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.

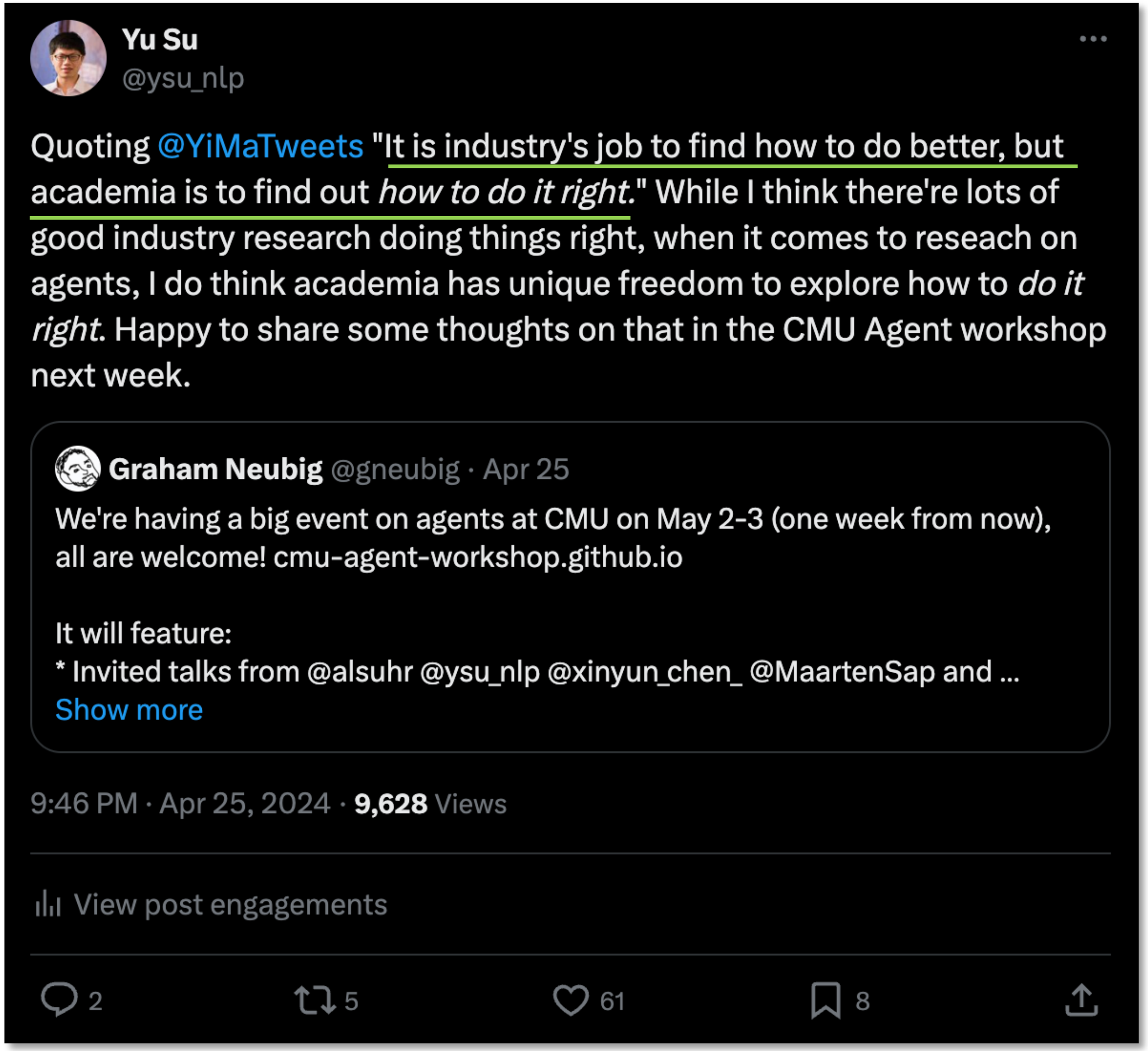
Sam Altman

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

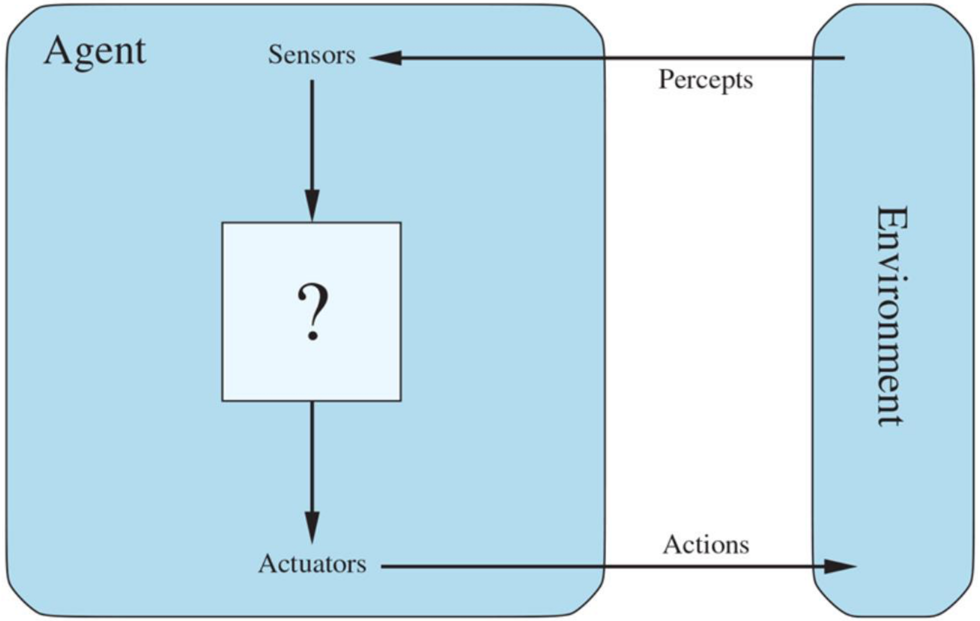
Quoting @YiMaTweets "It is industry's job to find how to do better, but academia is to find out *how to do it right.*" While I think there're lots of good industry research doing things right, when it comes to reseach on agents, I do think academia has unique freedom to explore how to *do it right*. 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

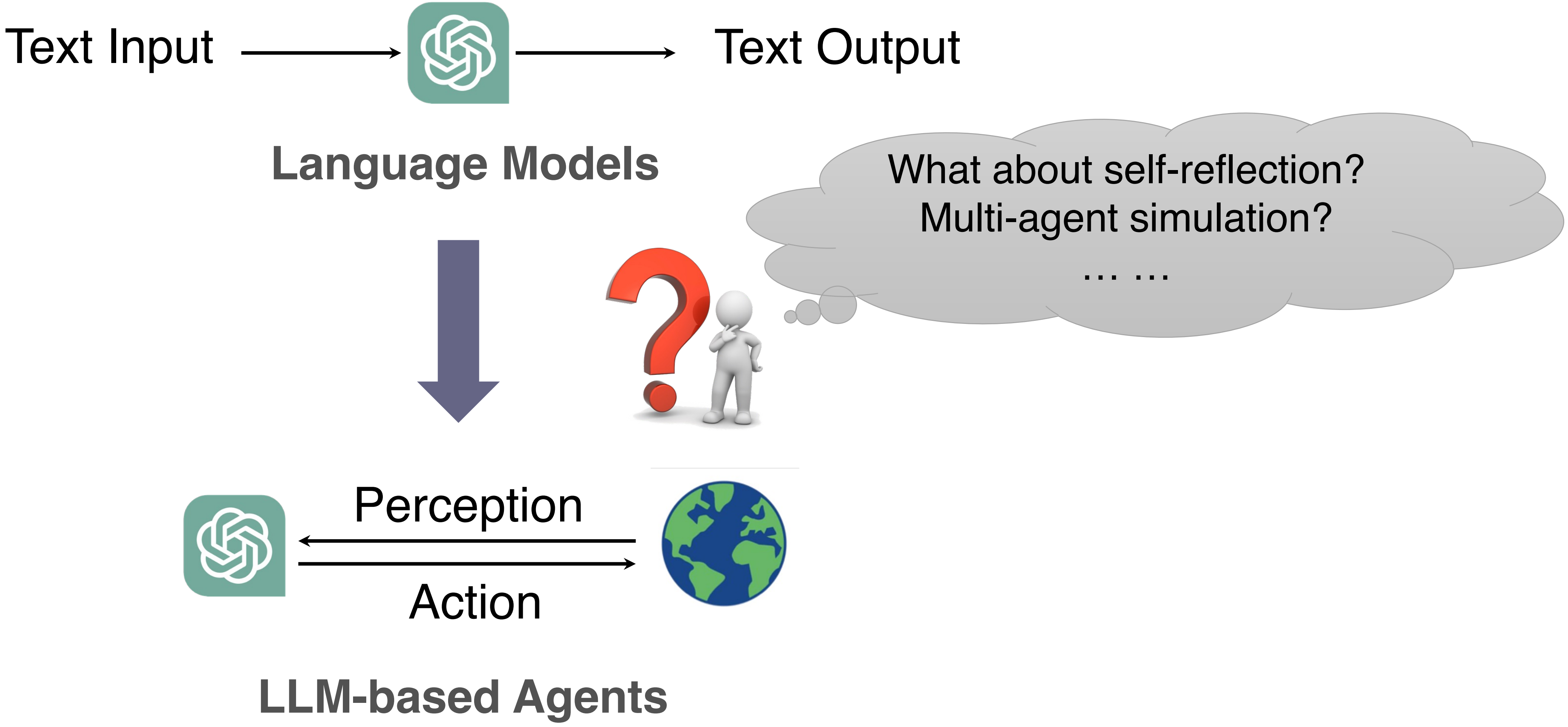
Why agents again?



“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?



Text Input → Text Output

Language Models

What about self-reflection?
Multi-agent simulation?
… …

Perception

Action

LLM-based Agents

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

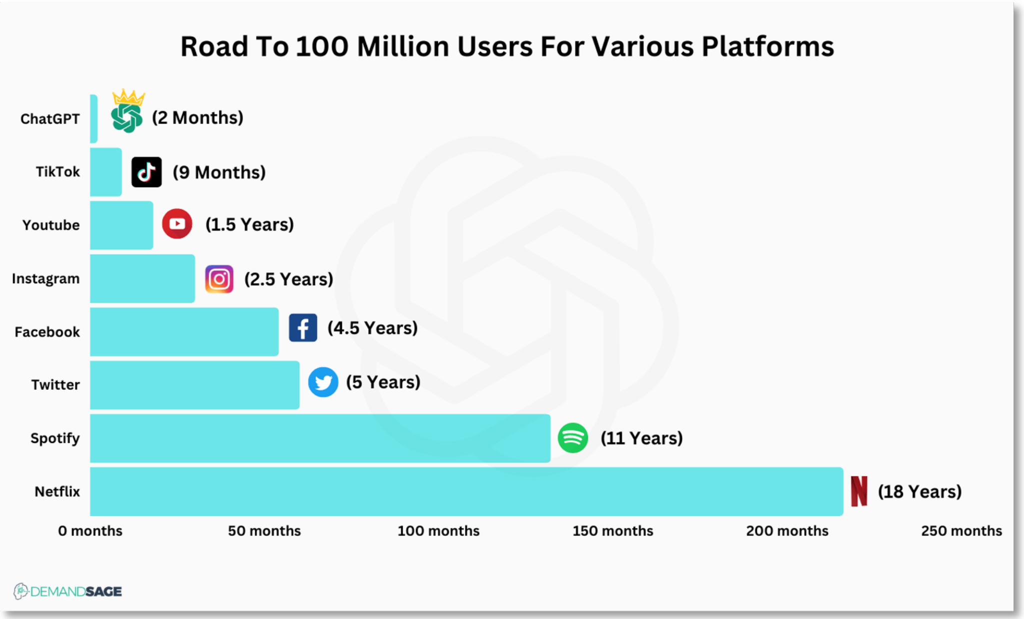
- Implications: All the same challenges faced by previous AI agents (e.g., *perception*, *reasoning*, *world models*, *planning*) still remain, but we need to re-examine them through the new lens of LLMs and tackle the new challenges and opportunities (e.g., *synthetic data*, *self-reflection*, *internalized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

What’s fundamentally different now?

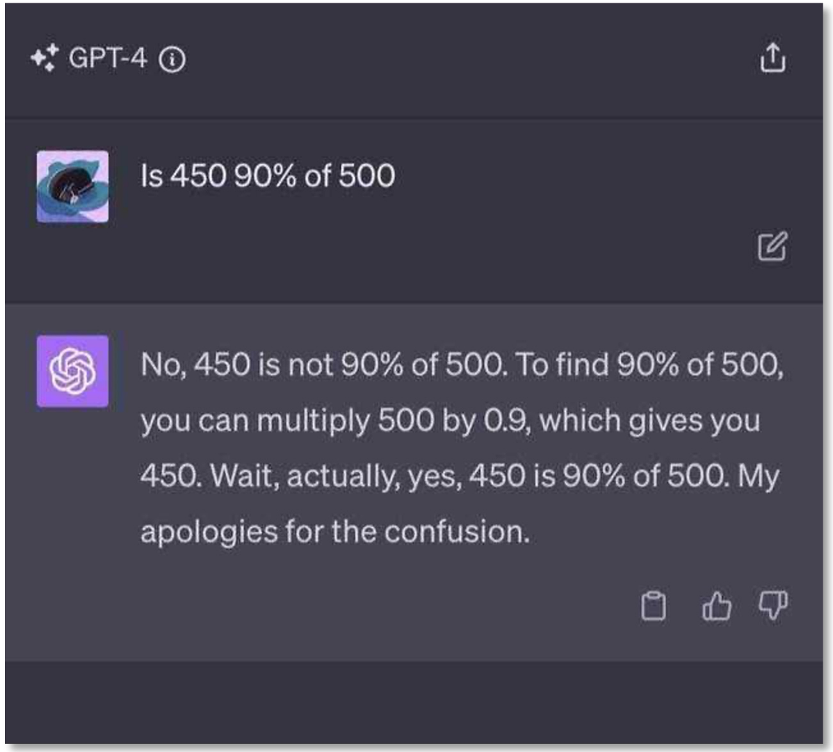
Contemporary AI agents, with integrated LLM(s), can *use language as a vehicle for reasoning and communication*

⬆ Instruction following, in-context learning, output customization

⬆ Reasoning (for better acting): state inferences, self-reflection, replanning, etc.



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

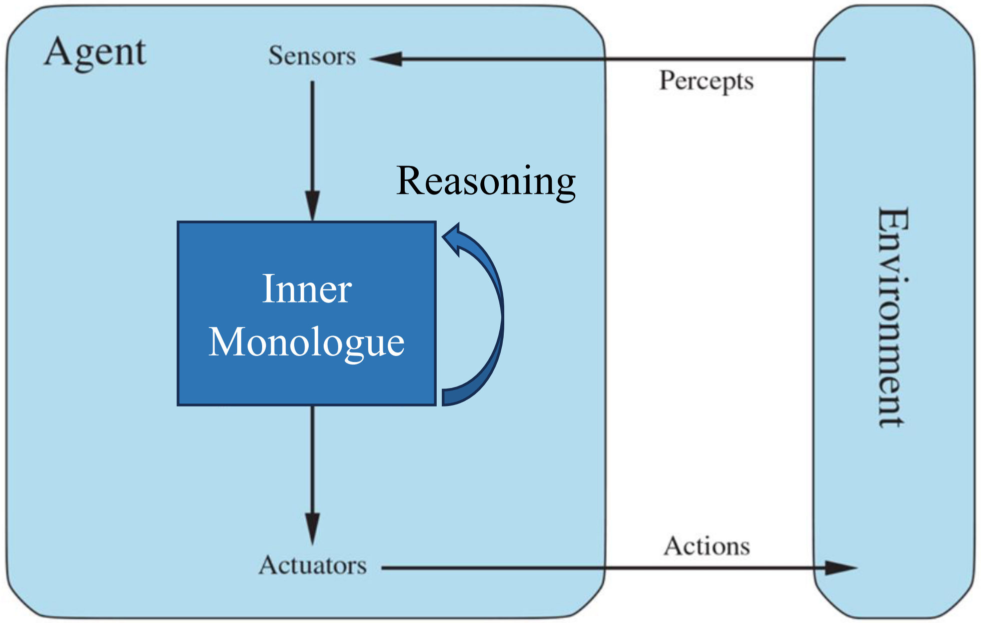


Is 450 90% of 500

No, 450 is not 90% of 500. To find 90% of 500, you can multiply 500 by 0.9, which gives you 450. Wait, actually, yes, 450 is 90% of 500. My apologies for the confusion.

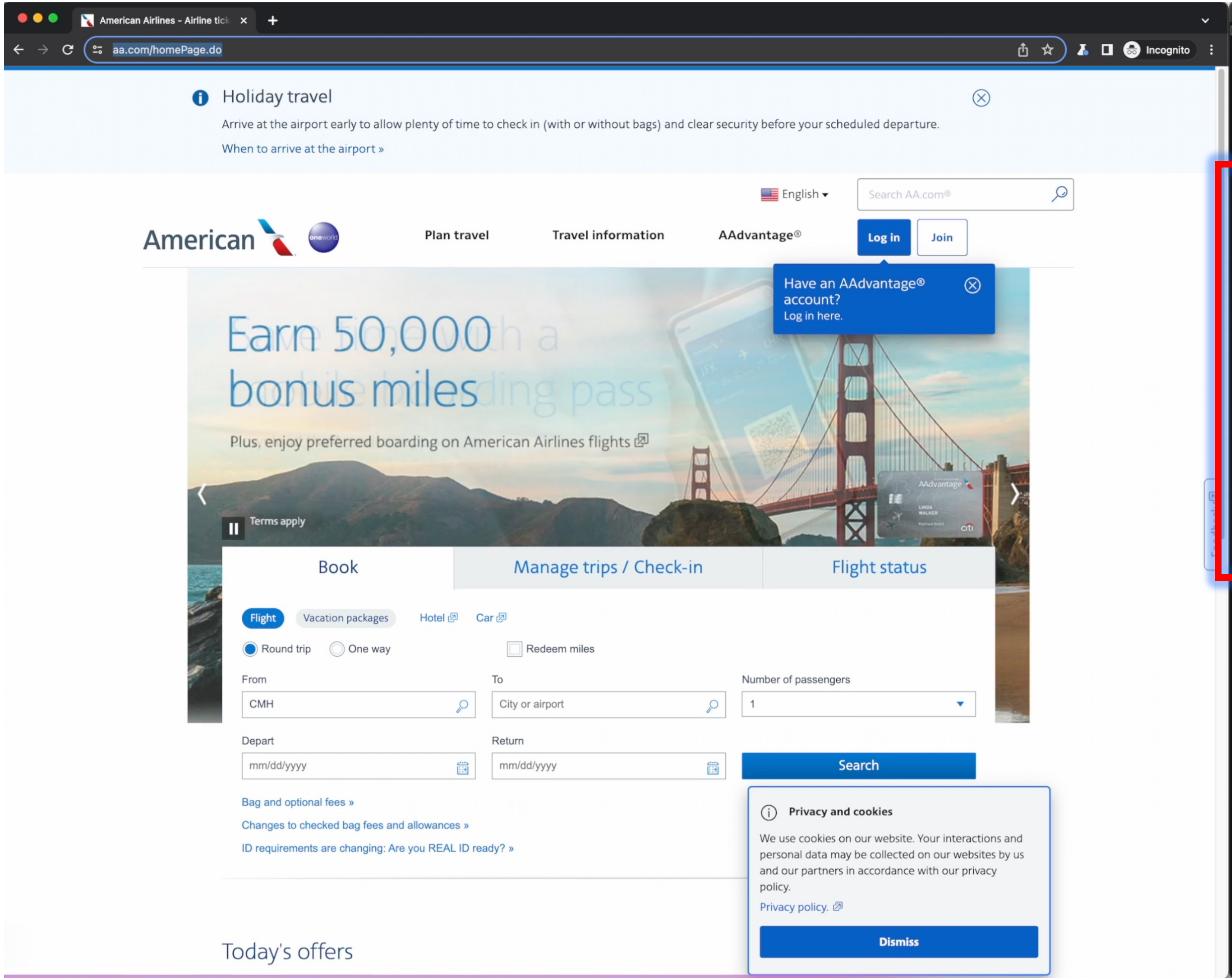
https://www.reddit.com/r/ChatGPT/comments/16jvl4x/wait_actually_yes/

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 tick

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 »

Have an AAdvantage® account? Log in here.

Earn 50,000 bonus miles

Plus, enjoy preferred boarding on American Airlines flights

From: CMH

To: City or airport

Number of passengers: 1

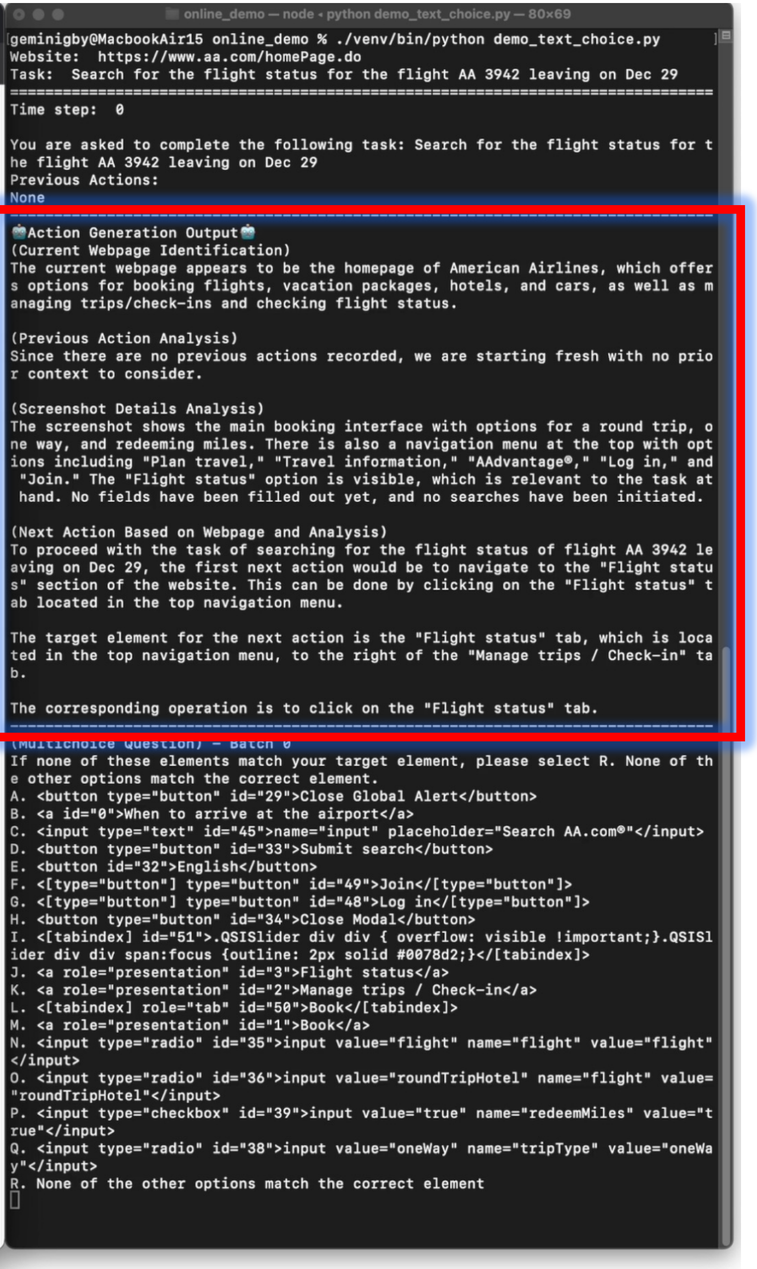
Depart: mm/dd/yyyy

Return: mm/dd/yyyy

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.

