

# LLM Agents and Tool Use

## IE686 Large Language Models and Agents





# Credits

- This slide set is based on slides from
  - Shunyu Yao
  - Yankai Lin
  - Yang Deng, An Zhang et al.
- Many thanks to all of you!



# Outline

- **Recap: Prompt Engineering and Efficient Adaptation**
- What is an Agent?
- Tool Usage for LLMs
- The ReAct Paradigm
- Unified Framework for LLM Agents
- Evaluating Agents



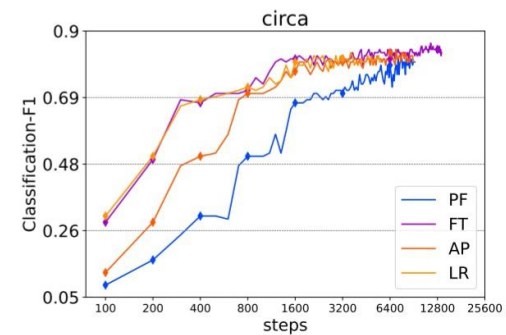
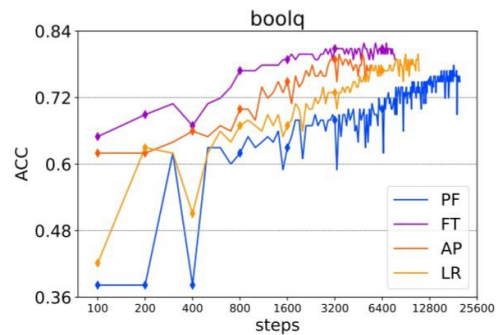
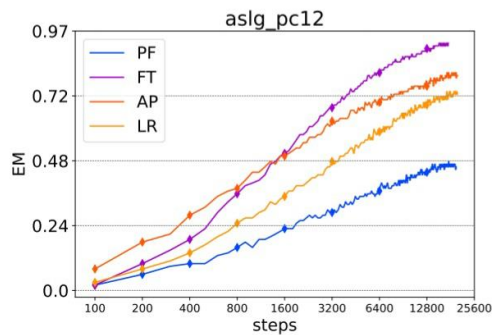
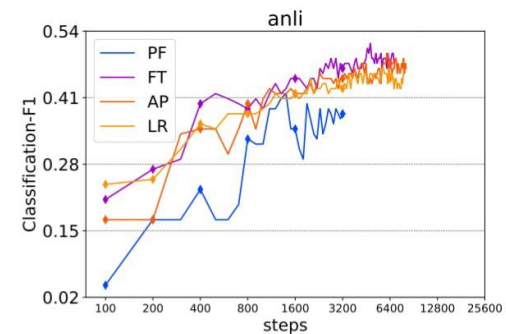
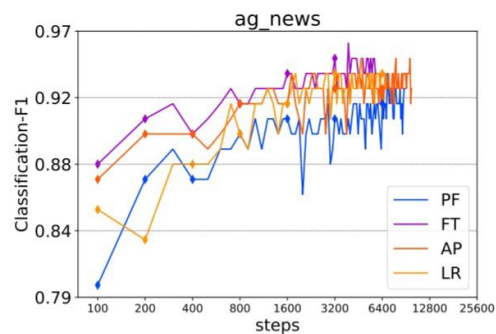
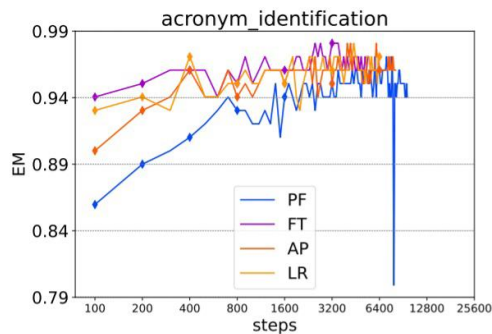
# Recap: Prompting

- For many tasks, supervised fine-tuning data may not be available or may be costly to obtain
- Due to **emergent abilities** coupled with instruction tuning, we can simply prompt or instruct models to do a task!
- Prompts are written in **natural language**
- Prompting is **non-invasive**:
  - No additional parameters are introduced
  - No tuning of existing parameters
  - No need to inspect model's embeddings



# Recap: Fine-tuning Methods

- Given enough data and computing resources
- Overall performance on T5-base: Full fine-tuning > LoRA > Adapters > Prefix Tuning > Prompt Tuning



Ding, N., et al., 2022. Delta tuning: A comprehensive study of parameter efficient methods for pre-trained language models. *arXiv preprint arXiv:2203.06904*.



