COMP 3361 Natural Language Processing

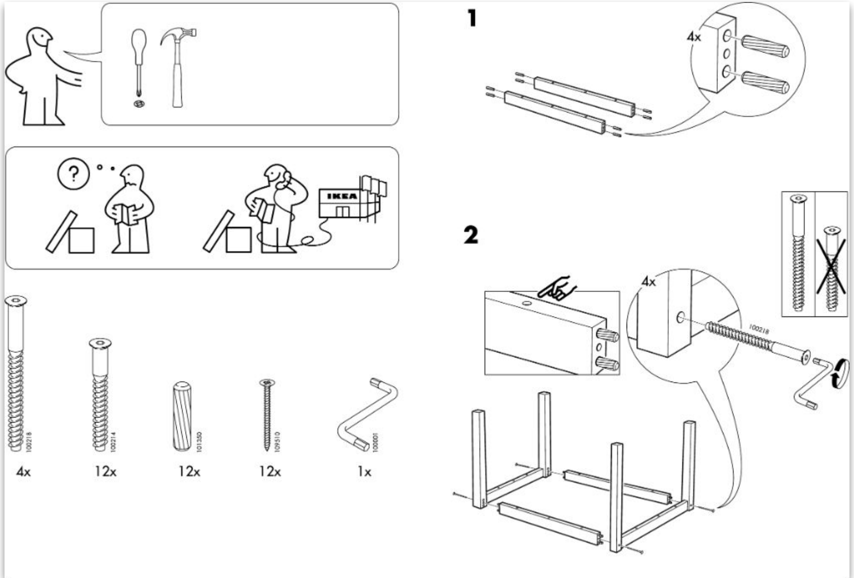
Lecture 17: LLMs/VLMs as Agents

Spring 2024

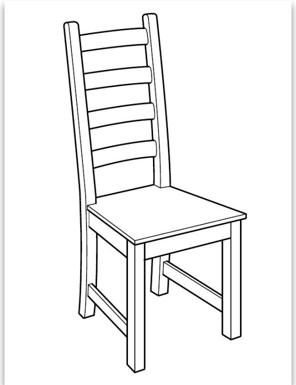
Announcements

- Two invited talks this Friday, April 12
    - Solving Real-World Tasks with AI Agents (by Shuyan Zhou, CMU)
    - Instruction tuning for LLMs (by Yizhong Wang, UW)

IKEA furniture assembly



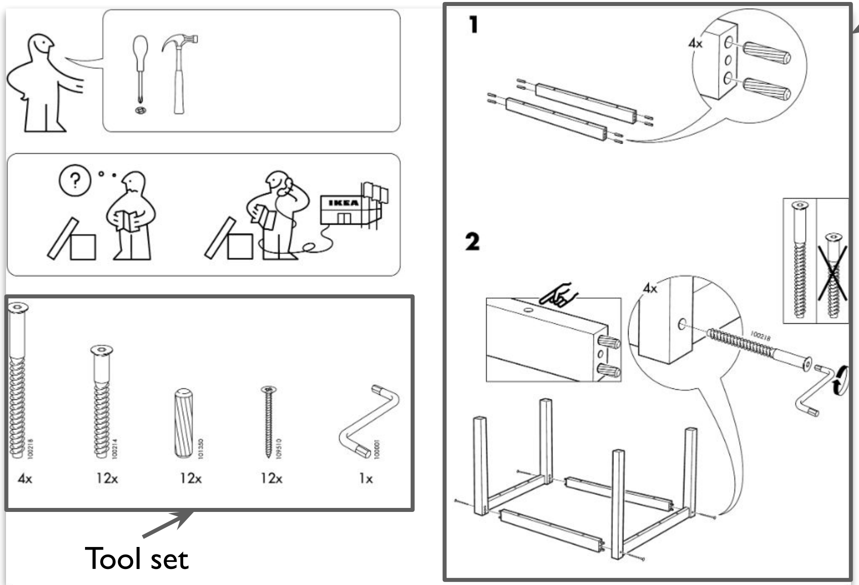
Assembly instructions



Assembled chair

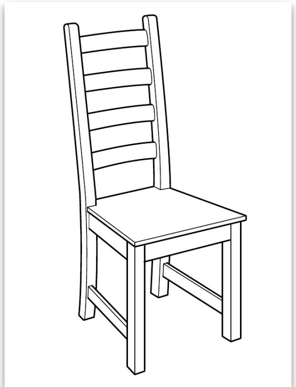
Planning with tools

Step-by-step plans



Tool set

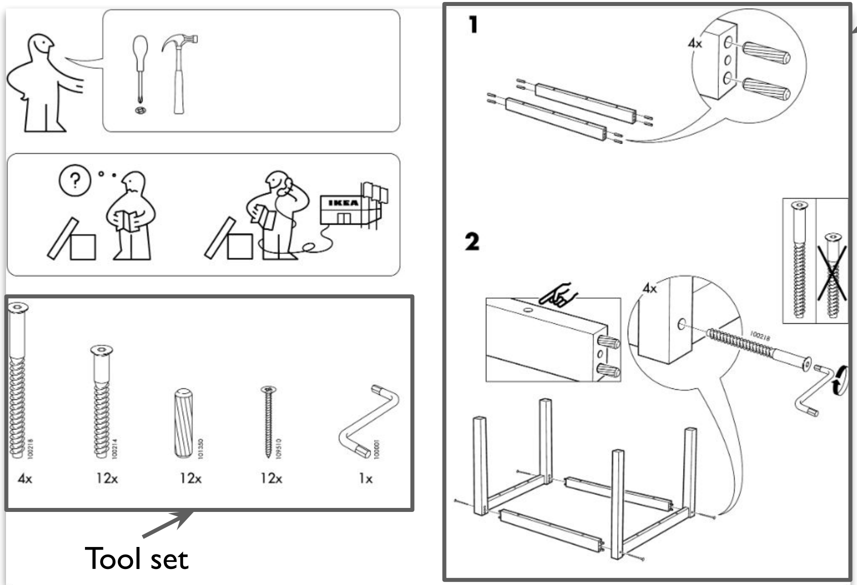
Assembly instructions



Assembled chair

Grounding plans into actions in physical world

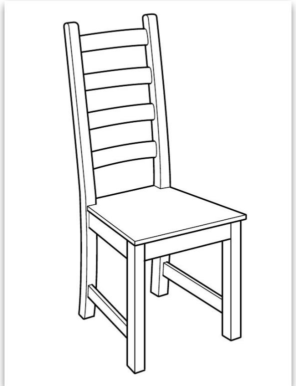
Step-by-step plans



Tool set

Assembly instructions

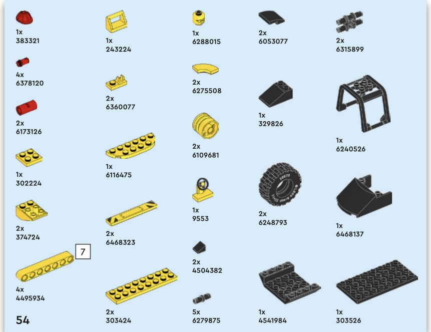
Grounding



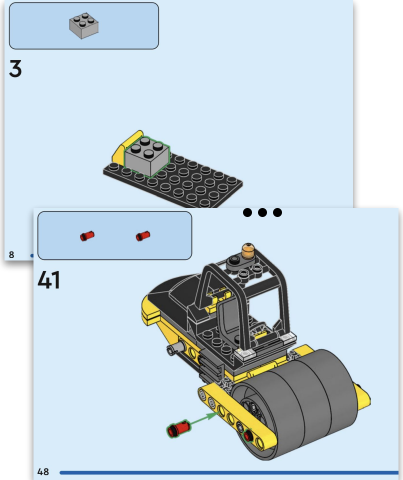
Assembled chair

https://www.ikea.com/us/en/customer-service/services/taskrabbit-and-ikea-partnership-pubddab78f4

Lego building in physical world



Building environments



Building instructions

Grounding

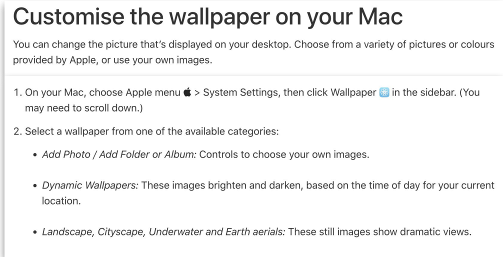
Finished steamroller!

https://www.lego.com/en-us/service/buildinginstructions/60401

Computer tasks in digital world

Task instruction: how do I change my Mac desktop background?

macOS environment



Customise the wallpaper on your Mac

You can change the picture that's displayed on your desktop. Choose from a variety of pictures or colours provided by Apple, or use your own images.

1. On your Mac, choose Apple menu > System Settings, then click Wallpaper in the sidebar. (You may need to scroll down.)
2. Select a wallpaper from one of the available categories:
    - *Add Photo / Add Folder or Album:* Controls to choose your own images.
    - *Dynamic Wallpapers:* These images brighten and darken, based on the time of day for your current location.
    - *Landscape, Cityscape, Underwater and Earth aerials:* These still images show dramatic views.

•••

Control instructions

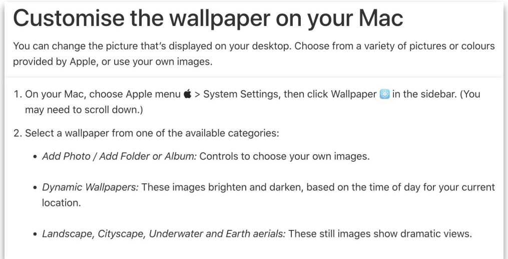
macOS w. new wallpaper!

https://support.apple.com/en-hk/guide/mac-help/mchlp3013/mac

Grounding plans into actions in digital world

Task instruction: how do I change my Mac desktop background?

macOS environment



Customise the wallpaper on your Mac

You can change the picture that's displayed on your desktop. Choose from a variety of pictures or colours provided by Apple, or use your own images.

1. On your Mac, choose Apple menu > System Settings, then click Wallpaper in the sidebar. (You may need to scroll down.)
2. Select a wallpaper from one of the available categories:
   - *Add Photo / Add Folder or Album:* Controls to choose your own images.
   - *Dynamic Wallpapers:* These images brighten and darken, based on the time of day for your current location.
   - *Landscape, Cityscape, Underwater and Earth aerials:* These still images show dramatic views.

•••

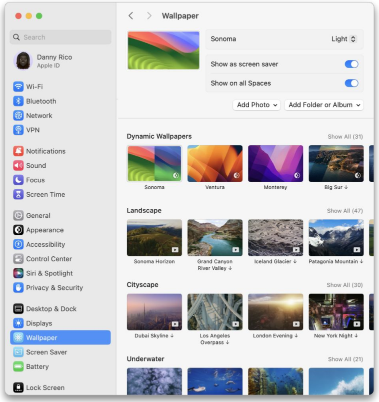
Control instructions

Grounding

Actions

mouse

keyboard

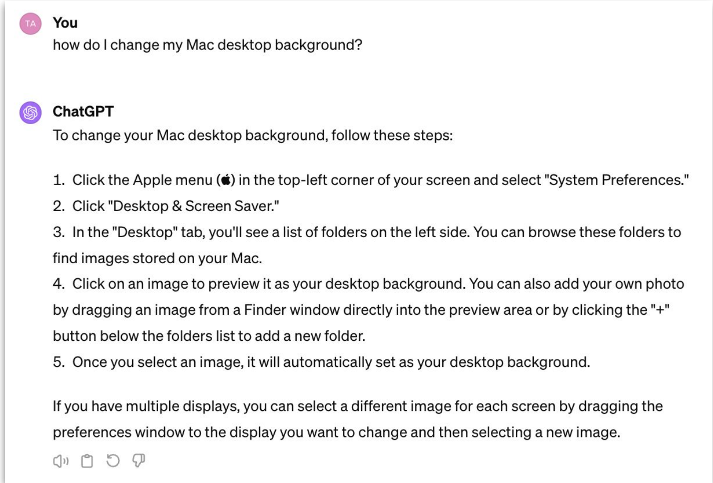


macOS w. new wallpaper!

https://support.apple.com/en-hk/guide/mac-help/mchlp3013/mac

Can LLMs/VLMs be used for these tasks?

Yes and **No!**

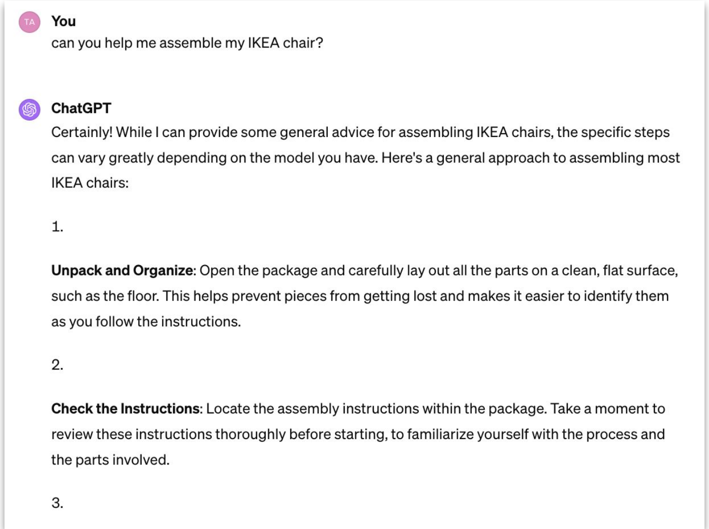


You
how do I change my Mac desktop background?