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



online_demo — node ◂ python demo_text_choice.py — 80×69

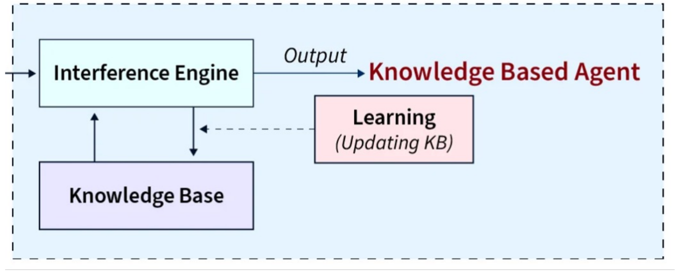
```
geminigby@MacbookAir15 online_demo % ./venv/bin/python demo_text_choice.py
Website:  https://www.aa.com/homePage.do
Task:  Search for the flight status for the flight AA 3942 leaving on Dec 29
================================================================================
Time step:  0

You are asked to complete the following task: Search for the flight status for t
he 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 offer
s options for booking flights, vacation packages, hotels, and cars, as well as m
anaging trips/check-ins and checking flight status.

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

(Screenshot Details Analysis)
The screenshot shows the main booking interface with options for a round trip, o
ne way, and redeeming miles. There is also a navigation menu at the top with opt
ions 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 le
aving on Dec 29, the first next action would be to navigate to the "Flight statu
s" section of the website. This can be done by clicking on the "Flight status" t
ab located in the top navigation menu.

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

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 th
e 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" placeholder="Search AA.com®"</input>
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;}.QSISl
ider div 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. <a role="presentation" id="1">Book</a>
N. <input type="radio" id="35">input value="flight" name="flight" value="flight"
</input>
O. <input type="radio" id="36">input value="roundTripHotel" name="flight" value=
"roundTripHotel"</input>
P. <input type="checkbox" id="39">input value="true" name="redeemMiles" value="t
rue"</input>
Q. <input type="radio" id="38">input value="oneWay" name="tripType" value="oneWa
y"</input>
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



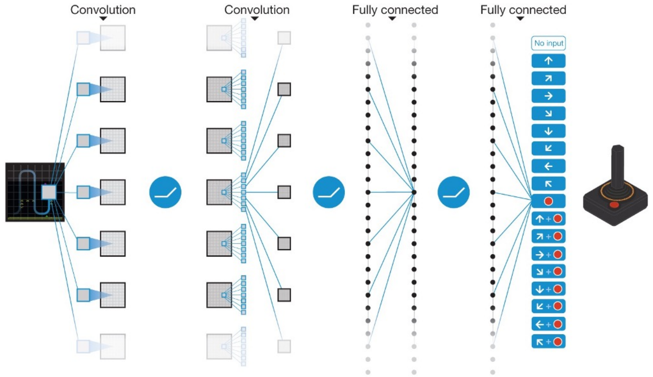
Interference Engine

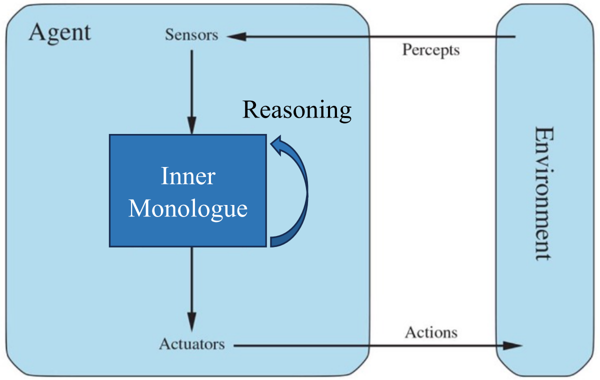
Output

Knowledge Based Agent

Learning
(Updating KB)

Knowledge Base





Agent

Sensors

Percepts

Reasoning

Inner Monologue

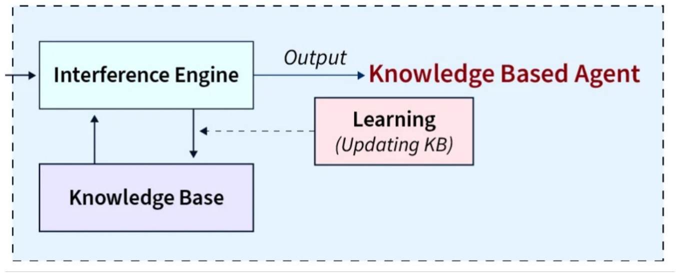
Actuators

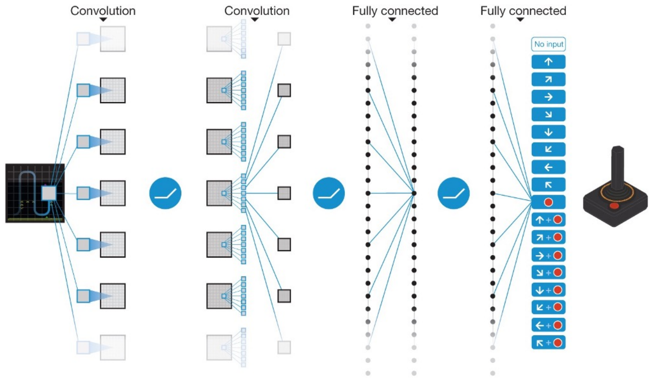
Actions

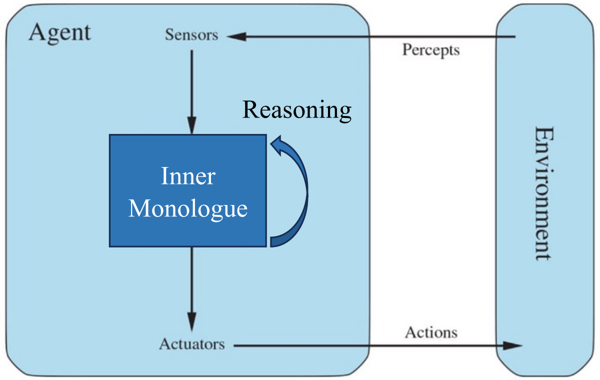
Environment

	Logical Agent	**Neural Agent**	**Language Agent**
Expressiveness			
Reasoning			
Adaptivity			

Evolution of AI agents



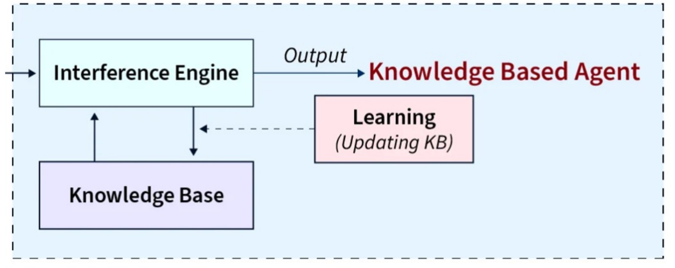


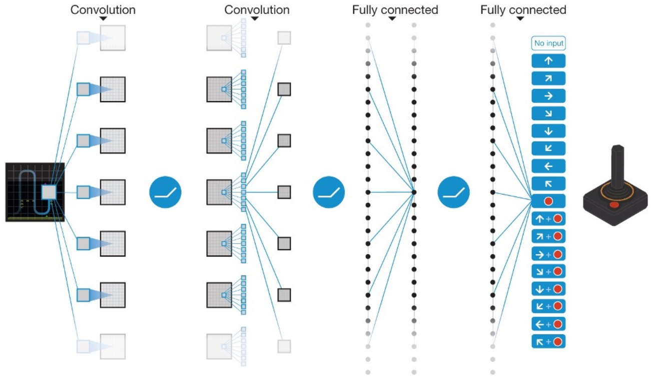


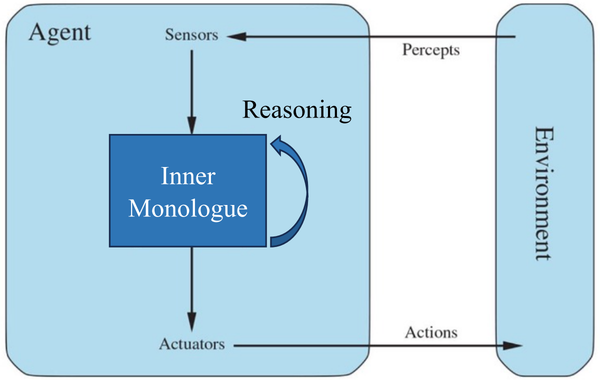
Logical Agent **Neural Agent** **Language Agent**

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

Evolution of AI agents

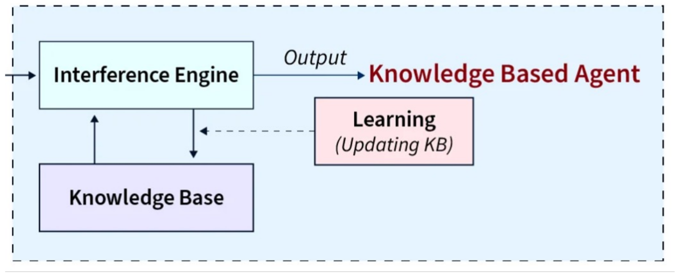


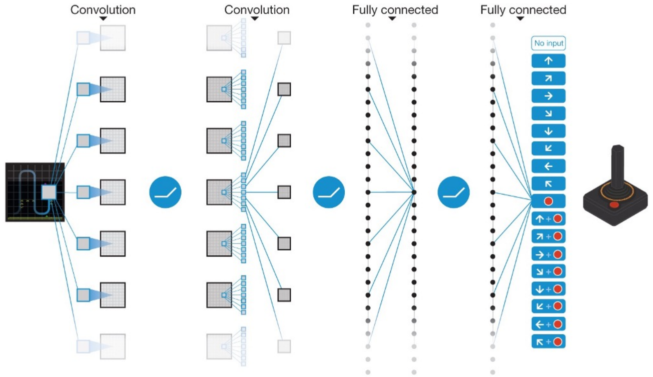


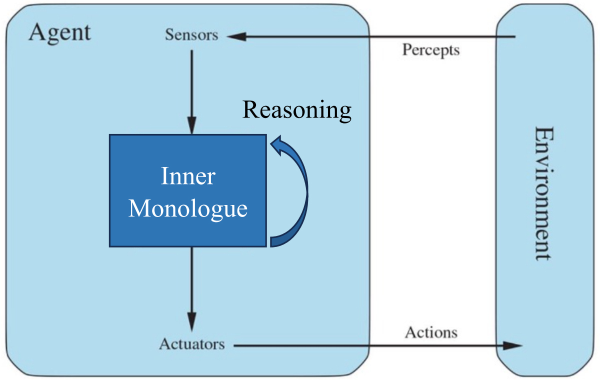


	Logical Agent	Neural Agent	Language Agent
Expressiveness	Low bounded by the logical language	Medium anything a (small-ish) NN can encode	
Reasoning	Logical inferences sound, explicit, rigid	Parametric inferences stochastic, implicit, rigid	
Adaptivity	Low bounded by knowledge curation	Medium data-driven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

Evolution of AI agents

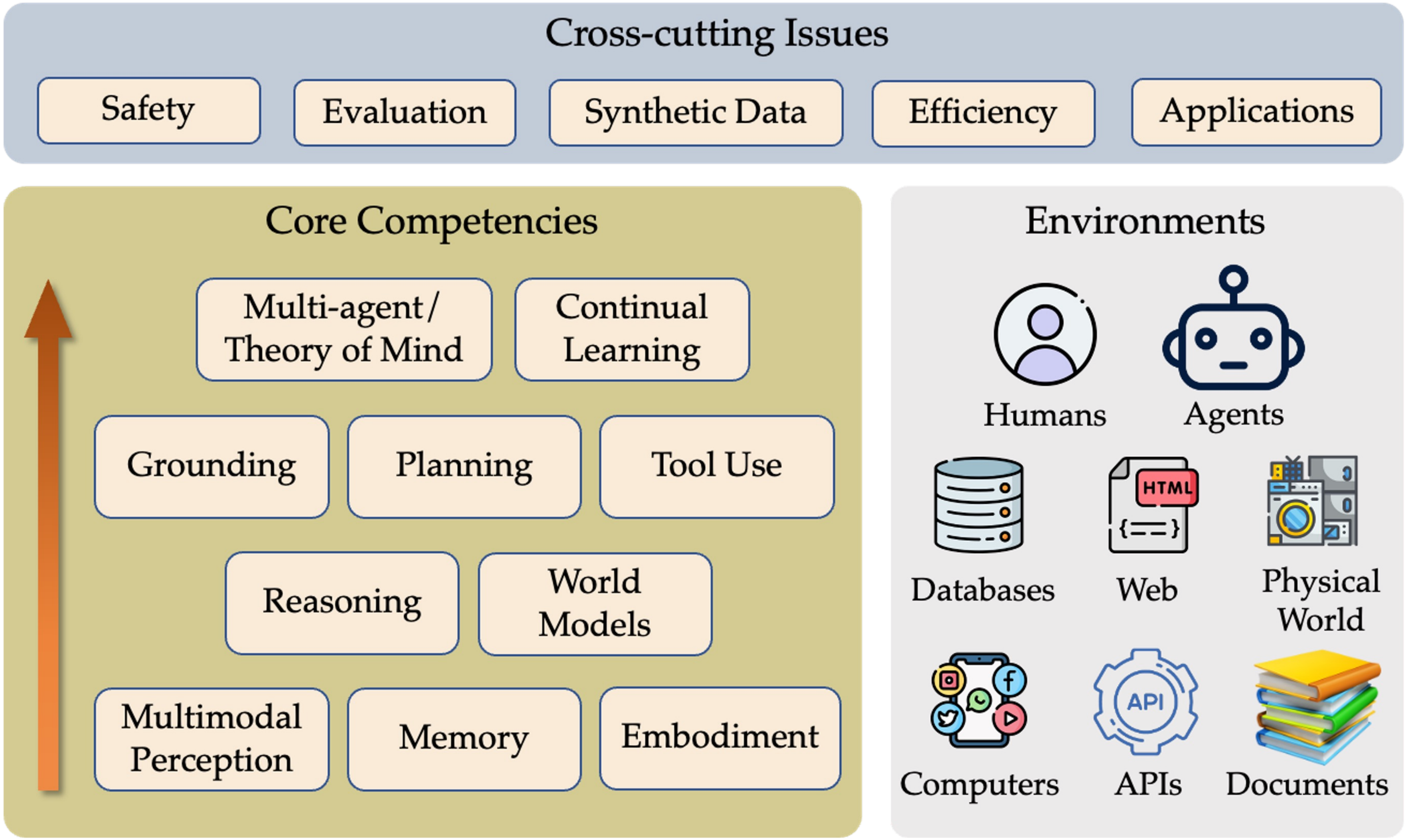






	Logical Agent	**Neural Agent**	**Language Agent**
Expressiveness	Low bounded by the logical language	Medium anything a (small-ish) NN can encode	High almost anything, esp. those verbalizable
Reasoning	Logical inferences sound, explicit, rigid	Parametric inferences stochastic, implicit, rigid	Language-based 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



Cross-cutting Issues

Safety

Evaluation

Synthetic Data

Efficiency

Applications

Core Competencies

Multi-agent/ Theory of Mind

Continual Learning

Grounding

Planning

Tool Use

Reasoning

World Models

Multimodal Perception

Memory

Embodiment

Environments

Humans

Agents

Databases

Web

Physical World

Computers

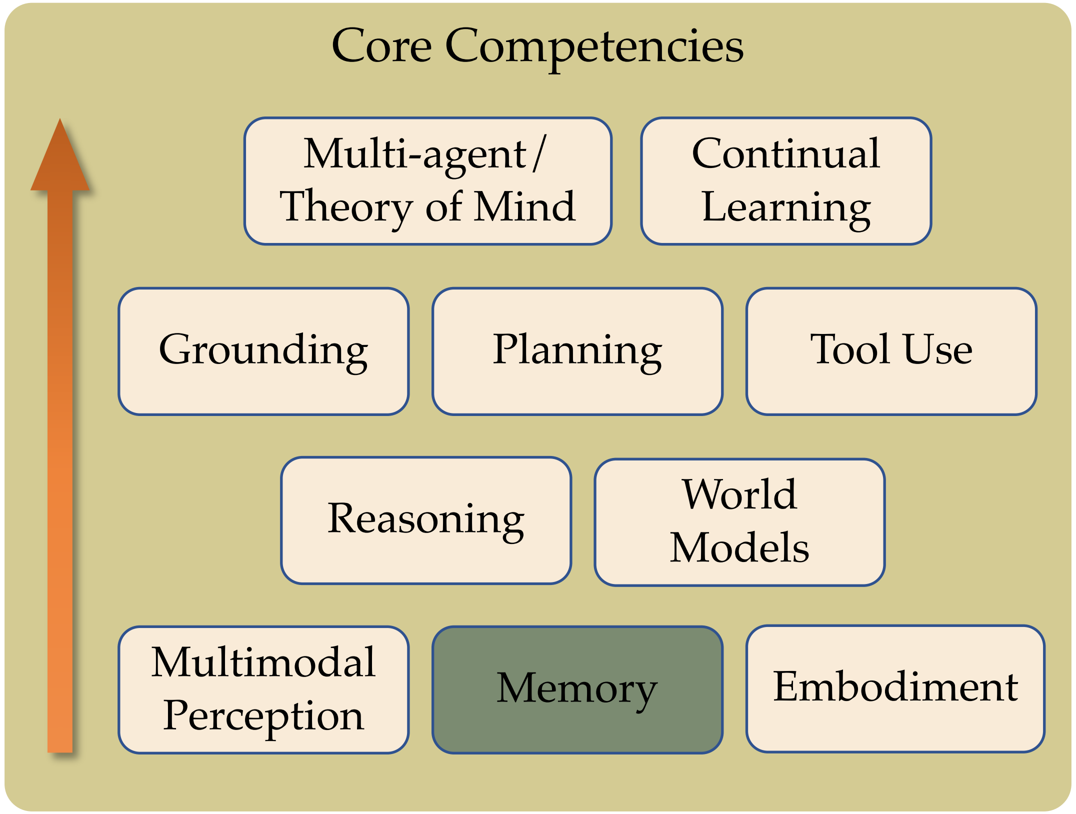
APIs

Documents

Outline

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

Memory


Core Competencies