# Recap: Evaluating LLMs

- **Benchmark-based evaluation**
  - Format problem into prompt and generate result
  - Parse result and compute standard metrics like accuracy
  - Good for close-ended evaluation
- **Model-based evaluation**
  - Use LLM like GPT-4 as surrogate for human evaluation
  - Shown to achieve high agreement with human evaluators
- **Human-based evaluation**
  - Human evaluators judge answer of LLMs
    - Pair-wise comparison of two answers from different models
    - Single-answer grading: score a single answer from an LLM
  - Good for open-ended evaluation

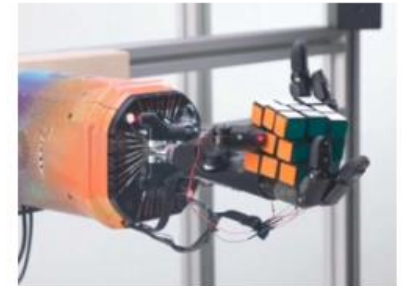


# Outline

- Recap: Prompt Engineering and Efficient Adaptation
- **What is an Agent?**
- Tool Usage for LLMs
- The ReAct Paradigm
- Unified Framework for LLM Agents
- Evaluating Agents



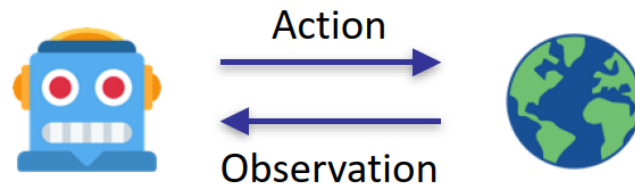
# What is an Agent?



- LLM-powered Agents are artificial entities that enhance LLMs with essential capabilities enabling them to sense their environment, make decisions, and take actions.



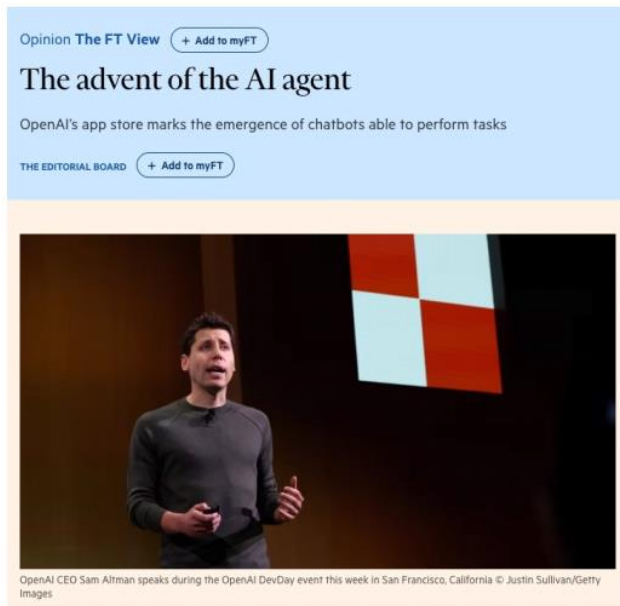
# What is an Agent?



- An “intelligent” system that interacts with some “environment”
  - Physical environments: robot, autonomous car, ...
  - Digital environments: DQN for Atari, Siri, AlphaGo
  - Humans as environment: Chatbots