ChatGPT
To change your Mac desktop background, follow these steps:

1. Click the Apple menu () in the top-left corner of your screen and select "System Preferences."
2. Click "Desktop & Screen Saver."
3. In the "Desktop" tab, you'll see a list of folders on the left side. You can browse these folders to find images stored on your Mac.
4. Click on an image to preview it as your desktop background. You can also add your own photo by dragging an image from a Finder window directly into the preview area or by clicking the "+" button below the folders list to add a new folder.
5. Once you select an image, it will automatically set as your desktop background.

If you have multiple displays, you can select a different image for each screen by dragging the preferences window to the display you want to change and then selecting a new image.



You
can you help me assemble my IKEA chair?

ChatGPT
Certainly! While I can provide some general advice for assembling IKEA chairs, the specific steps can vary greatly depending on the model you have. Here's a general approach to assembling most IKEA chairs:

1.

Unpack and Organize: Open the package and carefully lay out all the parts on a clean, flat surface, such as the floor. This helps prevent pieces from getting lost and makes it easier to identify them as you follow the instructions.

2.

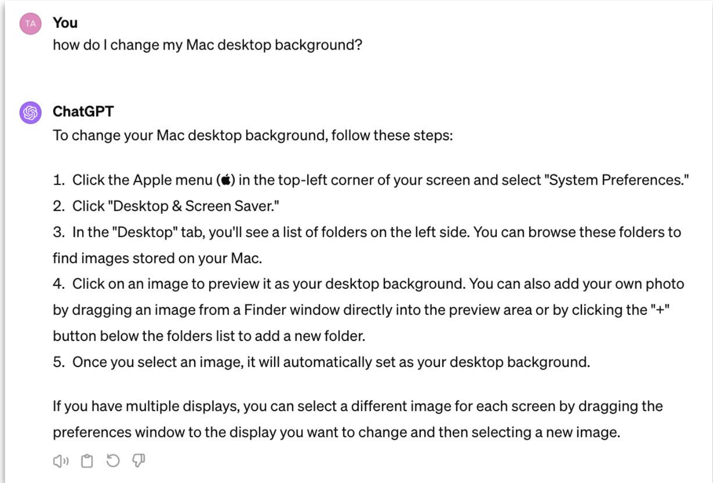
Check the Instructions: Locate the assembly instructions within the package. Take a moment to review these instructions thoroughly before starting, to familiarize yourself with the process and the parts involved.

3.

Can LLMs/VLMs be used for these tasks?

Yes and No!

ChatGPT is able to generate step-by-step plans



You

how do I change my Mac desktop background?

ChatGPT

To change your Mac desktop background, follow these steps:

1. Click the Apple menu () in the top-left corner of your screen and select "System Preferences."
2. Click "Desktop & Screen Saver."
3. In the "Desktop" tab, you'll see a list of folders on the left side. You can browse these folders to find images stored on your Mac.
4. Click on an image to preview it as your desktop background. You can also add your own photo by dragging an image from a Finder window directly into the preview area or by clicking the "+" button below the folders list to add a new folder.
5. Once you select an image, it will automatically set as your desktop background.

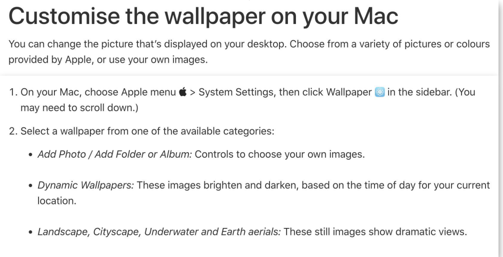
If you have multiple displays, you can select a different image for each screen by dragging the preferences window to the display you want to change and then selecting a new image.

Task instruction:
how do I change my Mac desktop background?

↓

ChatGPT

↓



Customise the wallpaper on your Mac

You can change the picture that's displayed on your desktop. Choose from a variety of pictures or colours provided by Apple, or use your own images.

1. On your Mac, choose Apple menu > System Settings, then click Wallpaper in the sidebar. (You may need to scroll down.)
2. Select a wallpaper from one of the available categories:
   - *Add Photo / Add Folder or Album:* Controls to choose your own images.
   - *Dynamic Wallpapers:* These images brighten and darken, based on the time of day for your current location.
   - *Landscape, Cityscape, Underwater and Earth aerials:* These still images show dramatic views.

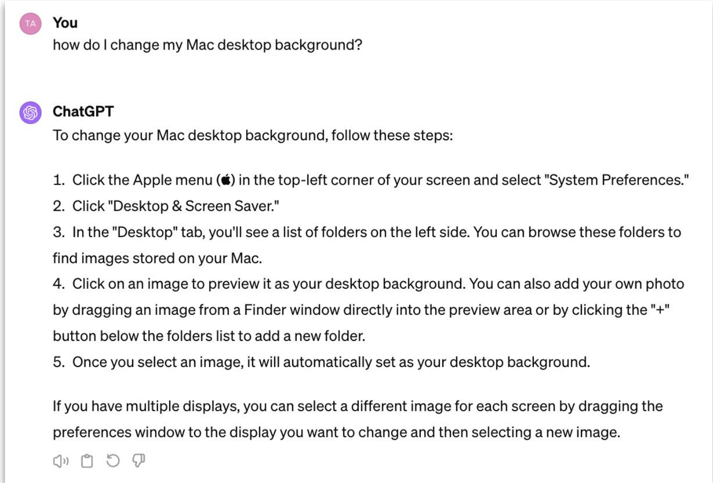
•••

Control instructions

Can LLMs/VLMs be used for these tasks?

Yes and **No!**

However, ChatGPT **cannot** execute tasks on your Mac by grounding plans into actions.



You

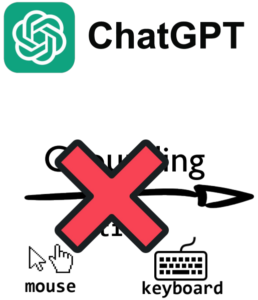
how do I change my Mac desktop background?

ChatGPT

To change your Mac desktop background, follow these steps:

1. Click the Apple menu () in the top-left corner of your screen and select "System Preferences."
2. Click "Desktop & Screen Saver."
3. In the "Desktop" tab, you'll see a list of folders on the left side. You can browse these folders to find images stored on your Mac.
4. Click on an image to preview it as your desktop background. You can also add your own photo by dragging an image from a Finder window directly into the preview area or by clicking the "+" button below the folders list to add a new folder.
5. Once you select an image, it will automatically set as your desktop background.

If you have multiple displays, you can select a different image for each screen by dragging the preferences window to the display you want to change and then selecting a new image.

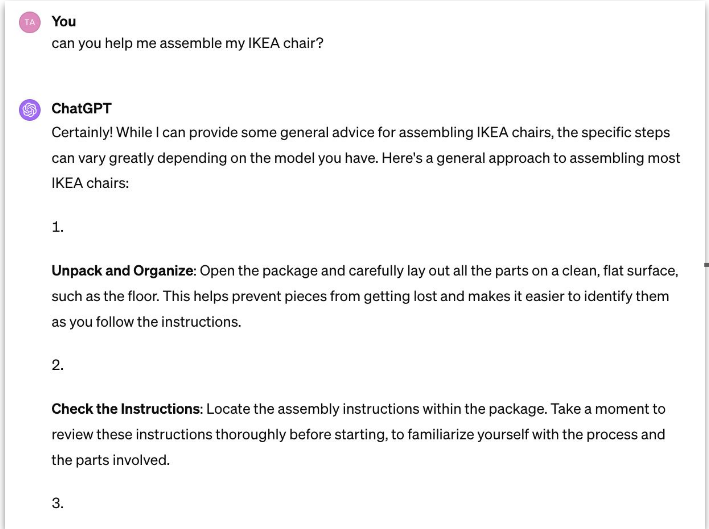


macOS w. new wallpaper!

Can LLMs/VLMs be used for these tasks?

Yes and **No!**

At times, ChatGPT also **cannot** generate step-by-step plans w.o interacting in the environment



You

can you help me assemble my IKEA chair?

ChatGPT

Certainly! While I can provide some general advice for assembling IKEA chairs, the specific steps can vary greatly depending on the model you have. Here's a general approach to assembling most IKEA chairs:

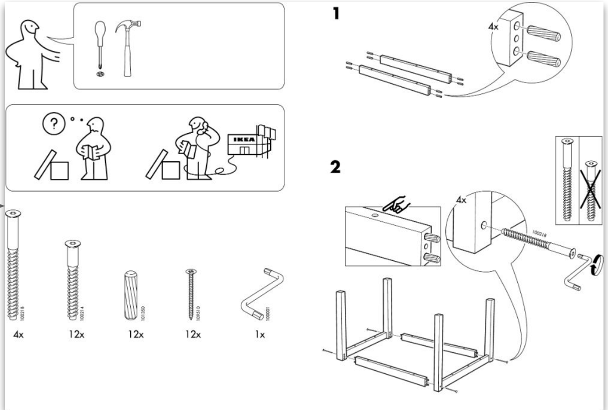
1.

Unpack and Organize: Open the package and carefully lay out all the parts on a clean, flat surface, such as the floor. This helps prevent pieces from getting lost and makes it easier to identify them as you follow the instructions.

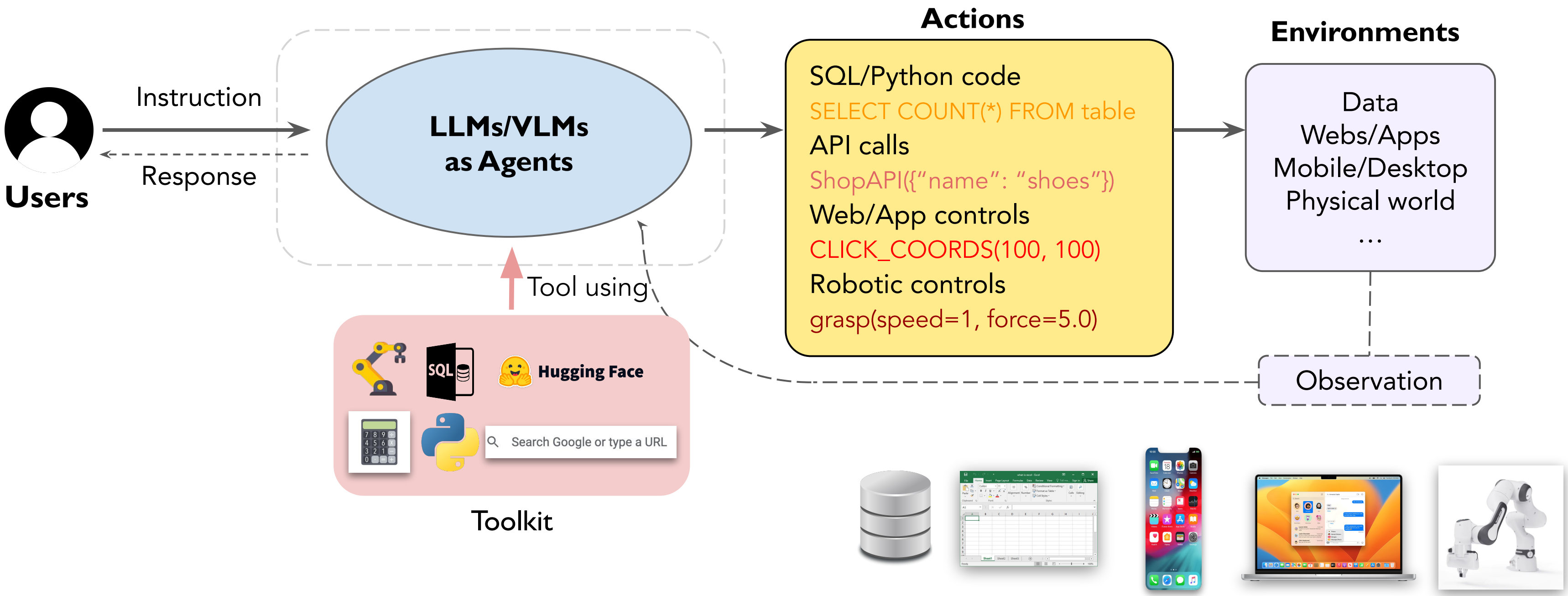
2.

Check the Instructions: Locate the assembly instructions within the package. Take a moment to review these instructions thoroughly before starting, to familiarize yourself with the process and the parts involved.

3.