Multi-agent / Theory of Mind
Continual Learning
Grounding
Planning
Tool Use
Reasoning
World Models
Multimodal Perception
Memory
Embodiment

NeurIPS 2024

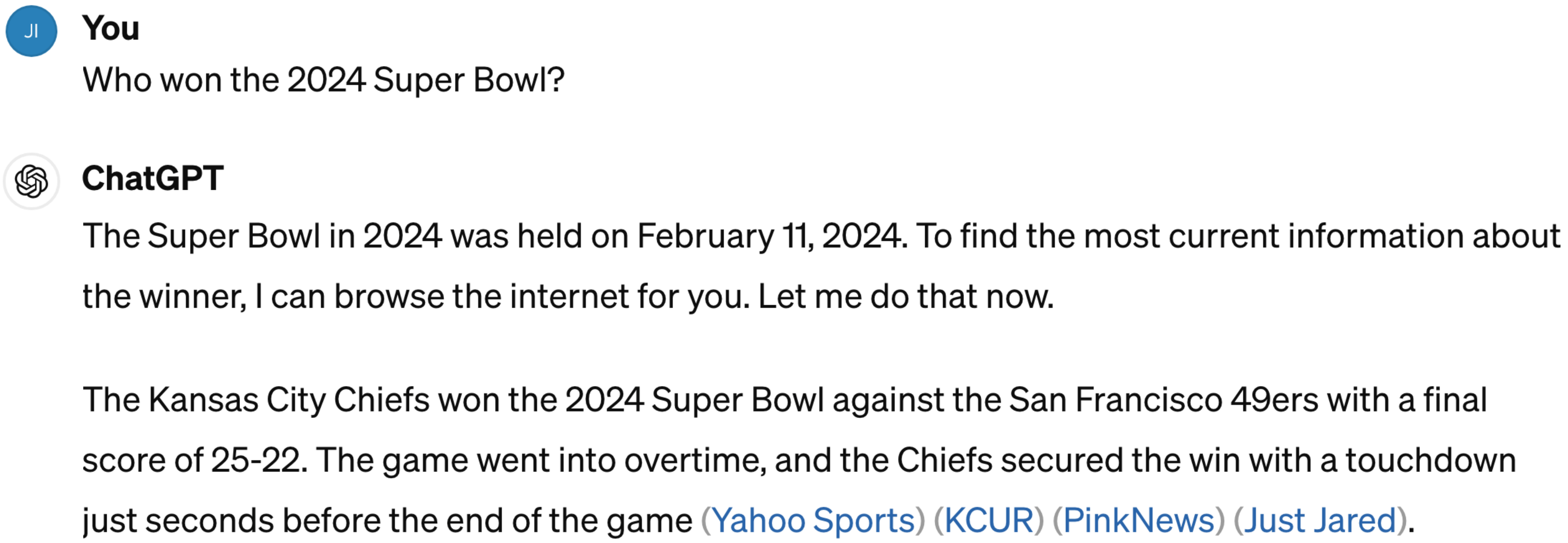
HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models

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

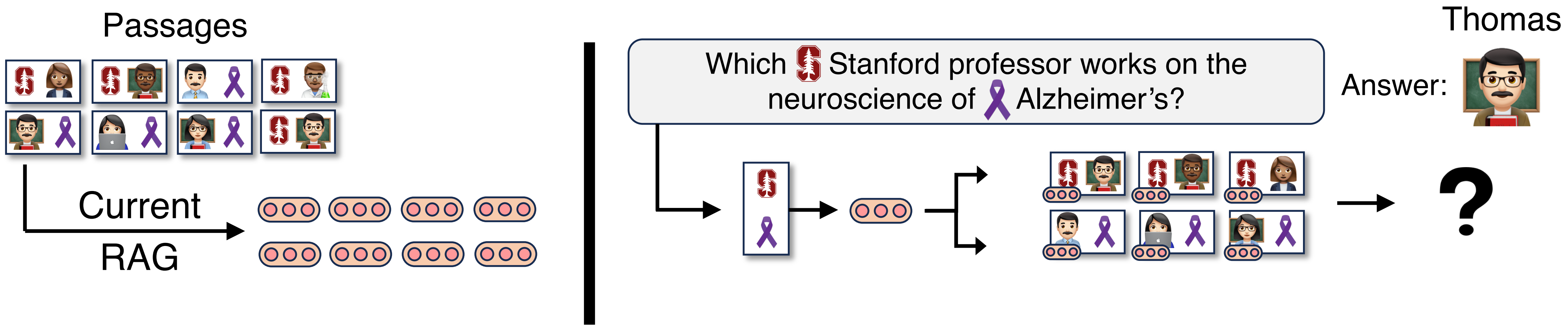
Passages

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

Thomas

Answer:

Current RAG does not always work



Passages

Current RAG

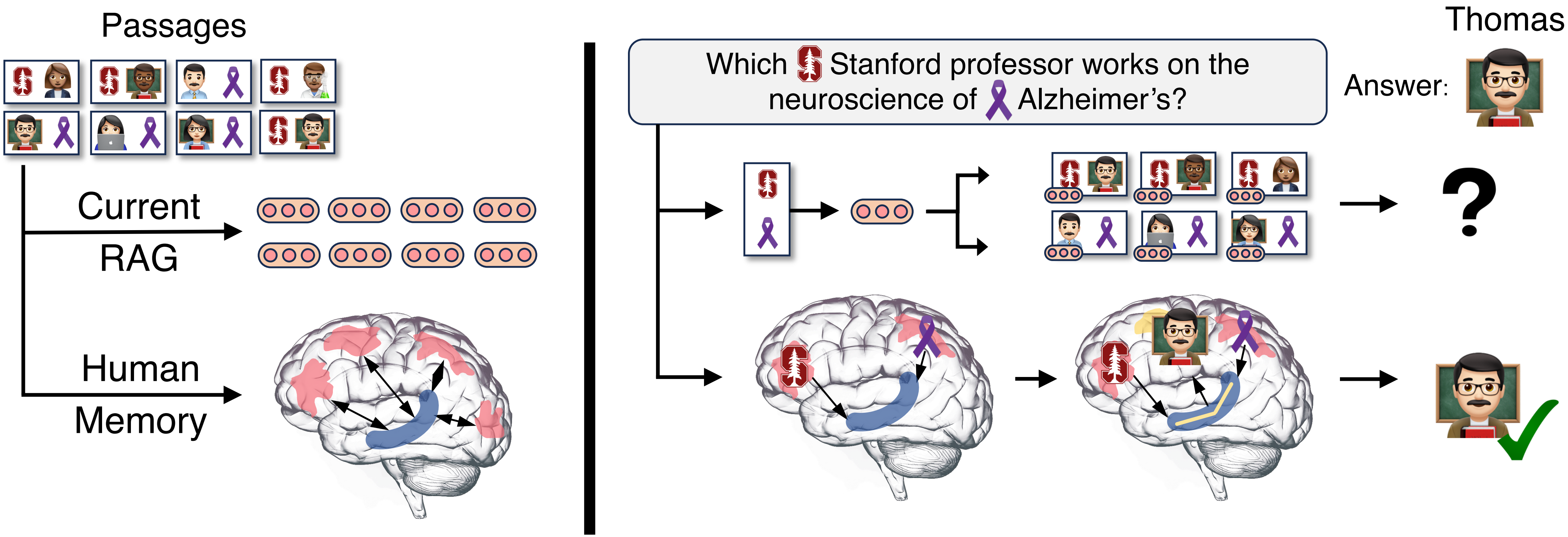
Which Stanford professor works on the neuroscience of Alzheimer's?

Thomas

Answer:

?

Current RAG does not always work



Passages

Current RAG

Human Memory

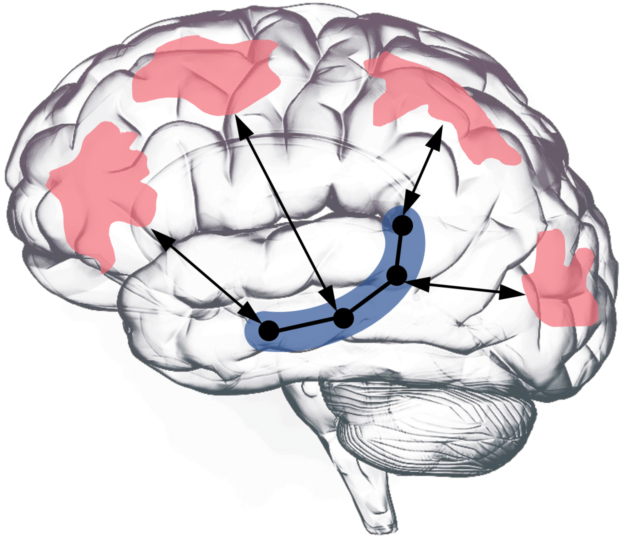
Which Stanford professor works on the neuroscience of Alzheimer's?

Thomas

Answer:

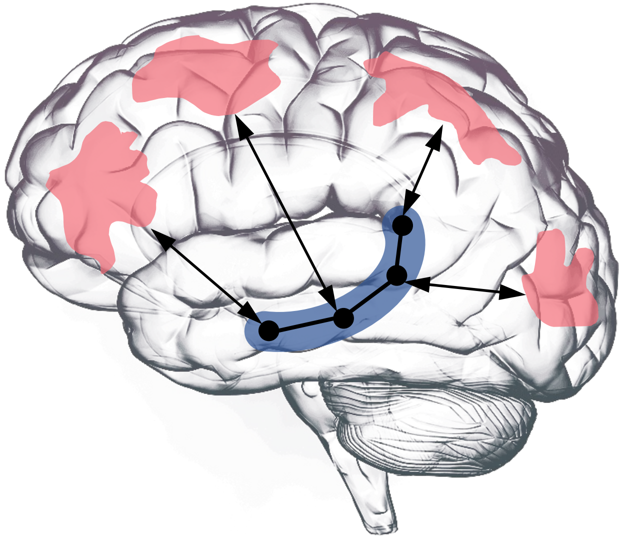
Long-term memory in humans

- 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.)

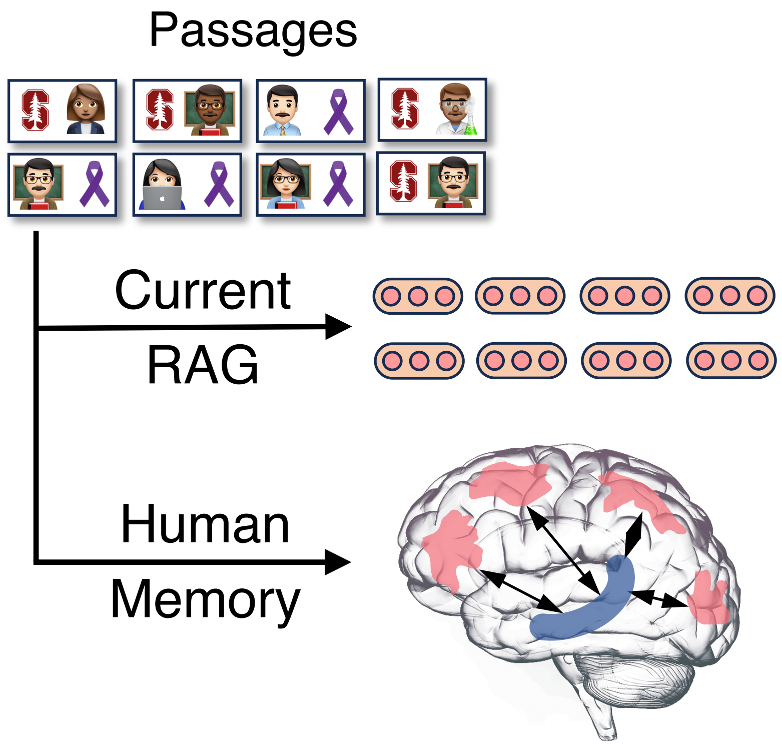


Long-term memory in humans

- 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)



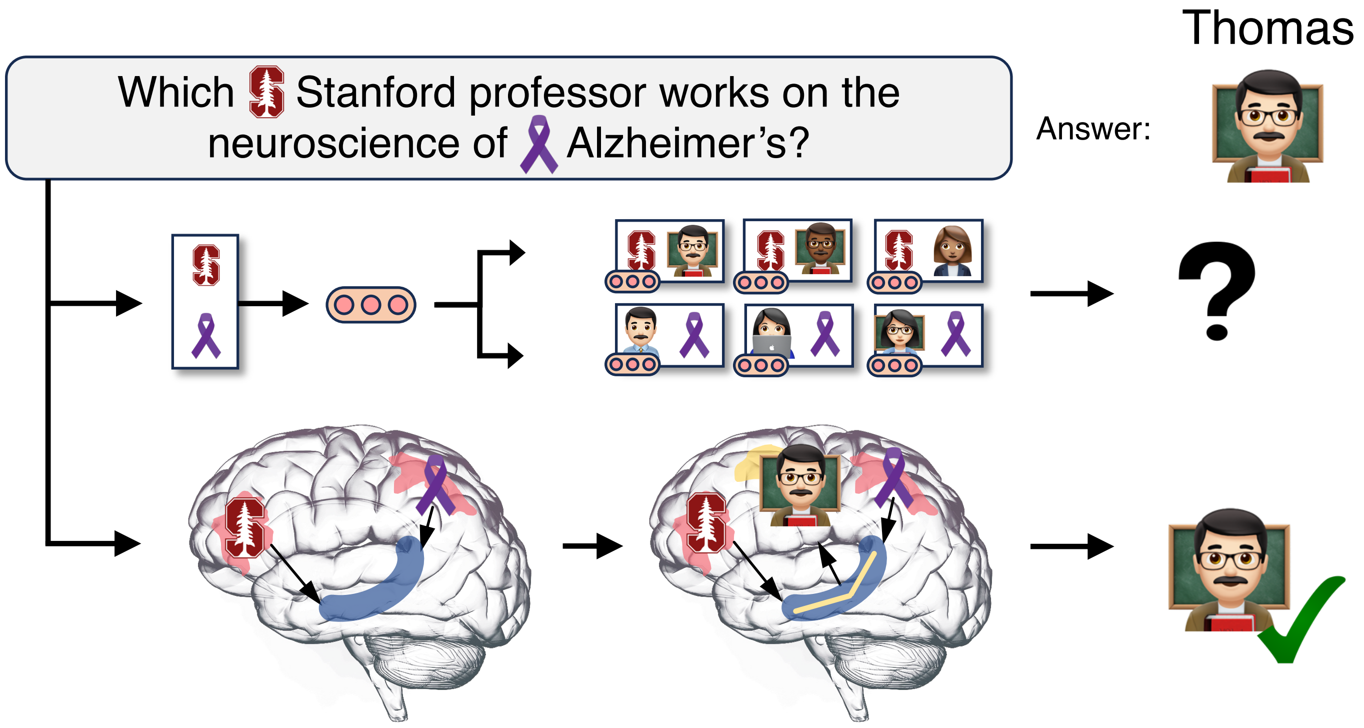
HippoRAG:
Neurobiologically-inspired LTM for LLMs



Passages

Current RAG

Human Memory

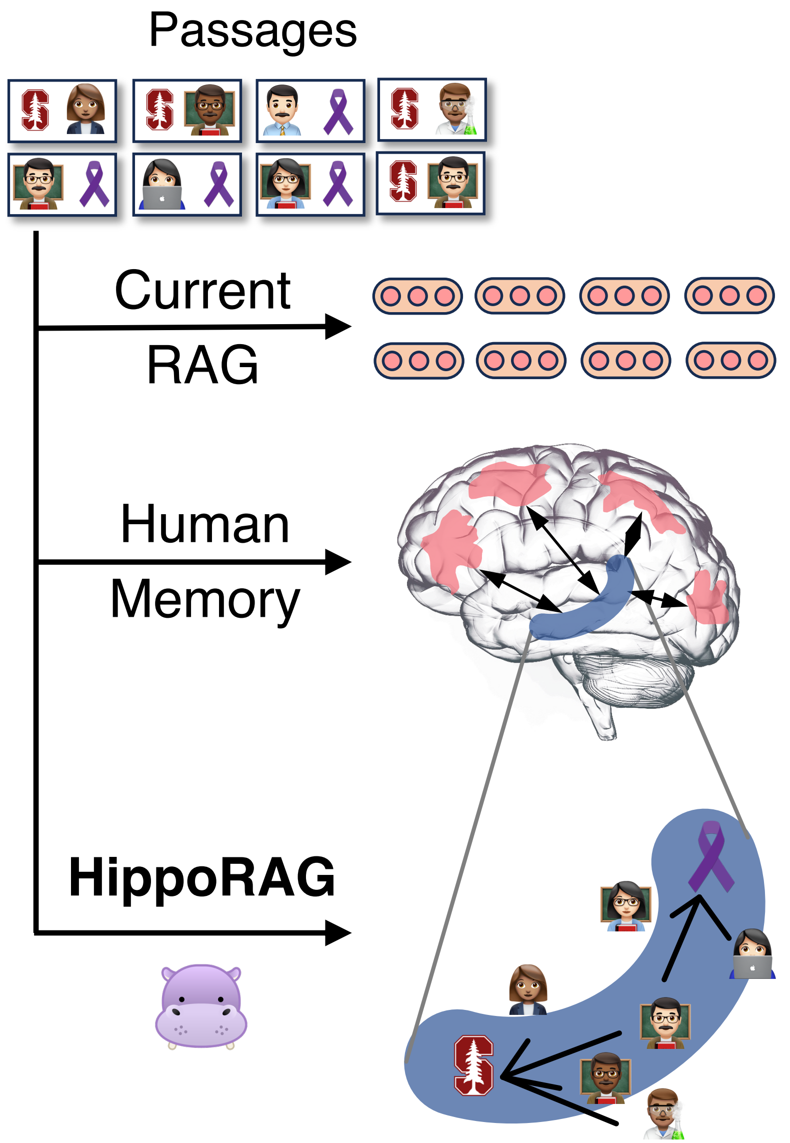


Which Stanford professor works on the neuroscience of Alzheimer's?

Thomas

Answer:

HippoRAG:
Neurobiologically-inspired LTM for LLMs

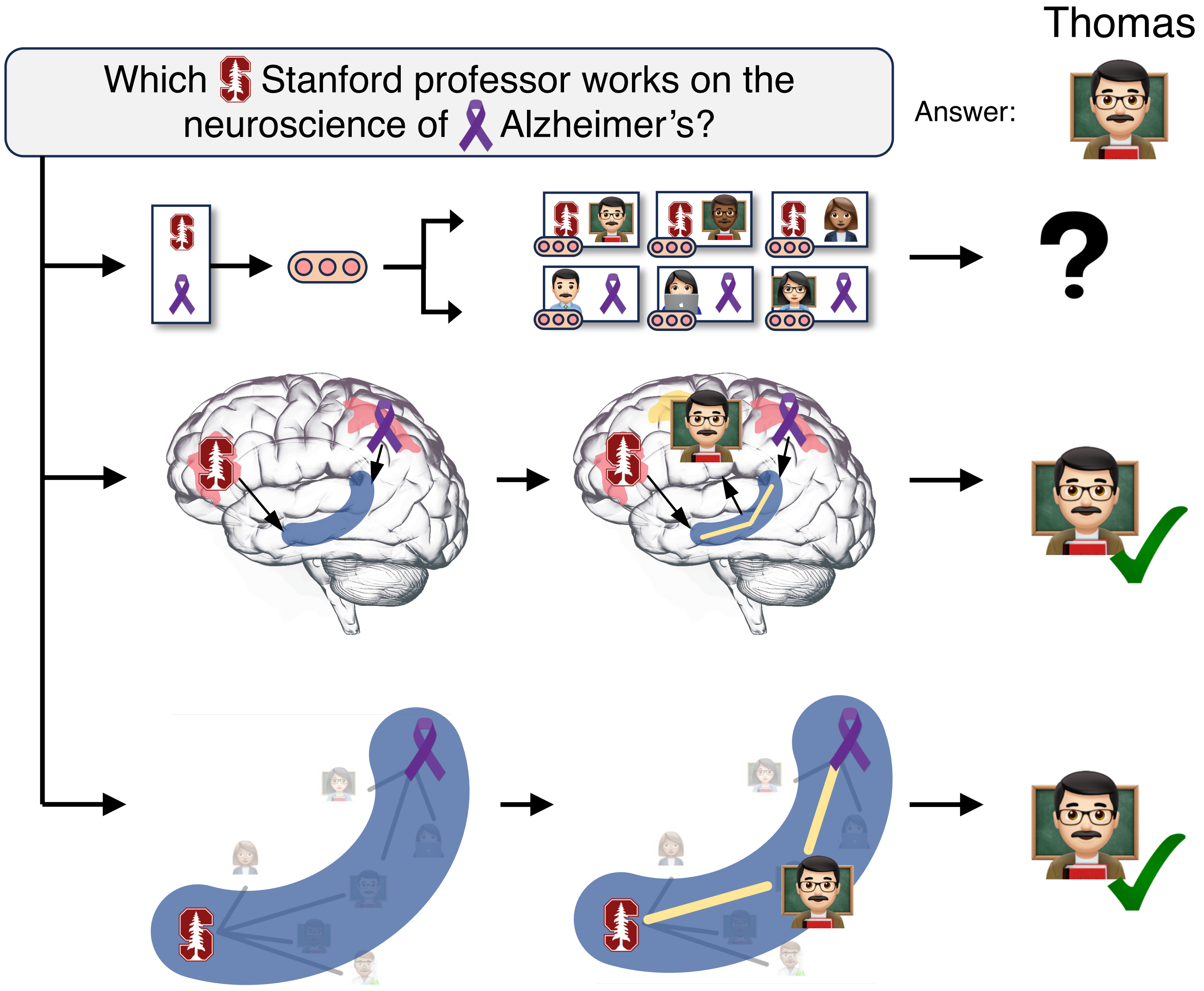


Passages

Current RAG

Human Memory

HippoRAG

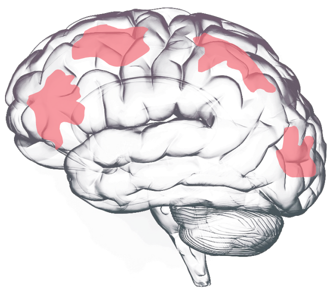


Which Stanford professor works on the neuroscience of Alzheimer's?

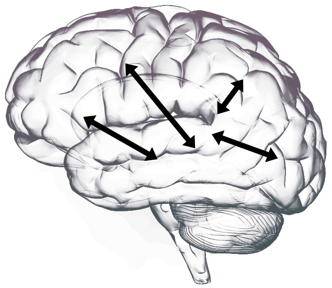
Answer:

Thomas

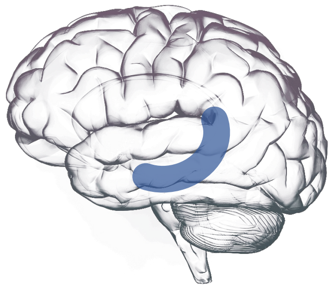
HippoRAG & the three components of LTM



Neocortex
Perception, linguistic abilities & reasoning

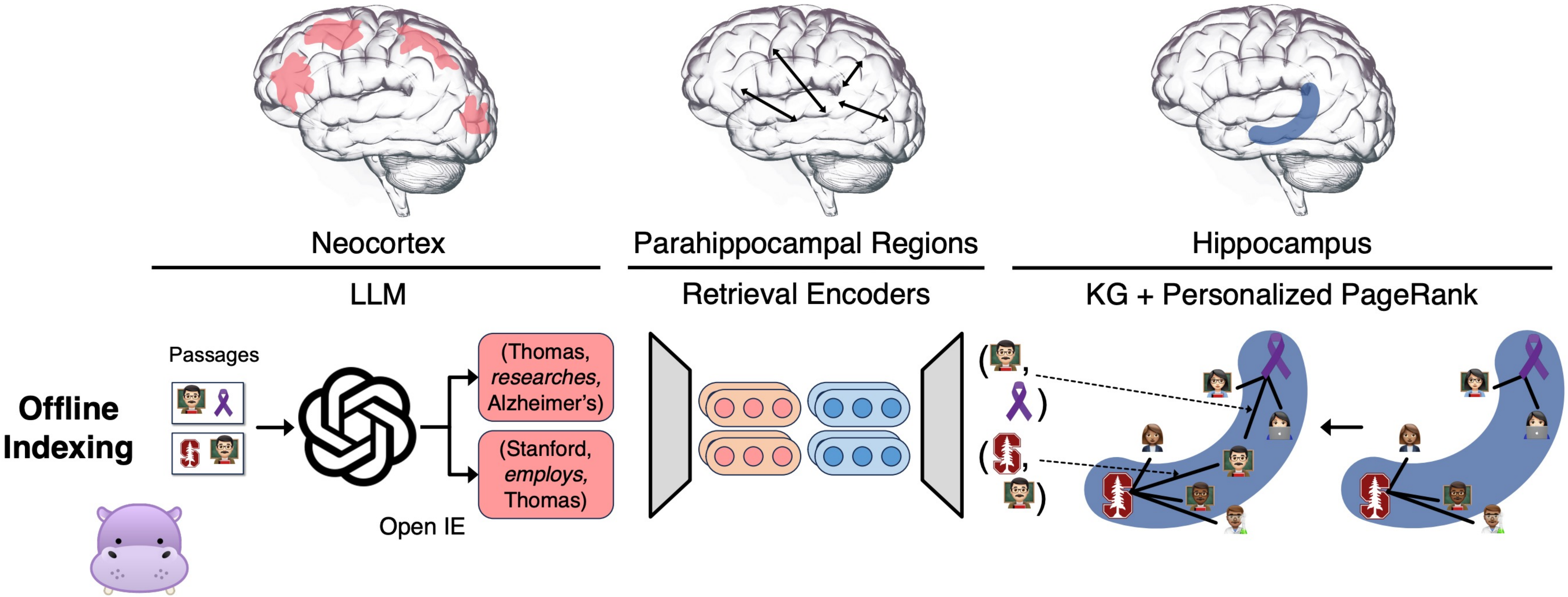


Parahippocampus
Bridge between areas; working memory



Hippocampus
Indexing & auto-associative memory

HippoRAG & the three components of LTM



Neocortex

LLM

Parahippocampal Regions

Retrieval Encoders

Hippocampus

KG + Personalized PageRank

Offline Indexing

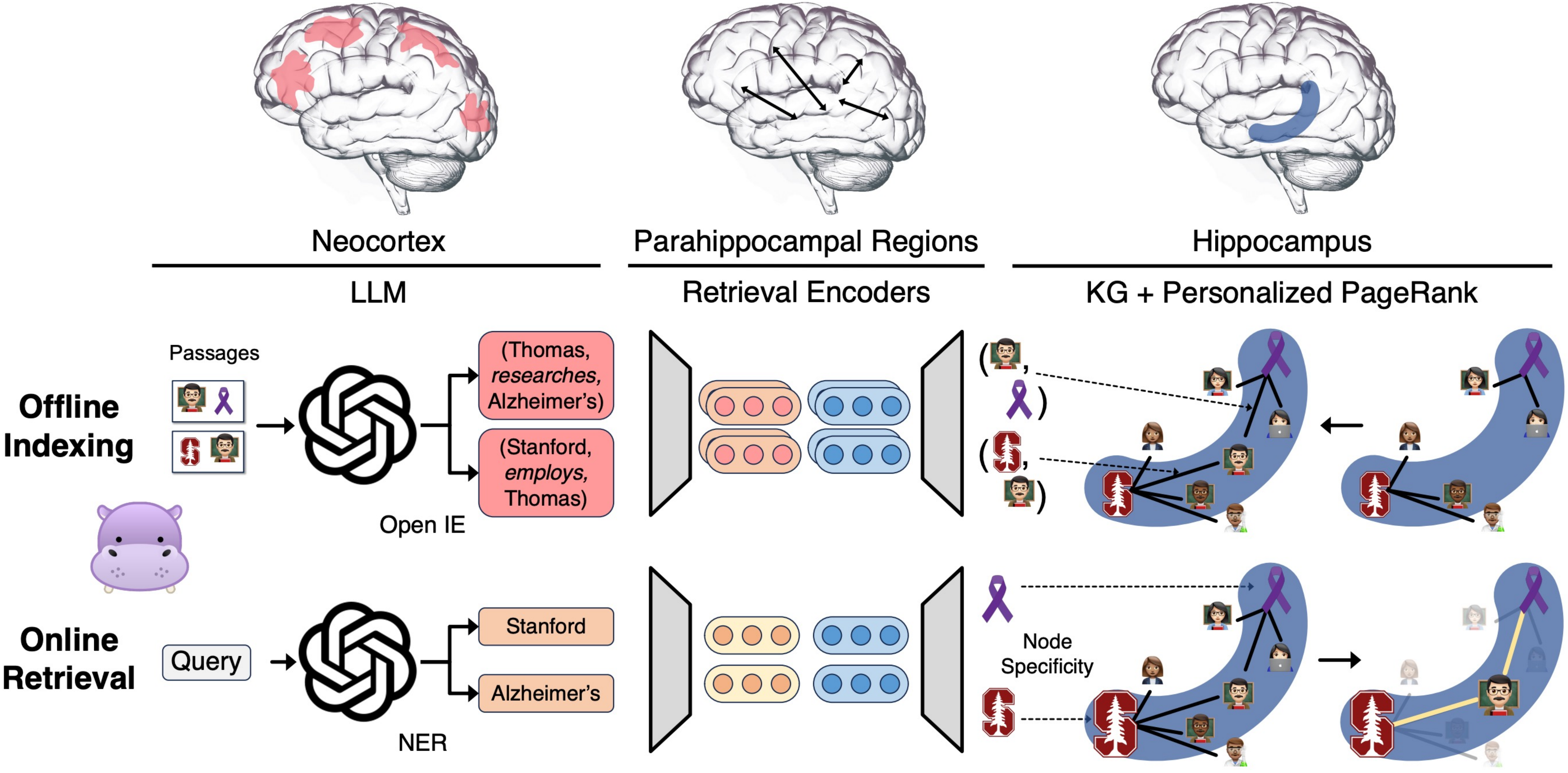
Passages

Open IE

(Thomas, *researches,* Alzheimer's)

(Stanford, *employs,* Thomas)

HippoRAG & the three components of LTM



Neocortex

LLM

Parahippocampal Regions

Retrieval Encoders

Hippocampus

KG + Personalized PageRank

Offline Indexing

Passages

(Thomas, *researches,* Alzheimer's)

(Stanford, *employs,* Thomas)

Open IE

Online Retrieval

Query

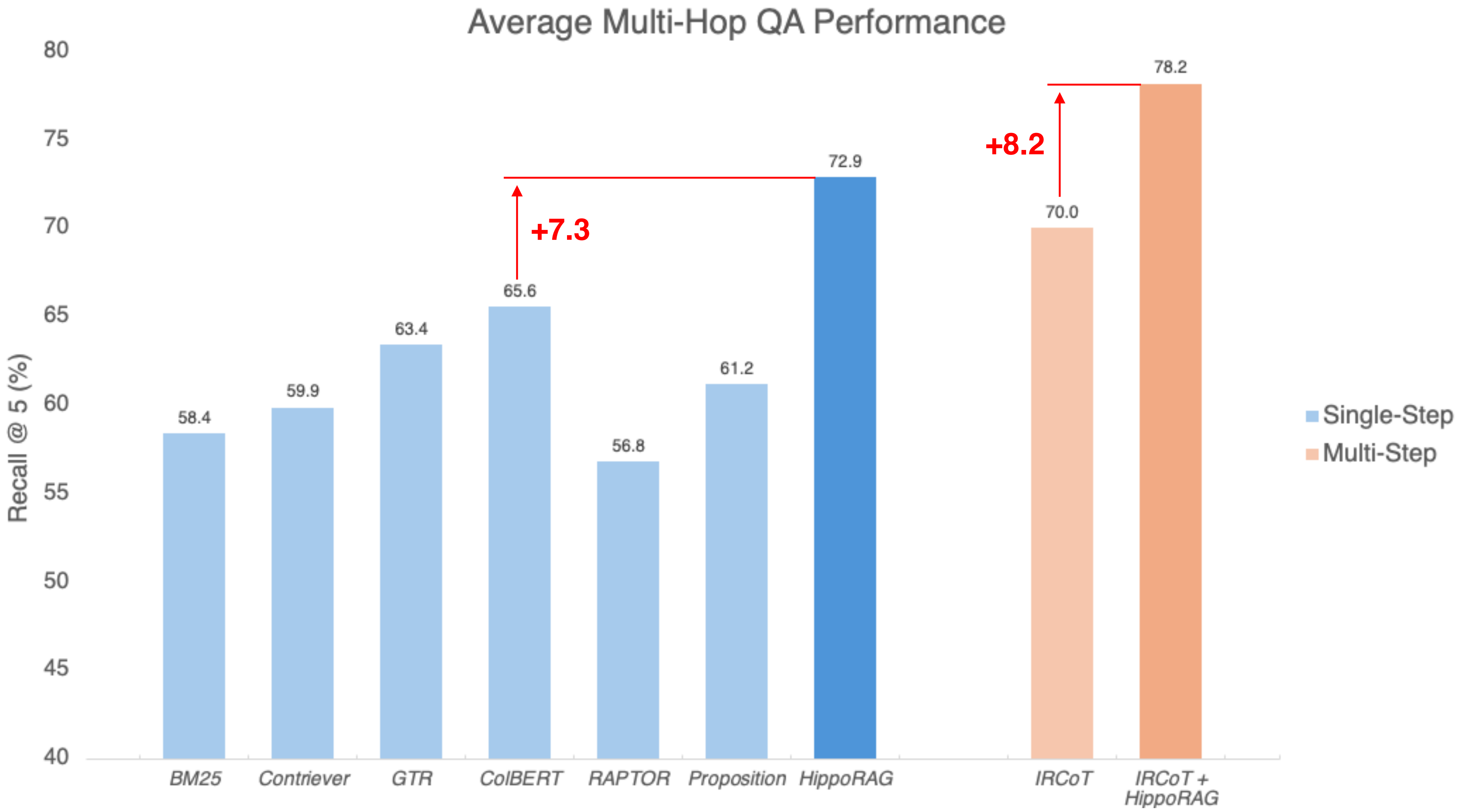
Stanford

Alzheimer's

NER

Node Specificity

HippoRAG is a SoTA memory retriever



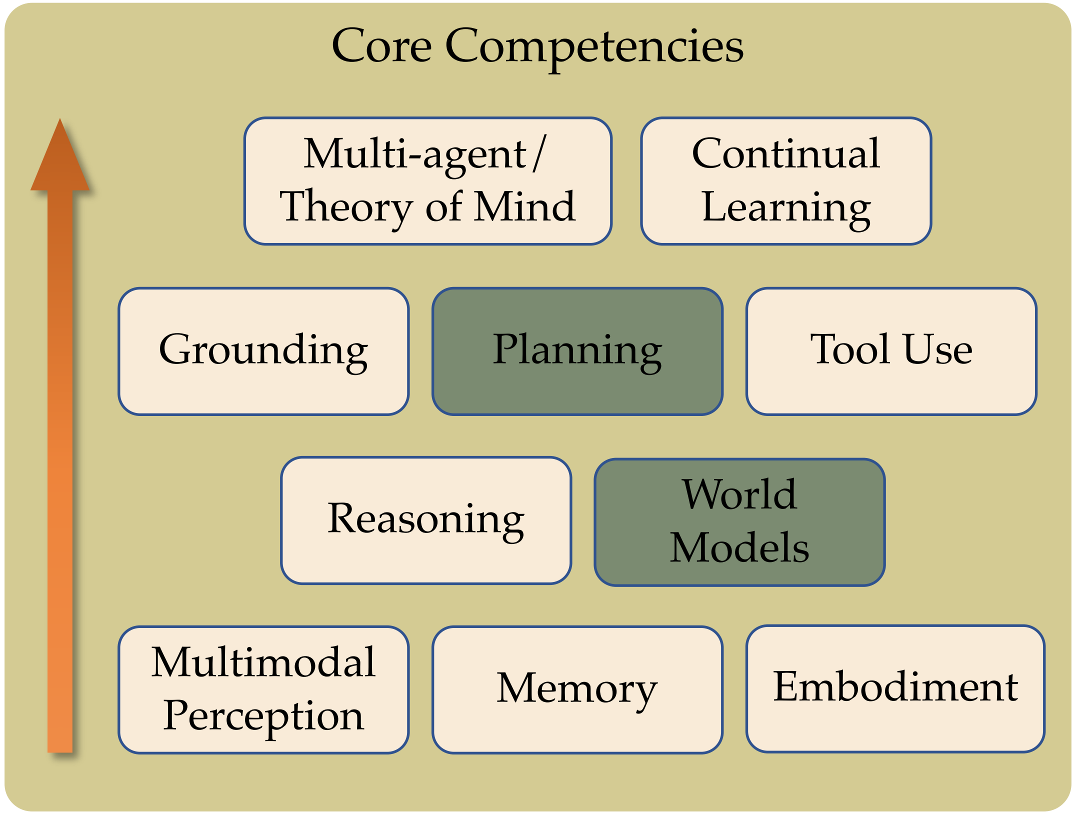
Practical aspects of HippoRAG

- 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



Core Competencies

Multi-agent / Theory of Mind

Continual Learning

Grounding

Planning

Tool Use

Reasoning

World Models

Multimodal Perception

Memory

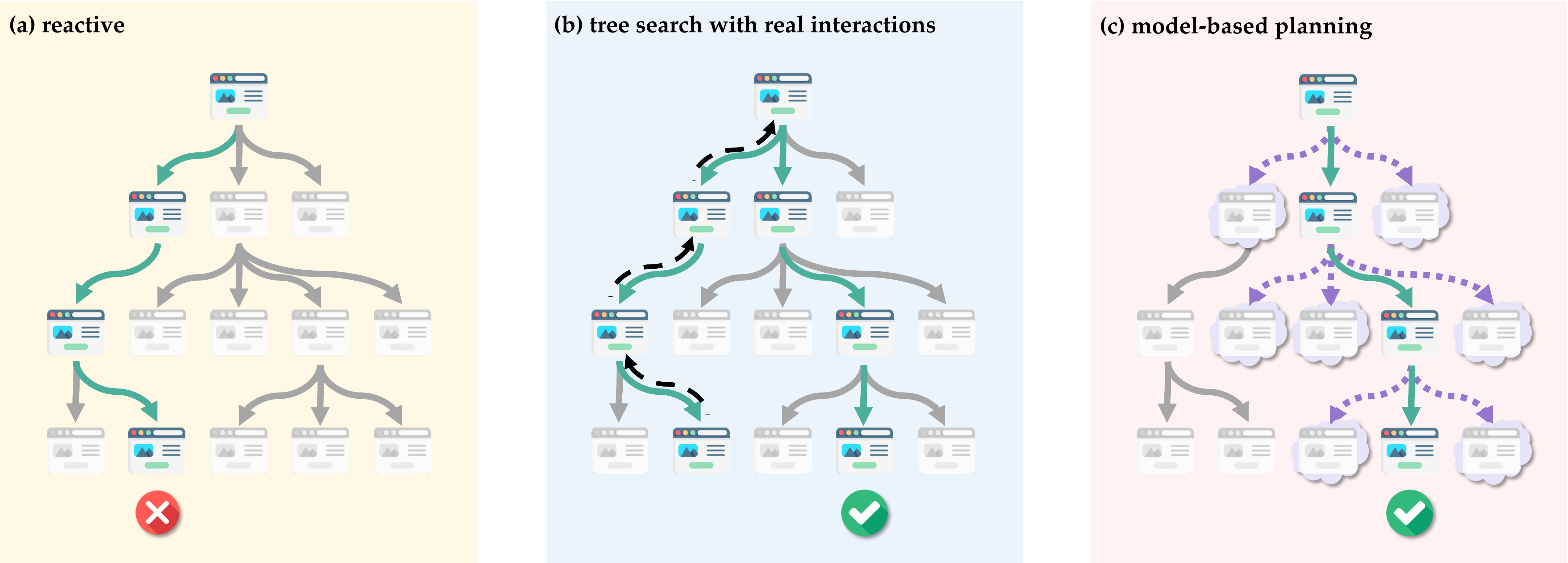
Embodiment

https://arxiv.org/abs/2411.06559

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



🙂 fast, easy to implement

😐 greedy, short-sighted

🙂 systematic exploration

😐 irreversible actions,
unsafe, slow

🙂 faster, safer,
systematic exploration

😐 how to get a world model?

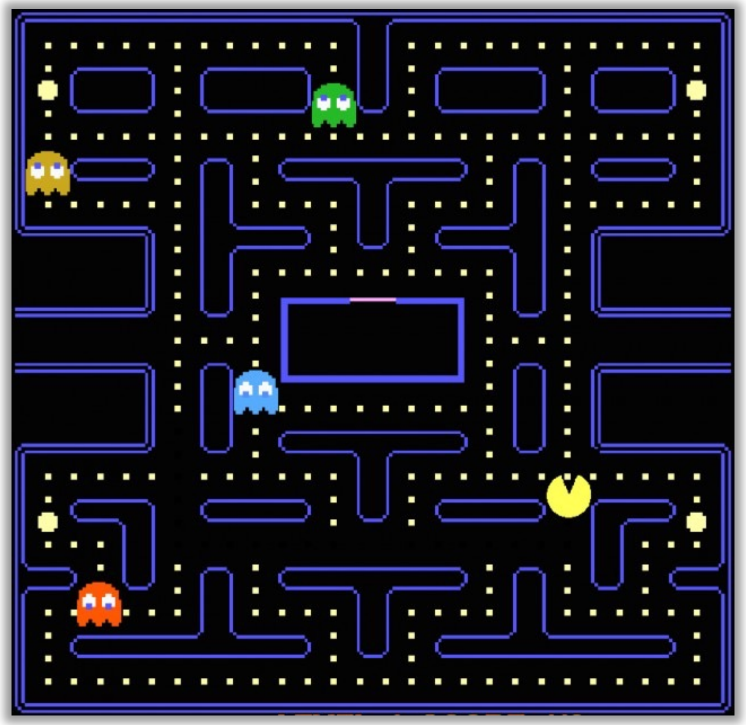
What's … a world model?

A computational model of environment transition dynamics

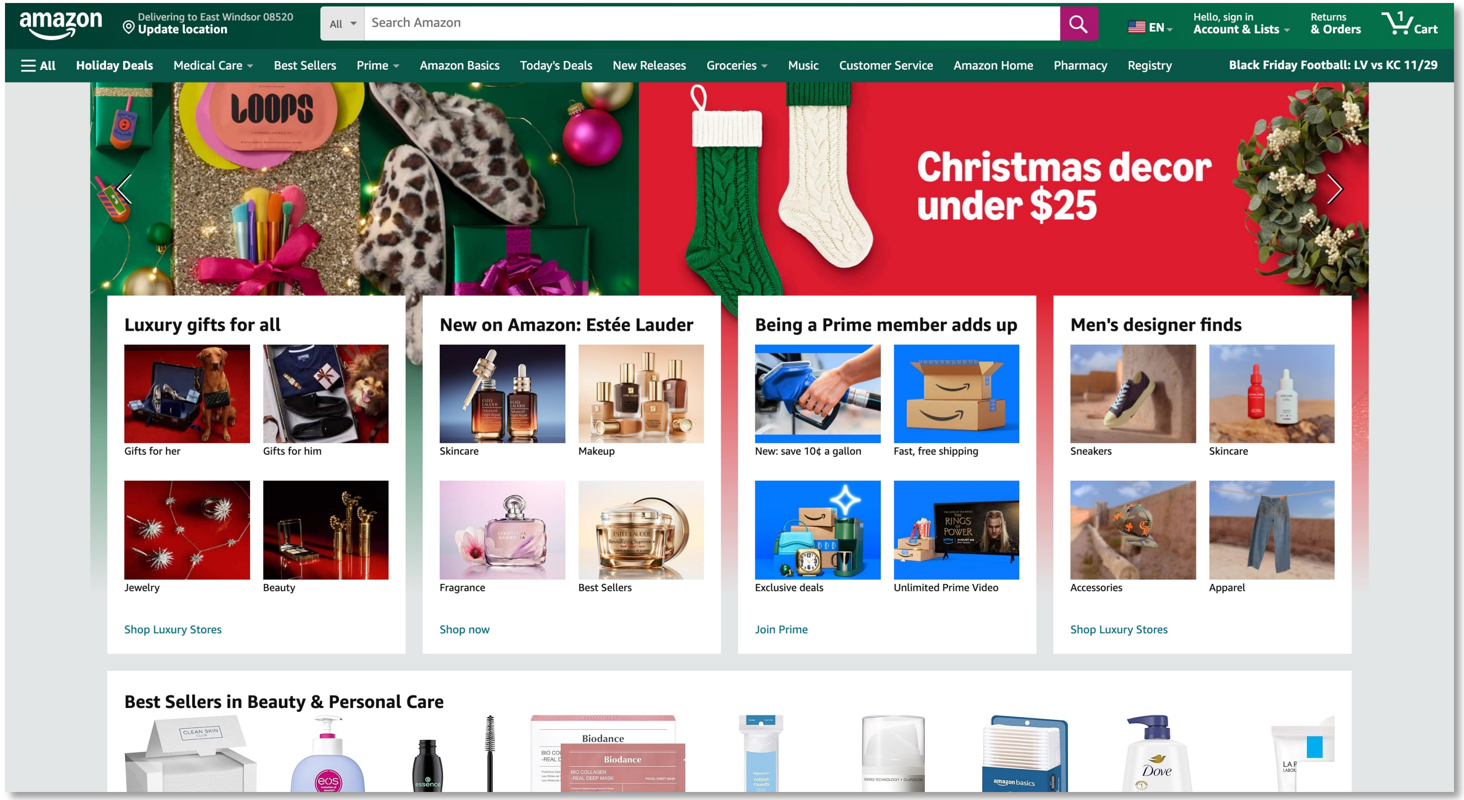
$$\hat{T}: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$$

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

Why hasn't it been done already?

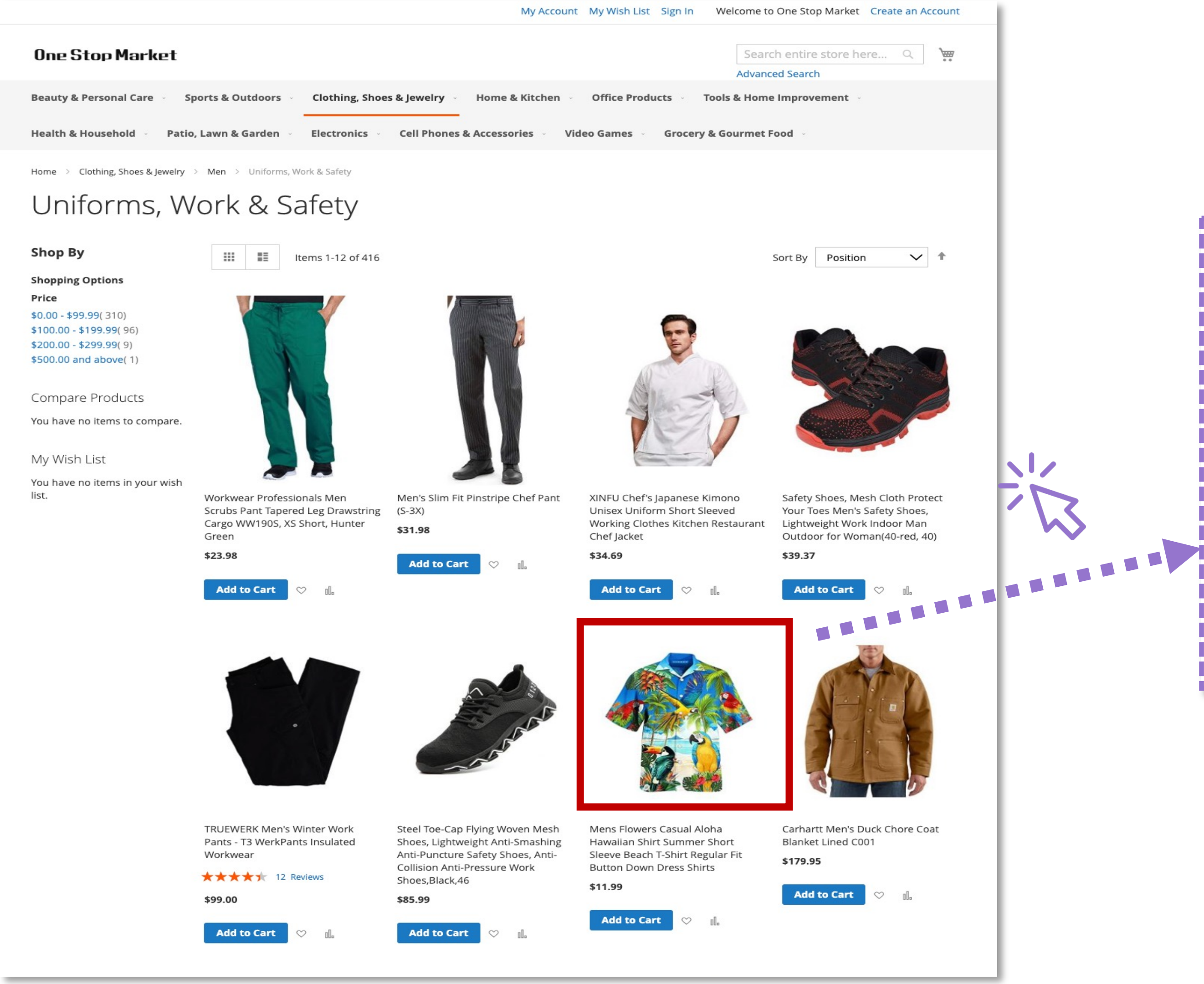


V.S



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

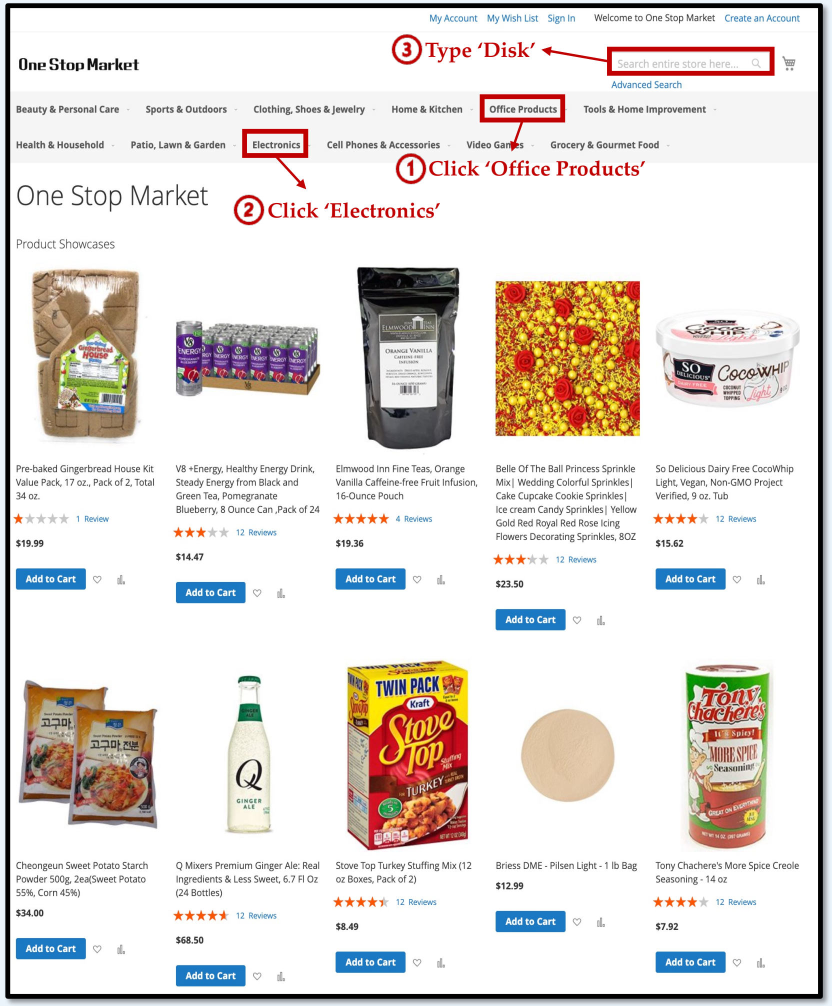
LLMs can predict state transitions


My Account
My Wish List
Sign In
Welcome to One Stop Market
Create an Account
One Stop Market
Search entire store here...
Advanced Search
Beauty & Personal Care
Sports & Outdoors
Clothing, Shoes & Jewelry
Home & Kitchen
Office Products
Tools & Home Improvement
Health & Household
Patio, Lawn & Garden
Electronics
Cell Phones & Accessories
Video Games
Grocery & Gourmet Food
Home > 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
