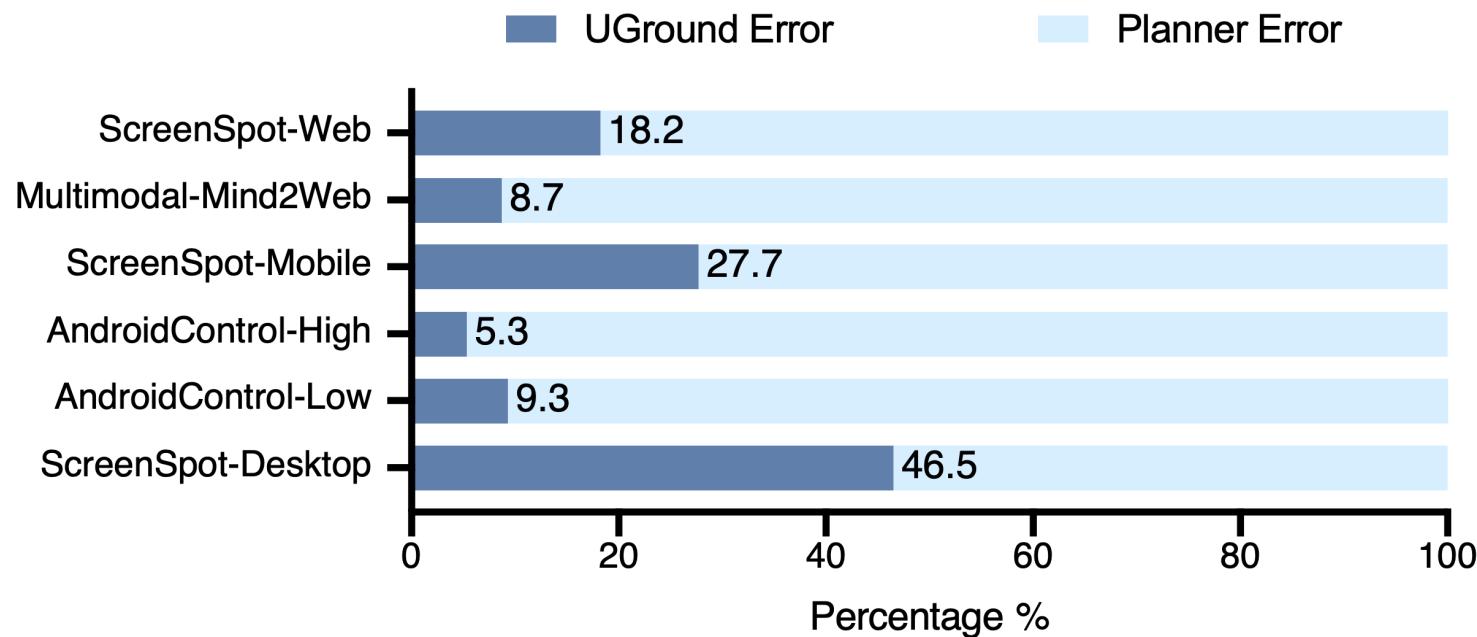


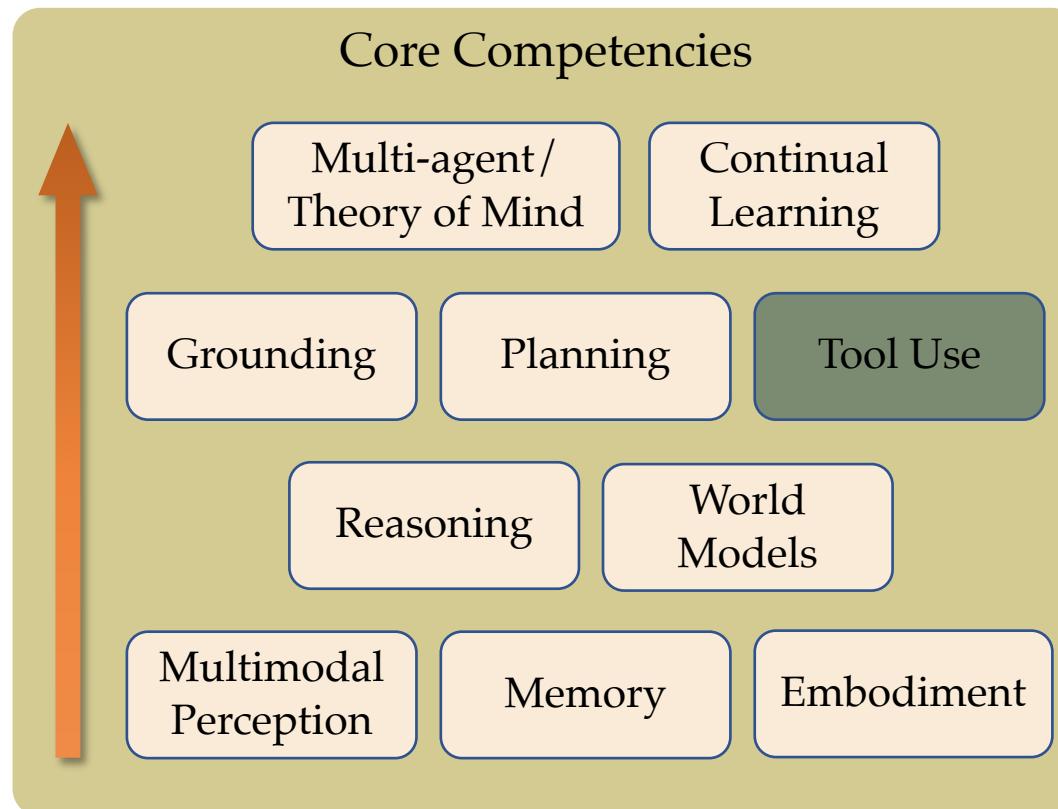
Figure 1: Examples of agent tasks across platforms and performance on **GUI grounding** (♣: ScreenSpot), **offline agent** (♠: Multimodal-Mind2Web, AndroidControl, and OmniAct), and **online agent benchmarks** (♥: Mind2Web-Live and AndroidWorld) when using GPT-4 as the planner.

