The Intelligence Feedback Loop: From Biological Inspiration to Augmented Cognition

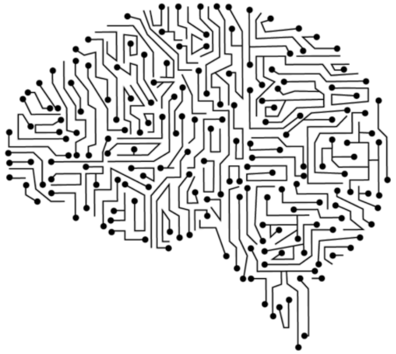
Yu Su

The Ohio State University

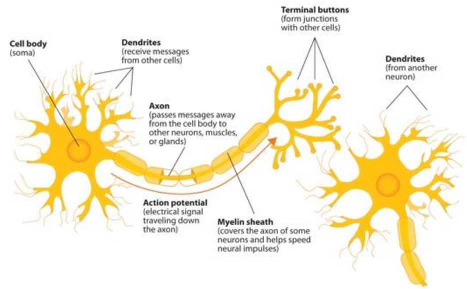
Outline

- The Intelligence Feedback Loop: Introduction
- Augmented Cognition: Computer Use Agents
- Biological Inspiration: Long-term Memory
- Future Directions

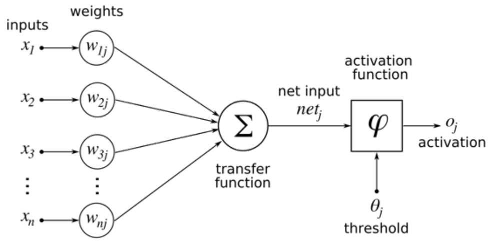
The intelligence feedback loop



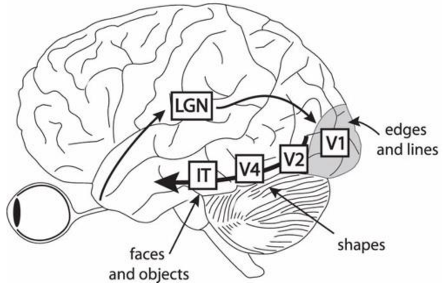
Biological inspiration in AI


Cell body
(soma)
Dendrites
(receive messages from other cells)
Terminal buttons
(form junctions with other cells)
Dendrites
(from another neuron)
Axon
(passes messages away from the cell body to other neurons, muscles, or glands)
Action potential
(electrical signal traveling down the axon)
Myelin sheath
(covers the axon of some neurons and helps speed neural impulses)

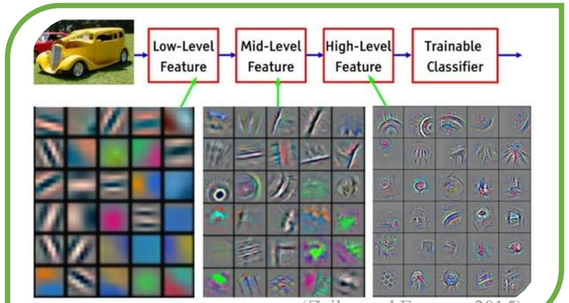
Neuron



McCulloch and Pitts (1943) developed the artificial neuron and showed that a network of such simple neurons can perform logical computations.



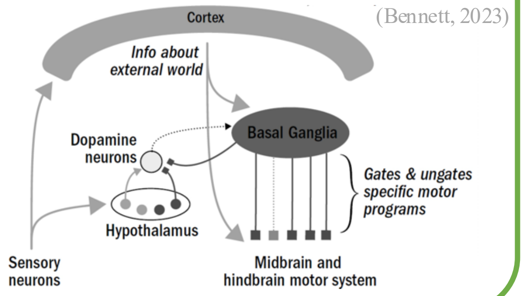
Hierarchical Representation Convolutional Neural Network



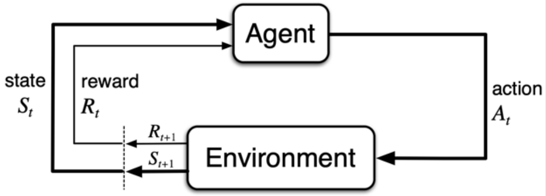
(Zeiler and Fergus., 2015)

Fukushima (1980) developed the convolutional neural network (Neocognitron) inspired by Hubel and Wiesel (1962); Lecun et al. (1989, 1998) extended it with backprop (among other changes).

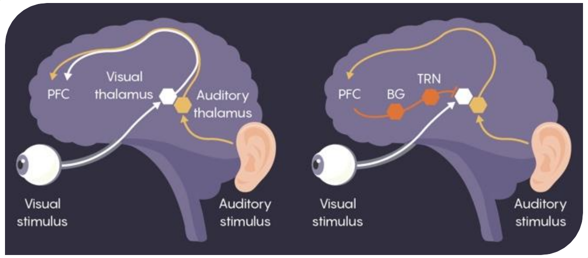
(Bennett, 2023)


Cortex
Info about external world
Basal Ganglia
Dopamine neurons
Hypothalamus
Gates & ungates specific motor programs
Sensory neurons
Midbrain and hindbrain motor system

Reinforcement Learning (Temporal Difference Learning)

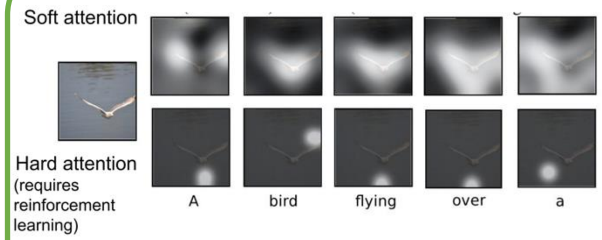


Sutton (1988) B.A. in psychology, formalized temporal difference learning. It inspired Schultz et al. (1997) to discover dopamine reward prediction errors in the brain.



quantamagazine.org

Selective Attention

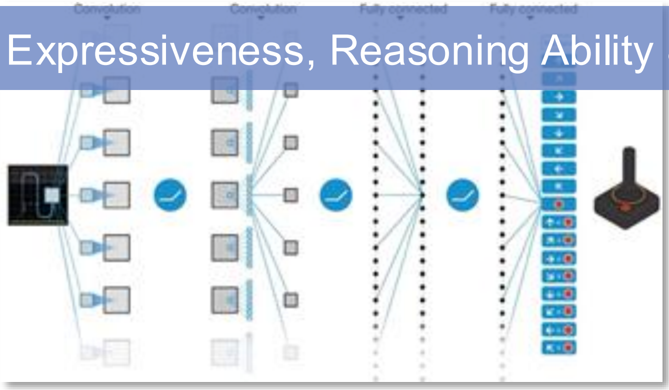


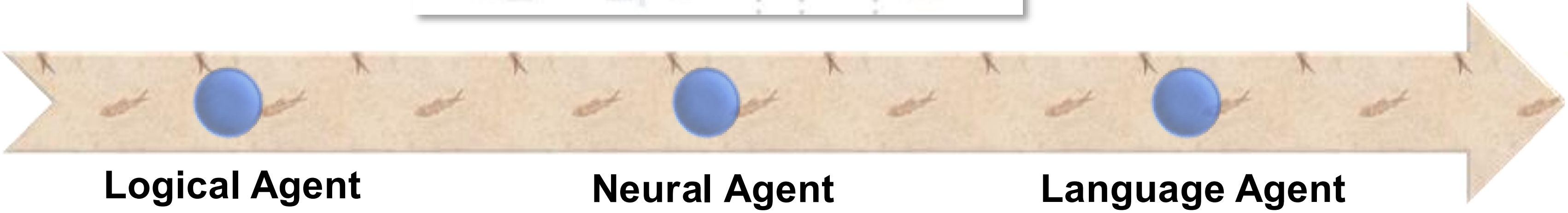
(Xu et al., 2015)

Bahdanau et al. (2015) developed attention in modern neural networks "*sort of inspired by translation exercises ... Your gaze shifts back and forth between source and target sequence as you translate.*"

A new evolutionary stage of machine intelligence

Increasing Expressiveness, Reasoning Ability & Adaptivity

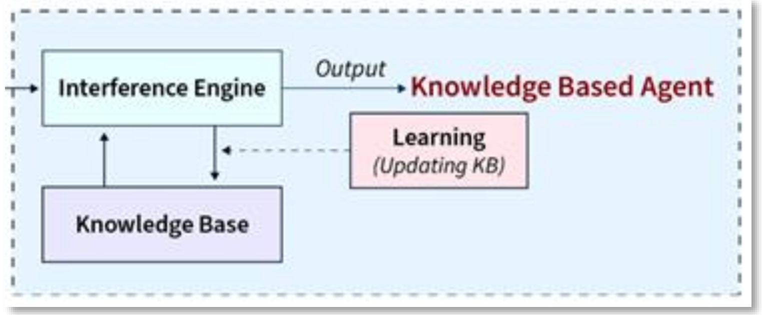




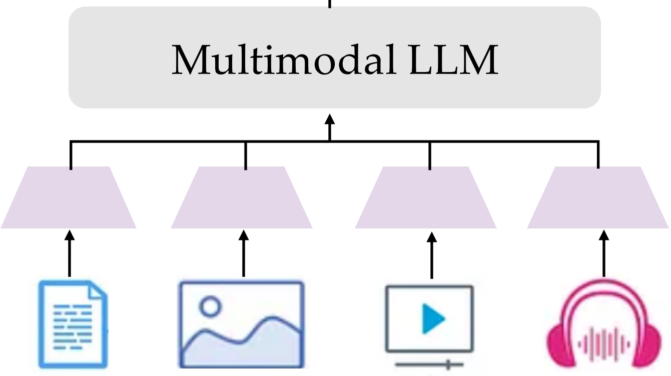
Logical Agent

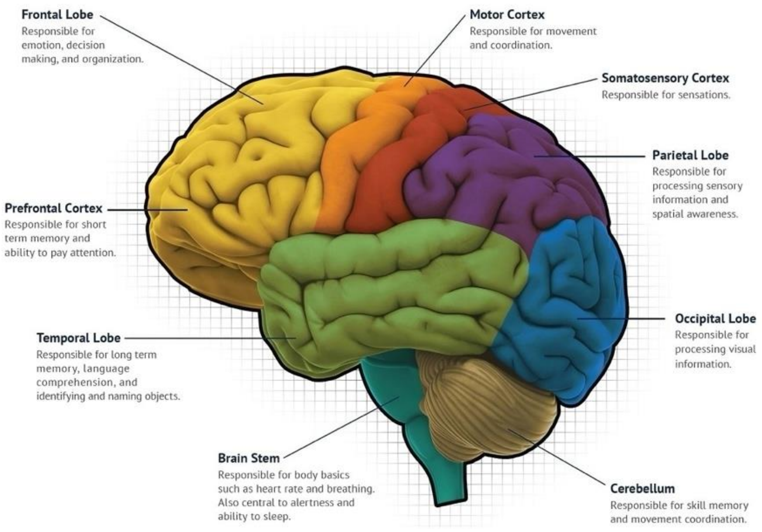
Neural Agent

Language Agent



"It looks like we are on the Amazon homepage. I'll search for 'foldable strollers' as requested."