Solution: LLMs/VLMs as agents



Users

Instruction

Response

LLMs/VLMs as Agents

Tool using

Toolkit

Actions

SQL/Python code

`SELECT COUNT(*) FROM table`

API calls

`ShopAPI({"name": "shoes"})`

Web/App controls

`CLICK_COORDS(100, 100)`

Robotic controls

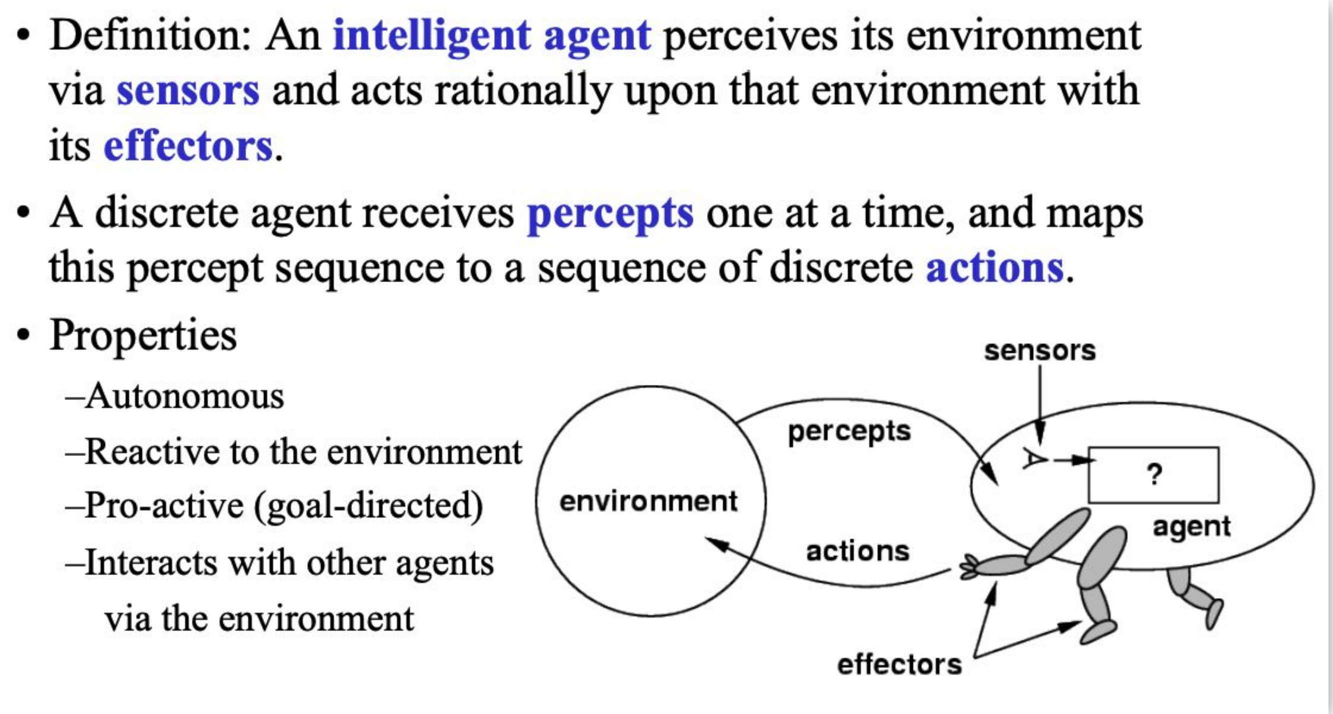
`grasp(speed=1, force=5.0)`

Environments

Data
Webs/Apps
Mobile/Desktop
Physical world
…

Observation

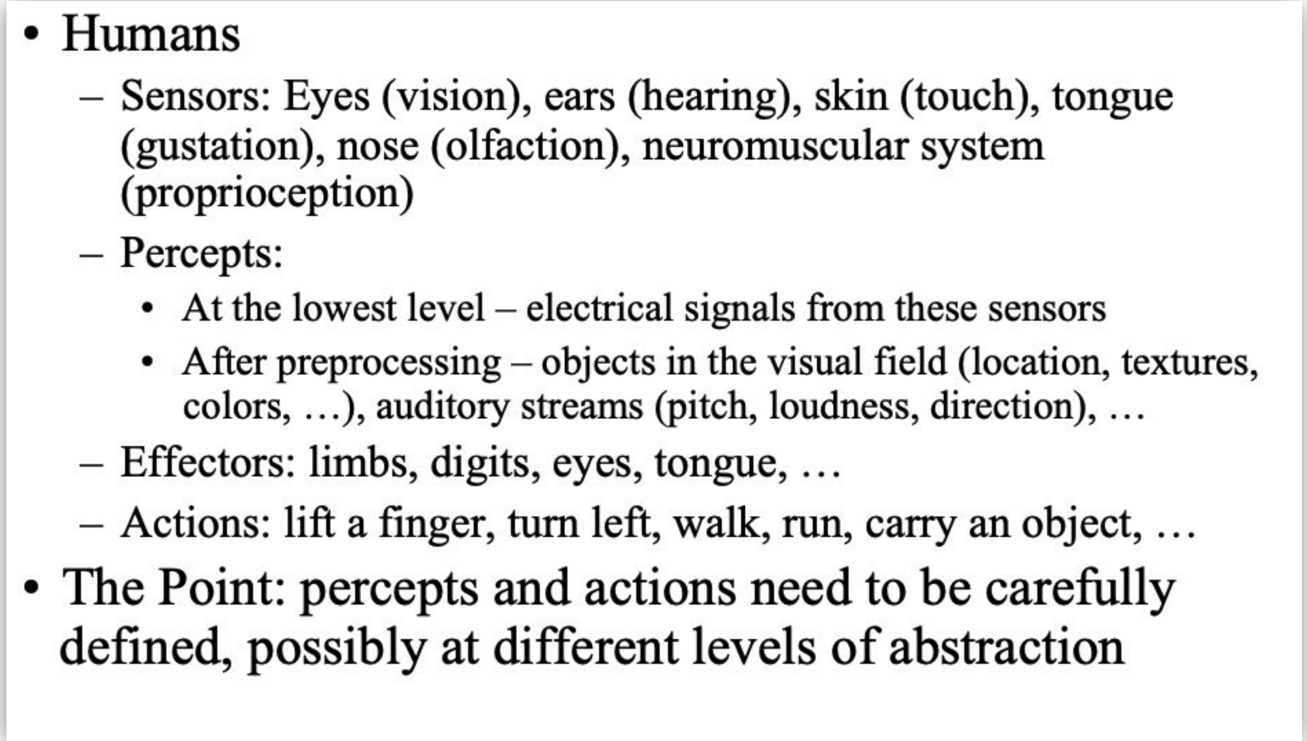
Wait…what is an intelligent agent?



- Definition: An **intelligent agent** perceives its environment via **sensors** and acts rationally upon that environment with its **effectors**.
- A discrete agent receives **percepts** one at a time, and maps this percept sequence to a sequence of discrete **actions**.
- Properties
  - –Autonomous
  - –Reactive to the environment
  - –Pro-active (goal-directed)
  - –Interacts with other agents via the environment

https://cs.brynmawr.edu/Courses/cs372/spring2012/slides/02_IntelligentAgents.pdf

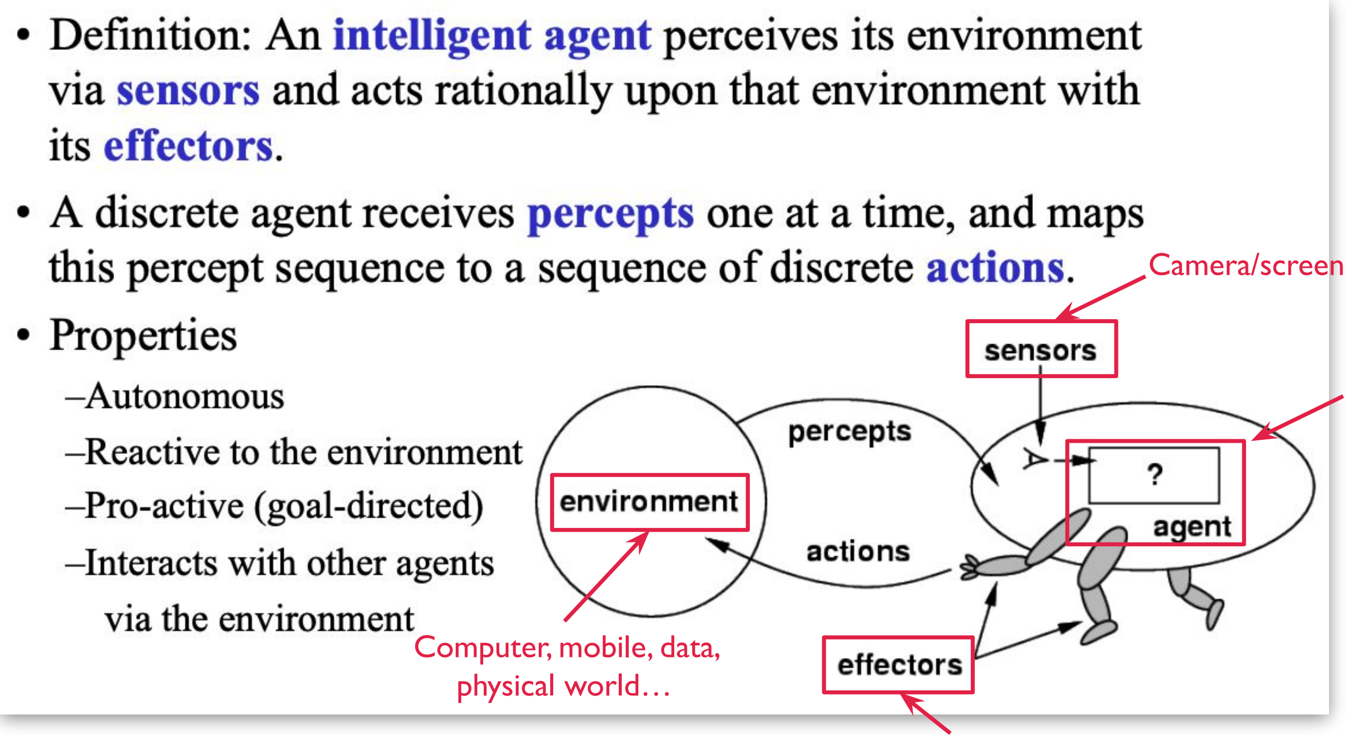
What are sensors/percepts and effectors/actions?



- Humans
  - – Sensors: Eyes (vision), ears (hearing), skin (touch), tongue (gustation), nose (olfaction), neuromuscular system (proprioception)
  - – Percepts:
    - At the lowest level – electrical signals from these sensors
    - After preprocessing – objects in the visual field (location, textures, colors, …), auditory streams (pitch, loudness, direction), …
  - – Effectors: limbs, digits, eyes, tongue, …
  - – Actions: lift a finger, turn left, walk, run, carry an object, …
- The Point: percepts and actions need to be carefully defined, possibly at different levels of abstraction

https://cs.brynmawr.edu/Courses/cs372/spring2012/slides/02_IntelligentAgents.pdf

LLMs/VLMs as agents



- Definition: An **intelligent agent** perceives its environment via **sensors** and acts rationally upon that environment with its **effectors**.
- A discrete agent receives **percepts** one at a time, and maps this percept sequence to a sequence of discrete **actions**.
- Properties
  - –Autonomous
  - –Reactive to the environment
  - –Pro-active (goal-directed)
  - –Interacts with other agents via the environment

Camera/screenshots…

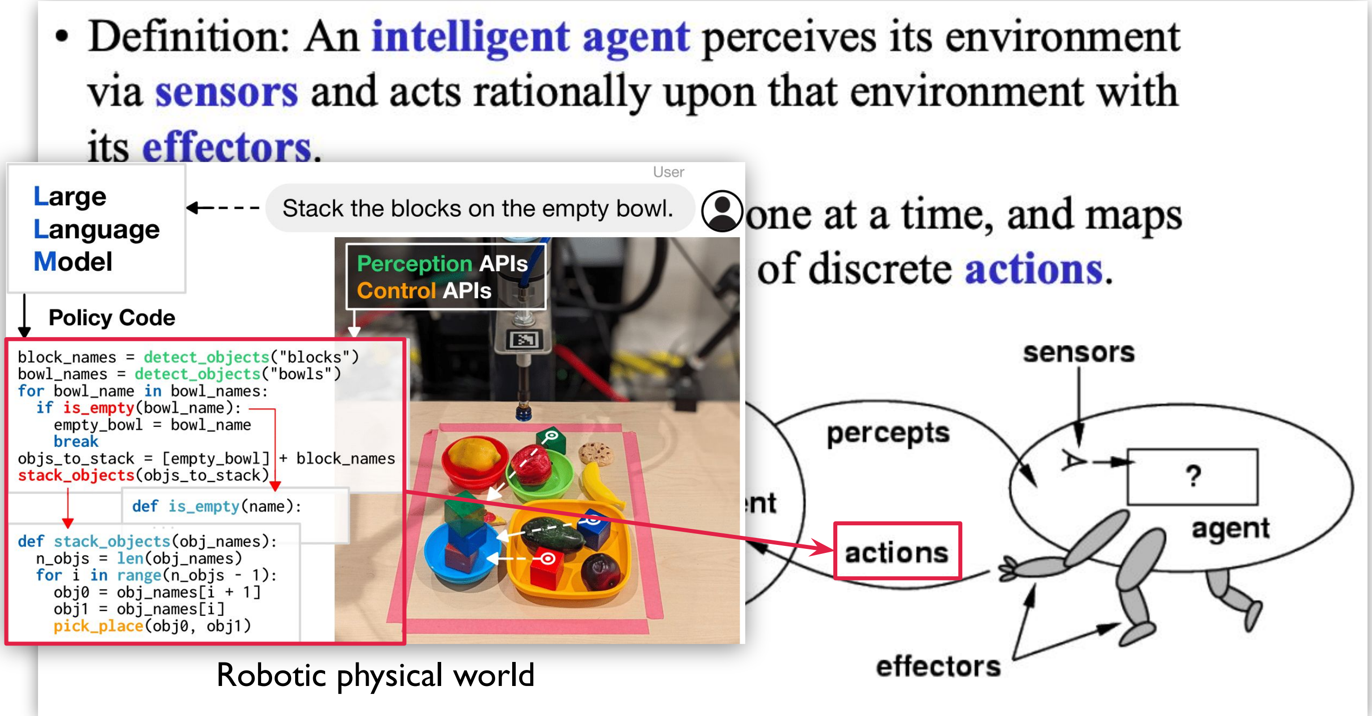
LLMs/VLMs as brains of agents

Computer, mobile, data, physical world…

Robots/interpreters…

https://cs.brynmawr.edu/Courses/cs372/spring2012/slides/02_IntelligentAgents.pdf

LLMs/VLMs as agents



- Definition: An **intelligent agent** perceives its environment via **sensors** and acts rationally upon that environment with its **effectors**.

one at a time, and maps of discrete **actions**.

