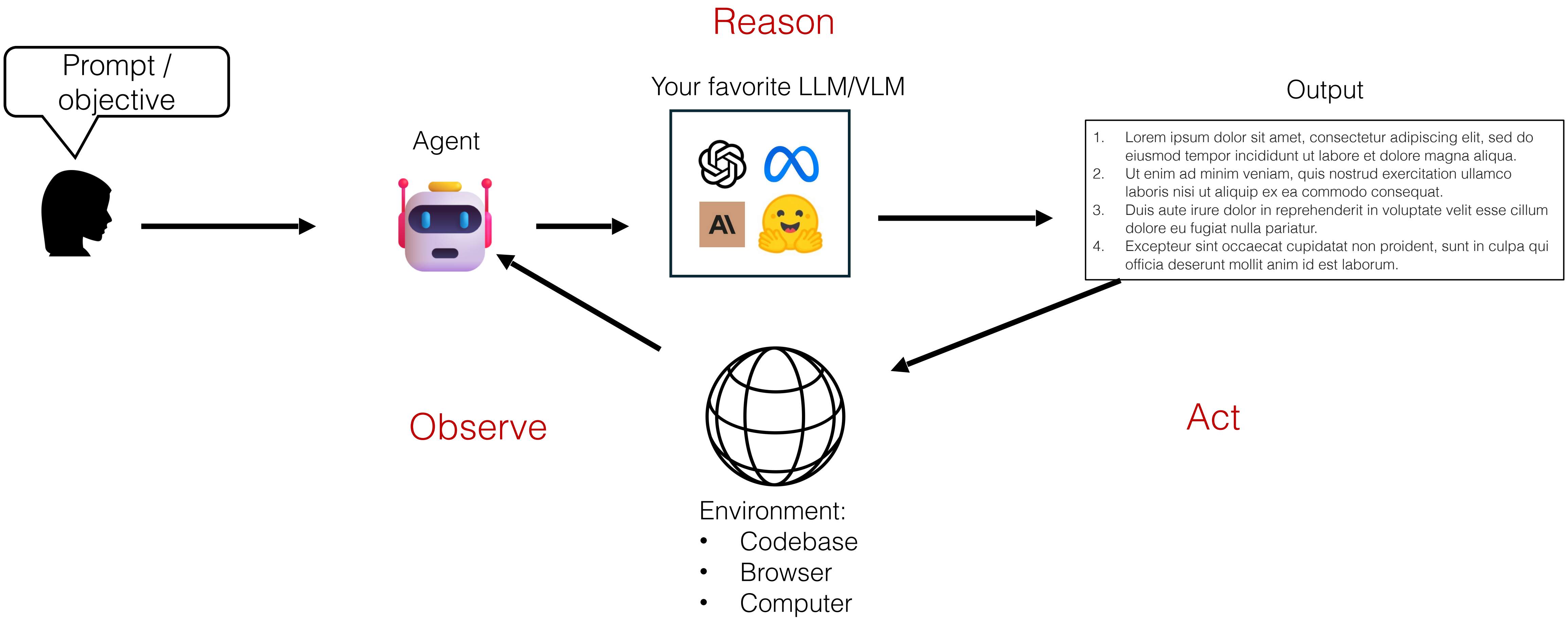
Windows Agent Arena: Evaluating Multi-Modal OS Agents at Scale

https://microsoft.github.io/WindowsAgentArena/

Rogerio Bonatti, Dan Zhao,
Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu,
Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Jang, Zack Hui

What are agents? What can they do?



Reason

Prompt / objective

Agent

Your favorite LLM/VLM

Output

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.
2. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.
3. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.
4. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Observe

Act

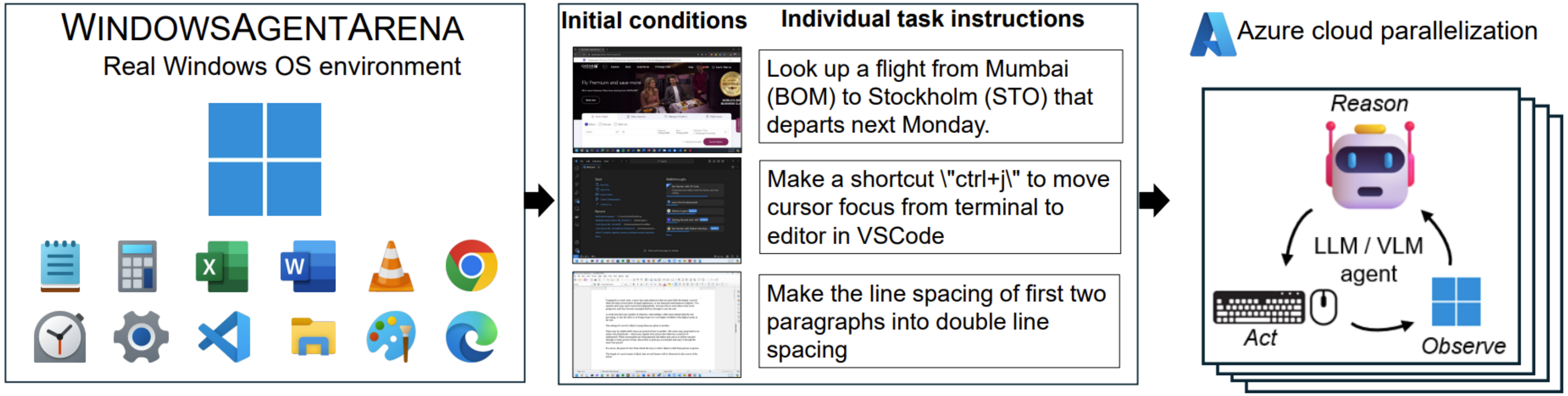
Environment:

- Codebase
- Browser
- Computer

Challenge 1: Generalist OS agents are a superset of browser and domain-specific agents.

But where’s Windows OS?

First agent benchmark for Windows OS


WINDOWSAGENTARENA
Real Windows OS environment
Initial conditions
Individual task instructions
Azure cloud parallelization
Reason
LLM / VLM agent
Act
Observe

Look up a flight from Mumbai (BOM) to Stockholm (STO) that departs next Monday.