Workwear Professionals Men Scrubs Pant Tapered Leg Drawstring Cargo WW190S, 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-red, 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
$85.99
Add to Cart
Mens Flowers Casual Aloha Hawaiian Shirt Summer Short Sleeve Beach T-Shirt Regular Fit Button Down Dress Shirts
$11.99
Add to Cart
Carhartt Men's Duck Chore Coat Blanket Lined C001
$179.95
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'

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| Wedding Colorful Sprinkles| Cake Cupcake Cookie Sprinkles| Ice cream Candy Sprinkles| Yellow Gold Red Royal Red Rose Icing Flowers Decorating Sprinkles, 8OZ
12 Reviews
$23.50

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

Cheongeun Sweet Potato Starch Powder 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

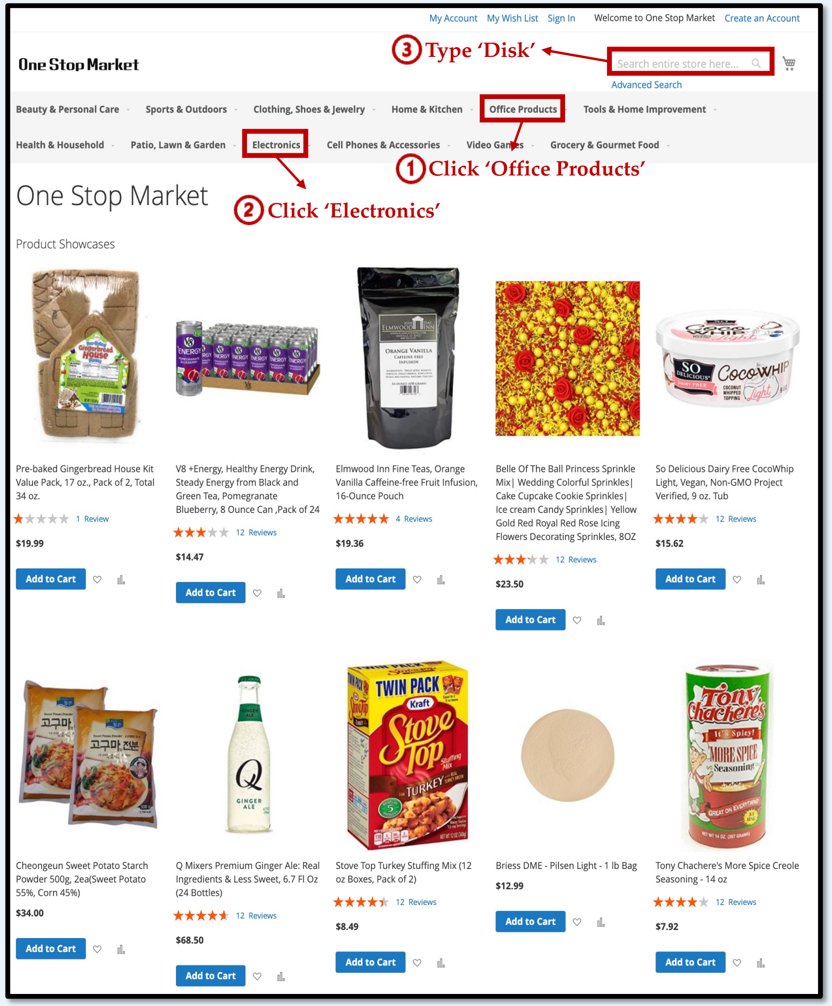
Stove Top Turkey Stuffing Mix (12 oz Boxes, Pack of 2)
12 Reviews
$8.49

Briess DME - Pilsen Light - 1 lb Bag
$12.99

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

WebDreamer: model-based planner for web agents

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



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

One Stop Market

Search entire store here...

Advanced Search

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

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

③ Type 'Disk'

① Click 'Office Products'

② Click 'Electronics'

One Stop Market

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 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 Red Royal Red 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
12 Reviews
$15.62
Add to Cart

Cheongeun Sweet Potato Starch Powder 500g, 2ea(Sweet Potato 55%, Corn 45%)
$34.00
Add to Cart

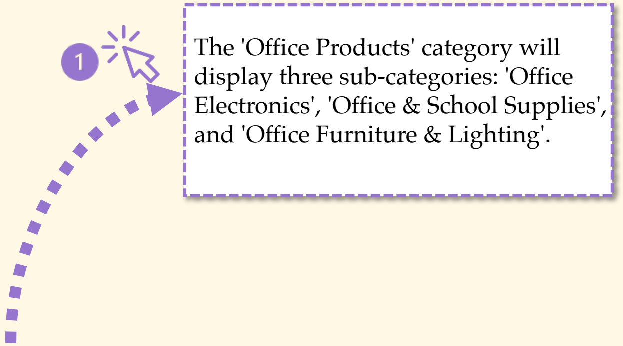
Q Mixers Premium Ginger Ale: Real Ingredients & Less Sweet, 6.7 Fl Oz (24 Bottles)
12 Reviews
$68.50
Add to Cart

Stove Top Turkey Stuffing Mix (12 oz Boxes, Pack of 2)
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

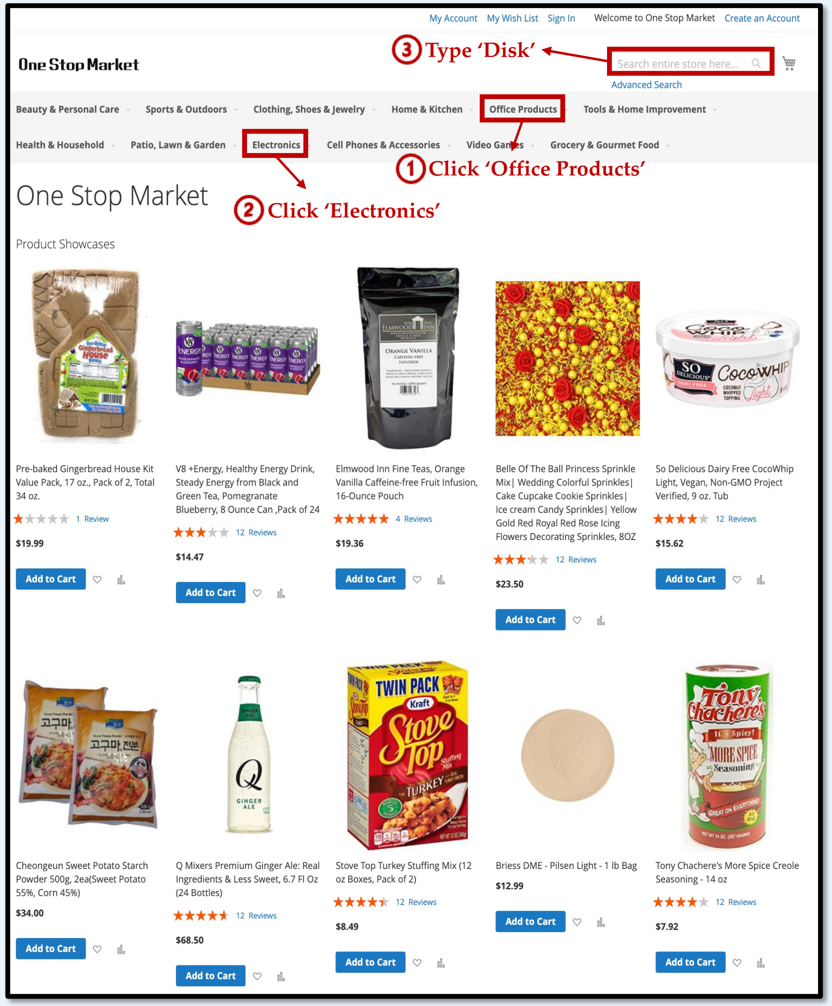
Stage I: Simulation



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.

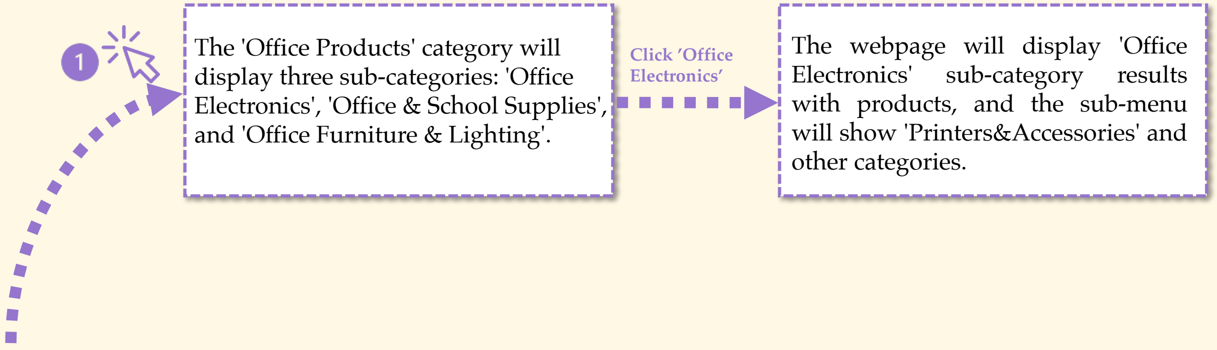

My Account
My Wish List
Sign In
Welcome to One Stop Market
Create an Account
One Stop Market
Search entire store here...
Advanced Search
Beauty & Personal Care
Sports & Outdoors
Clothing, Shoes & Jewelry
Home & Kitchen
Office Products
Tools & Home Improvement
Health & Household
Patio, Lawn & Garden
Electronics
Cell Phones & Accessories
Video Games
Grocery & Gourmet Food
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| Wedding Colorful Sprinkles| Cake Cupcake Cookie Sprinkles| Ice cream Candy Sprinkles| Yellow Gold Red Royal Red Rose Icing Flowers Decorating Sprinkles, 8OZ
12 Reviews
$23.50
So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project Verified, 9 oz. Tub
12 Reviews
$15.62
Cheongeun Sweet Potato Starch Powder 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
Stove Top Turkey Stuffing Mix (12 oz Boxes, Pack of 2)
12 Reviews
$8.49
Briess DME - Pilsen Light - 1 lb Bag
$12.99
Tony Chachere's More Spice Creole Seasoning - 14 oz
12 Reviews
$7.92
Add to Cart

③ Type 'Disk'

① Click 'Office Products'

② Click 'Electronics'

Stage I: Simulation



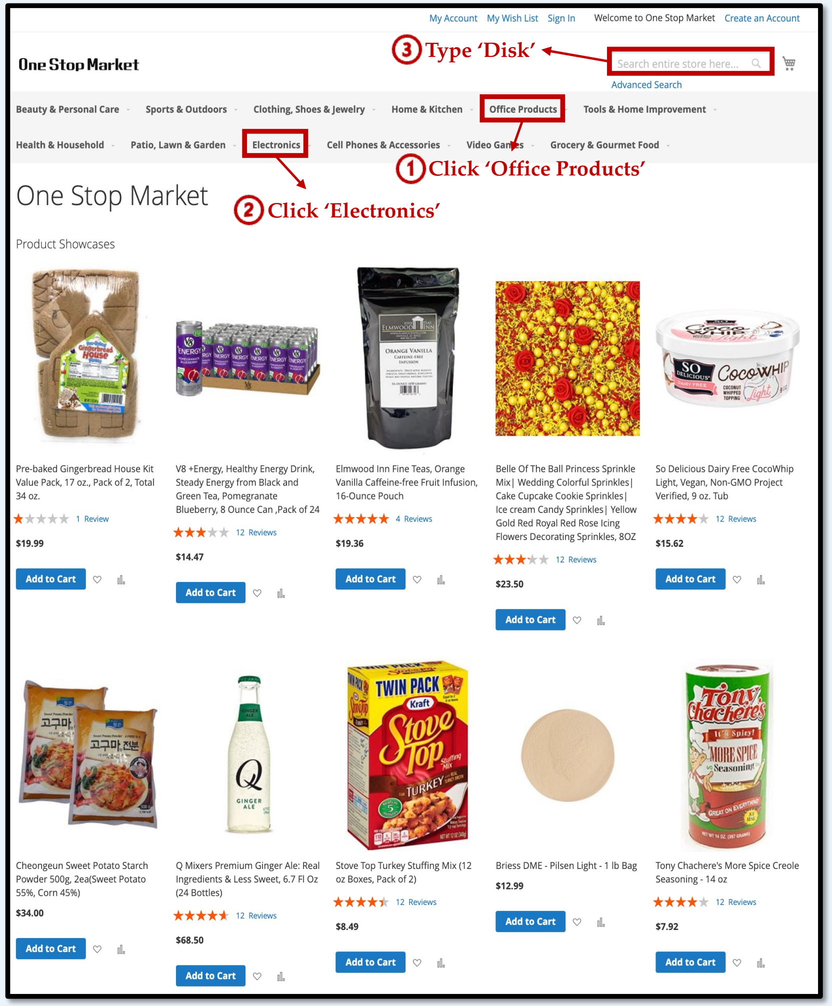
The 'Office Products' category will display three sub-categories: 'Office Electronics', 'Office & School Supplies', and 'Office Furniture & Lighting'.

Click 'Office Electronics'

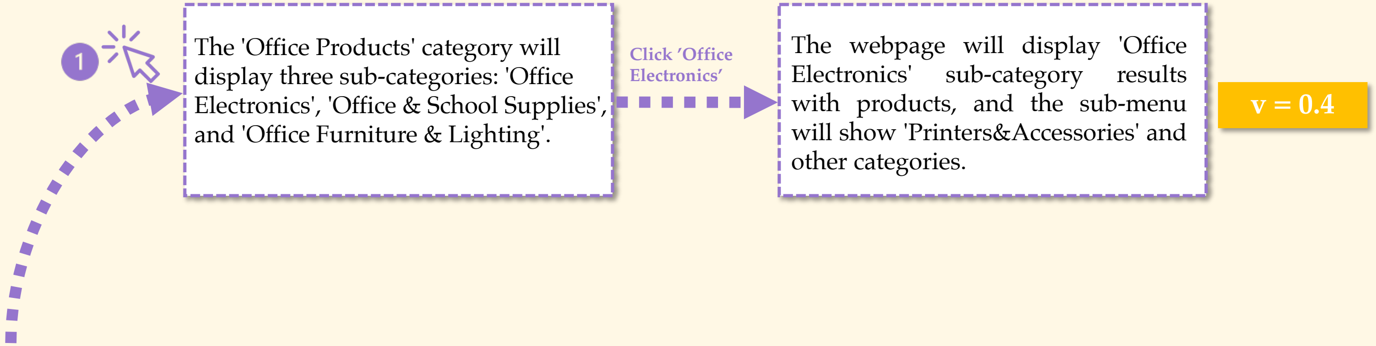
The webpage will display 'Office Electronics' sub-category results with products, and the sub-menu will show 'Printers&Accessories' and other categories.

WebDreamer: model-based planner for web agents

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


My Account
My Wish List
Sign In
Welcome to One Stop Market
Create an Account
One Stop Market
③ Type 'Disk'
Search entire store here...
Advanced Search
Beauty & Personal Care
Sports & Outdoors
Clothing, Shoes & Jewelry
Home & Kitchen
Office Products
Tools & Home Improvement
Health & Household
Patio, Lawn & Garden
Electronics
Cell Phones & Accessories
Video Games
Grocery & Gourmet Food
① Click 'Office Products'
② Click 'Electronics'
One Stop Market