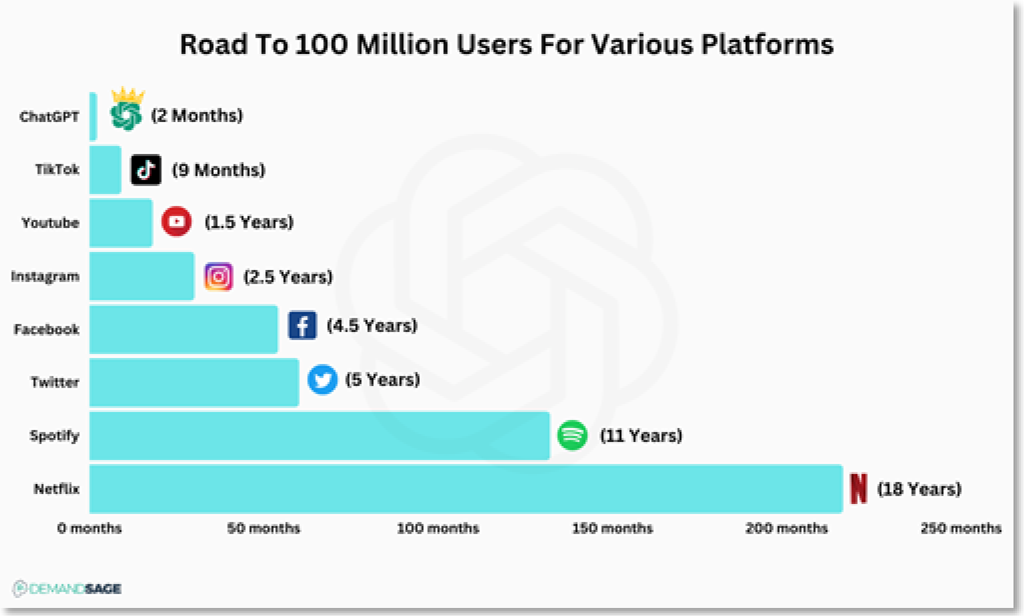

Frontal Lobe
Responsible for emotion, decision making, and organization.
Motor Cortex
Responsible for movement and coordination.
Somatosensory Cortex
Responsible for sensations.
Parietal Lobe
Responsible for processing sensory information and spatial awareness.
Prefrontal Cortex
Responsible for short term memory and ability to pay attention.
Occipital Lobe
Responsible for processing visual information.
Temporal Lobe
Responsible for long term memory, language comprehension, and identifying and naming objects.
Brain Stem
Responsible for body basics such as heart rate and breathing. Also central to alertness and ability to sleep.
Cerebellum
Responsible for skill memory and movement coordination.

Language agents: a new generation of AI agents

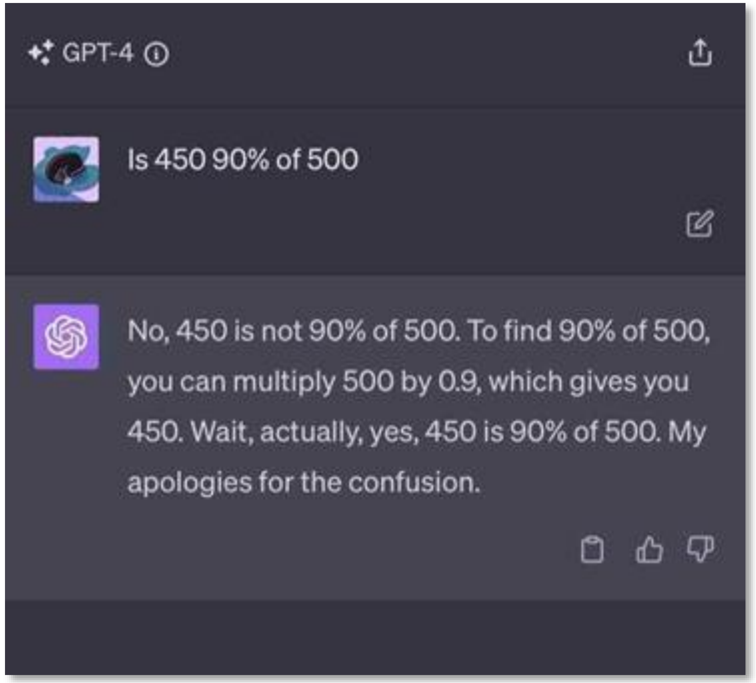
Contemporary AI agents, with integrated LLM(s), can *use language as a vehicle for reasoning and communication*

⬆ Instruction following, in-context learning, output customization

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



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

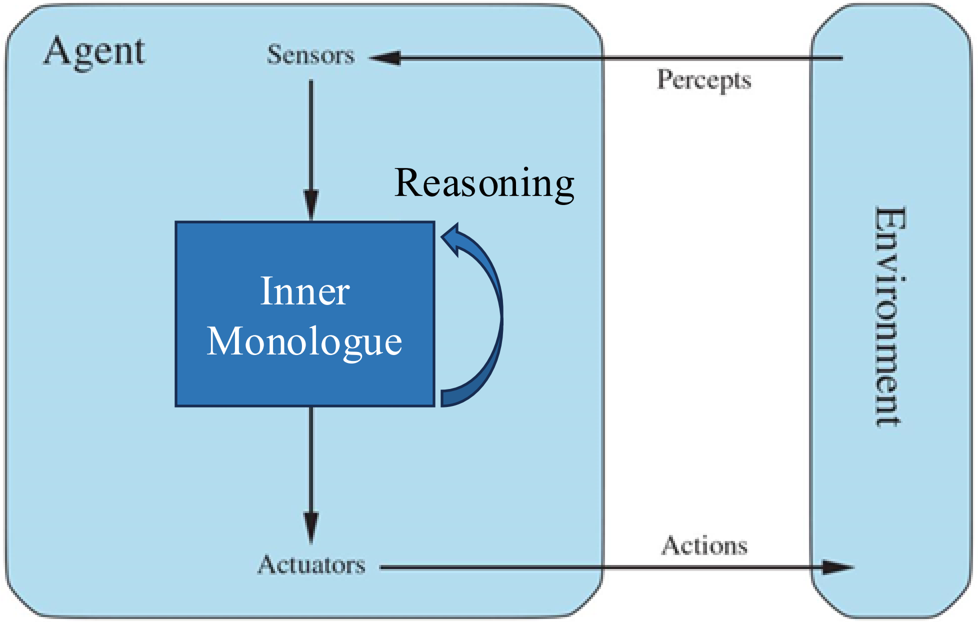


Is 450 90% of 500

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/

Reconciling with the classic view of agents



Adapted from Russell and Norvig (2020)

- 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 multimodal perception and using language for communication

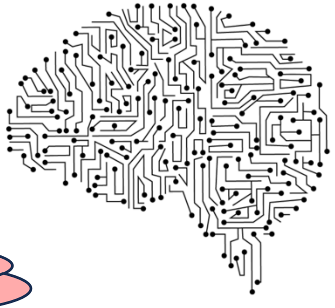
Language agents augment human cognition

A typical human (per minute)

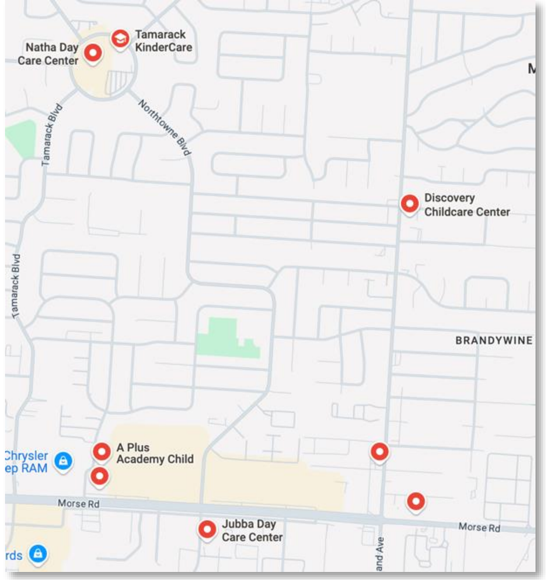
- reads 250 words
- thinks 400 words (inner monologue)
- clicks 40 times meaningfully
- types 55 words

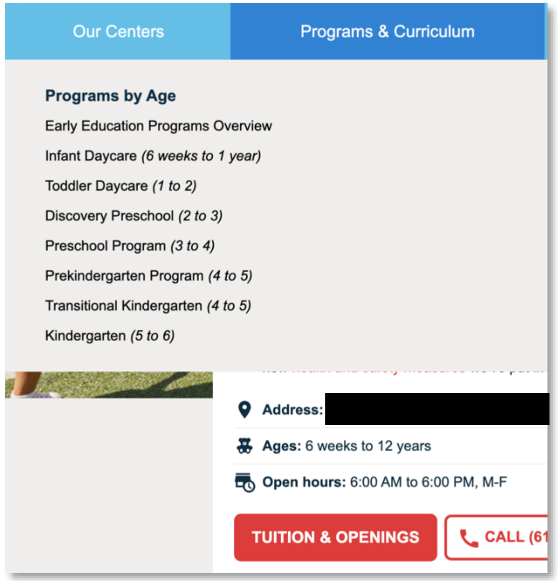
Which daycares within a 10-minute drive from my home can take a 1-year-old and provide meals?

Great! Book a tour for me!



Search → Read → Reason → Act

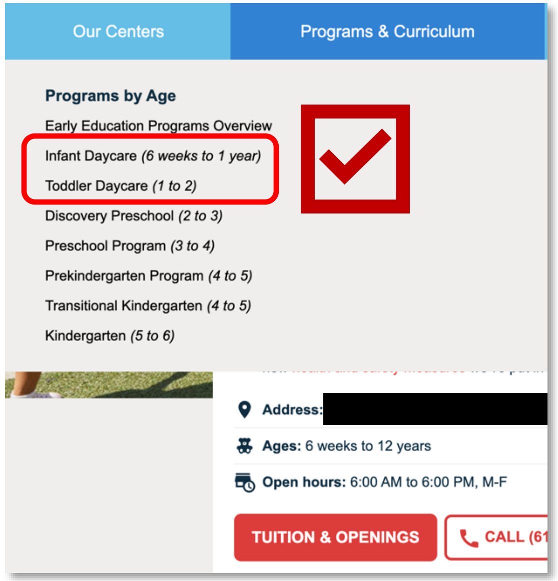




Programs by Age

Early Education Programs Overview
Infant Daycare *(6 weeks to 1 year)*
Toddler Daycare *(1 to 2)*
Discovery Preschool *(2 to 3)*
Preschool Program *(3 to 4)*
Prekindergarten Program *(4 to 5)*
Transitional Kindergarten *(4 to 5)*
Kindergarten *(5 to 6)*

Address: [redacted]
Ages: 6 weeks to 12 years
Open hours: 6:00 AM to 6:00 PM, M-F

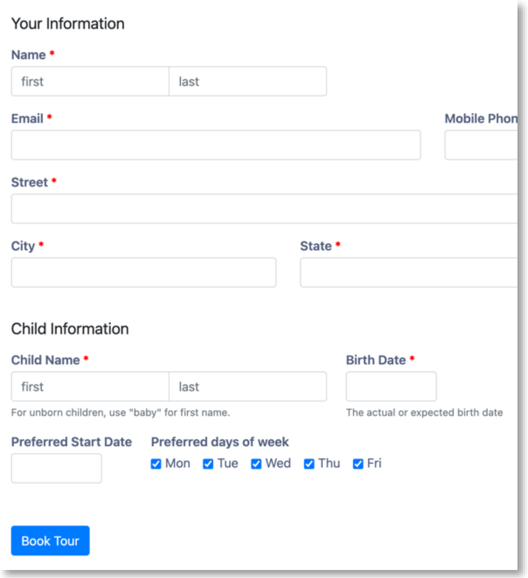


Programs by Age

Early Education Programs Overview
Infant Daycare *(6 weeks to 1 year)*
Toddler Daycare *(1 to 2)*
Discovery Preschool *(2 to 3)*
Preschool Program *(3 to 4)*
Prekindergarten Program *(4 to 5)*
Transitional Kindergarten *(4 to 5)*
Kindergarten *(5 to 6)*

✓

Address: [redacted]
Ages: 6 weeks to 12 years
Open hours: 6:00 AM to 6:00 PM, M-F



Your Information

Name *
first | last

Email *
Mobile Phon

Street *

City *
State *

Child Information

Child Name *
first | last
For unborn children, use "baby" for first name.