Make a shortcut \"ctrl+j\" to move cursor focus from terminal to editor in VSCode

Make the line spacing of first two paragraphs into double line spacing

Windows:

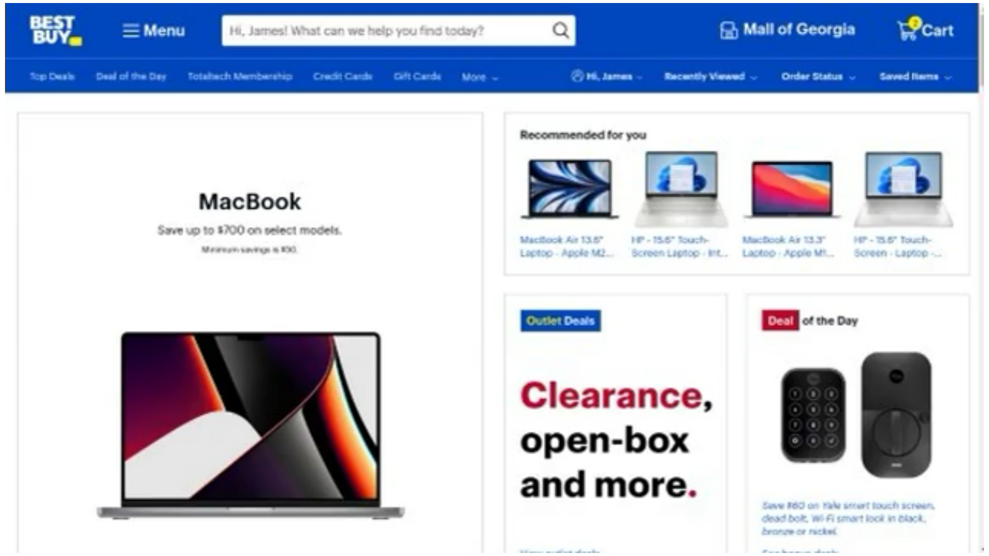
- Coding
- Command line
- Browser
- Documents
- Business
- Media/ entertainment
- Communication
- Photo/video editing
- Gaming
- Learning
- Multi-app tasks

Challenge 2: You can’t improve what you can’t measure. But how do we measure a computer agent’s performance?

- Static trajectory matching

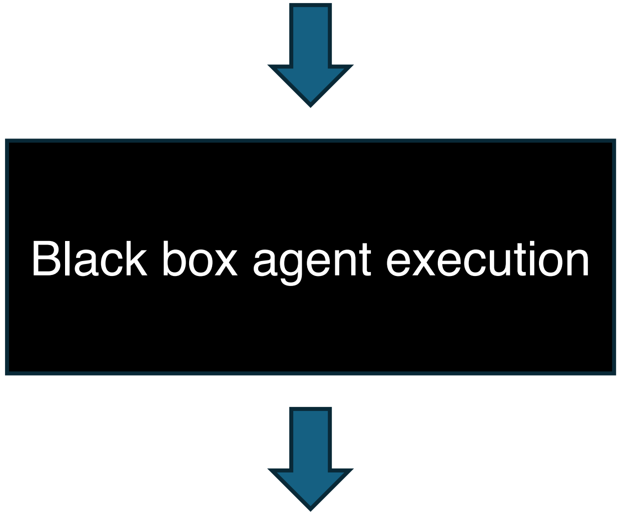
Task: finance a blue iPhone 13 with 256gb

Human demonstration (Mind2Web benchmark):


BEST BUY
Menu
Hi, James! What can we help you find today?
Mall of Georgia
Cart
Top Deals
Deal of the Day
Totaltech Membership
Credit Cards
Gift Cards
More
Hi, James
Recently Viewed
Order Status
Saved Items
MacBook
Save up to $700 on select models.
Minimum savings is $100.
Recommended for you
MacBook Air 13.6" Laptop - Apple M2...
HP - 15.6" Touch-Screen Laptop - Int...
MacBook Air 13.3" Laptop - Apple M1...
HP - 15.6" Touch-Screen - Laptop -...
Outlet Deals
Clearance, open-box and more.
Deal of the Day
Save $60 on Yale smart touch screen, dead bolt, Wi-Fi smart lock in black, bronze or nickel.

- Execution-based evaluation

Task: finance a blue iPhone 13 with 256gb



Black box agent execution

Trajectory is irrelevant. Success iff:

- iPhone added to cart
- iPhone memory = 256gb
- Payment method is financing

Challenge 3: execution-based evaluation is (very) slow

- Total eval time = # tasks * # steps * # seconds per step
  (20) (30 s)

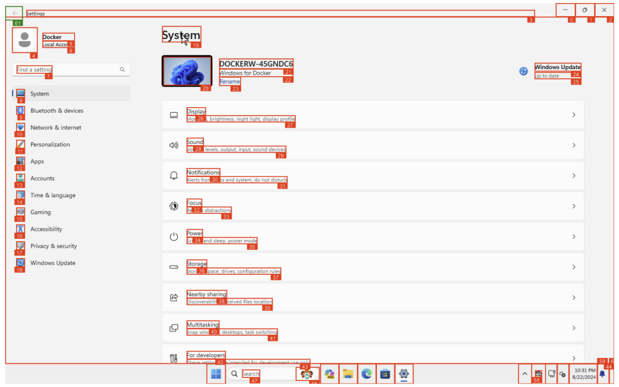
		Series	Parallel
Benchmark sizes today	100 tasks	17 hours	10 min
	1,000 tasks	1 week	10 min
Ideal benchmark size	10,000 tasks	2.5 months	10 min
Ideal dataset size	1M trajectories	19 years	10 min

Windows Agent Arena contributions

- First agent benchmark for Windows OS
- Execution-based evaluation
- 150+ tasks across 11 domains (browser, 1P/3P apps)
- Parallel evaluation in Azure in under 20min
- Open-source benchmark. Check it out!

Problem definition

- POMDP: $S, \mathcal{O}, \mathcal{A}, T, \mathcal{R}$
- Observation:



Pixels
+UIA

- Action: ⌨ 🖱
- Transition function $T : S \times \mathcal{A} \rightarrow S$
- Reward: $\mathcal{R} : S \times \mathcal{A} \rightarrow \mathbb{R}$