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 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 Red Royal Red 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
12 Reviews
$15.62
Add to Cart
Cheongeun Sweet Potato Starch Powder 500g, 2ea(Sweet 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
Stove Top Turkey Stuffing Mix (12 oz Boxes, Pack of 2)
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

Stage I: Simulation



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

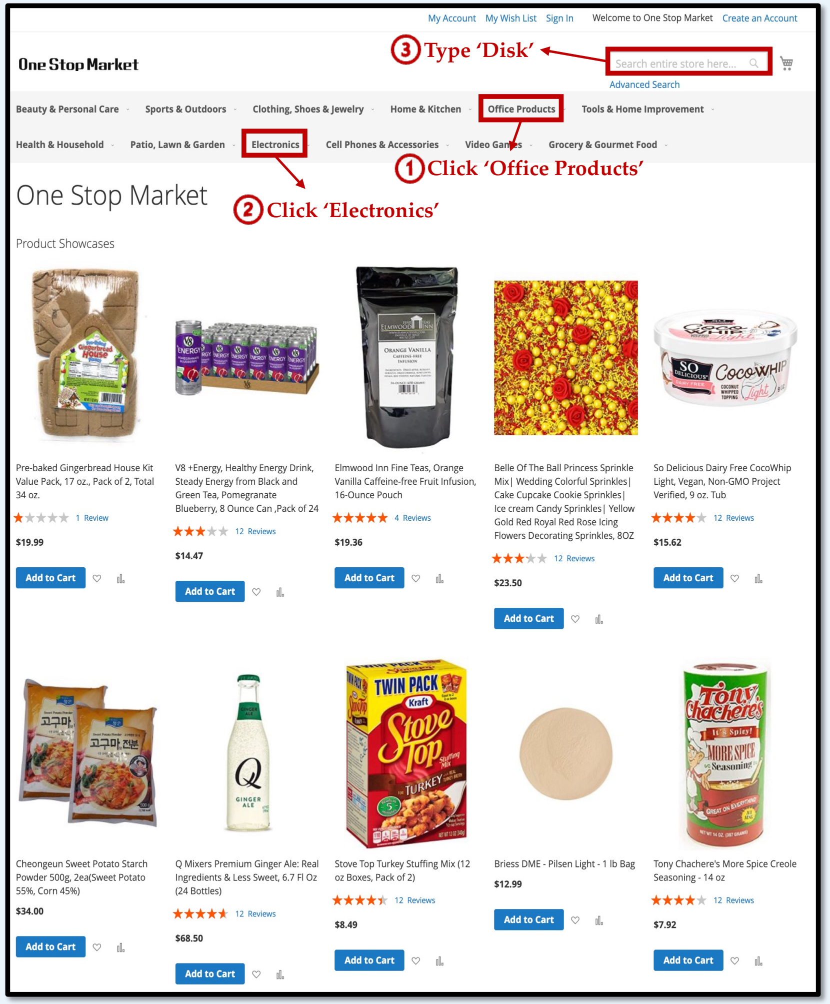
Click 'Office Electronics'

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

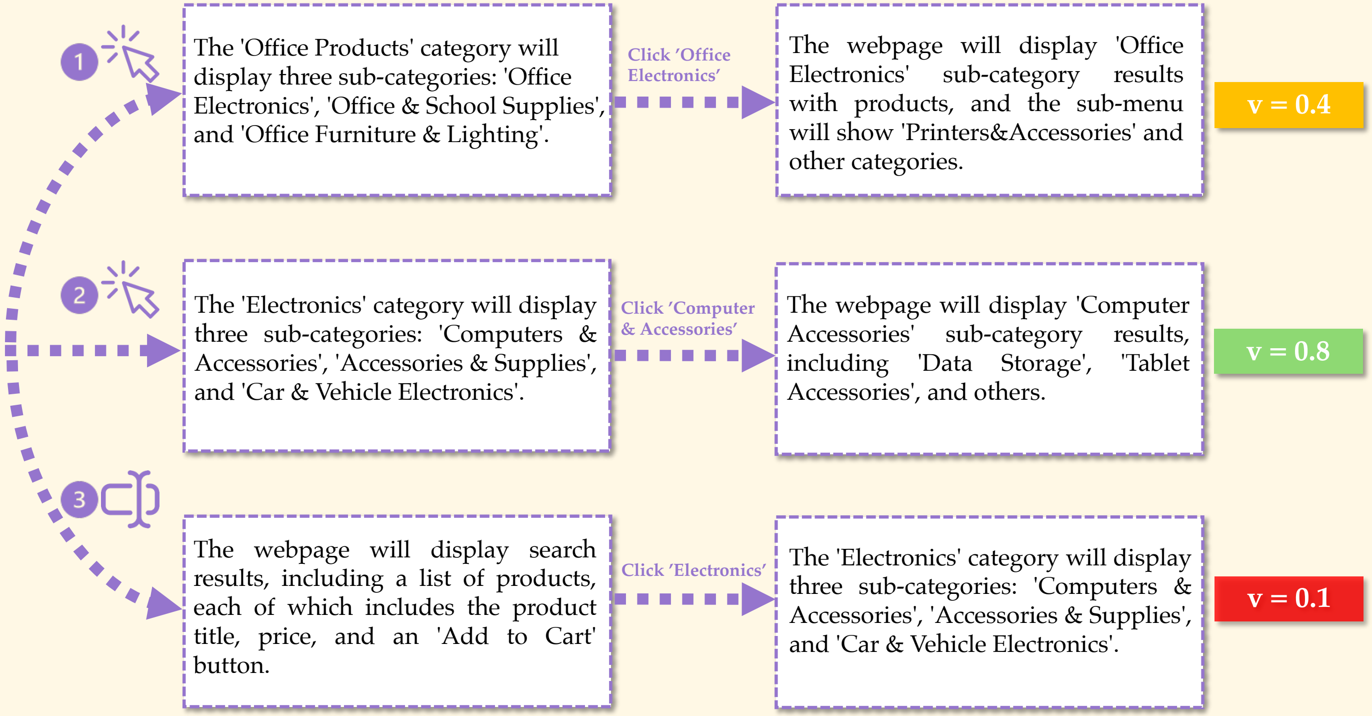
v = 0.4

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.

v = 0.4

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

Click 'Computer & Accessories'

The webpage will display 'Computer Accessories' sub-category results, including 'Data Storage', 'Tablet Accessories', and others.

v = 0.8

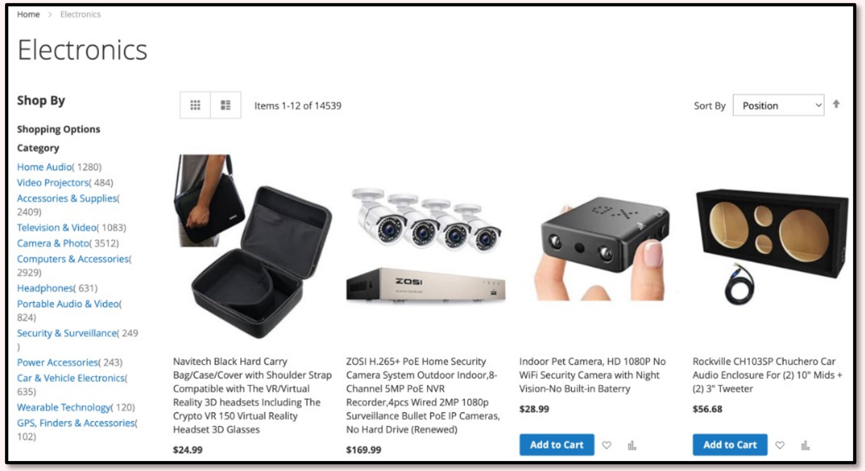
The webpage will display search results, including a list of products, each of which includes the product title, price, and an 'Add to Cart' button.

Click 'Electronics'

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

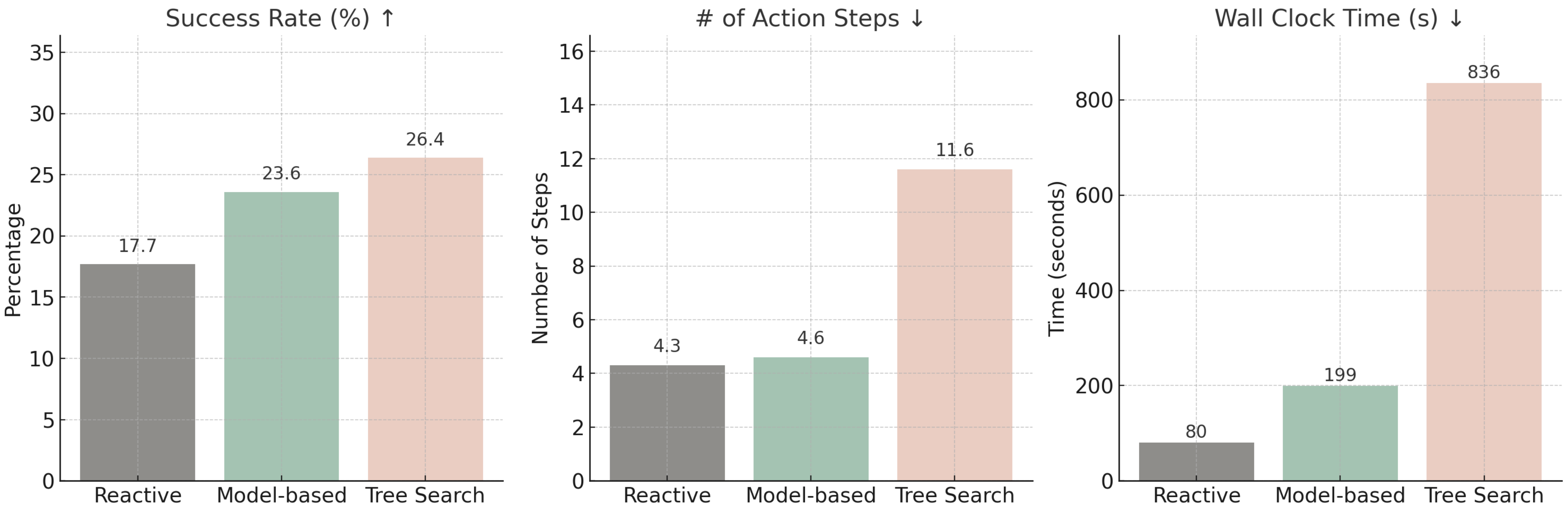
v = 0.1

Stage II: Execution


Home > Electronics
Electronics
Shop By
Shopping Options
Category
Home Audio(1280)
Video Projectors(484)
Accessories & Supplies(2409)
Television & Video(1083)
Camera & Photo(3512)
Computers & Accessories(2929)
Headphones(631)
Portable Audio & Video(824)
Security & Surveillance(249)
Power Accessories(243)
Car & Vehicle Electronics(635)
Wearable Technology(120)
GPS, Finders & Accessories(102)
Items 1-12 of 14539
Sort By
Position
Navitech Black Hard Carry Bag/Case/Cover with Shoulder Strap Compatible with The VR/Virtual Reality 3D headsets Including The Crypto VR 150 Virtual Reality Headset 3D Glasses
$24.99
ZOSI H.265+ PoE Home Security Camera System Outdoor Indoor,8-Channel 5MP PoE NVR Recorder,4pcs Wired 2MP 1080p Surveillance Bullet PoE IP Cameras, No Hard Drive (Renewed)
$169.99
Indoor Pet Camera, HD 1080P No WiFi Security Camera with Night Vision-No Built-in Baterry
$28.99
Rockville CH103SP Chuchero Car Audio Enclosure For (2) 10" Mids + (2) 3" Tweeter
$56.68
Add to Cart

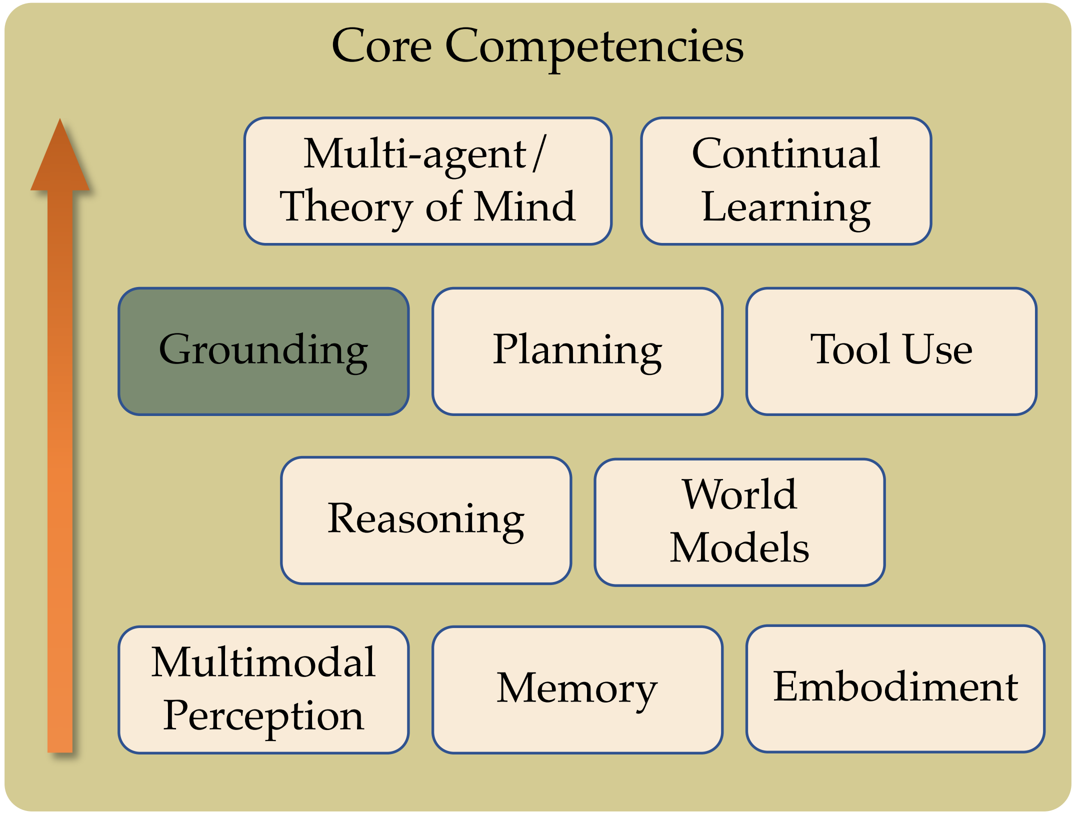
Results on VisualWebArena

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



Baselines are from Koh et al., (2024)

Grounding



Core Competencies

Multi-agent / Theory of Mind

Continual Learning

Grounding

Planning

Tool Use

Reasoning

World Models

Multimodal Perception

Memory

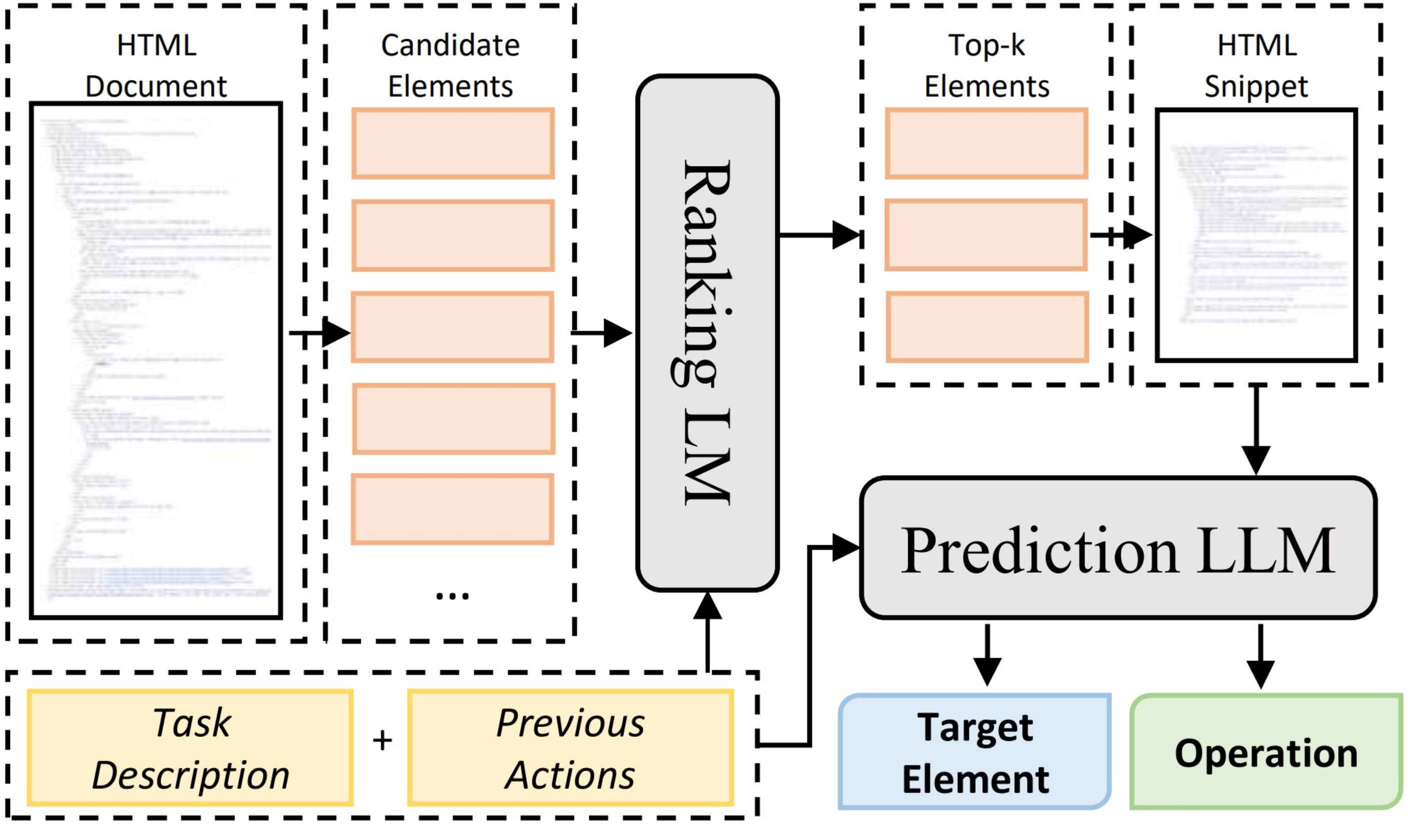
Embodiment

https://arxiv.org/abs/2410.05243

Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents

Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie,
Cheng Chang, Yiheng Shu, Huan Sun, Yu Su

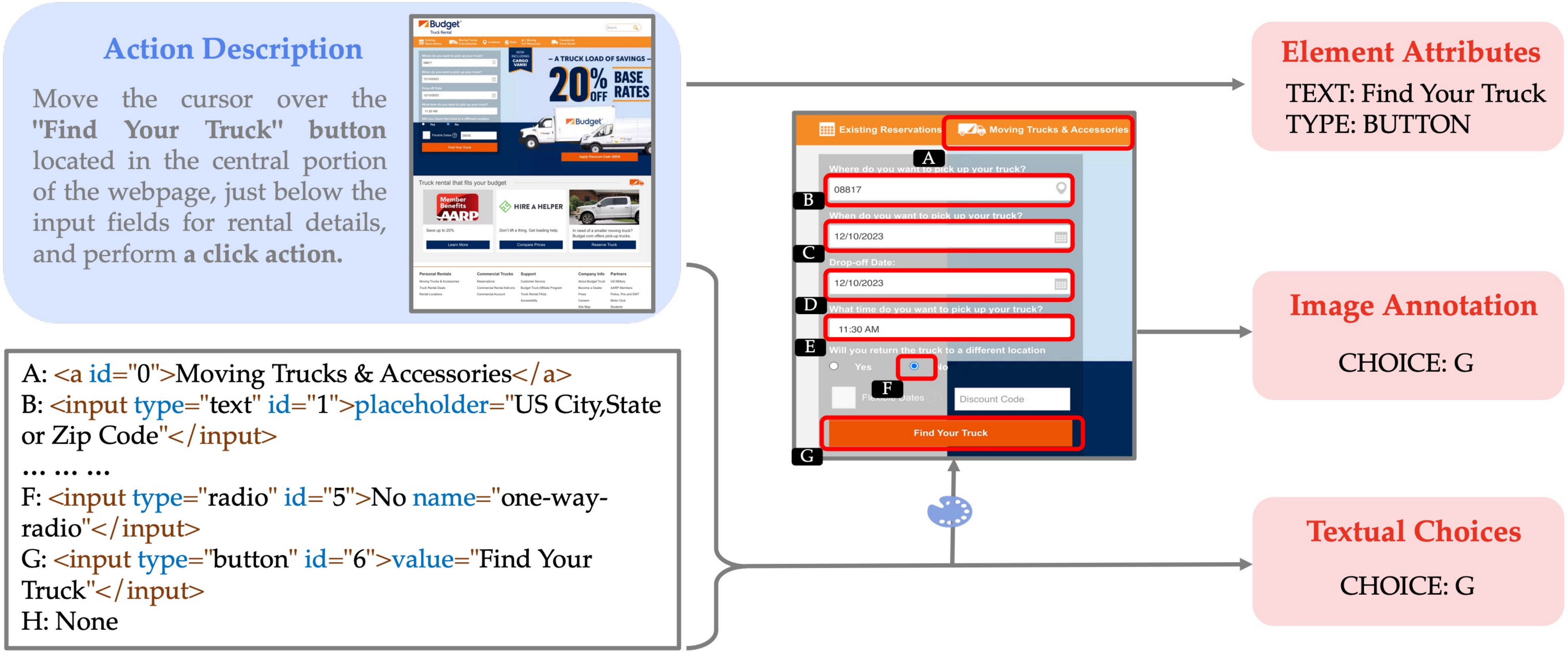
Embodiment of Web/GUI Agents: Evolution



Xiang Deng

	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: <a id="0">Moving Trucks & Accessories</a>
B: <input type="text" id="1">placeholder="US City,State or Zip Code"</input>
… … …
F: <input type="radio" id="5">No name="one-way-radio"</input>
G: <input type="button" id="6">value="Find Your Truck"</input>
H: None

Existing Reservations
Moving Trucks & Accessories
A
Where do you want to pick up your truck?
B
08817
When do you want to pick up your truck?
C
12/10/2023
Drop-off Date:
D
12/10/2023
What time do you want to pick up your truck?
E
11:30 AM
Will you return the truck to a different location
Yes
F
Flexible Dates
Discount Code
Find Your Truck
G

Element Attributes

TEXT: Find Your Truck
TYPE: BUTTON

Image Annotation

CHOICE: G

Textual Choices

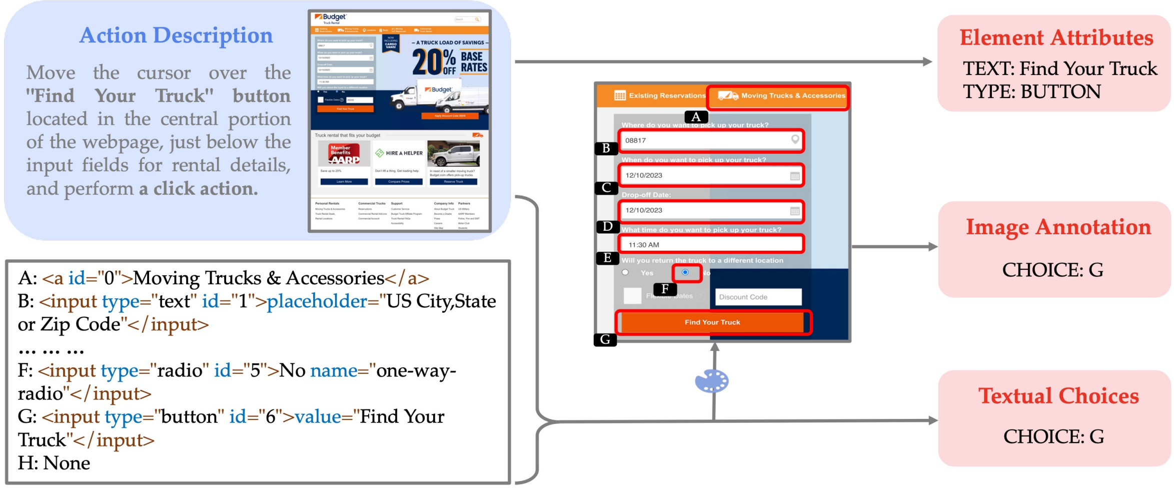
CHOICE: G

Boyuan Zheng

Boyu Gou

	Mind2Web (23' May)	SeeAct (24' Jan)
Sensory Inputs	HTML/DOM	Screenshot + DOM
Effectors	Multi-choice Selection	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: <a id="0">Moving Trucks & Accessories</a>
B: <input type="text" id="1">placeholder="US City,State or Zip Code"</input>
… … …
F: <input type="radio" id="5">No name="one-way-radio"</input>
G: <input type="button" id="6">value="Find Your Truck"</input>
H: None
```

