UIPro: Unleashing Superior Interaction Capability For GUI Agents

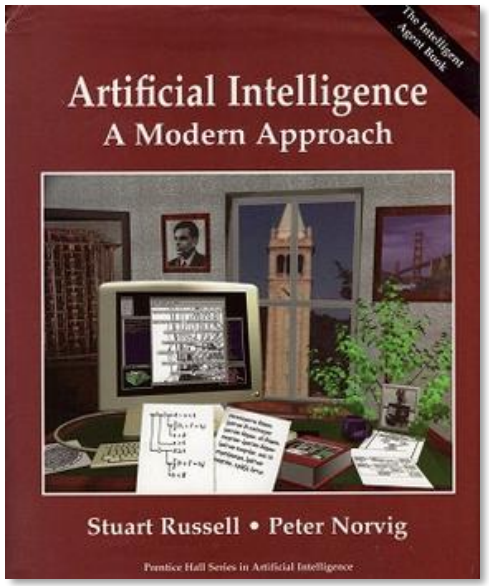
Hongxin Li[1,2,3,7] Jingran Su[5] Jingfan Chen[5] Zheng Ju[1,2,3] Yuntao Chen[4] Qing Li[5]
Zhaoxiang Zhang[1,2,3,6]
[1]UCAS, [2]NLPR, CASIA, [3]MAIS, CASIA, [4]Hong Kong Institute of Science & Innovation, CASIA
[5]PolyU, [6]Shanghai AI Lab, [7]StepFun

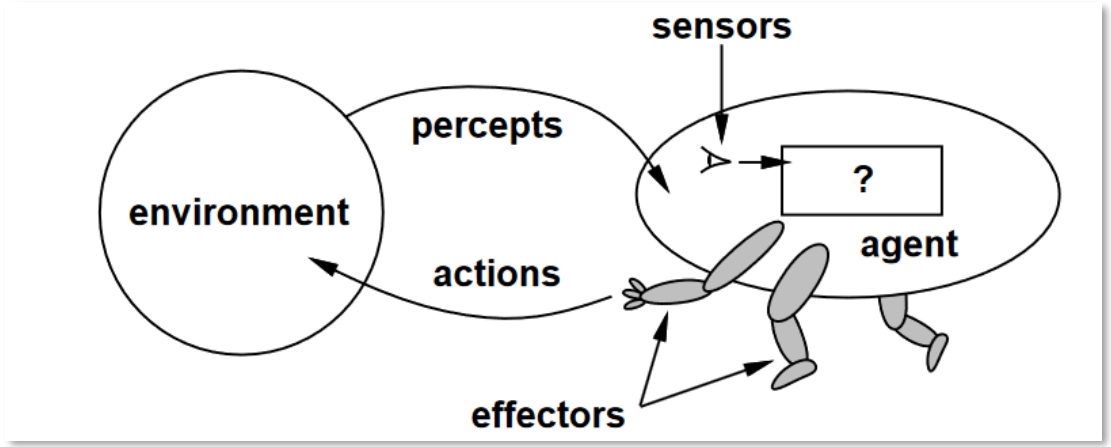
1 Research background - agents

"An agent is anything that can be thought of as perceiving its environment through sensors and acting on that environment through actuators."

"The hallmark of intelligence is **rational decision-making**, while artificial intelligence can be seen as the study and construction of rational agents"