Birth Date *
The actual or expected birth date

Preferred Start Date

Preferred days of week
☑ Mon ☑ Tue ☑ Wed ☑ Thu ☑ Fri

Outline

- The Intelligence Feedback Loop: Introduction
- Augmented Cognition: Computer Use Agents
- Biological Inspiration: Long-term Memory
- Future Directions

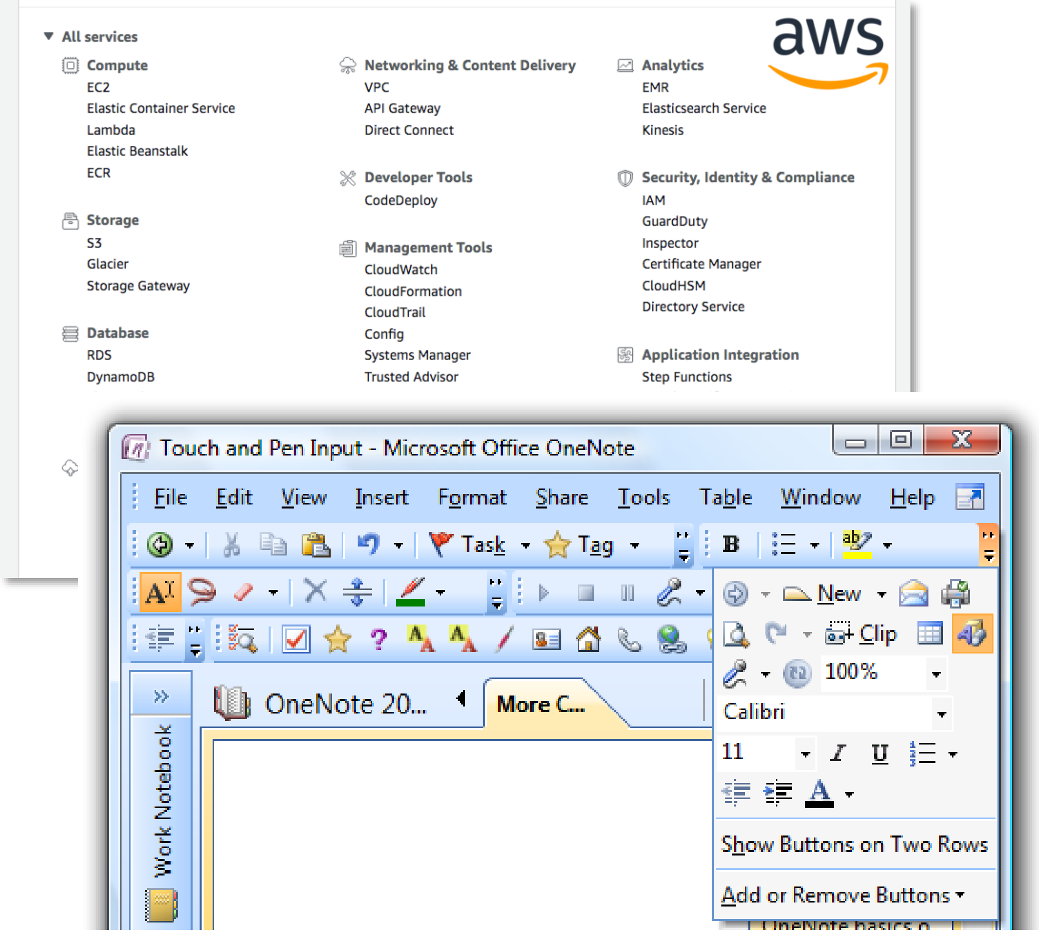
Increasing complexity of the digital world

Digital world overtaking physical world?

- **92% of jobs** require some type of digital skills
- An average user spends **6+ hours** online per day

Complexity exceeds cognitive capacity

- **1.2 billion websites**, **7 million mobile apps**
- **57% of employees** state that difficulty finding the right information is a top contributor to lagging productivity
- Information overload costs the U.S. economy **$900 billion** a year



All services

Compute
EC2
Elastic Container Service
Lambda
Elastic Beanstalk
ECR

Storage
S3
Glacier
Storage Gateway

Database
RDS
DynamoDB

Networking & Content Delivery
VPC
API Gateway
Direct Connect

Developer Tools
CodeDeploy

Management Tools
CloudWatch
CloudFormation
CloudTrail
Config
Systems Manager
Trusted Advisor

Analytics
EMR
Elasticsearch Service
Kinesis

Security, Identity & Compliance
IAM
GuardDuty
Inspector
Certificate Manager
CloudHSM
Directory Service

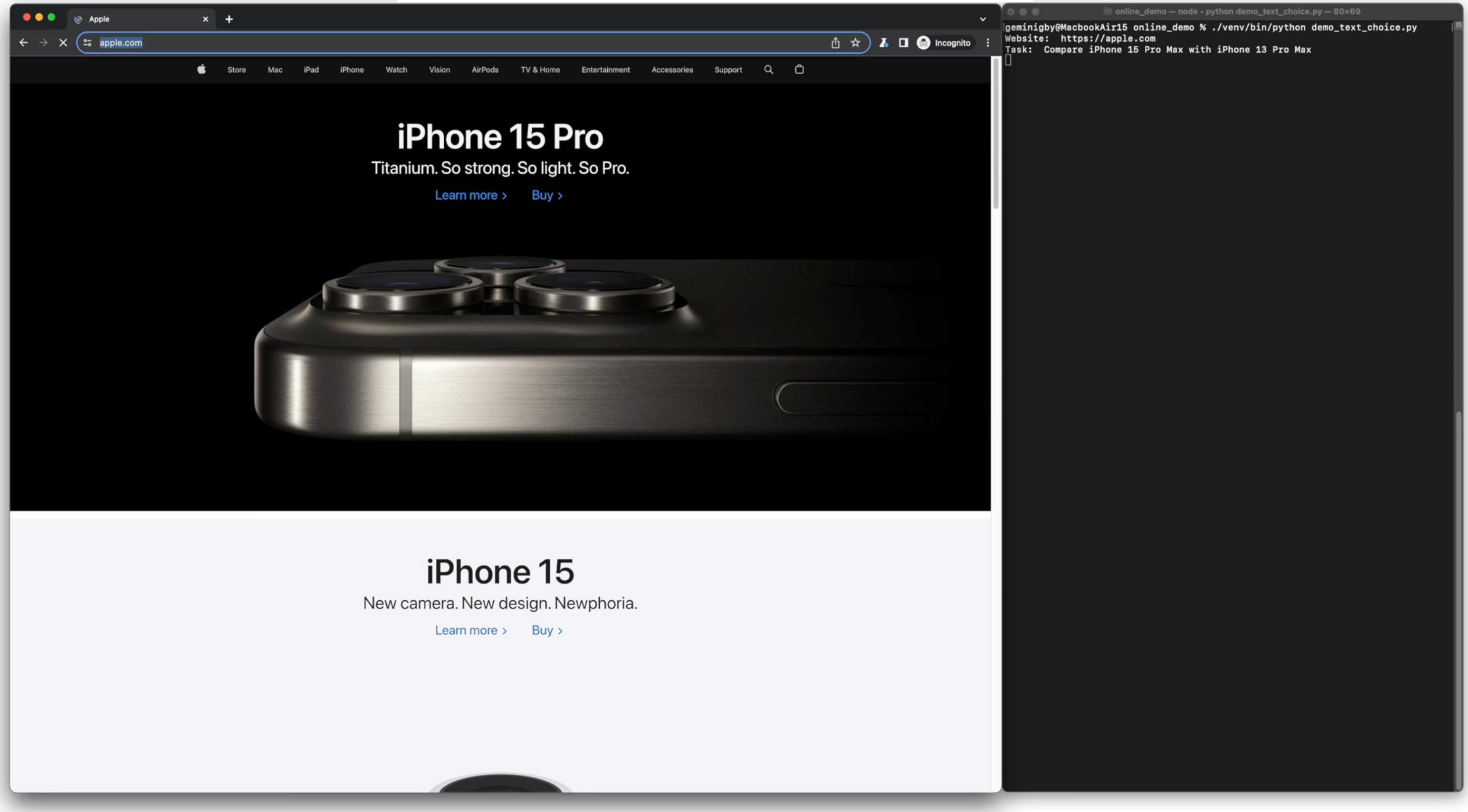
Application Integration
Step Functions

[1] https://nationalskillscoalition.org/resource/publications/closing-the-digital-skill-divide
[2] https://datareportal.com/global-digital-overview
[3] https://www.nasdaq.com/press-release/over-50-percent-of-knowledge-workers-cannot-find-the-information-they-need-at-work
[4] https://hbr.org/2009/09/death-by-information-overload

Agents that operate in the digital world

SeeAct

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



iPhone 15 Pro

Titanium. So strong. So light. So Pro.

Learn more › Buy ›

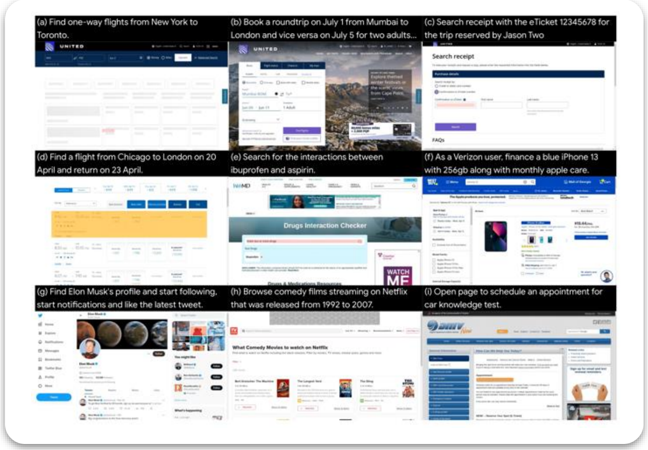
iPhone 15

New camera. New design. Newphoria.

Learn more › Buy ›

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

My changelog of computer use agents



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

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

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

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

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

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

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

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

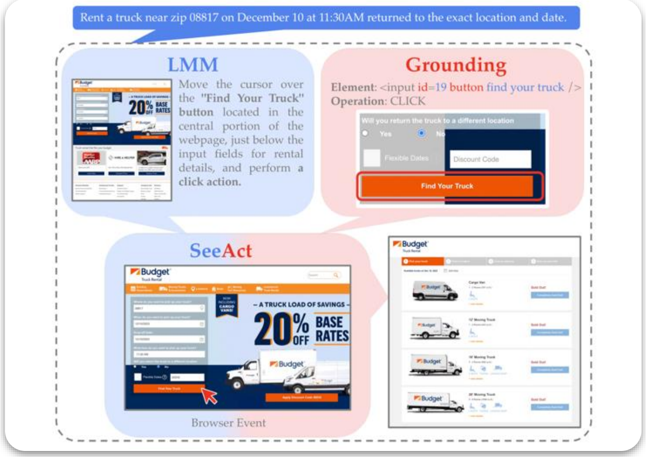
(i) Open page to schedule an appointment for car knowledge test.

