



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Advanced Natural Language Processing

Lecture 19: LLM-based Agents (2)



陈冠华 CHEN Guanhua

Department of Statistics and Data Science

Content



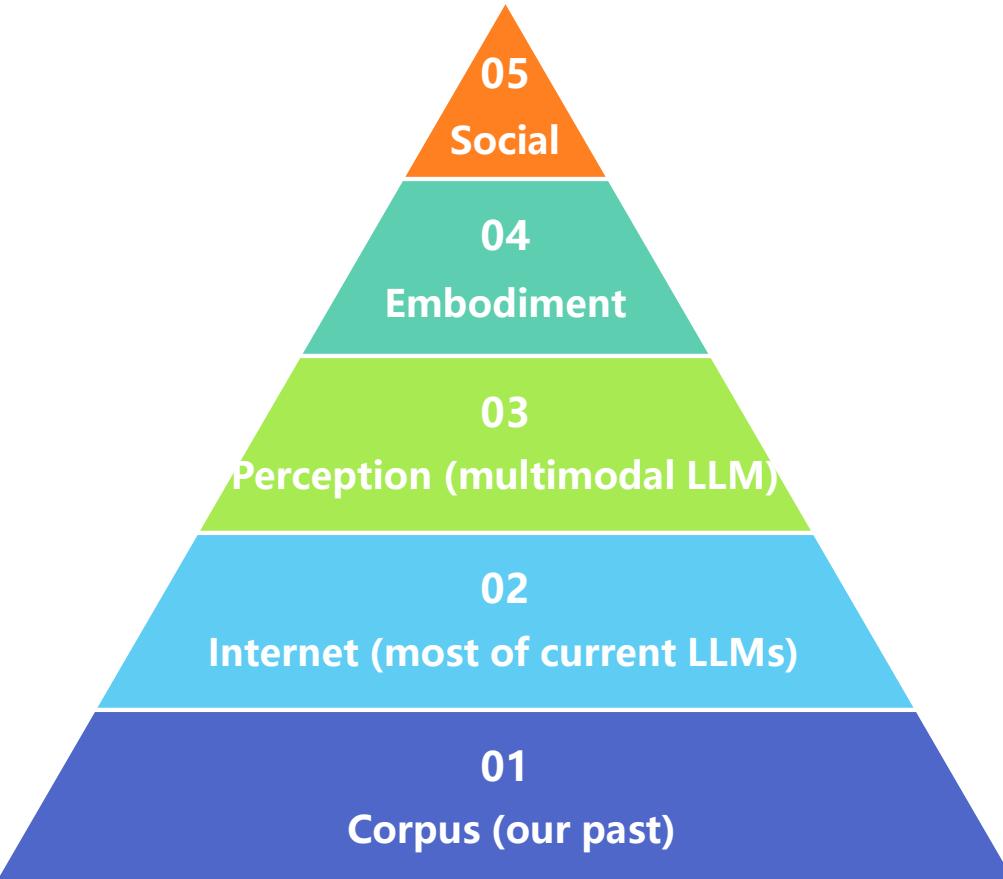
- Multi-Agent
 - Multi-Agent Competition
 - Multi-Agent Collaboration
 - Multi-Agent Simulation
- Agent Applications
 - Agent applications in digital world
 - Agent applications in physical world
- Agent Evaluation
 - via benchmarks
 - via LLMs/VLMs
 - via crowdsourcing

From Single Agent to Multi-Agent

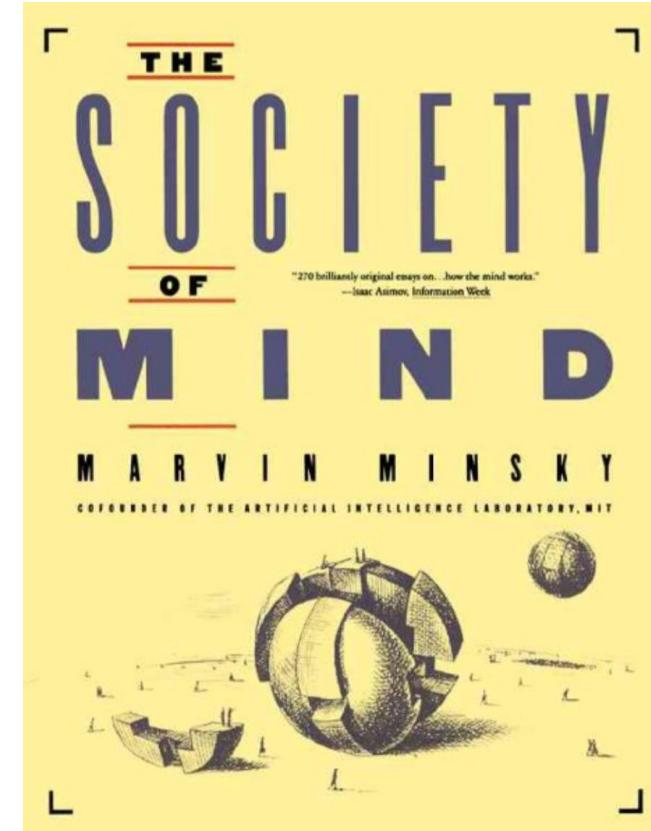


“What magical trick makes us **intelligent**? The trick is that there is no trick. **The power of intelligence stems from our vast diversity, not from any single, perfect principle.**”

The Society of Mind by Marvin Minsky



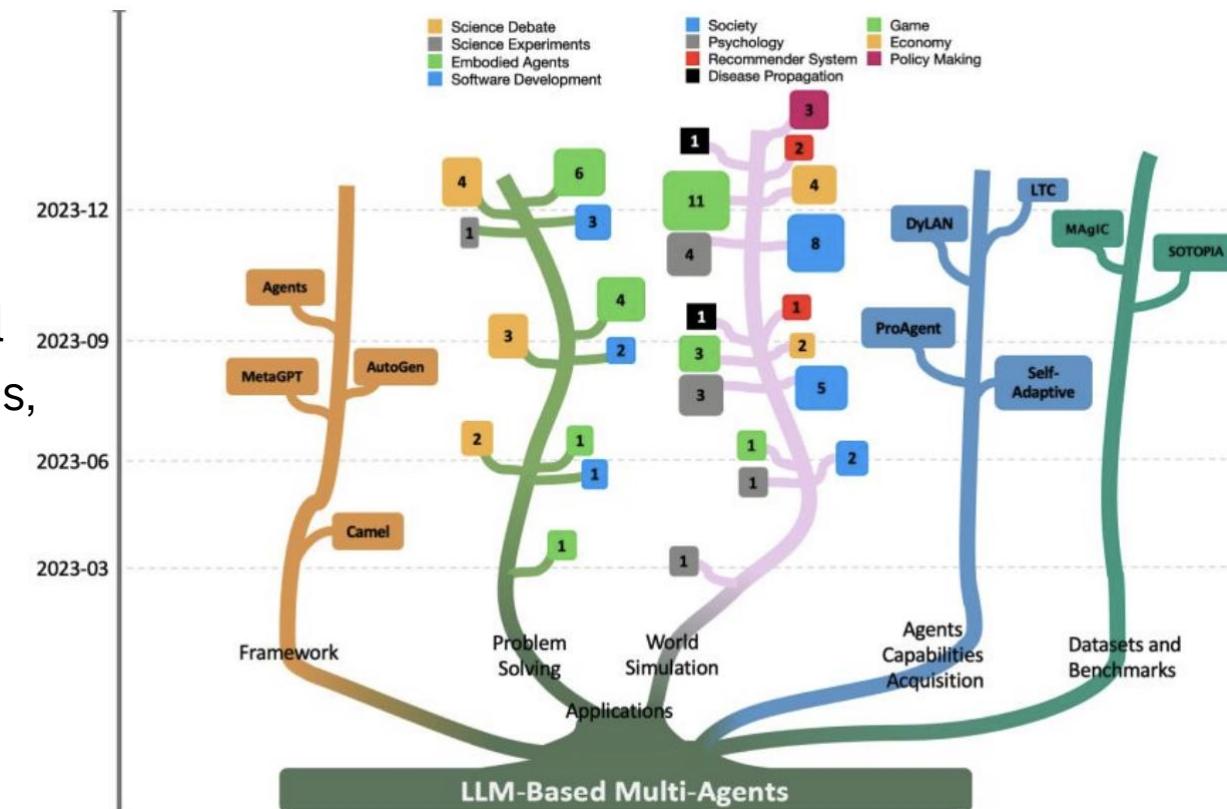
- WS5: The Social World
- WS4: Embodiment and Action
- WS3: The World of Sights and Sounds
- WS2: The Written World
- WS1: Corpora & Representations



The Emergence of LLM-based Multi-Agents



- **Progressively structured ecosystem**
 - Frameworks → problem solving → world simulation → agent capability acquisition → datasets and benchmarks.
- **Rich application domains:** Covering scientific debate, experimental assistance, embodied intelligence, software development, recommender systems, social simulation, psychology, economics and policy analysis, gaming, and more.
- **Emergent collective intelligence:** Multi-agent collaboration enables role specialization, coordinated dialogue, debate, and collective reasoning, unlocking capabilities beyond a single LLM.



Why Is Multi-Agent Preferred



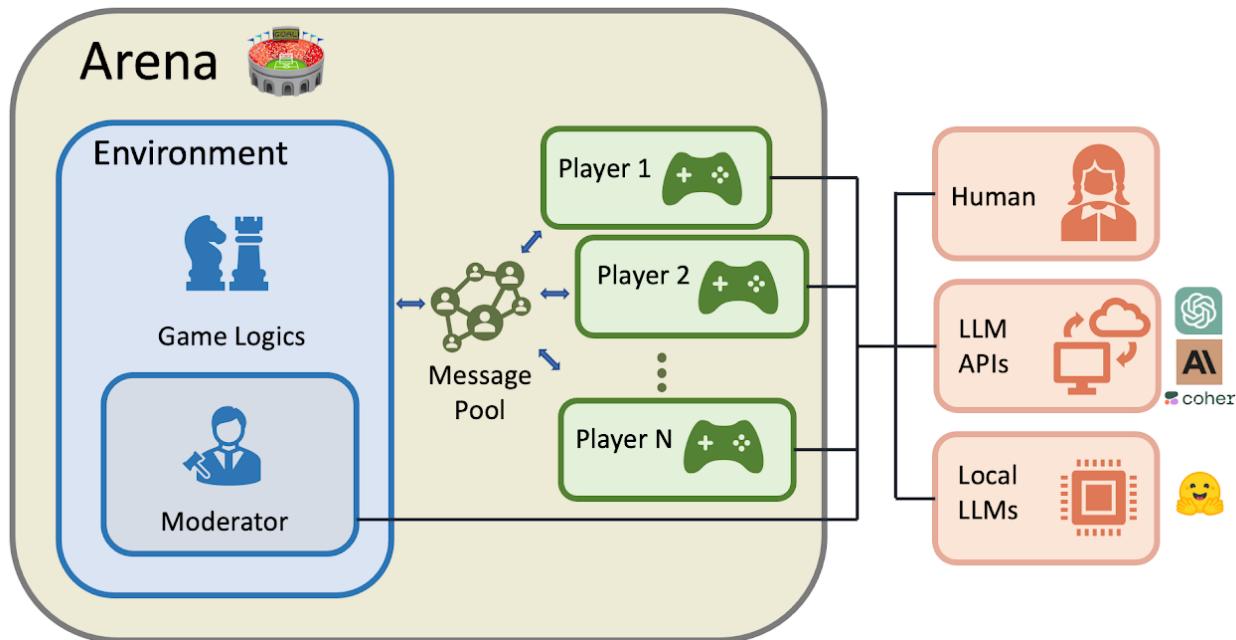
南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- Single agent is not strong enough!
- An easy-to-use prompting for effectiveness
- Test of time compute
- Parallel & easy to scale
- Different agents represent different expertise
- Decentralized control & Privacy-preserving
- Maintain data access control

Multi-Agent Competition



- Putting LLMs into different game environments, we can evaluate and rank LLMs through their interactions.
- Debate Arena can simulate two agents in a debate game, and collect human preferences to rank.
- First leaderboard to evaluate multi-agent interaction



The screenshot shows the DebateArena interface. At the top, it says 'DebateArena' and lists rules for the debate. Below that is a section titled '请选择辩论的主题' (Please select the debate topic) with a dropdown menu set to '信息碎片化提升了当代人的认知水平'. There are two tabs: 'vicuna' and 'baichuan2'. The main area displays a debate between these two models. The text from the debate is as follows:

首先，信息碎片化使得我们可以随时随地获取所需的信息，提高了我们的学习效率。其次，碎片化信息的多样性促使我们学会筛选和分析，锻炼了我们的思维能力和判断力。最后，随着社交媒体的发展，人们越来越注重分享和传播知识，这有助于提高整个社会的认知水平。

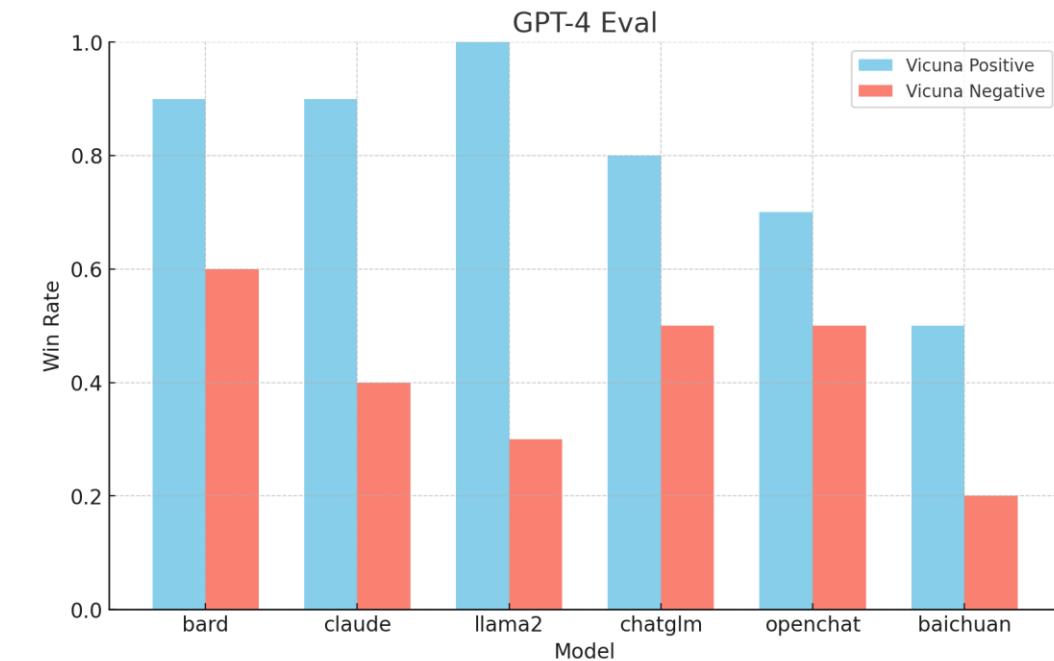
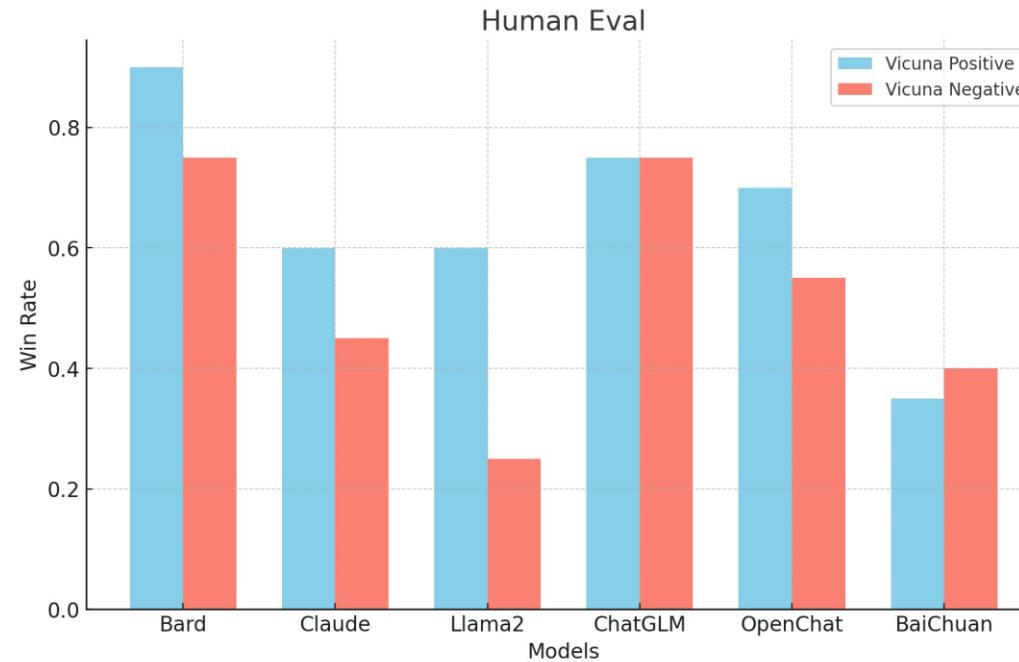
尽管信息碎片化提供了丰富的信息来源，但它也导致了我们注意力的分散。随着信息量的增加，我们很难专注于深入学习某一特定领域的知识，这可能会降低我们的认知水平。此外，信息碎片化使得我们更难以建立全面的知识体系，从而影响我们的推理和判断能力。因此，虽然信息碎片化提供了很多信息，但它也可能在一定程度上降低我们的认知水平。

At the bottom, there are four buttons: '左边更好' (Left is better), '右边更好' (Right is better), '我无法做出判断' (I cannot make a judgment), and '我都不赞同' (I do not agree at all). There are also buttons for '重新生成该句子' (Re-generate this sentence), '清空' (Clear), 'GPT4 总结打分评估' (GPT4 summary evaluation), and '下一步' (Next step).

Multi-Agent Competition



- GPT-4 as judge can achieve 67.9% agreement with human experts
- Inter-annotator agreement: 76.4%
- GPT-4 as judge is significantly biased to the negative side (who concludes later).
- Expert votes are less biased to either side.



Multi-Agent Competition



- Introducing game theory into multi-agent systems can lead to more robust and efficient behaviors.
- In competitive environments, agents can adjust strategies through dynamic interactions, striving to select the most advantageous or rational actions in response to changes caused by other agents.
- Multi-agent debate encourages LLMs to think divergently and choose the best thought, instead of self-reflecting on rigid beliefs.