User

Stack the blocks on the empty bowl.

Policy Code

```
block_names = detect_objects("blocks")
bowl_names = detect_objects("bowls")
for bowl_name in bowl_names:
  if is_empty(bowl_name):
    empty_bowl = bowl_name
    break
objs_to_stack = [empty_bowl] + block_names
stack_objects(objs_to_stack)
```

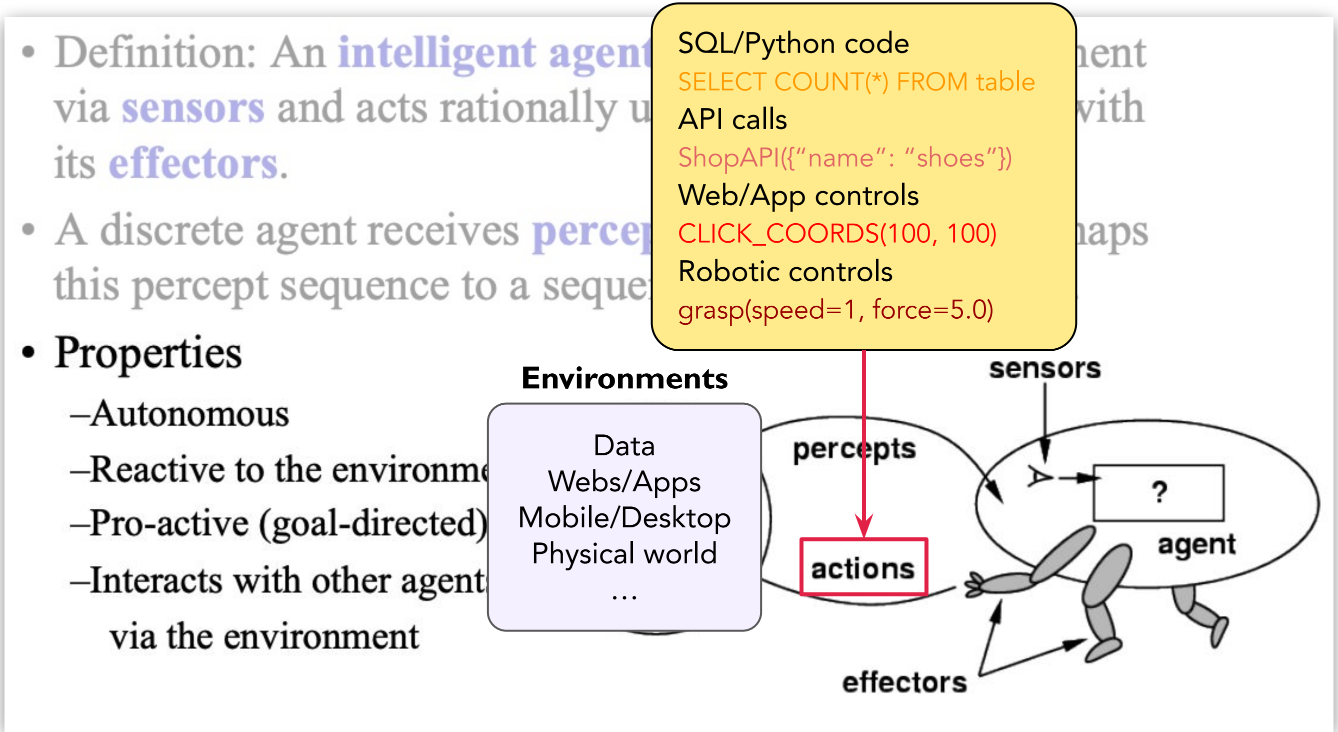
```
def is_empty(name):
  ...
```

```
def stack_objects(obj_names):
  n_objs = len(obj_names)
  for i in range(n_objs - 1):
    obj0 = obj_names[i + 1]
    obj1 = obj_names[i]
    pick_place(obj0, obj1)
```

Robotic physical world

https://cs.brynmawr.edu/Courses/cs372/spring2012/slides/02_IntelligentAgents.pdf
Code as Policies: Language Model Programs for Embodied Control

LLMs/VLMs as agents



- Definition: An **intelligent agent** ... nent via **sensors** and acts rationally u ... with its **effectors**.
- A discrete agent receives **percep** ... naps this percept sequence to a seque
- Properties
  - –Autonomous
  - –Reactive to the environme
  - –Pro-active (goal-directed)
  - –Interacts with other agent via the environment

sensors
percepts
?
agent
actions
effectors

Actions

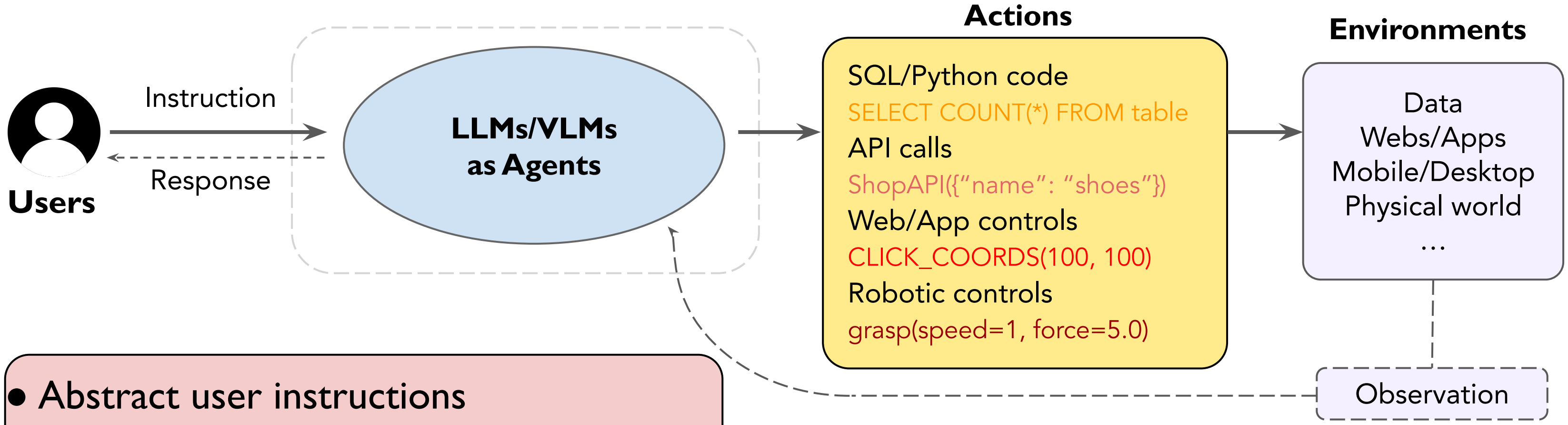
SQL/Python code
SELECT COUNT(*) FROM table
API calls
ShopAPI({"name": "shoes"})
Web/App controls
CLICK_COORDS(100, 100)
Robotic controls
grasp(speed=1, force=5.0)

Environments

Data
Webs/Apps
Mobile/Desktop
Physical world
...

https://cs.brynmawr.edu/Courses/cs372/spring2012/slides/02_IntelligentAgents.pdf

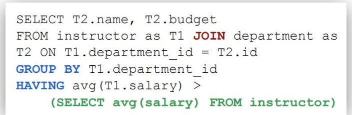
LLMs/VLMs as agents



- Abstract user instructions
- Utilize tools to expand capacities
- Explore complex unseen environments
- Multiple step planning and reasoning
- Follow feedback and self debug

Why LLMs/VLMs as agents?

Current popular interfaces



```
SELECT T2.name, T2.budget
FROM instructor as T1 JOIN department as
T2 ON T1.department_id = T2.id
GROUP BY T1.department_id
HAVING avg(T1.salary) >
    (SELECT avg(salary) FROM instructor)
```

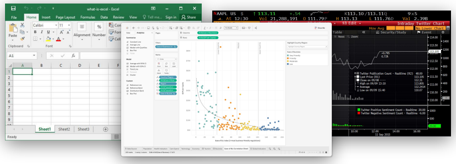
1. Programming languages

Users

Digital world

Data
Webs/Apps
Mobile/Desktop
…

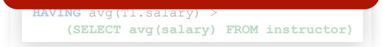
2. Graphical user interfaces



Why LLMs/VLMs as agents?

Problems with current popular interfaces

👎Hard to learn



```
HAVING avg(T1.salary) >
    (SELECT avg(salary) FROM instructor)
```

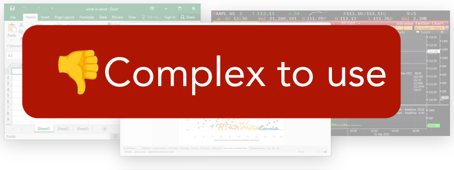
1. Programming languages

Users

Digital world

Data
Webs/Apps
Mobile/Desktop
…

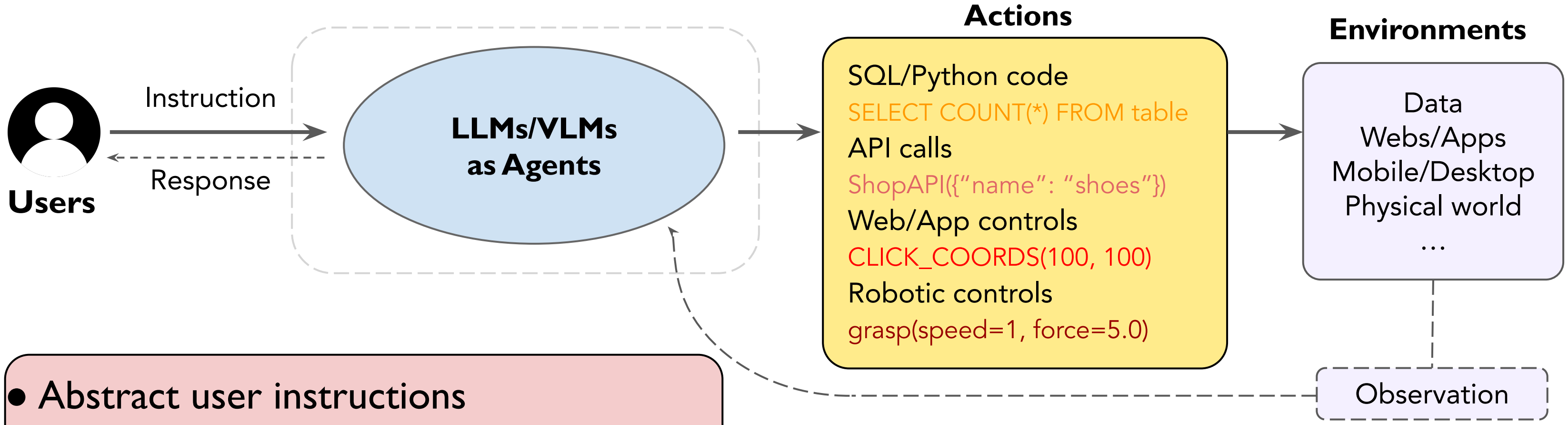
2. Graphical user interfaces



👎Complex to use

Why LLMs/VLMs as agents?

