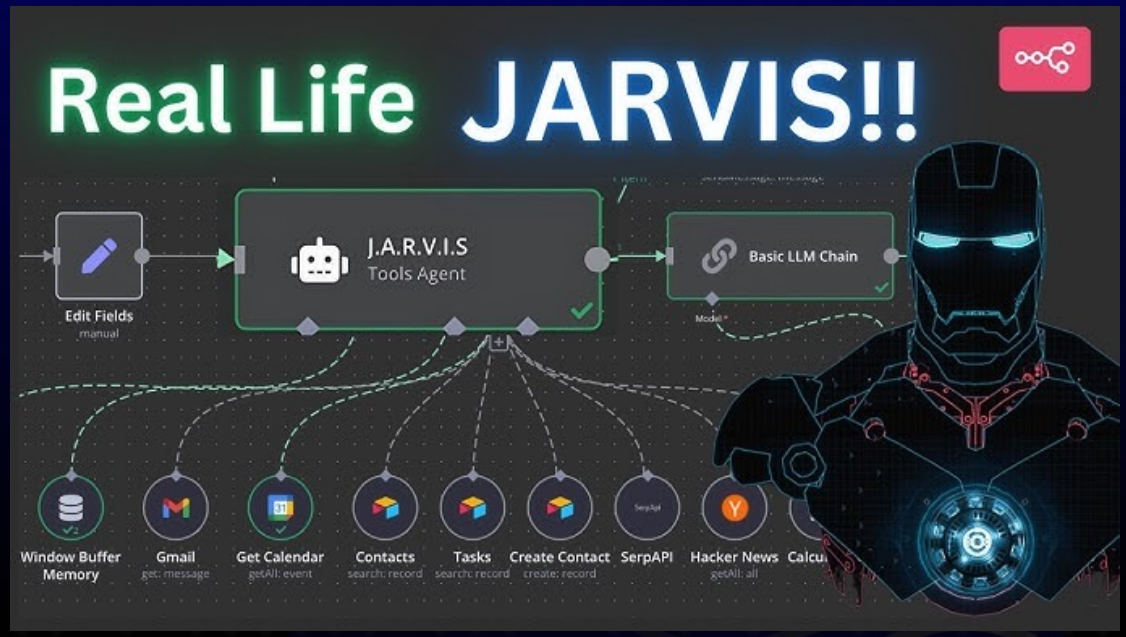
AI Insight Talk

OpenMMLab | Hugging Face | ModelScope 魔搭社区 | 知乎 | 机智流

From Digital Agents to AI Co-Scientists

Qiushi Sun
qiushisun.github.io
𝕏 @qiushi_sun

Digital Agents



The Feasibility of Jarvis AI from Marvel in Real Life

[1] OS-Copilot: Towards Generalist Computer Agents with Self-Improvement

Computer-Using Agents

Once out of reach, but we are turning it into reality.

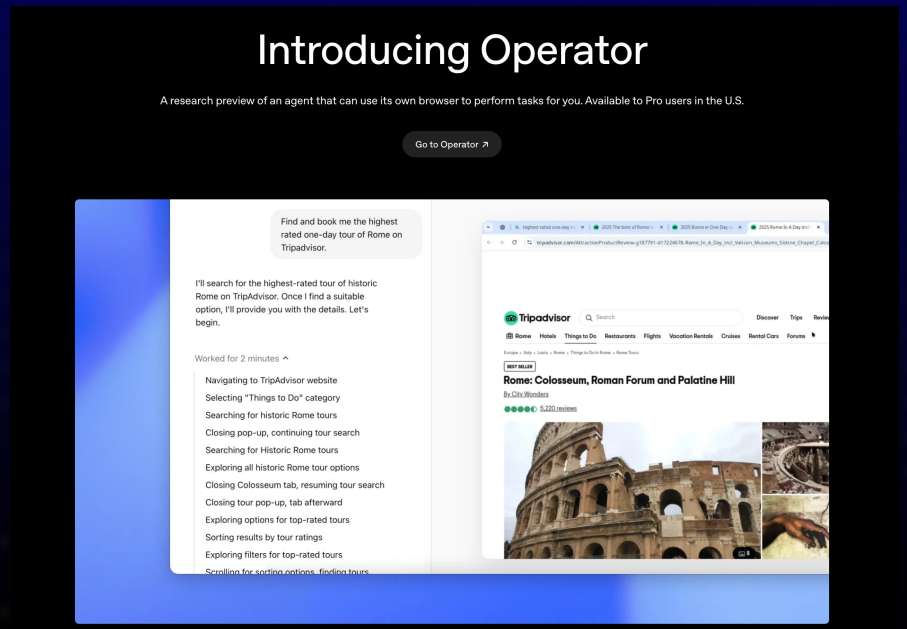
Computer-Using Agents

Both academia and industry are building computer-using agents

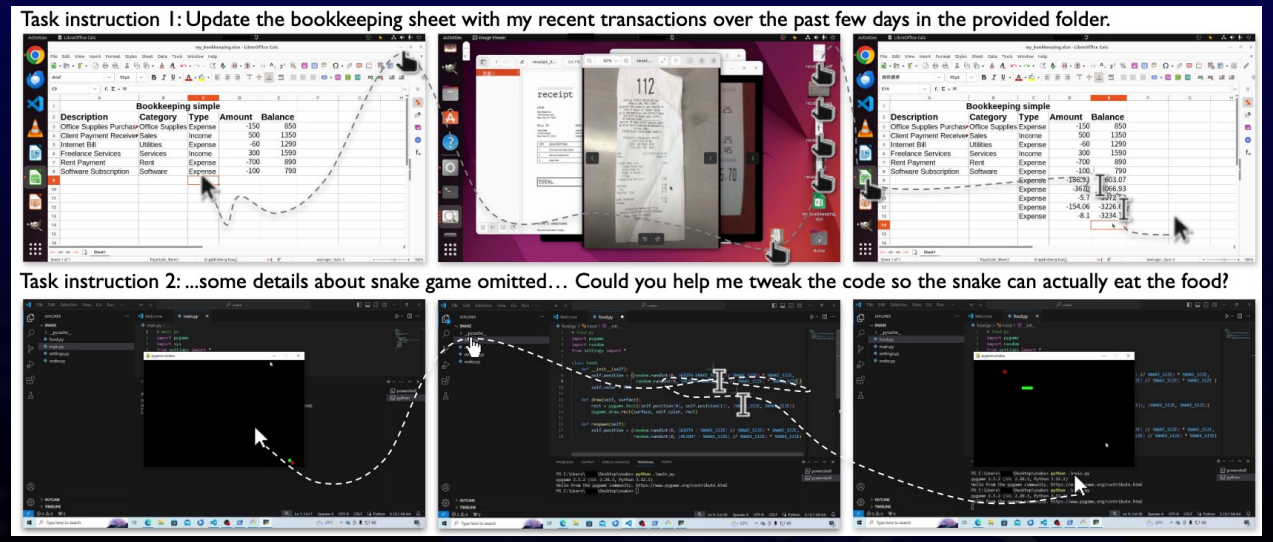
[2] Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku, 22 Oct 2024

Computer-Using Agents

Automating daily computer tasks



OpenAI Operator



Daily Computer Use

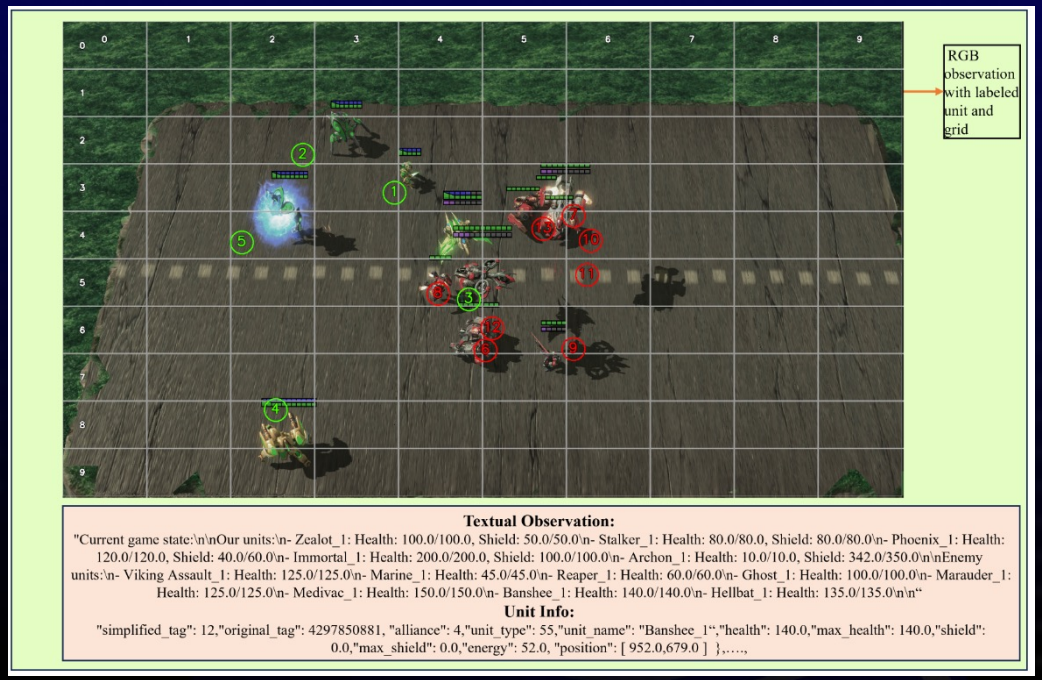
[3] Introducing Operator: A research preview of an agent that can use its own browser to perform tasks for you., Jan 23, 2025

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

Computer-Using Agents

Playing Games.

MineCraft

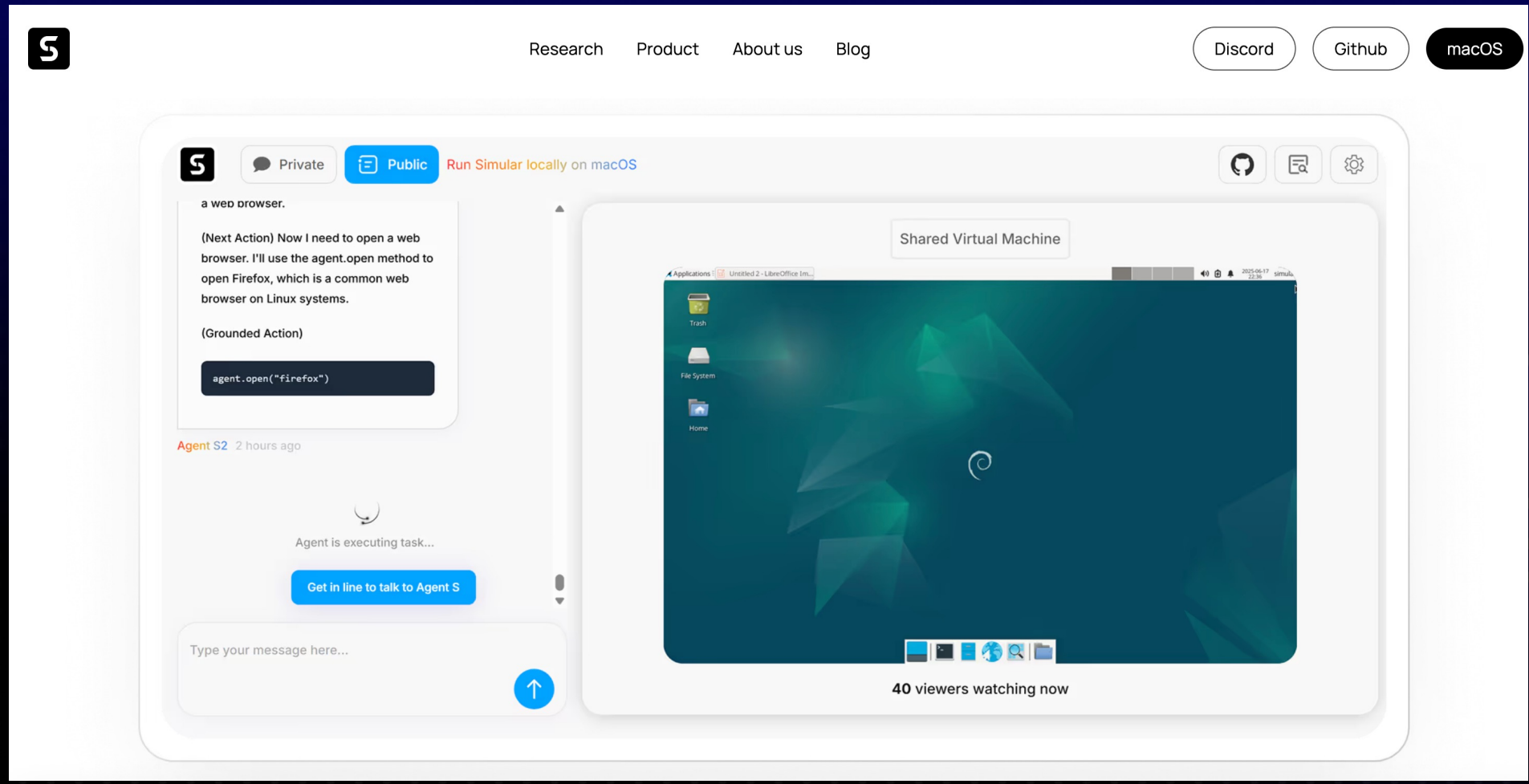


StarCraft II

[5] AVA: Attentive VLM Agent for Mastering StarCraft II

Computer-Using Agents

Playing Games.



Seminal works on Computer-Using Agents

Foundation Models

SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents, ACL 2024

OS-ATLAS: A Foundation Action Model for Generalist GUI Agents , ICLR 2025 Spotlight

Data

OS-Genesis: Automating GUI Agent Trajectory Construction via Reverse Task Synthesis , ACL 2025

Breaking the Data Barrier -- Building GUI Agents Through Task Generalization, COLM 2025

Algorithm

AgentStore: Scalable Integration of Heterogeneous Agents As Specialized Generalist Computer Assistant , ACL 2025

Evaluation

OS-MAP: How Far Can Computer Use Agents Go in Breadth and Depth?

Frontier Application

ScienceBoard: Evaluating Multimodal Autonomous Agents in Realistic Scientific Workflows

Seminal works on Computer-Using Agents

SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents, ACL 2024

OS-ATLAS: A Foundation Action Model for Generalist GUI Agents , ICLR 2025 Spotlight

OS-Genesis: Automating GUI Agent Trajectory Construction via Reverse Task Synthesis , ACL 2025

Breaking the Data Barrier -- Building GUI Agents Through Task Generalization, COLM 2025

AgentStore: Scalable Integration of Heterogeneous Agents As Specialized Generalist Computer Assistant , ACL 2025

OS-MAP: How Far Can Computer Use Agents Go in Breadth and Depth?

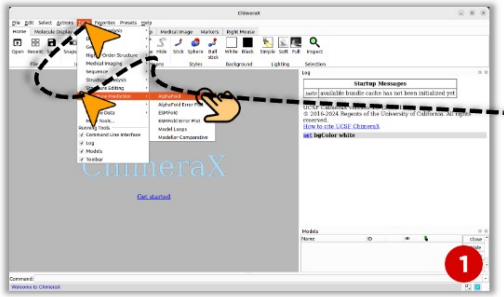
ScienceBoard: Evaluating Multimodal Autonomous Agents in Realistic Scientific Workflows

Frontier Application

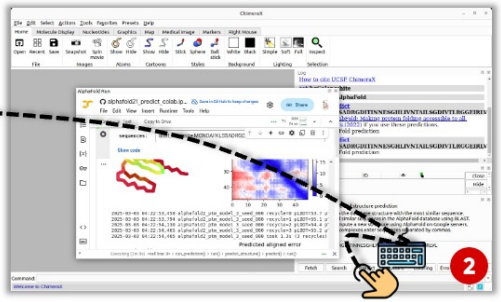
Computer-Using Agents

Automate scientific workflows, be your co-scientist!

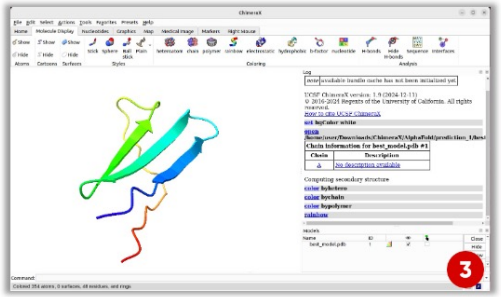
Instruction: Predict the protein structure for the amino acid sequence of 'MGND…' via AlphaFold in ChimeraX.



Step1: Toggle the widget of AlphaFold.

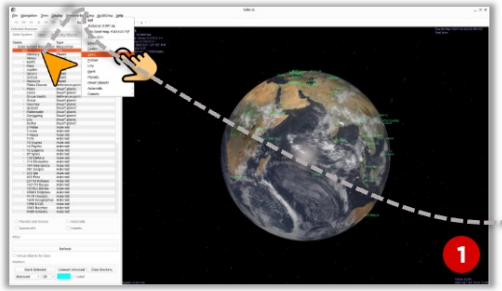


Step2: Input the given sequence and call out AlphaFold for structure prediction.

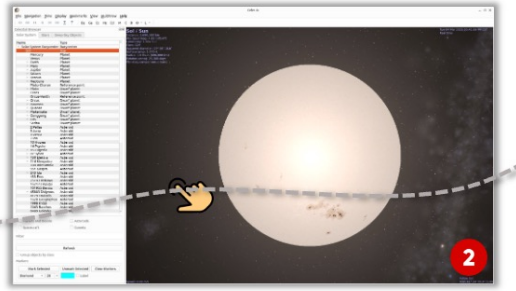


Step3: Wait until the prediction finished.

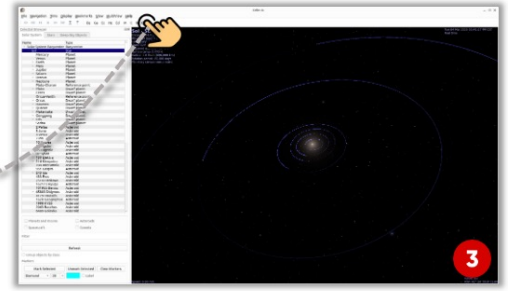
Instruction: Show planets' orbits of Solar System in Celestia.



Step1: Select the Sol and click 'Goto' in contect menu.



Step2: Slide the mouse wheel to move the camera away from Sol.