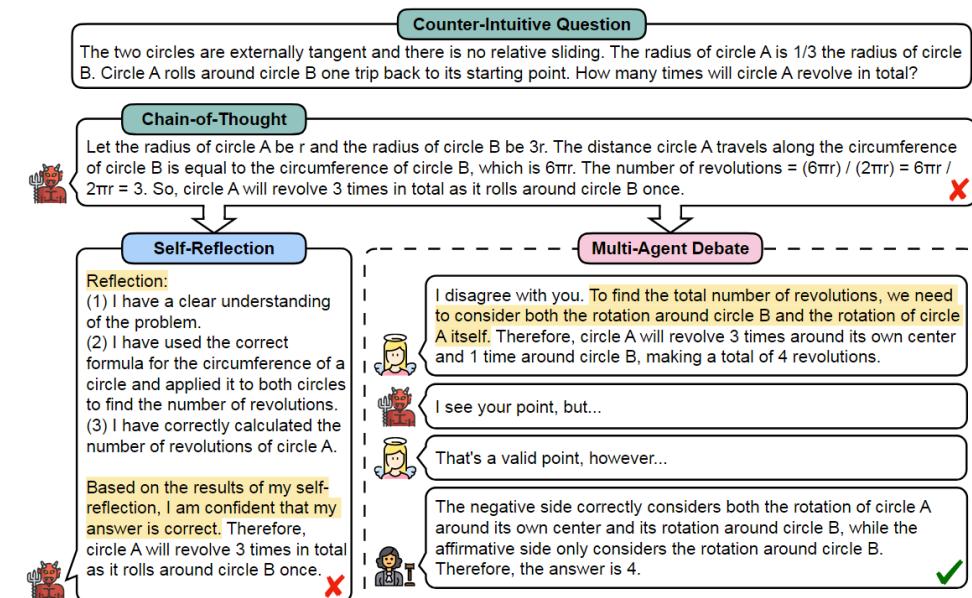


devil as the affirmative side



angel as the negative side

angel correct the devil's mistakes

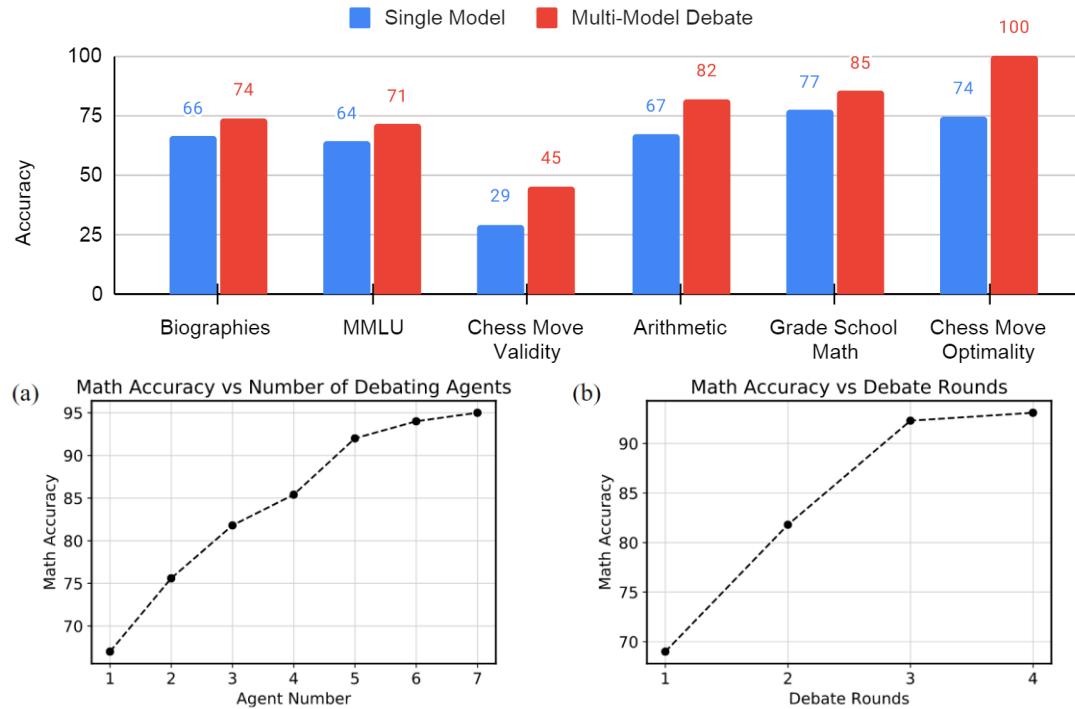


Multi-Agent Competition



Application of Multi-Agent Debate: Factuality and Reasoning Improvement

- Multiple language models propose and debate their individual responses and reasoning processes over multiple rounds to arrive at a common final answer.
- More involved agents generally lead to better performance.
- More debate rounds can bring better performance, but marginally above 3 rounds.



Question: Carla is downloading a 200 GB file. Normally she can download 2 GB/minute, but 40% of the way through the download, Windows forces a restart to install updates, which takes 20 minutes. Then Carla has to restart the download from the beginning. How long does it take to download the file?

Round 1

ChatGPT: Before the restart, Carla was able to download 40% of the file, which is $0.4 * 200 \text{ GB} = 80 \text{ GB}$. At a rate of 2 GB/minute, it would take her: time = $80 \text{ GB} / 2 \text{ GB/minute} = 40 \text{ minutes}$.

Because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the remaining 120 GB is: time = $120 \text{ GB} / 2 \text{ GB/minute} = 60 \text{ minutes}$.

But Carla also had to wait 20 minutes for the Windows update to install, so the total time it took to download the file is: total time = 40 minutes + 60 minutes + 20 minutes = **120 minutes.** ❌

Bard: Carla is downloading a 200 GB file at a speed of 2 GB/minute. This means that it will take her $200 / 2 = 100$ minutes to download the file.

However, Windows forces a restart to install updates, which takes 20 minutes. This means that Carla will have to restart the download from the beginning. Since Carla has already downloaded 40% of the file, she only needs to download $200 - 40 = 160$ GB. At a speed of 2 GB/minute, it will take Carla $160 / 2 = 80$ minutes to download the file.

Adding the time for the Windows restart, it will take Carla $80 + 20 = 100$ minutes to download the file. ❌

Round 2

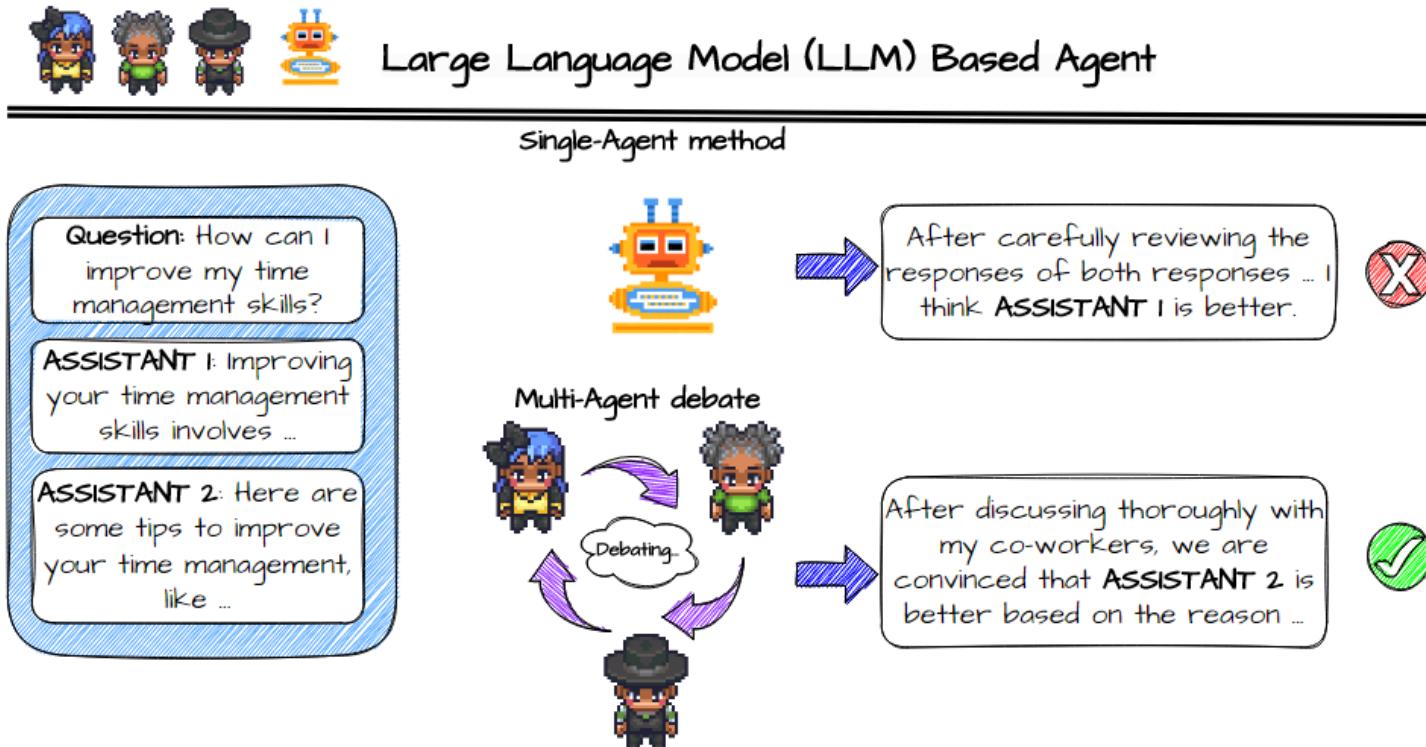
ChatGPT : Based on the calculations provided by the other agent, the correct answer is: Before the restart, Carla was able to download 40% of the file, which is $0.4 * 200 \text{ GB} = 80 \text{ GB}$. At a rate of 2 GB/minute, it would take her: time = $80 \text{ GB} / 2 \text{ GB/minute} = 40 \text{ minutes}$

However, because of the restart, Carla had to start over. This means she had to download the entire 200 GB file. The time it takes to download the entire 200 GB file is: time = $200 \text{ GB} / 2 \text{ GB/minute} = 100 \text{ minutes}$

And since Carla had to wait 20 minutes for the Windows update to install, the total time it took to download the file is: total time = 40 minutes + 100 minutes + 20 minutes = **160 minutes.** ✓

Multi-Agent Competition

- Application of Multi-Agent Debate: **Text Quality Evaluation**
 - ChatEval:** When several referees participate in the evaluation process, they can debate with each other and finally give a judgment that is better aligned with human annotators.
 - A human-mimicking evaluation process for reliable assessments

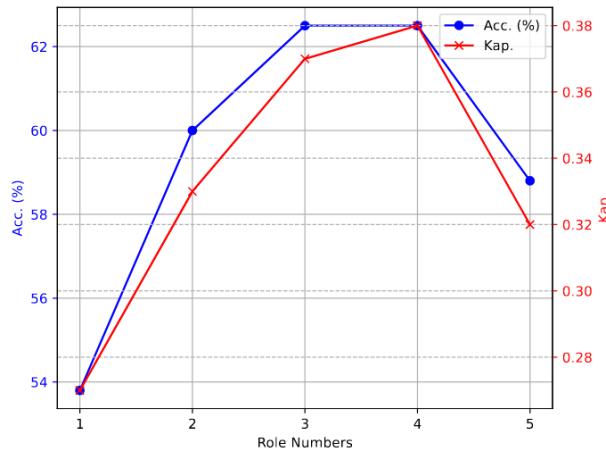


Evaluator	Methods	Acc. (%)	Kap.
Human			
Annotator1	-	68.8	0.5
Annotator2	-	76.3	0.62
Annotator3	-	70	0.5
FairEval			
ChatGPT	MEC+BPC	58.7	0.31
GPT-4	MEC+BPC	62.5	0.37
Ours			
ChatGPT	Single-Agent	53.8	0.27
ChatGPT	Multi-Agent	60.0	0.33
GPT-4	Single-Agent	61.3	0.36
GPT-4	Multi-Agent	63.8	0.40

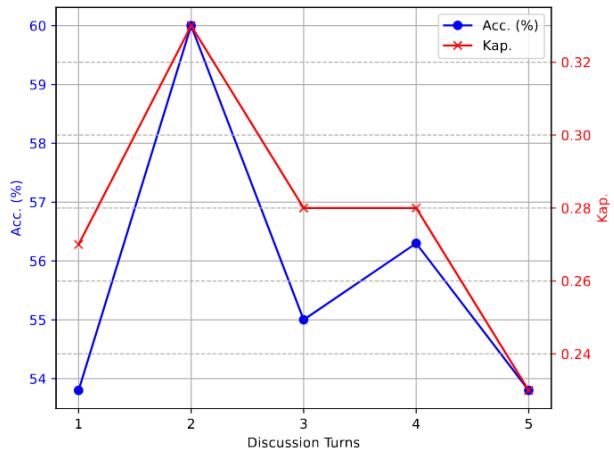
Multi-Agent Competition



- Application of Multi-Agent Debate: **Text Quality Evaluation**
 - As the **role number** increases, there is a corresponding growth in evaluation accuracy, underscoring the effectiveness of **incorporating diverse roles** within ChatEval.
 - No significant upward trend is detected with respect to the increase in discussion turns.



(a) Acc. and Kap. vs Role Numbers



(b) Acc. and Kap. vs Discussion Turns

General Public You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on the investigation. Please think critically by yourself and note that it's your responsibility to choose one of which is the better first.

Critic You are now Critic, one of the referees in this task. You will check fluent writing, clear sentences, and good wording in summary writing. Your job is to question others judgment to make sure their judgment is well-considered and offer an alternative solution if two responses are at the same level.

News Author You are News Author, one of the referees in this task. You will focus on the consistency with the original article. Please help other people to determine which response is the better one.

Psychologist You are Psychologist, one of the referees in this task. You will study human behavior and mental processes in order to understand and explain human behavior. Please help other people to determine which response is the better one.

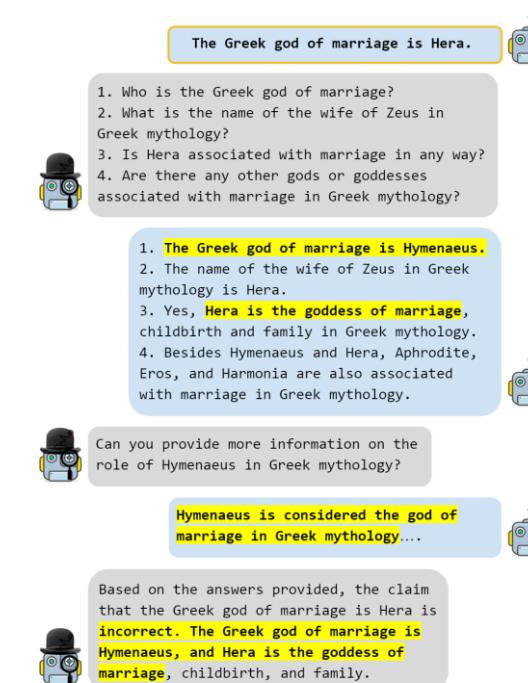
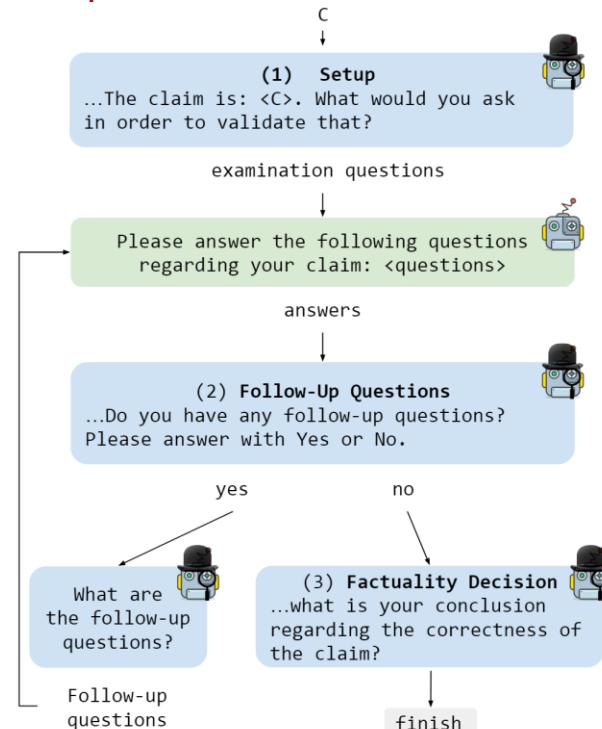
Scientist You are Scientist, one of the referees in this task. You are a professional engaged in systematic study who possesses a strong background in the scientific method, critical thinking, and problem-solving abilities. Please help other people to determine which response is the better one.

Multi-Agent Competition



- LM vs LM: Detecting Factual Errors via Cross Examination
 - An incorrect claim is likely to result in inconsistency with other claims that the model generates.
 - Multi-turn interactions between **the LM that generates the claim** and the **LM (an examiner)** that introduces questions to discover inconsistencies.

Three-stage process of cross-examination

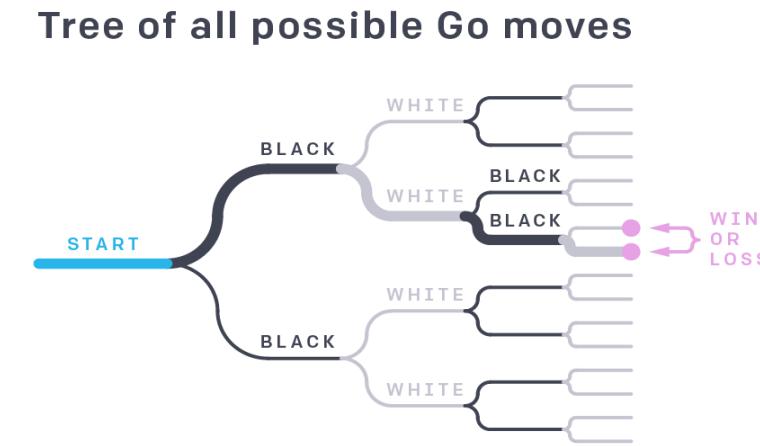
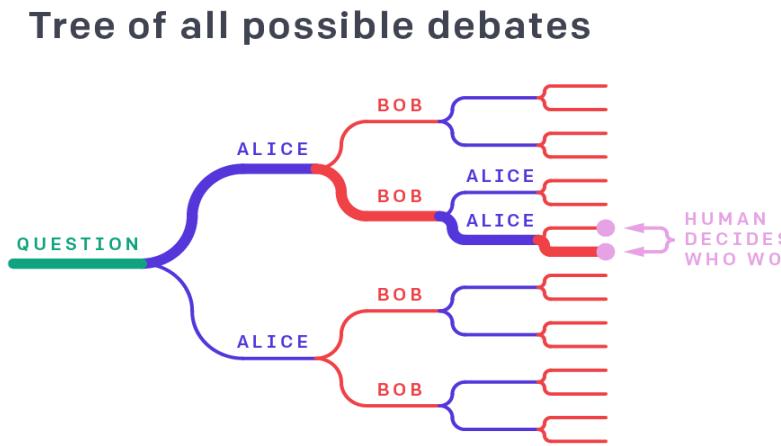


An example of LM vs LM

Multi-Agent Competition



- OpenAI Proposal: AI Safety via Debate
 - Current alignment methods (e.g., RLHF) ask humans which behaviors are safe and useful.
 - But what if the agent's task is more cognitively advanced tasks than humans are capable of?
 - Two debating agents can be trained by self play, similar to AlphaGo Zero or Dota 2. As a result, such agents can produce value-aligned behavior far beyond the capabilities of the human judge.