Group	Function

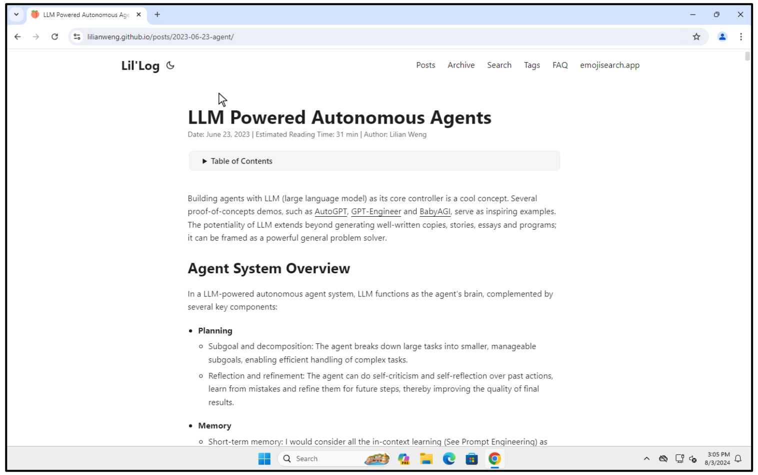
computer.mouse	move_id(id)
	move_abs(x,y)
	single_click()
	double_click()
	right_click()
	scroll(direction)

computer.
keyboard
	write(text)
	press(key)

computer.
clipboard
	copy_text(text)
	copy_image(image)
	paste()

computer.os	open_program(program)
computer. window_manager	switch_to_ application(window)

Task: *Computer, can you turn the webpage I'm looking at into a PDF file and put it on my main screen, you know, the Desktop?*



Lil'Log

Posts Archive Search Tags FAQ emojisearch.app

LLM Powered Autonomous Agents

Date: June 23, 2023 | Estimated Reading Time: 31 min | Author: Lilian Weng

▸ Table of Contents

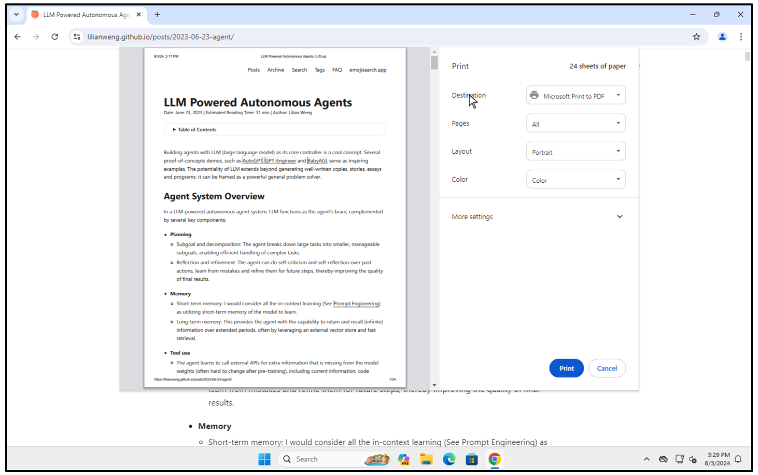
Building agents with LLM (large language model) as its core controller is a cool concept. Several proof-of-concepts demos, such as AutoGPT, GPT-Engineer and BabyAGI, serve as inspiring examples. The potential of LLM extends beyond generating well-written copies, stories, essays and programs; it can be framed as a powerful general problem solver.

Agent System Overview

In a LLM-powered autonomous agent system, LLM functions as the agent's brain, complemented by several key components:

- Planning
  - Subgoal and decomposition: The agent breaks down large tasks into smaller, manageable subgoals, enabling efficient handling of complex tasks.
  - Reflection and refinement: The agent can do self-criticism and self-reflection over past actions, learn from mistakes and refine them for future steps, thereby improving the quality of final results.
- Memory
  - Short-term memory: I would consider all the in-context learning (See Prompt Engineering) as

Step 1: computer.keyboard.press("ctrl+p")



Print 24 sheets of paper

Destination: Microsoft Print to PDF

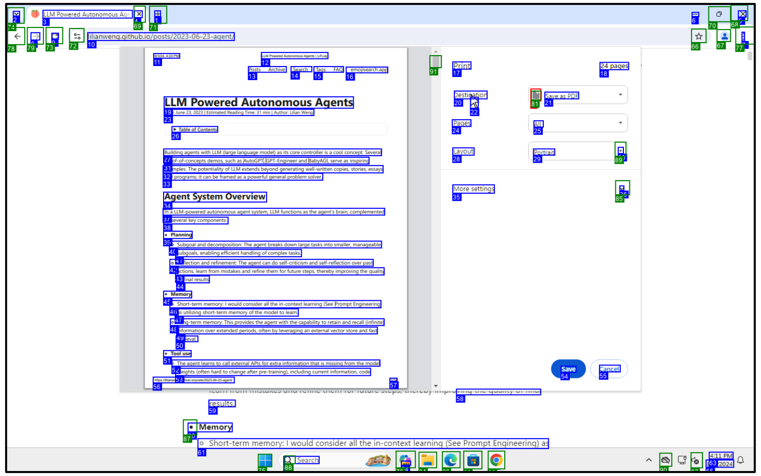
Pages: All

Layout: Portrait

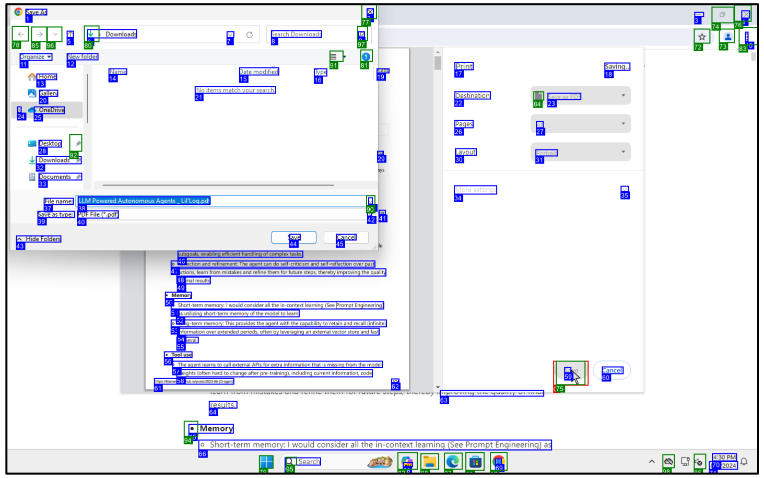
Color: Color

More settings

Step 2: # Move the focus to the 'Destination' dropdown
computer.keyboard.press("tab")
# Attempt to navigate the dropdown with the keyboard
computer.keyboard.press("down")