Mind2Web [NeurIPS'23 Spotlight]

First LLM-based web agent
Ecologically valid eval

SeeAct [ICML'24]

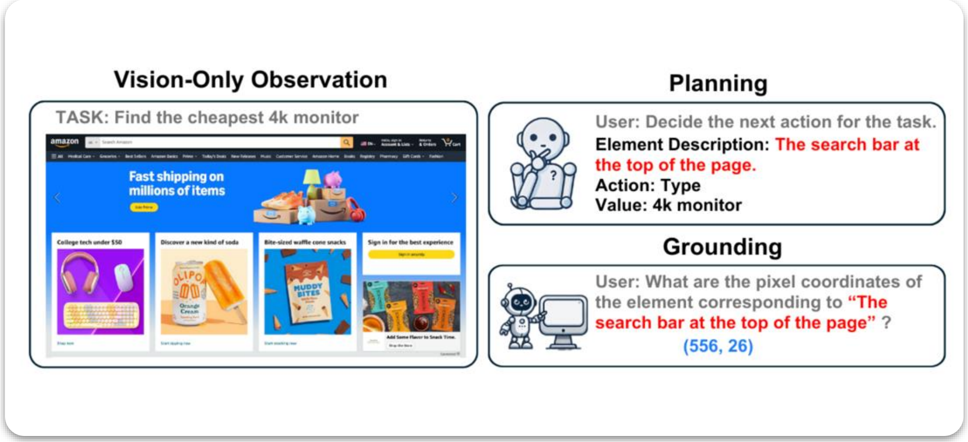
First generalist web agent with
visual perception



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

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

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



Vision-Only Observation

TASK: Find the cheapest 4k monitor

Planning

User: Decide the next action for the task.
Element Description: The search bar at the top of the page.
Action: Type
Value: 4k monitor

Grounding

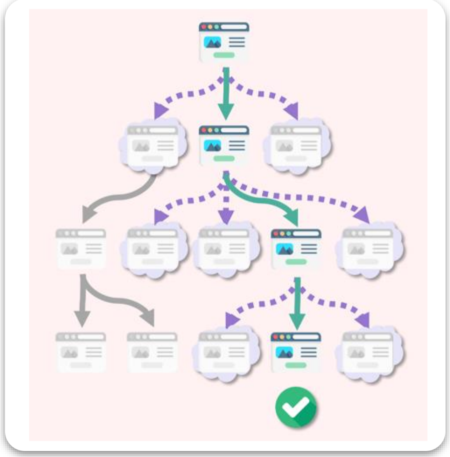
User: What are the pixel coordinates of the element corresponding to "The search bar at the top of the page" ?
(556, 26)

UGround [ICLR'25 Oral]

Pure vision-based agent
Human-like embodiment

World models
Model-based planning

WebDreamer



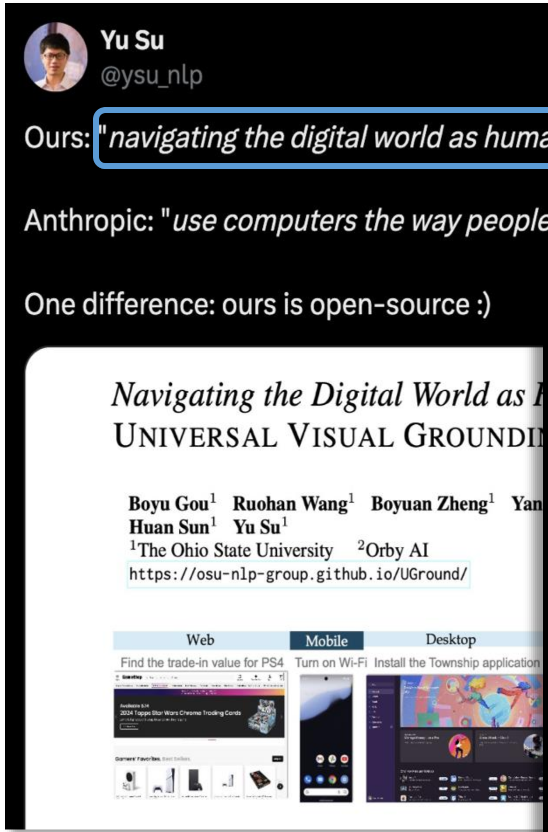
Agentic search
Continual learning
Safety

ICLR 2025 Oral

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

What is "*the way humans do?*"
And *why*?



Yu Su
@ysu_nlp

Ours: "*navigating the digital world as humans do*"

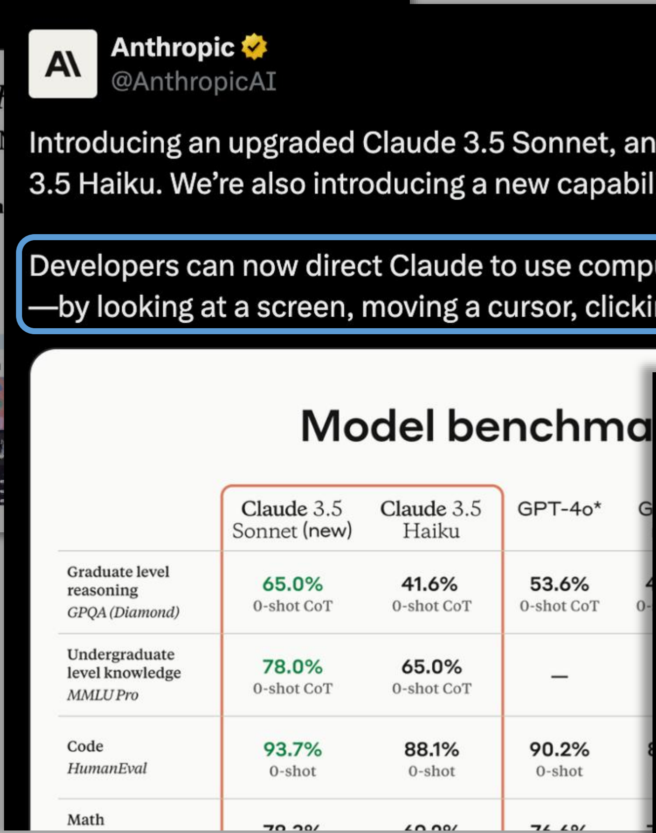
Anthropic: "*use computers the way people do*"

One difference: ours is open-source :)

Navigating the Digital World as H
UNIVERSAL VISUAL GROUNDI

Boyu Gou[1] Ruohan Wang[1] Boyuan Zheng[1] Yan
Huan Sun[1] Yu Su[1]
[1]The Ohio State University [2]Orby AI
https://osu-nlp-group.github.io/UGround/

Web
Mobile
Desktop
Find the trade-in value for PS4
Turn on Wi-Fi
Install the Township application



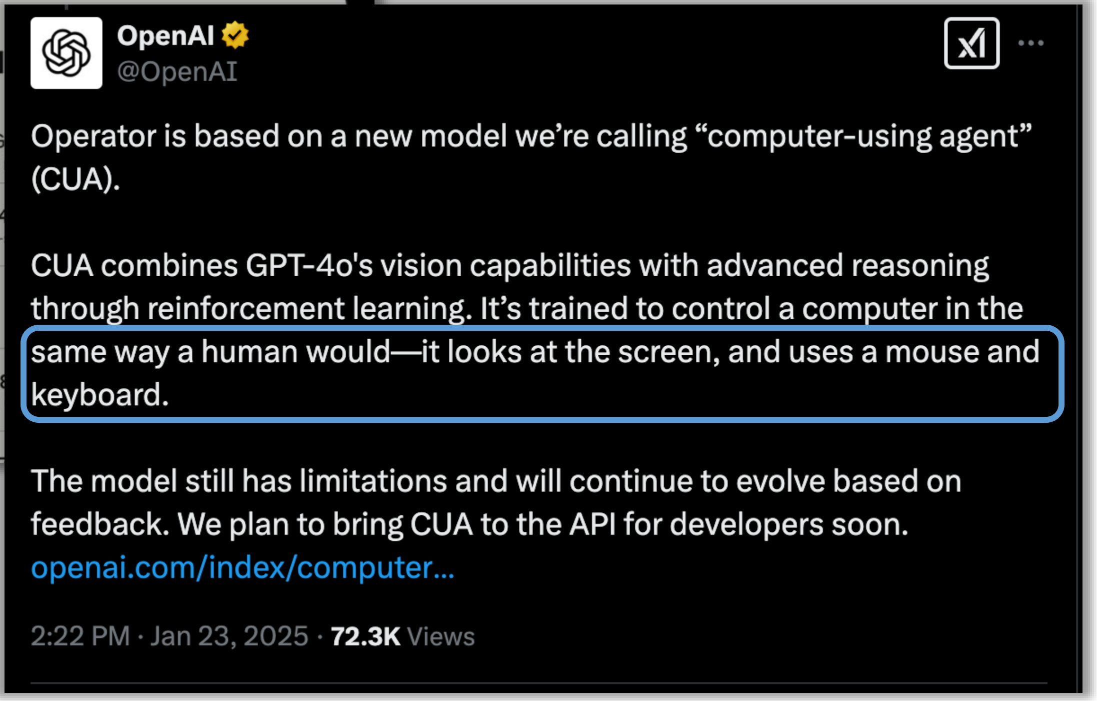
Anthropic
@AnthropicAI

Introducing an upgraded Claude 3.5 Sonnet, and a new model, Claude 3.5 Haiku. We're also introducing a new capability in beta: computer use.

Developers can now direct Claude to use computers the way people do—by looking at a screen, moving a cursor, clicking, and typing text.

Model benchma

	Claude 3.5 Sonnet (new)	Claude 3.5 Haiku	GPT-4o*	G
Graduate level reasoning *GPQA (Diamond)*	65.0% 0-shot CoT	41.6% 0-shot CoT	53.6% 0-shot CoT	4 0-
Undergraduate level knowledge *MMLU Pro*	78.0% 0-shot CoT	65.0% 0-shot CoT	—	
Code *HumanEval*	93.7% 0-shot	88.1% 0-shot	90.2% 0-shot	8
Math	78.3%	69.2%	76.6%	



OpenAI
@OpenAI

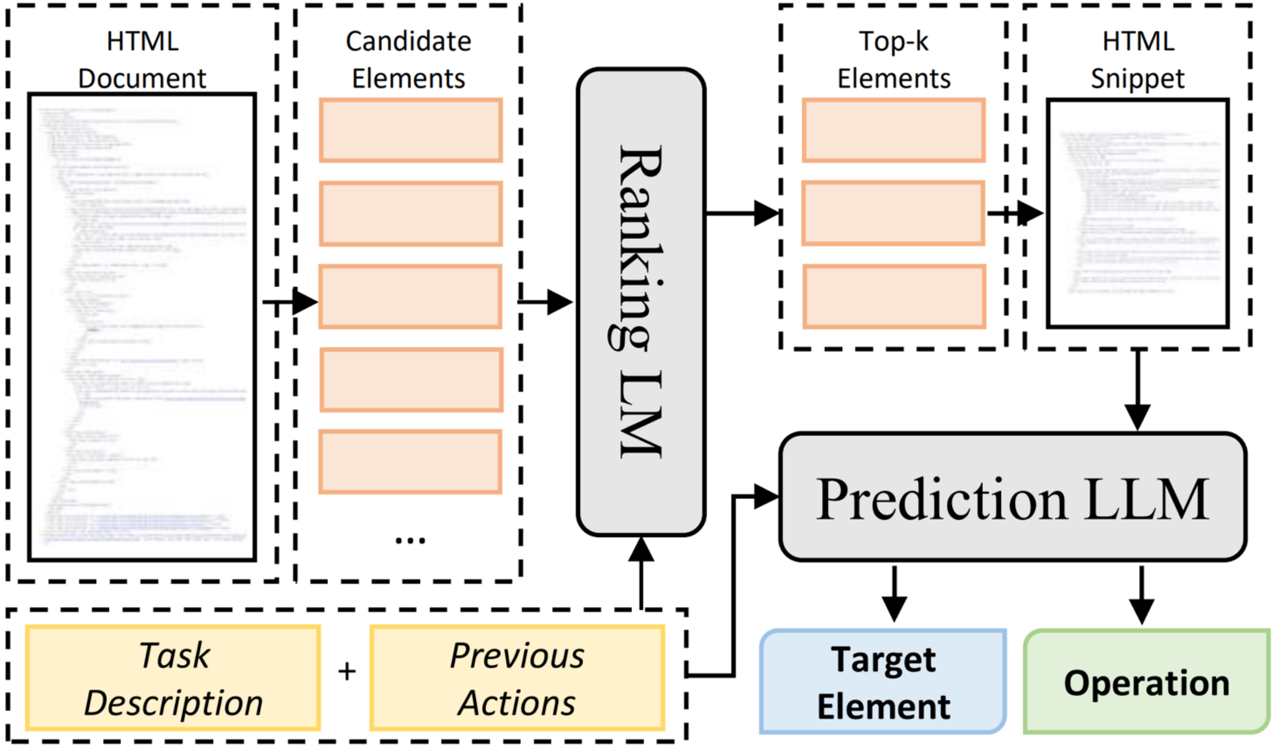
Operator is based on a new model we're calling "computer-using agent" (CUA).

CUA combines GPT-4o's vision capabilities with advanced reasoning through reinforcement learning. It's trained to control a computer in the same way a human would—it looks at the screen, and uses a mouse and keyboard.

The model still has limitations and will continue to evolve based on feedback. We plan to bring CUA to the API for developers soon.
openai.com/index/computer...