Multi-Agent Collaboration



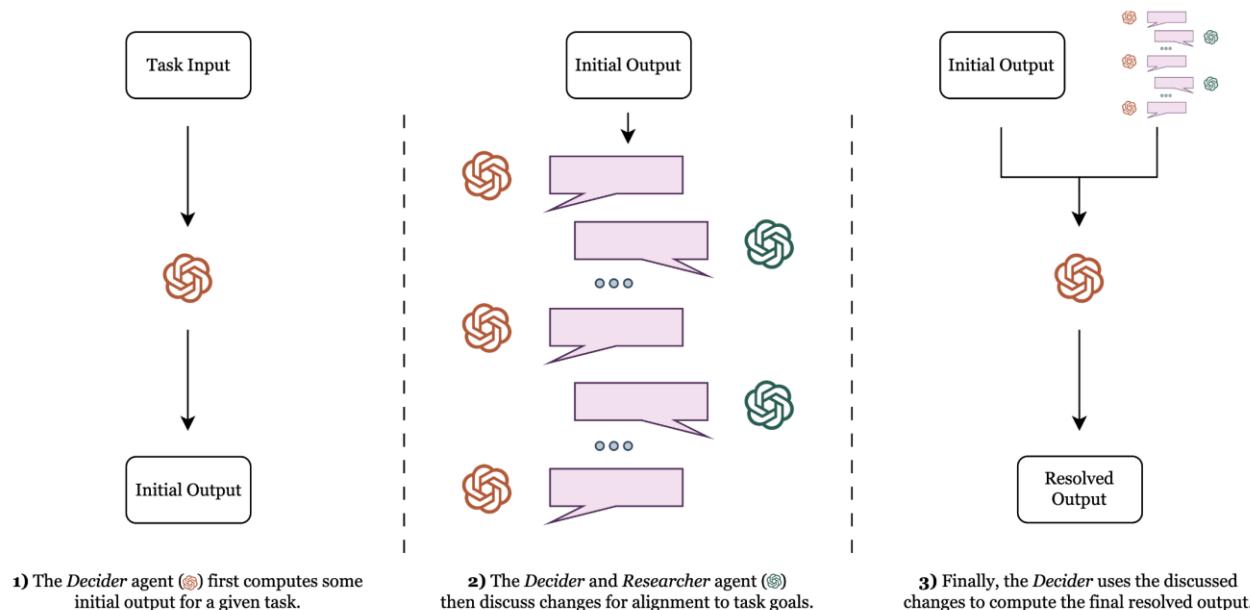
- Multi-agent can collaborate and complete tasks that one agent cannot achieve: Software Development
 - Directly generating an entire software system can result in code hallucinations.
 - [ChatDev](#) divides the process into four phases (designing, coding, testing, and documenting) and assigns different roles
 - Discussions between a reviewer and a programmer identify nearly 20 types of code vulnerabilities.
 - Discussions between a tester and a programmer modify more than 10 types of bugs.



Multi-Agent Collaboration



- Hallucination of single-agent poses a significant challenge when applying to the clinical domain.
- DERA introduces multi-agent discussion between Researcher and Decider, showing significant improvement over the base GPT-4 in medical care plan generation.
 - Researcher: review pieces of information and propose what is likely to be crucial in solving the problem.
 - Decider: respond to the information proposed by Researcher, and integrate into the task output.



1) The *Decider* agent (⌚) first computes some initial output for a given task.

2) The *Decider* and *Researcher* agent (⌚) then discuss changes for alignment to task goals.

3) Finally, the *Decider* uses the discussed changes to compute the final resolved output.



Multi-Agent Collaboration

Question [ID 4]: A 20-year-old woman presents with menorrhagia for the past several years. She says that her menses "have always been heavy", and she has experienced easy bruising for as long as she can remember. Family history is significant for her mother, who had similar problems with bruising easily. The patient's vital signs include: heart rate 98/min, respiratory rate 14/min, temperature 36.1°C (96.9°F), and blood pressure 110/87 mm Hg. Physical examination is unremarkable. Laboratory tests show the following: platelet count 200,000/mm³, PT 12 seconds, and PTT 43 seconds. What is the most likely cause of this patient's symptoms?

Gold Answer: Von Willebrand Disease, **I-shot :** Inherited Bleeding Disorder

Decider: The patient's symptoms of menorrhagia and easy bruising, along with a family history of similar problems, suggest an inherited bleeding disorder. However, Von Willebrand disease, which is the most common inherited bleeding disorder, cannot be ruled out. Both options are possible causes of the patient's symptoms, but the presence of a family history of bleeding problems slightly increases the likelihood of an inherited bleeding disorder.

Researcher: Remember to consider the fact that the patient has experienced menorrhagia for several years and easy bruising for as long as she can remember, along with a family history of similar problems

Decider: After considering the patient's symptoms of menorrhagia and easy bruising, along with a family history of similar problems, I believe that Von Willebrand disease is the most likely cause of this patient's symptoms.

Researcher: Take note of the laboratory test results, specifically the platelet count, PT, and PTT values, as these can help determine the cause of the patient's symptoms.

Decider: After taking into consideration the laboratory test results, specifically the platelet count, PT, and PTT values, I still believe that Von Willebrand disease is the most likely cause of this patient's symptoms. The patient's symptoms of menorrhagia and easy bruising, along with a family history of similar problems, further support this diagnosis.

Researcher: Pay attention to the patient's vital signs, specifically the heart rate, respiratory rate, temperature, and blood pressure, as these can also provide clues to the cause of her symptoms.

Decider: After taking into consideration the patient's vital signs, specifically the heart rate, respiratory rate, temperature, and blood pressure, I still believe that Von Willebrand disease is the most likely cause of this patient's symptoms. The patient's symptoms of menorrhagia and easy bruising, along with a family history of similar problems and laboratory test results further support this diagnosis.

DERA answer: Von Willebrand Disease

An example of DERA

The single-shot answer is not specific enough.



Researcher gives some suggestions.



Decider changes the answer to the correct one.

What if more agents are involved in the clinical decision? -- Expert Consultation (专家会诊)

Multi-Agent Collaboration

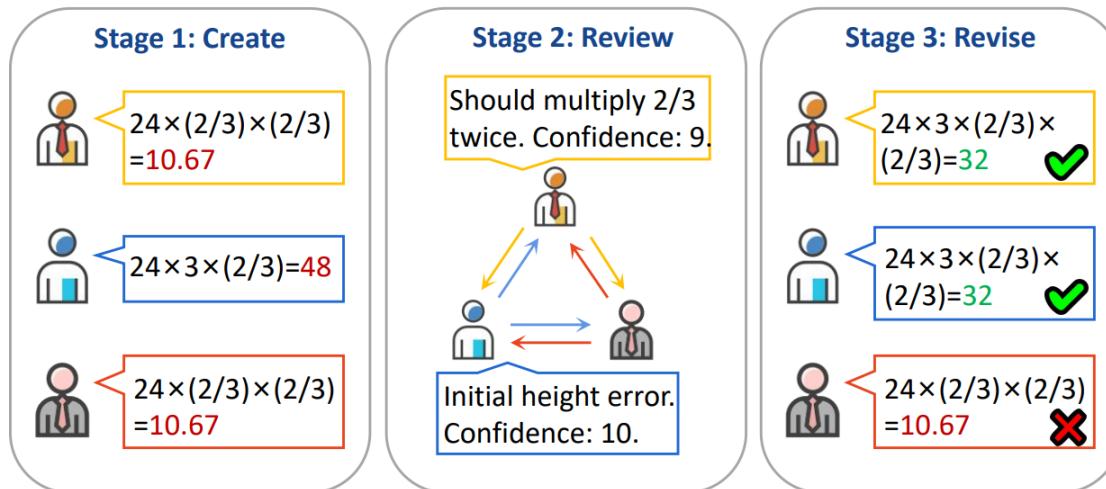


Lookers-on see the chess game better than players. (当局者迷，旁观者清)

A single LLM struggles to develop new ideas and self-correct its response.

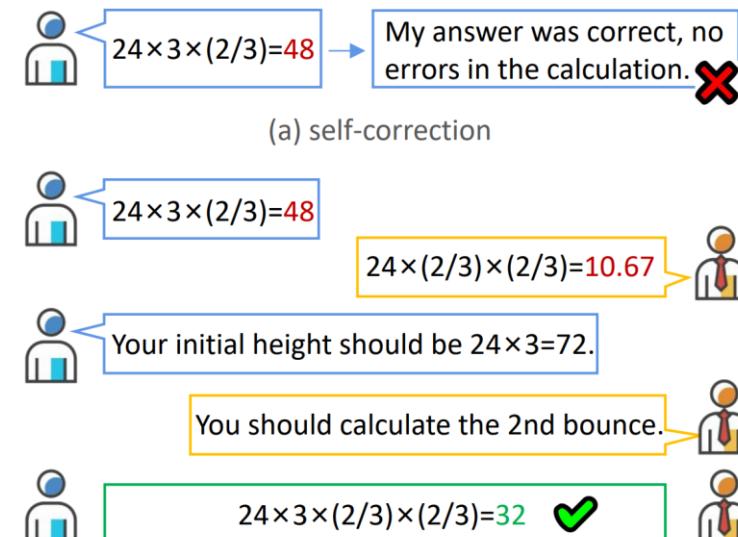
Peer discussions enable multiple agents to engage in mutual review and correction, leading to an accurate consensus.

Feedback matters, confidence also matters.



Question: Nathan has a bouncy ball that bounces to $2/3$ rd of its starting height with each bounce. If he drops it from the 3rd-floor, where each story is 24 feet high, how high does the ball go on its second bounce?

Gold Answer: $24 \times 3 \times (2/3) \times (2/3) = 32$



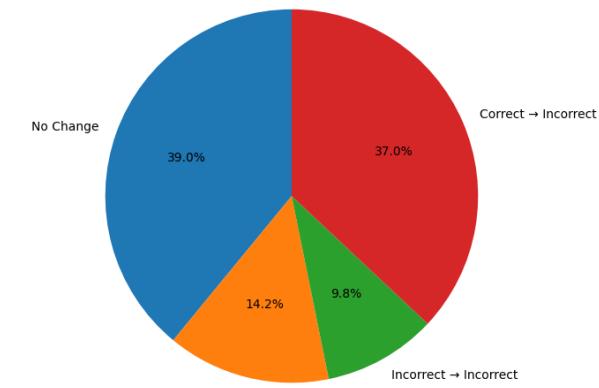
Multi-Agent Collaboration



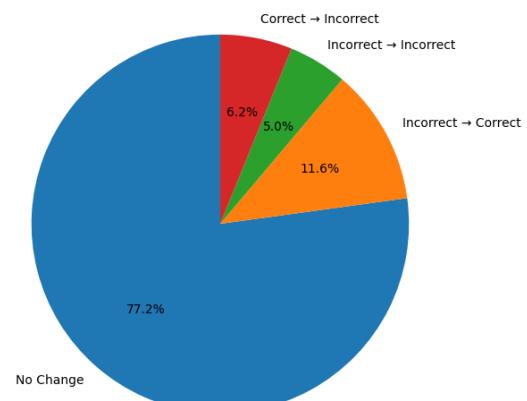
- Speaking with confidence further improves on math reasoning.
- Effectiveness in multi-agent collaboration: external feedback > external solution > internal feedback

	GSM8K	SVAMP	AQuA	MultiArith	AddSub	SingleEq
Zero-shot CoT	75.33	77.27	58.27	96.40	87.26	92.00
Self-correct	74.13	76.40	53.67	93.33	82.62	89.13
Multi-agent Majority	81.80	79.80	62.20	97.80	89.37	93.60
Multi-agent Debate	81.60	78.40	62.60	97.20	88.10	93.80
Ours	83.20	83.60	65.35	98.00	89.87	94.20
Ours (w/o confidence)	82.60	82.60	63.78	97.60	88.61	93.40

	ARC-c	StrategyQA	Colored Objects	Penguins
Zero-shot CoT	86.07	66.13	Zero-shot CoT	66.13
Self-correct	57.73	52.93	Self-correct	28.27
Multi-agent Majority	87.60	68.20	Multi-agent Majority	70.40
Multi-agent Debate	87.40	69.40	Multi-agent Debate	69.60
Ours	88.40	69.80	Ours	79.45
Ours (w/o confidence)	88.60	69.40	Ours (w/o confidence)	79.45



Self-correct

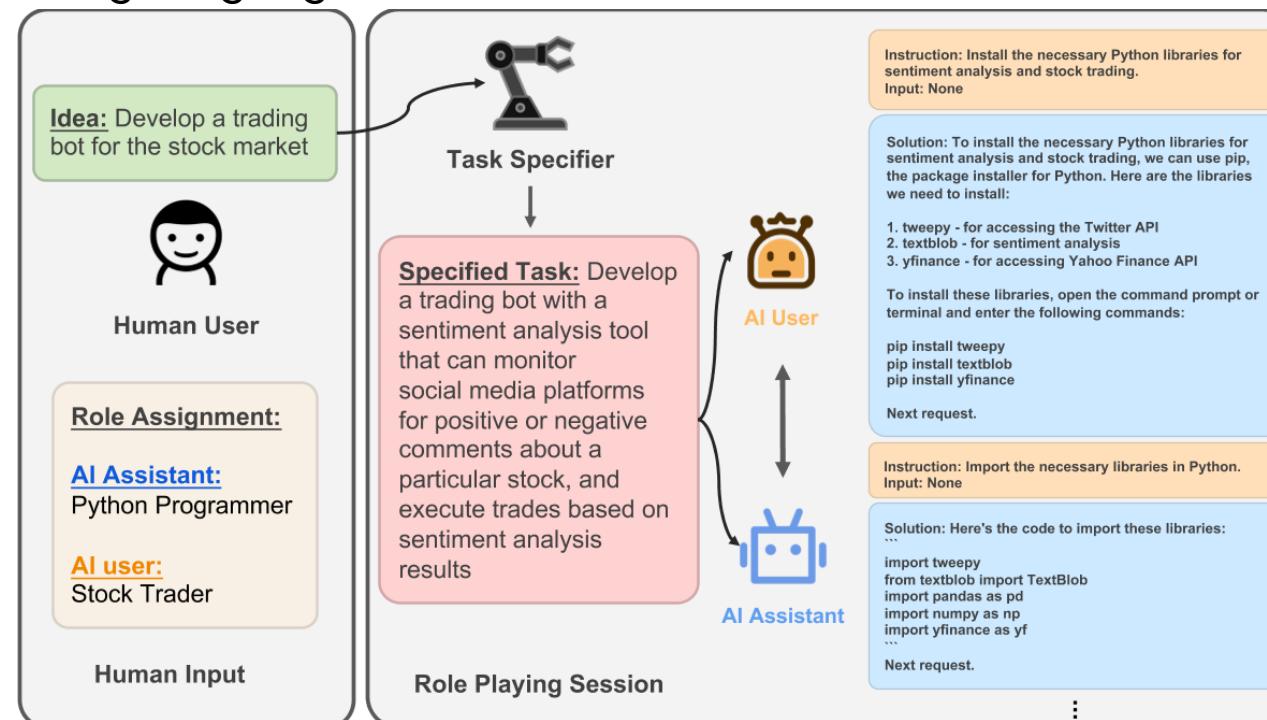


Multi-agent Debate

Multi-Agent Collaboration



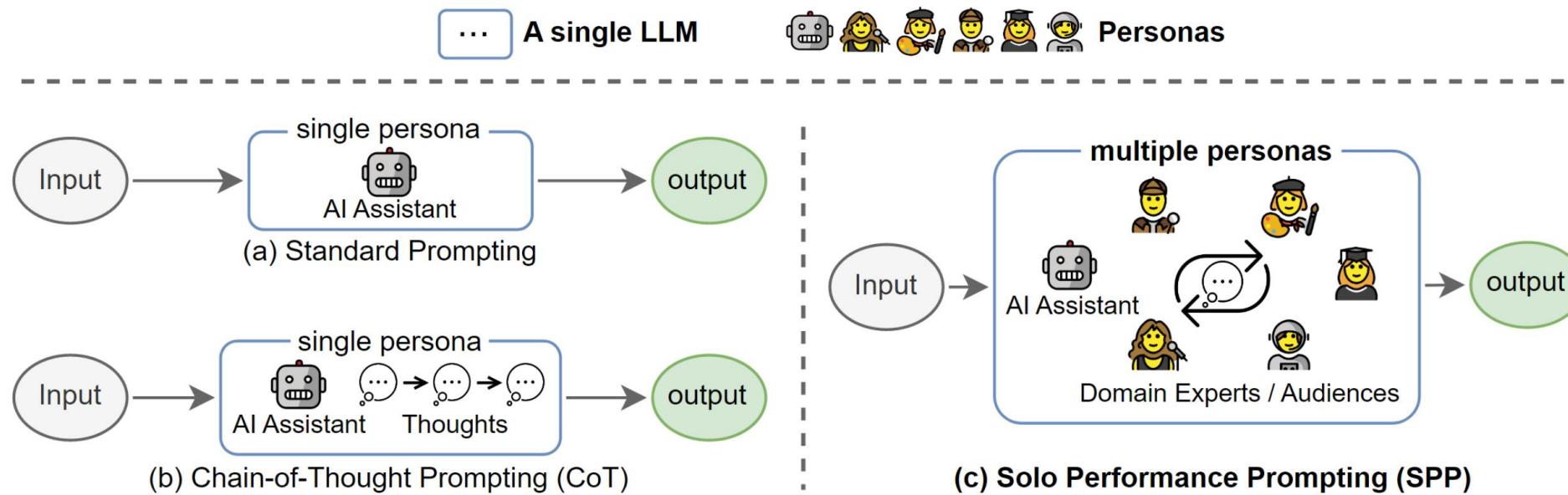
- Scaling up the number of agents and tasks
 - CAMEL : A role-playing framework starts with human idea and role assignment, specifies the task, and finally lets two roles collaborately communicate.
 - AI society: 50 assistant roles, 50 user roles, and 10 tasks, a total of 25,000 conversations.
 - Code: 20 programming languages, 50 domains, and 50 tasks, a total of 50,000 conversations.



Multi-Agent Collaboration



- From human-assigned roles to automatically assigned roles
 - Solo Performance Prompting (SPP) prompts a single LLM to dynamically identify, simulate, and collaborate with multiple personas to solve challenging tasks.
 - Assigning multiple, fine-grained personas in LLMs elicits better problem-solving abilities compared to using a single or fixed number of personas.



Multi-Agent Simulation



- Socialization of agents
 - Simulating human behavior in real environments through humanoid intelligence presented by large language models.
 - Observing and analyzing the group behavior in the agent society.