—— "Artificial Intelligence: A Modern Approach" by Stuart Russell

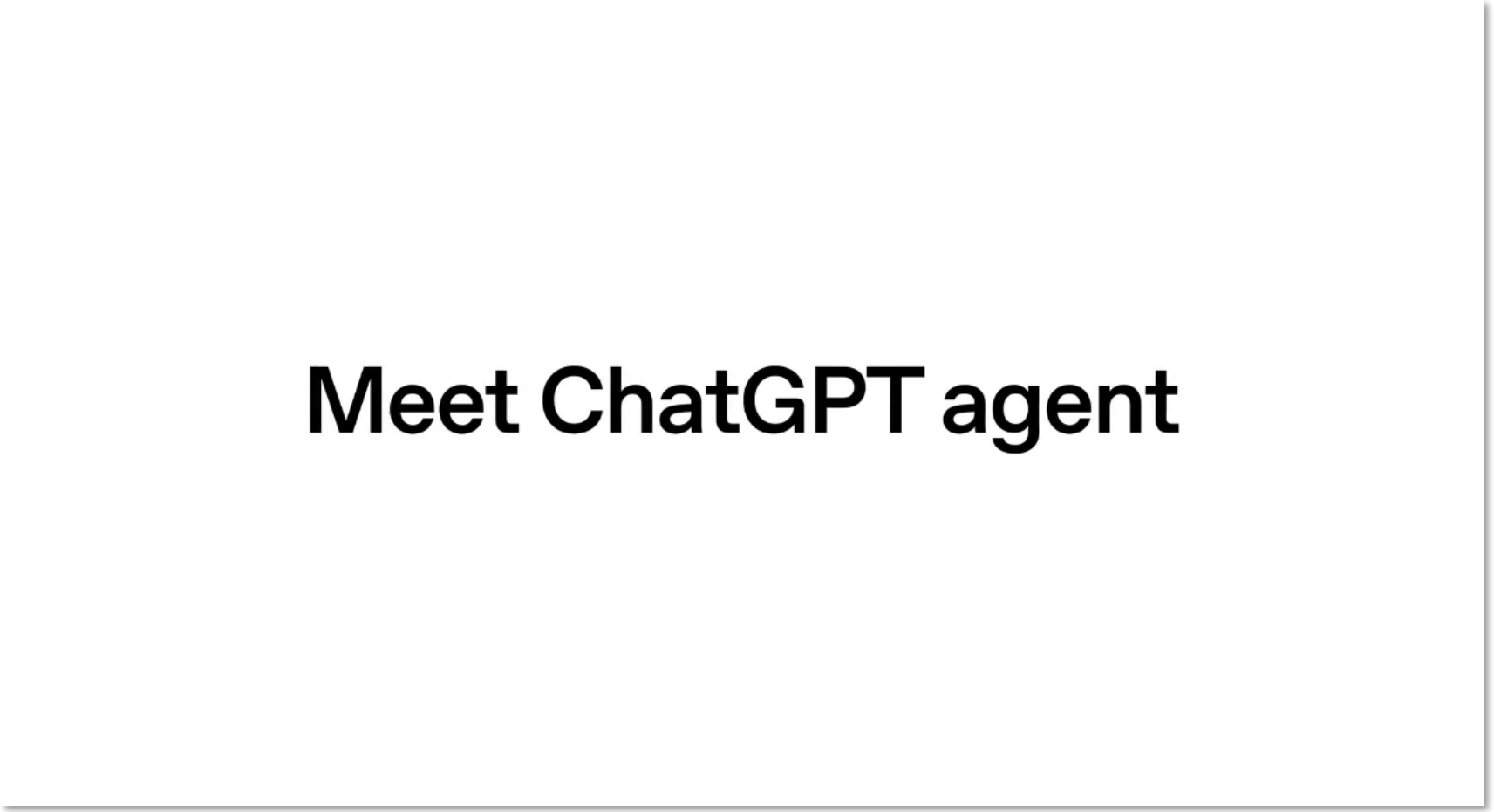




Agents and the environment

1 Research Background - Agents are no longer abstract concepts

- ChatGPT Agent can manipulate tables while also interacting visually with web pages.
- The model is specifically trained to recognize and apply the most suitable tools at each step, learning dynamically as it performs tasks – adapting the way it works by optimizing speed, accuracy, and efficiency.



1 Research background - The rise of large model agents

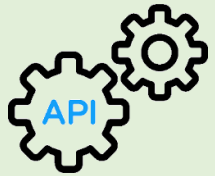
Thanks to the massive knowledge, instruction following ability and advanced post-training algorithms of large models, agents have been able to initially interact with the real world.