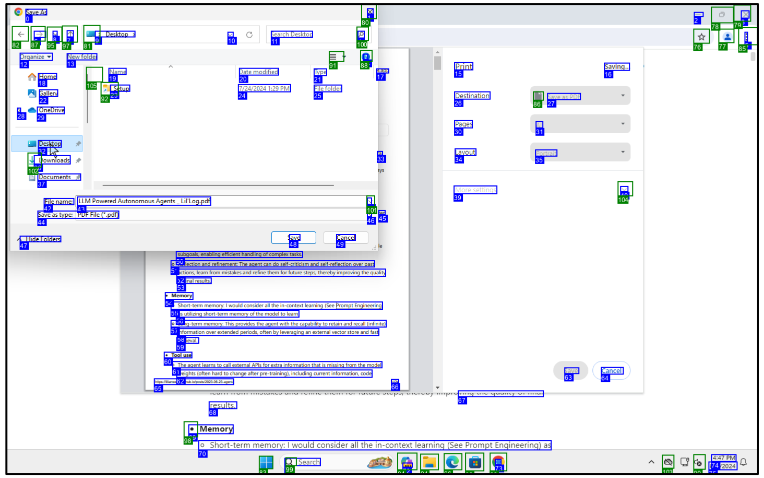
Step 3: # Move to the "Save" button
computer.mouse.move_id(id=54)
# Click to save the webpage as a PDF
computer.mouse.single_click()



File name: LLM Powered Autonomous Agents _ Lil'Log.pdf

Save as type: PDF File (*.pdf)

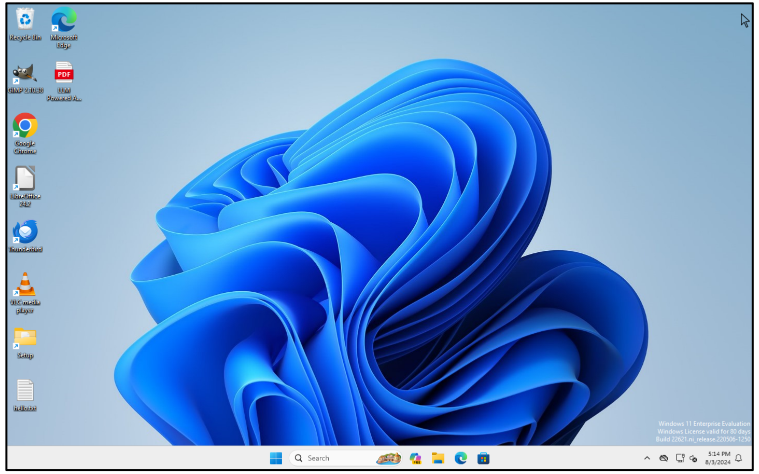
Step 4: # Move to "Desktop" (pinned) to change the location computer.mouse.move_id(id=28)
computer.mouse.single_click()



File name: LLM Powered Autonomous Agents _ Lil'Log.pdf

Save as type: PDF File (*.pdf)

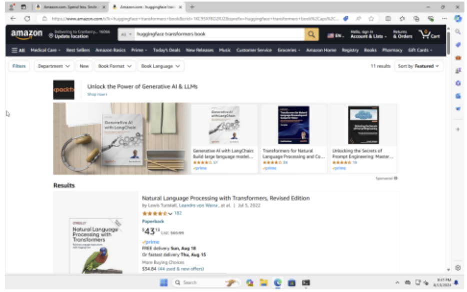
Step 5: # Move cursor to the "Save" button
computer.mouse.move_id(id=48)
# Click to save the webpage as a PDF to the Desktop

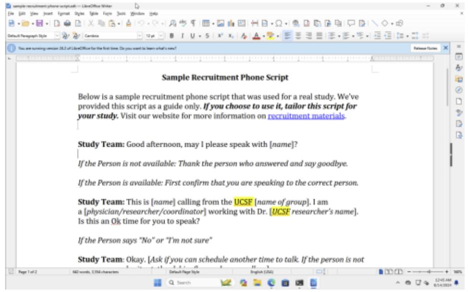


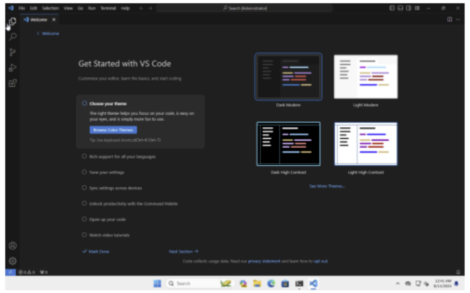
Result: Task successful, reward 1.0

✓ Execution-based evaluation

Initial State	Task Instruction	Evaluation Script (Simplified)

	"Can you help me clean up my computer by getting rid of all the tracking things that Amazon might have saved? I want to make sure my browsing is private and those sites don't remember me."	`cookie_data = get_cookie_data(env)`
		`rule = {"type":"rule",`
		`         "domains": ["amazon.com"]}`
		`is_cookie_deleted(cookie_data, rule)`

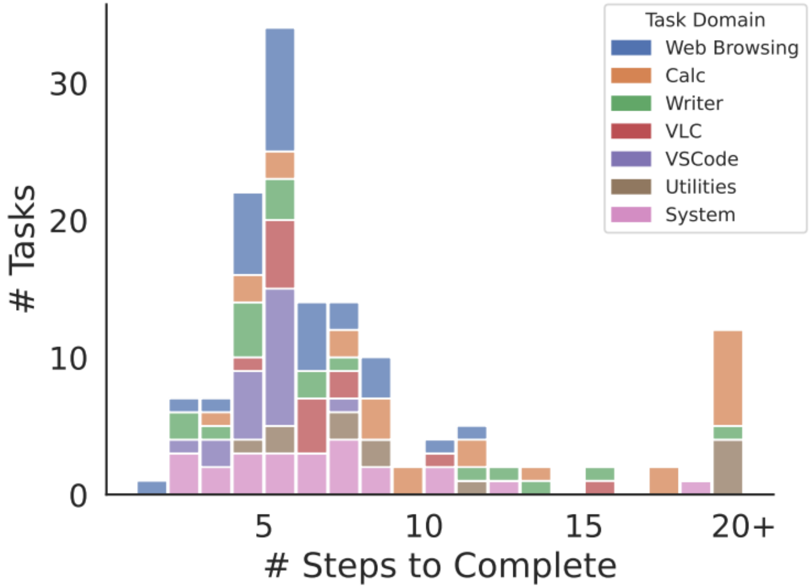
 **Sample Recruitment Phone Script**
Below is a sample recruitment phone script that was used for a real study. We've provided this script as a guide only. ***If you choose to use it, tailor this script for your study.*** Visit our website for more information on recruitment materials.
Study Team: Good afternoon, may I please speak with [*name*]?
If the Person is not available: Thank the person who answered and say goodbye.
If the Person is available: First confirm that you are speaking to the correct person.
Study Team: This is [*name*] calling from the UCSF [*name of group*]. I am a [*physician/researcher/coordinator*] working with Dr. [*UCSF researcher's name*]. Is this an Ok time for you to speak?
If the Person says "No" or "I'm not sure"
Study Team: Okay. [*Ask if you can schedule another time to talk. If the person is not*
	"I have been editing my document and some words that needed to be rewritten are highlighted in yellow. As I fix those words, please help me remove all highlight. I want to make sure that there is no highlighted word."	`vm_file = get_file(env, "file.doc")`
		`golden_file = get_file(drive,`
		`                        "file.doc")`
		`check_highlighted_words(vm_file,`
		`                         golden_file)`