# What is an Agent?



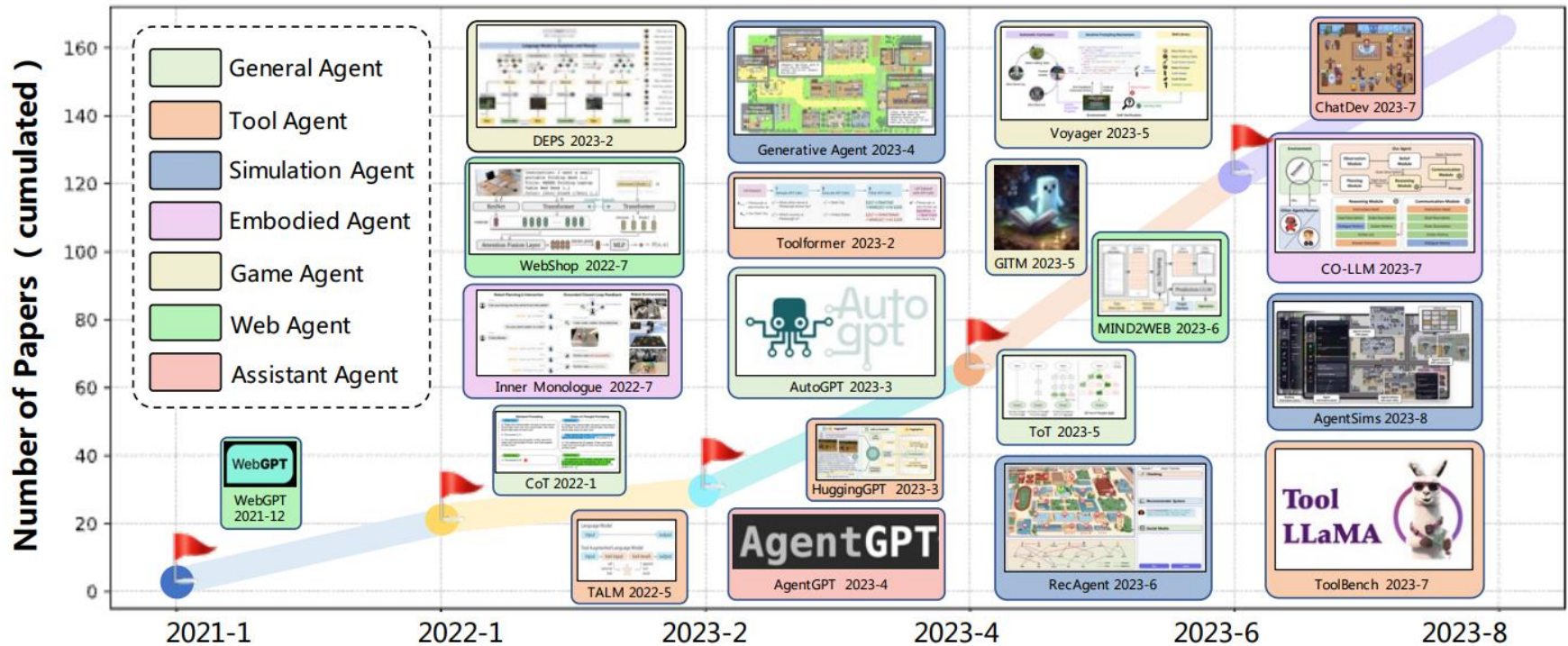
- **Sam Altman** said in one of his key notes: *“GPTs and Assistants are **precursors** to **agents**. They will gradually be able to plan and to perform more complex actions on your behalf. These are our first steps toward AI Agents.”*
- Bill Gates wrote in his Blog: *“**Agents** are not only going to change how everyone interacts with computers. They’re also going to **upend the software industry**, bringing about the biggest revolution in computing since we went from typing commands to tapping on icons.”*

Financial Times. [“The advent of the AI agent”](#)

GatesNotes. [“The Future of Agents: AI is about to completely change how you use computers”](#)