Error Analysis

- Most remaining errors are due to planning rather than grounding
- Grounding errors on mobile and desktop are higher, but that doesn't necessarily manifest in the agent performance



Tool Use



LLMs in the Imaginarium: Tool Learning through Simulated Trial and Error

**Boshi Wang, Hao Fang, Jason Eisner,
Benjamin Van Durme, Yu Su**



Microsoft



We don't master a tool by just reading the 'manual'

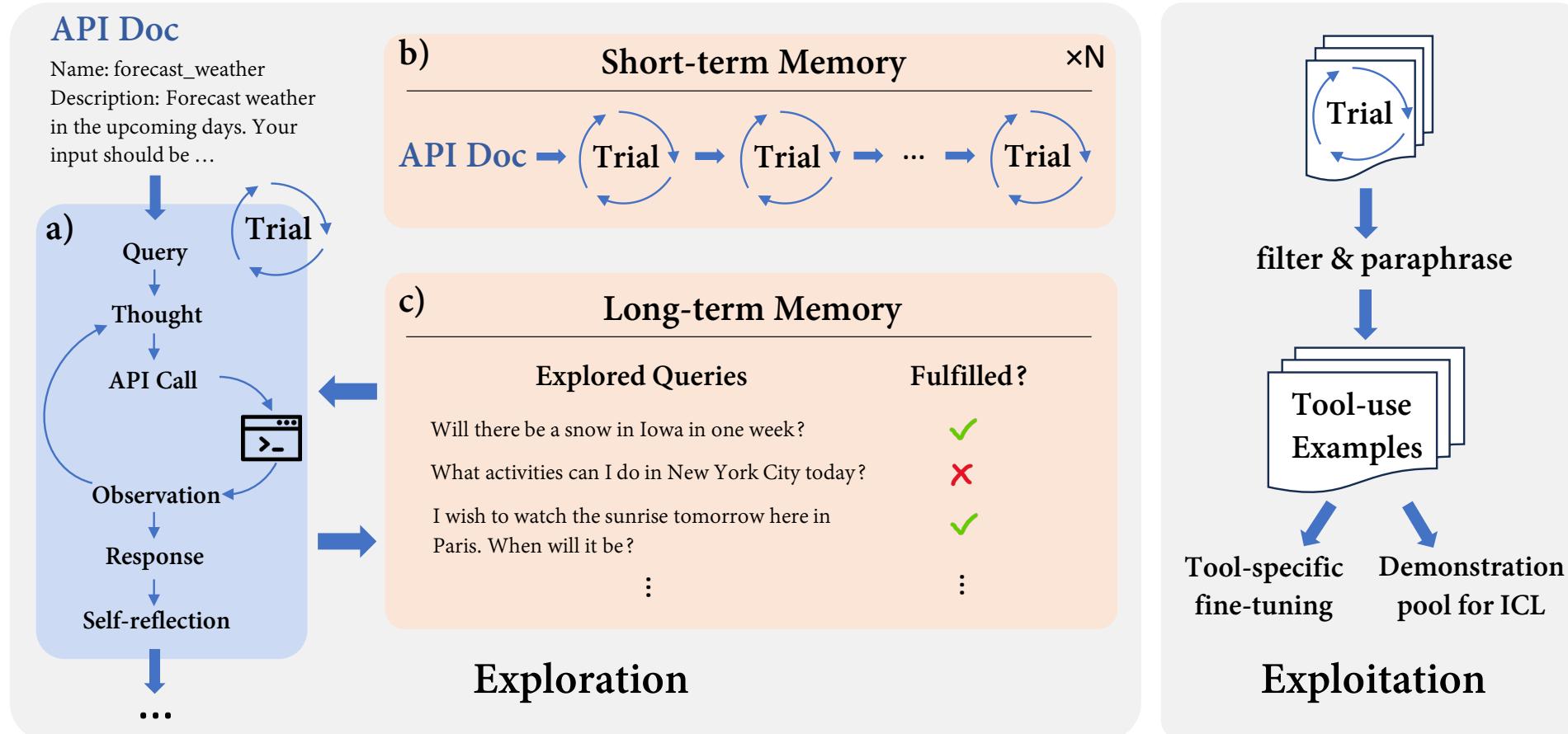


- Trial and error
- Imagination
- Memory

<https://www.youtube.com/watch?v=s2IBayVsbz8>

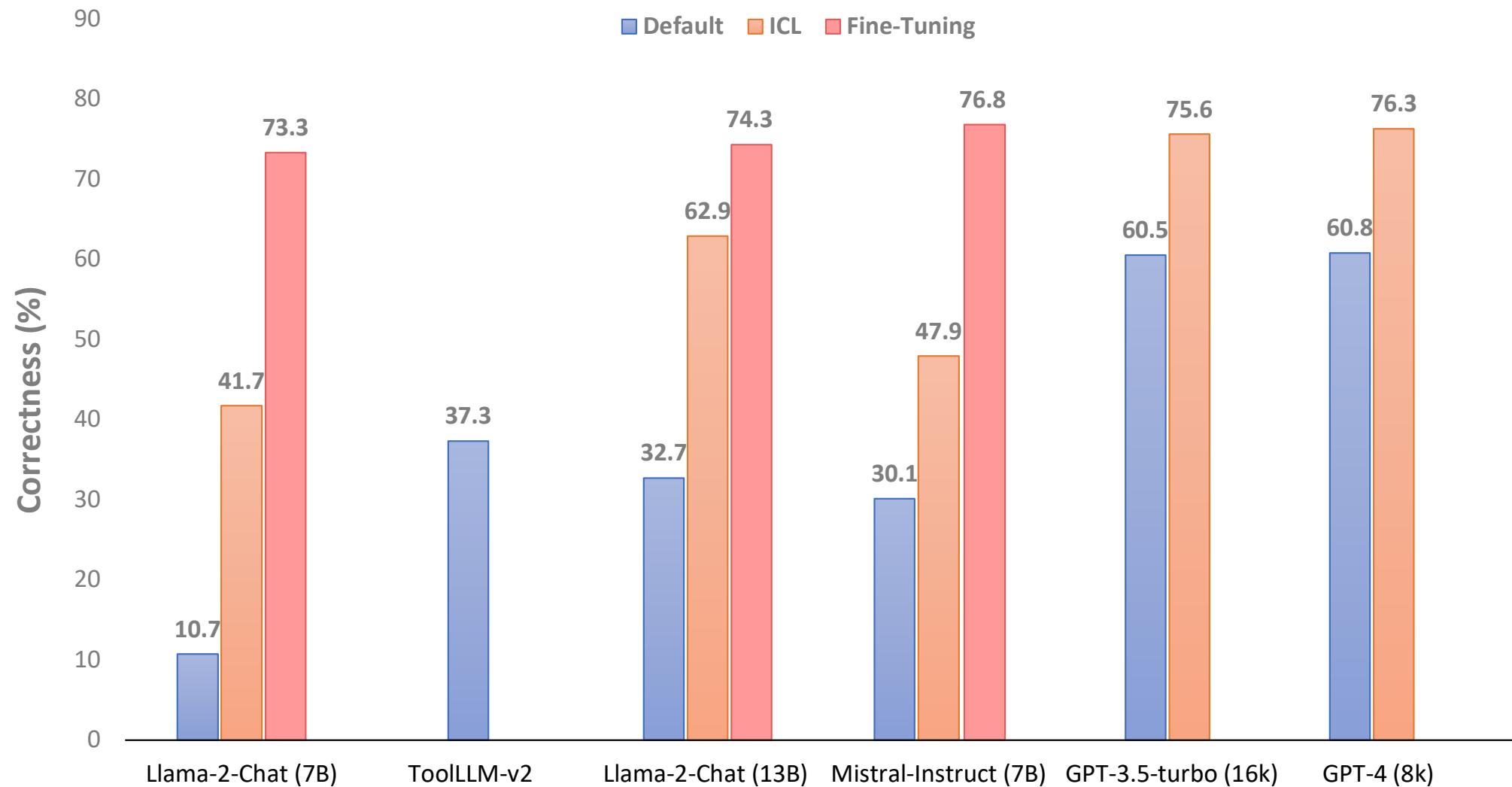
Tool learning through *simulated trial and error*

Synthesizing diverse **perception-decision-execution** data with LLMs



STE substantially improves tool use

ICL: Retrieve from STE synthesized examples. Fine-Tuning: Fine-tune LLM with STE examples