robot

Autonomous vehicles

Software tools

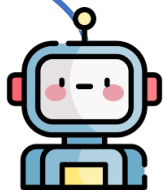
Automated operations



Decision-making

Real World

Multiple rounds of interaction



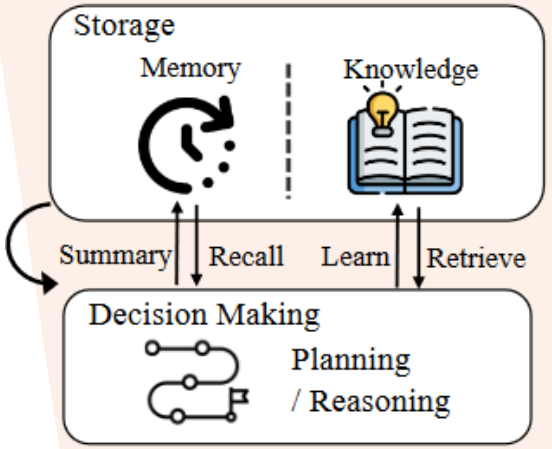
LLM/VLM agents

Input Processing

Multimodal information understanding

Basic competencies

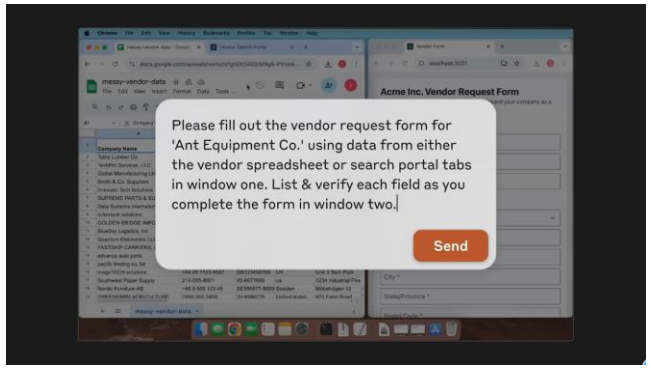
- Task breakdown
- Information analysis
- Reasoning and decision-making
- Reflection and Traceback



The Rise and Potential of Large Language Model Based Agents: A Survey, 2023

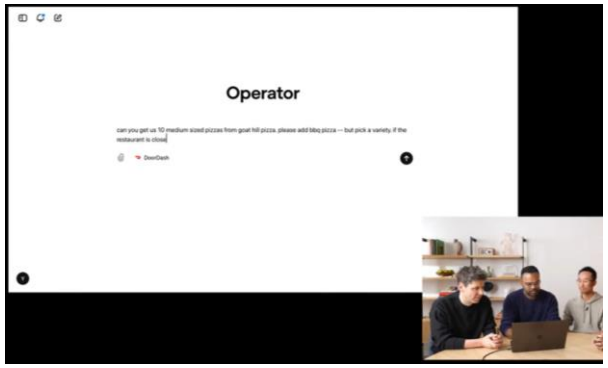
1 Research background - Digital agents have attracted attention

Digital agents are starting to gain attention due to their immense value in workflow automation