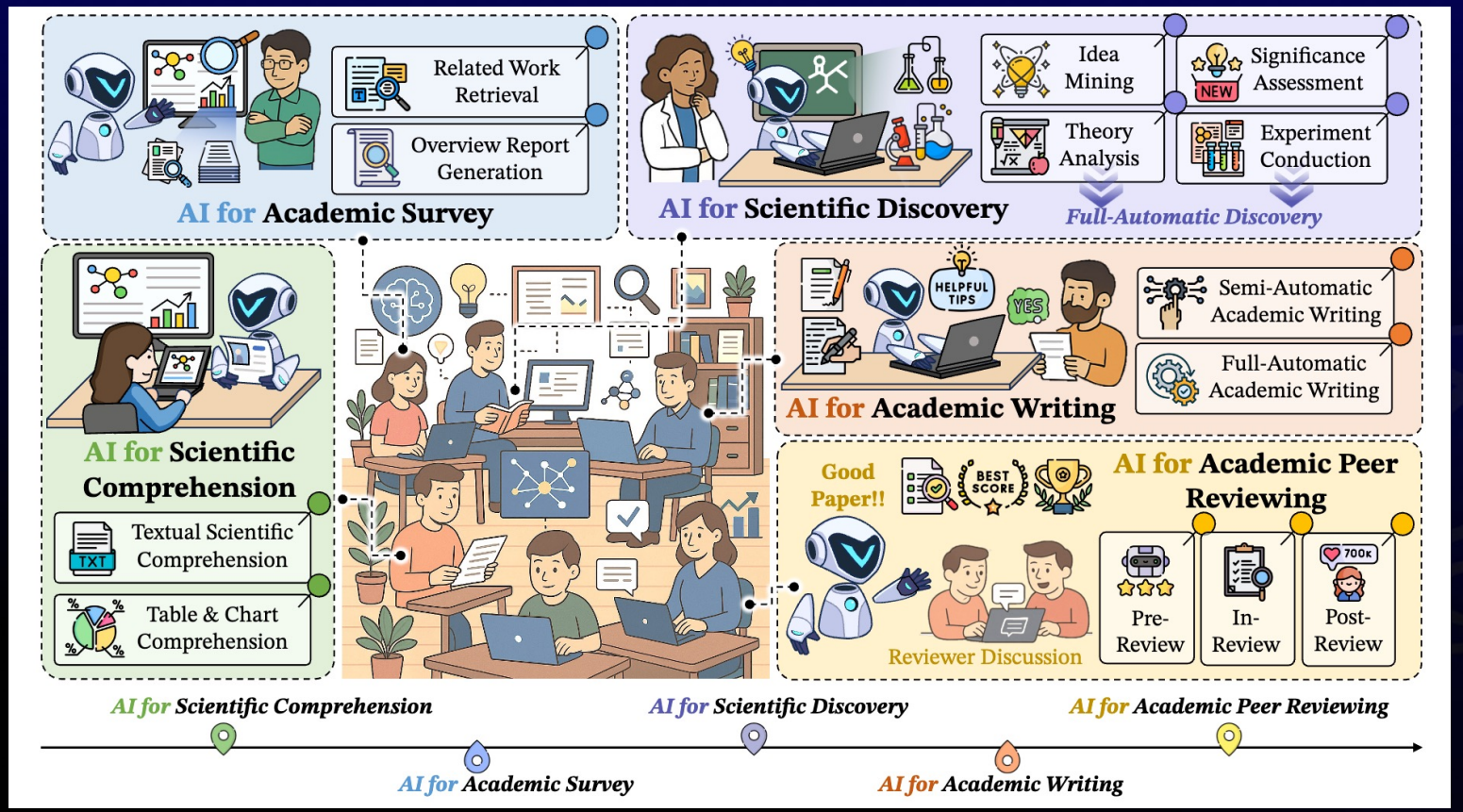
Step3: Click to show orbits of planets.

[6] ScienceBoard: Evaluating Multimodal Autonomous Agents in Realistic Scientific Workflows

Part2 | AI4Research

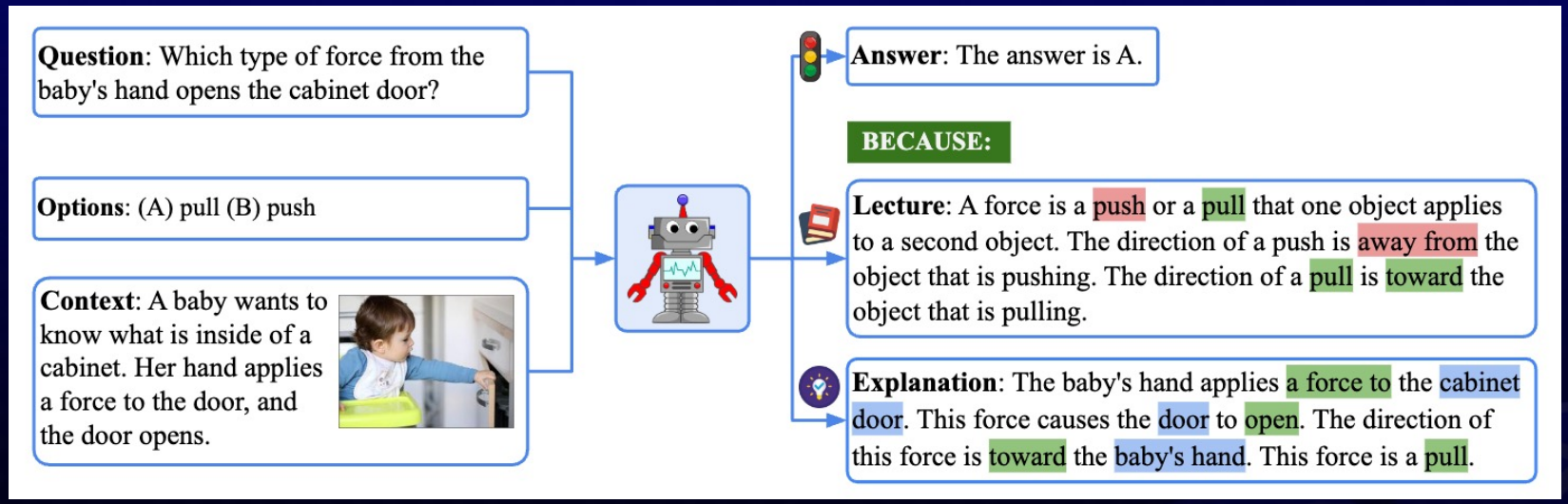
Backgrounds

AI4Research is a highly popular concept.



[7] AI4Research: A Survey of Artificial Intelligence for Scientific Research

Backgrounds: Pastoral Age



ScienceQA (NIPS 2022)

- Multimodal Reasoning (Chain-of-Thought)
- Natural science, language science, and social science
- 12k Grade school-level MCQ

[8] Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering, Lu, et al, NIPS 2022

Backgrounds: Pastoral Age

BioASQ-QA (Nature 2023)

- Designed for biomedical question answering
  - English questions, exact answers, and ideal summaries.
  - Supports information retrieval, passage retrieval, and natural language generation.
- Meets real information needs of biomedical experts.
- Annually expanded with new questions and answers.
- Available on Zenodo in JSON format.

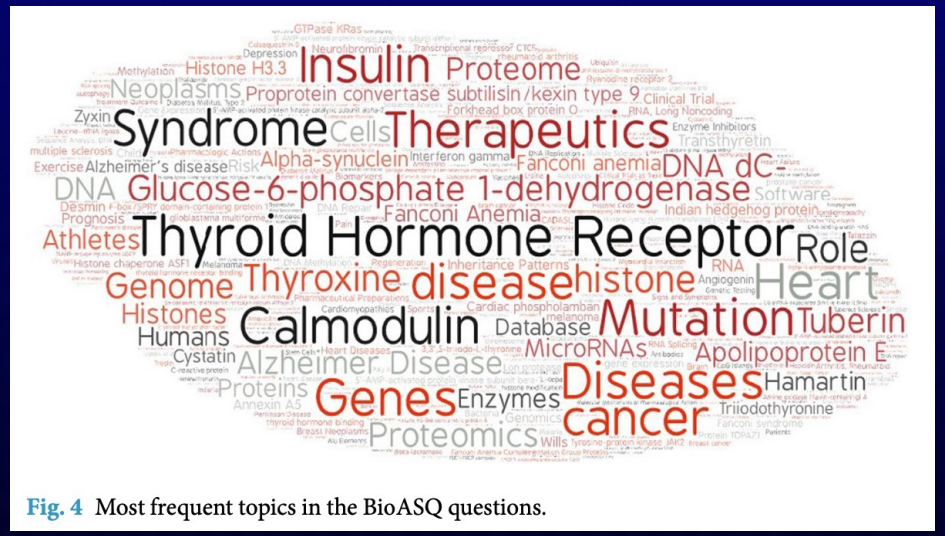


Fig. 4 Most frequent topics in the BioASQ questions.

MoleculeQA (ArXiv 2024)

- Evaluate Factual Accuracy in Molecular Comprehension
- 62K QA Pairs across 23K molecules
- MCQ problems (training set available)
- Textual-based

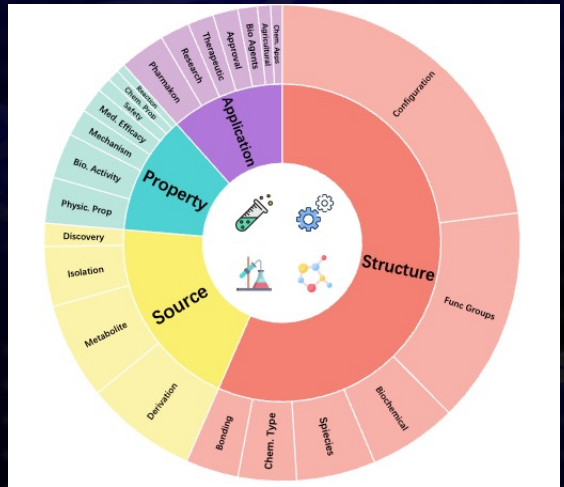


Figure 4: An overview of MoleculeQA topics distribution. Four coarse-grained aspects occupy the inner circle, and in the outer circle we list finer-grained non-leaf topics.

[9] BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Krithara et al, Nature 2023

[10] MoleculeQA: A Dataset to Evaluate Factual Accuracy in Molecular Comprehension, Lu, et al, ArXiv 2024

Backgrounds: Contemporary Era

SciCode (NIPS 2024)

- 16 subfields (e.g., math, physics, chem).
- 80 main problems, decomposed into 338 subproblems involving recall, reasoning, and code synthesis.
- Each question verified by 2 senior researchers to ensure scientific accuracy and relevance.

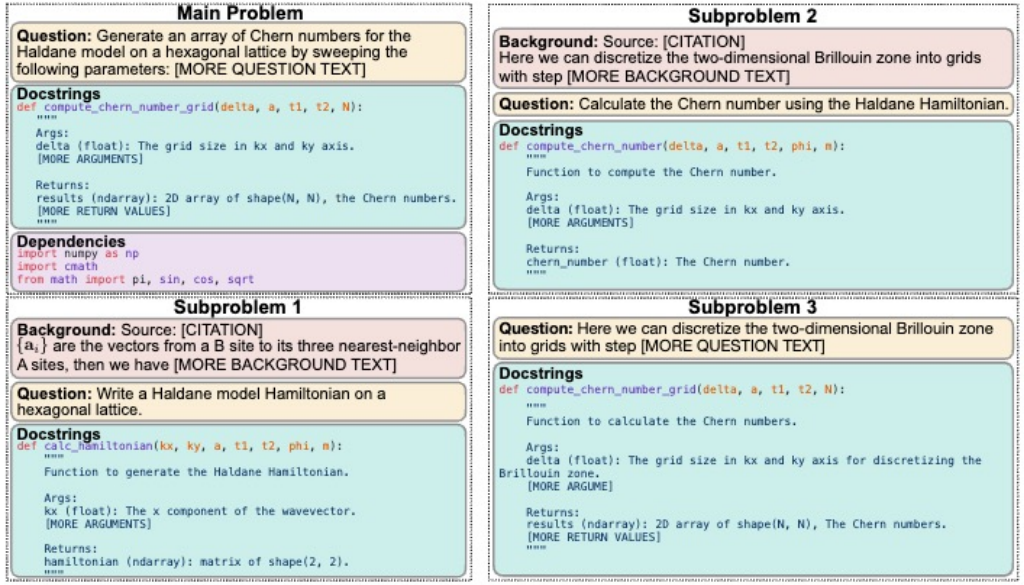
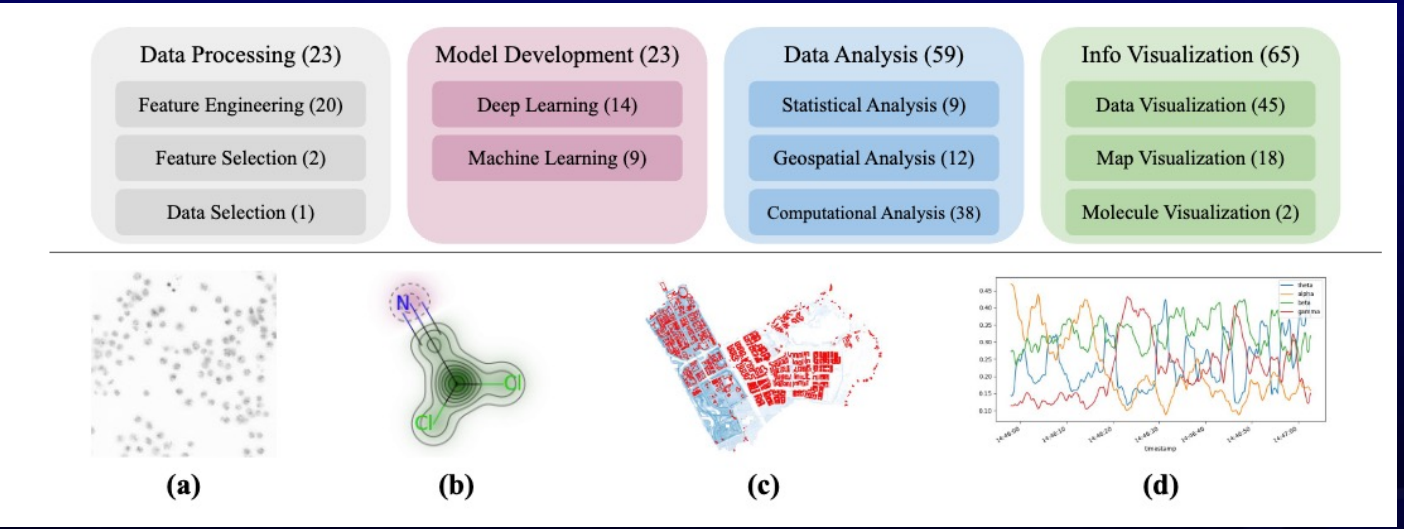


Figure 1: A SciCode main problem is decomposed into multiple smaller and easier subproblems. Docstrings specify the requirements and input-output formats. When necessary, scientific background knowledge is provided, written by our scientist annotators. The full problem is shown in subsection A.3

[11] SciCode: A Research Coding Benchmark Curated by Scientists, NIPS 2024

Backgrounds: Contemporary Era



ScienceAgentBench (ICLR 2025)

- Evaluating language **agents** in data-driven scientific discovery.
- 102 tasks from 44 peer-reviewed publications across four disciplines.
- Tasks require generating a self-contained Python program.
- Prompting based solutions: OpenHands CodeAct, and self-debug.

[12] ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery, ICLR 2025

Backgrounds: Contemporary Era

SCIENCE

AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms

14 MAY 2025

By AlphaEvolve team

Share

[13] AlphaEvolve: A coding agent for scientific and algorithmic discovery

Backgrounds: Contemporary Era

A lot of "AI Research" systems have been built…

2024-9-4

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu[1,2,*], Cong Lu[3,4,*], Robert Tjarko Lange[1,*], Jakob Foerster[2,†], Jeff Clune[3,4,5,†] and David Ha[1,†]
[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising

SciMON 🧪 : Scientific Inspiration Machines Optimized for Novelty

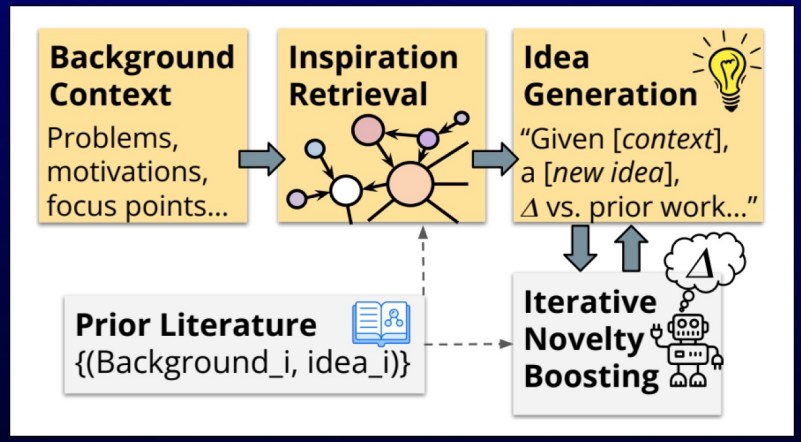
Qingyun Wang[1], Doug Downey[2], Heng Ji[1], Tom Hope[2,3]
[1] University of Illinois at Urbana-Champaign [2] Allen Institute for Artificial Intelligence (AI2)
[3] The Hebrew University of Jerusalem
{tomh,doug}@allenai.org, {qingyun4,hengji}@illinois.edu

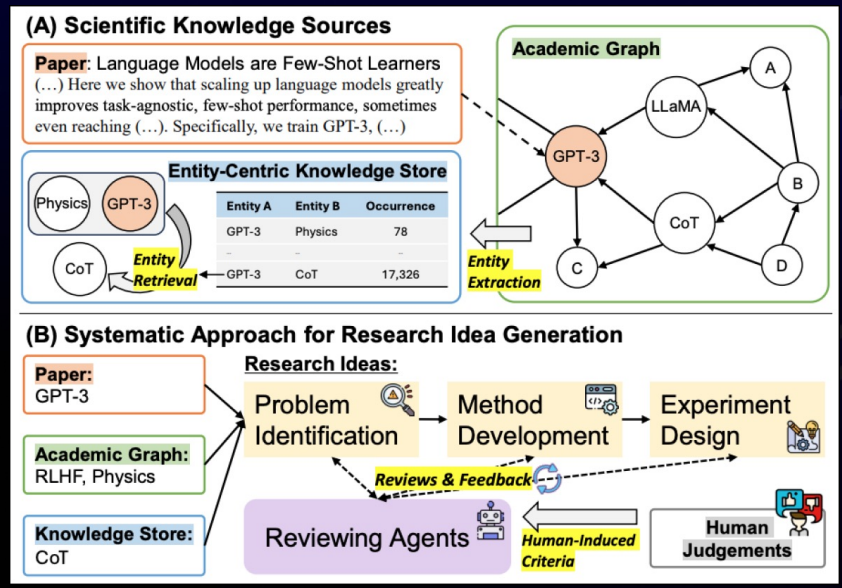
ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models

Jinheon Baek[1] Sujay Kumar Jauhar[2] Silviu Cucerzan[2] Sung Ju Hwang[1,3]
KAIST[1] Microsoft Research[2] DeepAuto.ai[3]
{jinheon.baek, sjhwang82}@kaist.ac.kr {sjauhar, silviu}@microsoft.com

Automated Peer Reviewing in Paper SEA: Standardization, Evaluation, and Analysis

Jianxiang Yu[♡*], Zichen Ding[♡*], Jiaqi Tan[♡], Kangyang Luo[♡], Zhenmin Weng[♡], Chenghua Gong[♡], Long Zeng[♡], Renjing Cui[♡], Chengcheng Han[♡], Qiushi Sun[♢], Zhiyong Wu[♢], Yunshi Lan[♡], Xiang Li[♡†]
[♡] East China Normal University, Shanghai, China
[♢] Shanghai AI Laboratory, Shanghai, China
sea.ecnu@gmail.com
https://ecnu-sea.github.io/





Thinking

Currently, AI acted as an "Analyzer," helping with idea thinking data analysis, writing, and visualization.