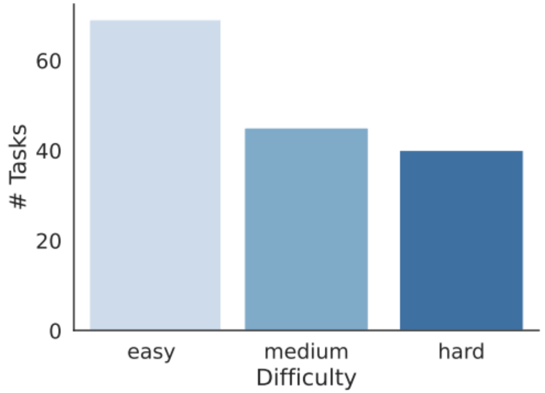
	"Please help me modify the setting of VS Code to keep my cursor focused on the debug console when debugging in VS Code, instead of automatically focusing back on the Editor.".	`vscode_data = get_vscode_config(env)`
		`rule = {"expected":`
		` {"debug.focusEditorOnBreak": false}`
		`      }`
		`check_json_settings(vscode_data, rule)`

⊕ 154 tasks across 12 domains

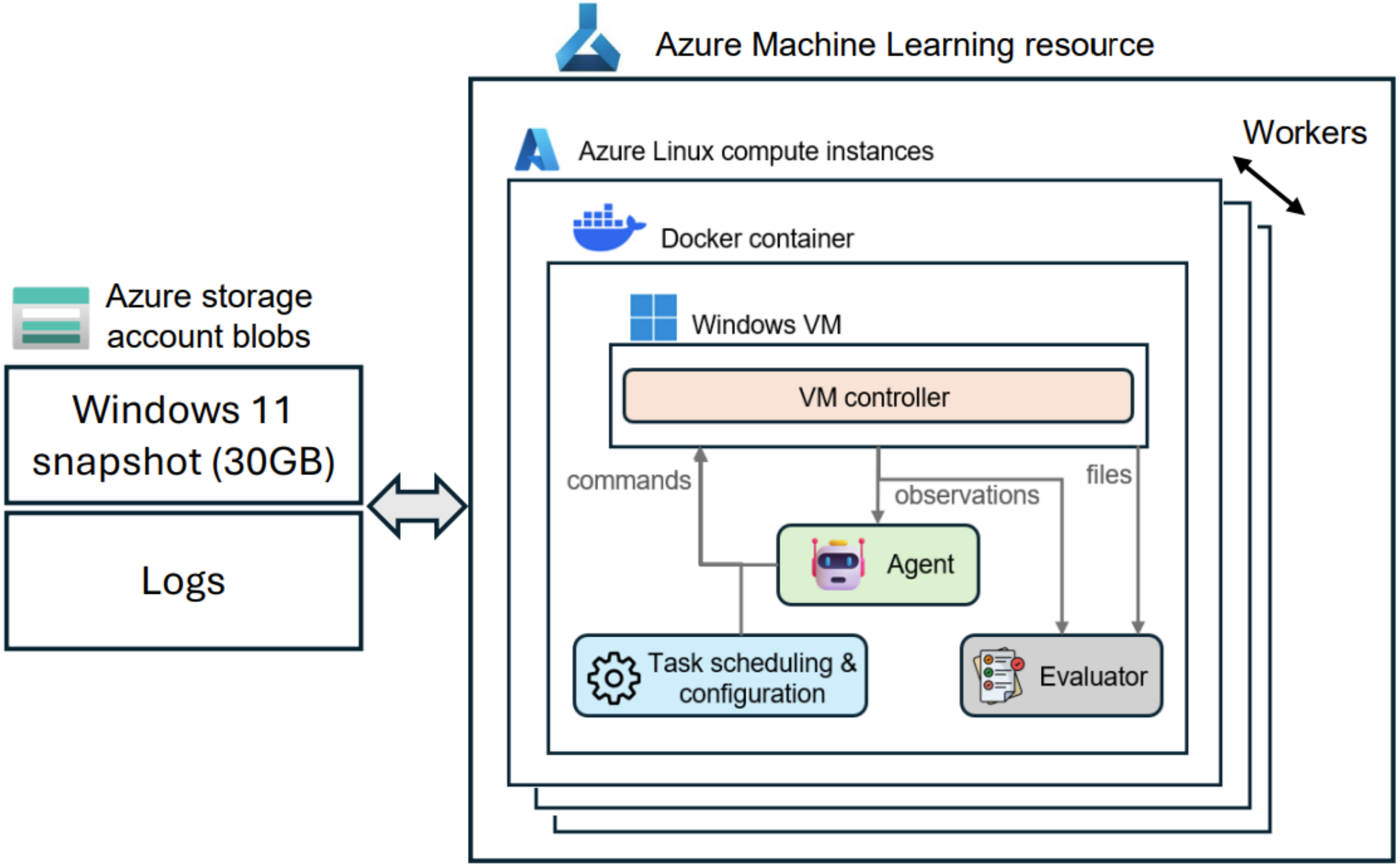
Unassisted human performance = 74% success rate

Domain	Activity Examples	# Tasks	%
Office	Libreoffice Writer (Editing, writing) and LibreOffice Calc (spreadsheets, plotting, data manipulation)	43	28%
Web Browsing	Online shopping & search, customizing app settings with Edge and Chrome	30	19%
Windows System	File Explorer (navigate, customize files), Windows Settings, customization	24	16%
Coding	Editing code, customizing IDE, installing extensions with VSCode	24	16%
Media & Video	Watching videos in VLC, listening to music, settings	21	14%
Windows Utilities	Miscellaneous tasks with Notepad, Clock, and Paint	12	8%
Total		**154**	**100%**