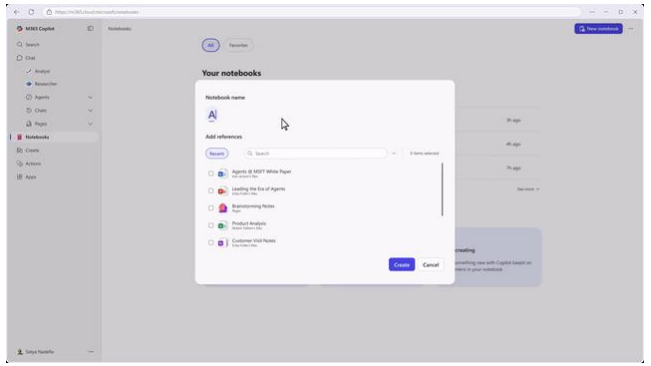
Take over software office processes

Claude



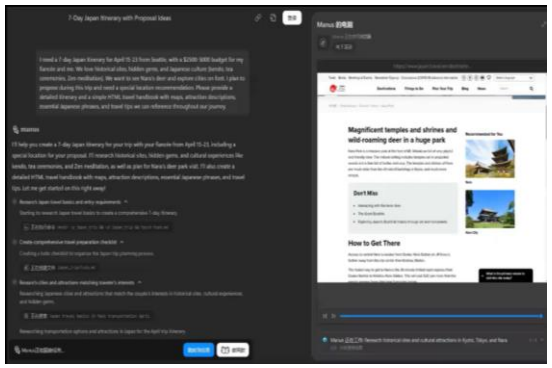
Automation of information

OpenAI Operator



Office personal assistant

Copilot



Personal life convenience assistant

manus
The general AI agent

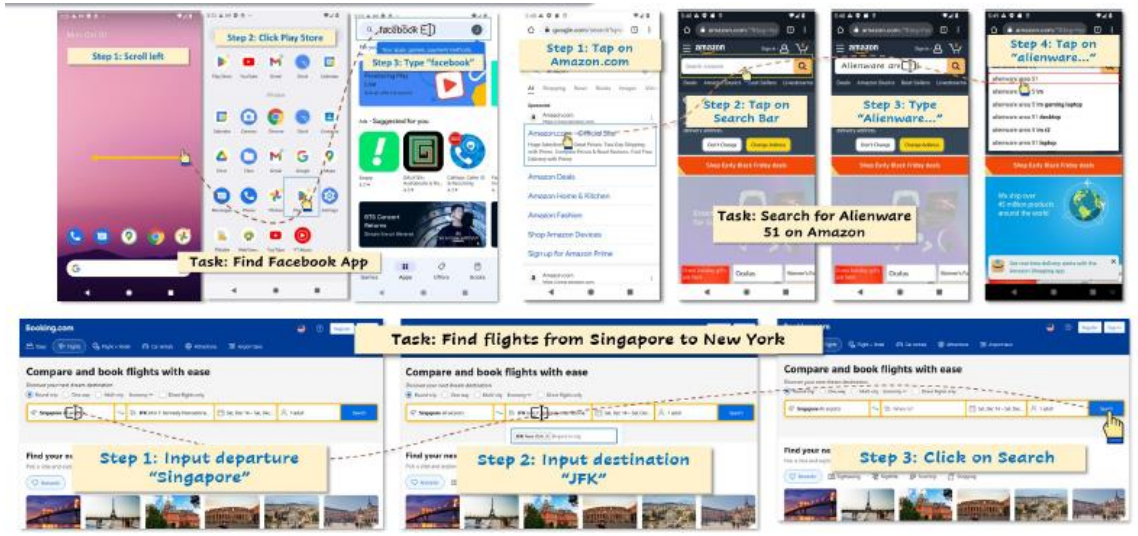
Typical representatives are OpenAI Operator, which combines the multimodal understanding capabilities of the commercial closed-source model (GPT-4o) and the agent reasoning interaction behavior through reinforcement learning

2 Our Agent - UIPro

GUI Operation Across Multiple Platforms

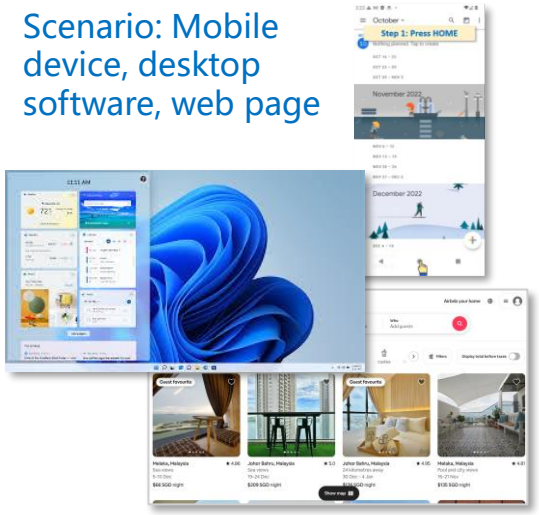
Background: The generalizable interaction ability of agents relies on large-scale pre-training data, but the existing datasets are not uniform and isolated, and most of them are only for a single platform.

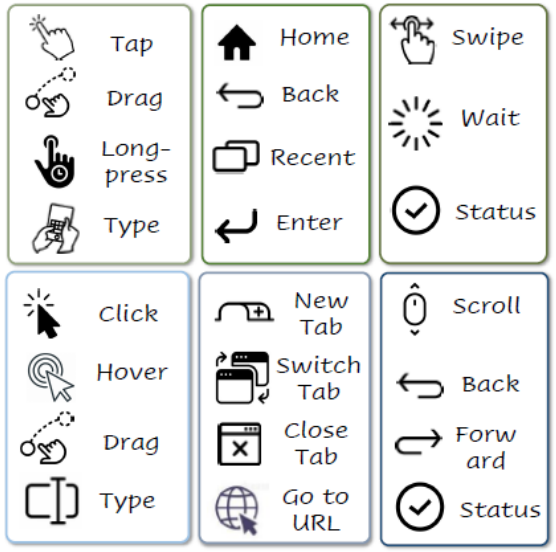
Methods: A set of systematic heterogeneous data cleaning and sorting processes is proposed, and multiple data sources of different formats are merged to realize the integration of cross-platform agent training data



Step 1. The system cleans and organizes data from different sources.