Text-based Representations: Limitations

- *Noisy and incomplete*
  - **95.9%** of home pages have accessibility conformance errors
  - Avg. **56.8** errors per page[1]
- *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	Screenshot + DOM
Effectors	Multi-choice Selection	Multi-choice Selection

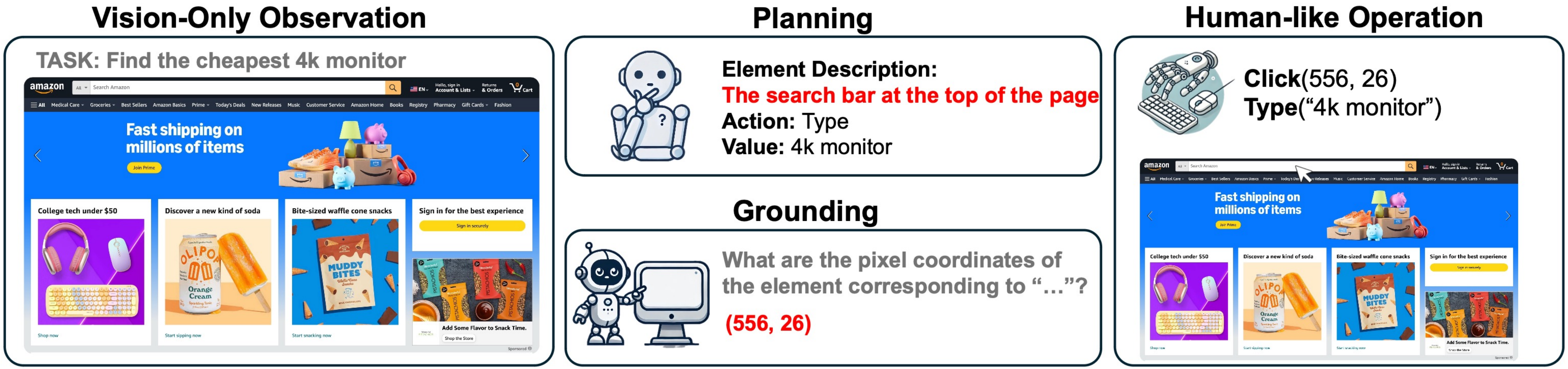
---

[1] https://webaim.org/projects/million/

Embodiment of Web/GUI Agents: Evolution

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

Boyu Gou


amazon
Search Amazon
Fast shipping on millions of items
Join Prime
College tech under $50
Discover a new kind of soda
Bite-sized waffle cone snacks
Sign in for the best experience
Sign in securely
Add Some Flavor to Snack Time.
Shop the Store
Shop now
Start sipping now
Start snacking now

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)

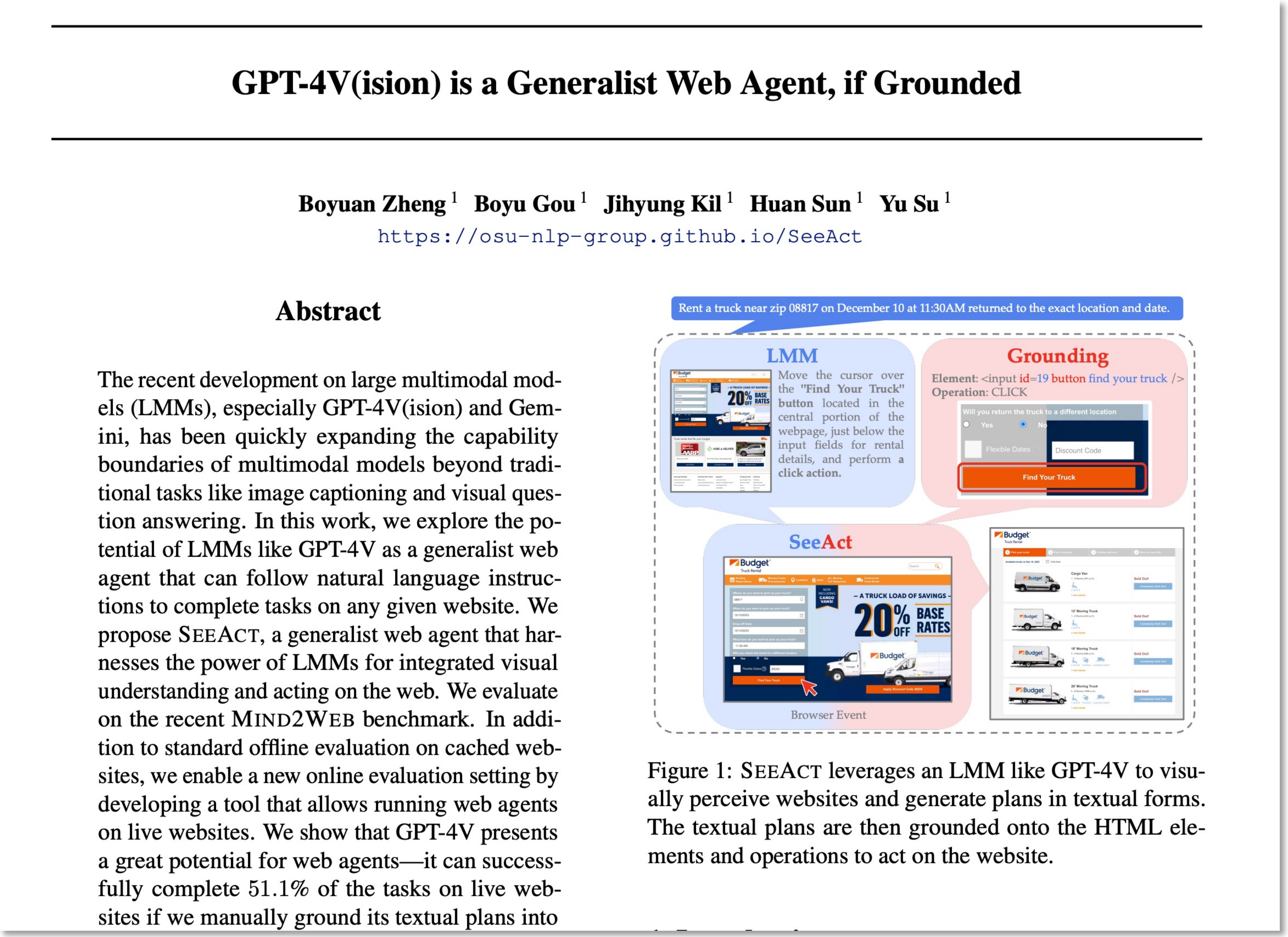
Human-like Operation

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

	Mind2Web (23' May)	SeeAct (24' Jan)	SeeAct-V (24' Aug)
Sensory Inputs	HTML/DOM	Screenshot + DOM	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 Zheng[1] **Boyu Gou**[1] **Jihyung Kil**[1] **Huan Sun**[1] **Yu Su**[1]
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

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

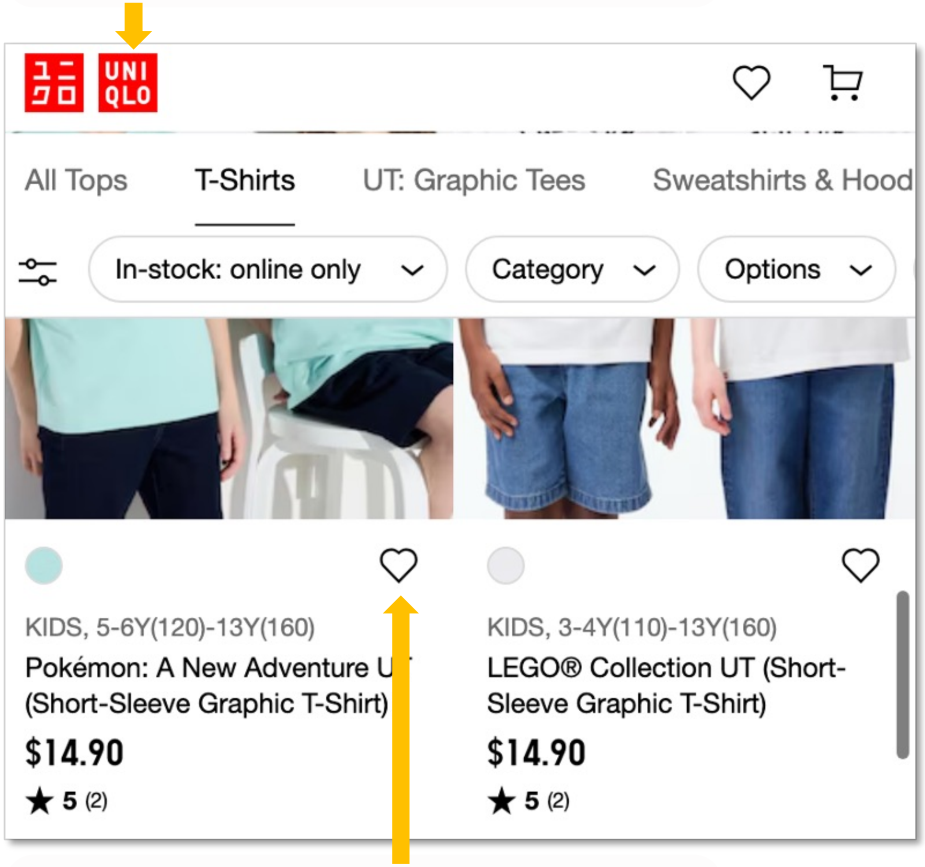
Move the cursor over the **"Find Your Truck" button** located in the central portion of the webpage, just below the input fields for rental details, and perform **a click action.**

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

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

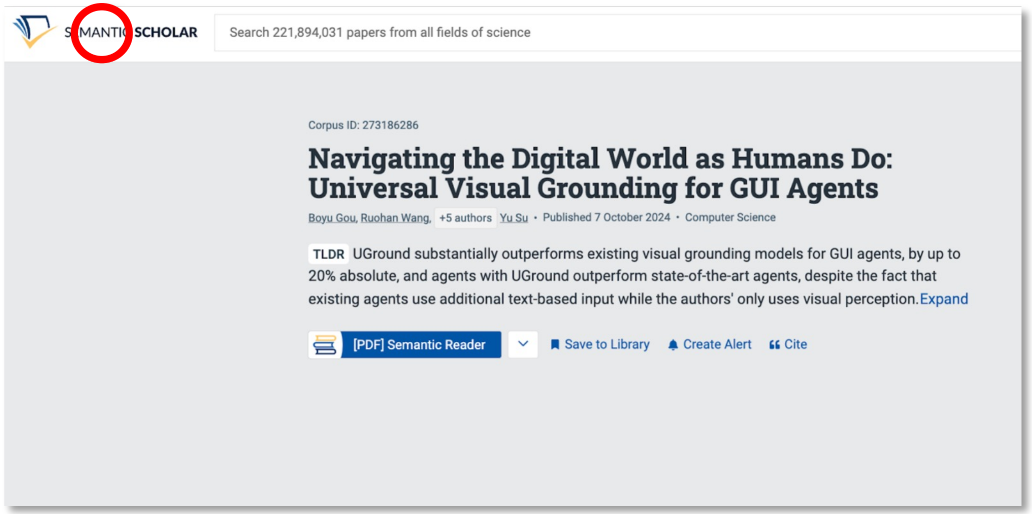
1. Red icon labeled “UNIQLO”
2. Button at the top left corner
3. Navigate back to the homepage



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



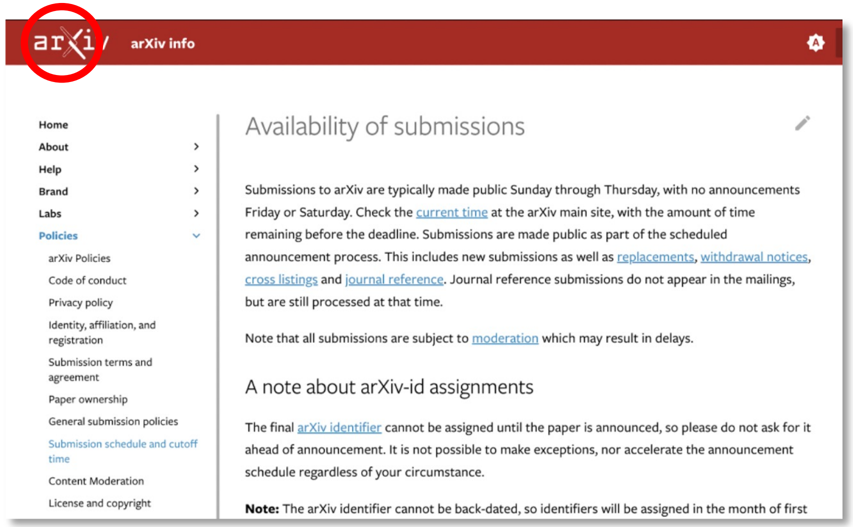
Corpus ID: 273186286

Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents

Boyu Gou, Ruohan Wang, +5 authors Yu Su • Published 7 October 2024 • Computer Science

TLDR UGround substantially outperforms existing visual grounding models for GUI agents, by up to 20% absolute, and agents with UGround outperform state-of-the-art agents, despite the fact that existing agents use additional text-based input while the authors' only uses visual perception. Expand

"Go to homepage"



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.

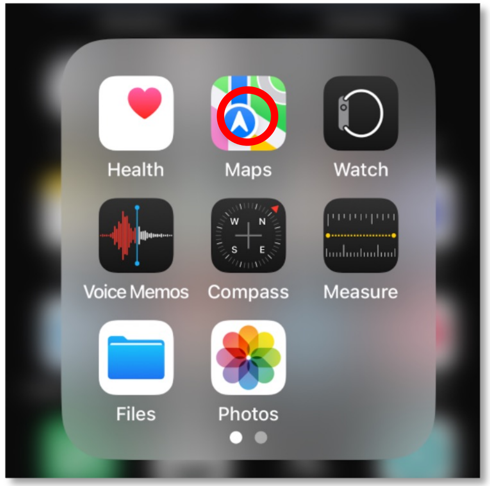
Note that all submissions are subject to moderation which may result in delays.