Can AI evolve into an "Executor" that helps (1) formulate a plan, (2) directly operates scientific software via GUI or CLI, and (3) even generates some reports?

The answer is YES—with the emergence of computer-using agents.

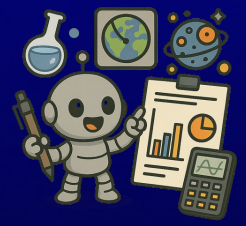
Let's move beyond QA and Coding to actively performing some research tasks!

From Digital Agents to AI Co-Scientists

Part3 | ScienceBoard

AI Insight Talk

OpenMMLab | Hugging Face | ModelScope 魔搭社区 | 知乎 | 机智流



ScienceBoard: Evaluating Multimodal Autonomous Agents in Realistic Scientific Workflows

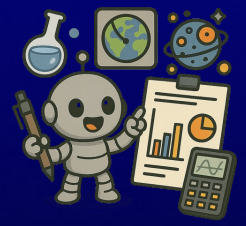
Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, Jianing Wang, Qintong Li, Xiangru Tang, Tianbao Xie, Xiachong Feng, Xiang Li, Ben Kao, Wenhai Wang, Biqing Qi, Lingpeng Kong, Zhiyong Wu

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Preprint / WUCA @ ICML 2025 Oral

AI Insight Talk

OpenMMLab | Hugging Face | ModelScope 魔搭社区 | 知乎 | 机智流



ScienceBoard: Evaluating Multimodal Autonomous Agents in Realistic Scientific Workflows

Qiushi Sun, Zhoumianze Liu, Chang Ma, Zichen Ding, Fangzhi Xu, Zhangyue Yin, Haiteng Zhao, Zhenyu Wu, Kanzhi Cheng, Zhaoyang Liu, Jianing Wang, Qintong Li, Xiangru Tang, Tianbao Xie, Xiachong Feng, Xiang Li, Ben Kao, Wenhai Wang, Biqing Qi, Lingpeng Kong, Zhiyong Wu

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Preprint / WUCA @ ICML 2025 Oral

Language Agents -> Computer-using Agents

Let's start with some background on computer-using agents.

Remark: For computer-using agents, both GUI and CLI represent distinct approaches.

In ScienceBoard, we primarily focus on GUI-based interaction, complemented by CLI support.

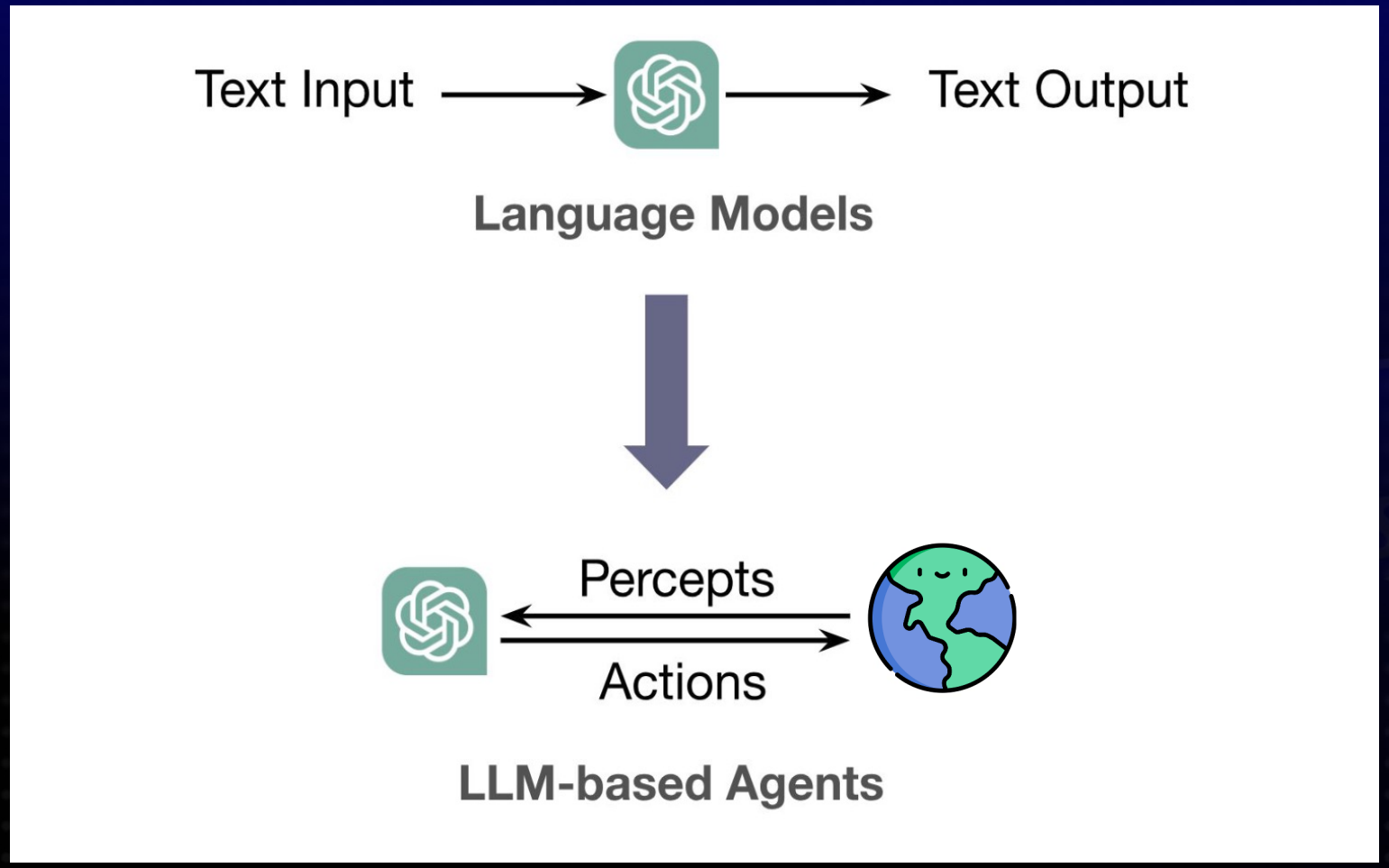
GUI Agents

Intelligent agents that operate within GUI environments, leveraging LLMs as their core inference and cognitive engine to generate, plan, and execute actions in a flexible and adaptive manner.

[15] API Agents vs. GUI Agents: Divergence and Convergence, arXiv 2025

Language Agents

Computer-using agents are language agents.



Language Agents -> Computer-using Agents

Agents are promising, but building powerful computer-using agents is challenging:

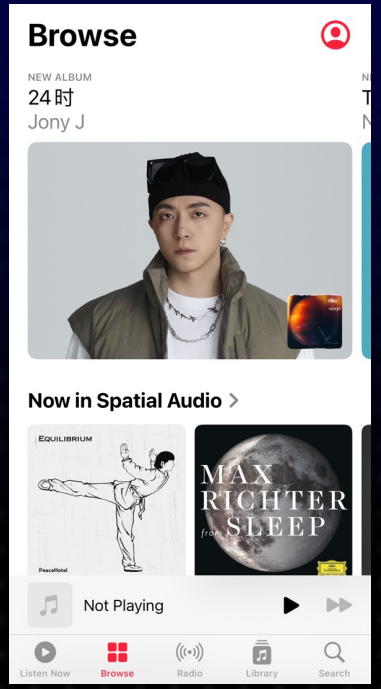
1. Agents need to follow human instructions.
2. Agents need to perform planning and action.
3. Agents need to perceive envs. and the applications they are interacting with.

Language Agents -> Computer-using Agents

What are “actions”

Language Agents -> Computer-using Agents

Typical action: GUI grounding – the capacity to accurately locate screen elements based on instructions, e.g., CLICK.



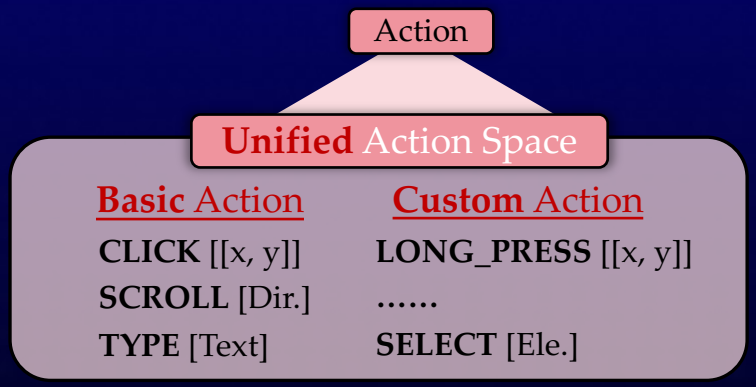
In order to view the new album of Jony J, where should I click?

GPT-4o (an earlier version): hmmm… Sorry I don't know. ❌

SeeClick: (0.49, 0.40) ✅

[16] SeeClick: Harnessing GUI Grounding for Advanced Visual GUI Agents, ACL 2024

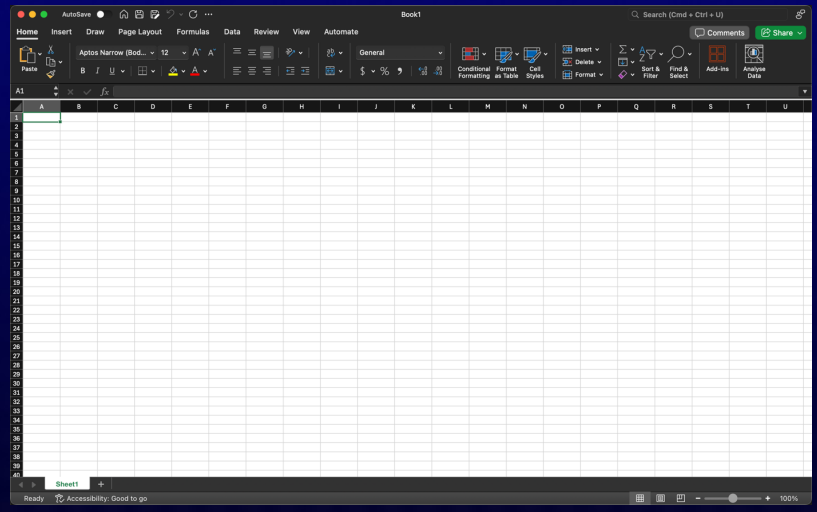
Computer-using Agents



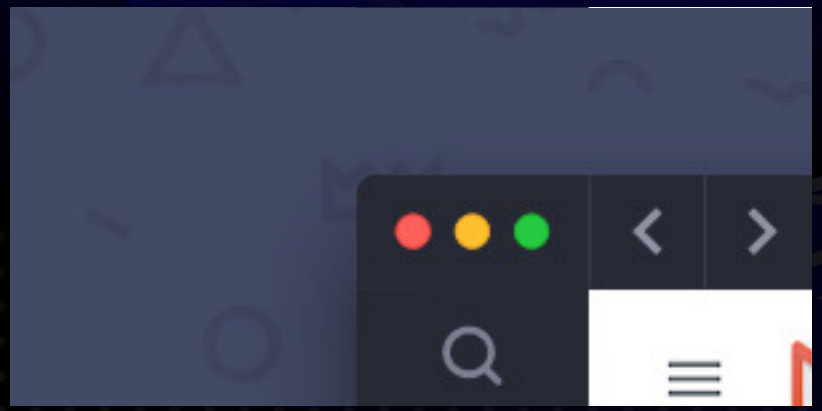
V.S. CLI

1. Screenshot is information complete for agentic tasks

2. GUI action space is much smaller and is shared across platforms/apps



"Close the app"



ScienceBoard Infra

So to achieve our goal, we need an environment that allows agents to actively interact.

1. Supports native multimodal interaction
2. Fully compatible with coding and conversational research assistance
3. Enables rigorous validation

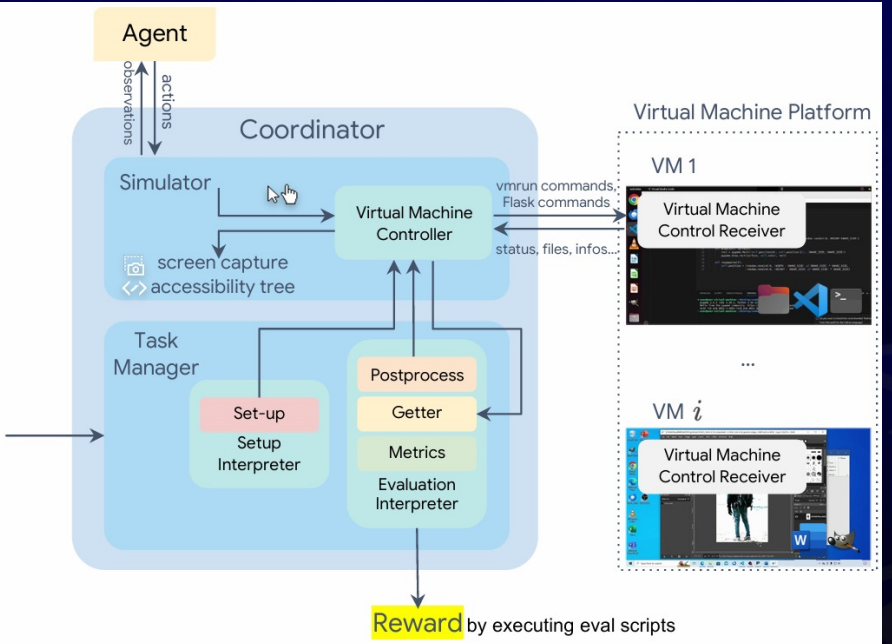
A playground—a virtual machine pre-installed with well-adapted scientific software.

ScienceBoard Infra

We build upon the OSWorld infrastructure for GUI interaction.

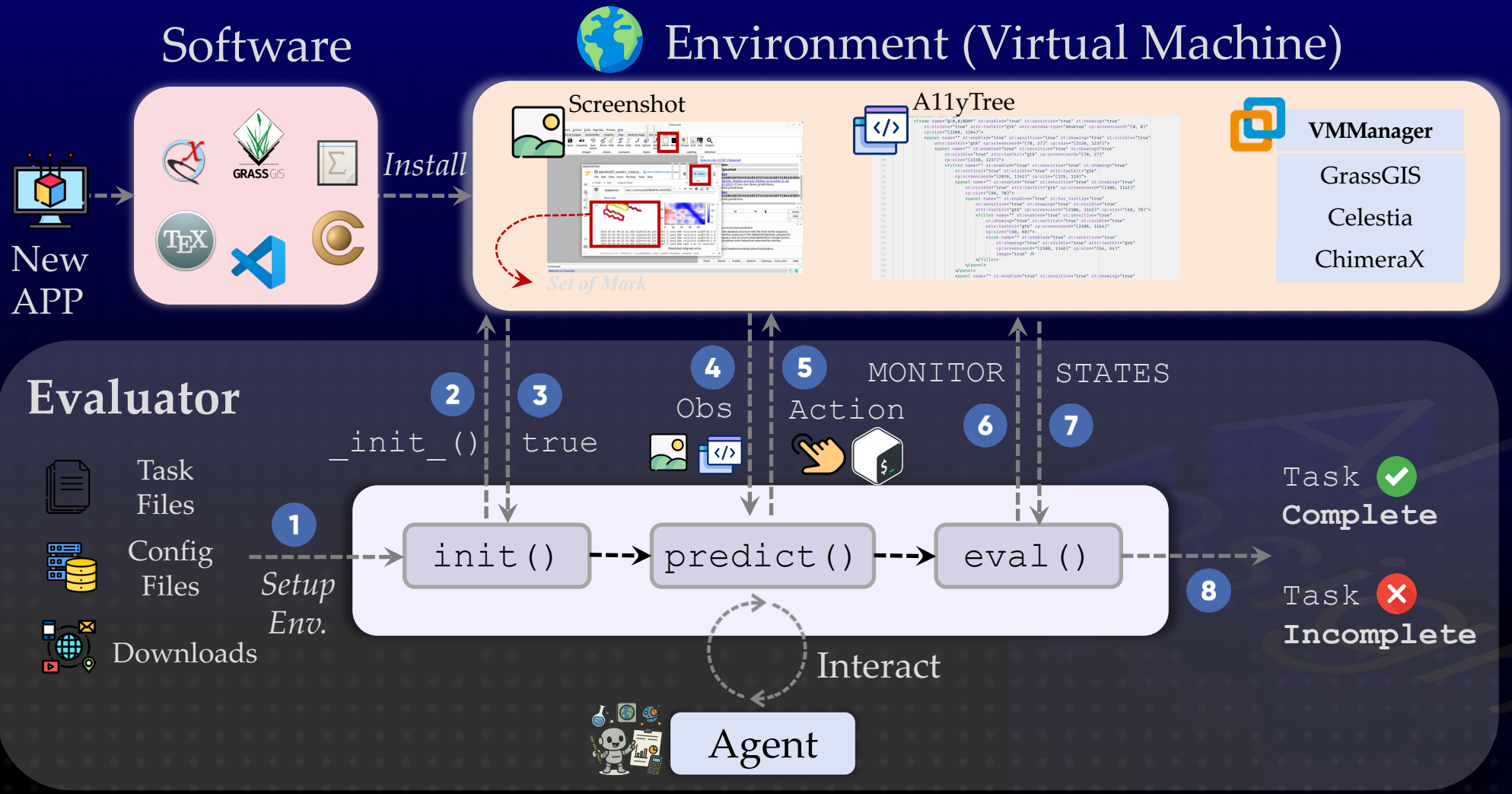
Config

```
{ "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",... }]}]
}
```



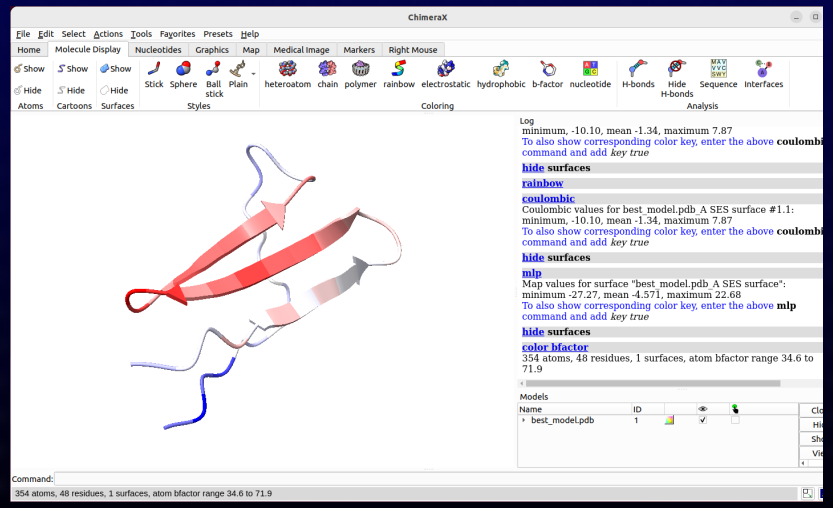
For CLI, we enable interaction by modifying the software itself and leveraging VSCode.