2:22 PM · Jan 23, 2025 · 72.3K Views

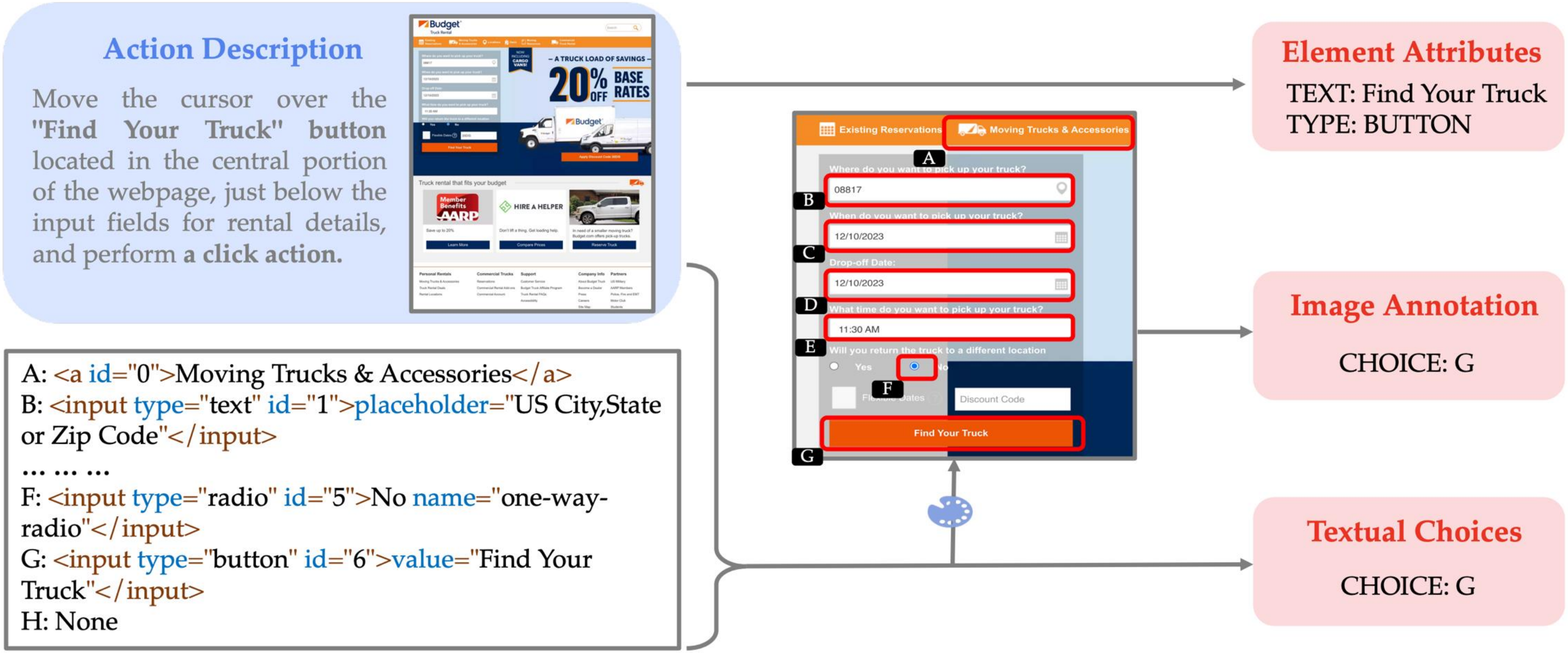
Embodiment of computer use agents: evolution



	Mind2Web	
Sensory Inputs	HTML/DOM	
Effectors	Multi-choice Selection	

[NeurIPS'23]

Embodiment of computer use 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?
12/10/2023
C
Drop-off Date:
12/10/2023
D
What time do you want to pick up your truck?
11:30 AM
E
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

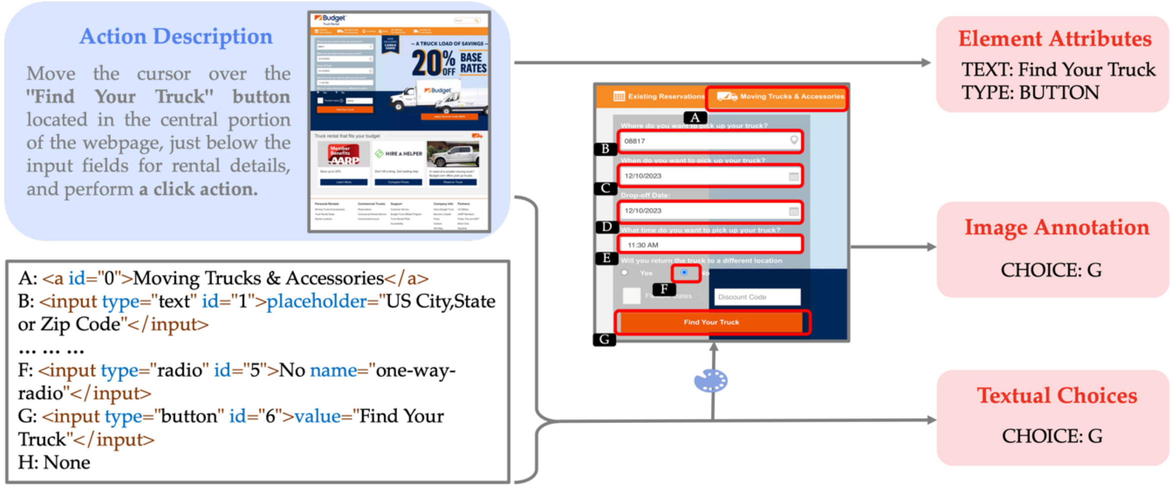
CHOICE: G

Textual Choices

CHOICE: G

	Mind2Web	SeeAct
Sensory Inputs	HTML/DOM	Screenshot + DOM
Effectors	Multi-choice Selection	Multi-choice Selection
	[NeurIPS’23]	[ICML’24]

Embodiment of computer use 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.
Element Attributes
TEXT: Find Your Truck
TYPE: BUTTON
Image Annotation
CHOICE: G
Textual Choices
CHOICE: G
A
B
C
D
E
F
G
Find Your Truck

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

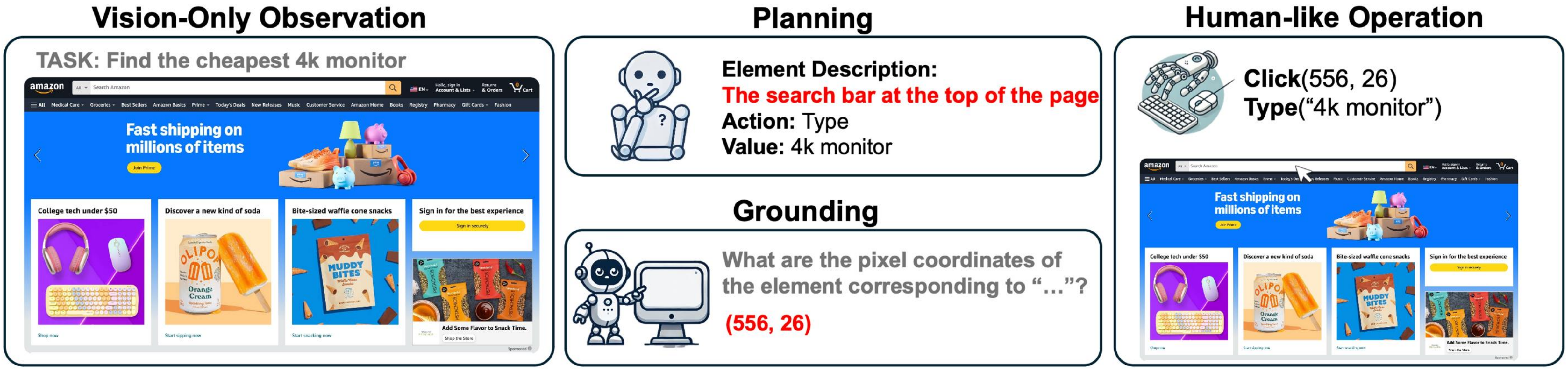
- *Noisy and/or incomplete*
    - **95.9%** of home pages have accessibility conformance errors
    - Avg. **56.8** errors per page[1]
- *Additional input increases latency and inference costs*
    - Consuming more tokens
    - Difficult and time-consuming to get
    - Compounding over long horizon

	Mind2Web	SeeAct	
Sensory Inputs	HTML/DOM	Screenshot + DOM	
Effectors	Multi-choice Selection	Multi-choice Selection	
	[NeurIPS'23]	[ICML'24]	

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

Embodiment of computer use agents: evolution

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



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

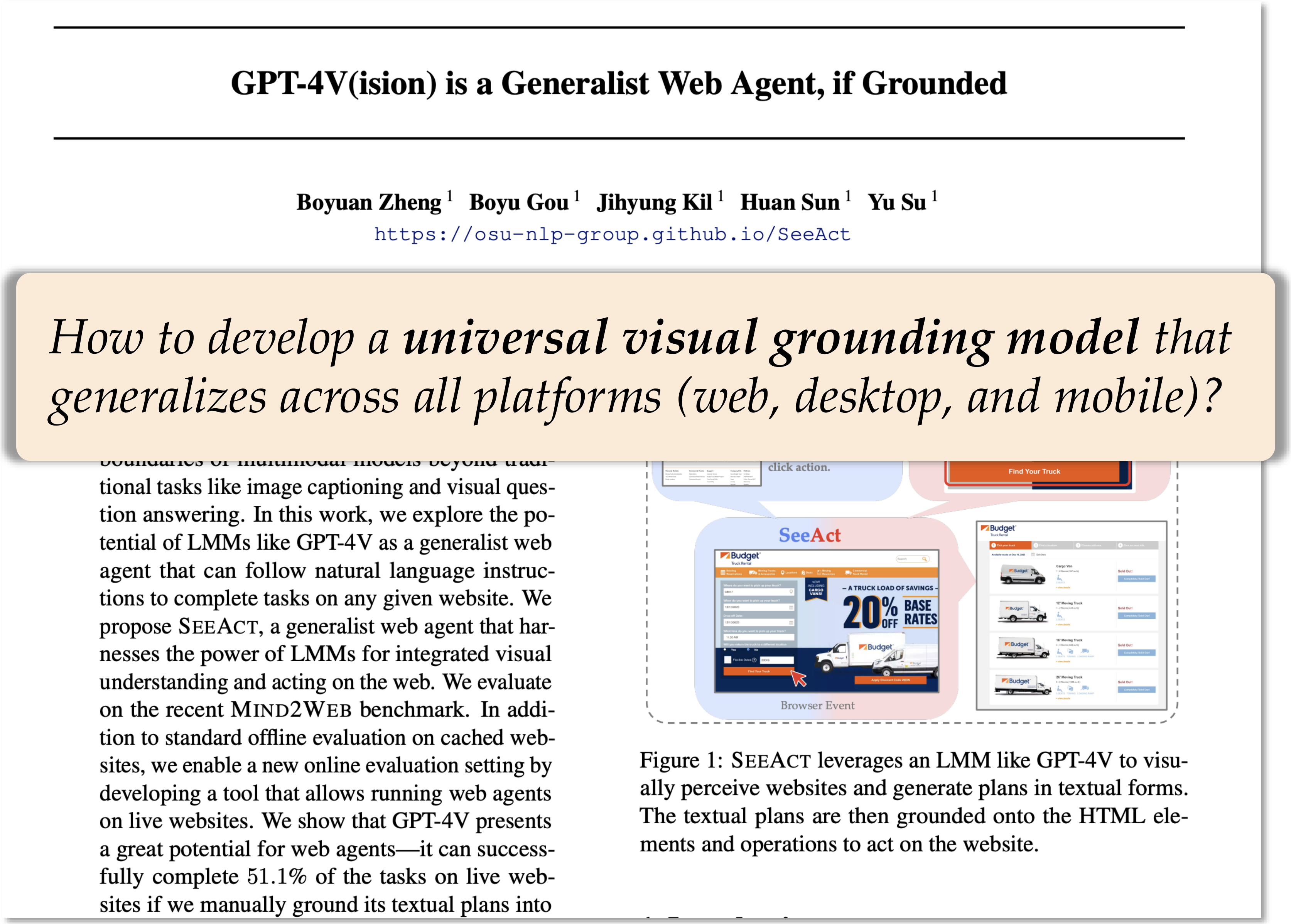
Click(556, 26)