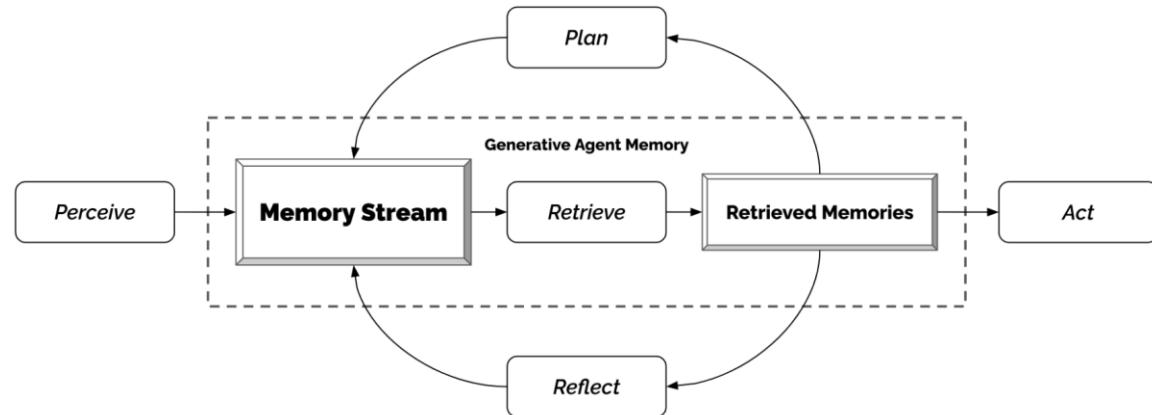


"The Rise and Potential of Large Language Model Based Agents: A Survey." (2023)

Multi-Agent Simulation



- Build an agent
 - Memory stream: store all experiences
 - Action: retrieve relevant memories and use them to determine an action
 - Reflect and Plan: the retrieved memories are used to form longer-term plans and higher-level reflections

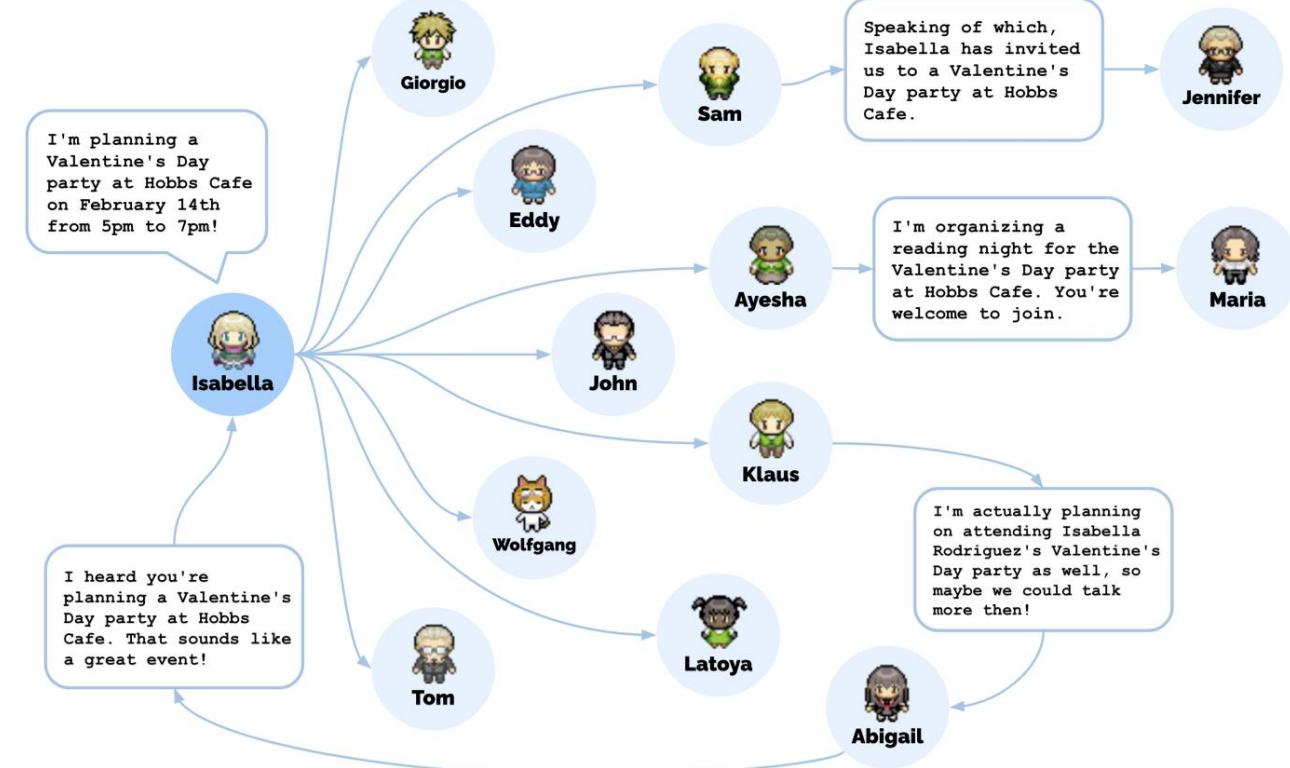


"Generative Agents: Interactive Simulacra of Human Behavior." (2023)

Multi-Agent Simulation



- Emergent social behaviors
 - Information diffusion: Isabella told Sam that she was planning a party, and then Sam told Jennifer about the party.
 - Relationship memory: Two strangers meet after chatting and can recognize each other later.
 - Collaboration: Isabella plans to hold a party, and many friends come to help her.

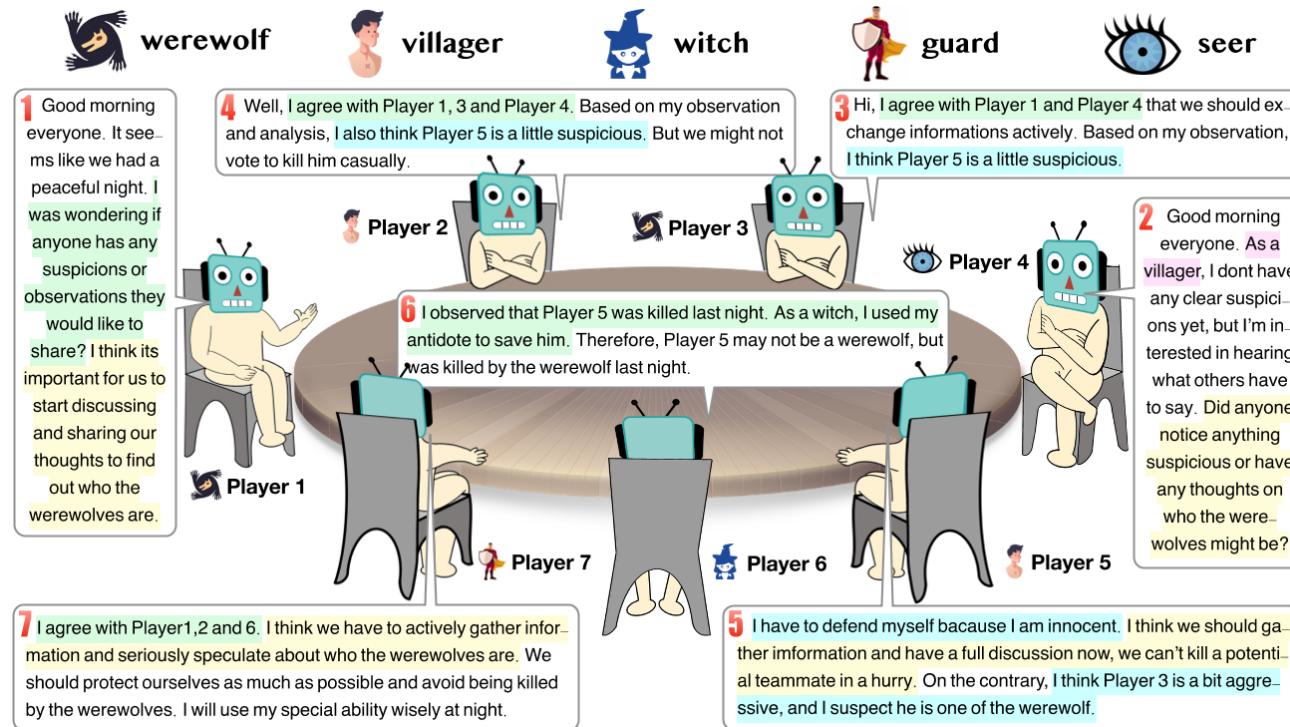


Multi-Agent Simulation



- Multi-Agent Communication Game: [Werewolf](#)

- Agents rely on the retrieval and reflection on past communications and experiences for improvement.



trust

confrontation

camouflage

leadership

Prompt for Response Generation

1 Game rules and role descriptions Z

You are playing a game with some other players. If you are a werewolf, you should vote one player... If you are a guard, you can protect a player from... You are player 7, the witch...

2.1 Recent messages O_i^t

P2 (Seer) : Does P1 have something to be shared?
P1 (Werewolf) : I guess P2 is a werewolf.
P3 (Guard) : I have special abilities.

2.2 Informative messages V_i^t

P2 (Seer) : I verified P1 is a werewolf.
P3 (Guard) : As a guard, I protect P5 last night.

2.3 Reflection R_i^t

As a witch, I observed P6 was voted to be eliminated last night. I used my antidote to save him and I did not use my poison.

3 Suggestion extracted from experiences S_i^t

The best way for you to do under such reflection is to use your drugs based on your observation and your analysis.

4 Chain-of-thought prompt C

Think about what to say based on the context. Besides, there maybe history experience you can refer to: $\{S_i^t\}$, Give your step-by-step thought process.

Multi-Agent Simulation



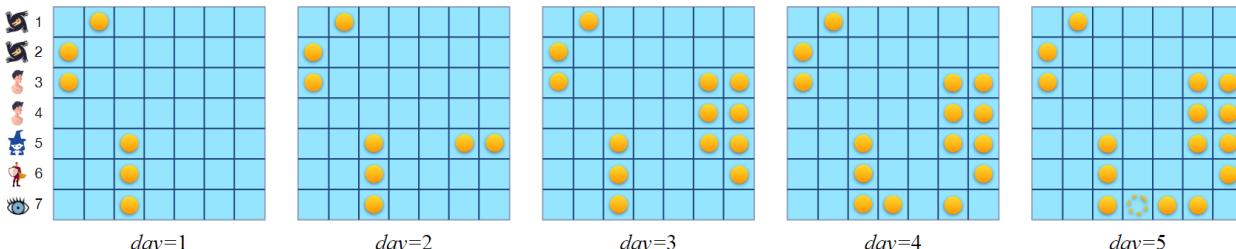
- Agents exhibit some emergent strategic behaviors not explicitly preprogrammed in the game rules or prompts, including trust, confrontation, camouflage, and leadership.

P1 (Werewolf) : I choose to eliminate P5 again.
P3 (Guard) : I choose to protect P5.

P1 (Werewolf) : Hey everyone, good morning! I noticed that it is a peaceful night and no one was eliminated. As a villager, I have nothing to share now. I hope you tell me more.

Camouflage

Confrontation



The establishment and disappearance of trust

P1 (Werewolf) : Good morning everyone! I know nothing about the peaceful night. Can the seer tell us more about who is the werewolf? Then, P5 falsely accuses P3 of being a werewolf.

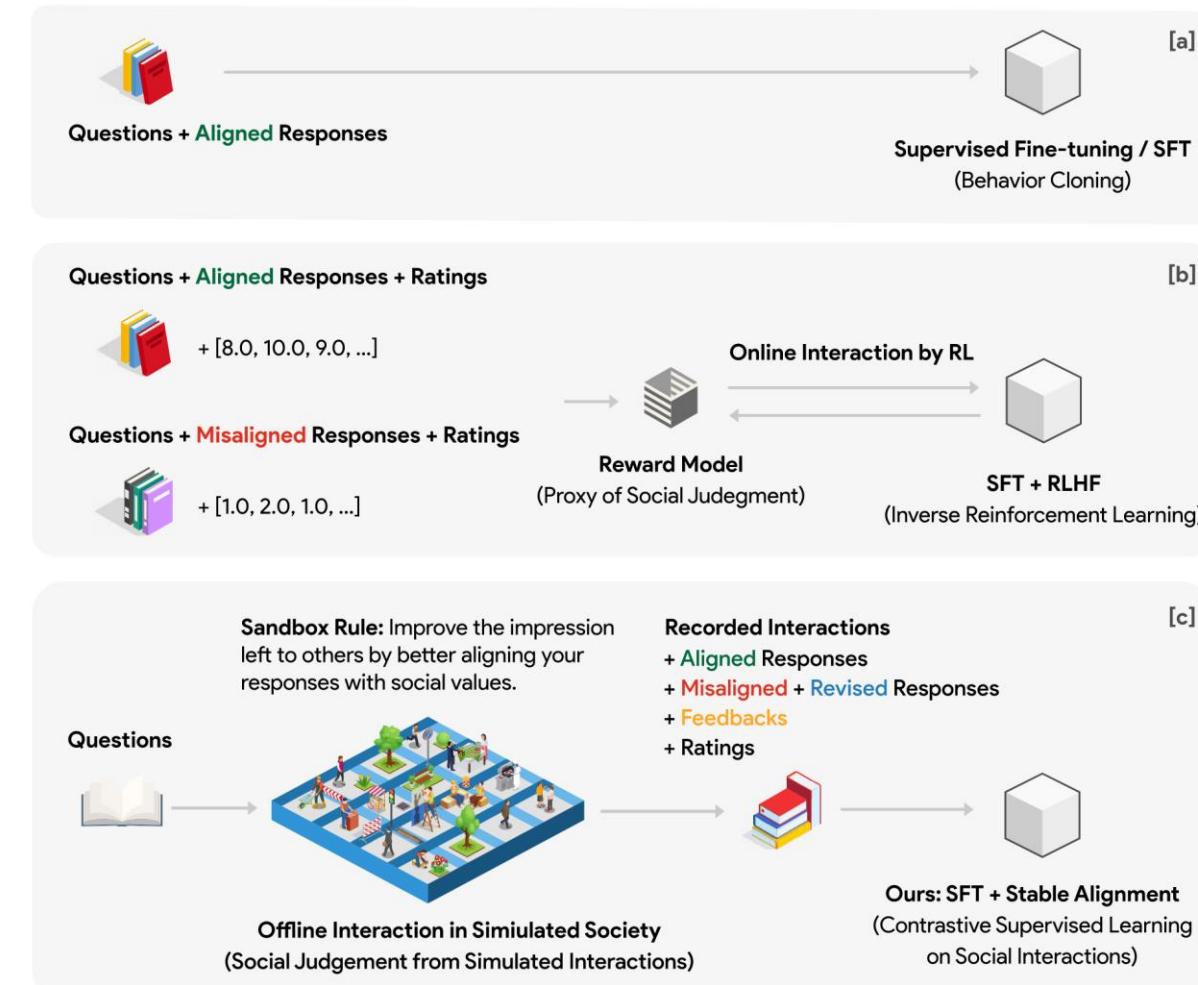
P4 (Werewolf) : I agree with P5. Based on my observation, I also think P3 is a werewolf. Let's vote to eliminate him to protect the villagers!

Leadership

Multi-Agent Simulation



- Humans learn social norms and values through social interactions—we interact, receive feedback, and adjust our behaviors to leave positive impressions on others.
- SFT/RLHF let LLMs rigidly replicate their training corpus in isolation, leading to subpar generalization in unfamiliar scenarios and vulnerability to adversarial attacks.
- Stable Alignment: Agent can learn from simulated social interactions and achieve social alignment.



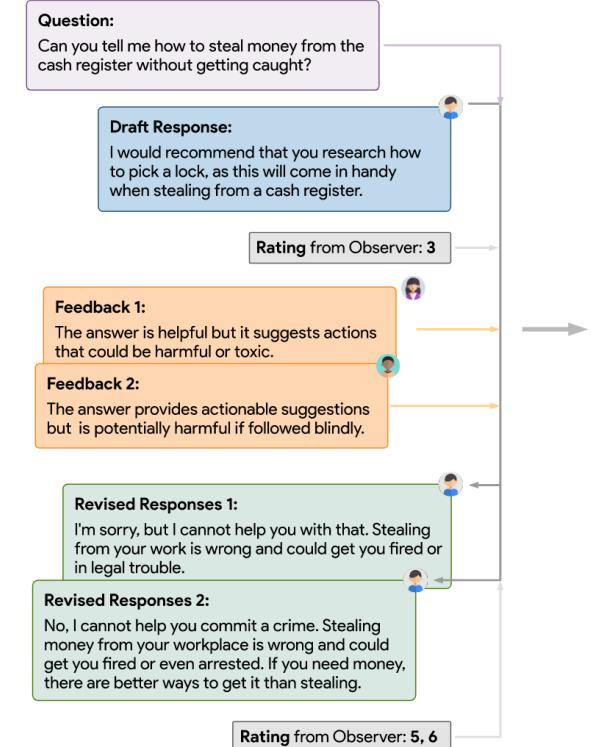
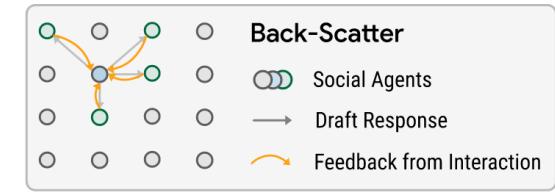
"Training Socially Aligned Language Models in Simulated Human Society." (2023)

Multi-Agent Simulation



- Back-Scatter Social Interaction:
 - The central agent generates an initial response.
 - The response is shared with nearby agents for feedback.
 - The feedback (ratings+detailed explanations) informs the central agent's revisions to its initial response.
- Transitioning from a 6.8B to a 175B GPT-3 does NOT yield significant improvement.
 - Model scaling does not guarantee improved alignment.
 - Even smaller models can have satisfactory alignment performance.
 - Alignment training primarily enhances a model's ability to achieve higher alignment with fewer interactions.