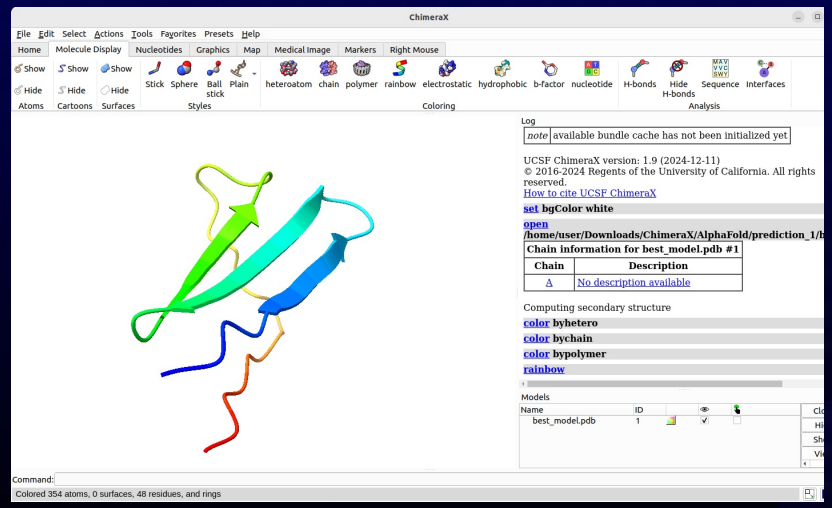
[4] OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

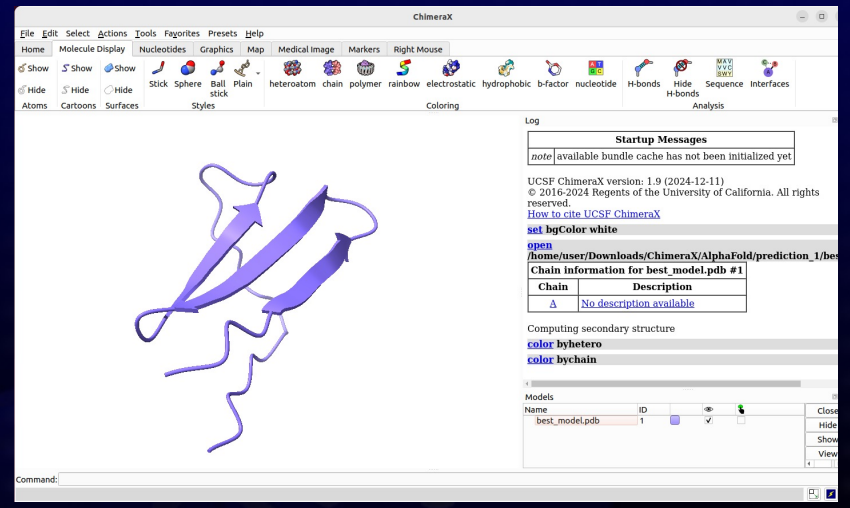

Software
Environment (Virtual Machine)
New APP
Install
Screenshot
Set of Mark
A11yTree
VMManager
GrassGIS
Celestia
ChimeraX
Evaluator
Task Files
Config Files
Downloads
1
Setup Env.
2
init()
3
true
4
Obs
5
Action
MONITOR
6
7
STATES
init()
predict()
eval()
8
Task Complete
Task Incomplete
Interact
Agent

The Dilemma of Evaluation

Evaluation is harder than we expect, let's take visualization as an example





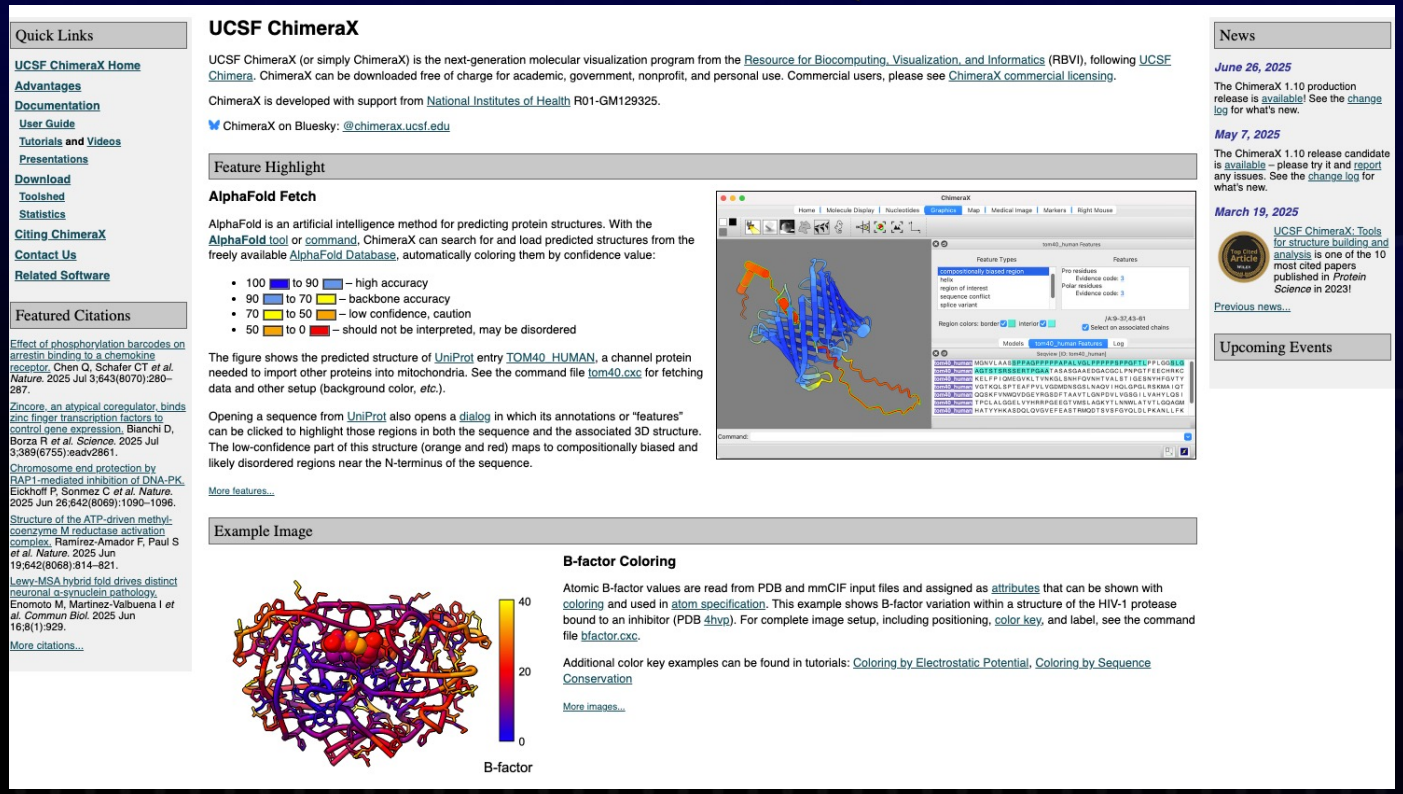


It is impossible to "match"

How to Evaluate?

We rely on internal states.

By modifying the software, we access intermediate runtime states and enable precise state-based evaluation, e.g., UCSF ChimeraX



Quick Links

UCSF ChimeraX Home
Advantages
Documentation
User Guide
Tutorials and Videos
Presentations
Download
Toolshed
Statistics
Citing ChimeraX
Contact Us
Related Software

Featured Citations

Effect of phosphorylation barcodes on arrestin binding to a chemokine receptor. Chen Q, Schafer CT *et al. Nature.* 2025 Jul 3;643(8070):280–287.

Zincore, an atypical coregulator, binds zinc finger transcription factors to control gene expression. Bianchi D, Borza R *et al. Science.* 2025 Jul 3:389(6755):eadv2861.

Chromosome end protection by RAP1-mediated inhibition of DNA-PK. Eickhoff P, Sonmez C *et al. Nature.* 2025 Jun 26;642(8069):1090–1096.

Structure of the ATP-driven methyl-coenzyme M reductase activation complex. Ramírez-Amador F, Paul S *et al. Nature.* 2025 Jun 19;642(8068):814–821.

Lewy-MSA hybrid fold drives distinct neuronal α-synuclein pathology. Enomoto M, Martinez-Valbuena I *et al. Commun Biol.* 2025 Jun 16;8(1):929.

More citations...

UCSF ChimeraX

UCSF ChimeraX (or simply ChimeraX) is the next-generation molecular visualization program from the Resource for Biocomputing, Visualization, and Informatics (RBVI), following UCSF Chimera. ChimeraX can be downloaded free of charge for academic, government, nonprofit, and personal use. Commercial users, please see ChimeraX commercial licensing.

ChimeraX is developed with support from National Institutes of Health R01-GM129325.

ChimeraX on Bluesky: @chimerax.ucsf.edu

Feature Highlight

AlphaFold Fetch

AlphaFold is an artificial intelligence method for predicting protein structures. With the AlphaFold tool or command, ChimeraX can search for and load predicted structures from the freely available AlphaFold Database, automatically coloring them by confidence value:

- 100 to 90 – high accuracy
- 90 to 70 – backbone accuracy
- 70 to 50 – low confidence, caution
- 50 to 0 – should not be interpreted, may be disordered

The figure shows the predicted structure of UniProt entry TOM40_HUMAN, a channel protein needed to import other proteins into mitochondria. See the command file tom40.cxc for fetching data and other setup (background color, *etc.*).

Opening a sequence from UniProt also opens a dialog in which its annotations or "features" can be clicked to highlight those regions in both the sequence and the associated 3D structure. The low-confidence part of this structure (orange and red) maps to compositionally biased and likely disordered regions near the N-terminus of the sequence.

More features...

Example Image

B-factor Coloring

Atomic B-factor values are read from PDB and mmCIF input files and assigned as attributes that can be shown with coloring and used in atom specification. This example shows B-factor variation within a structure of the HIV-1 protease bound to an inhibitor (PDB 4hvp). For complete image setup, including positioning, color key, and label, see the command file bfactor.cxc.

Additional color key examples can be found in tutorials: Coloring by Electrostatic Potential, Coloring by Sequence Conservation

More images...

40
20
0
B-factor

News

June 26, 2025

The ChimeraX 1.10 production release is available! See the change log for what's new.

May 7, 2025

The ChimeraX 1.10 release candidate is available – please try it and report any issues. See the change log for what's new.

March 19, 2025

UCSF ChimeraX: Tools for structure building and analysis is one of the 10 most cited papers published in *Protein Science* in 2023!

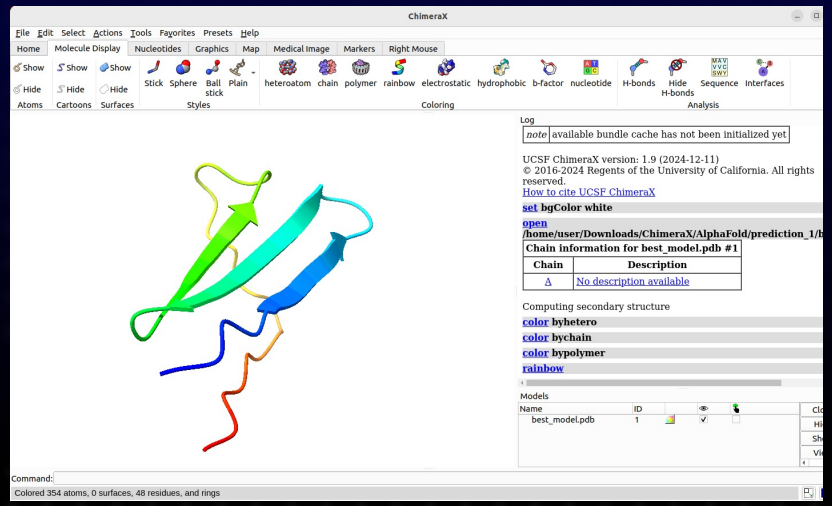
Previous news...

Upcoming Events

How to Evaluate?

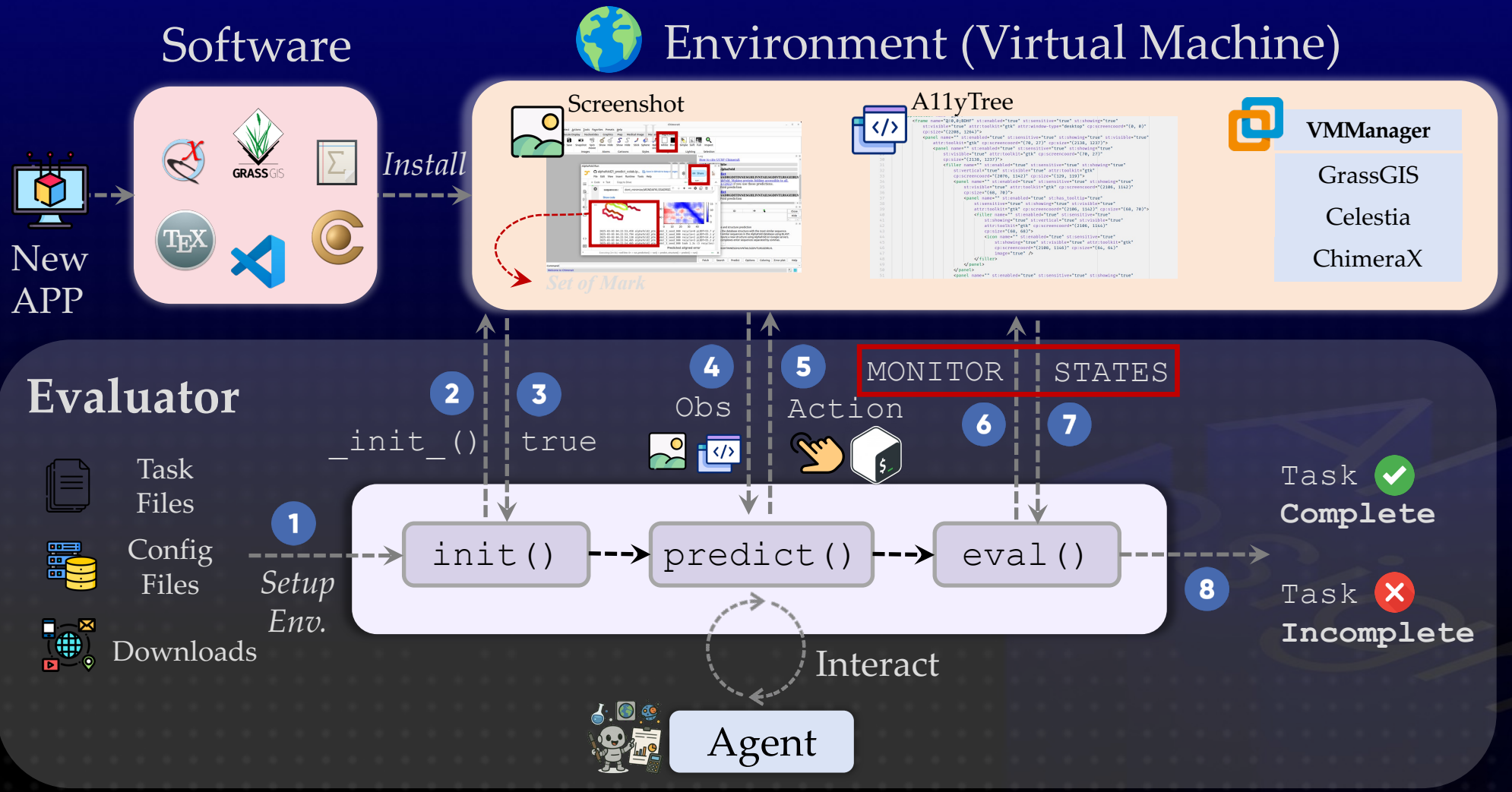
We rely on internal states.

By modifying the software, we access intermediate runtime states and enable precise state-based evaluation, e.g., UCSF ChimeraX

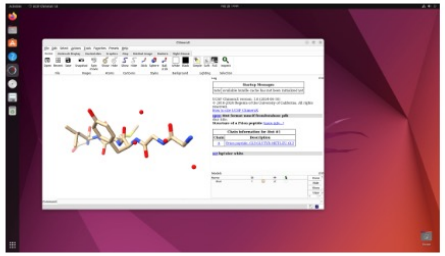
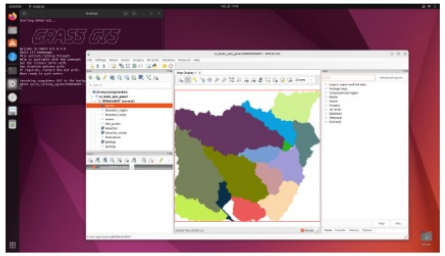
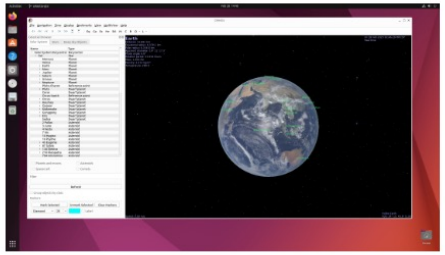


```
_state_containers.tools._tool_instances.#
0.tool_window._ToolWindow__toolkit.main_w
indow.tool_instance_to_windows.<chimerax.
log.tool.Log object at
0x0000020AA8023B50>.#0.tool_instance.page
_source
```

ScienceBoard Infra



Evaluation

Initial State	Instruction	Evaluation Script (Simplified)
	*Select all water molecules and draw their centroids with radius of 1Å in ChimeraX.*	`{ "type":"info","key":"sell", "value":["atom id #!1/A:201@O idatm_type O3" "...",] },{ "type":"states", "find":"lambda k,v:k.endswith('._name')", "key":"lambda k:'..._atoms_drawing'", "value":"[[13.0012 1.7766 21.3672 1.]]" }`
	*Display and ONLY display the layer of 'boundary_region' in Grass GIS.*	`{ "type":"info", "key":"lambda dump:len(dump['layers'])", "value":1 },{"type":"info" "key":"lambda dump:dump['layers'][0]['name']", "value":"boundary_region@PERMANENT" }`
	*Set the Julian date to 2400000 in Celestia.*	`{ "type":"info", "key":"simTime", "value":2400000, "pred":"lambda left, right:abs(left-right) < 1", }`

Evaluation

"Approach to the Earth and display a solar eclipse in Celestia."

```
"evaluate": [
    {
        "type": "info",
        "key": "lambda dump: dump['entity']['Earth']['distance']",
        "value": 0,
        "pred": "lambda key, value: abs(key - value) < 450000"
    },
    {
        "type": "info",
        "key": "lambda dump: dump['entity']['Sol']['visible']",
        "value": false
    },
    {
        "type": "info",
        "key": "lambda dump: dump['entity']['Moon']['visible']",
        "value": true
    },
    {
        "type": "info",
        "key": "lambda dump: (s := dump['entity']['Sol']['position'], e := dump['entity']['Earth']['position'], m
        := dump['entity']['Moon']['position'], mv := [m[i] - e[i] for i in range(3)], sv := [s[i] - e[i] for i in
        range(3)], dp := mv[0] * sv[0] + mv[1] * sv[1] + mv[2] * sv[2], _mv := __import__('math').sqrt(mv[0]**2 + mv
        [1]**2 + mv[2]**2), _sv := __import__('math').sqrt(sv[0]**2 + sv[1]**2 + sv[2]**2), dp / (_mv * _sv))[-1]",
        "value": 0.99,
        "pred": "lambda key, value: key > value"
    }
]
```

Next

Now agents can freely explore the environment and execute any actions they choose.