Scenario: Mobile device, desktop software, web page





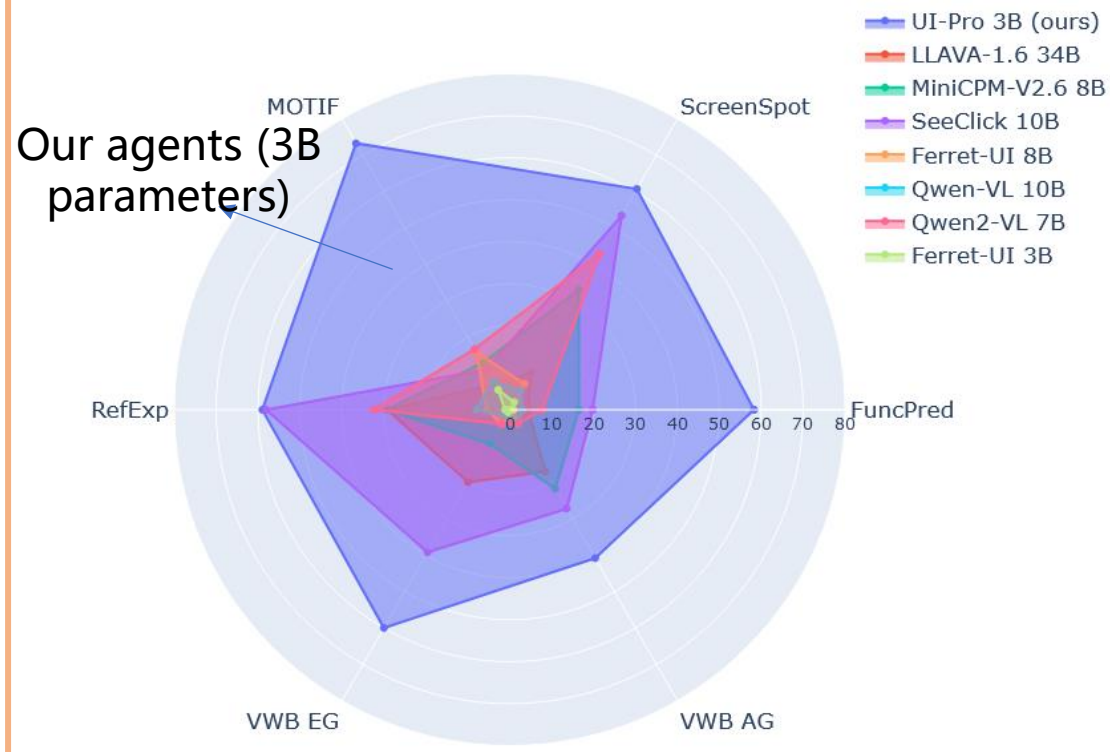
Step 2. Unify the action space of different digital interface platforms, so that the interaction skills learned by agents can be generalized across platforms

Method	Scenarios	#Annos	#TaskTypes
Screen2Words (Google)	Mobile	112k	1
Wid. Captioning (Google)	Mobile	163k	1
SeeClick (NJU)	Web, Mobile	5.3M	6
OS-ATLAS	Web, Mobile,Desktop	13.6M	5
Ours	Multi Platforms	**>20M**	**>13**

Results: Our unified datasets provide a much larger scale of annotation and task richness than previous datasets