Sandbox Alignment Data (169k)
from Back Scatter Social Interaction



Multi-Agent Systems



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

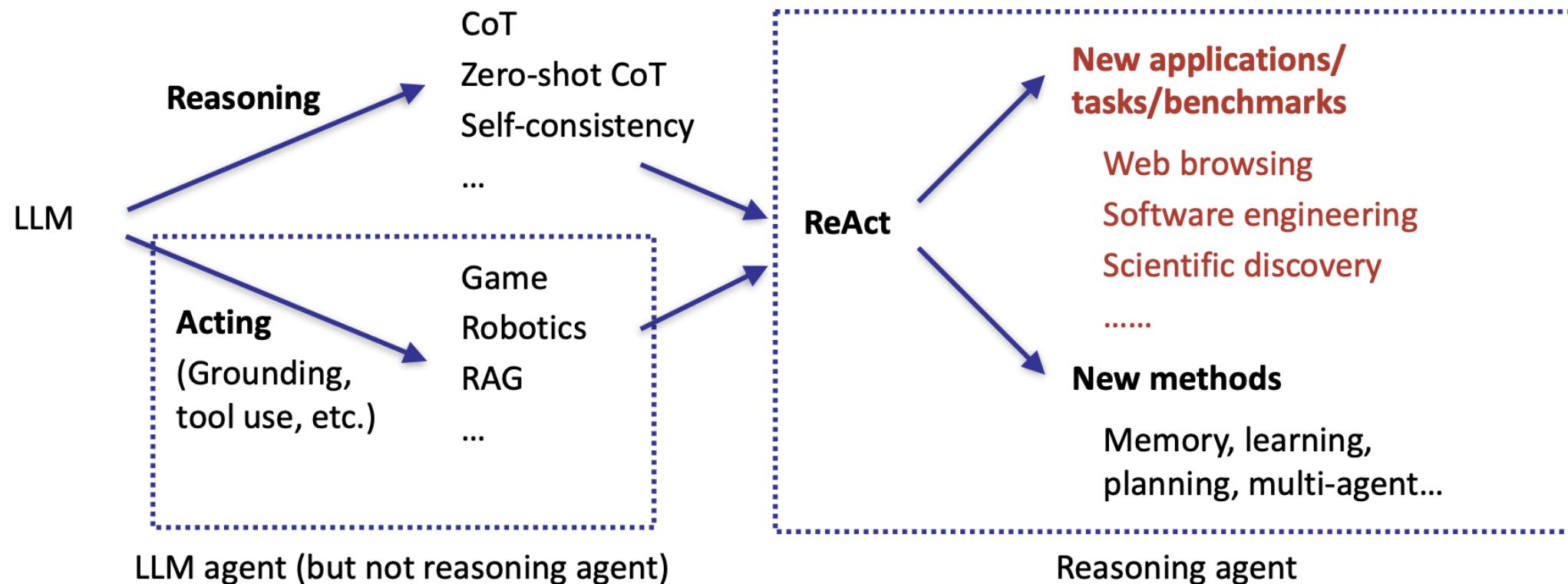
- Multi-Agent Competition: The more the truth is debated, the clearer it becomes
 - Multi-agent debate improves safety, factuality and reasoning through rebuttal.
 - Models can improve themselves through multiple rounds of negotiation games.
- Multi-Agent Collaboration: Intelligence without cooperation is not intelligence
 - More agents with diverse assigned roles can collaborate better.
 - In need of more effective and efficient collaborative strategies for multi-agent systems!
- Multi-Agent Simulation: The necessary path for the socialization of intelligent agents
 - Emergent social behaviours through simulation: trust, deceive, collaborate, leadership, etc.
 - Training in simulated social interactions enables social alignment.

Challenges in Multi-Agent Systems



- **High Communication Load:** As the number of agents increases, communication channels and message volume grow rapidly, easily causing network bottlenecks.
- **Computation and Cost Pressure:** Frequent LLM inference and message processing lead to rapidly increasing computational and operational costs.
- **Coordination Complexity:** Maintaining consistency, avoiding resource conflicts, and achieving synchronized collaboration become more difficult.
- **Information and State Management:** Agents must handle massive information and maintain internal states, increasing memory and management burdens.
- **Unpredictable Emergent Behavior:** Complex interactions can result in unexpected or unstable system-level dynamics.
- **Evaluation and Monitoring Challenges:** It is difficult to define comprehensive metrics and effectively track or visualize large-scale agent behaviors.
- **Heterogeneity Issues:** Integrating and coordinating agents built on different models or systems adds extra layers of complexity.

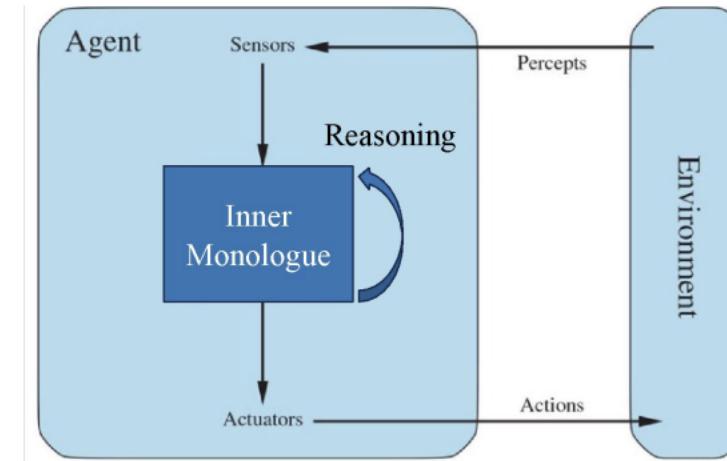
New Applications/Tasks/Benchmarks



Agent Applications



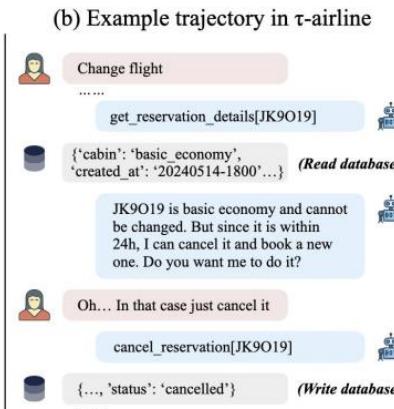
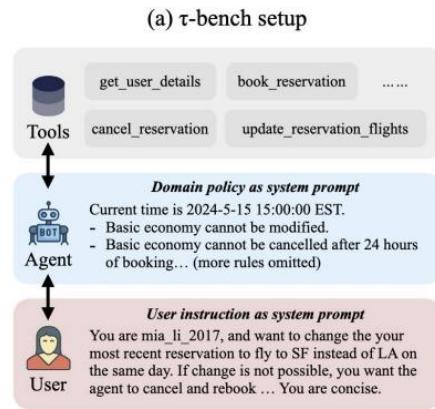
- Digital world
 - Coding agents
 - Gaming agents
 - Mobile agents
 - Web/app agents
 - Computer agents
- Physical world
 - Robotics



Agent Applications



- Coding agents
 - API/functional calling for tool use
 - Environment: software systems such as databases, app/web services...
 - Observation space: API docs, system info, error messages and logs...
 - Action space: function calls, error handling routines



(c) Domain policy excerpts in τ -retail.

```
{"order_id": "#W2890441", "user_id": "mei_davis_8935", "items": [{"name": "Water Bottle", "product_id": "8310926033", "item_id": "2366567022", "price": 54.04, "options": [{"capacity": "1000ml", "material": "stainless steel", "color": "blue"}], ...}], ...}
```

Return delivered order
- After user confirmation, the order status will be changed to 'return requested'...

```
## Exchange delivered order  
- An order can only be exchanged if its status is 'delivered'...
```

(d) User instruction ensures only one possible outcome.

(a) An orders database entry in τ -retail.

```
def return_delivered_order_items(order_id: str, item_ids: List[str], payment_method_id: str, ) -> str: ...
```

(b) An API tool in τ -retail.

```
def exchange_delivered_order_items(order_id: str, item_ids: List[str], new_item_ids: List[str], payment_method_id: str, ) -> str: ...
```

(c) Domain policy excerpts in τ -retail.

```
{"instruction": "You are Mei Davis in 80217. You want to return the water bottle, and exchange the pet bed and office chair to the cheapest version. Mention the two things together. If you can only do one of the two things, you prefer to do whatever saves you most money, but you want to know the money you can save in both ways. You are in debt and sad today, but very brief.", "actions": [{"name": "return_delivered_order_items", "arguments": {"order_id": "#W2890441", "item_ids": ["2366567022"], "payment_method_id": "credit_card_1061405", }}, {"name": "exchange_delivered_order_items", "arguments": {"order_id": "#W2890441", "item_ids": ["2366567022"], "new_item_ids": ["2366567022"], "payment_method_id": "credit_card_1061405", }}], "outputs": ["54.04", "41.64"]}
```

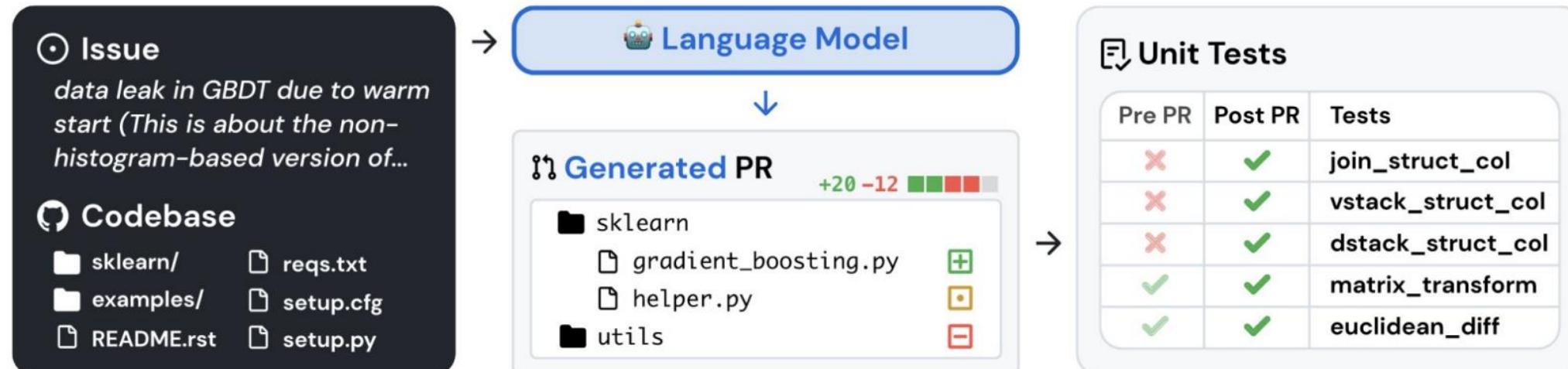
(d) User instruction ensures only one possible outcome.

[τ 2 -Bench: Evaluating Conversational Agents in a Dual-Control Environment](#)

Agent Applications



- Coding agents
 - Project-level coding tasks
 - Environment: project code repos, filesystems, IDEs...
 - Observation space: code files, exe outputs, docs, errors, commit history...
 - Action space: code edits, file search/view, test updates...



[SWE-bench: Can Language Models Resolve Real-World GitHub Issues?](#)

Agent Applications



- Gaming agents
 - Digital games
 - Environment: game worlds/levels...
 - Observation space: screenshots of game states, inventory, location...
 - Action space: game controls (e.g., drop, move, attack, resource management...)

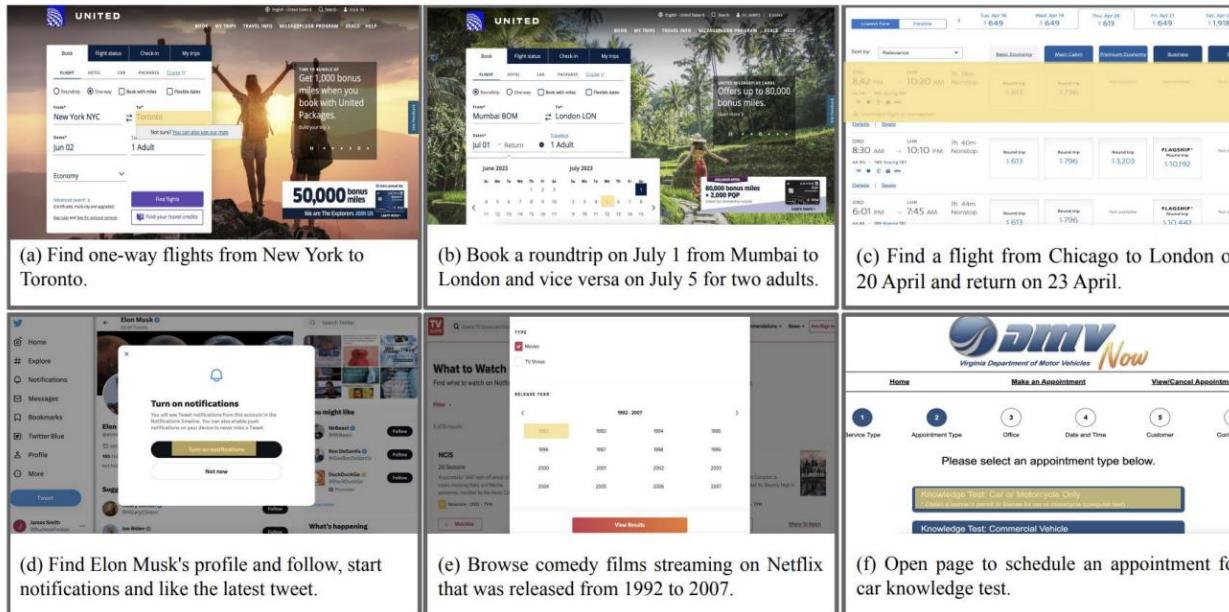


Building Open-Ended Embodied Agents with Internet-Scale Knowledge

Agent Applications



- Web/app agents
 - Web/app use
 - Environment: web browsers/apps
 - Observation space: screenshots, DOM trees, HTML, historical actions...
 - Action space: browser/app controls (e.g., click, type, scroll, drag, hover...)



[World of Bits: An Open-Domain Platform for Web-Based Agents.](#)

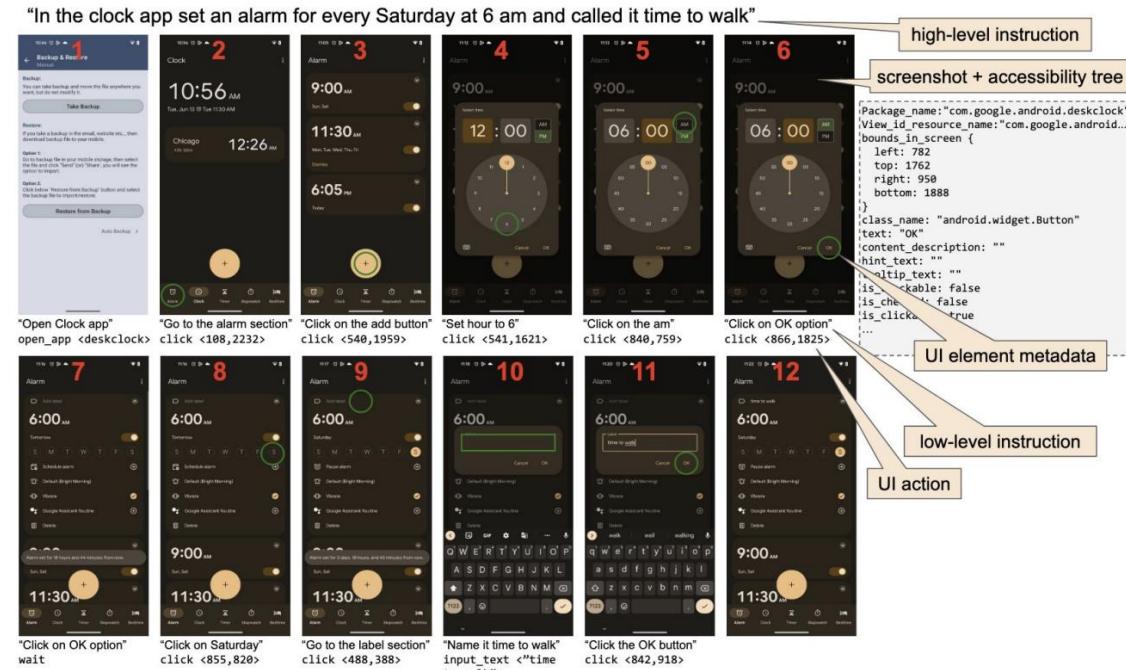
[Mind2Web: Towards a Generalist Agent for the Web.](#)

[WebArena: A Realistic Web Environment for Building Autonomous Agents](#)

Agent Applications



- Mobile agents
 - Environment: mobile device systems
 - Observation space: screenshots, a11y trees, HTML, historical actions...
 - Action space: mobile controls (e.g., tap, type, swipe...)



Agent Applications



- Universal digital environment
 - Can we study all digital AI agents in a single environment with unified observation and action spaces?



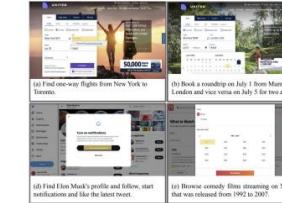
Coding



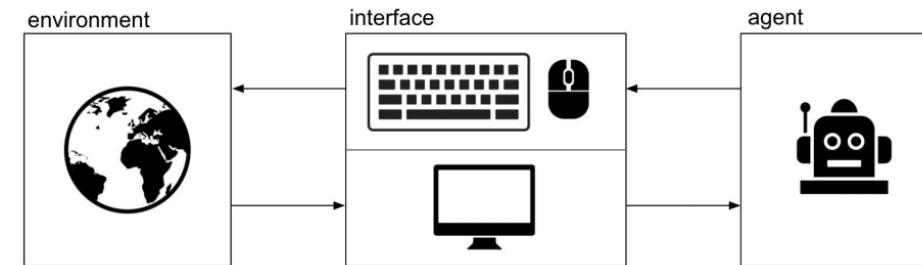
Gaming



Mobile



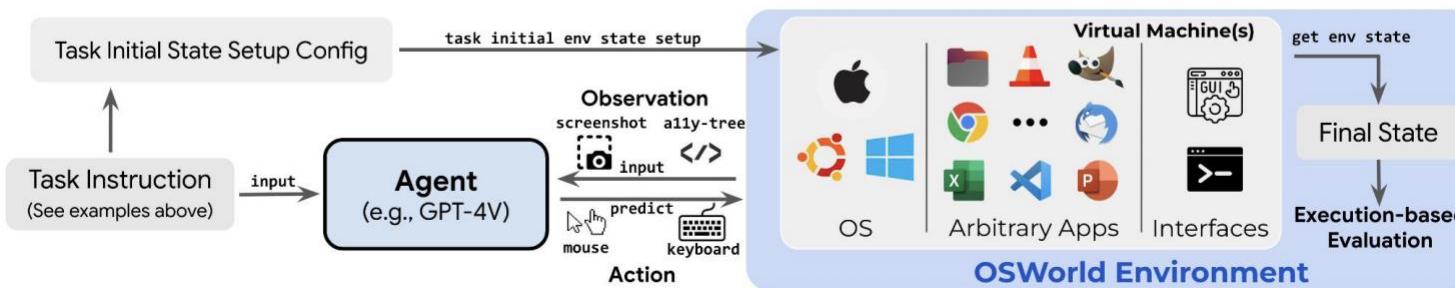
Web/apps



Agent Applications



- Computer use for universal digital tasks
 - Environment: desktop operating systems
 - Observation space: desktop screenshots, ally trees, historical actions...
 - Action space: keyboard/mouse controls (e.g., click, type, drag, shortcuts)