But how well do these agents actually perform to automate science tasks?

What kind of benchmark is needed to truly evaluate their capabilities?

ScienceBoard Benchmark

We aim to build a benchmark with:

1. Real-world tasks that human actually perform
2. Coverage across multiple disciplines
3. Graded difficulty levels
4. Support for cross-application workflows
5. Cross “modality” GUI + CLI

And more

ScienceBoard Benchmark

Coverage:

Biochem, GIS, Astronomy, Algebra, ATP, Documentation

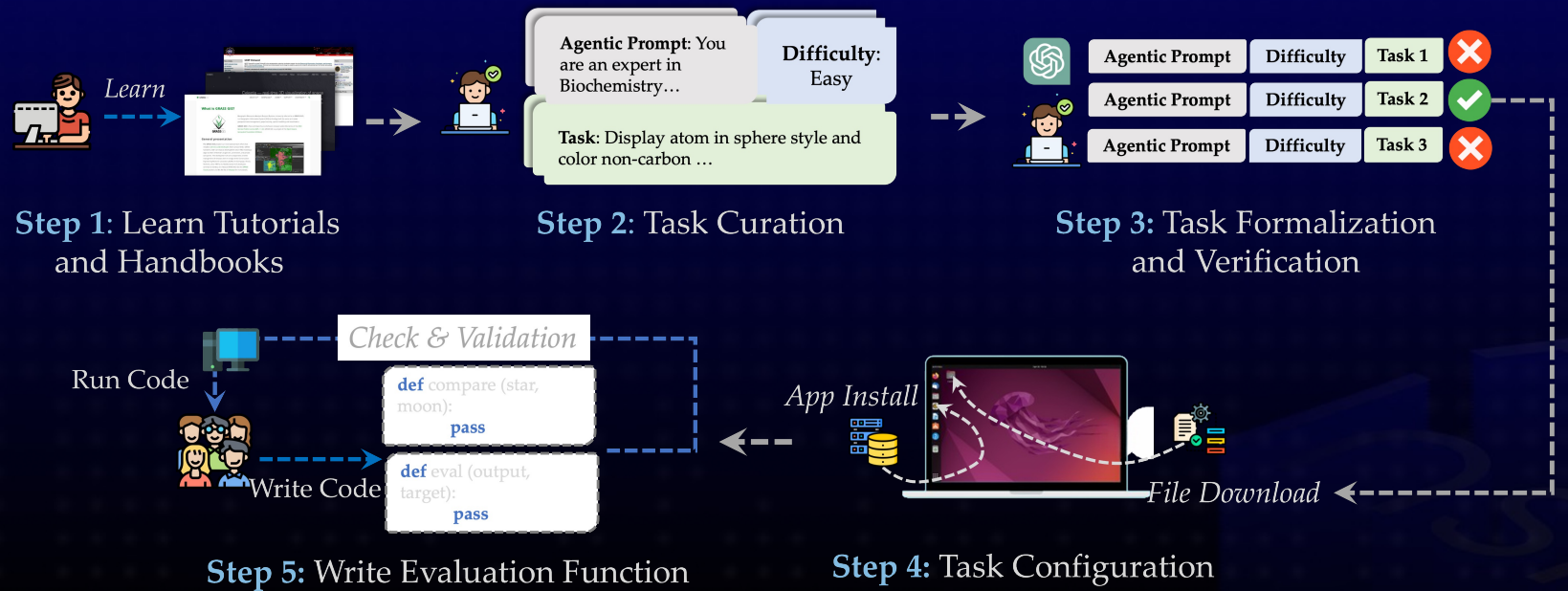
ScienceBoard Benchmark

Our criteria for selecting software:

1. Stable operation on Ubuntu
2. Open-source, allowing for adaptation
3. Accessible a11ytree, enabling text-based agents to participate (e.g., o1-mini)
4. License

ScienceBoard Benchmark

How the benchmark is built



Annotation pipeline

AI Insight Talk

ScienceBoard Benchmark

OpenMMLab | Hugging Face | ModelScope 魔搭社区 | 知乎 | 机智流

Celestia Users Guide

by Frank Gregorio

- MS Word document (zipped) or Read Online
- Документ MS Word или PDF файл или Читать онлайн
- Document PDF (pour Celestia 1.6.0)
- PDF Datei (für Celestia 1.6.0-1)
- Documento MS Word e OpenOffice (per Celestia 1.6.1, archivio zip)
- PDF 文档 (Celestia 1.6.1)

Celestia Key Chart

by The Learning Technologies Project Office of NASA

- PNG Image
- PNG изображение
- PNG 图像

CEL Scripting Guide

by Don Goyette

- MS Word document or Read Online
- Документ PDF (архив ZIP)

SSC File Scripting Guide

- PDF document (zipped)

Celestia tutorials

ChimeraX Quick Start Guide

UCSF ChimeraX is the next-generation visualization program from the Resource for Biocomputing, Visualization, and Informatics at UC San Francisco, following Chimera. See also: ChimeraX tutorials

Many ChimeraX actions require typing commands. The help for a specific command can be shown with the **help** command (for example, **help style**). Other ways to interact with the program include:

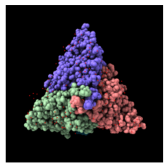
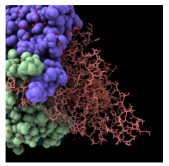
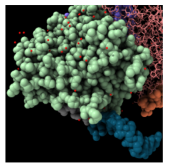
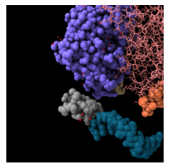
- clicking toolbar icons, optionally after making a selection
- graphical tools
- context menus shown by right-click (Ctrl-click on Mac, Alt-click on Windows trackpad)

Command-Execution Links

Clicking command links in the examples below will execute them in ChimeraX *if this page is shown in the ChimeraX internal browser*, such as with **Help... Quick Start Guide** in the ChimeraX menu.

Example Atomic-Structure Commands

Example structure: Protein DataBank 2BBV, black beetle virus capsid

open 2bbv color bychain	style /b stick	Mouse drag to move. color /n teal	hide /c
			
File is fetched from the PDB in mmCIF format and cached locally. Lighting with shadows.	*Change chain b to stick style.*	*Rotate by dragging, translate by dragging with middle mouse button or with **option** key pressed (Mac) (more...).*	*Hide chain c atoms.*

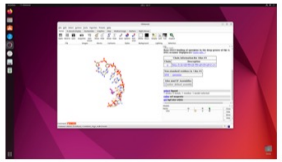
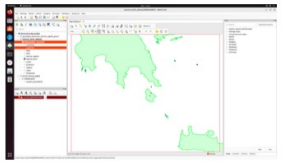
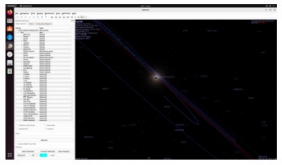
ribbon /c	Mouse click with **ctrl** key pressed to select an atom, or command select /N:4@C5'	Press up-arrow key, or command select up	color sel gold select clear

ChimeraX tutorials

ScienceBoard Benchmark

Develop a large set of evaluation scripts

Manually validate their correctness.

Initial State	Instruction	Evaluation Script (Simplified)
	*Select all ligand(s) and color them into magenta in ChimeraX.*	`{ "type": "info", "key": "sel", "value": ["atom id /A:9@N1 idatm_type N3+", ... ] },{ "type": "info", "key": "rescolor /A", "value": ["#1/A:1 color #d2b48c", ... ] }`
	*There is a point located in the Mediterranean Sea. Please find and delete it.*	`{ "type": "db", "cmd": "v.to.db", "kwargs": { "flags": "p", "map": "countries@PERMANENT", "type": "point", "option": "coor" }, "key": "lambda out: out.strip()", "value": "cat|x|y|z\n...|8.348947891274|0", "pred": "lambda key, value: key == value" }`
	*Approach to the Earth and display a solar eclipse in Celestia.*	`{ "type": "info", "key": "lambda ...['Earth']['distance']", "value": 0, "pred": "lambda k, v: abs(k - v) < 450000" },{ "type": "info", "key": "lambda ...['Sol']['visible']", "value": false },{ "type": "info", "key": "lambda ...['Moon']['visible']", "value": true },{ "type": "info", "key": "lambda ...", "value": 0.99, "pred": "lambda key, value: key > value" }`
	`theorem TP_3 [TopologicalSpace X] [TopologicalSpace Y] (f : X -> Y) (Z : Set X) (h₁ : Continuous f) (h₂ : IsConnected Z) : IsConnected {y : Y | ∃ z ∈ Z, f z = y} := by sorry`	`{ "type": "placeholder" }`

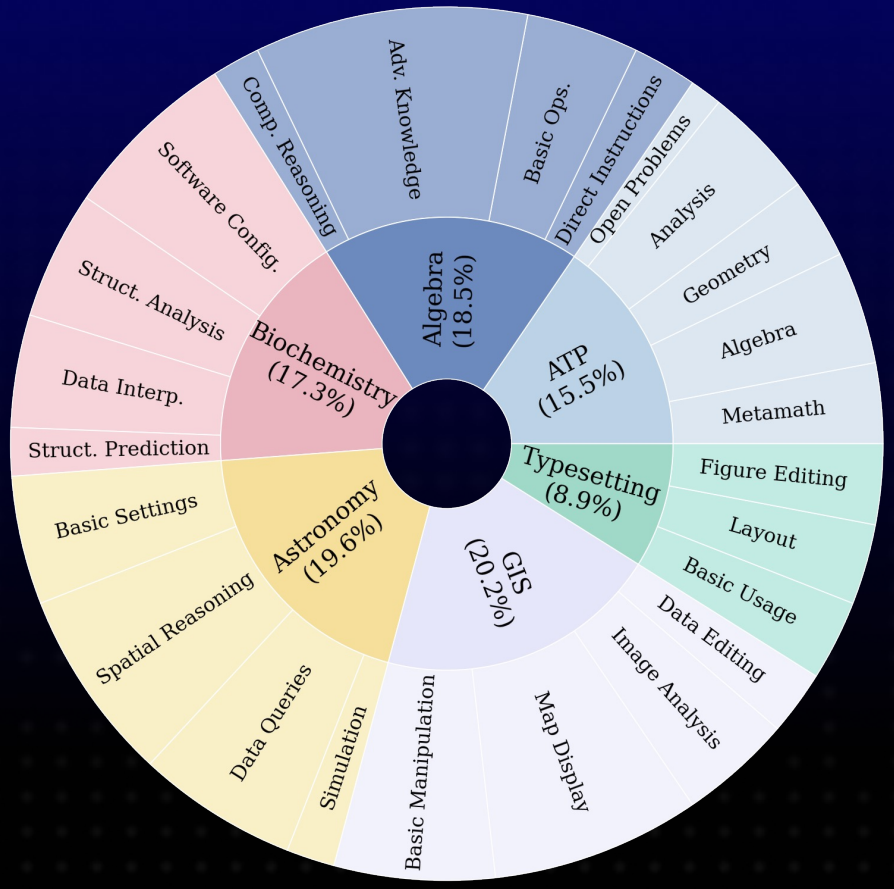
ScienceBoard Benchmark

Task Type	Statistics
Total Tasks	**169 (100%)**
- GUI	38 (22.5%)
- CLI	33 (19.5%)
- GUI + CLI	98 (58.0%)
Difficulty	
- Easy	91 (53.8%)
- Medium	48 (28.4%)
- Hard	28 (16.6%)
- Open Problems	2 (1.2%)
Instructions	
Avg. Length of Task Instructions	20.0
Avg. Length of Agentic Prompt	374.9
Execution	
Avg. Steps	9.0
Avg. Time Consumption	124(s)

Evaluate autonomous computer-using agents in realistic scientific workflows.

Tasks require complex tool usage, scientific reasoning, and multi-step GUI/CLI operations

ScienceBoard Benchmark



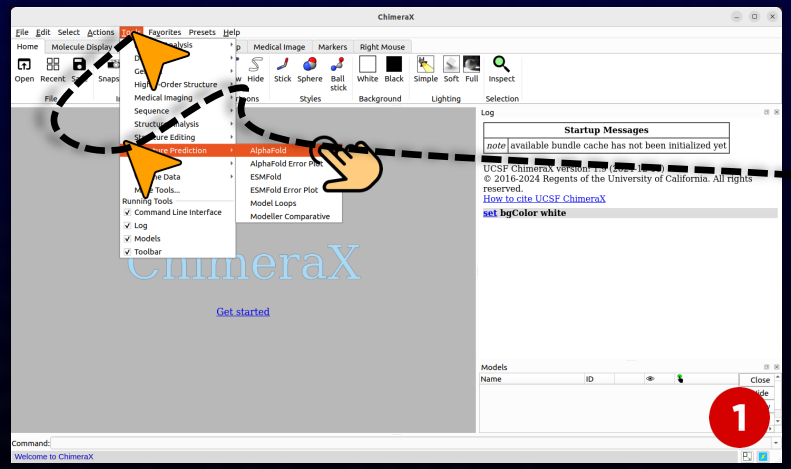
1. 169 high-quality tasks across 6 domains: Biochemistry, Algebra, Theorem Proving, GIS, Astronomy, Documentation.

2. Tasks require GUI operation, visual/textual reasoning, tool use, coding, spatial understanding.

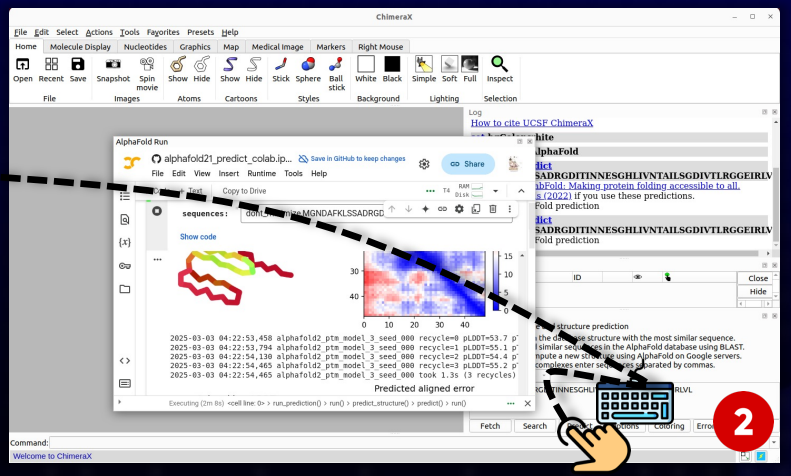
3. CLI-only, GUI-only, hybrid workflows.

Use Cases

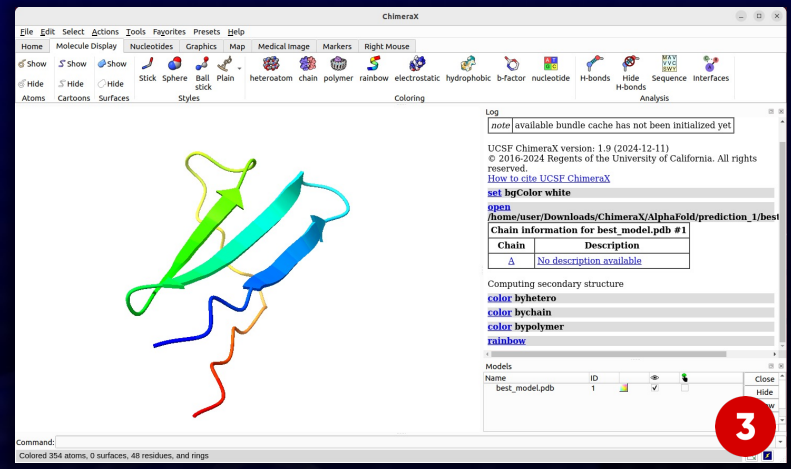
Instruction: Predict the protein structure for the amino acid sequence of 'MGND…' via AlphaFold in ChimeraX.



Step1: Toggle the widget of AlphaFold



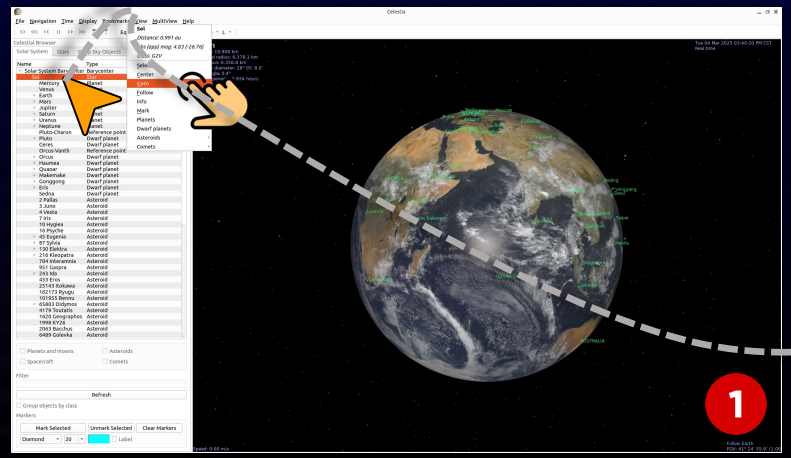
Step2: Input the given sequence and call

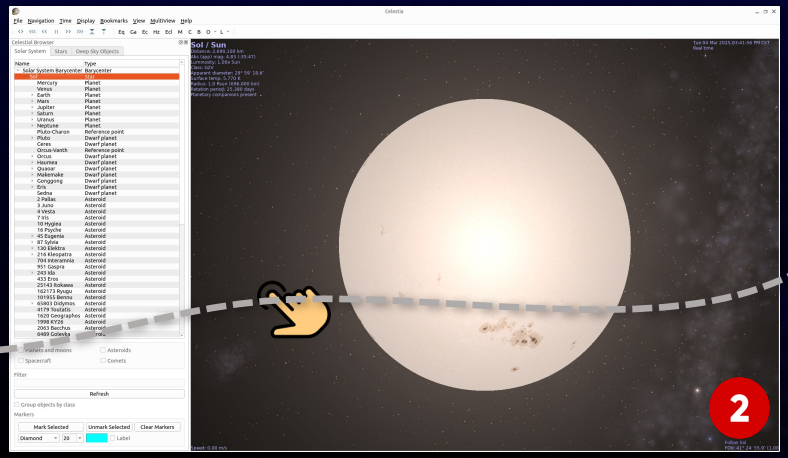


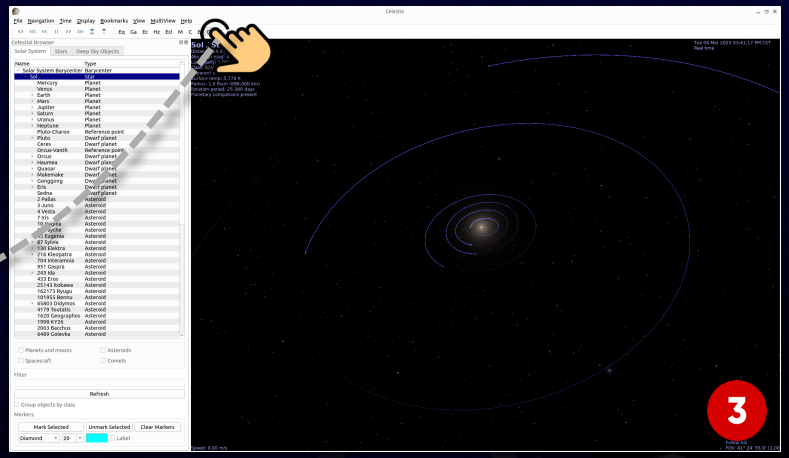
Step3: Wait until the prediction finished

Use Cases

Instruction: Show planets' orbits of Solar System in Celestia.