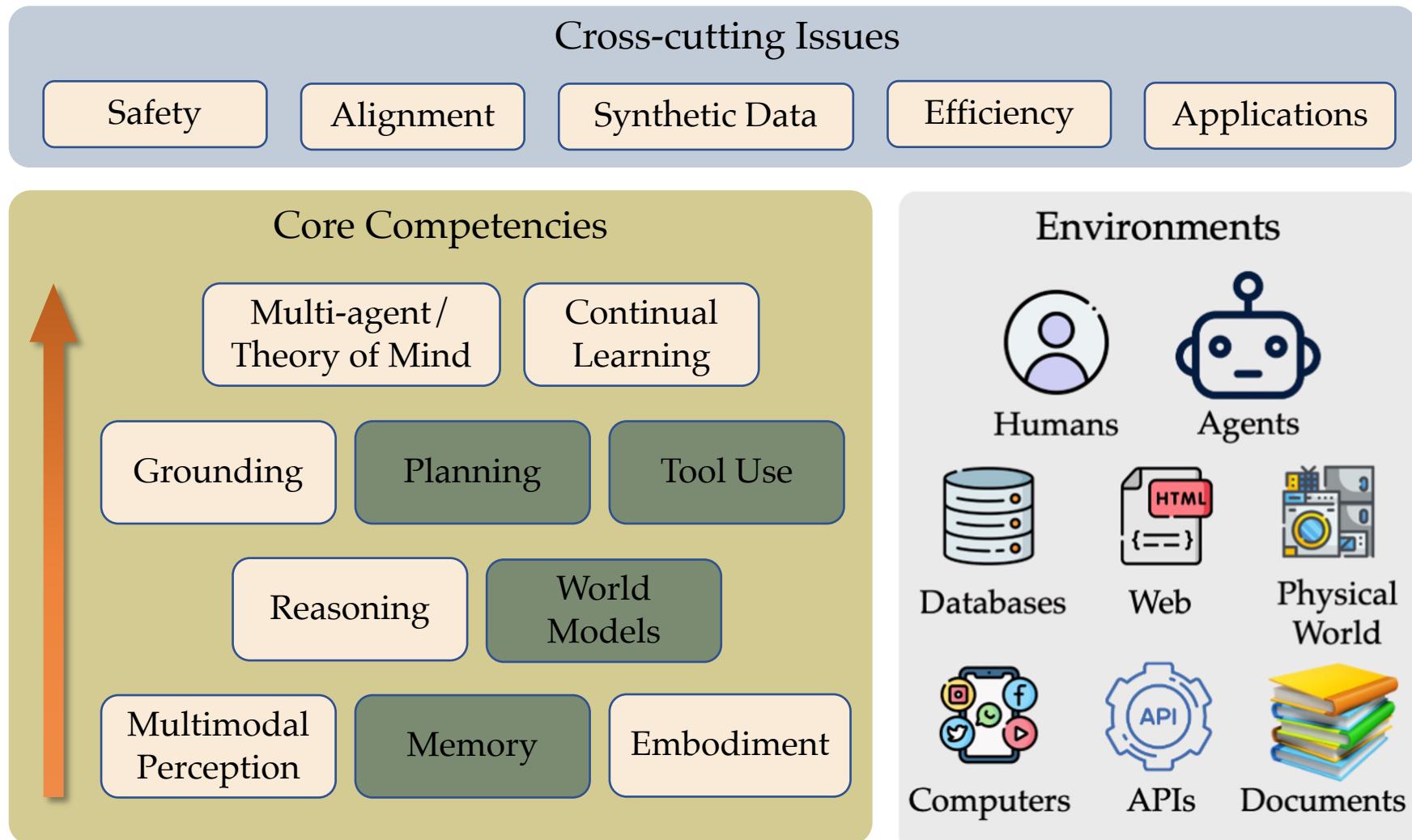
A few remarks

- Language agents are still AI agents. The same challenges remain. But the new capability of using language for reasoning and communication allows us to re-examine these challenges through a new, powerful lens.
- Natural selection is a real magical power. Over billions of years of evolution, natural selection has found numerous good designs ('local optima') that *just make sense*; we should take some inspirations!
- Tons of new venues for interesting research. We are barely scratching the surface...