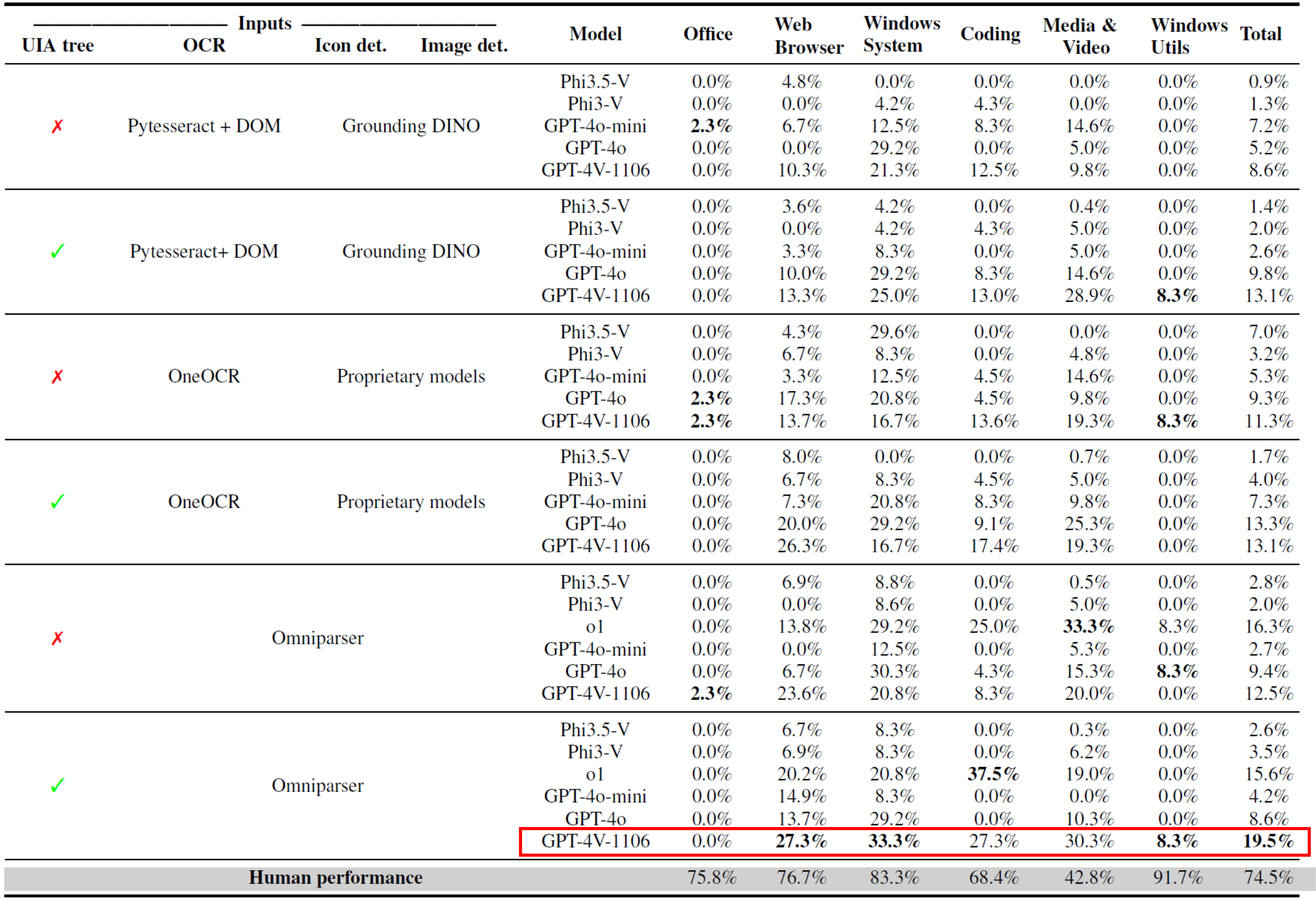


Secure cloud parallelization in Azure ML


Azure storage account blobs
Windows 11 snapshot (30GB)
Logs
Azure Machine Learning resource
Azure Linux compute instances
Workers
Docker container
Windows VM
VM controller
commands
observations
files
Agent
Task scheduling & configuration
Evaluator

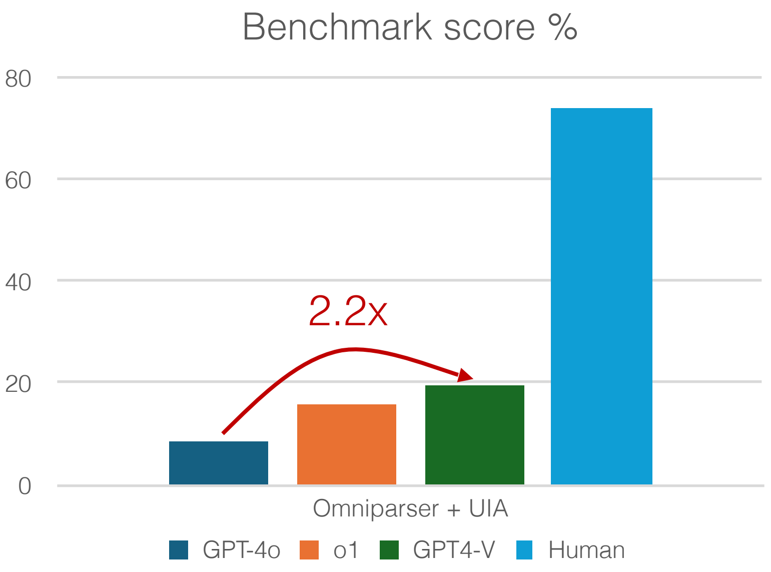
- Golden snapshot copied for each VM
- Full separation between instances
- Security: no external VM commands