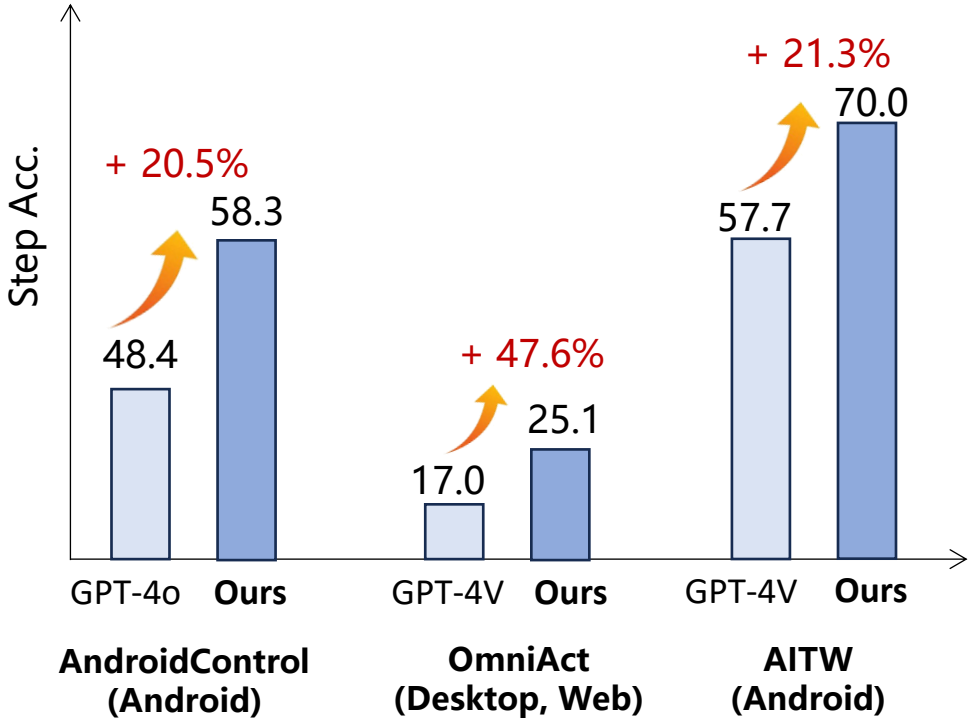
2 Our Agent - UIPro

GUI Element Positioning: Locate tiny elements on GUI screenshots based on element descriptions



Thanks to high-quality data, our agents surpass existing open-source methods on multiple elemental positioning benchmarks

Agent Task: Predict interaction with the GUI according to the user's task requirements



Thanks to the unification of data and action space, the success rate of our agent's interaction tasks on multiple platforms far exceeds that of commercial closed-source models

2 Demos

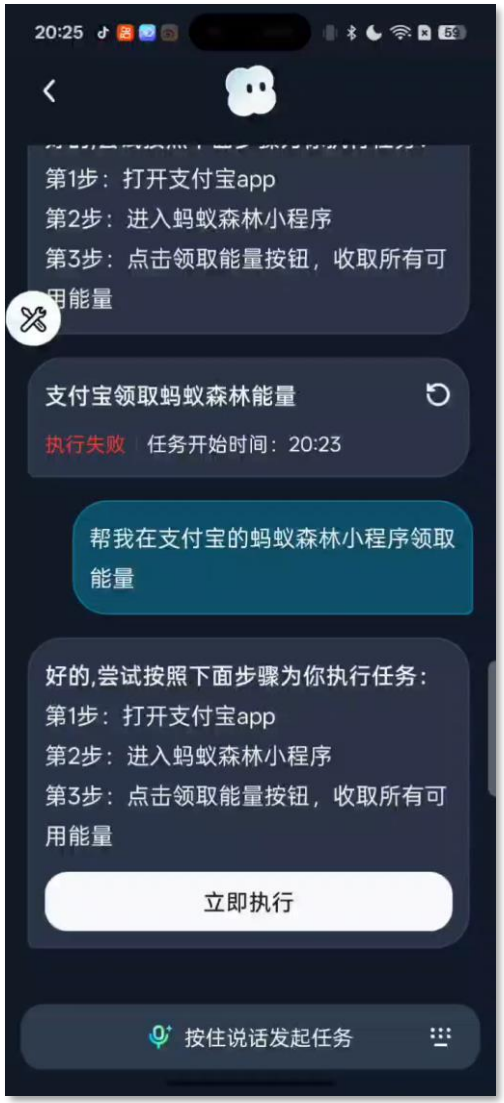
Our approach is used in the StepFun Agent Platform to complete daily tasks such as information retrieval on multiple types of mobile devices

Search for information before buying a car

Search for travel guides in a large number of documents

2 Demos

Help your child find extracurricular tutoring exercises



Complete the Ant Forest Daily Watering Task

Conclusion

1. We introduces UIPro, a generalist GUI agent with superior GUI interaction capability.
2. By curating extensive multi-platform and multi-task GUI interaction data and the proposed unified action space, UIPro exhibits superior performance on multiple GUI interaction and grounding benchmarks.
3. We hope our curation programs and the cleaned dataset will facilitate further research and development in the GUI agent domain.

Repo: https://github.com/ZJULiHongxin/UIPro