A note about arXiv-id assignments

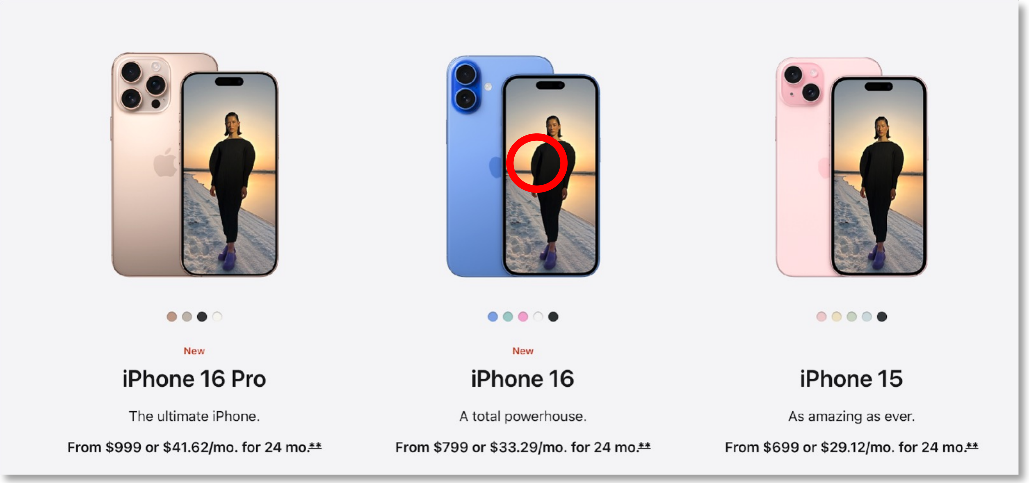
The final arXiv identifier cannot be assigned until the paper is announced, so please do not ask for it ahead of announcement. It is not possible to make exceptions, nor accelerate the announcement schedule regardless of your circumstance.

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

"Go to homepage"



"Open Maps"



New

iPhone 16 Pro

The ultimate iPhone.

From $999 or $41.62/mo. for 24 mo.**

New

iPhone 16

A total powerhouse.

From $799 or $33.29/mo. for 24 mo.**

iPhone 15

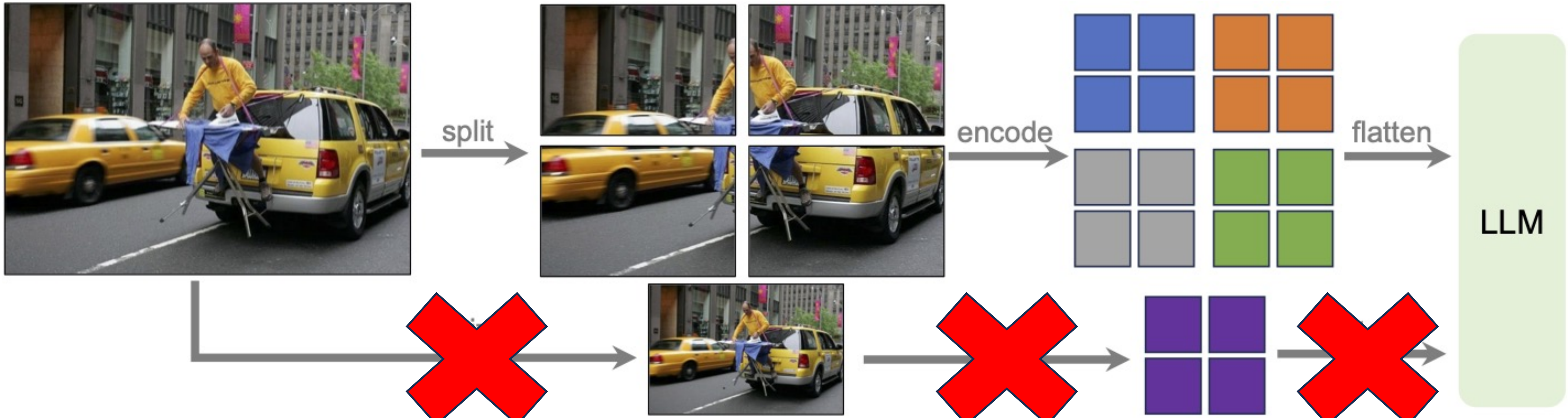
As amazing as ever.

From $699 or $29.12/mo. for 24 mo.**

"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

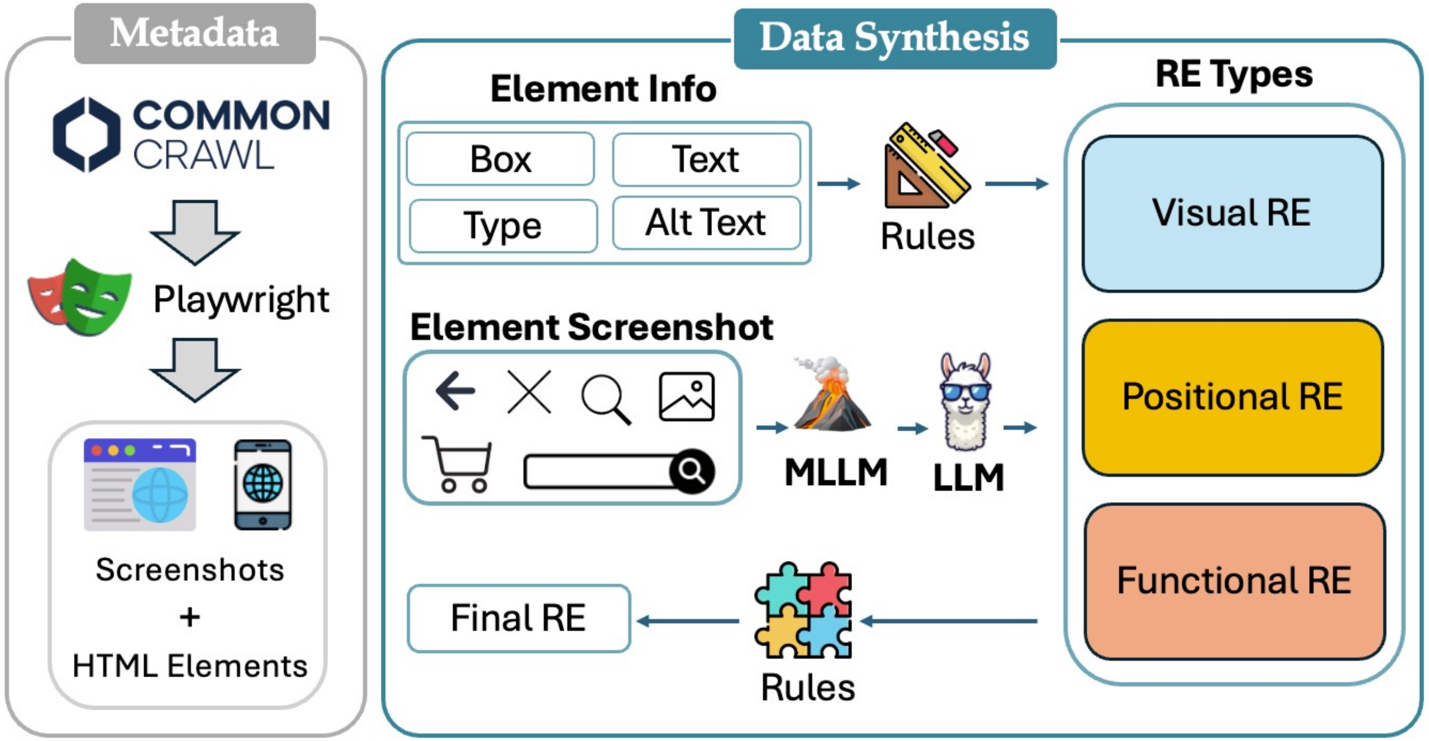


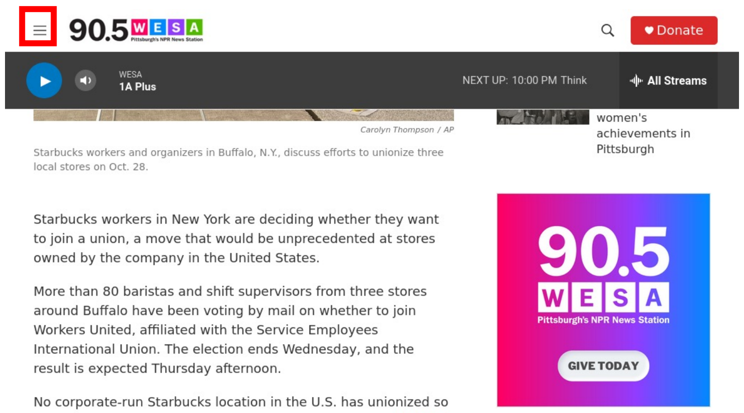
Synthetic Data is Key

- 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





Starbucks workers and organizers in Buffalo, N.Y., discuss efforts to unionize three local stores on Oct. 28.

Starbucks workers in New York are deciding whether they want to join a union, a move that would be unprecedented at stores owned by the company in the United States.

More than 80 baristas and shift supervisors from three stores around Buffalo have been voting by mail on whether to join Workers United, affiliated with the Service Employees International Union. The election ends Wednesday, and the result is expected Thursday afternoon.

No corporate-run Starbucks location in the U.S. has unionized so

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.

LLM Rewrite

three-line menu icon

Rules

Absolute Position: Top Left Corner
Relative Position: to the left of "Donate"
…

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

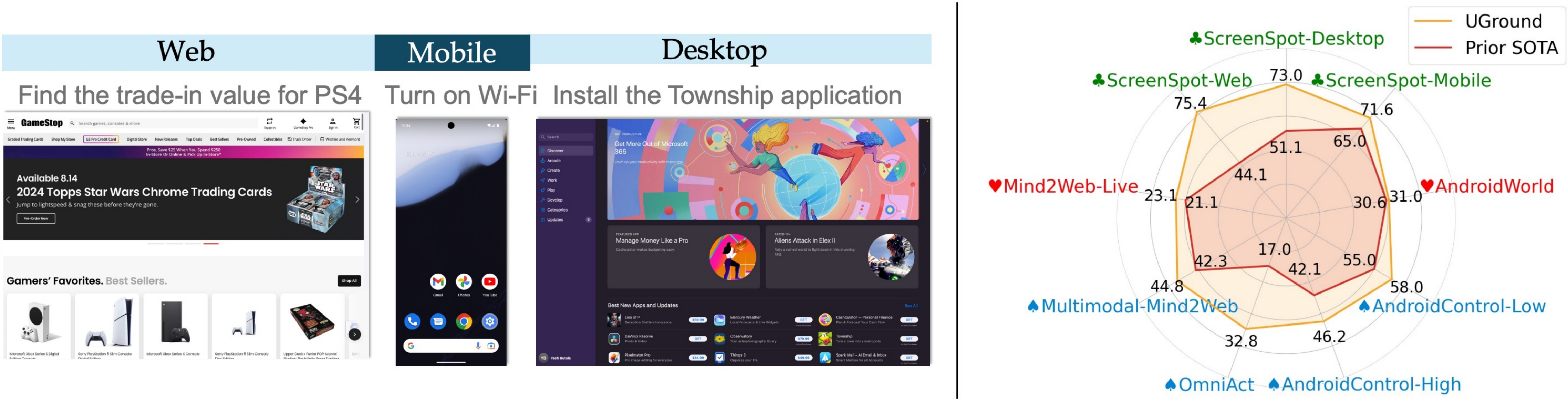
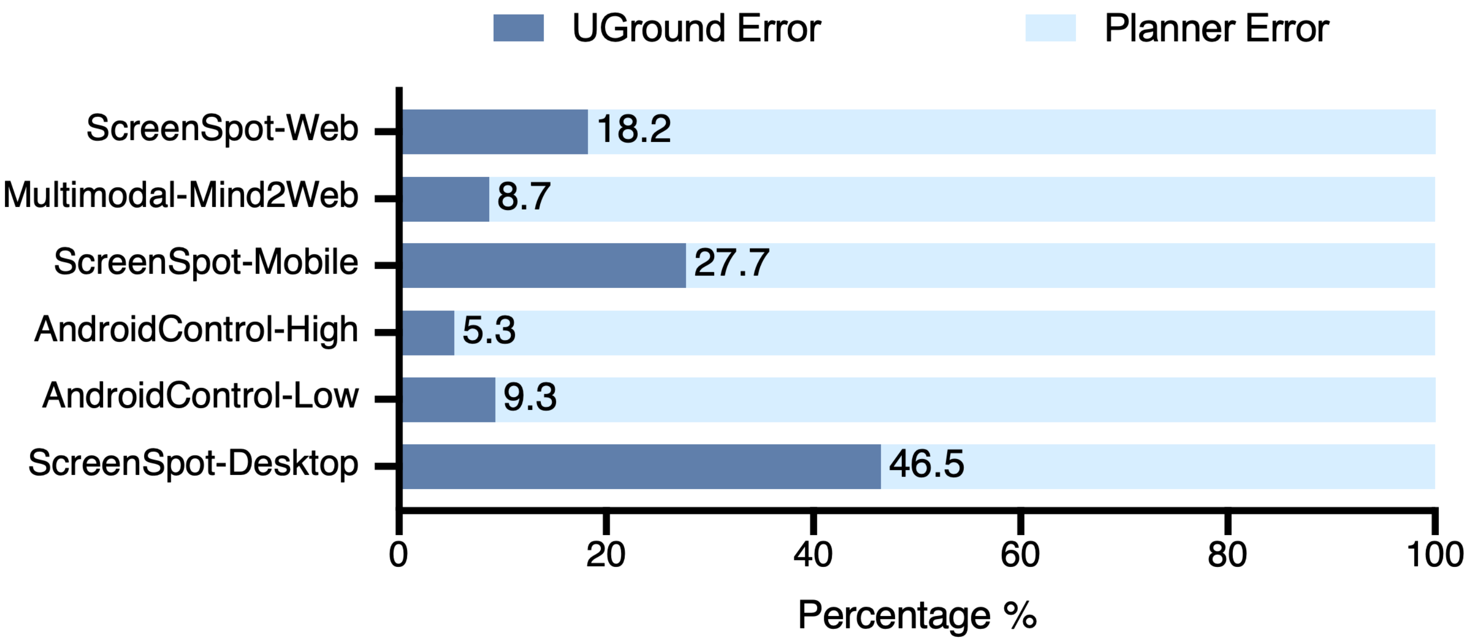


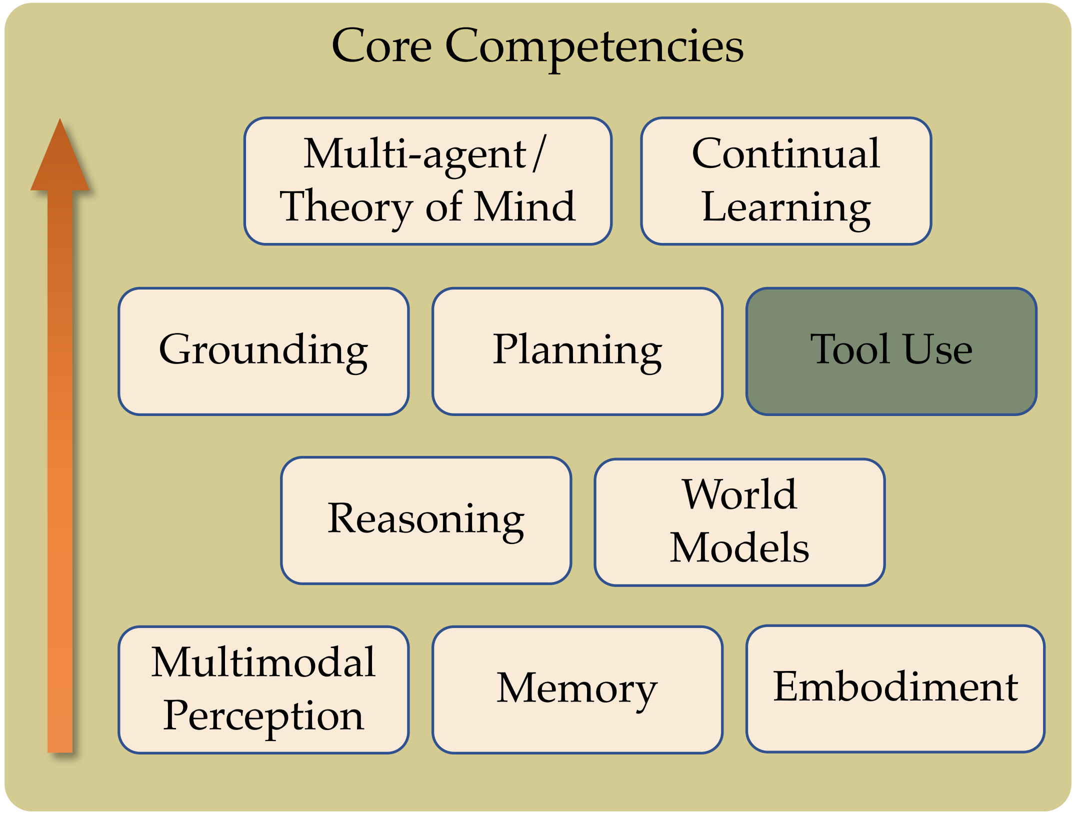
Figure 1: Examples of agent tasks across platforms and performance on **GUI grounding** (♣: ScreenSpot), **offline agent** (♠: Multimodal-Mind2Web, AndroidControl, and OmniAct), and **online agent benchmarks** (♥: Mind2Web-Live and AndroidWorld) when using GPT-4 as the planner.

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



Core Competencies

Multi-agent / Theory of Mind

Continual Learning

Grounding

Planning

Tool Use

Reasoning

World Models

Multimodal Perception

Memory

Embodiment

ACL 2024

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

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

We don’t master a tool by just reading the ‘manual’

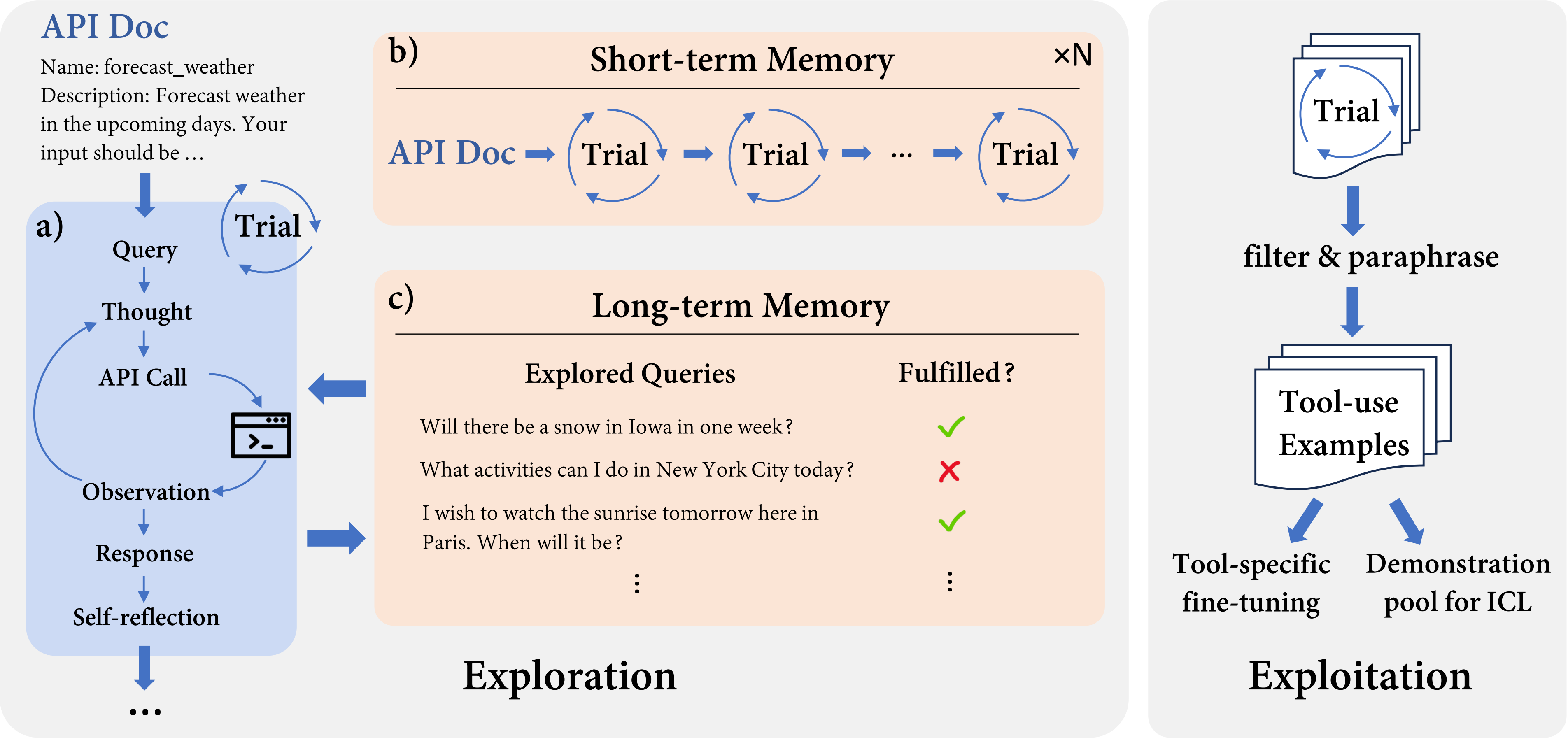
this too and humans once they reach
about age five

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

- Trial and error
- Imagination
- Memory

Tool learning through *simulated trial and error*

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



API Doc

Name: forecast_weather
Description: Forecast weather in the upcoming days. Your input should be …

a)

Trial

Query
Thought
API Call
Observation
Response
Self-reflection

…

b) Short-term Memory ×N

API Doc → Trial → Trial → … → Trial

c) Long-term Memory

Explored Queries	Fulfilled?
Will there be a snow in Iowa in one week?	✓
What activities can I do in New York City today?	✗
I wish to watch the sunrise tomorrow here in Paris. When will it be?	✓
⋮	⋮

Exploration

Trial

filter & paraphrase

Tool-use Examples

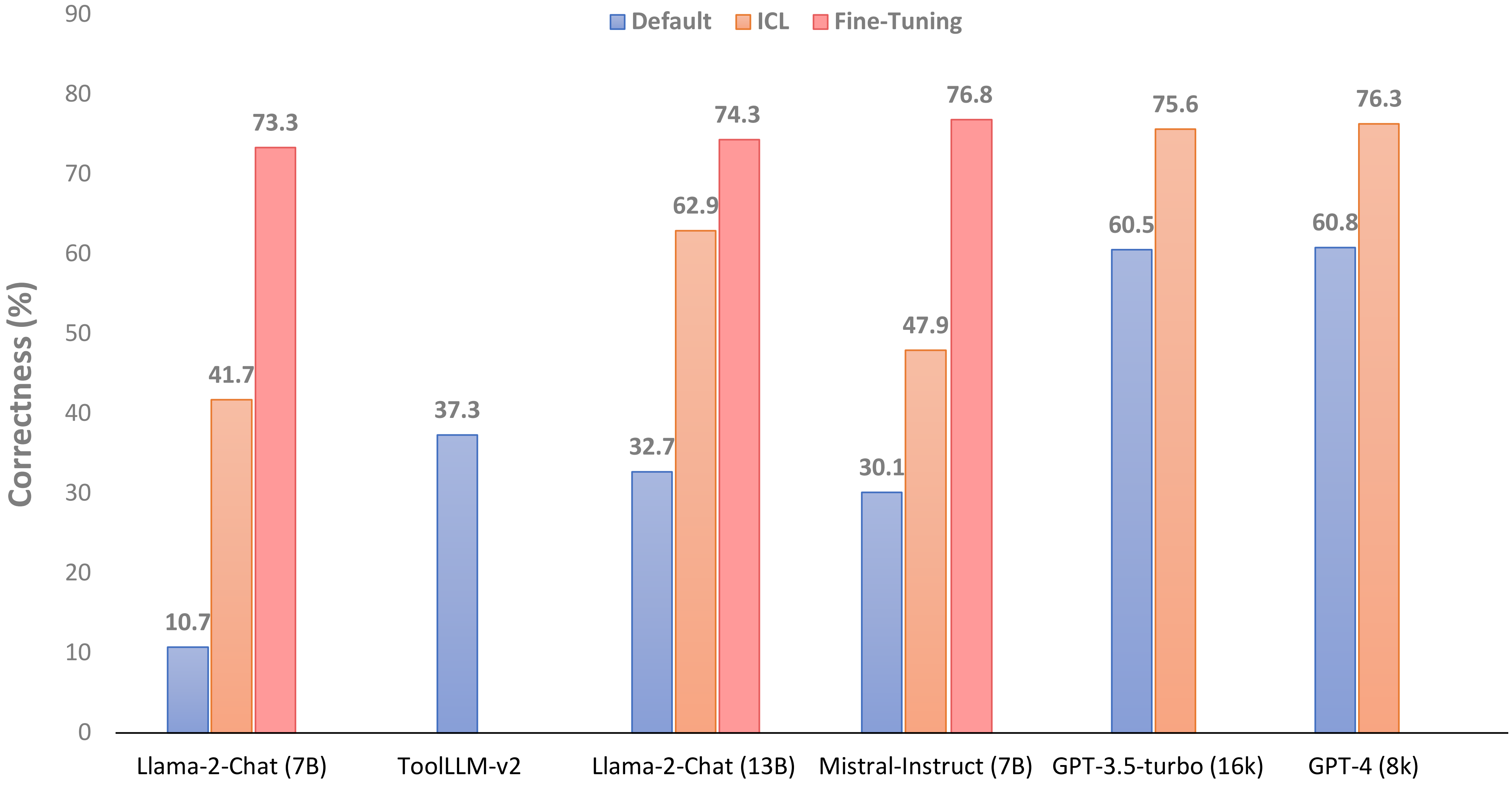
Tool-specific fine-tuning

Demonstration pool for ICL

Exploitation

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.