Results

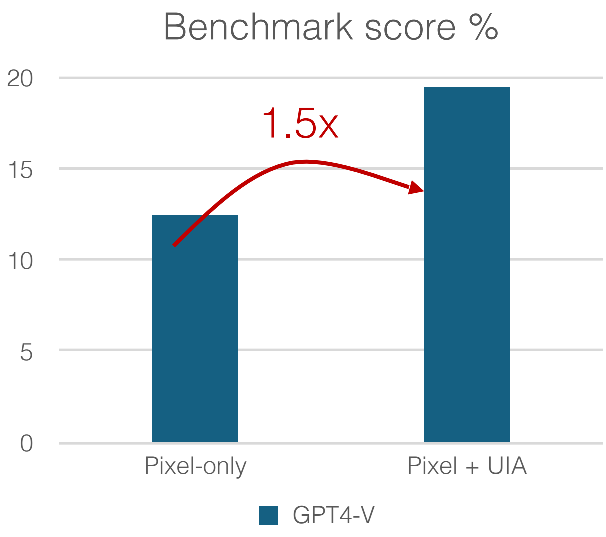


Inputs				Model	Office	Web Browser	Windows System	Coding	Media & Video	Windows Utils	Total
UIA tree	**OCR**	**Icon det.**	**Image det.**								
✗	Pytesseract + DOM	Grounding DINO		Phi3.5-V	0.0%	4.8%	0.0%	0.0%	0.0%	0.0%	0.9%
				Phi3-V	0.0%	0.0%	4.2%	4.3%	0.0%	0.0%	1.3%
				GPT-4o-mini	**2.3%**	6.7%	12.5%	8.3%	14.6%	0.0%	7.2%
				GPT-4o	0.0%	0.0%	29.2%	0.0%	5.0%	0.0%	5.2%
				GPT-4V-1106	0.0%	10.3%	21.3%	12.5%	9.8%	0.0%	8.6%
✓	Pytesseract+ DOM	Grounding DINO		Phi3.5-V	0.0%	3.6%	4.2%	0.0%	0.4%	0.0%	1.4%
				Phi3-V	0.0%	0.0%	4.2%	4.3%	5.0%	0.0%	2.0%
				GPT-4o-mini	0.0%	3.3%	8.3%	0.0%	5.0%	0.0%	2.6%
				GPT-4o	0.0%	10.0%	29.2%	8.3%	14.6%	0.0%	9.8%
				GPT-4V-1106	0.0%	13.3%	25.0%	13.0%	28.9%	**8.3%**	13.1%
✗	OneOCR	Proprietary models		Phi3.5-V	0.0%	4.3%	29.6%	0.0%	0.0%	0.0%	7.0%
				Phi3-V	0.0%	6.7%	8.3%	0.0%	4.8%	0.0%	3.2%
				GPT-4o-mini	0.0%	3.3%	12.5%	4.5%	14.6%	0.0%	5.3%
				GPT-4o	**2.3%**	17.3%	20.8%	4.5%	9.8%	0.0%	9.3%
				GPT-4V-1106	**2.3%**	13.7%	16.7%	13.6%	19.3%	**8.3%**	11.3%
✓	OneOCR	Proprietary models		Phi3.5-V	0.0%	8.0%	0.0%	0.0%	0.7%	0.0%	1.7%
				Phi3-V	0.0%	6.7%	8.3%	4.5%	5.0%	0.0%	4.0%
				GPT-4o-mini	0.0%	7.3%	20.8%	8.3%	9.8%	0.0%	7.3%
				GPT-4o	0.0%	20.0%	29.2%	9.1%	25.3%	0.0%	13.3%
				GPT-4V-1106	0.0%	26.3%	16.7%	17.4%	19.3%	0.0%	13.1%
✗	Omniparser			Phi3.5-V	0.0%	6.9%	8.8%	0.0%	0.5%	0.0%	2.8%
				Phi3-V	0.0%	0.0%	8.6%	0.0%	5.0%	0.0%	2.0%
				o1	0.0%	13.8%	29.2%	25.0%	**33.3%**	8.3%	16.3%
				GPT-4o-mini	0.0%	0.0%	12.5%	0.0%	5.3%	0.0%	2.7%
				GPT-4o	0.0%	6.7%	30.3%	4.3%	15.3%	**8.3%**	9.4%
				GPT-4V-1106	**2.3%**	23.6%	20.8%	8.3%	20.0%	0.0%	12.5%
✓	Omniparser			Phi3.5-V	0.0%	6.7%	8.3%	0.0%	0.3%	0.0%	2.6%
				Phi3-V	0.0%	6.9%	8.3%	0.0%	6.2%	0.0%	3.5%
				o1	0.0%	20.2%	20.8%	**37.5%**	19.0%	0.0%	15.6%
				GPT-4o-mini	0.0%	14.9%	8.3%	0.0%	0.0%	0.0%	4.2%
				GPT-4o	0.0%	13.7%	29.2%	0.0%	10.3%	0.0%	8.6%
				GPT-4V-1106	0.0%	**27.3%**	**33.3%**	27.3%	30.3%	**8.3%**	**19.5%**
Human performance					75.8%	76.7%	83.3%	68.4%	42.8%	91.7%	74.5%

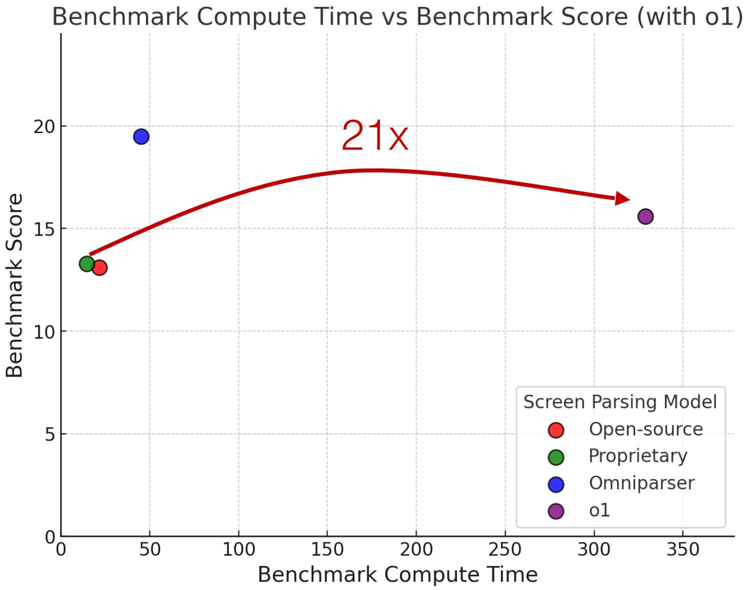
A tale in 3 plots:



GPT4-V >> GPT4-o
(for agents)