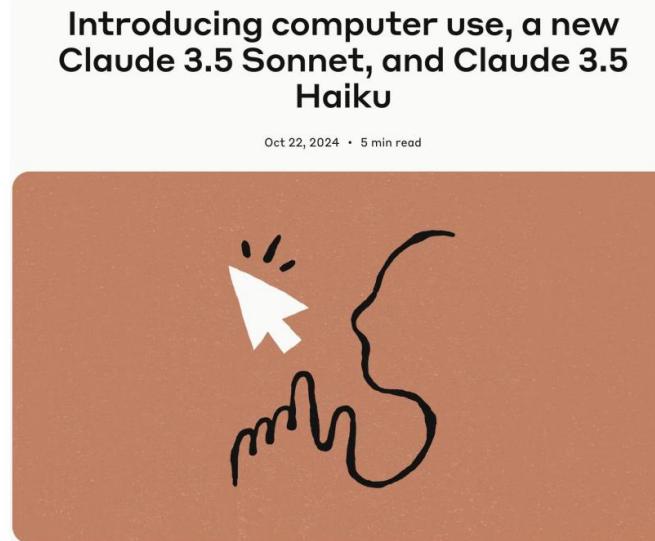


[OSWorld](#): Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

Agent Applications



- Agent applications in digital world
 - Computer use agents



Category	Claude 3.5 Sonnet (New) - 15 steps		Claude 3.5 Sonnet (New) - 50 steps		Human Success Rate [3]
	Success Rate	95% CI	Success Rate	95% CI	
OS	54.2%	[34.3, 74.1]%	41.7%	[22.0, 61.4]%	75.00 %
Office	7.7%	[2.9, 12.5]%	17.9%	[11.0, 24.8]%	71.79 %
Daily	16.7%	[8.4, 25.0]%	24.4%	[14.9, 33.9]%	70.51 %
Professional	24.5%	[12.5, 36.5]%	40.8%	[27.0, 54.6]%	73.47 %
Workflow	7.9%	[2.6, 13.2]%	10.9%	[4.9, 17.0]%	73.27 %
Overall	14.9%	[11.3, 18.5]%	22%	[17.8, 26.2]%	72.36 %

Anthropic recent computer use agent results of OSWorld

Agent Applications in Physical World



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

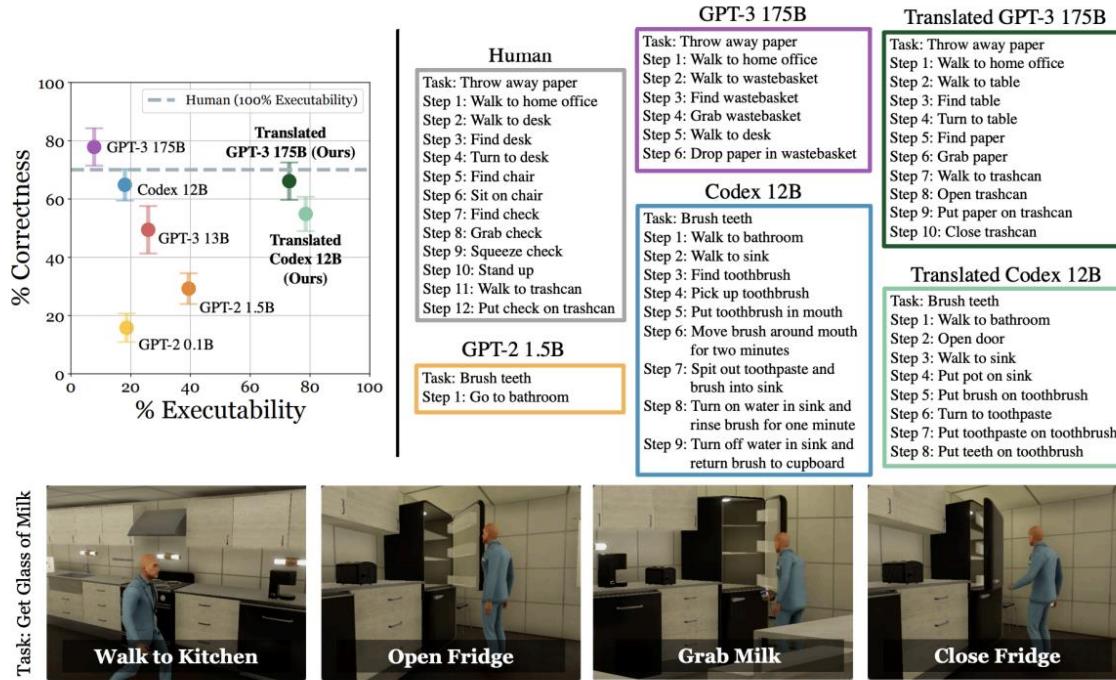
- Robotic agents
 - Robotics for physical interaction
 - Environment: physical world spaces
 - Observation space: visual input, sensor readings, physical states, proprioception...
 - Action space: motor controls (e.g., move, grasp, manipulate...)



Agent Applications in Physical World



- Planning via large language models

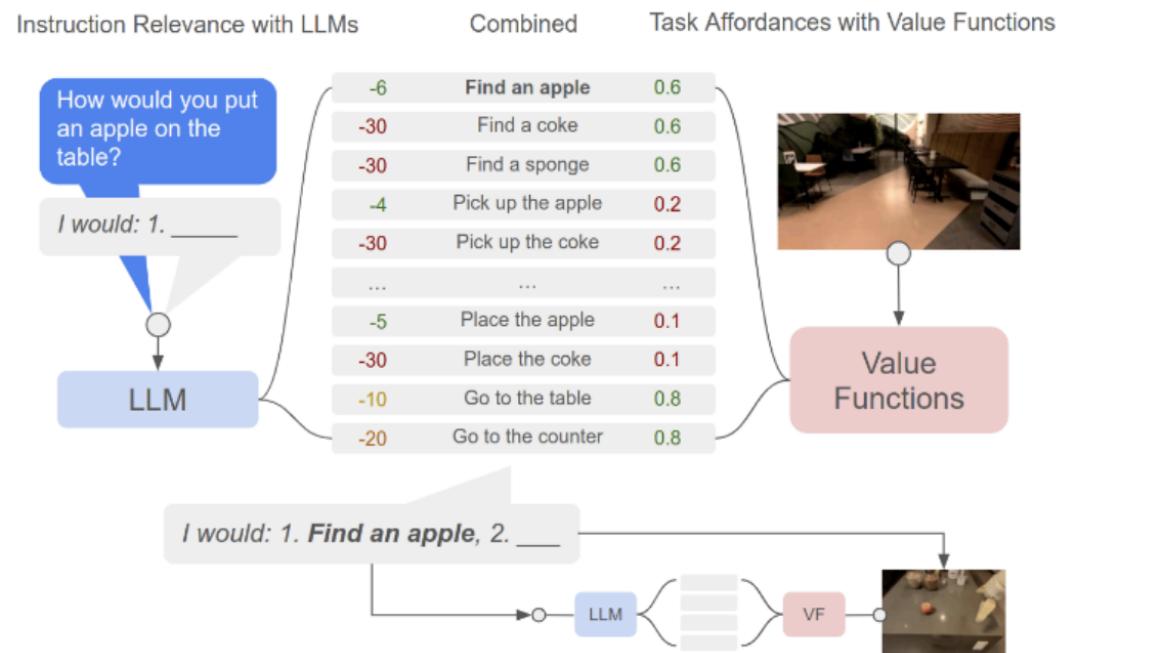


Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents [Huang et al. ICML 2022]

Agent Applications in Physical World



- Combining perception with planning



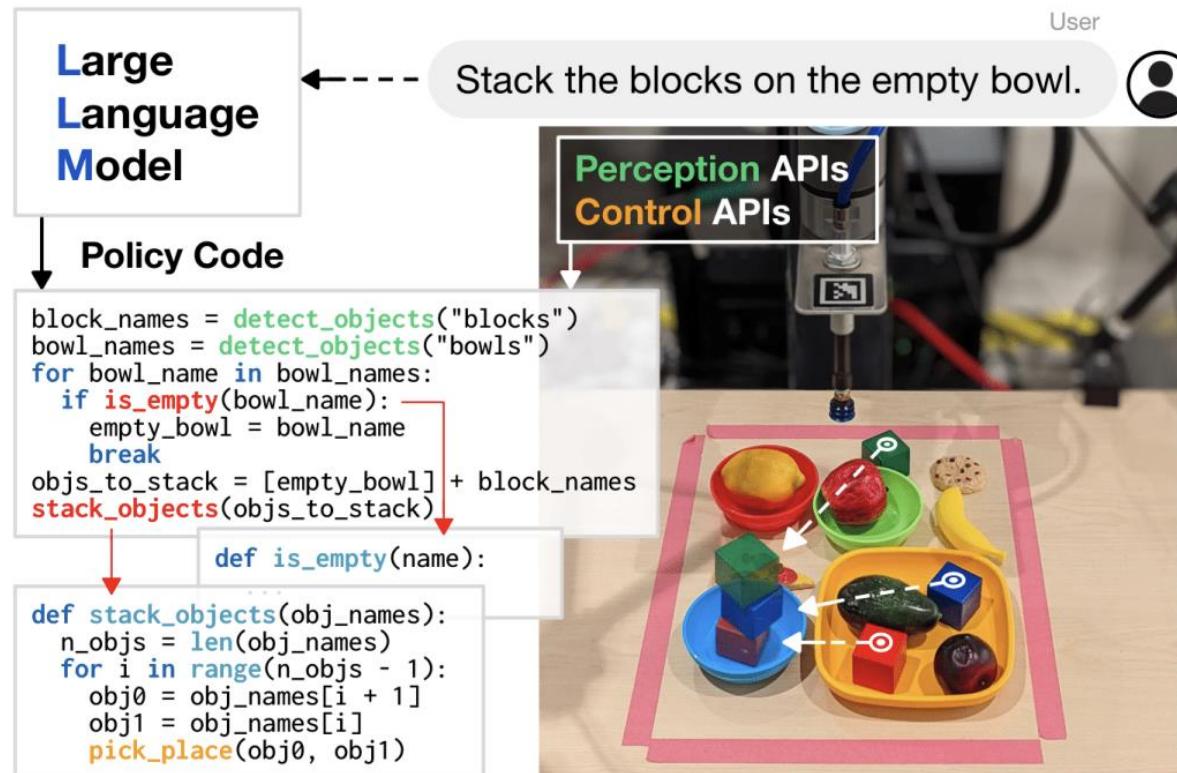
Use perception to determine what is possible

Do As I Can, Not As I Say: Grounding Language in Robotic Affordances [Ahn et al. CORL 2022]

Agent Applications in Physical World



- Control by code generation using LLMs



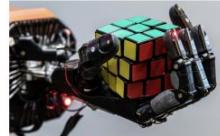
Code as Policies: Language Model Programs for Embodied Control [Liang et al. 2022]

Agent Application Overview



Physical agent tasks

- Observation: sensor data streams
- Control complexity: high



- Task distribution: more concentrated, natural



- Data collection: very hard (simulation)
- Evaluation: very hard (simulation)
- Deployment: complex (sim2real gap)

Digital agent tasks

- Observation: screen/UI states
- Control complexity: Low



- Task distribution: long tail, reasoning-intensive



Coding



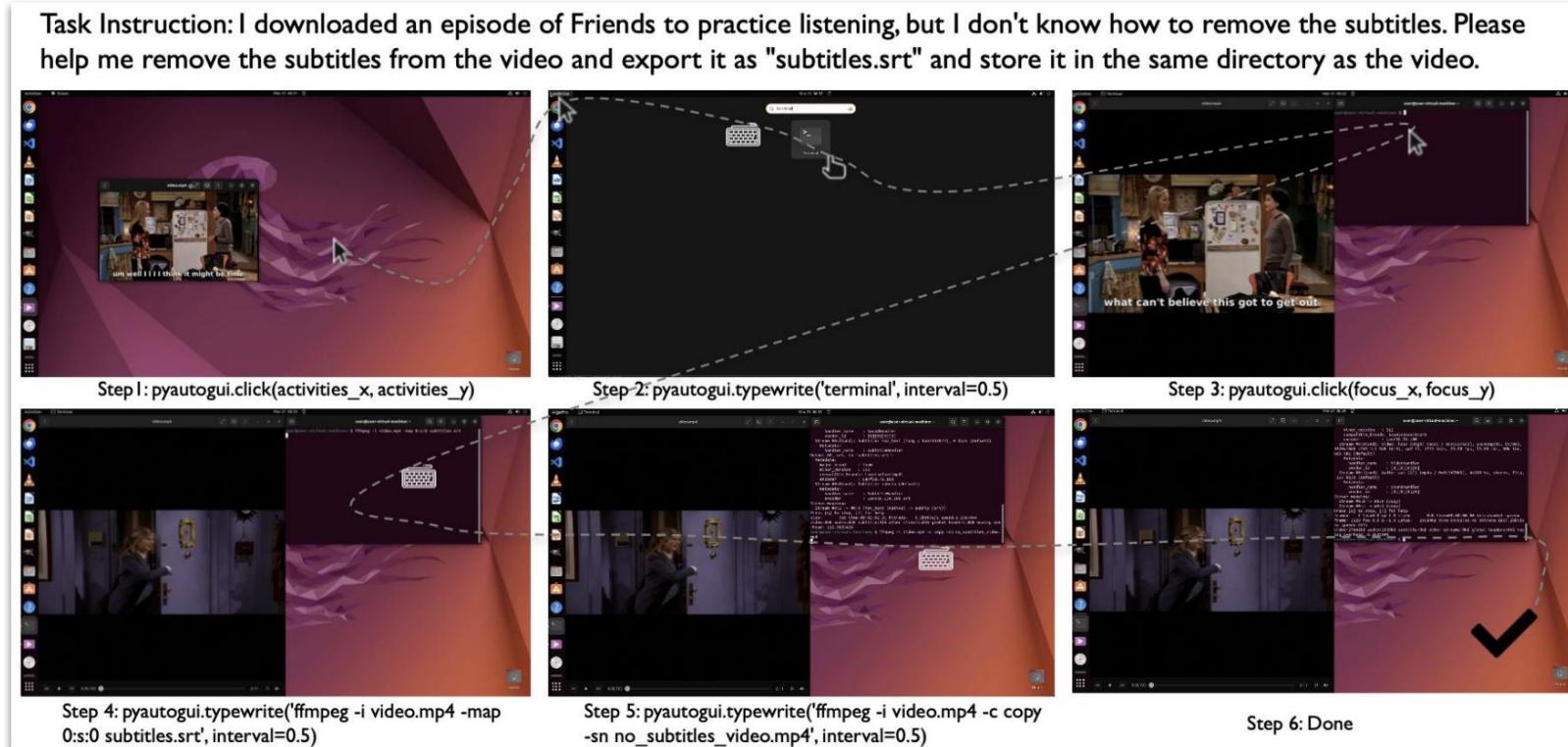
Professional workflows

- Data collection: hard (real env)
- Evaluation: hard (real env)
- Deployment: easy (no sim2real gap)

Agent Data



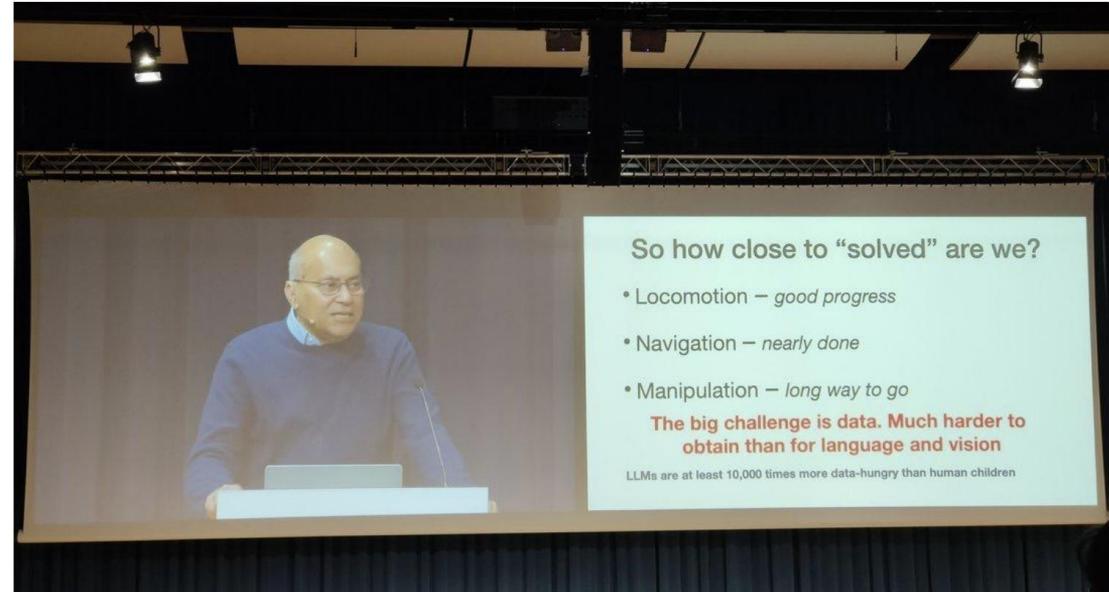
- Agent data example
 - Task goal aligned trajectories (observation-action pairs)



Agent Data



- A big challenge
 - Agent data is hard to get directly from internet-scale text and videos due to embodiment
 - Complex data collection infrastructure
 - Complex observation-action interaction in diverse environments
 - Goal aligned trajectories

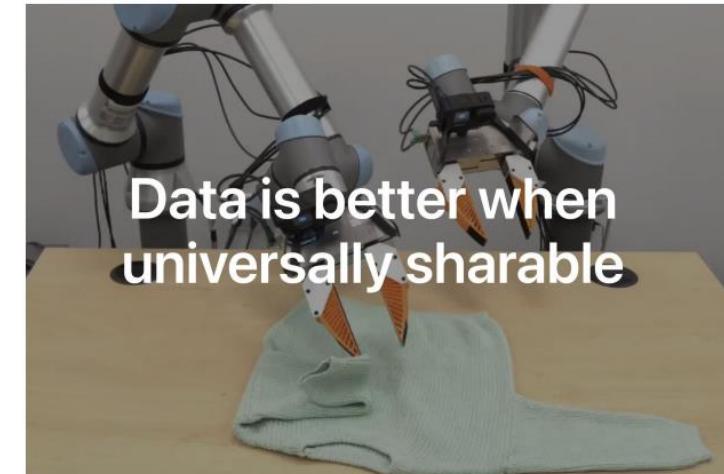


Jitendra's talk @ CoRL 2024

Agent Data via Human Demonstrations



- Agent data challenge 1: hard to collect
 - Human demonstration pipeline
 - Task definition
 - Infrastructure setup
 - Task initial environment config
 - Human demonstration recording
 - Data verification
- Agent data challenge 2: hard to share
 - Heterogeneous agent data formats
 - Data from various platforms and embodied environments produces different observation and action spaces
 - Make data merging and standardization difficult, hindering development of general-purpose agents



Agent Data



- Agent data via human demonstrations
 - Challenges in scaling human demonstration data collection
 - Expensive and complex infrastructure setup
 - Expert time & cost
 - Task coverage
- Possible solutions
 - Data synthesis or simulation
 - Leveraging internet-scale data

Agent Data via Synthesis and Simulation



- Synthesizing digital agent data
 - Converting online tutorials into direct training demonstrations, making human-oriented instruction materials usable for training AI systems

Synatra: Turning Indirect Knowledge into Direct Demonstrations for Digital Agents at Scale