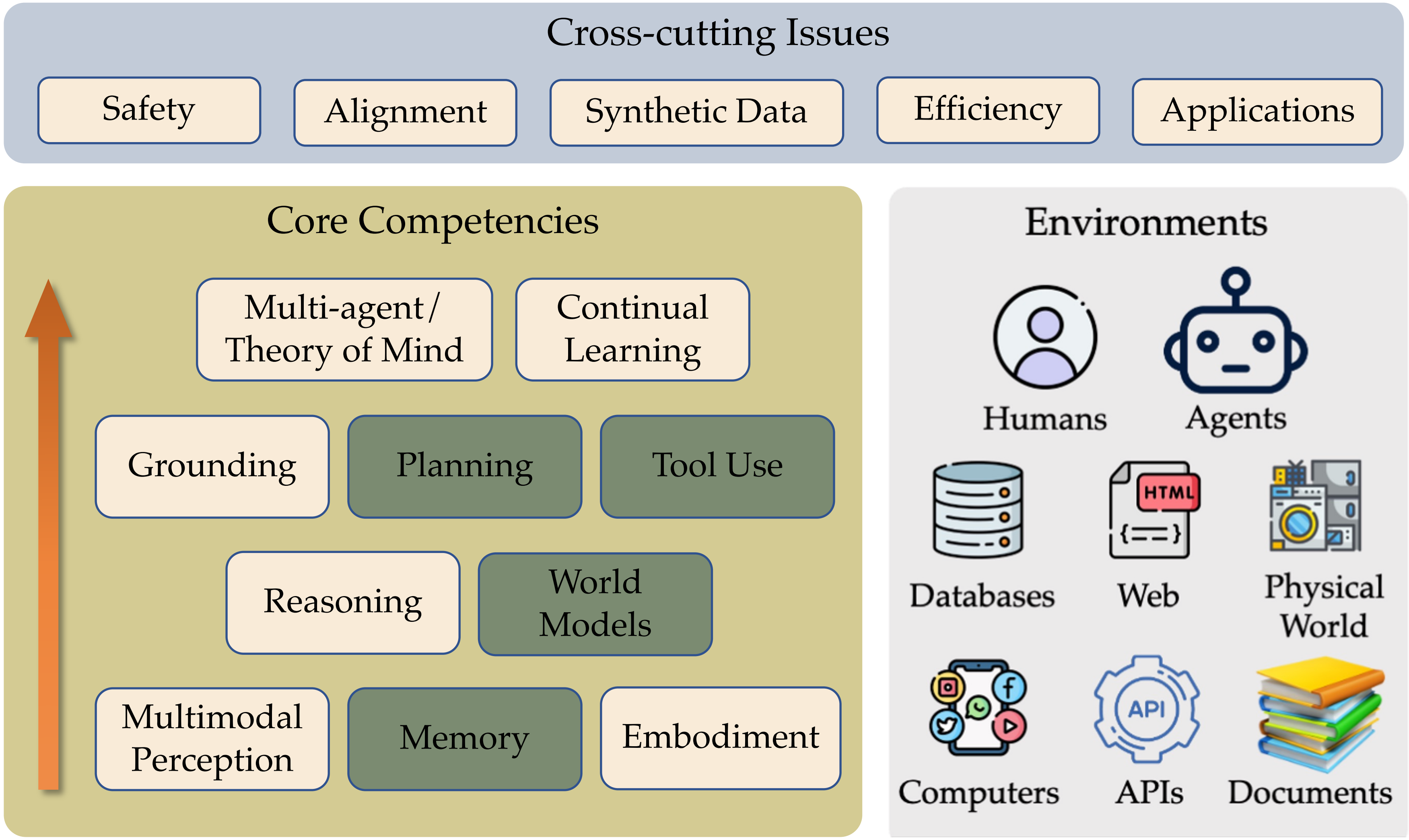
- Natural selection is a real magical power. Over billions of years of evolution, natural selection has found numerous good designs (‘local optima’) that *just make sense*; we should take some inspirations!

- 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



Cross-cutting Issues

- Safety
- Alignment
- Synthetic Data
- Efficiency
- Applications

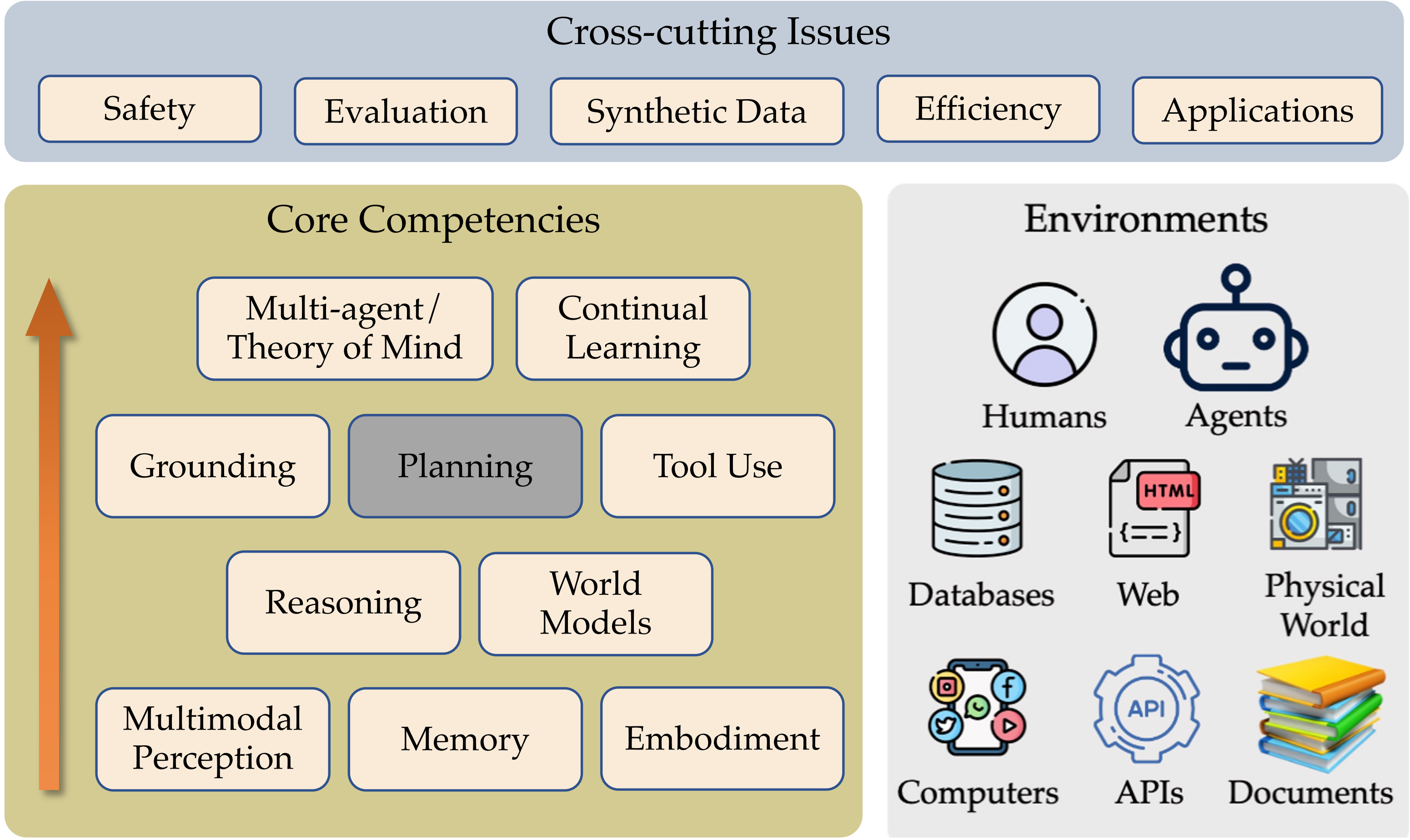
Core Competencies

- Multi-agent/ Theory of Mind
- Continual Learning
- Grounding
- Planning
- Tool Use
- Reasoning
- World Models
- Multimodal Perception
- Memory
- Embodiment

Environments

- Humans
- Agents
- Databases
- Web
- Physical World
- Computers
- APIs
- Documents

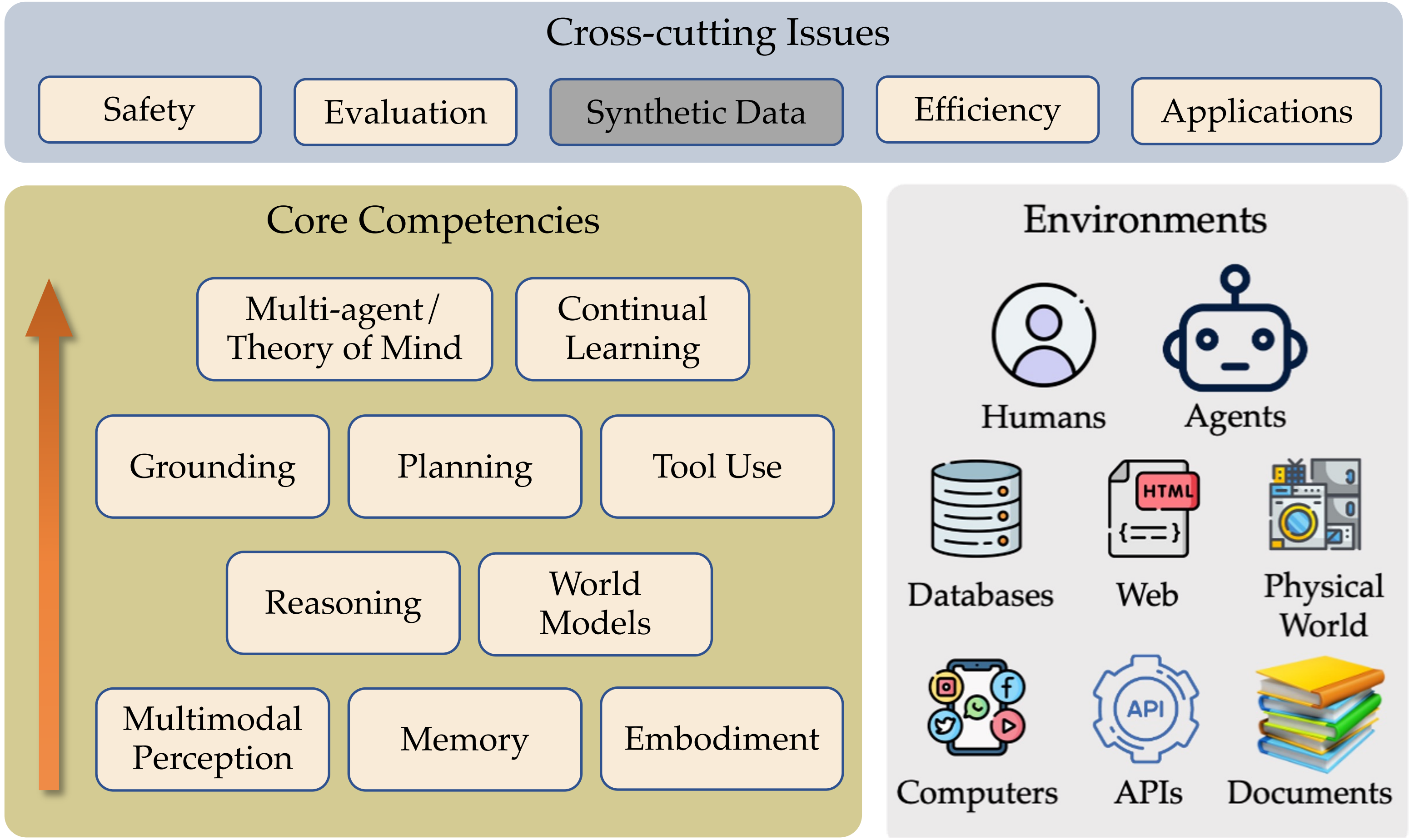
Future directions for language agents



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

Chen et al., *When is Tree Search Useful for LLM Planning? It Depends on the Discriminator*, ACL 2024
Gu et al., *Is Your LLM Secretly a World Model of the Internet? Model-based Planning for Web Agents*, arXiv 2024

Future directions for language agents



Cross-cutting Issues

Safety

Evaluation

Synthetic Data

Efficiency

Applications

Core Competencies

Multi-agent/ Theory of Mind

Continual Learning

Grounding

Planning

Tool Use

Reasoning

World Models

Multimodal Perception

Memory

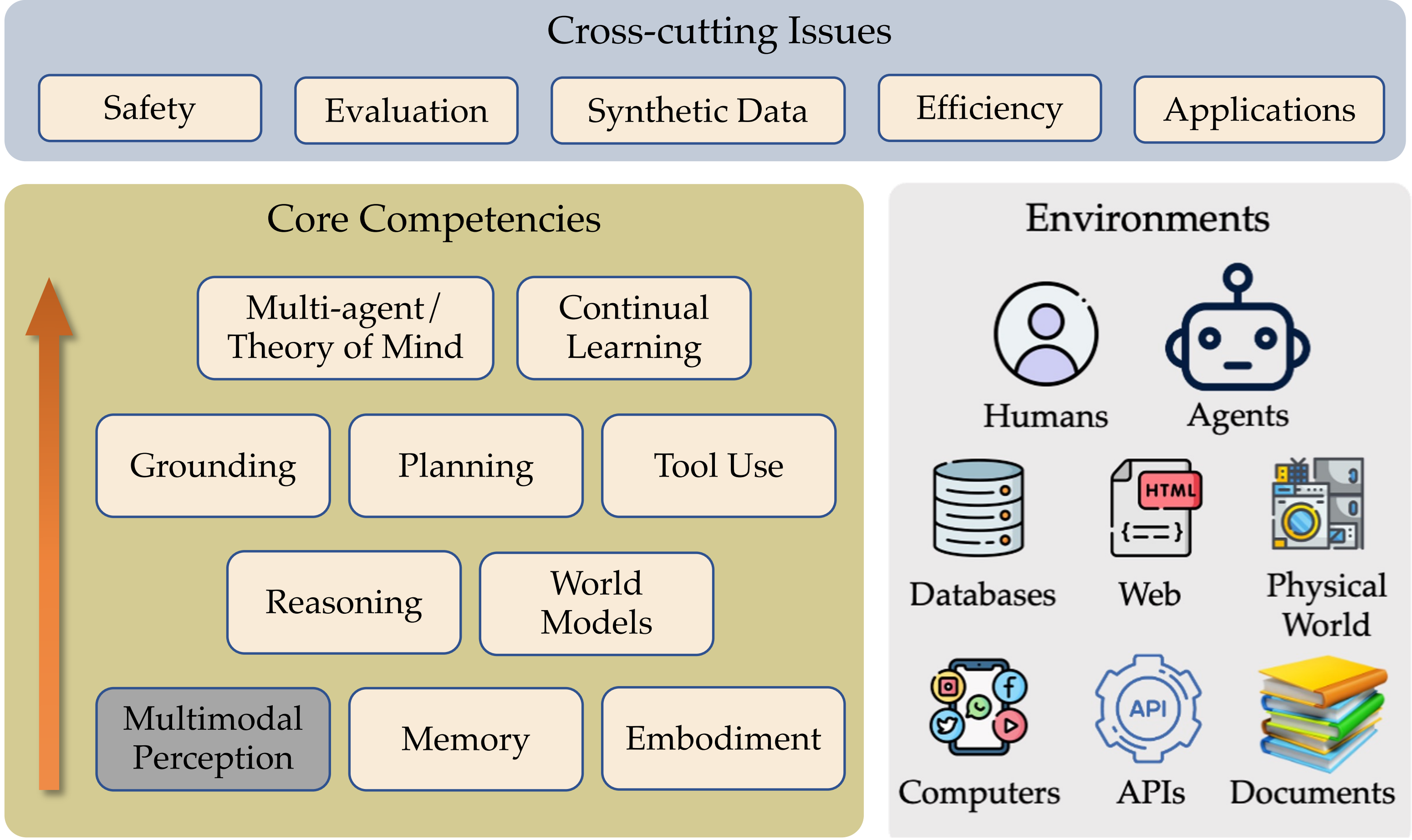
Embodiment

Environments

- 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

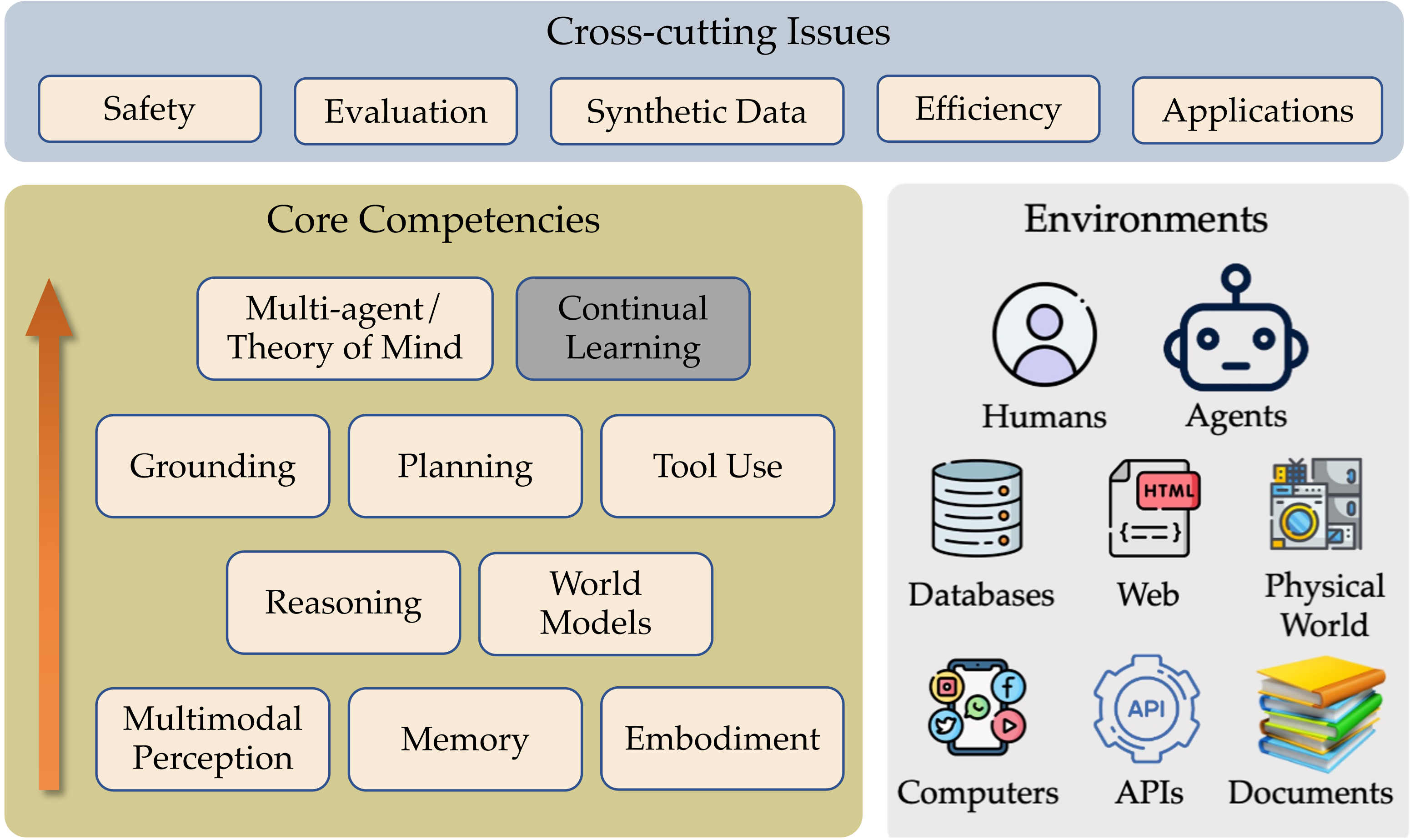
Wang et al., *LLMs in the Imaginarium: Tool Learning through Simulated Trial and Error*, ACL 2024
Gou et al., *Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents*, arXiv 2024

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



Cross-cutting Issues

Safety

Evaluation

Synthetic Data

Efficiency

Applications

Core Competencies

Multi-agent / Theory of Mind

Continual Learning

Grounding

Planning

Tool Use

Reasoning

World Models

Multimodal Perception

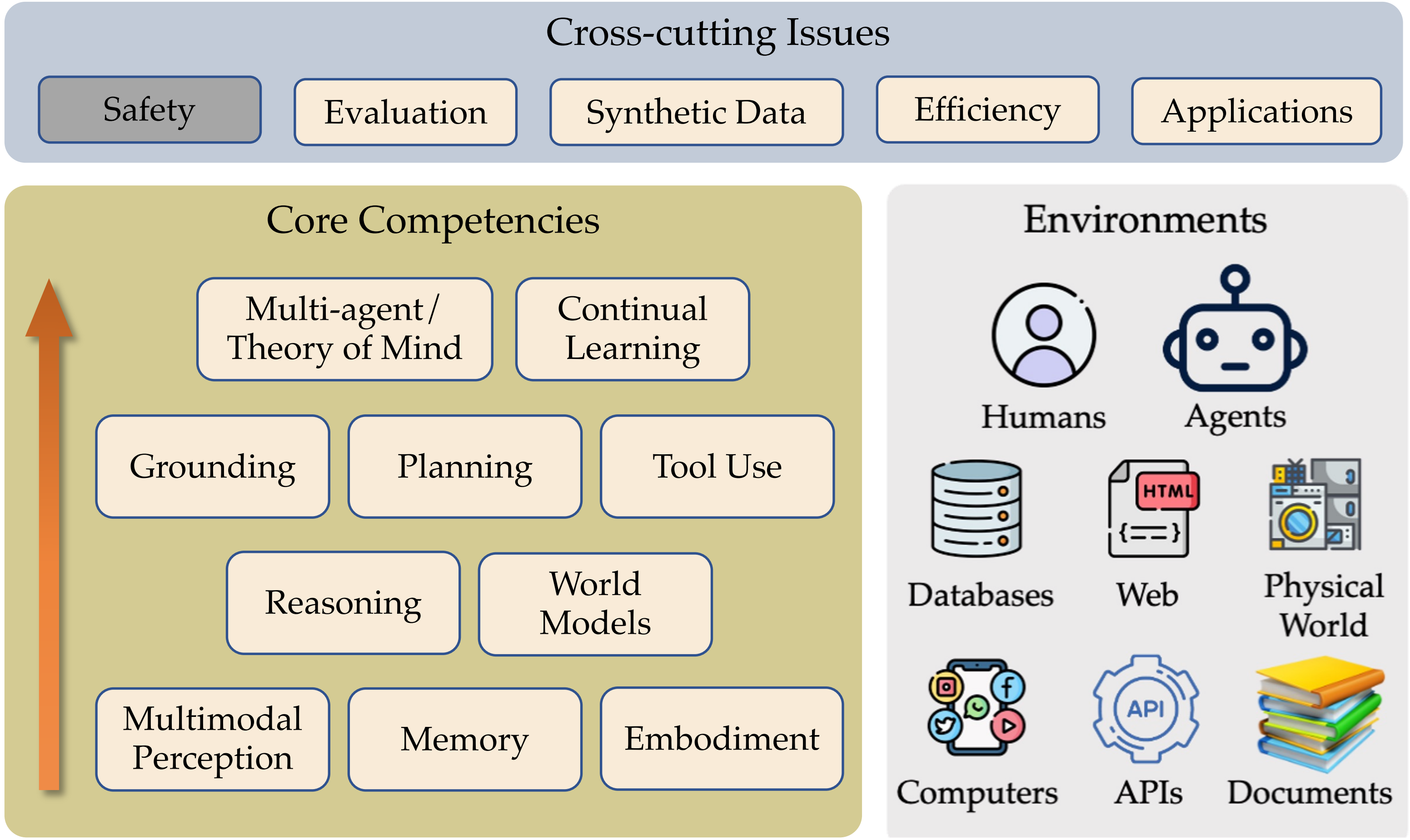
Memory

Embodiment

Environments

- 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 inherent 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*)

(Upcoming) EMNLP 2024 Tutorial

Language Agents: Foundations, Prospects, and Risks

Yu Su, Diyi Yang, Shunyu Yao, Tao Yu

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

OHIO STATE

Stanford University

PRINCETON UNIVERSITY

OpenAI

香 港 大 學
THE UNIVERSITY OF HONG KONG

Acknowledgements

NSF

NIH

DEVCOM
ARMY RESEARCH
LABORATORY

amazon

Imageomics
BIO
ML

ICICLE
DEMOCRATIZING AI

THE OHIO STATE UNIVERSITY
TRANSLATIONAL DATA
ANALYTICS INSTITUTE

cisco

rby AI

Apple

Walmart

Thanks &