Type("4k monitor")

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

	Mind2Web	SeeAct	SeeAct-V
Sensory Inputs	HTML/DOM	Screenshot + DOM	Screenshot Only
Effectors	Multi-choice Selection	Multi-choice Selection	Pixel-level Operations
	[NeurIPS'23]	[ICML'24]	[ICLR'25]

Visual grounding was 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

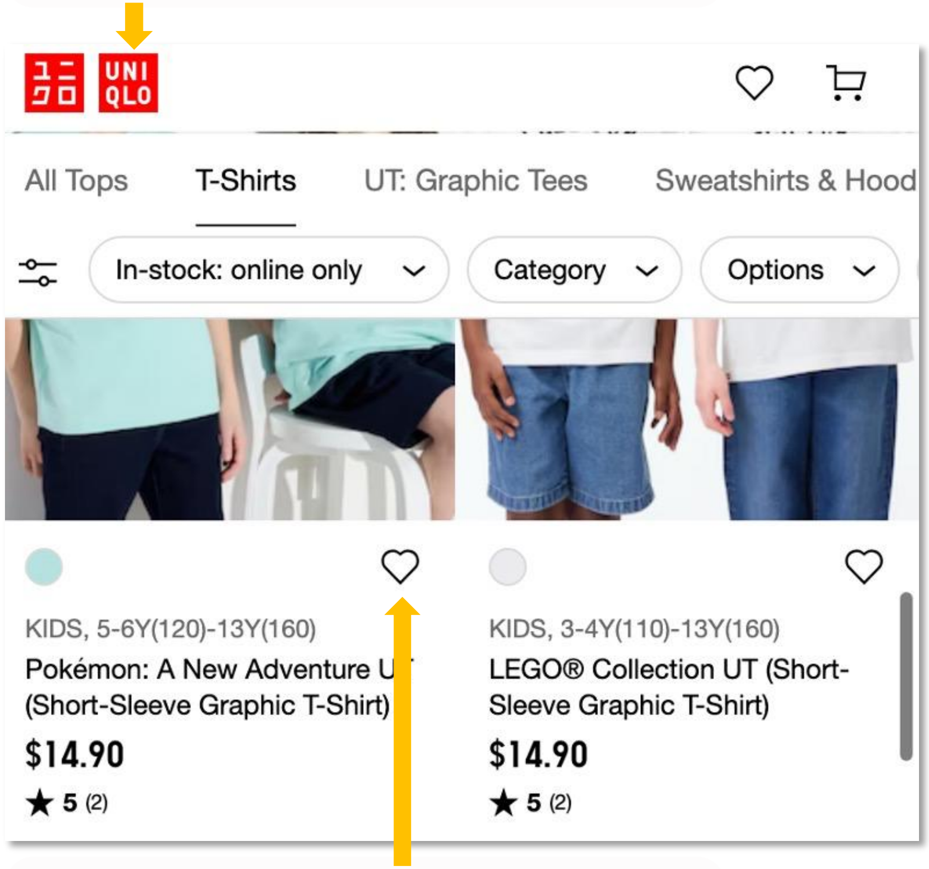
boundaries of multimodal models beyond traditional tasks like image captioning and visual question answering. In this work, we explore the potential of LMMs like GPT-4V as a generalist web agent that can follow natural language instructions to complete tasks on any given website. We propose SEEACT, a 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.

*How to develop a **universal visual grounding model** that generalizes across all platforms (web, desktop, and mobile)?*

Referring expressions for GUIs are diverse

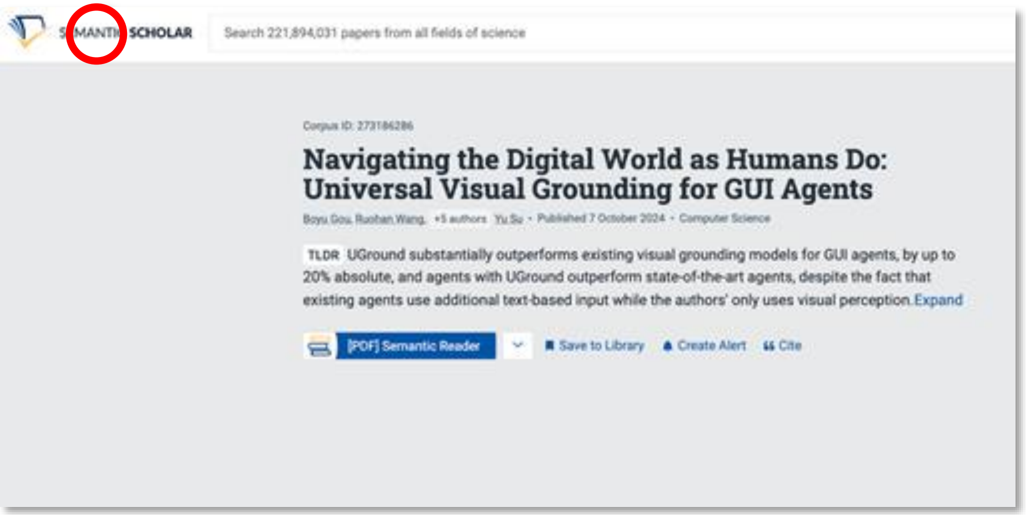
1. Red icon labeled “UNIQLO”
2. Button at the top left corner
3. Navigate back to the homepage



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



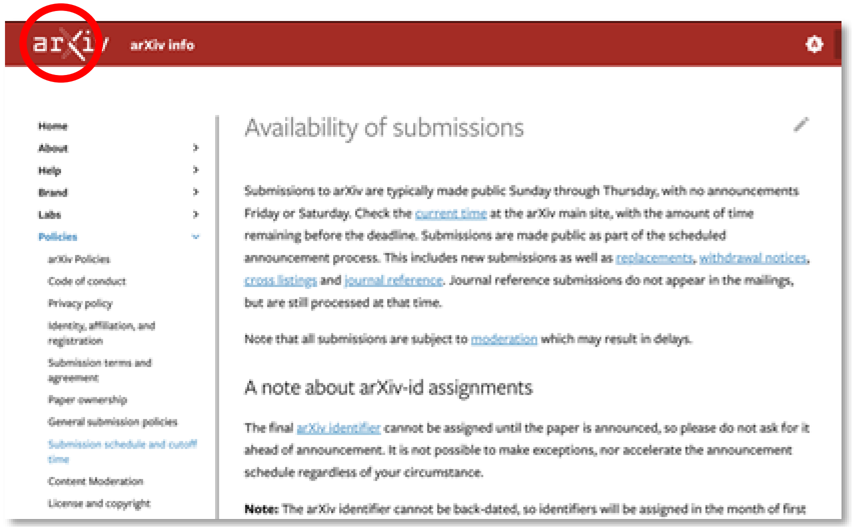
Corpus ID: 273186286

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.

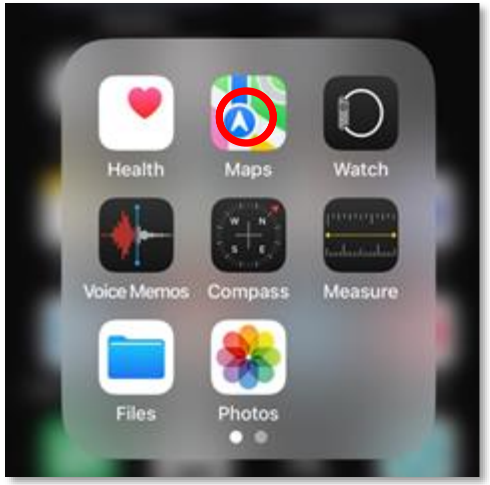
Note that all submissions are subject to moderation which may result in delays.

A note about arXiv-id assignments

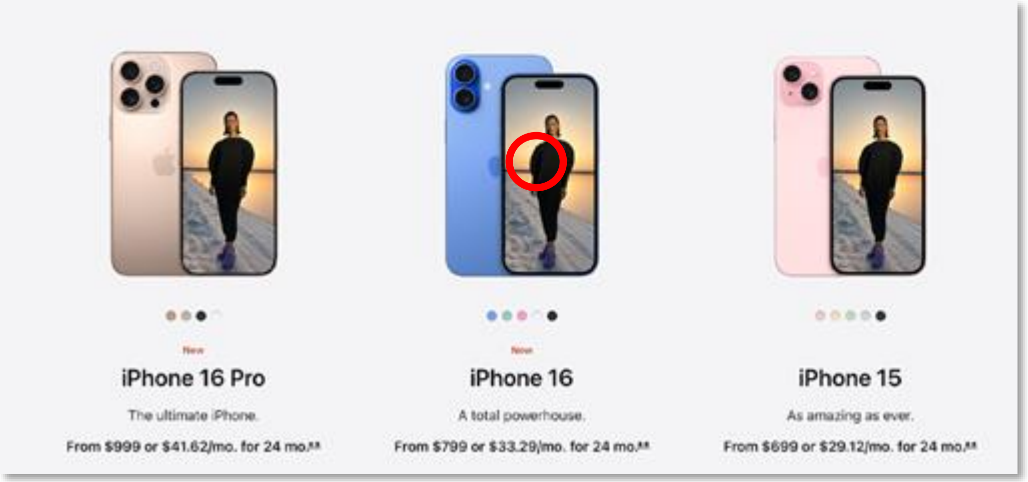
The final arXiv identifier cannot be assigned until the paper is announced, so please do not ask for it ahead of announcement. It is not possible to make exceptions, nor accelerate the announcement schedule regardless of your circumstance.

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

"Go to homepage"



"Open Maps"

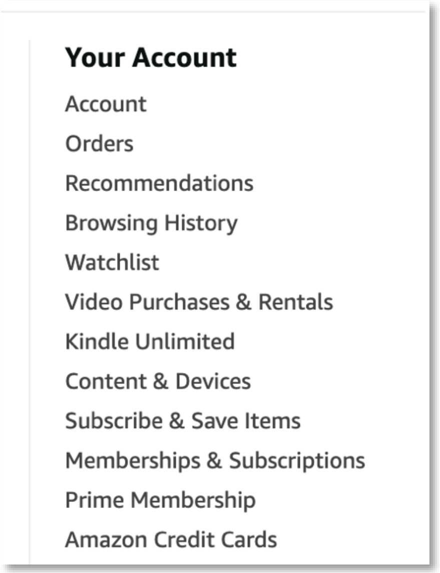


New
iPhone 16 Pro
The ultimate iPhone.
From $999 or $41.62/mo. for 24 mo.**
New
iPhone 16
A total powerhouse.
From $799 or $33.29/mo. for 24 mo.**
iPhone 15
As amazing as ever.
From $699 or $29.12/mo. for 24 mo.**

"iPhone 16"

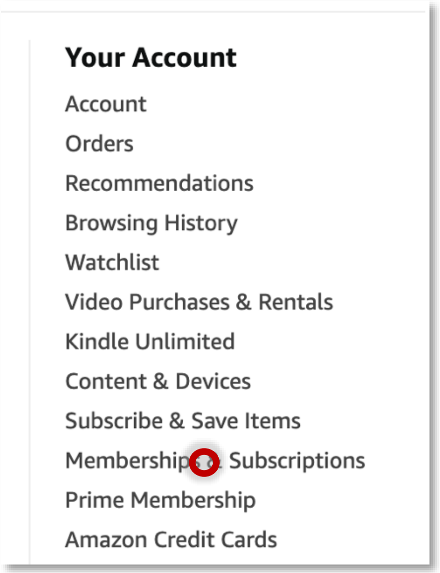
Synthetic data is key for agent learning