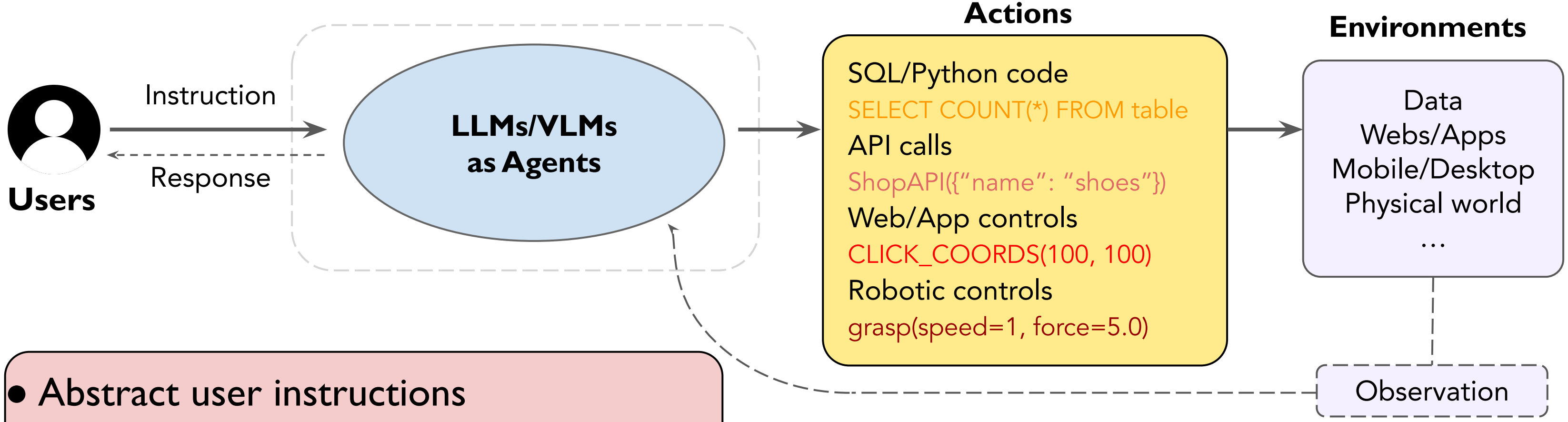
LLMs/VLMs as agents enable users to interact with environments through natural language, making the process more intuitive and accessible.



- Abstract user instructions
- Utilize tools to expand capacities
- Explore complex unseen environments
- Multiple step planning and reasoning
- Follow feedback and self debug

LLMs/VLMs as agents

Key research direction: transforming (“grounding”) language instructions into code or actions executable in real-world environments



Users

Instruction

Response

LLMs/VLMs as Agents

Actions

SQL/Python code

`SELECT COUNT(*) FROM table`

API calls

`ShopAPI({“name”: “shoes”})`

Web/App controls

`CLICK_COORDS(100, 100)`

Robotic controls

`grasp(speed=1, force=5.0)`

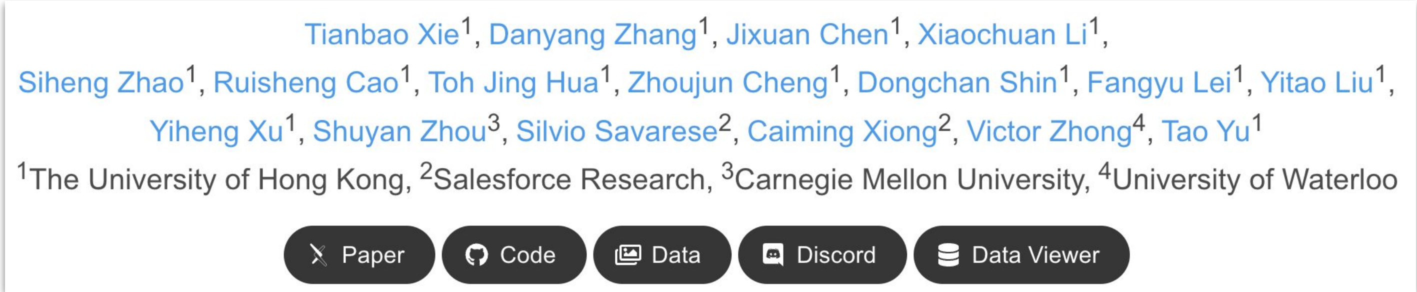
Environments

Data
Webs/Apps
Mobile/Desktop
Physical world
…

Observation

- Abstract user instructions
- Utilize tools to expand capacities
- Explore complex unseen environments
- Multiple step planning and reasoning
- Follow feedback and self debug

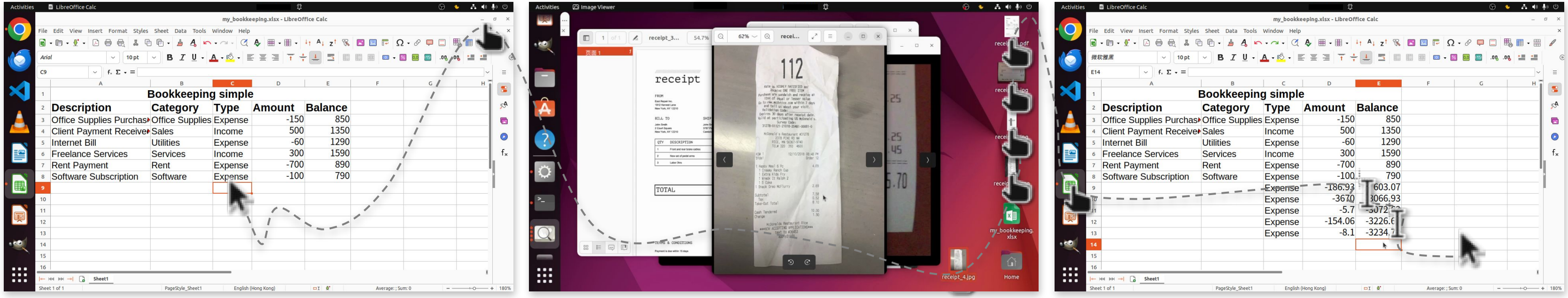
OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments


Tianbao Xie[1], Danyang Zhang[1], Jixuan Chen[1], Xiaochuan Li[1],
Siheng Zhao[1], Ruisheng Cao[1], Toh Jing Hua[1], Zhoujun Cheng[1], Dongchan Shin[1], Fangyu Lei[1], Yitao Liu[1],
Yiheng Xu[1], Shuyan Zhou[3], Silvio Savarese[2], Caiming Xiong[2], Victor Zhong[4], Tao Yu[1]
[1]The University of Hong Kong, [2]Salesforce Research, [3]Carnegie Mellon University, [4]University of Waterloo

Project website: https://os-world.github.io

OSWorld task examples

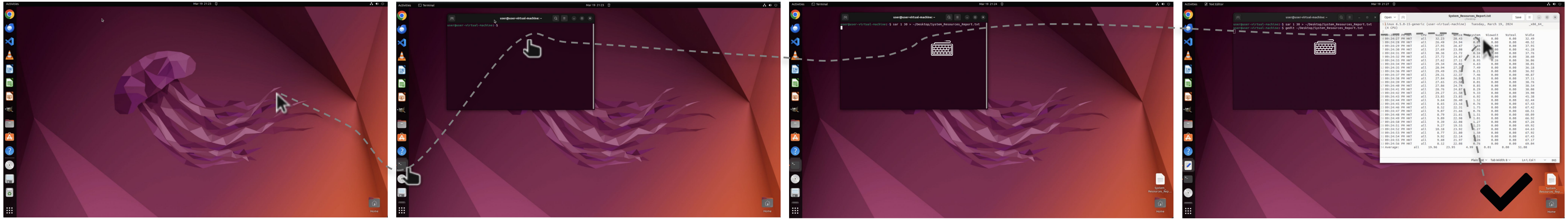
Task instruction 1: Update the bookkeeping sheet with my recent transactions over the past few days in the provided folder.



my_bookkeeping.xlsx - LibreOffice Calc

	A	B	C	D	E
1	Bookkeeping simple				
2	Description	Category	Type	Amount	Balance
3	Office Supplies Purchas	Office Supplies	Expense	-150	850
4	Client Payment Receive	Sales	Income	500	1350
5	Internet Bill	Utilities	Expense	-60	1290
6	Freelance Services	Services	Income	300	1590
7	Rent Payment	Rent	Expense	-700	890
8	Software Subscription	Software	Expense	-100	790
9			Expense	-186.93	603.07
10			Expense	-3670	3066.93
11			Expense	-5.7	-3072.[illegible]
12			Expense	-154.06	-3226.[illegible]
13			Expense	-8.1	-3234.[illegible]

Task Instruction 2: monitor the system CPU for 30s and output the results



```
user@user-virtual-machine:~$ sar 1 30 > ~/Desktop/System_Resources_Report.txt
user@user-virtual-machine:~$ gedit ~/Desktop/System_Resources_Report.txt
```

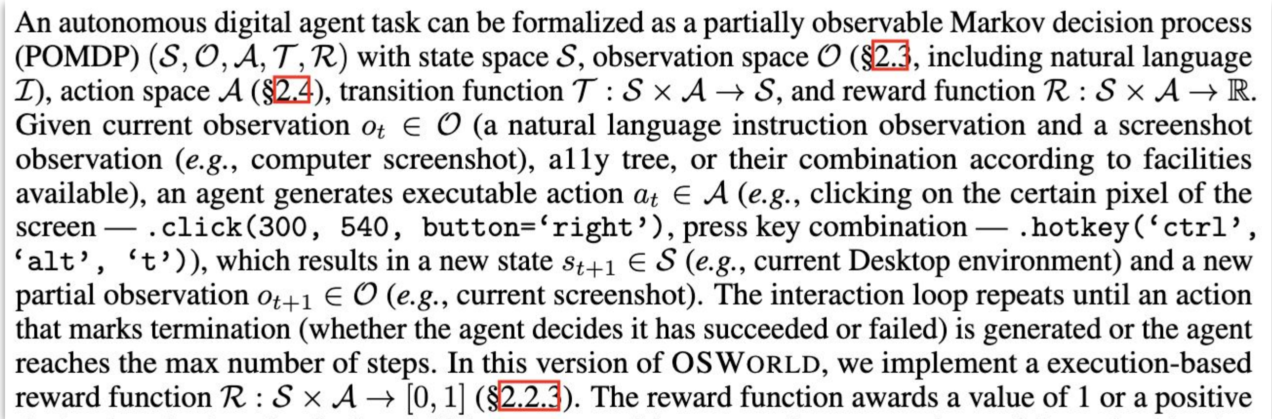
pyautogui.click(terminal_x, terminal_y)

pyautogui.click(focus_x, focus_y)

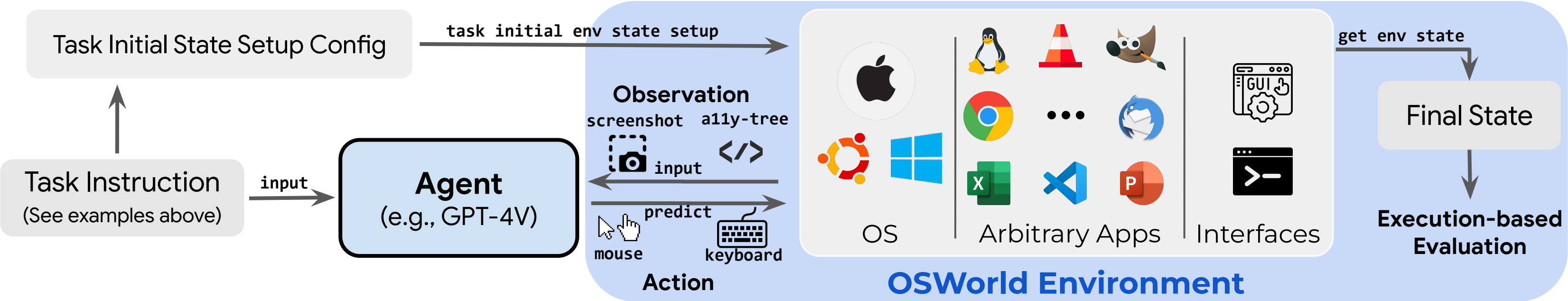
pyautogui.typewrite('sar 1 30 > …', interval=0.5)

Done

OSWorld agent task definition

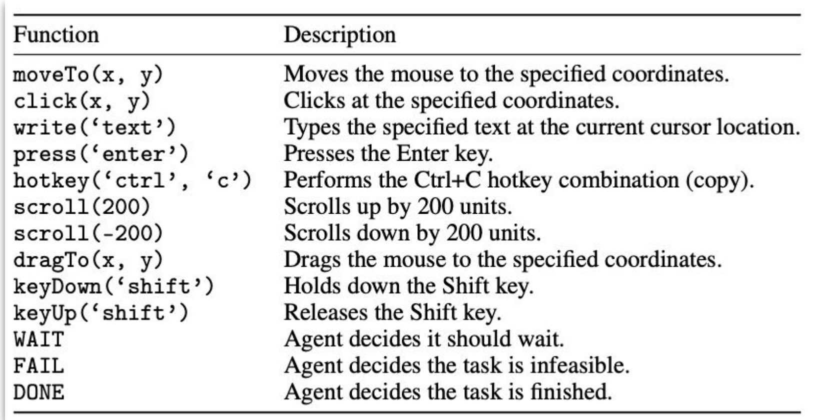


An autonomous digital agent task can be formalized as a partially observable Markov decision process (POMDP) $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{R})$ with state space $\mathcal{S}$, observation space $\mathcal{O}$ (§2.3, including natural language $\mathcal{I}$), action space $\mathcal{A}$ (§2.4), transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, and reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. Given current observation $o_t \in \mathcal{O}$ (a natural language instruction observation and a screenshot observation (*e.g.*, computer screenshot), `a11y` tree, or their combination according to facilities available), an agent generates executable action $a_t \in \mathcal{A}$ (*e.g.*, clicking on the certain pixel of the screen — `.click(300, 540, button=‘right’)`, press key combination — `.hotkey(‘ctrl’, ‘alt’, ‘t’)`), which results in a new state $s_{t+1} \in \mathcal{S}$ (*e.g.*, current Desktop environment) and a new partial observation $o_{t+1} \in \mathcal{O}$ (*e.g.*, current screenshot). The interaction loop repeats until an action that marks termination (whether the agent decides it has succeeded or failed) is generated or the agent reaches the max number of steps. In this version of OSWORLD, we implement a execution-based reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ (§2.2.3). The reward function awards a value of 1 or a positive