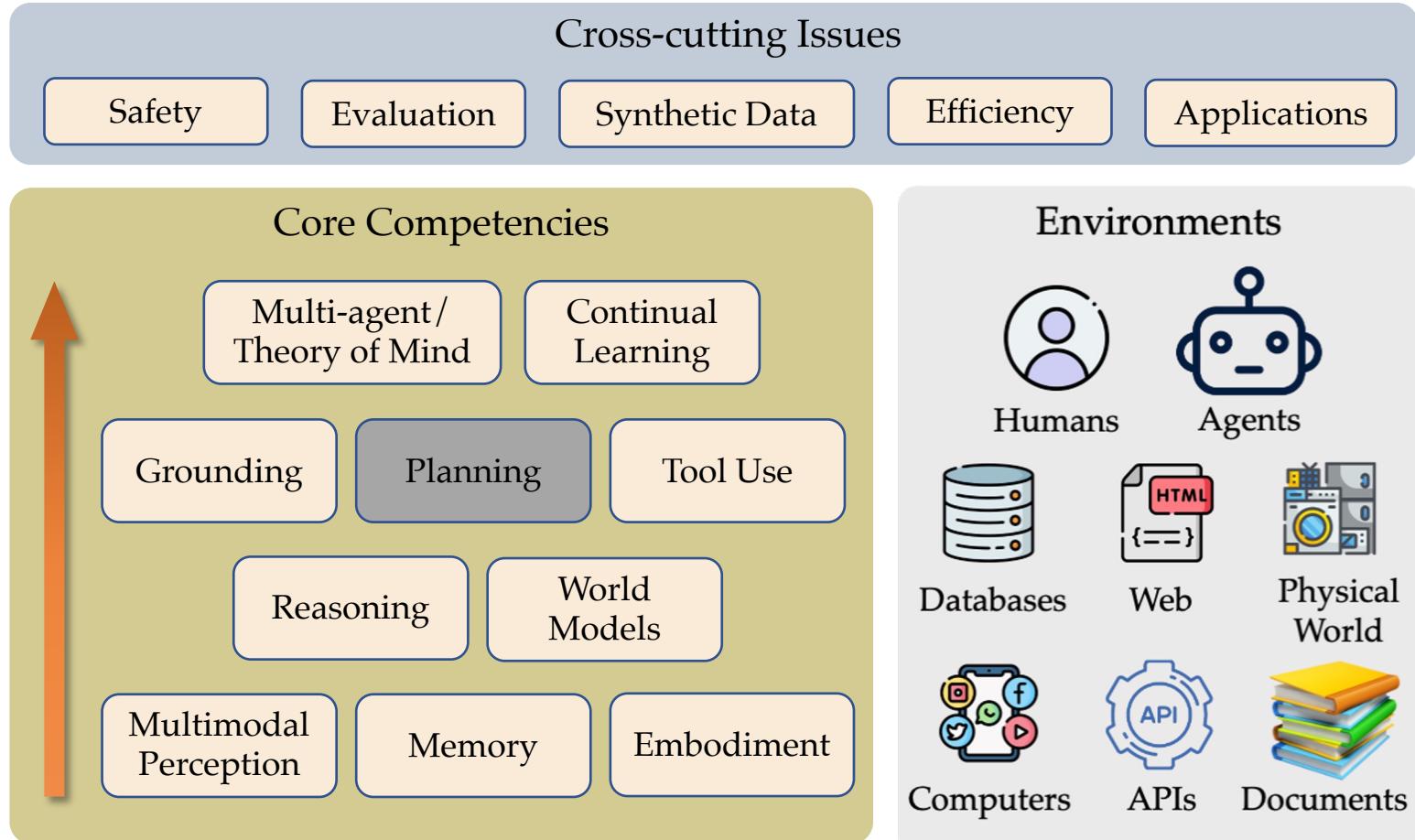
Outline

- Language Agents: Introduction
- The Power of Language: Memory, Planning, Grounding, Tool Use
- Future Directions