UIA / DOM boosts
pixel models

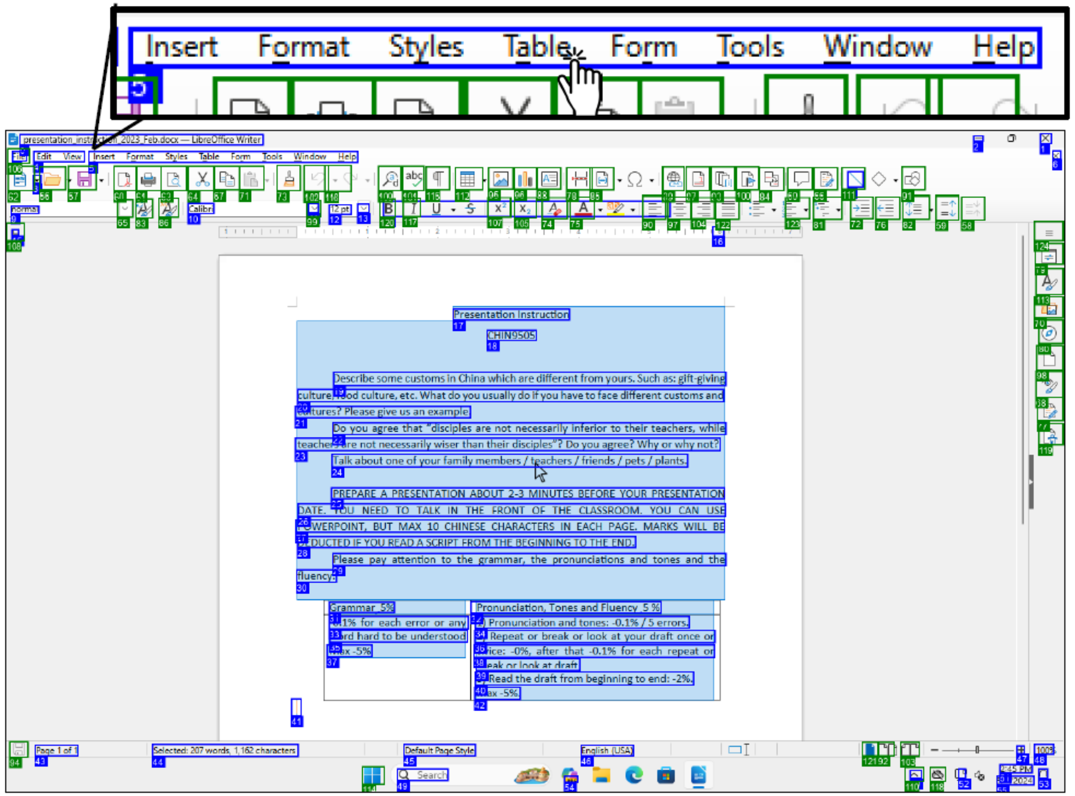


Model speed matters
(even longer without
our parallelization)

Where do agents fail?

Task: *Convert all uppercase text to lowercase*

Step 3: computer.mouse.move_id(id=5) # Move to the 'Format' menu option
computer.mouse.single_click() # Open the 'Format' menu



Presentation Instruction

CHIN9505

Describe some customs in China which are different from yours. Such as: gift-giving culture, food culture, etc. What do you usually do if you have to face different customs and cultures? Please give us an example.

Do you agree that "disciples are not necessarily inferior to their teachers, while teachers are not necessarily wiser than their disciples"? Do you agree? Why or why not?

Talk about one of your family members / teachers / friends / pets / plants.

PREPARE A PRESENTATION ABOUT 2-3 MINUTES BEFORE YOUR PRESENTATION DATE. YOU NEED TO TALK IN THE FRONT OF THE CLASSROOM. YOU CAN USE POWERPOINT, BUT MAX 10 CHINESE CHARACTERS IN EACH PAGE. MARKS WILL BE DEDUCTED IF YOU READ A SCRIPT FROM THE BEGINNING TO THE END.

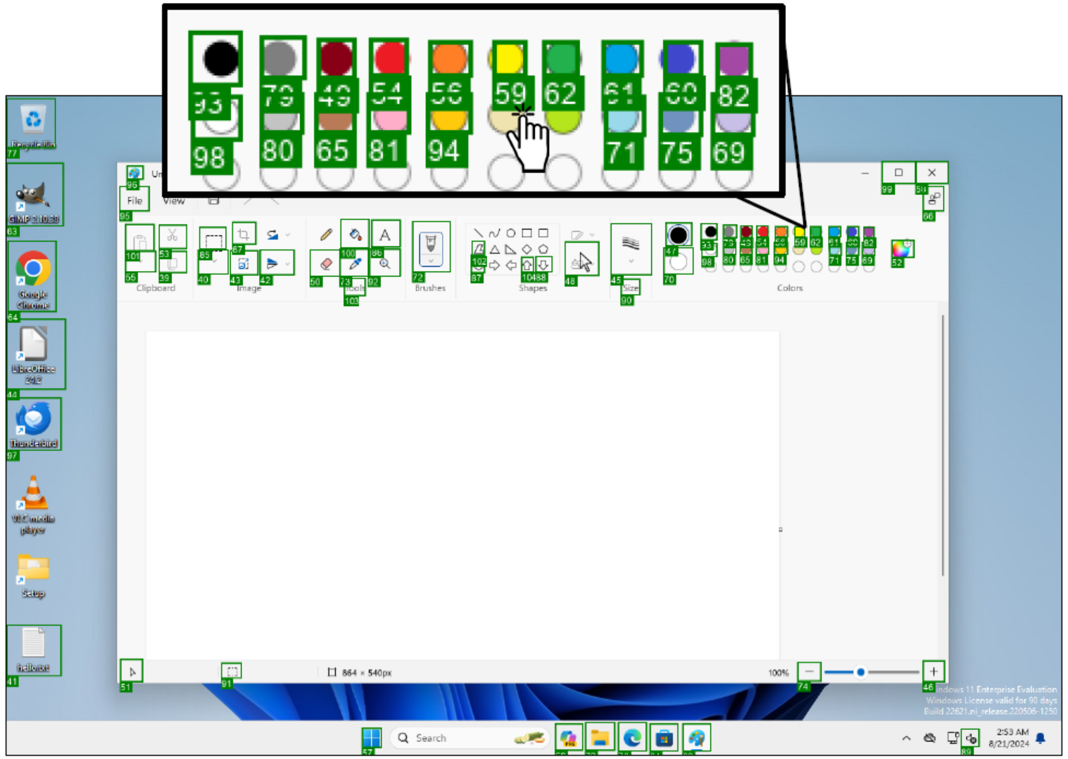
Please pay attention to the grammar, the pronunciations and tones and the fluency.

Grammar 5%	Pronunciation, Tones and Fluency 5 %
-0.1% for each error or any word hard to be understood Max -5%	1) Pronunciation and tones: -0.1% / 5 errors. 2) Repeat or break or look at your draft once or twice: -0%, after that -0.1% for each repeat or break or look at draft 3) Read the draft from beginning to end: -2%. Max -5%.

Page 1 of 1 Selected: 207 words, 1,162 characters Default Page Style English (USA)

Task: *Open Paint and draw a red circle*

Step 7: computer.mouse.move_id(id=59) # Move to the red color on the palette
computer.mouse.single_click() # Click to select the red color



Screen parsing errors

Visual-language misalignment

Thank you!

Feel free to reach out to dz1158@nyu.edu

Hope to see you at our poster at ICML