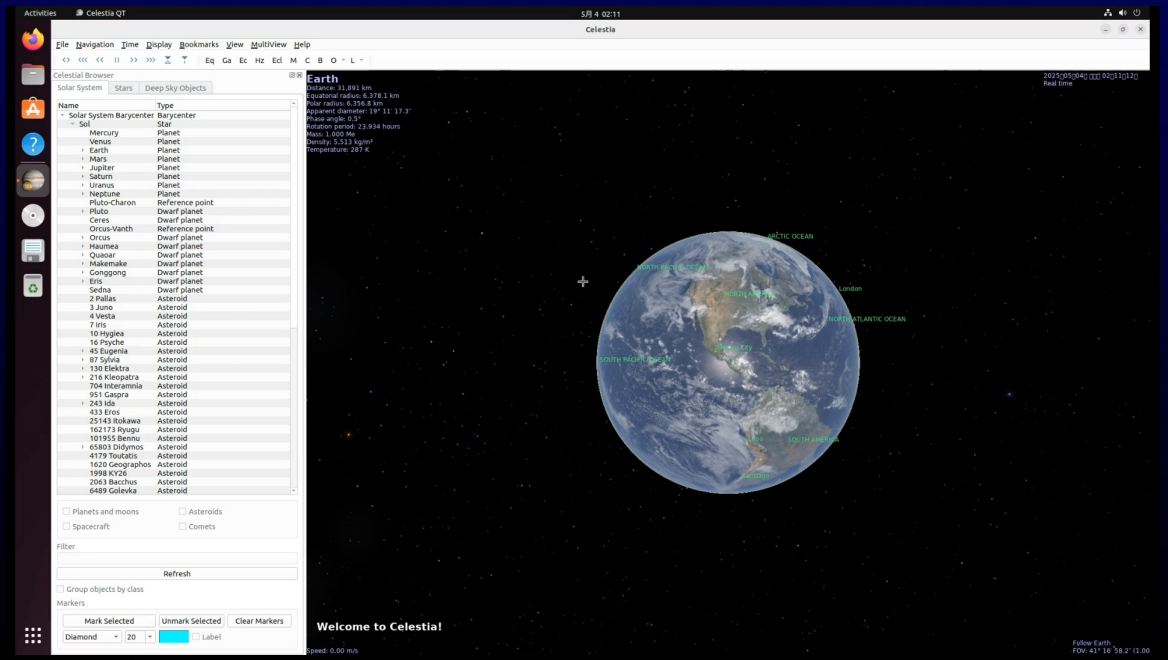






Step3: Click to show orbits of planets

Use Cases



Display constellations with their boundries in Celestia.

I see the current state of Celestia, where the Earth is being displayed in the main view. The task is to display constellations with their boundaries in Celestia.

Looking at the accessibility tree, I can see there are several checkboxes including "C" which likely stands for Constellations and "B" which likely stands for Boundaries. These checkboxes are located in the main toolbar and need to be checked to display the constellations and their boundaries.

Currently, neither the constellations nor their boundaries are visible in the screenshot. I need to click on both the "C" checkbox and "B" checkbox to enable these features.

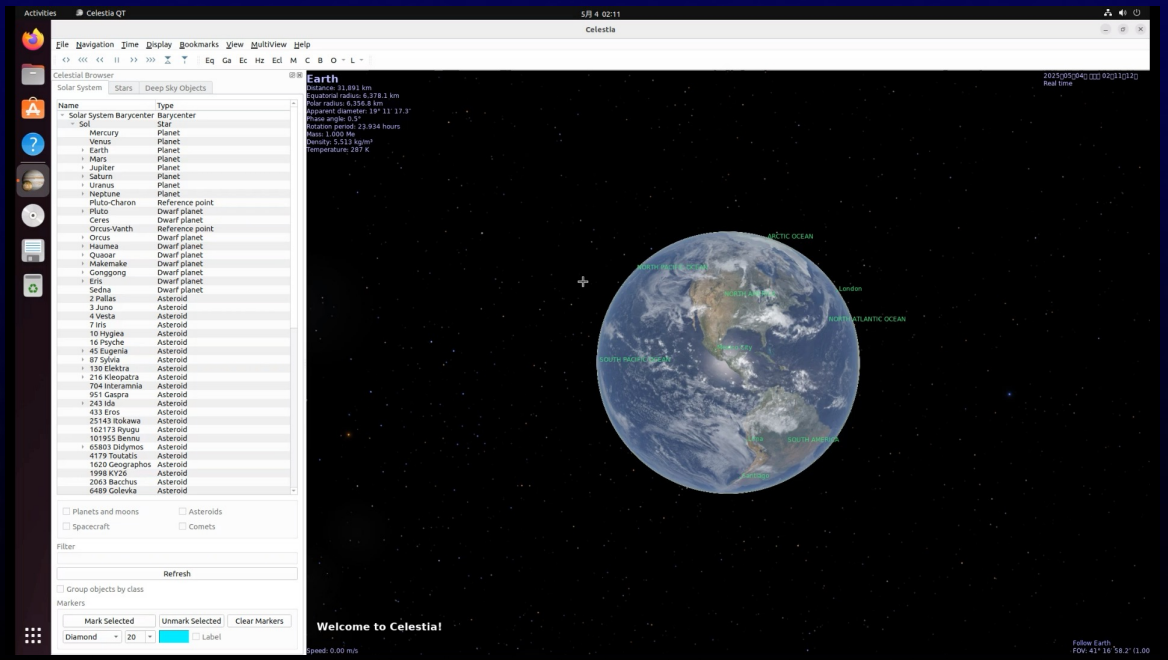
```python
import pyautogui
import time

# Click on the "C" checkbox to enable constellations
pyautogui.click(570, 93)
time.sleep(0.5)

# Click on the "B" checkbox to enable constellation boundaries
pyautogui.click(594, 93)
time.sleep(0.5)

# Zoom out a bit to get a better view of the constellations
pyautogui.press('minus')
time.sleep(0.5)
pyautogui.press('minus')
```

Use Cases



Looking at the screenshot, I can see that my previous action to click on "C" (Constellation) and "B" (Boundary) checkboxes worked correctly. I can see constellation lines displayed around the Earth (the blue lines showing the constellation patterns), and I can also see the boundary lines (the red/pink lines showing constellation boundaries).

To get a better view of the constellations and their boundaries, I should zoom out more and move away from Earth to see the full night sky. This will give a clearer view of all the constellations.

```python
import pyautogui
import time
# Zoom out more to get a better view of the sky
pyautogui.press('minus')
time.sleep(0.5)
pyautogui.press('minus')
time.sleep(0.5)
pyautogui.press('minus')
time.sleep(0.5)

# Use the right arrow key to rotate away from Earth
pyautogui.keyDown('right')
time.sleep(2)
pyautogui.keyUp('right')

# Move slightly upward to get a better view of the sky
pyautogui.keyDown('up')
time.sleep(1)
pyautogui.keyUp('up')
```

Evaluation: Observation

Proprietary Models

Opensource LLM / VLMs

GUI Action Models

OS-Atlas

NLP

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

[17] UI-TARS: Pioneering Automated GUI Interaction with Native Agents

[18] GUI-Actor: Coordinate-Free Visual Grounding for GUI Agents

Evaluation: Observation

Screenshot: Vision-only 

A11ytree: Text-only 

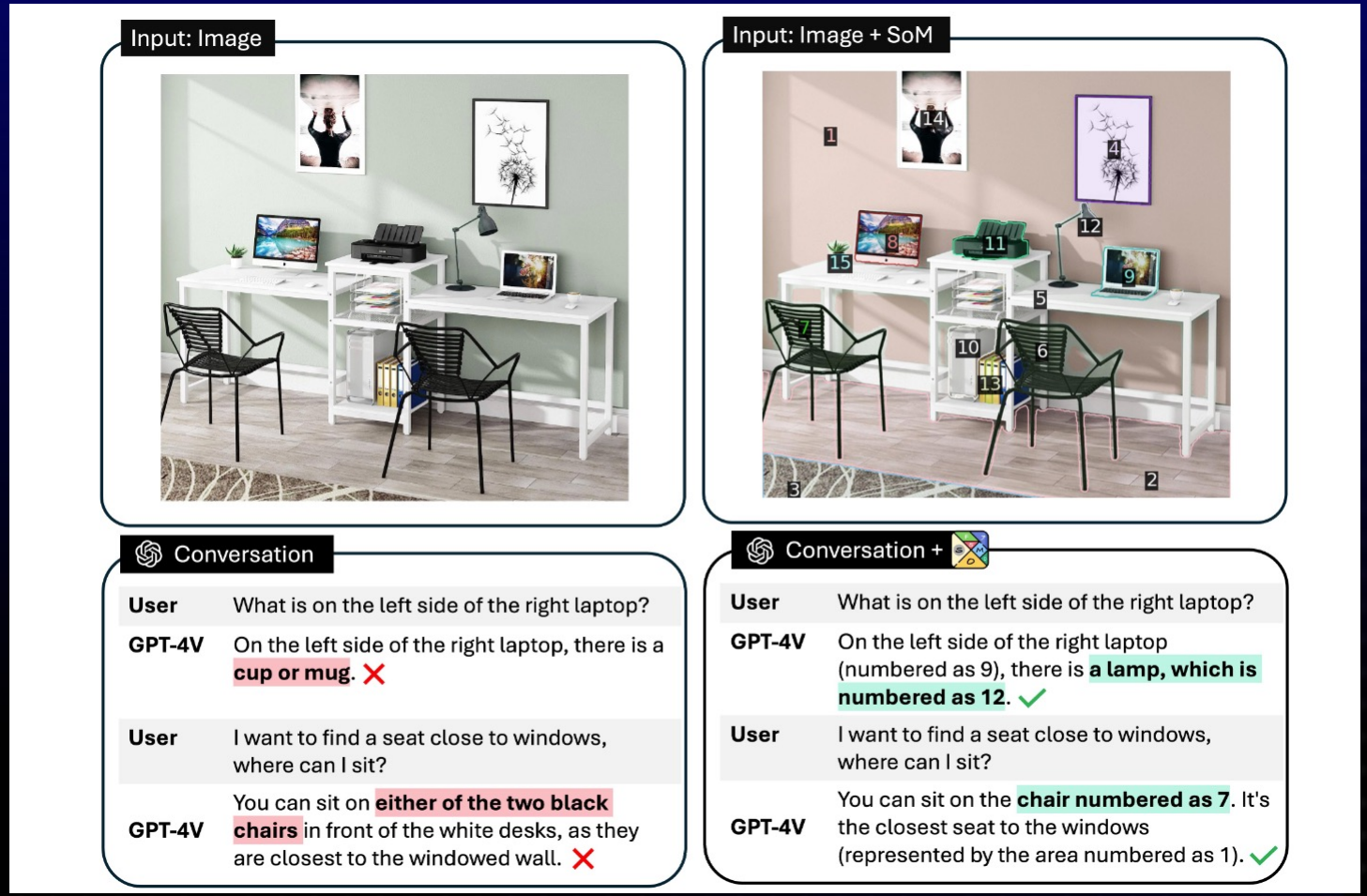
Screenshot + A11ytree: Hybird  

Set-of-Mark 

Evaluation: Observation

Set-of-Mark ? “segmentation”

Partition an image into regions at different levels of granularity + label



[19] Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V

Evaluation: General Setting

Our agents will generate ReACT-Style output

Step 1: To begin the molecular visualization task, I need to move the mouse to the ChimeraX icon.

action: `{"action_type": "moveTo", "x": 721, "y": 190}`

Step 2: I need to launch it.

action: `{"action_type": "click", "x": 251, "y": 300}`

Step 3: I move the mouse to the “Open” option and click to open the file dialog.

action: `{"action_type": "moveTo", "x": 120, "y": 90}`

...

Step n: I move to the “ball-and-stick” style option and click to apply it.

...

[20] ReAct: Synergizing Reasoning and Acting in Language Models, ICLR 2023

[21] OS-Genesis: Automating GUI Agent Trajectory Construction via Reverse Task Synthesis, ACL 2025

Evaluation: General Setting

Overall success rate remains low (avg. ~15%)

Performance varies among domains

Best results achieved with combined Screenshot + a11ytree setting

Table 3: Success rates on SCIENCEBOARD. We present the performance of each agent backbone across different scientific domains under various observation settings. Proprietary Models, Open-Source VLMs / LLMs, and GUI Action Model are distinguished by color.

Observations	Model	Success Rate (↑)						
		Algebra	Biochem	GIS	ATP	Astron	Doc	Overall
Screenshot	GPT-4o	3.23%	0.00%	0.00%	0.00%	0.00%	6.25%	1.58%
	Claude-3.7-Sonnet	9.67%	37.93%	2.94%	0.00%	6.06%	6.25%	10.48%
	Gemini-2.0-Flash	6.45%	3.45%	2.94%	0.00%	0.00%	6.06%	3.15%
	Qwen2.5-VL-72B	22.58%	27.59%	5.88%	0.00%	9.09%	12.50%	12.94%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	0.00%	6.25%	2.69%
	UI-TARS-1.5-7B	12.90%	13.79%	0.00%	0.00%	6.06%	0.00%	2.69%
a11ytree	GPT-4o	12.90%	20.69%	2.94%	0.00%	6.06%	0.00%	7.10%
	Claude-3.7-Sonnet	19.35%	34.48%	2.94%	3.85%	12.12%	0.00%	12.12%
	Gemini-2.0-Flash	9.68%	17.24%	0.00%	0.00%	0.00%	0.00%	4.49%
	o3-mini	16.13%	20.69%	2.94%	3.85%	15.15%	6.25%	10.84%
	Qwen2.5-VL-72B	9.68%	10.34%	2.94%	0.00%	3.03%	0.00%	4.33%
	InternVL3-78B	3.23%	3.45%	0.00%	0.00%	0.00%	0.00%	1.11%
Screenshot + a11ytree	GPT-4o	22.58%	37.93%	2.94%	7.69%	3.03%	12.50%	14.45%
	Claude-3.7-Sonnet	12.90%	41.37%	8.82%	3.85%	9.09%	18.75%	15.79%
	Gemini-2.0-Flash	16.13%	24.14%	2.94%	0.00%	18.18%	12.50%	12.32%
	Qwen2.5-VL-72B	16.13%	20.69%	2.94%	0.00%	18.18%	12.50%	11.74%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	3.03%	6.25%	3.20%
Set-of-Mark	GPT-4o	6.45%	3.45%	0.00%	0.00%	3.03%	12.50%	4.24%
	Claude-3.7-Sonnet	16.13%	31.03%	5.88%	0.00%	6.06%	12.50%	11.93%
	Gemini-2.0-Flash	3.23%	0.00%	0.00%	0.00%	3.03%	6.25%	2.09%
	Qwen2.5-VL-72B	6.45%	6.90%	2.94%	0.00%	3.03%	12.50%	6.36%
	QvQ-72B-Preview	0.00%	0.00%	2.94%	0.00%	3.03%	0.00%	0.49%
	InternVL3-78B	3.23%	6.90%	2.94%	0.00%	0.00%	0.00%	2.18%
Human Performance		74.19%	68.97%	55.88%	42.31%	51.52%	68.75%	60.27%

Evaluation: General Setting

Significant performance gaps across domains!

Agents perform much better in biochemistry and algebra compared to other fields.

Why? "Tutorial learning"

We see this as a key opportunity for the future development of science agents!

Table 3: Success rates on SCIENCEBOARD. We present the performance of each agent backbone across different scientific domains under various observation settings. Proprietary Models, Open-Source VLMs / LLMs, and GUI Action Model are distinguished by color.

Observations	Model	Success Rate (↑)						
		Algebra	Biochem	GIS	ATP	Astron	Doc	Overall
Screenshot	GPT-4o	3.23%	0.00%	0.00%	0.00%	0.00%	6.25%	1.58%
	Claude-3.7-Sonnet	9.67%	37.93%	2.94%	0.00%	6.06%	6.25%	10.48%
	Gemini-2.0-Flash	6.45%	3.45%	2.94%	0.00%	0.00%	6.06%	3.15%
	Qwen2.5-VL-72B	22.58%	27.59%	5.88%	0.00%	9.09%	12.50%	12.94%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	0.00%	6.25%	2.69%
	UI-TARS-1.5-7B	12.90%	13.79%	0.00%	0.00%	6.06%	0.00%	2.69%
a11ytree	GPT-4o	12.90%	20.69%	2.94%	0.00%	6.06%	0.00%	7.10%
	Claude-3.7-Sonnet	19.35%	34.48%	2.94%	3.85%	12.12%	0.00%	12.12%
	Gemini-2.0-Flash	9.68%	17.24%	0.00%	0.00%	0.00%	0.00%	4.49%
	o3-mini	16.13%	20.69%	2.94%	3.85%	15.15%	6.25%	10.84%
	Qwen2.5-VL-72B	9.68%	10.34%	2.94%	0.00%	3.03%	0.00%	4.33%
	InternVL3-78B	3.23%	3.45%	0.00%	0.00%	0.00%	0.00%	1.11%
Screenshot + a11ytree	GPT-4o	22.58%	37.93%	2.94%	7.69%	3.03%	12.50%	14.45%
	Claude-3.7-Sonnet	12.90%	41.37%	8.82%	3.85%	9.09%	18.75%	15.79%
	Gemini-2.0-Flash	16.13%	24.14%	2.94%	0.00%	18.18%	12.50%	12.32%
	Qwen2.5-VL-72B	16.13%	20.69%	2.94%	0.00%	18.18%	12.50%	11.74%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	3.03%	6.25%	3.20%
Set-of-Mark	GPT-4o	6.45%	3.45%	0.00%	0.00%	3.03%	12.50%	4.24%
	Claude-3.7-Sonnet	16.13%	31.03%	5.88%	0.00%	6.06%	12.50%	11.93%
	Gemini-2.0-Flash	3.23%	0.00%	0.00%	0.00%	3.03%	6.25%	2.09%
	Qwen2.5-VL-72B	6.45%	6.90%	2.94%	0.00%	3.03%	12.50%	6.36%
	QvQ-72B-Preview	0.00%	0.00%	2.94%	0.00%	3.03%	0.00%	0.49%
	InternVL3-78B	3.23%	6.90%	2.94%	0.00%	0.00%	0.00%	2.18%
Human Performance		74.19%	68.97%	55.88%	42.31%	51.52%	68.75%	60.27%

Evaluation: General Setting

ATP tasks remain particularly challenging. Why?

Because agents struggle to balance normal operations, coding skills, and highly logical reasoning.