We are just standing at the dawn of a long journey

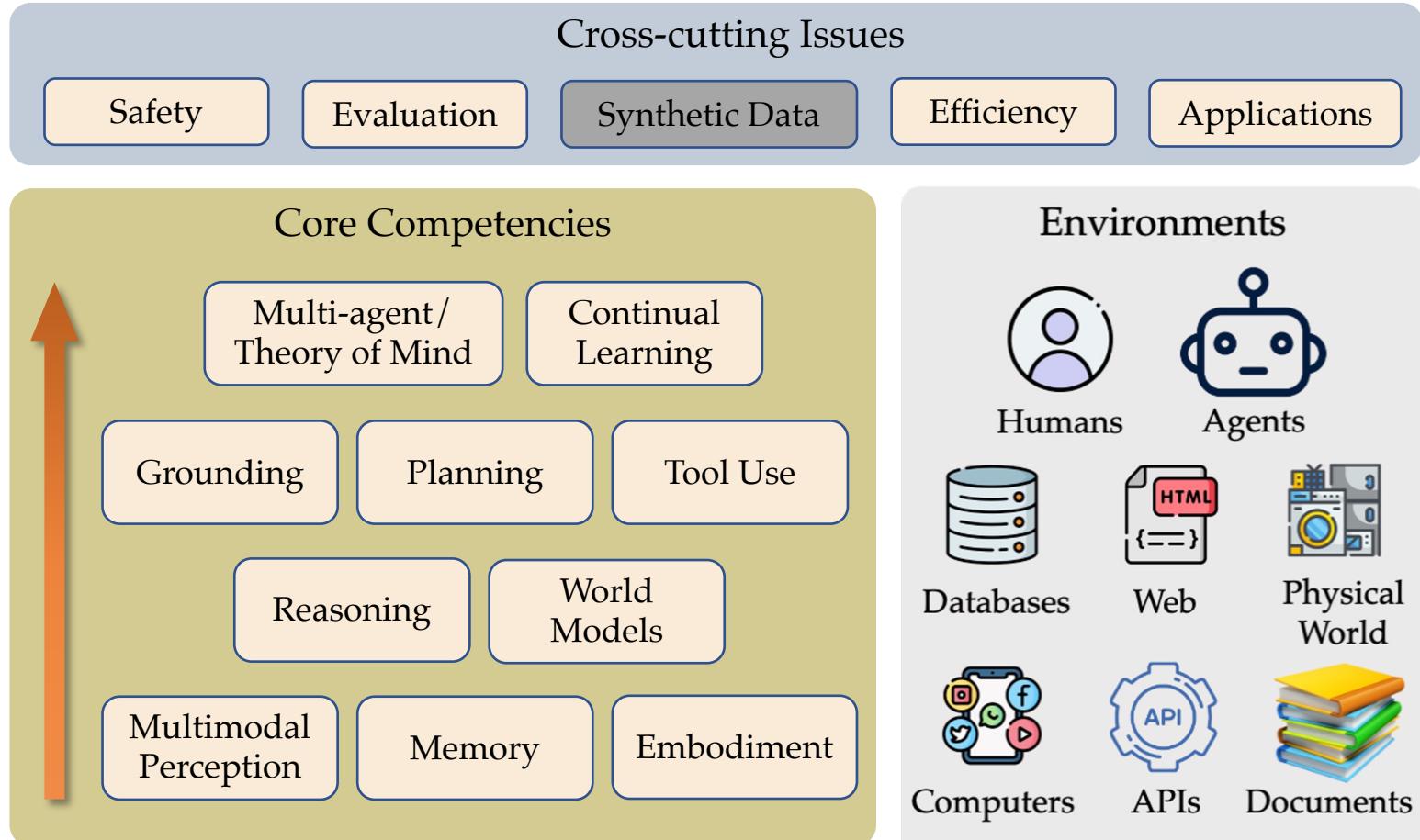


Future directions for language agents



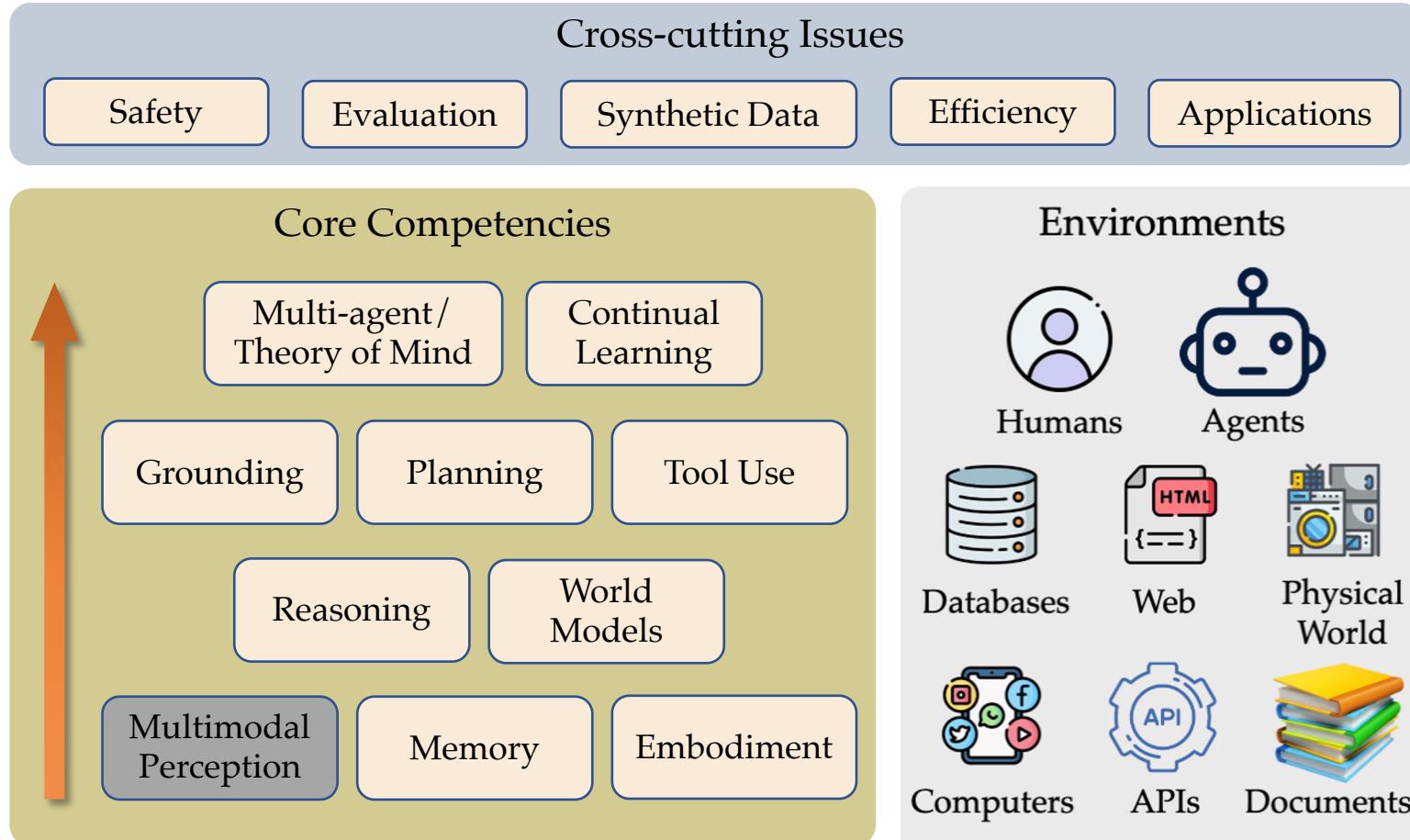
- *Hierarchical planning?*
- *How far can (tree) search go?*
- *How to evaluate (partial) plans? Value functions? Reward models?*
- *Speculative planning?*