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



Your Account
- Account
- Orders
- Recommendations
- Browsing History
- Watchlist
- Video Purchases & Rentals
- Kindle Unlimited
- Content & Devices
- Subscribe & Save Items
- Memberships & Subscriptions
- Prime Membership
- Amazon Credit Cards

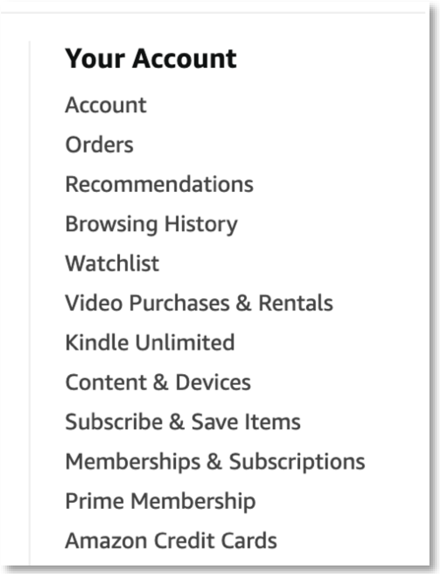
I want to renew my Prime membership, so I should click “Memberships & Subscriptions”



Your Account
- Account
- Orders
- Recommendations
- Browsing History
- Watchlist
- Video Purchases & Rentals
- Kindle Unlimited
- Content & Devices
- Subscribe & Save Items
- Memberships & Subscriptions
- Prime Membership
- Amazon Credit Cards

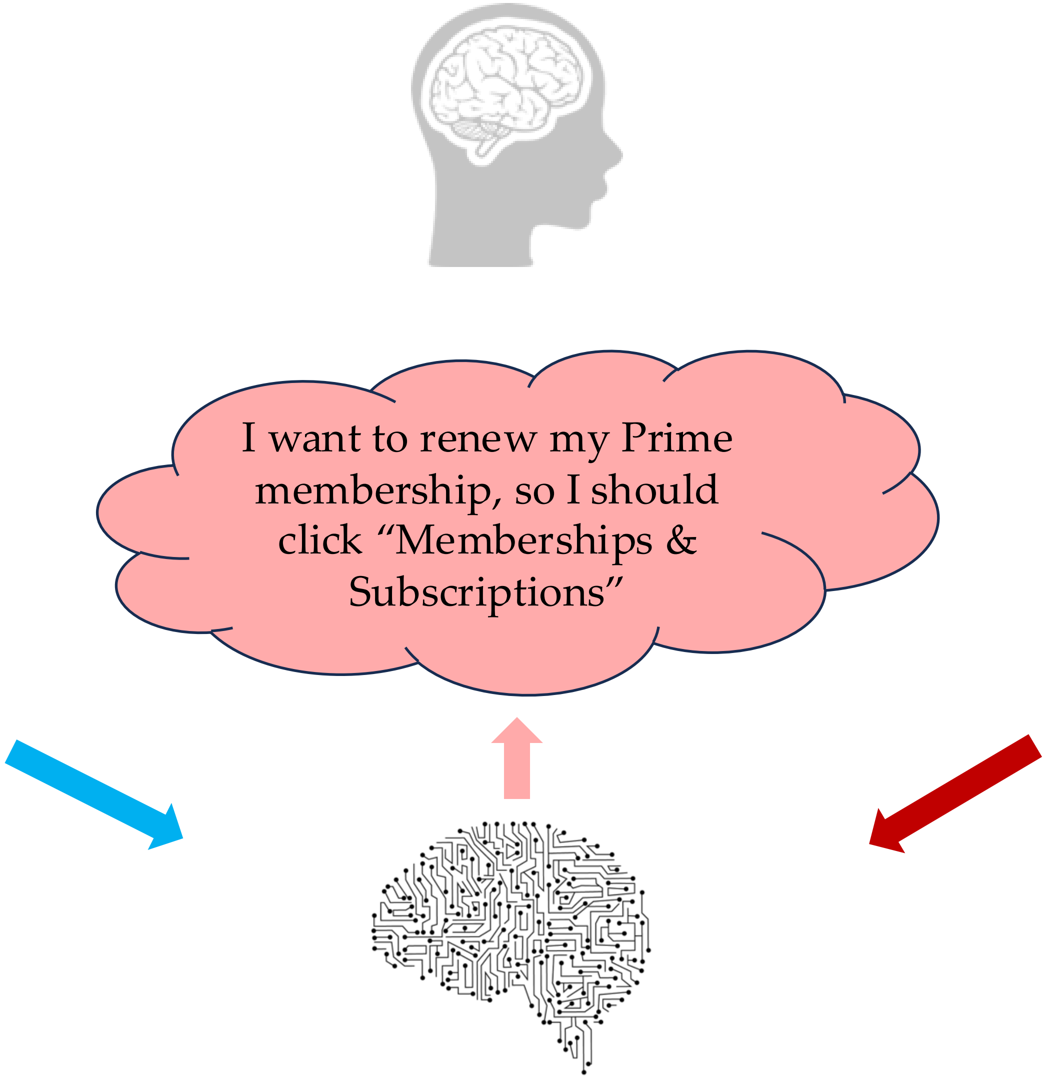
Synthetic data is key for agent learning

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

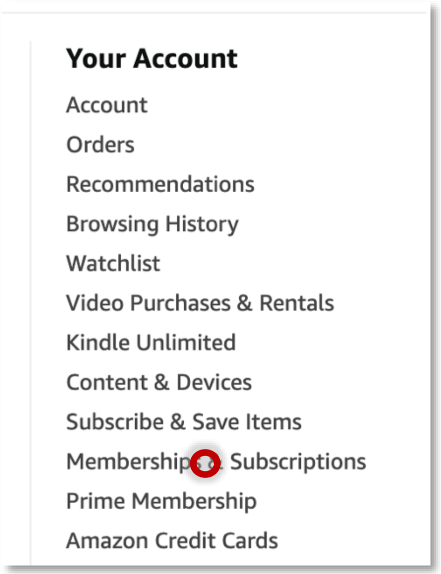


Your Account

- Account
- Orders
- Recommendations
- Browsing History
- Watchlist
- Video Purchases & Rentals
- Kindle Unlimited
- Content & Devices
- Subscribe & Save Items
- Memberships & Subscriptions
- Prime Membership
- Amazon Credit Cards



I want to renew my Prime membership, so I should click "Memberships & Subscriptions"



Your Account

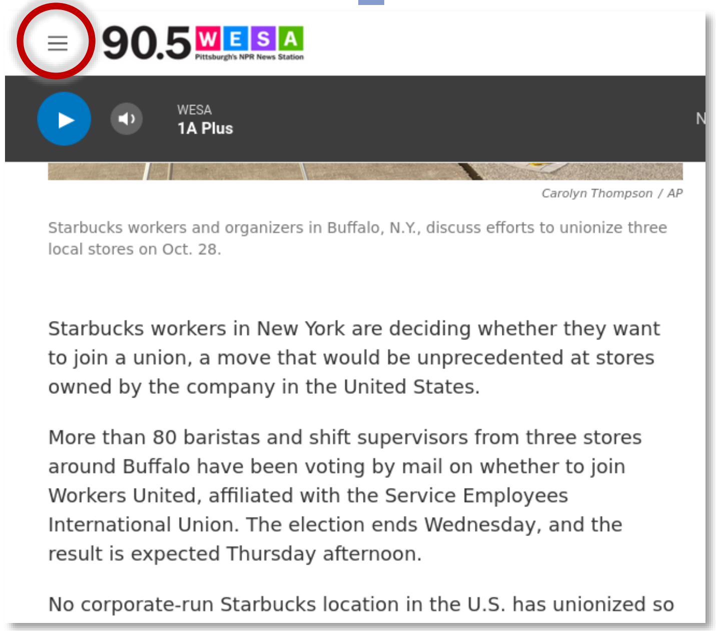
- Account
- Orders
- Recommendations
- Browsing History
- Watchlist
- Video Purchases & Rentals
- Kindle Unlimited
- Content & Devices
- Subscribe & Save Items
- Memberships & Subscriptions
- Prime Membership
- Amazon Credit Cards

Screenshot

HTML

type: button
aria-label: menu
alt-text: …
…

Metadata

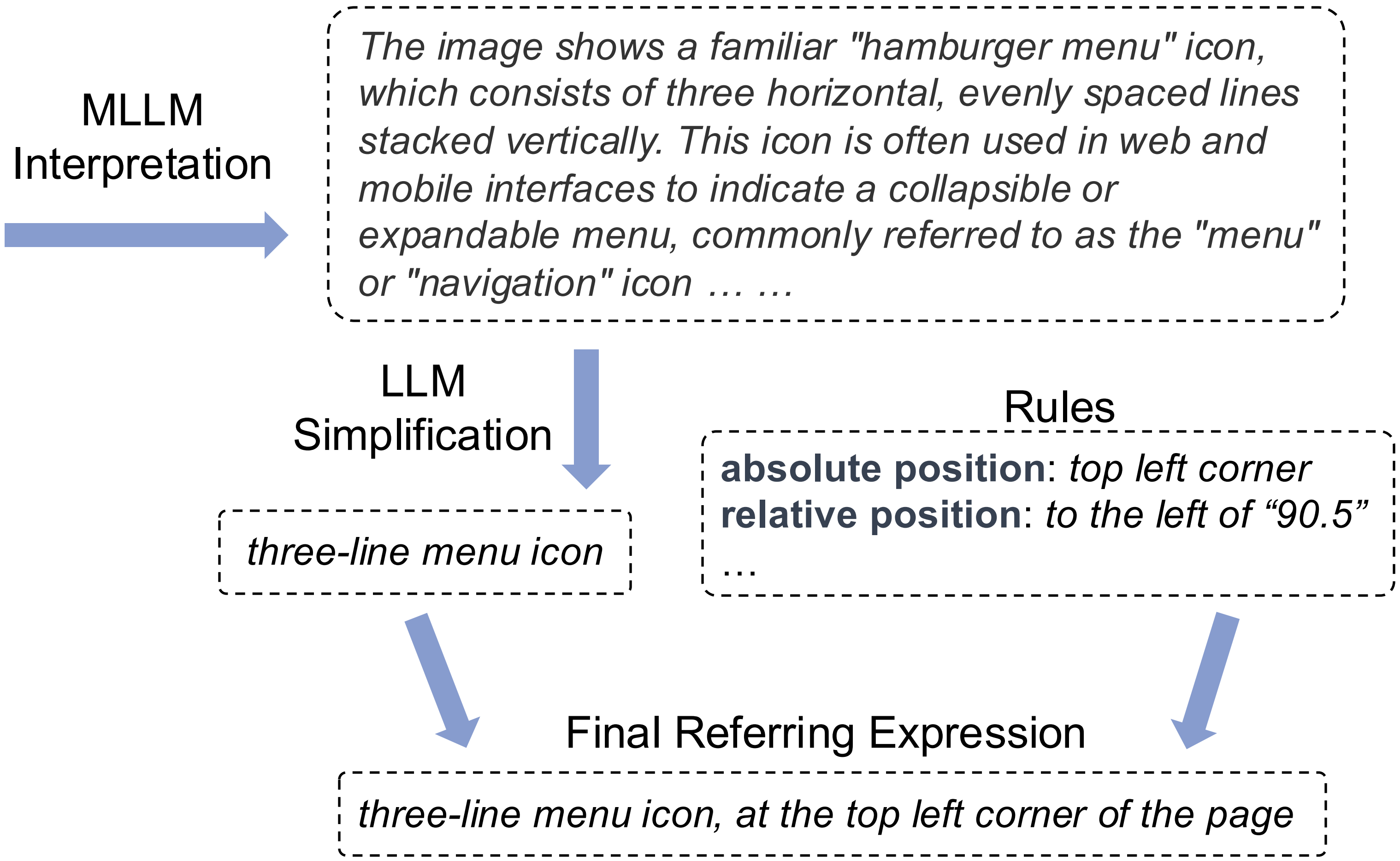


Starbucks workers and organizers in Buffalo, N.Y., discuss efforts to unionize three local stores on Oct. 28.