Table 3: Success rates on SCIENCEBOARD. We present the performance of each agent backbone across different scientific domains under various observation settings. Proprietary Models, Open-Source VLMs / LLMs, and GUI Action Model are distinguished by color.

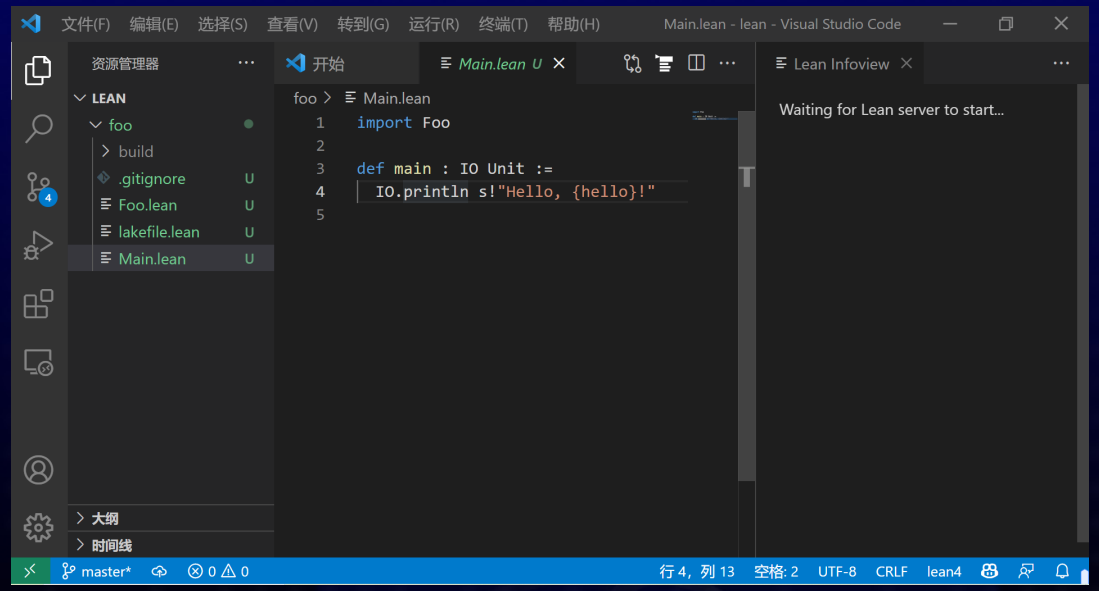
Observations	Model	Success Rate (↑)						
		Algebra	Biochem	GIS	ATP	Astron	Doc	Overall
Screenshot	GPT-4o	3.23%	0.00%	0.00%	0.00%	0.00%	6.25%	1.58%
	Claude-3.7-Sonnet	9.67%	37.93%	2.94%	0.00%	6.06%	6.25%	10.48%
	Gemini-2.0-Flash	6.45%	3.45%	2.94%	0.00%	0.00%	6.06%	3.15%
	Qwen2.5-VL-72B	22.58%	27.59%	5.88%	0.00%	9.09%	12.50%	12.94%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	0.00%	6.25%	2.69%
	UI-TARS-1.5-7B	12.90%	13.79%	0.00%	0.00%	6.06%	0.00%	2.69%
a11ytree	GPT-4o	12.90%	20.69%	2.94%	0.00%	6.06%	0.00%	7.10%
	Claude-3.7-Sonnet	19.35%	34.48%	2.94%	3.85%	12.12%	0.00%	12.12%
	Gemini-2.0-Flash	9.68%	17.24%	0.00%	0.00%	0.00%	0.00%	4.49%
	o3-mini	16.13%	20.69%	2.94%	3.85%	15.15%	6.25%	10.84%
	Qwen2.5-VL-72B	9.68%	10.34%	2.94%	0.00%	3.03%	0.00%	4.33%
	InternVL3-78B	3.23%	3.45%	0.00%	0.00%	0.00%	0.00%	1.11%
Screenshot + a11ytree	GPT-4o	22.58%	37.93%	2.94%	7.69%	3.03%	12.50%	14.45%
	Claude-3.7-Sonnet	12.90%	41.37%	8.82%	3.85%	9.09%	18.75%	15.79%
	Gemini-2.0-Flash	16.13%	24.14%	2.94%	0.00%	18.18%	12.50%	12.32%
	Qwen2.5-VL-72B	16.13%	20.69%	2.94%	0.00%	18.18%	12.50%	11.74%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	3.03%	6.25%	3.20%
Set-of-Mark	GPT-4o	6.45%	3.45%	0.00%	0.00%	3.03%	12.50%	4.24%
	Claude-3.7-Sonnet	16.13%	31.03%	5.88%	0.00%	6.06%	12.50%	11.93%
	Gemini-2.0-Flash	3.23%	0.00%	0.00%	0.00%	3.03%	6.25%	2.09%
	Qwen2.5-VL-72B	6.45%	6.90%	2.94%	0.00%	3.03%	12.50%	6.36%
	QvQ-72B-Preview	0.00%	0.00%	2.94%	0.00%	3.03%	0.00%	0.49%
	InternVL3-78B	3.23%	6.90%	2.94%	0.00%	0.00%	0.00%	2.18%
Human Performance		74.19%	68.97%	55.88%	42.31%	51.52%	68.75%	60.27%

Evaluation: General Setting

ATP tasks remain particularly challenging. Why?

Hard to perform human-like operations

e.g., Autocomplete



Evaluation: General Setting

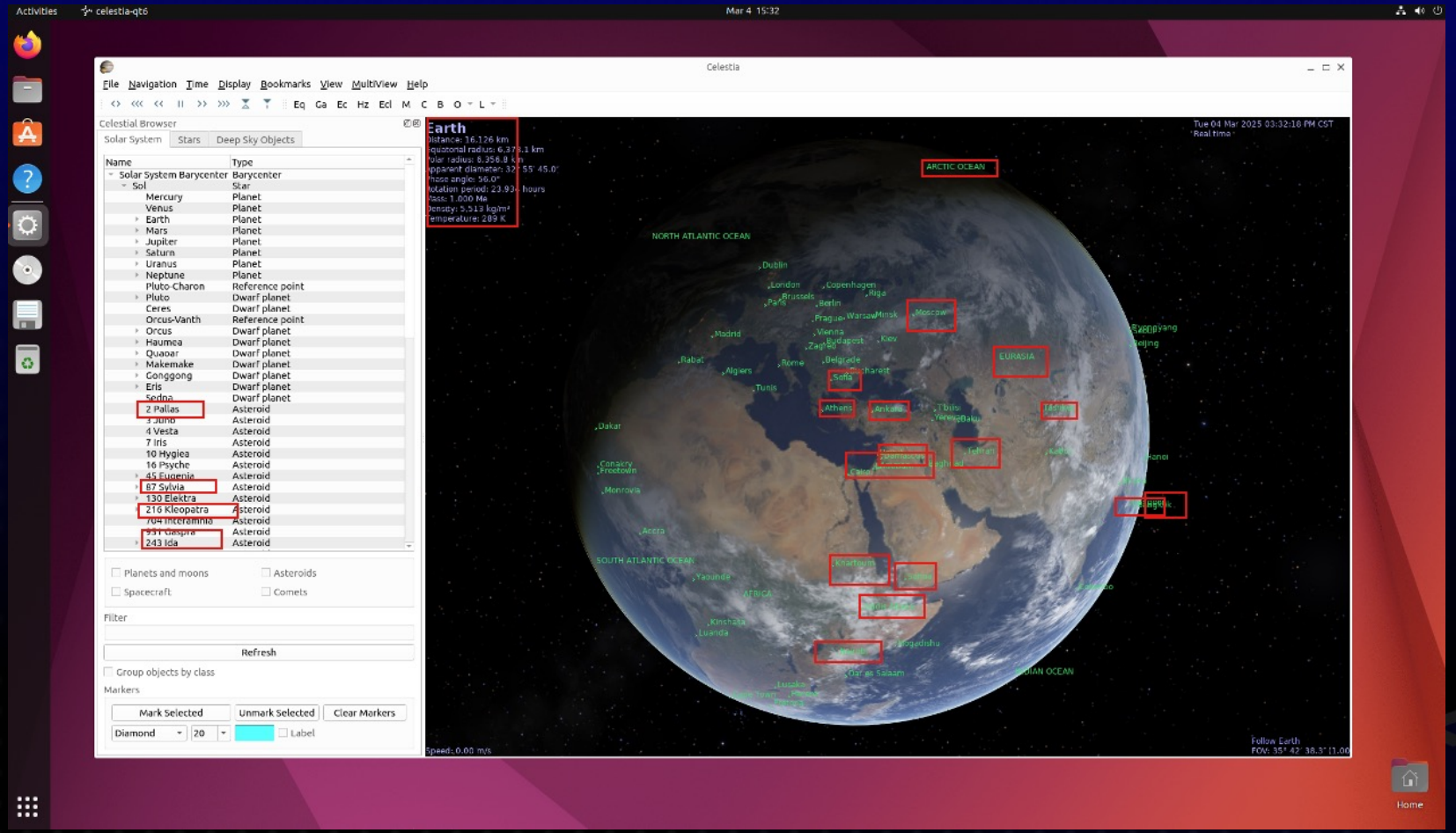
SoM? Does not fit all scenarios!

Table 3: Success rates on SCIENCEBOARD. We present the performance of each agent backbone across different scientific domains under various observation settings. Proprietary Models, Open-Source VLMs / LLMs, and GUI Action Model are distinguished by color.

Observations	Model	Success Rate (↑)						
		Algebra	Biochem	GIS	ATP	Astron	Doc	Overall
Screenshot	GPT-4o	3.23%	0.00%	0.00%	0.00%	0.00%	6.25%	1.58%
	Claude-3.7-Sonnet	9.67%	37.93%	2.94%	0.00%	6.06%	6.25%	10.48%
	Gemini-2.0-Flash	6.45%	3.45%	2.94%	0.00%	0.00%	6.06%	3.15%
	Qwen2.5-VL-72B	22.58%	27.59%	5.88%	0.00%	9.09%	12.50%	12.94%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	0.00%	6.25%	2.69%
	UI-TARS-1.5-7B	12.90%	13.79%	0.00%	0.00%	6.06%	0.00%	2.69%
a11ytree	GPT-4o	12.90%	20.69%	2.94%	0.00%	6.06%	0.00%	7.10%
	Claude-3.7-Sonnet	19.35%	34.48%	2.94%	3.85%	12.12%	0.00%	12.12%
	Gemini-2.0-Flash	9.68%	17.24%	0.00%	0.00%	0.00%	0.00%	4.49%
	o3-mini	16.13%	20.69%	2.94%	3.85%	15.15%	6.25%	10.84%
	Qwen2.5-VL-72B	9.68%	10.34%	2.94%	0.00%	3.03%	0.00%	4.33%
	InternVL3-78B	3.23%	3.45%	0.00%	0.00%	0.00%	0.00%	1.11%
Screenshot + a11ytree	GPT-4o	22.58%	37.93%	2.94%	7.69%	3.03%	12.50%	14.45%
	Claude-3.7-Sonnet	12.90%	41.37%	8.82%	3.85%	9.09%	18.75%	15.79%
	Gemini-2.0-Flash	16.13%	24.14%	2.94%	0.00%	18.18%	12.50%	12.32%
	Qwen2.5-VL-72B	16.13%	20.69%	2.94%	0.00%	18.18%	12.50%	11.74%
	InternVL3-78B	6.45%	3.45%	0.00%	0.00%	3.03%	6.25%	3.20%
Set-of-Mark	GPT-4o	6.45%	3.45%	0.00%	0.00%	3.03%	12.50%	4.24%
	Claude-3.7-Sonnet	16.13%	31.03%	5.88%	0.00%	6.06%	12.50%	11.93%
	Gemini-2.0-Flash	3.23%	0.00%	0.00%	0.00%	3.03%	6.25%	2.09%
	Qwen2.5-VL-72B	6.45%	6.90%	2.94%	0.00%	3.03%	12.50%	6.36%
	QvQ-72B-Preview	0.00%	0.00%	2.94%	0.00%	3.03%	0.00%	0.49%
	InternVL3-78B	3.23%	6.90%	2.94%	0.00%	0.00%	0.00%	2.18%
Human Performance		74.19%	68.97%	55.88%	42.31%	51.52%	68.75%	60.27%

Evaluation: General Setting

"Cognitive burden"



Evaluation: Modular Setting

Planner + Action Model

Step 1: To begin the molecular visualization task, I need to move the mouse to the ChimeraX icon.

action: `{"action_type": "moveTo", "x": 721, "y": 190}`

Step 2: I need to launch it.

action: `{"action_type": "click", "x": 251, "y": 300}`

Step 3: I move the mouse to the “Open” option and click to open the file dialog.

action: `{"action_type": "moveTo", "x": 120, "y": 90}`

...

Step n: I move to the “ball-and-stick” style option and click to apply it.

...

[20] ReAct: Synergizing Reasoning and Acting in Language Models, ICLR 2023

[21] OS-Genesis: Automating GUI Agent Trajectory Construction via Reverse Task Synthesis, ACL 2025

Evaluation: Modular Setting

GPT-4o as the planner + GUI model

Clear performance improvement (up to ~20% SR)

Separating planning and action offers a promising direction!

Table 4: Success rates of different VLM agent combinations under the planner + grounding model setting on SCIENCEBOARD. The observation setting used in this experiment is screenshot. Colors denote Proprietary Models, Open-Source VLMs and GUI Action Models.

Planner	Grounding Model	Success Rate (↑)				
		Algebra	Biochem	GIS	Astron	Overall
GPT-4o	OS-Atlas-Pro-7B	6.25%	10.34%	0.00%	3.03%	4.92%
	UGround-V1-7B	0.00%	3.45%	0.00%	3.03%	1.62%
	Qwen2.5-VL-72B	12.50%	34.48%	11.76%	9.09%	16.96%
	UI-TARS-72B	3.23%	10.34%	5.88%	6.06%	6.38%
	GUI-Actor-7B	21.88%	44.83%	2.94%	12.12%	20.44%
GPT-4o		3.23%	0.00%	0.00%	0.00%	0.81%

Analysis

CLI is very helpful.

Finding: (V)LMs tend to prefer completing tasks via CLI when possible.

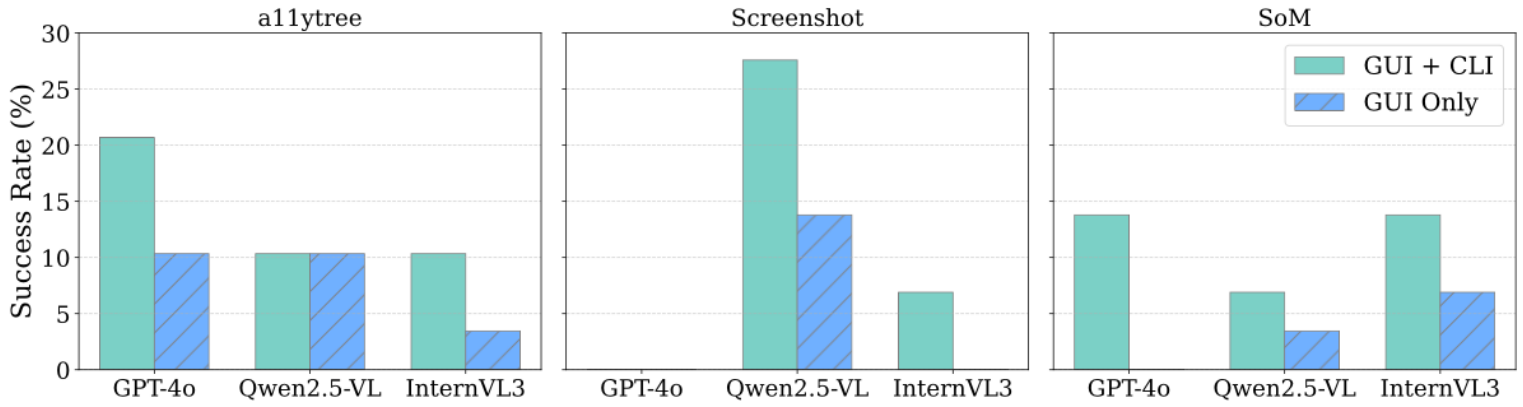


Figure 8: Extended analysis of Vision-Only vs. Hybrid Interface.

More analysis available in the paper!