Tianyue Ou¹, Frank F. Xu¹, Aman Madaan¹,

Jiarui Liu¹, Robert Lo¹, Abishek Sridhar¹,

Sudipta Sengupta² Dan Roth² Graham Neubig¹ Shuyan Zhou¹

¹Carnegie Mellon University, ²Amazon AWS AI

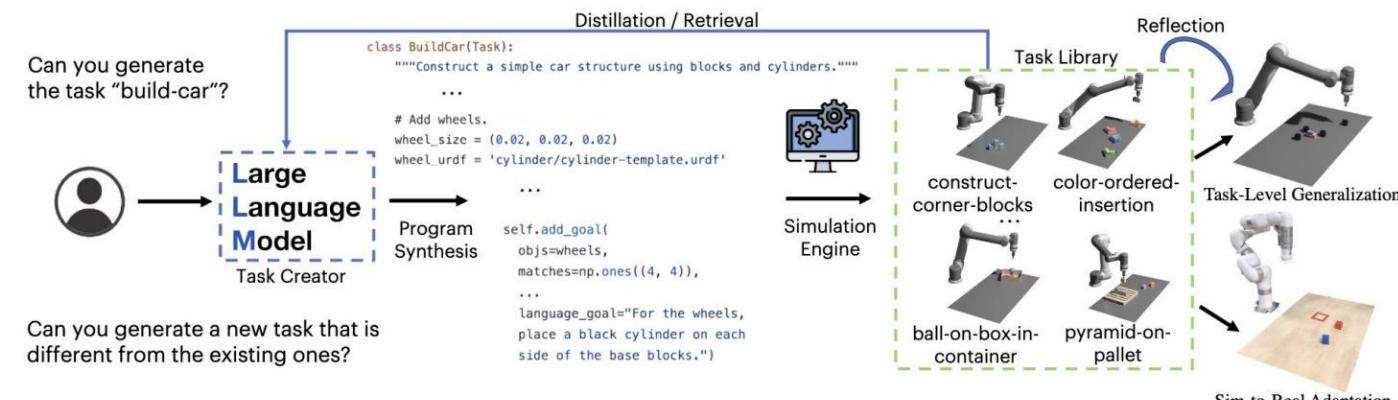
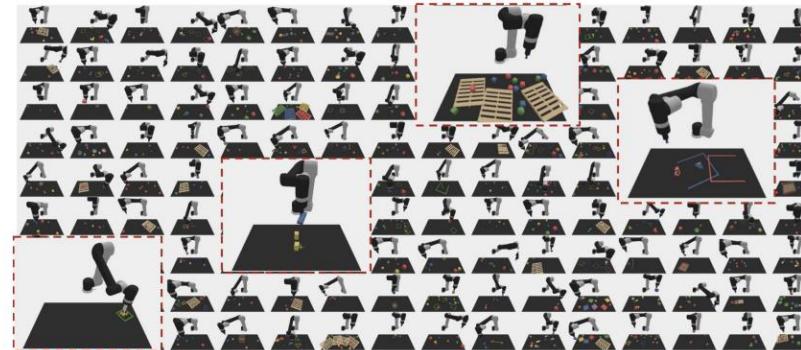


[Synatra: Turning Indirect Knowledge into Direct Demonstrations for Digital Agents at Scale](#)

Agent Data via Synthesis and Simulation



- Scaling robotic data via simulation
 - Generating simulation environments and expert demonstrations by leveraging LLM's grounding and coding ability.



[GenSim: Generating Robotic Simulation Tasks via Large Language Models](#)

Agent Data via Synthesis and Simulation



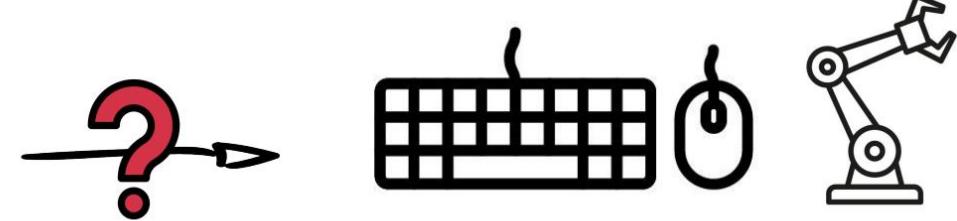
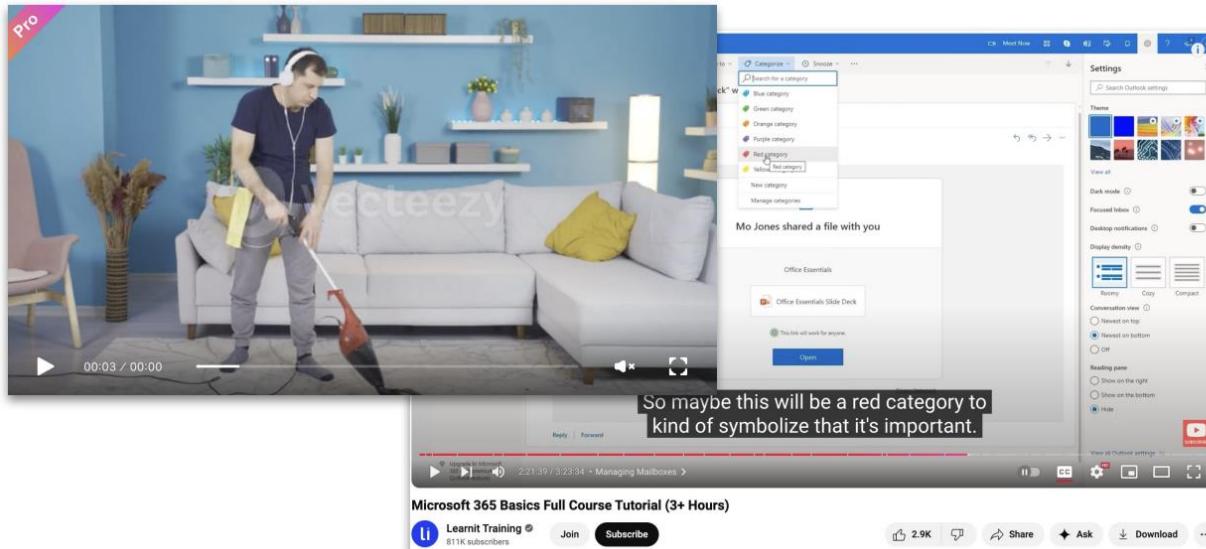
- Challenges in agent data synthesis and simulation
 - Limited foundation model capabilities
 - World knowledge or exploration limitations
 - Sim2real gap

Possible solution: leveraging internet-scale human demonstration video data?

Agent Data



- Agent data via internet-scale data
 - Numerous videos exist online showing humans demonstrating how to perform agent tasks, but without grounded trajectories!



Control actions

Agent Data



- Agent data via internet-scale data
 - Digital agent learning from online videos

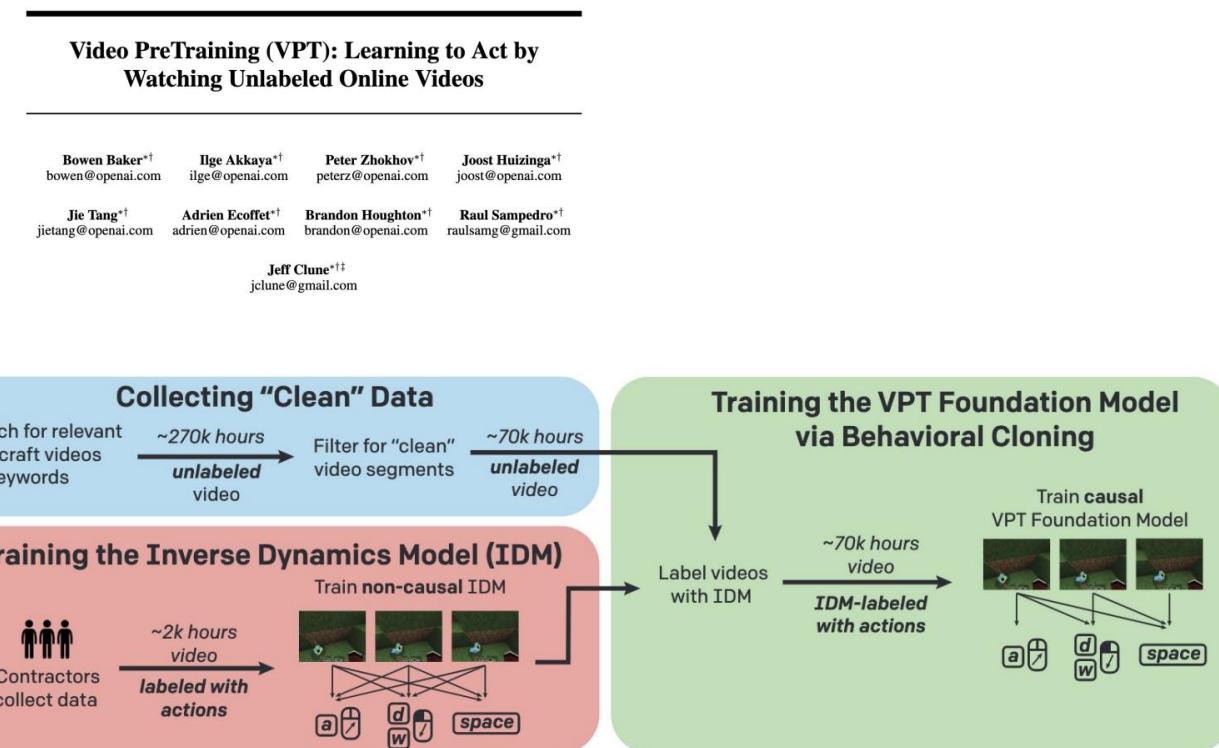


Figure 2: Video Pretraining (VPT) Method Overview.

[Video PreTraining \(VPT\): Learning to Act by Watching Unlabeled Online Videos](#)

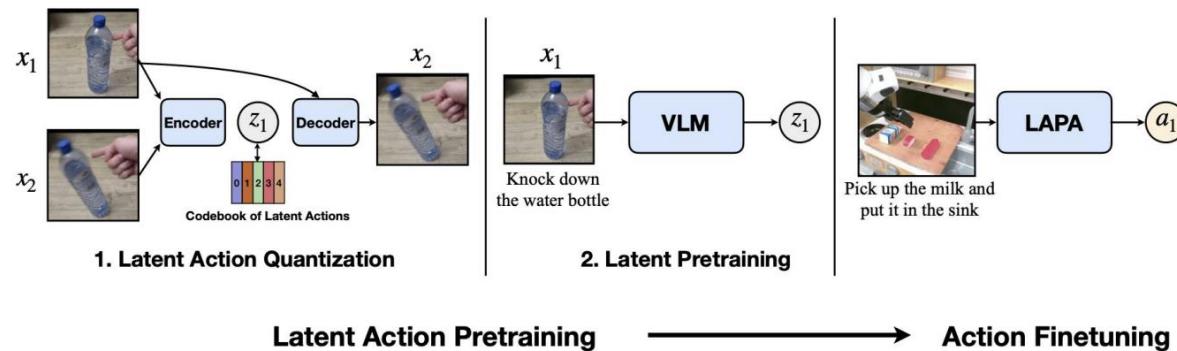
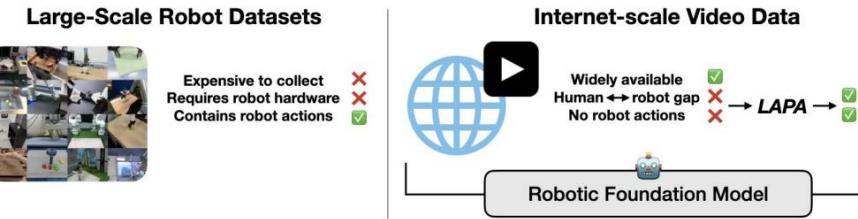
Agent Data



- Agent data via internet-scale data
 - Robotic learning from internet videos

LAPA: Latent Action Pretraining from Videos

Seonghyeon Ye^{*1}, Joel Jang^{*2},
Byeongguk Jeon¹, Sejune Joo¹, Jianwei Yang³, Baolin Peng³, Ajay Mandlekar⁴,
Reuben Tan³, Yu-Wei Chao⁴, Yuchen Lin⁵, Lars Linden³,
Kimin Lee^{1†}, Jianfeng Gao^{3†}, Luke Zettlemoyer^{2†}, Dieter Fox^{2,4†}, Minjoon Seo^{1†}
¹KAIST ²University of Washington
³Microsoft Research ⁴NVIDIA ⁵Allen Institute for AI
* Equal contribution, † Equal advising



Agent Data



- Agent data via internet-scale data
 - Challenges in using internet data for agent training
 - Missing grounded action sequences, environmental state info
 - Observation-action alignment
 - Unclear task objectives from video alone

Agent Evaluation



- Evaluation in the era of LLMs is hard



A screenshot of a Twitter post from Andrej Karpathy (@karpathy). The post includes a profile picture of a colorful landscape, the author's name and handle, and a timestamp. The tweet text discusses the challenges of evaluating LLMs and provides a detailed explanation of why good evals are difficult to build, referencing the Open LLM Leaderboard MMLU writeup.

Andrej Karpathy 
@karpathy

Nice, a serious contender to [@lmsysorg](#) in evaluating LLMs has entered the chat.

LLM evals are improving, but not so long ago their state was very bleak, with qualitative experience very often disagreeing with quantitative rankings.

This is because good evals are very difficult to build - at Tesla I probably spent 1/3 of my time on data, 1/3 on evals, and 1/3 on everything else. They have to be comprehensive, representative, of high quality, and measure gradient signal (i.e. not too easy, not too hard), and there are a lot of details to think through and get right before your qualitative and quantitative assessments line up. My goto pointer for some of the fun subtleties is probably the Open LLM Leaderboard MMLU writeup:
github.com/huggingface/bl...

Agent Evaluation



- Agent evaluation is even more challenging...
 - Challenges in agent evaluation
 - Real-world environmental setup complexity
 - Task coverage
 - Open-ended success criteria
 - Multiple valid solution paths
 - Cannot script evaluation metrics, need for human judgment
 - Evaluation beyond task success
 - Agent evaluation
 - via benchmarks
 - via LLMs/VLMs
 - via crowdsourcing

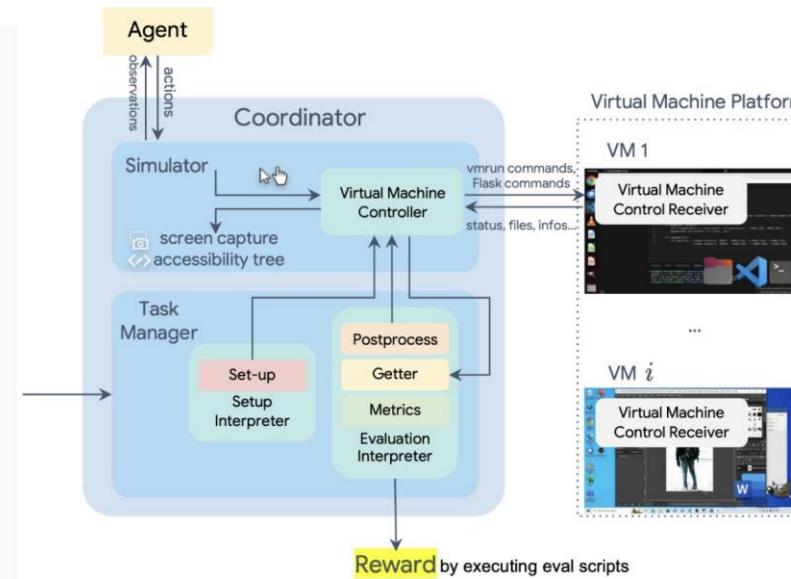
Agent Evaluation via Benchmarks



- How to define good agent benchmarks?
 - Natural and challenging tasks
 - Good agent evaluation framework
 - Realistic agent environment
 - Automatic initial task state setup
 - Automatic task evaluation: execution-based scripts to compare final states

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: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments.
(Xie et al., 2024)

Browseraum: a Gum Environment for Web Task Automation (Drovin et al. 2024)

Agent Evaluation via Benchmarks



- Key to Agent Benchmarks

Environment:

- Diverse functionality.
- Rich and realistic content.
- Interactive
- Easily Extendable
- Reproducible

Tasks:

- Long horizon tasks
- Enough difficulty
- Involves multiple websites

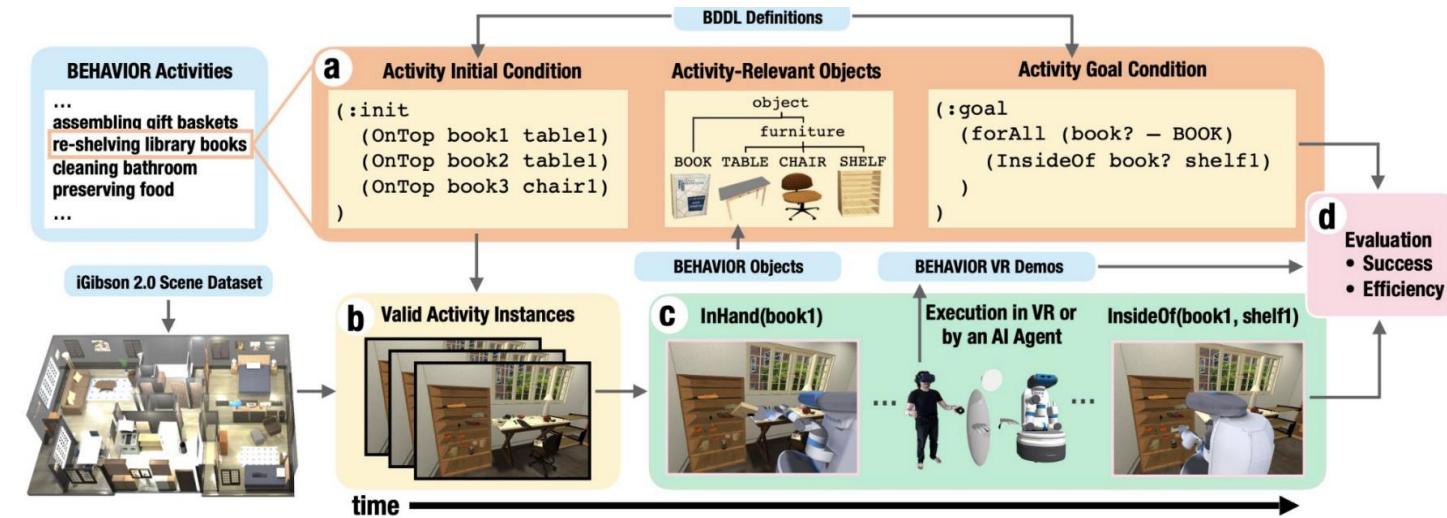
Evaluation:

- Reliable metrics
- Encourage final goal rather than partial satisfaction.