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



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 … …

LLM Simplification

three-line menu icon

Rules

absolute position: *top left corner*
relative position: *to the left of "90.5"*
…

Final Referring Expression

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

9M (screenshot, refer. expression, coordinates) triplets over **773K** web screenshots

Minimalist design is the most generalizable

- Most comprehensive evaluation on six agent benchmarks
- SeeAct-V + UGround outperforms prior art despite its minimalist design
- Generalize to desktop/mobile even though most data is from the web

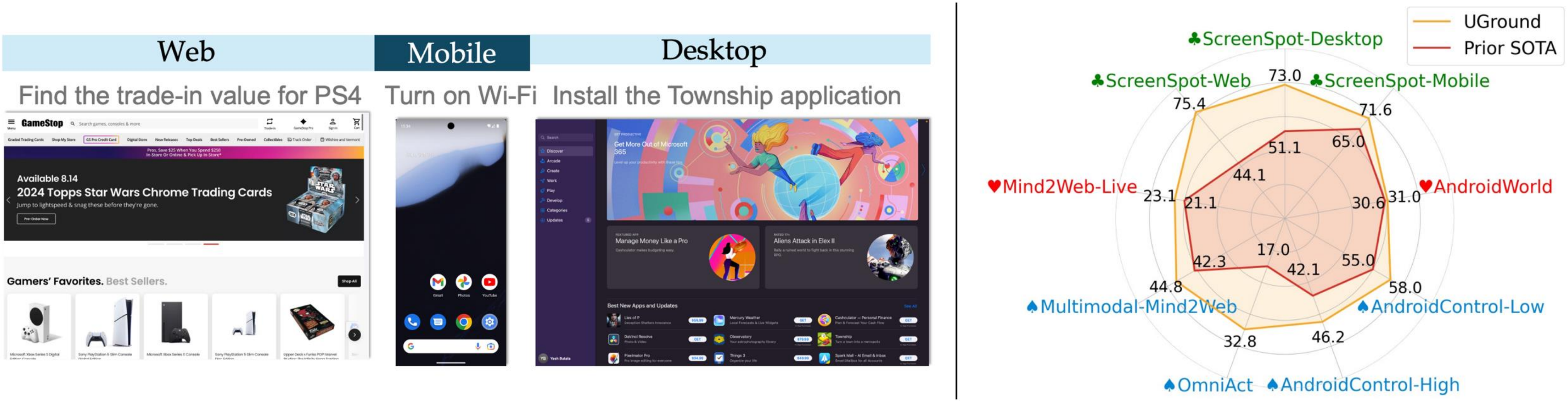
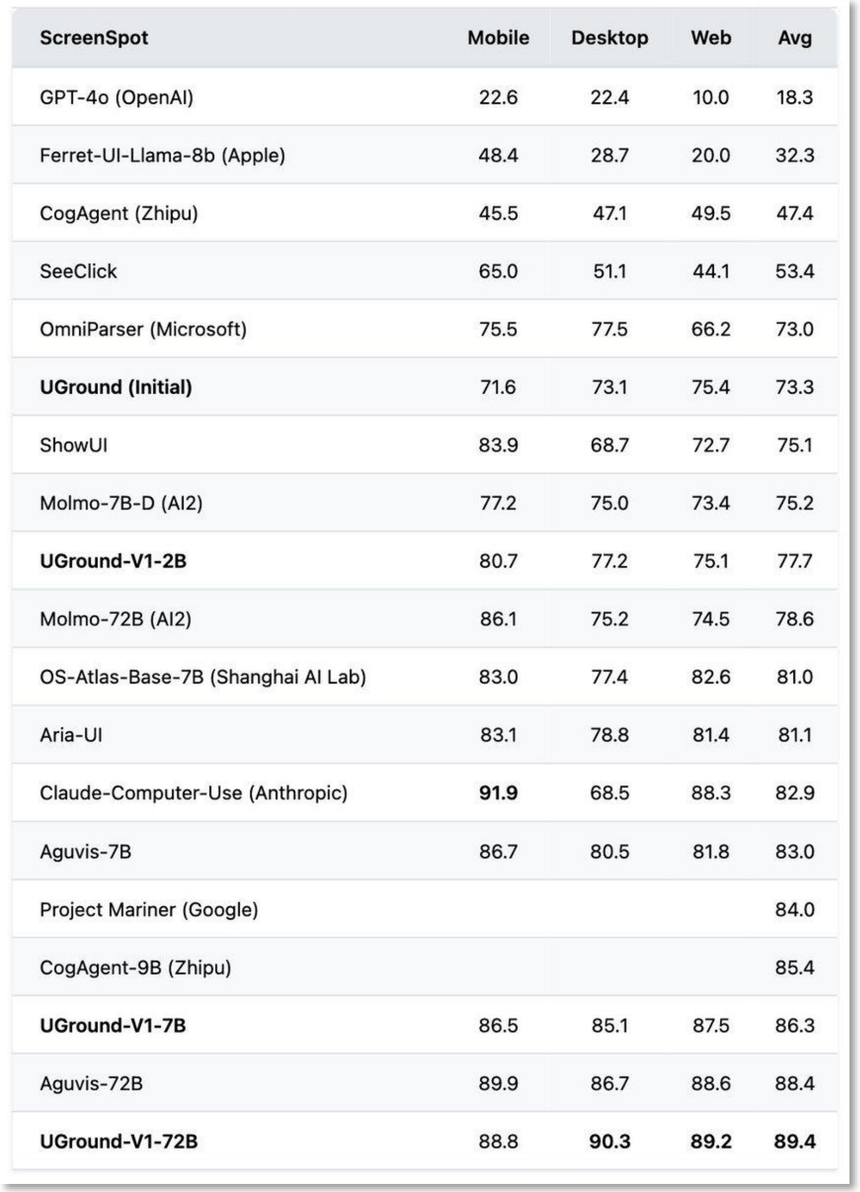


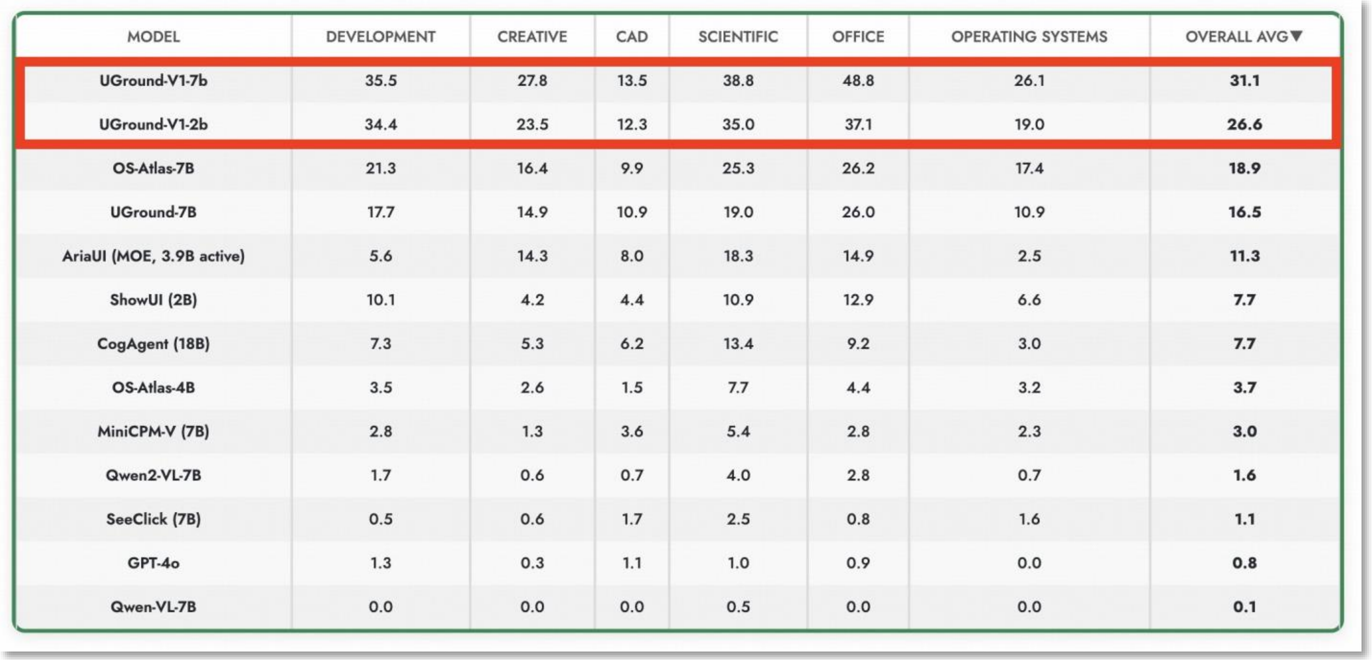
Figure 1: Examples of agent tasks across platforms and performance on **GUI grounding** (♣: ScreenSpot), **offline agent** (♠: Multimodal-Mind2Web, AndroidControl, and OmniAct), and **online agent benchmarks** (♥: Mind2Web-Live and AndroidWorld) when using GPT-4 as the planner.

Remarkable effectiveness of synthetic data

- Same data + Qwen2-VL (instead of Llava-NeXT)
- 95% data from web + 5% Android. 0% desktop data



ScreenSpot	Mobile	Desktop	Web	Avg
GPT-4o (OpenAI)	22.6	22.4	10.0	18.3
Ferret-UI-Llama-8b (Apple)	48.4	28.7	20.0	32.3
CogAgent (Zhipu)	45.5	47.1	49.5	47.4
SeeClick	65.0	51.1	44.1	53.4
OmniParser (Microsoft)	75.5	77.5	66.2	73.0
UGround (Initial)	71.6	73.1	75.4	73.3
ShowUI	83.9	68.7	72.7	75.1
Molmo-7B-D (AI2)	77.2	75.0	73.4	75.2
UGround-V1-2B	80.7	77.2	75.1	77.7
Molmo-72B (AI2)	86.1	75.2	74.5	78.6
OS-Atlas-Base-7B (Shanghai AI Lab)	83.0	77.4	82.6	81.0
Aria-UI	83.1	78.8	81.4	81.1
Claude-Computer-Use (Anthropic)	**91.9**	68.5	88.3	82.9
Aguvis-7B	86.7	80.5	81.8	83.0
Project Mariner (Google)				84.0
CogAgent-9B (Zhipu)				85.4
UGround-V1-7B	86.5	85.1	87.5	86.3
Aguvis-72B	89.9	86.7	88.6	88.4
UGround-V1-72B	88.8	**90.3**	**89.2**	**89.4**



MODEL	DEVELOPMENT	CREATIVE	CAD	SCIENTIFIC	OFFICE	OPERATING SYSTEMS	OVERALL AVG▼
UGround-V1-7b	35.5	27.8	13.5	38.8	48.8	26.1	**31.1**
UGround-V1-2b	34.4	23.5	12.3	35.0	37.1	19.0	**26.6**
OS-Atlas-7B	21.3	16.4	9.9	25.3	26.2	17.4	**18.9**
UGround-7B	17.7	14.9	10.9	19.0	26.0	10.9	**16.5**
AriaUI (MOE, 3.9B active)	5.6	14.3	8.0	18.3	14.9	2.5	**11.3**
ShowUI (2B)	10.1	4.2	4.4	10.9	12.9	6.6	**7.7**
CogAgent (18B)	7.3	5.3	6.2	13.4	9.2	3.0	**7.7**
OS-Atlas-4B	3.5	2.6	1.5	7.7	4.4	3.2	**3.7**
MiniCPM-V (7B)	2.8	1.3	3.6	5.4	2.8	2.3	**3.0**
Qwen2-VL-7B	1.7	0.6	0.7	4.0	2.8	0.7	**1.6**
SeeClick (7B)	0.5	0.6	1.7	2.5	0.8	1.6	**1.1**
GPT-4o	1.3	0.3	1.1	1.0	0.9	0.0	**0.8**
Qwen-VL-7B	0.0	0.0	0.0	0.5	0.0	0.0	**0.1**

ScreenSpot

ScreenSpot-Pro
(Professional Desktop Software)