Leaderboard

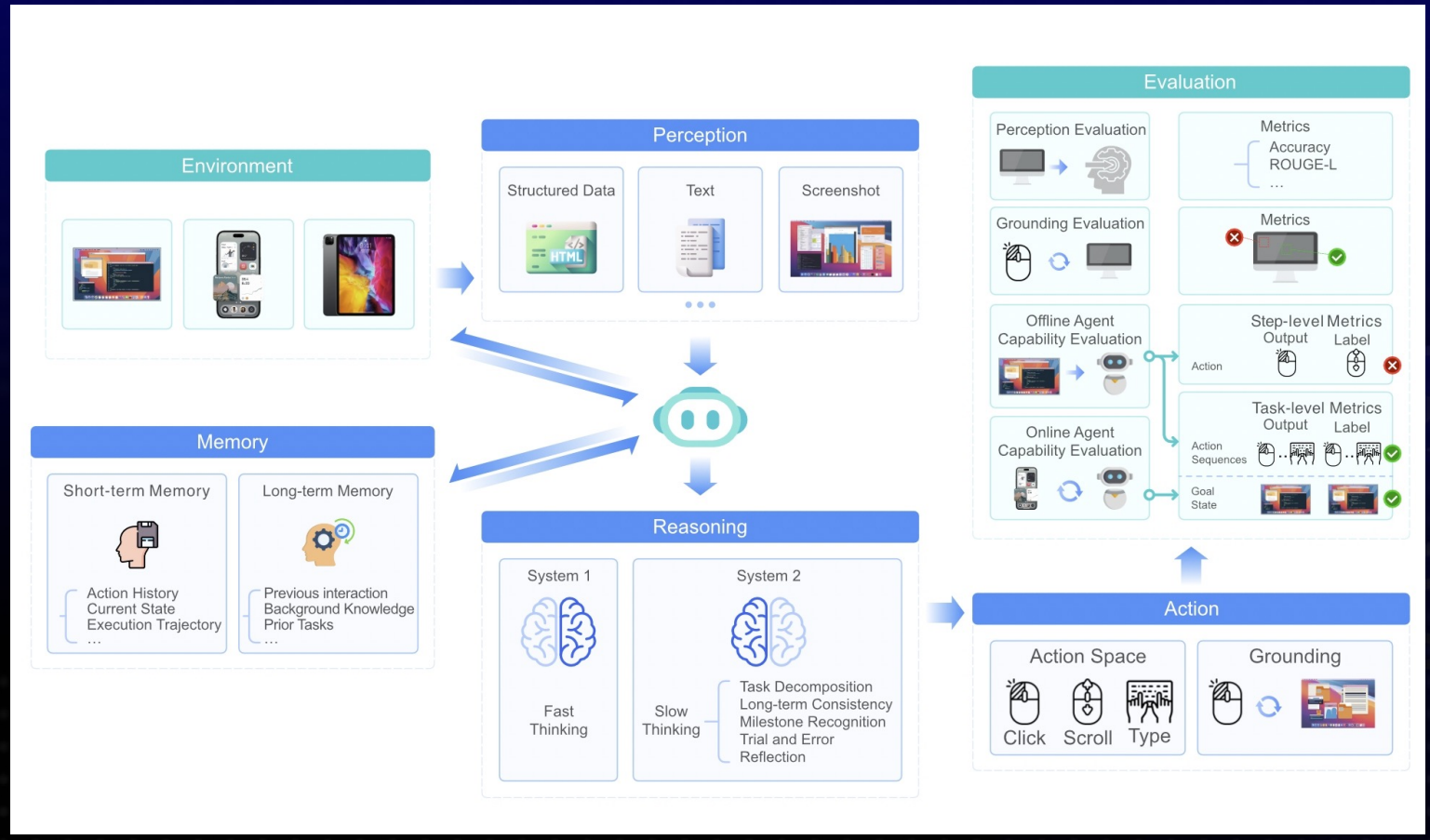
Screenshot | A11y Tree | Screenshot + A11y Tree | Set of Marks | Search by keywords

O..	Settings	% Acc ↓	% Alg	% Biochem	% GIS	% ATP	% Astron	% Doc
	Calude-3.7-Sonnet w/ screenshot…	15.79	12.90	41.37	8.82	3.85	9.09	18.75
	GPT-4o (2024-08-06) w/ screensh…	14.45	22.58	37.93	2.94	7.69	3.03	12.50
	GPT-4o (2024-08-06) w/ set_of_m…	14.45	6.45	3.45	0.00	0.00	3.03	12.50
	Qwen2.5-VL-72B w/ screenshot	12.94	22.58	27.59	5.88	0.00	9.09	12.50
	Gemini-2.0-Flash w/ screenshot+a…	12.32	16.13	24.14	2.94	0.00	18.18	12.50
	Calude-3.7-Sonnet w/ a11y_tree	12.12	19.35	34.48	2.94	3.85	12.12	0.00
	Calude-3.7-Sonnet w/ set_of_marks	11.93	16.13	31.03	5.88	0.00	6.06	12.50
	Qwen2.5-VL-72B w/ screenshot+a…	11.74	16.13	20.69	2.94	0.00	18.18	12.50
	o3-mini (2025-01-31) w/ a11y_tree	10.84	16.13	20.69	2.94	3.85	15.15	6.25
	Calude-3.7-Sonnet w/ screenshot	10.48	9.67	37.93	2.94	0.00	6.06	6.25
	GPT-4o (2024-08-06) w/ a11y_tree	7.10	12.90	20.69	2.94	0.00	0.00	6.06
	Qwen2.5-VL-72B w/ set_of_marks	6.36	6.45	6.90	2.94	0.00	3.03	12.50
	UI-TARS-1.5 w/ screenshot	5.92	12.90	13.79	0.00	0.00	6.06	0.00
	Gemini-2.0-Flash w/ a11y_tree	4.49	9.68	17.24	0.00	0.00	0.00	0.00
	Qwen2.5-VL-72B w/ a11y_tree	4.33	9.68	10.34	2.94	0.00	3.03	0.00
	InternVL3-78B w/ screenshot+a11…	3.20	6.45	3.45	0.00	0.00	3.03	6.25
	Gemini-2.0-Flash w/ screenshot	3.15	6.45	3.45	2.94	0.00	0.00	6.06

https://qiushisun.github.io/ScienceBoard-Home/

Towards High Value Area

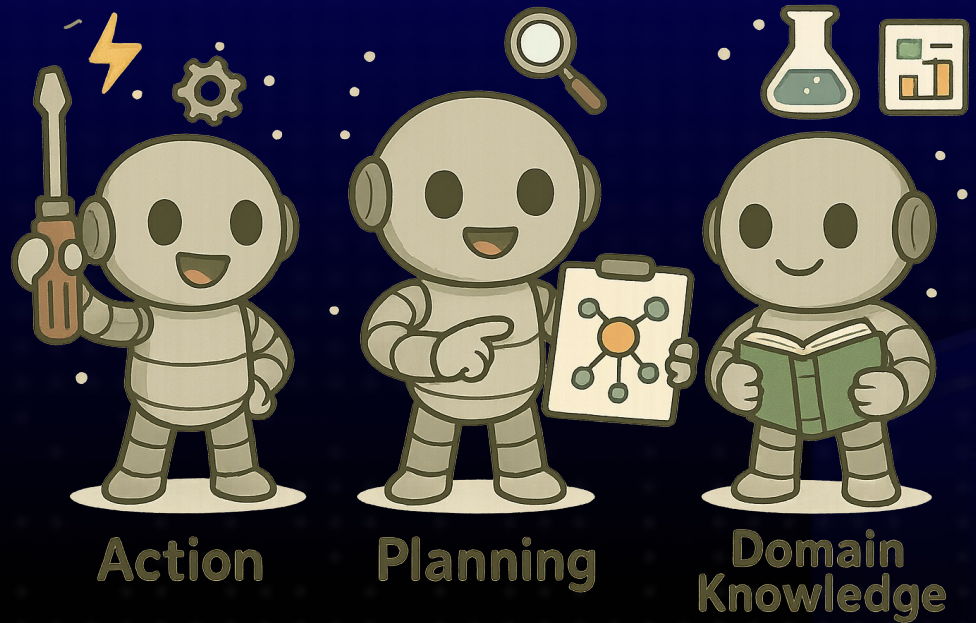
Recap: Core capabilities and evaluation for GUI agents



[22] UI-TARS: Pioneering Automated GUI Interaction with Native Agents

Towards High Value Area

For science agents, we need to strike a balance!



Some Limitations

1. The current evaluation uses binary (0/1) scoring; allowing partial credit for intermediate steps may better reflect real-world scenarios.

Challenge: exploration space!

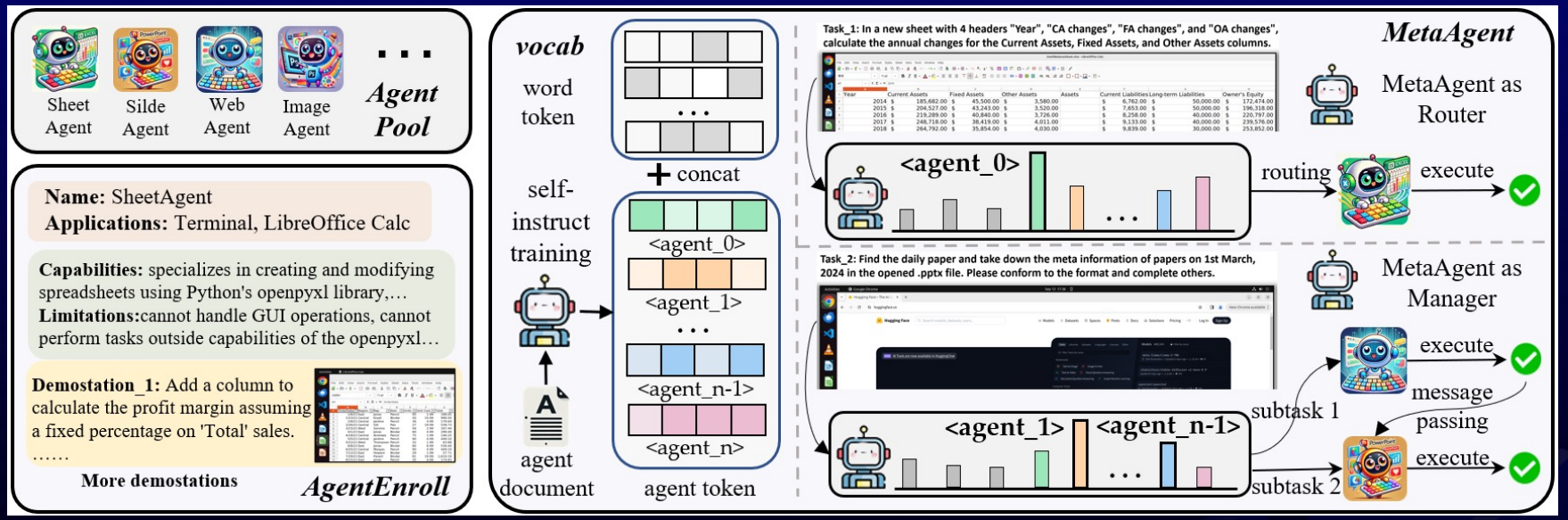
2. All evaluated software is open-source distribute; incorporating commercial software remains challenging.

Challenge: evaluation!

[23] AgentBoard: An Analytical Evaluation Board of Multi-turn LLM Agents, NIPS 2024 Oral

Part4 | Future Direction

Future Directions



Example: Heterogeneous Agents As Specialized Generalist Computer Assistant

AGENTSTORE: SCALABLE INTEGRATION OF HETEROGENEOUS AGENTS AS SPECIALIZED GENERALIST COMPUTER ASSISTANT

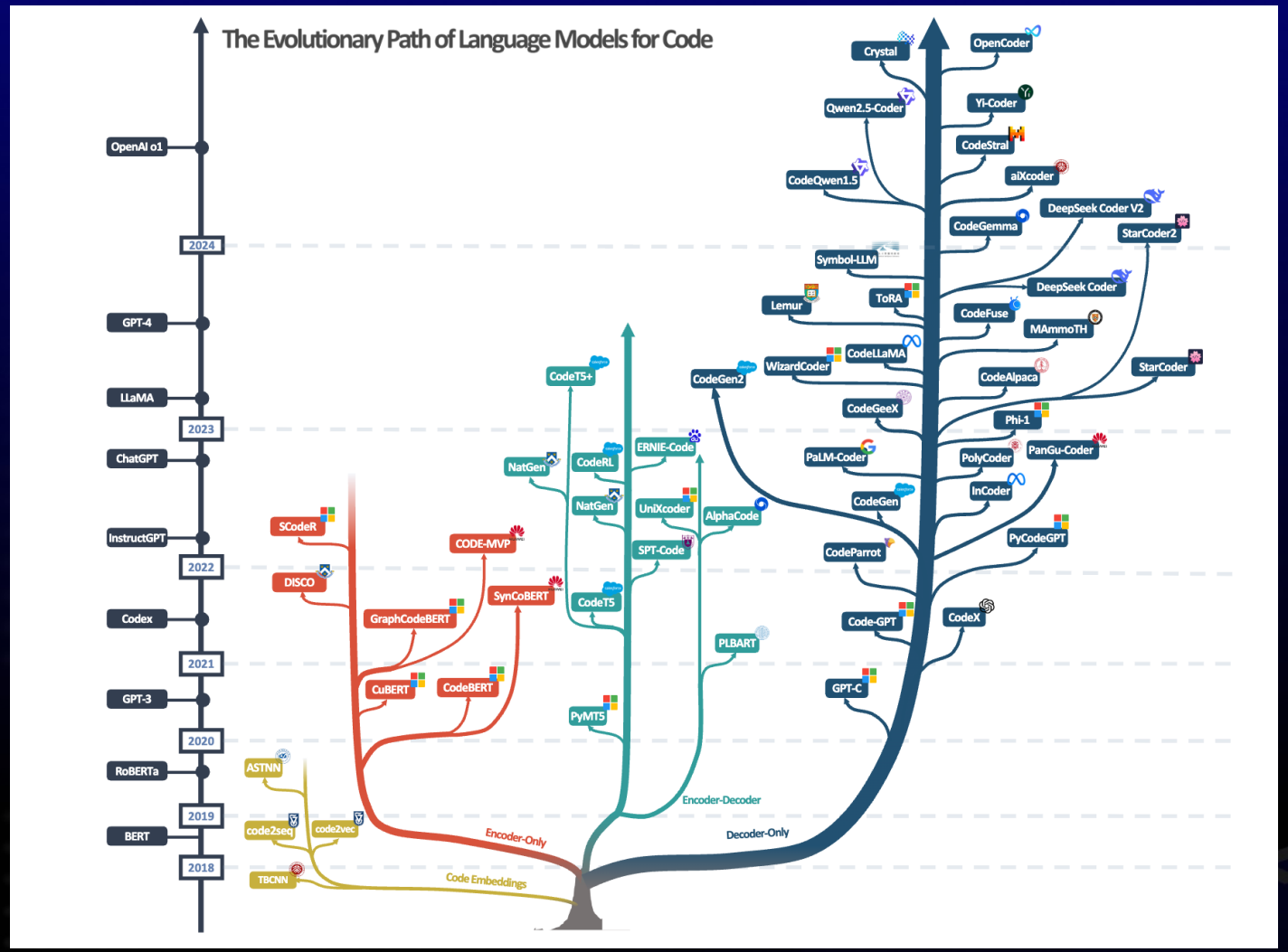
Chengyou Jia[1,2]*, **Minnan Luo**[1]✉, **Zhuohang Dang**[1], **Qiushi Sun**[2,3], **Fangzhi Xu**[1,2], **Junlin Hu**[2], **Tianbao Xie**[3], **Zhiyong Wu**[2]✉
[1]Xi'an Jiaotong University, [2]Shanghai AI Lab, [3]The University of Hong Kong
cp3jia@stu.xjtu.edu.cn, wuzhiyong@pjlab.org.cn

[24] AgentStore: Scalable Integration of Heterogeneous Agents As Specialized Generalist Computer Assistant, ACL 2025

Future Directions

Integration with CodeLLMs?

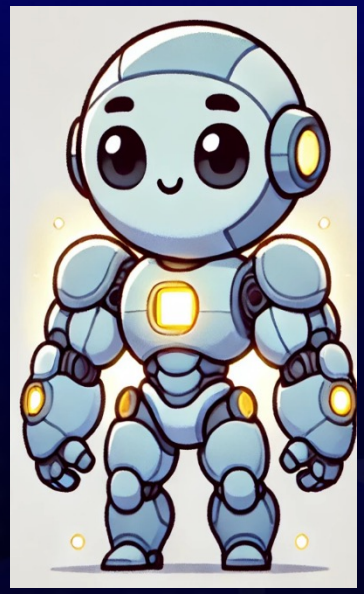
This enables the adoption of more data-driven methods, like ScienceAgentBench.



[25] A Survey of Neural Code Intelligence: Paradigms, Advances and Beyond

Future Directions

Mid-Training



Potential solution: Mid-training?

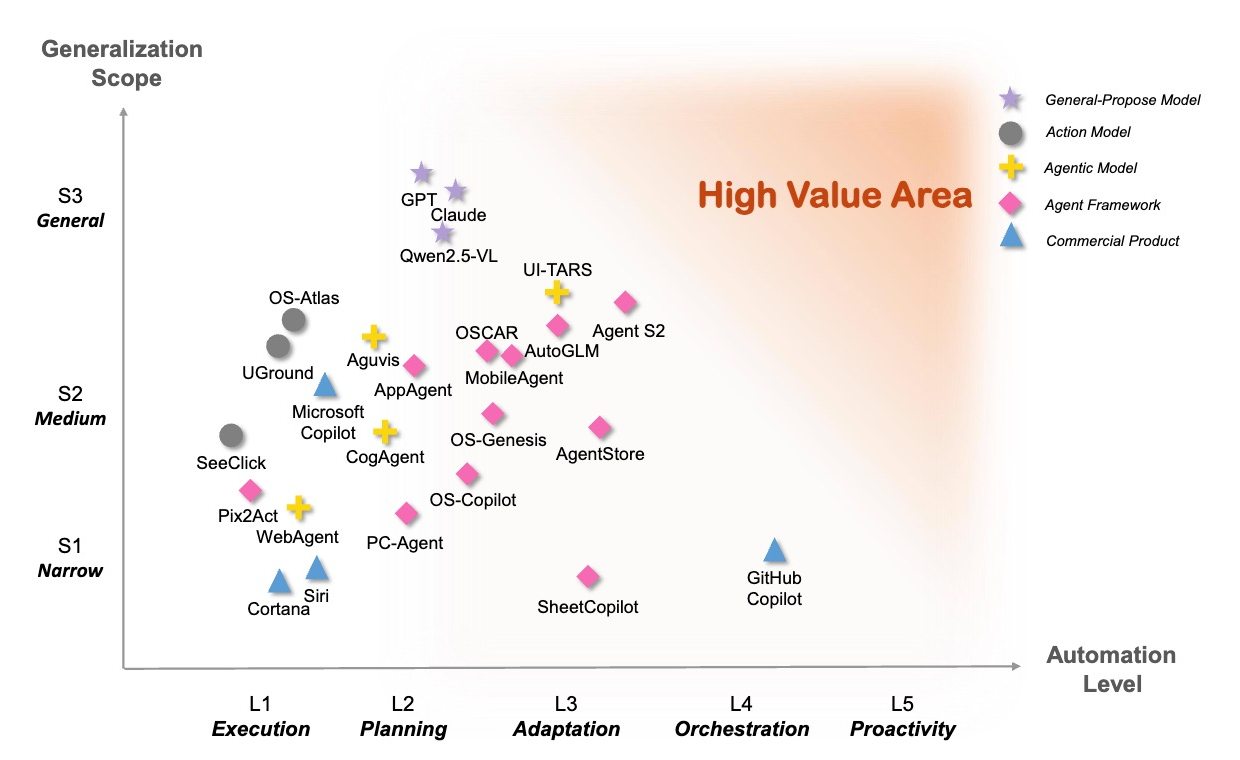
Breaking the Data Barrier – Building GUI Agents Through Task Generalization

Junlei Zhang[*◇☆] **Zichen Ding**[*♠] **Chang Ma**[♣] **Zijie Chen**[◇☆] **Qiushi Sun**[♣]
Zhenzhong Lan[☆] **Junxian He**[★]
[◇]Zhejiang University [☆] Westlake University [♠] Shanghai AI Laboratory
[♣] The University of Hong Kong [★]HKUST

[26] Breaking the Data Barrier -- Building GUI Agents Through Task Generalization, COLM 2025

Future Directions

Towards High Value Area



OS-MAP: How Far Can Computer Use Agents Go in Breadth and Depth?

Preprint / WUCA @ ICML 2025

Towards High Value Area

1. Operating robotic arms?

2. Controlling exoskeletons?

3. Utilizing highly specialized scientific software?

4. …

We are just standing at the dawn of a long journey!

中文解读（ScienceBoard）

ScienceBoard

Evaluating Multimodal Autonomous Agents in Realistic Scientific Workflows

Introducing ScienceBoard, a first-of-its-kind evaluation platform for multimodal agents in ***scientific workflows***. ScienceBoard is characterized by the following core features:

- **Pioneering Application**: ScienceBoard is the first to bring computer-using agents into the domain of scientific discovery, enabling autonomous research assistants across disciplines.
- **Realistic Environment**: We provide a dynamic, visually grounded virtual environment integrated with professional scientific software, supporting both GUI and CLI interaction in real-time workflows.
- **Challenging Benchmark**: A new benchmark of 169 rigorously validated tasks across 6 core domains is introduced, capturing real-world challenges.
- **Comprehensive Evaluations**: We presents systematic evaluations across a wide range of agents powered by LLMs, VLMs, and GUI action models.

arXiv | Code | Data | VM Snapshot

https://qiushisun.github.io/ScienceBoard-Home/

We are just standing at the dawn of a long journey!

Thanks for listening!

Contact: qiushisun@connect.hku.hk