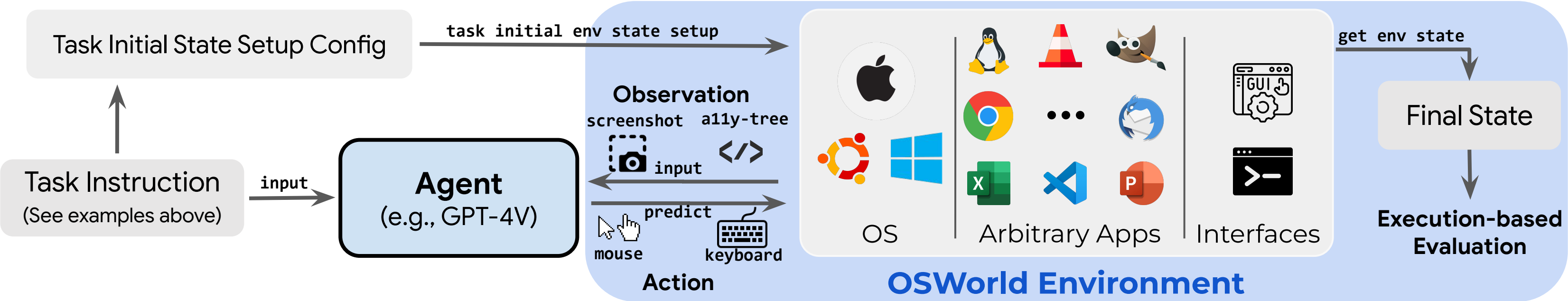


OSWorld agent task action space

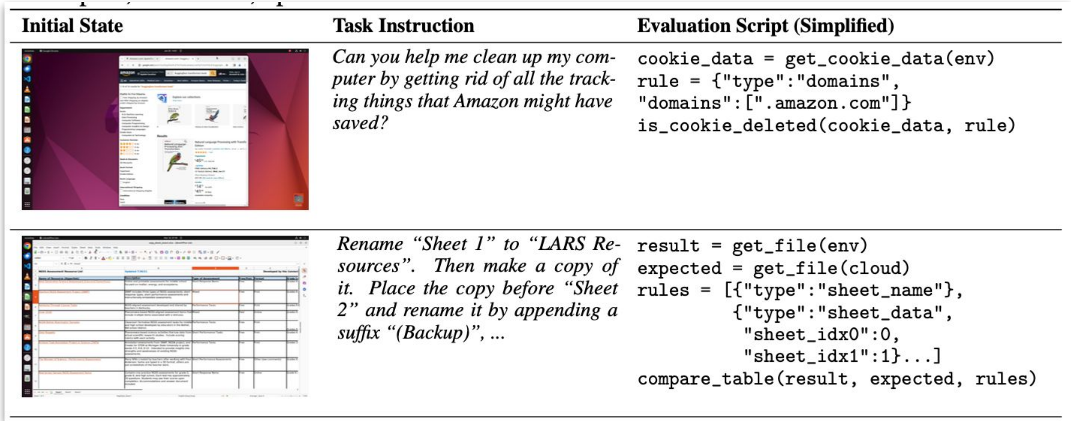
Some examples of the
mouse and keyboard actions



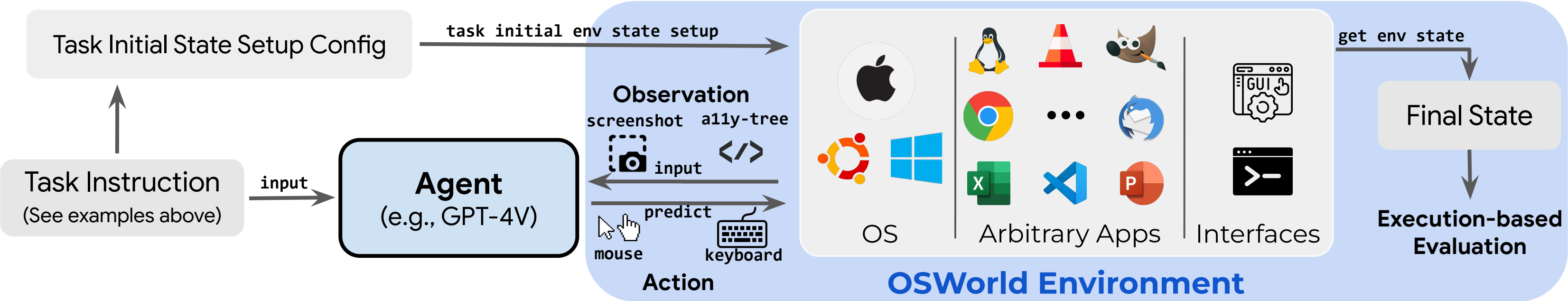
Function	Description
`moveTo(x, y)`	Moves the mouse to the specified coordinates.
`click(x, y)`	Clicks at the specified coordinates.
`write('text')`	Types the specified text at the current cursor location.
`press('enter')`	Presses the Enter key.
`hotkey('ctrl', 'c')`	Performs the Ctrl+C hotkey combination (copy).
`scroll(200)`	Scrolls up by 200 units.
`scroll(-200)`	Scrolls down by 200 units.
`dragTo(x, y)`	Drags the mouse to the specified coordinates.
`keyDown('shift')`	Holds down the Shift key.
`keyUp('shift')`	Releases the Shift key.
`WAIT`	Agent decides it should wait.
`FAIL`	Agent decides the task is infeasible.
`DONE`	Agent decides the task is finished.



OSWorld agent task evaluation



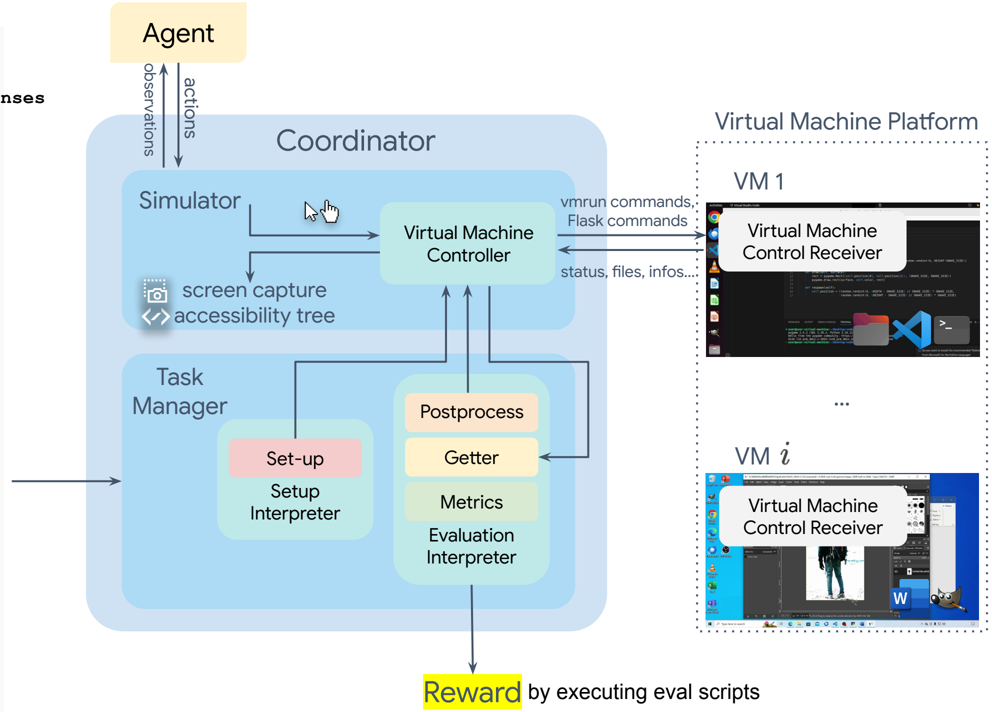
<table>
<tr><th>Initial State</th><th>Task Instruction</th><th>Evaluation Script (Simplified)</th></tr>
<tr><td></td><td><em>Can you help me clean up my computer by getting rid of all the tracking things that Amazon might have saved?</em></td><td><pre>cookie_data = get_cookie_data(env)
rule = {"type":"domains",
"domains":[".amazon.com"]}
is_cookie_deleted(cookie_data, rule)</pre></td></tr>
<tr><td></td><td><em>Rename "Sheet 1" to "LARS Resources". Then make a copy of it. Place the copy before "Sheet 2" and rename it by appending a suffix "(Backup)", ...</em></td><td><pre>result = get_file(env)
expected = get_file(cloud)
rules = [{"type":"sheet_name"},
         {"type":"sheet_data",
          "sheet_idx0":0,
          "sheet_idx1":1}...]
compare_table(result, expected, rules)</pre></td></tr>
</table>



OSWorld environment

Config

```
{ "instruction": "Please update my bookkeeping sheet with the
recent transactions from the provided folder, detailing my expenses
over the past few days.",
  "config": [{"type": "download",
      "parameters": {"files": [
{"path": "/home/user/Desktop/my_bookkeeping.xlsx",
 "url": "https://drive.google.com/uc?id=xxxx"},
{"path": "/home/user/Desktop/receipt_0.jpeg",
 "url": "https://drive.google.com/uc?id=xxxx"},…]}},
    {"type": "open",
      "parameters": {  "path":
"/home/user/Desktop/my_bookkeeping.xlsx"}}],
   "evaluator": {"postconfig": [{"type": "activate_window",
      "parameters": {"window_name": "my_bookkeeping.xlsx -
LibreOffice Calc",... ],
    "result": {"type": "vm_file",
      "path": "/home/user/Desktop/my_bookkeeping.xlsx",
      "dest": "my_bookkeeping.xlsx"},
    "expected": {"type": "cloud_file",
      "path": "https://drive.google.com/uc?id=xxx",
      "dest": "my_bookkeeping_gold.xlsx" },
    "func": "compare_table",
    "options": {
      "rules": [{
          "type": "sheet_fuzzy",
          "sheet_idx0": "RNSheet1",
          "sheet_idx1": "ENSheet1",
          "rules": [ {"range": [ "A1:A8",... }]}]
}
```



OSWorld benchmark dataset

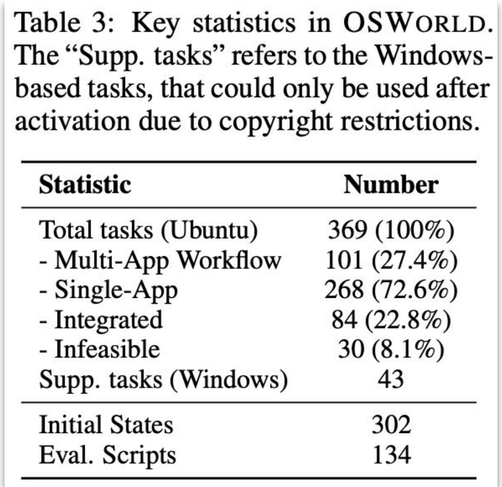


Table 3: Key statistics in OSWORLD. The "Supp. tasks" refers to the Windows-based tasks, that could only be used after activation due to copyright restrictions.