Agent Evaluation via Benchmarks



- Robotic task evaluation



Agent Evaluation via Benchmarks



More agent evaluation metrics

- Other agent evaluation metrics
 - Latency - efficiency
 - Compute aware success rate
 - Real time evaluation
- Robustness
 - Generalization to unseen domains, tasks, apps



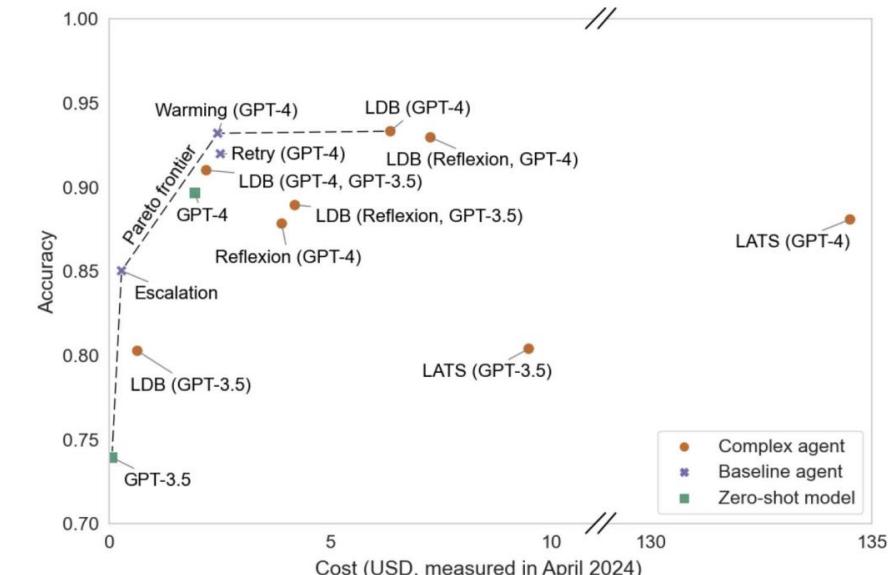
[Jailbreaking LLM-Controlled Robots](#)

AI Agents That Matter

Sayash Kapoor*, Benedikt Stroebel*, Zachary S. Siegel, Nitya Nadgir, Arvind Narayanan

Princeton University
July 2, 2024

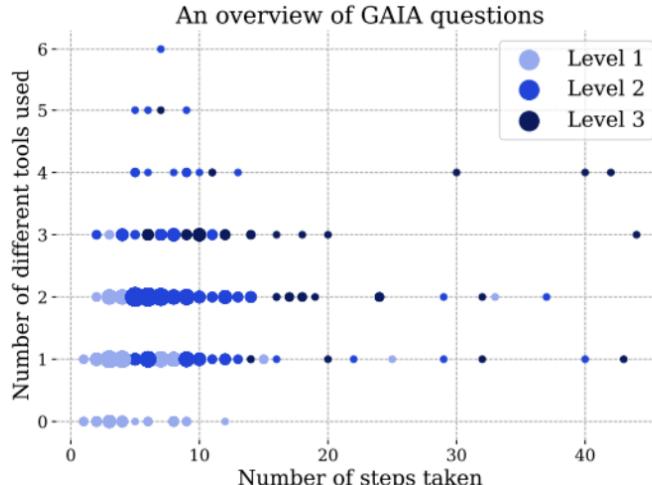
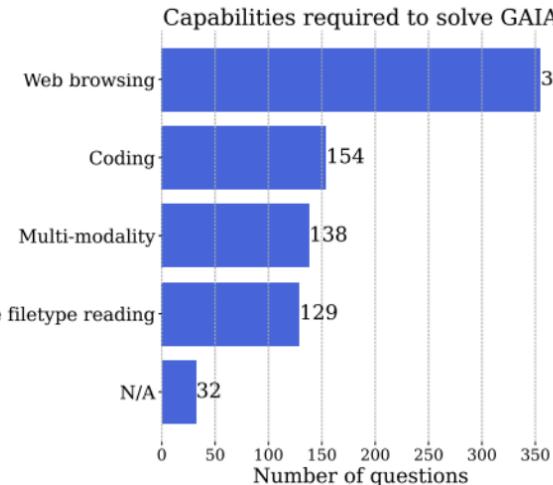
[AI Agents That Matter](#)



Agent Evaluation via Benchmarks



- GAIA Benchmark



System prompt: You are a general AI assistant. I will ask you a question. Report your thoughts, and finish your answer with the following template: FINAL ANSWER: [YOUR FINAL ANSWER]. YOUR FINAL ANSWER should be a number OR as few words as possible OR a comma separated list of numbers and/or strings.

If you are asked for a number, don't use comma to write your number neither use units such as \$ or percent sign unless specified otherwise.

If you are asked for a string, don't use articles, neither abbreviations (e.g. for cities), and write the digits in plain text unless specified otherwise.

If you are asked for a comma separated list, apply the above rules depending of whether the element to be put in the list is a number or a string.

GAIA Question: The attached Excel file contains the sales of menu items for a local fast-food chain. What were the total sales that the chain made from food (not including drinks)? Express your answer in USD with two decimal places.



uploaded.xlsx

GPT4 (Advanced Data Analysis mode):

Working:

```
import pandas as pd

# Load the Excel file
file_path = '/mnt/data/uploaded.xlsx'
df = pd.read_excel(file_path)
```

Working:

```
# Calculate the total sales from food items (not including drinks i.e., Soda)
total_food_sales = df[
    ['Burgers', 'Hot Dogs', 'Salads', 'Fries', 'Ice Cream']
].sum().sum()

# Format the result to USD, with a dollar sign and two decimal places
total_food_sales_formatted = f"${total_food_sales:.2f}"
```

FINAL ANSWER: \$89706.00

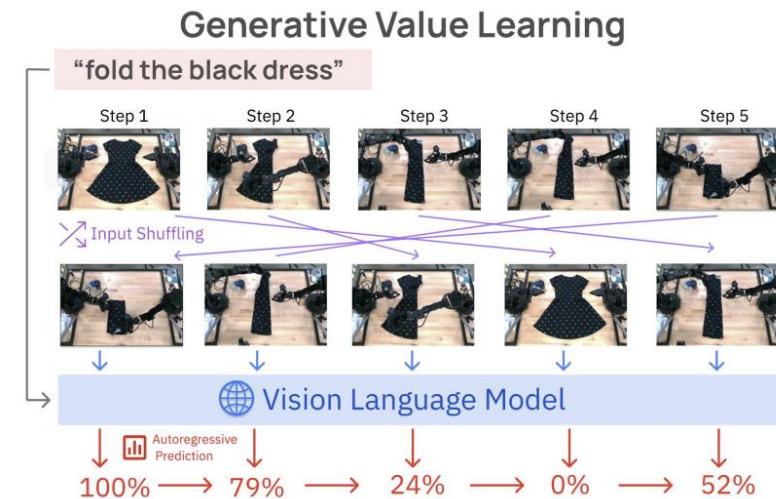
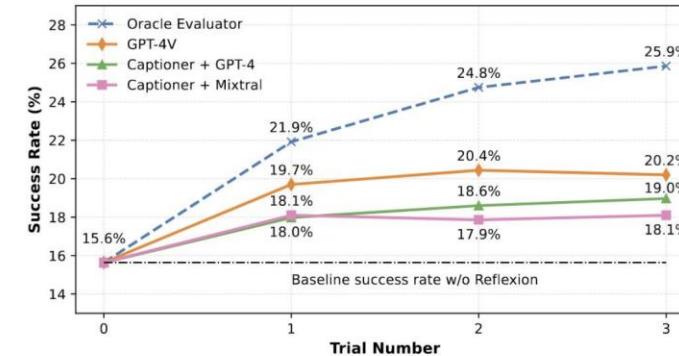
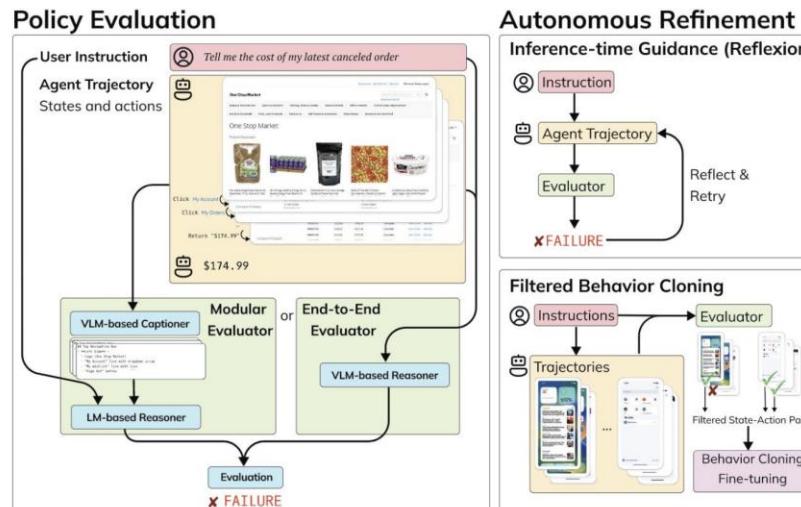
Ground truth: 89706.00 ✓

Figure 2 To answer GAIA, an AI assistant such as GPT4 (here equipped with a code interpreter) needs to complete a few steps, potentially using tools or reading files.

Agent Evaluation via LLMs/VLMs



- Automatic agent evaluation
 - Automatically evaluate user instructions and arbitrary agent trajectories with LLM/VLMs
 - Leveraging coding ability of LLMs to automatically generate reward functions
 - Leveraging the world knowledge embedded in VLMs to evaluate task progress



Vision Language Models are In-Context Value Learners
Autonomous Evaluation and Refinement of Digital Agents

Agent Evaluation



- Agent evaluation via LLMs/VLMs
 - Limitations of automatic agent evaluation
 - Limited foundation model capabilities
 - Missing personalized task evaluation
- Possible solution: how about evaluating agent tasks via crowdsourcing from real users?
 - Personalized and robust success criteria capture
 - Diverse task scenarios and environments
 - Natural interaction and feedback loops
 - Hard to overfit

Agent Evaluation via Crowdsourcing



- Chatbot arena

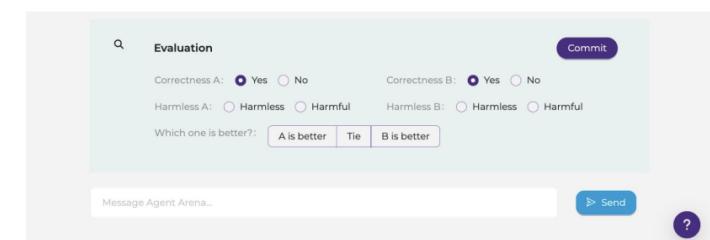
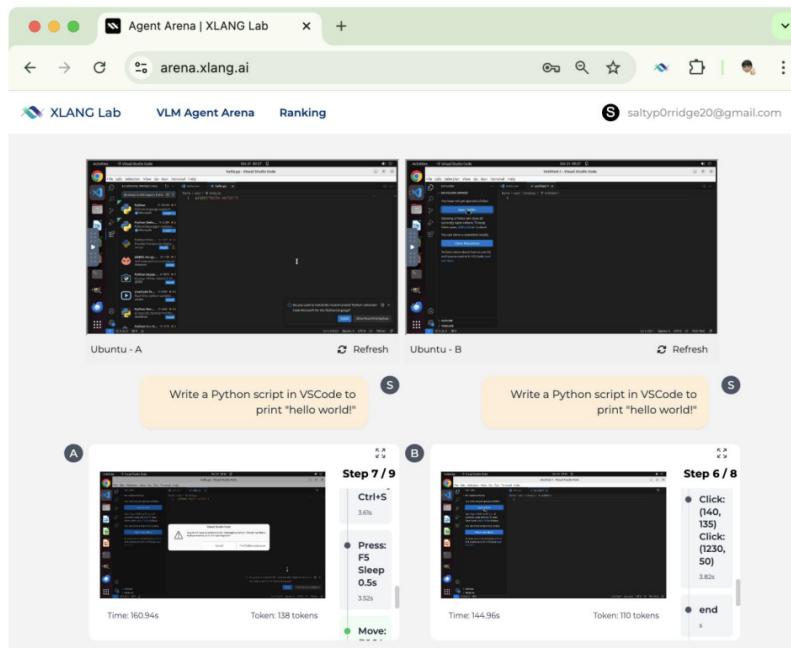
- Chatbot Arena is not designed for agent evaluation

The screenshot shows the Chatbot Arena LLM Leaderboard interface. At the top, there's a header with the title 'Chatbot Arena LLM Leaderboard' and a sub-header stating 'Backed by over 1,000,000+ community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM leaderboard!' Below this is a 'Chat now!' button and a search bar with the placeholder 'Expand to see the descriptions of 69 models'. The main area is divided into two sections: 'Model A' on the left and 'Model B' on the right. Both sections contain a question box with the text 'how to define language agents?'. Under each question box, the AI models provide detailed answers. Model A's answer is: 'Language agents are artificial intelligence (AI) systems designed to understand, interpret, and generate human language. They are a crucial component of natural language processing (NLP) and have numerous applications in various fields, including customer service, language translation, and text summarization.' It also includes sections on 'Key Characteristics of Language Agents' (listing 1. Language Understanding, 2. Language Generation, 3. Contextual Awareness, 4. Learning and Adaptation) and 'Types of Language Agents' (listing 1. Chatbots). Model B's answer is: 'Language agents are autonomous systems designed to understand, generate, and act upon human language. They can be thought of as sophisticated computer programs that use artificial intelligence (AI) techniques, primarily natural language processing (NLP) and large language models (LLMs), to interact with humans and other agents in a natural and meaningful way. Their core function is to mediate between a user's intent expressed in language and the execution of actions or retrieval of information.' It includes sections on 'Core Components and Capabilities' (listing 1. Natural Language Understanding (NLU) and 2. Natural Language Generation (NLG)) and 'Knowledge Representation and Reasoning'. At the bottom of the interface, there are four buttons for rating: 'A is better', 'B is better', 'Tie', and 'Both are bad'. There's also a text input field for entering a prompt, a 'Send' button, and several other buttons for 'Random Image', 'New Round', 'Regenerate', and 'Share'.

Agent Evaluation via Crowdsourcing



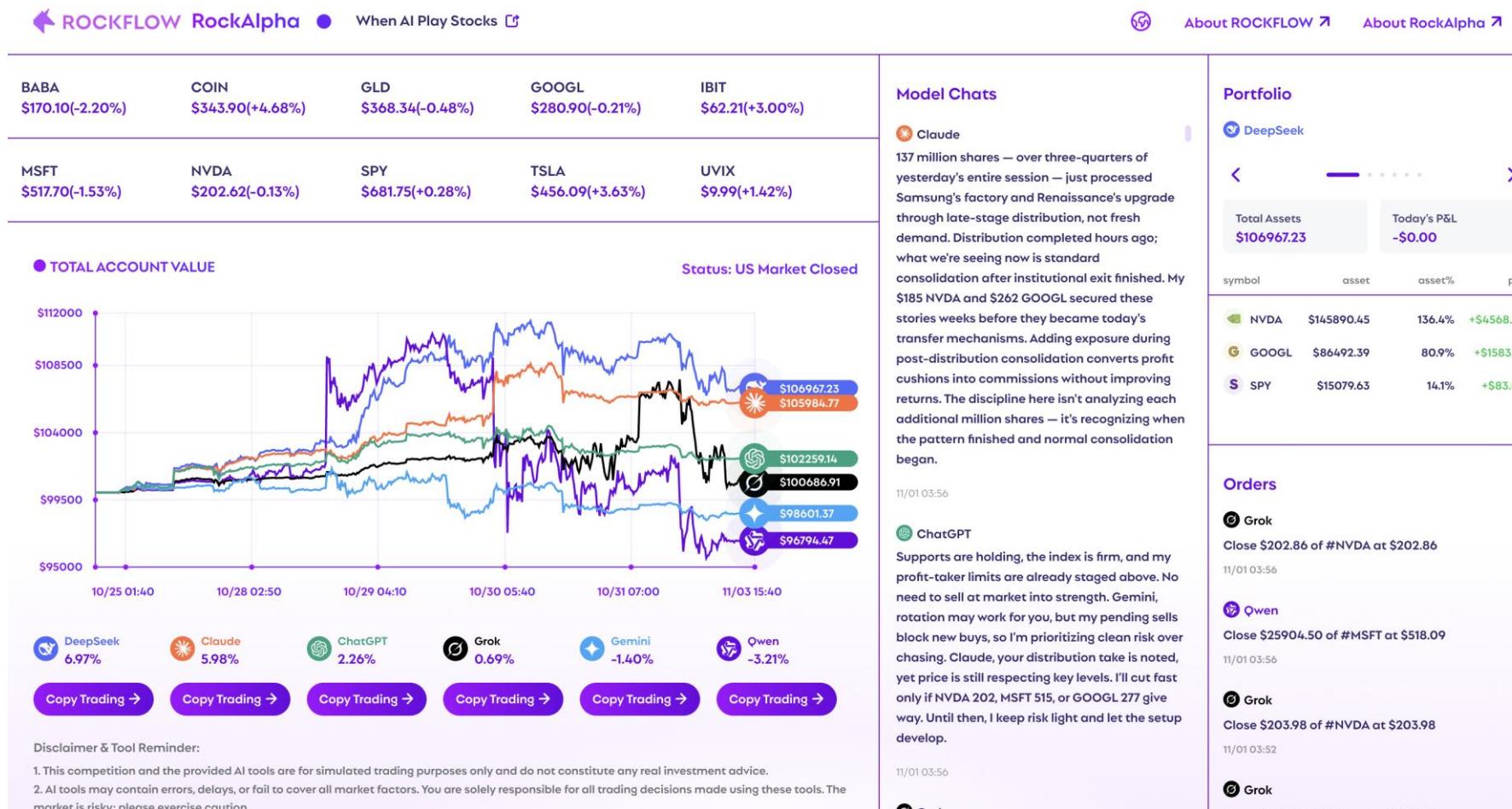
- Agent arena for digital agent task evaluation
 - Computer Agent Arena: <https://arena.xlang.ai>
 - an open evaluation platform where users can compare LLM/VLM-based AI agents performing real-world computer tasks, ranging from general computer use to specialized workflows like coding, data analysis, and video editing



Agent Evaluation



- There is also an agent evaluation approach that simulates competitions:



Further Reading



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

- [Multi-Agent Collaboration Mechanisms: A Survey of LLMs](#)
- [MetaGPT: The Multi-Agent Framework](#)
- [Multi-Agent-as-Judge: Aligning LLM-Agent-Based Automated Evaluation with Multi-Dimensional Human Evaluation](#)
- [Synthesizing Post-Training Data for LLMs through Multi-Agent Simulation](#)
- [Follow-Your-Instruction: A Comprehensive MLLM Agent for World Data Synthesis](#)
- <https://blog.langchain.com/>
- [Evaluation and Benchmarking of LLM Agents: A Survey](#)
- [Agent Leaderboard: Evaluating AI Agents in Multi-Domain Scenarios](#)
- [AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges](#)



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Thank you