Future directions for language agents



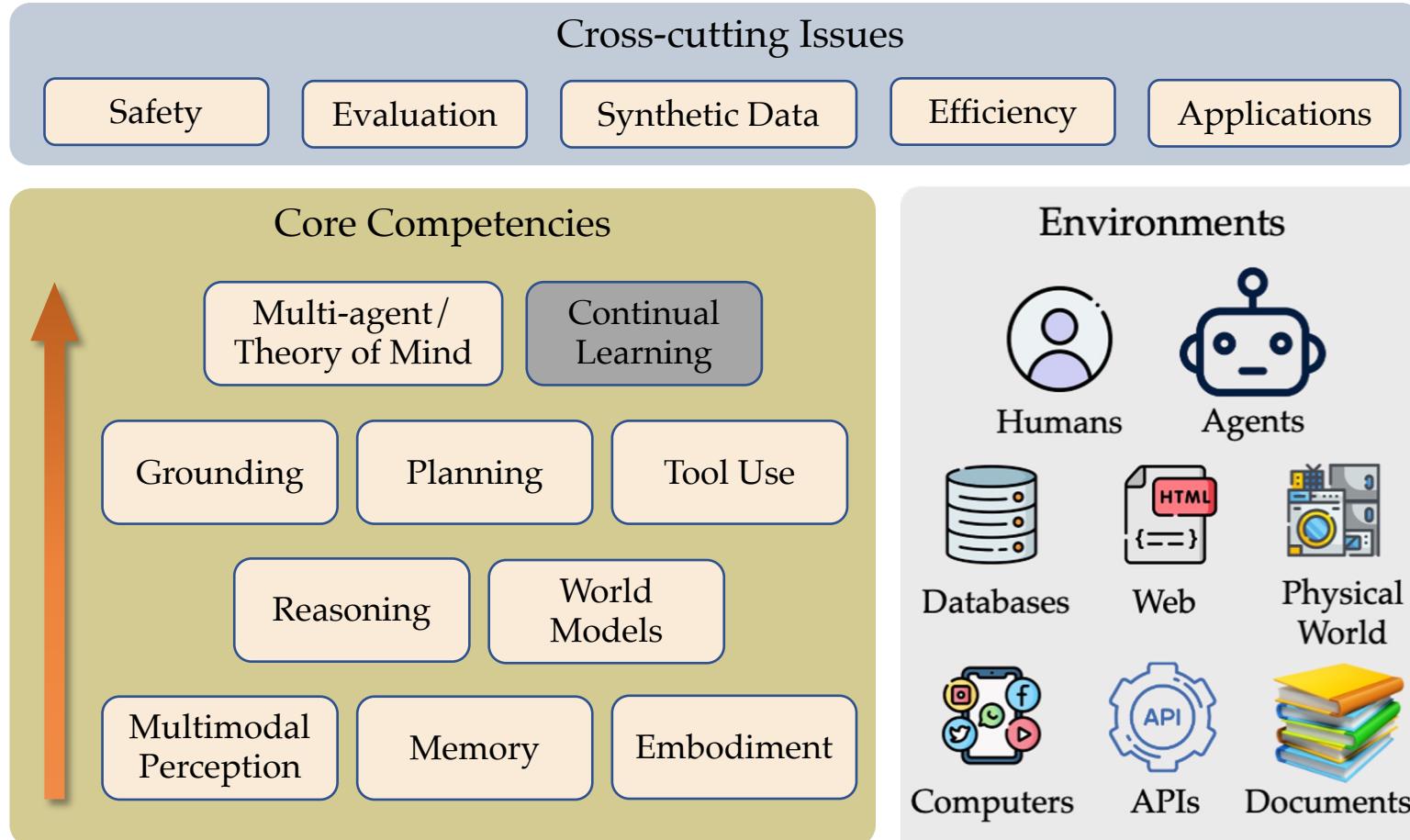
- Agents need to learn **perception-decision-execution** capabilities
- Data on the Internet is mostly artifacts from such processes, not capturing the processes *per se*
- Synthesizing data with LLMs provides a possibility to uncover (some of) these hidden processes