Statistic	Number
Total tasks (Ubuntu)	369 (100%)
- Multi-App Workflow	101 (27.4%)
- Single-App	268 (72.6%)
- Integrated	84 (22.8%)
- Infeasible	30 (8.1%)
Supp. tasks (Windows)	43
Initial States	302
Eval. Scripts	134

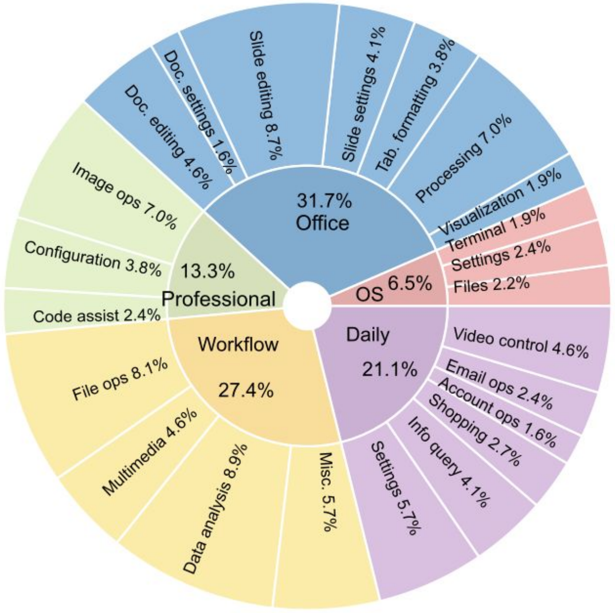
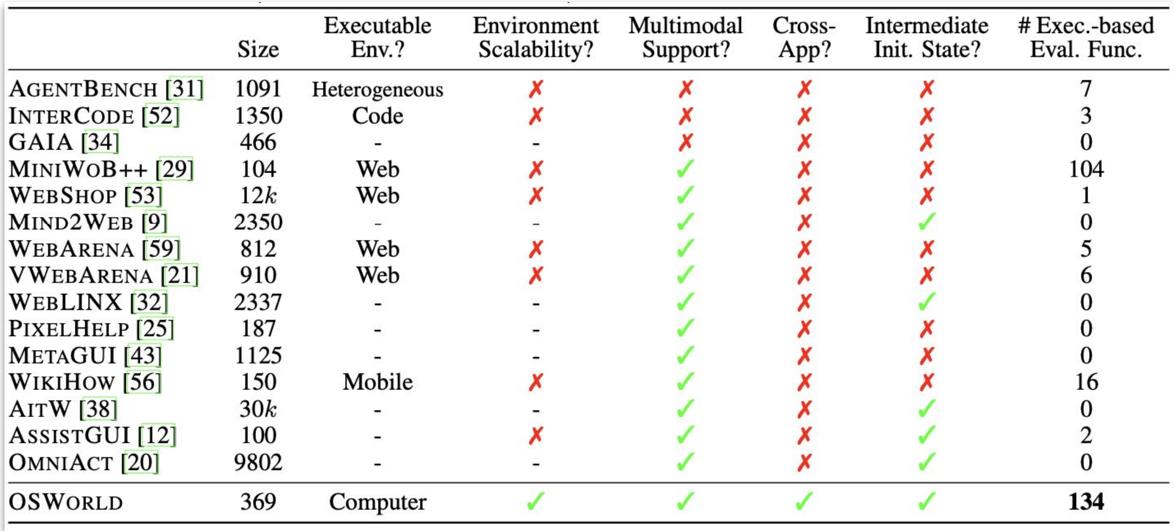


Figure 3: Distribution of task instructions in OS-WORLD based on the app domains and operation types to showcase the content intuitively.

OSWorld benchmark dataset



	Size	Executable Env.?	Environment Scalability?	Multimodal Support?	Cross-App?	Intermediate Init. State?	# Exec.-based Eval. Func.
AgentBench [31]	1091	Heterogeneous	✗	✗	✗	✗	7
InterCode [52]	1350	Code	✗	✗	✗	✗	3
GAIA [34]	466	-	-	✗	✗	✗	0
MiniWoB++ [29]	104	Web	✗	✓	✗	✗	104
WebShop [53]	12*k*	Web	✗	✓	✗	✗	1
Mind2Web [9]	2350	-	-	✓	✗	✓	0
WebArena [59]	812	Web	✗	✓	✗	✗	5
VWebArena [21]	910	Web	✗	✓	✗	✗	6
WebLINX [32]	2337	-	-	✓	✗	✓	0
PixelHelp [25]	187	-	-	✓	✗	✗	0
MetaGUI [43]	1125	-	-	✓	✗	✗	0
WikiHow [56]	150	Mobile	✗	✓	✗	✗	16
AitW [38]	30*k*	-	-	✓	✗	✓	0
AssistGUI [12]	100	-	✗	✓	✗	✓	2
OmniAct [20]	9802	-	-	✓	✗	✓	0
OSWorld	369	Computer	✓	✓	✓	✓	**134**

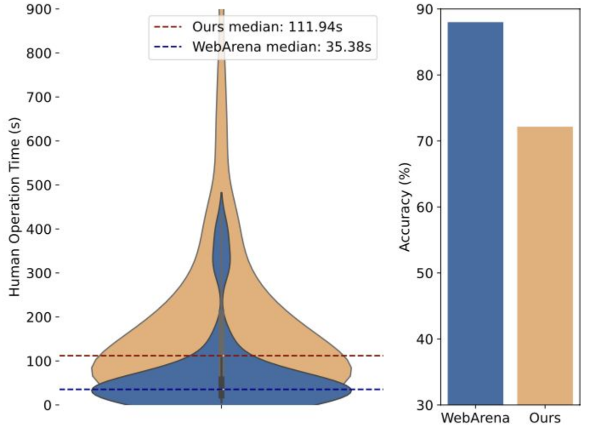
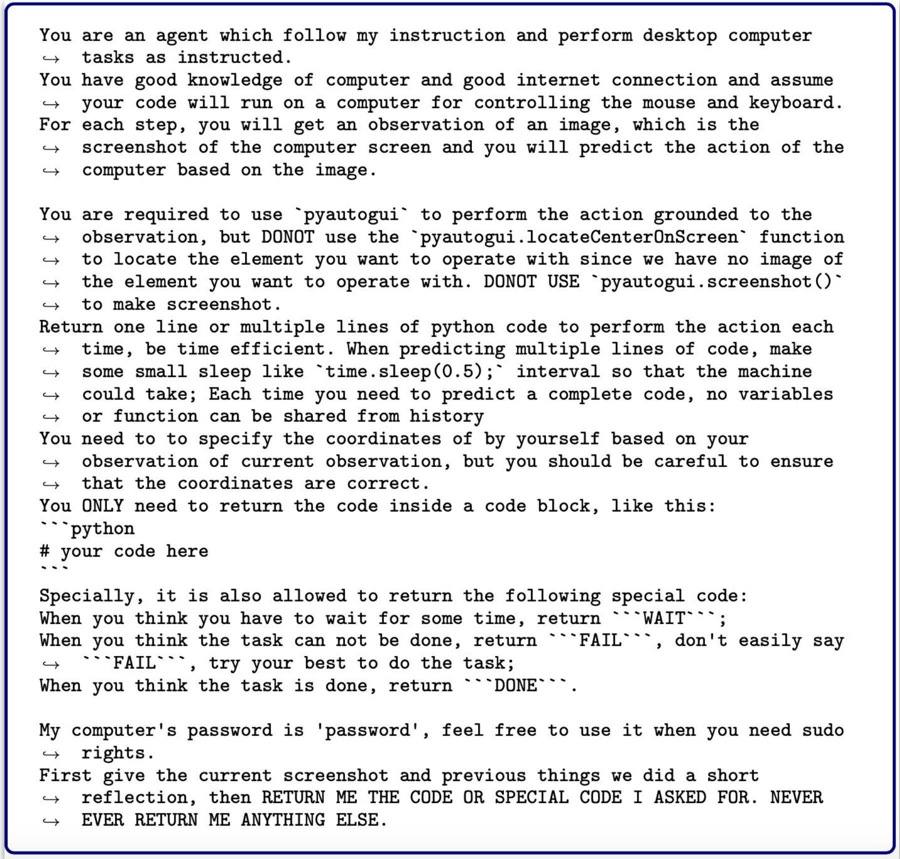


Figure 4: Human operation time and accuracy on OSWorld and WebArena.

LLM/VLM agent baselines

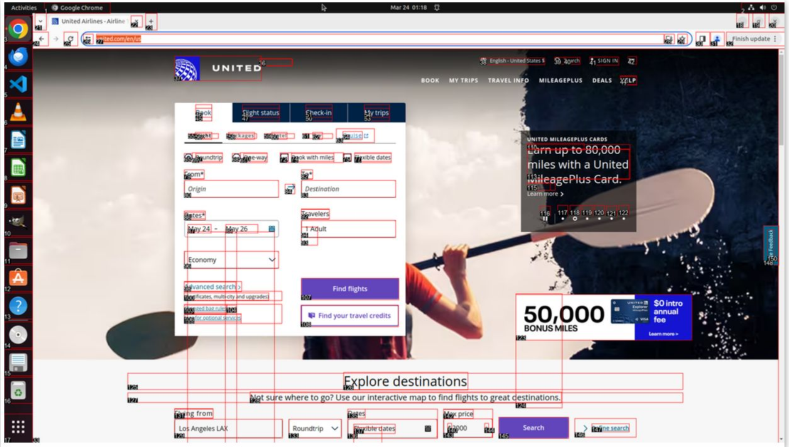
- LLMs and VLMs from Mixtral and CogAgent (open-source), and GPT4, Gemini-pro, and Claude-3 (closed-source) as agents.
- Prompt details (see right - much more complex prompting pipelines)
- Temperature of 1.0 and top-p of 0.9
- Providing the most recent 3 observations and actions as history context for each step.



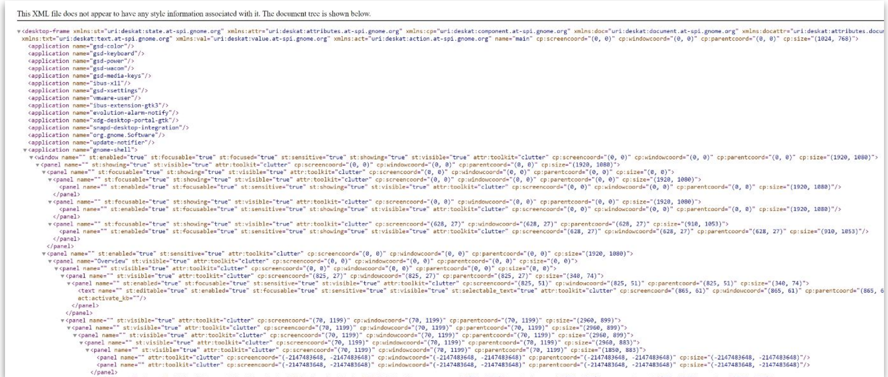
~~~
You are an agent which follow my instruction and perform desktop computer
↪  tasks as instructed.
You have good knowledge of computer and good internet connection and assume
↪  your code will run on a computer for controlling the mouse and keyboard.
For each step, you will get an observation of an image, which is the
↪  screenshot of the computer screen and you will predict the action of the
↪  computer based on the image.

You are required to use `pyautogui` to perform the action grounded to the
↪  observation, but DONOT use the `pyautogui.locateCenterOnScreen` function
↪  to locate the element you want to operate with since we have no image of
↪  the element you want to operate with. DONOT USE `pyautogui.screenshot()`
↪  to make screenshot.
Return one line or multiple lines of python code to perform the action each
↪  time, be time efficient. When predicting multiple lines of code, make
↪  some small sleep like `time.sleep(0.5);` interval so that the machine
↪  could take; Each time you need to predict a complete code, no variables
↪  or function can be shared from history
You need to to specify the coordinates of by yourself based on your
↪  observation of current observation, but you should be careful to ensure
↪  that the coordinates are correct.
You ONLY need to return the code inside a code block, like this:
```python
# your code here
```
Specially, it is also allowed to return the following special code:
When you think you have to wait for some time, return ```WAIT```;
When you think the task can not be done, return ```FAIL```, don't easily say
↪  ```FAIL```, try your best to do the task;
When you think the task is done, return ```DONE```.

My computer's password is 'password', feel free to use it when you need sudo
↪  rights.
First give the current screenshot and previous things we did a short
↪  reflection, then RETURN ME THE CODE OR SPECIAL CODE I ASKED FOR. NEVER
↪  EVER RETURN ME ANYTHING ELSE.
~~~

Input settings for LLM/VLM agent baselines

- Accessibility tree
- Screenshot
- Screenshot + accessibility tree
- Set-of-Marks



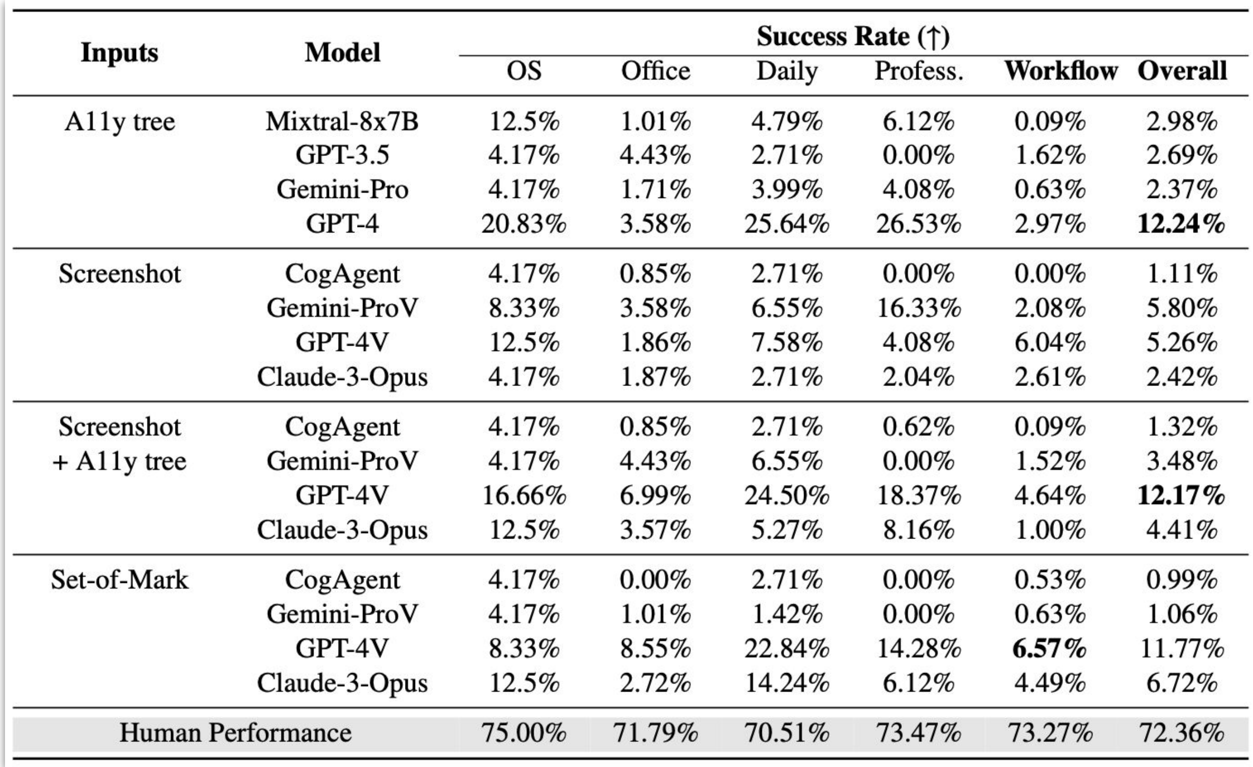
Set-of-Marks


This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<desktop-frame xmlns:st="uri:deskat:state.at-spi.gnome.org" xmlns:attr="uri:deskat:attributes.at-spi.gnome.org" xmlns:cp="uri:deskat:component.at-spi.gnome.org" xmlns:doc="uri:deskat:document.at-spi.gnome.org" xmlns:docattr="uri:deskat:attributes.docu
 xmlns:txt="uri:deskat:text.at-spi.gnome.org" xmlns:val="uri:deskat:value.at-spi.gnome.org" xmlns:act="uri:deskat:action.at-spi.gnome.org" name="main" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1024, 768)">
   <application name="gsd-color"/>
   <application name="gsd-keyboard"/>
   <application name="gsd-power"/>
   <application name="gsd-wacom"/>
   <application name="gsd-media-keys"/>
   <application name="ibus-x11"/>
   <application name="gsd-xsettings"/>
   <application name="vmware-user"/>
   <application name="ibus-extension-gtk3"/>
   <application name="evolution-alarm-notify"/>
   <application name="xdg-desktop-portal-gtk"/>
   <application name="snapd-desktop-integration"/>
   <application name="org.gnome.Software"/>
   <application name="update-notifier"/>
 ▼<application name="gnome-shell">
   ▼<window name="" st:enabled="true" st:focusable="true" st:focused="true" st:sensitive="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1920, 1080)">
     ▼<panel name="" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1920, 1080)">
       ▼<panel name="" st:focusable="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(0, 0)">
         ▼<panel name="" st:focusable="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1920, 1080)">
             <panel name="" st:enabled="true" st:focusable="true" st:sensitive="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1920, 1080)"/>
           </panel>
         ▼<panel name="" st:focusable="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1920, 1080)">
             <panel name="" st:enabled="true" st:focusable="true" st:sensitive="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1920, 1080)"/>
           </panel>
         ▼<panel name="" st:focusable="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(628, 27)" cp:windowcoord="(628, 27)" cp:parentcoord="(628, 27)" cp:size="(910, 1053)">
             <panel name="" st:enabled="true" st:focusable="true" st:sensitive="true" st:showing="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(628, 27)" cp:windowcoord="(628, 27)" cp:parentcoord="(628, 27)" cp:size="(910, 1053)"/>
           </panel>
         </panel>
       ▼<panel name="" st:enabled="true" st:sensitive="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(1920, 1080)">
         ▼<panel name="Overview" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(0, 0)">
           ▼<panel name="" st:visible="true" attr:toolkit="clutter" cp:screencoord="(0, 0)" cp:windowcoord="(0, 0)" cp:parentcoord="(0, 0)" cp:size="(0, 0)">
             ▼<panel name="" st:visible="true" attr:toolkit="clutter" cp:screencoord="(825, 27)" cp:windowcoord="(825, 27)" cp:parentcoord="(825, 27)" cp:size="(340, 74)">
               ▼<panel name="" st:enabled="true" st:focusable="true" st:sensitive="true" st:visible="true" attr:toolkit="clutter" cp:screencoord="(825, 51)" cp:windowcoord="(825, 51)" cp:parentcoord="(825, 51)" cp:size="(340, 74)">
                   <text name="" st:editable="true" st:enabled="true" st:focusable="true" st:sensitive="true" st:visible="true" st:selectable_text="true" attr:toolkit="clutter" cp:screencoord="(865, 61)" cp:windowcoord="(865, 61)" cp:parentcoord="(865, 6
                   act:activate_kb=""/>
                 </panel>
               </panel>
             ▼<panel name="" st:visible="true" attr:toolkit="clutter" cp:screencoord="(70, 1199)" cp:windowcoord="(70, 1199)" cp:parentcoord="(70, 1199)" cp:size="(2960, 899)">
               ▼<panel name="" st:visible="true" attr:toolkit="clutter" cp:screencoord="(70, 1199)" cp:windowcoord="(70, 1199)" cp:parentcoord="(70, 1199)" cp:size="(2960, 899)">
                 ▼<panel name="" st:visible="true" attr:toolkit="clutter" cp:screencoord="(70, 1199)" cp:windowcoord="(70, 1199)" cp:parentcoord="(70, 1199)" cp:size="(2960, 899)">
                   ▼<panel name="" st:visible="true" attr:toolkit="clutter" cp:screencoord="(70, 1199)" cp:windowcoord="(70, 1199)" cp:parentcoord="(70, 1199)" cp:size="(2960, 883)">
                     ▼<panel name="" st:visible="true" attr:toolkit="clutter" cp:screencoord="(70, 1199)" cp:windowcoord="(70, 1199)" cp:parentcoord="(70, 1199)" cp:size="(1850, 883)">
                         <panel name="" attr:toolkit="clutter" cp:screencoord="(-2147483648, -2147483648)" cp:windowcoord="(-2147483648, -2147483648)" cp:parentcoord="(-2147483648, -2147483648)" cp:size="(-2147483648, -2147483648)"/>
                         <panel name="" attr:toolkit="clutter" cp:screencoord="(-2147483648, -2147483648)" cp:windowcoord="(-2147483648, -2147483648)" cp:parentcoord="(-2147483648, -2147483648)" cp:size="(-2147483648, -2147483648)"/>
                       </panel>
```

Accessibility tree

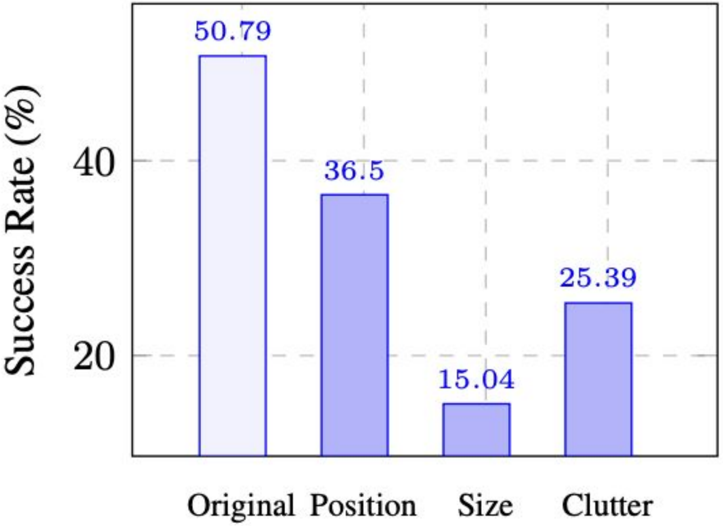
Results of LLM/VLM agent baselines



Inputs	**Model**	**Success Rate (↑)**					
		OS	Office	Daily	Profess.	**Workflow**	**Overall**
A11y tree	Mixtral-8x7B	12.5%	1.01%	4.79%	6.12%	0.09%	2.98%
	GPT-3.5	4.17%	4.43%	2.71%	0.00%	1.62%	2.69%
	Gemini-Pro	4.17%	1.71%	3.99%	4.08%	0.63%	2.37%
	GPT-4	20.83%	3.58%	25.64%	26.53%	2.97%	**12.24%**
Screenshot	CogAgent	4.17%	0.85%	2.71%	0.00%	0.00%	1.11%
	Gemini-ProV	8.33%	3.58%	6.55%	16.33%	2.08%	5.80%
	GPT-4V	12.5%	1.86%	7.58%	4.08%	6.04%	5.26%
	Claude-3-Opus	4.17%	1.87%	2.71%	2.04%	2.61%	2.42%
Screenshot + A11y tree	CogAgent	4.17%	0.85%	2.71%	0.62%	0.09%	1.32%
	Gemini-ProV	4.17%	4.43%	6.55%	0.00%	1.52%	3.48%
	GPT-4V	16.66%	6.99%	24.50%	18.37%	4.64%	**12.17%**
	Claude-3-Opus	12.5%	3.57%	5.27%	8.16%	1.00%	4.41%
Set-of-Mark	CogAgent	4.17%	0.00%	2.71%	0.00%	0.53%	0.99%
	Gemini-ProV	4.17%	1.01%	1.42%	0.00%	0.63%	1.06%
	GPT-4V	8.33%	8.55%	22.84%	14.28%	**6.57%**	11.77%
	Claude-3-Opus	12.5%	2.72%	14.24%	6.12%	4.49%	6.72%
Human Performance		75.00%	71.79%	70.51%	73.47%	73.27%	72.36%

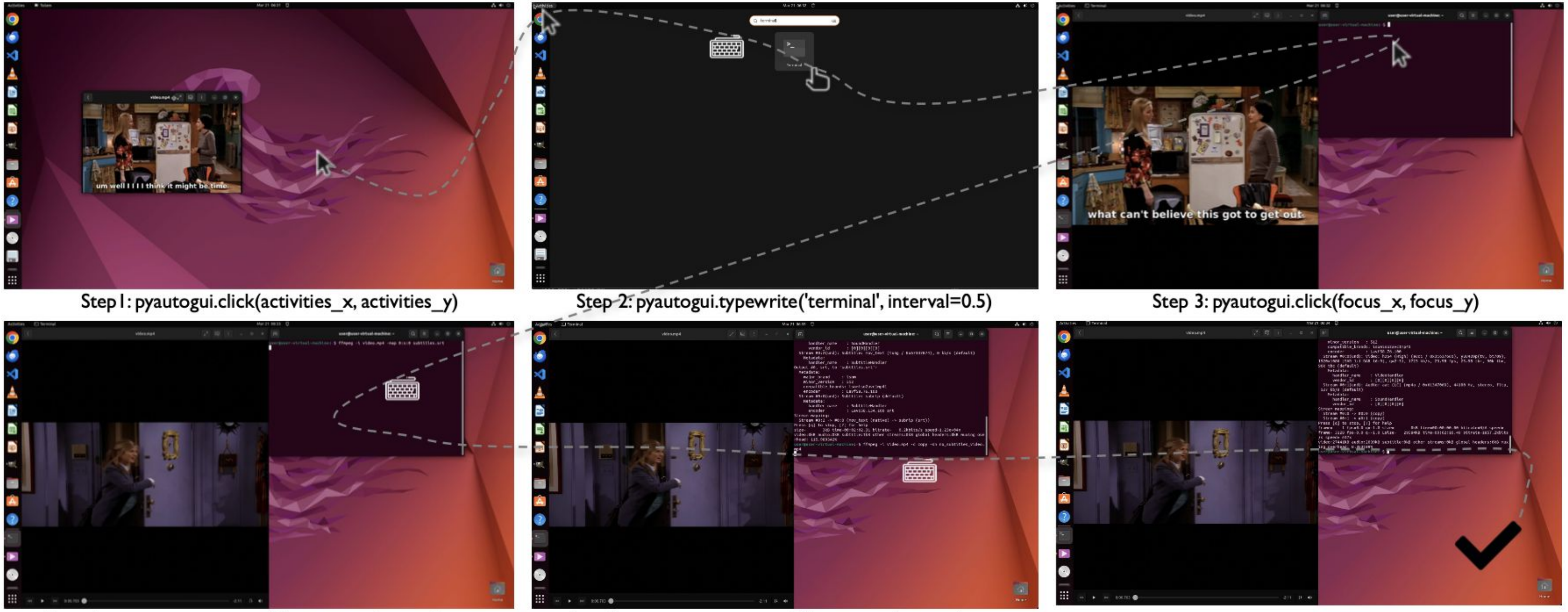
Analysis of LLM/VLM agent baselines

- Analysis example 1:VLM agents struggle with perturbation of position and size of application windows and irrelevant information.
- See paper for more interesting analysis.



Success case of LLM/VLM agent baselines

Task Instruction: I downloaded an episode of Friends to practice listening, but I don't know how to remove the subtitles. Please help me remove the subtitles from the video and export it as "subtitles.srt" and store it in the same directory as the video.


um well I I I think it might be time
what can't believe this got to get out

Step 1: pyautogui.click(activities_x, activities_y)

Step 2: pyautogui.typewrite('terminal', interval=0.5)

Step 3: pyautogui.click(focus_x, focus_y)

Step 4: pyautogui.typewrite('ffmpeg -i video.mp4 -map 0:s:0 subtitles.srt', interval=0.5)

Step 5: pyautogui.typewrite('ffmpeg -i video.mp4 -c copy -sn no_subtitles_video.mp4', interval=0.5)

Step 6: Done