Future directions for language agents



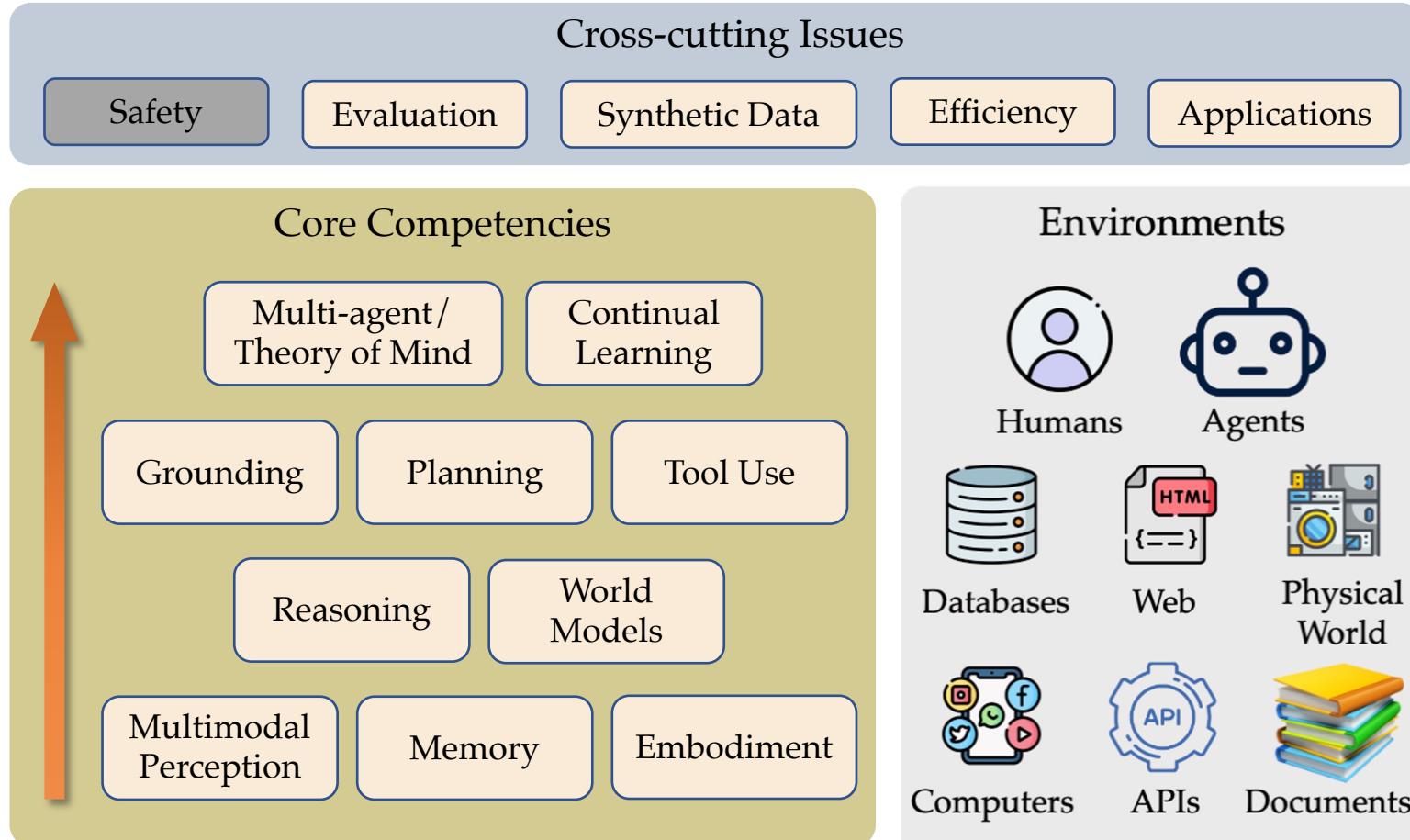
- Existing MLLMs are not optimized for GUI understanding.
 - Resolution/grounding/fine-grained understanding
- Current MLLMs do not seem to fuse visual and language well. New arch or training recipes are needed.

Future directions for language agents



- Currently, the field is transitioning from *prompting* to *behavior cloning / supervised fine-tuning*
- Behavior cloning is unlikely to be sufficient for generalist agents; they need to explore the environments and *trial and error*
- Challenges from open action space, reward model, and safety

Future directions for language agents



- Agent safety research is far behind agent development and deployment
- Language agents inherit all the safety risks of LLMs (e.g., *bias, fairness, hallucination, privacy, transparency*), amplify some of them (e.g., *workforce displacement*), and bring many new ones (e.g., *irreversible actions*)

Language Agents: Foundations, Prospects, and Risks

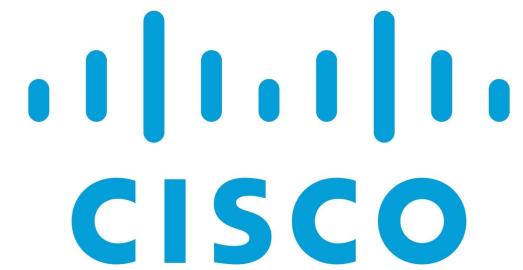
Yu Su, Diyi Yang, Shunyu Yao, Tao Yu

<https://language-agent-tutorial.github.io/>



香港大學
THE UNIVERSITY OF HONG KONG

Acknowledgements



Thanks &