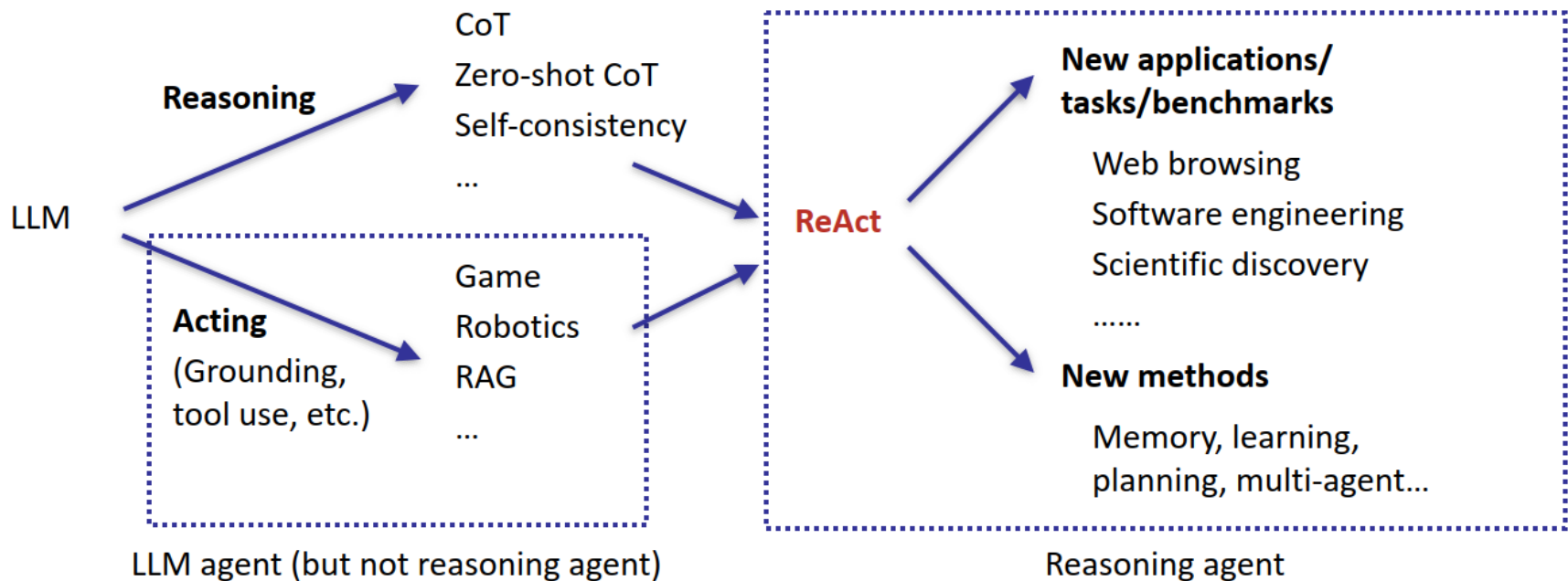


# LLM Agents over Time





# A brief history of LLM agents





# Outline

- Recap: Prompt Engineering and Efficient Adaptation
- What is an Agent?
- **Tool Usage for LLMs**
- The ReAct Paradigm
- Unified Framework for LLM Agents
- Evaluating Agents



# Example Task: Question Answering

Q: what is  $1 + 2$ ?



A: 3

Q: Janet's ducks lay 16 eggs per day. She eats three for breakfast every morning and bakes muffins for her friends every day with four. She sells the remainder for \$2 per egg. How much does she make every day?



Requires reasoning

Q: who is the latest UK PM?



Requires knowledge

Q: what is the prime factorization of 34324329?



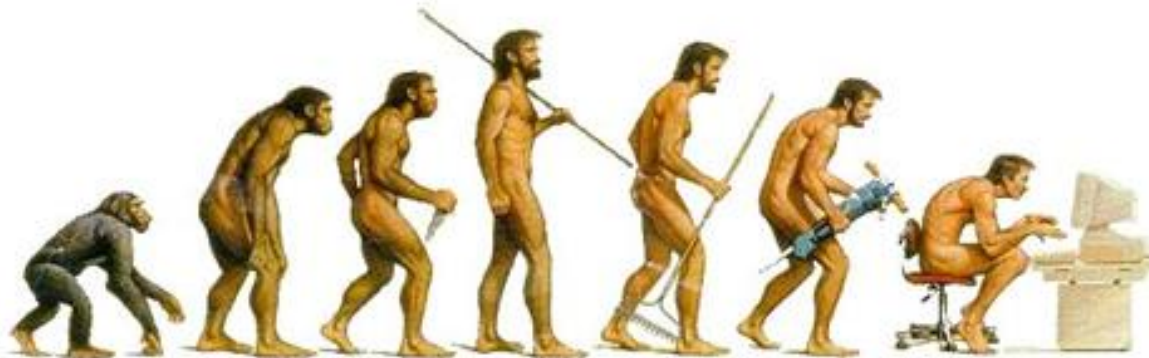
Requires computation

- Various solutions were developed for the different QA tasks



# Supporting LLMs with Tools

- How did humanity develop over time to where we are now?
- An important factor: Usage of Tools
  - Spears, the plow, electricity, computers, ...
  - ➔ Today we have many complex tools to help us solve problems, e.g. calculators, search engines, ...



[Image Source](#)

Mialon, G., et al. 2023, Augmented Language Models: a Survey. *Transactions on Machine Learning Research*.



# Example: Code Generation for Computational Problems

Question: In Fibonacci sequence, it follows the rule that each number is equal to the sum of the preceding two numbers. Assuming the first two numbers are 0 and 1, what is the 50th number in Fibonacci sequence?

The first number is 0, the second number is 1, therefore, the third number is  $0+1=1$ . The fourth number is  $1+1=2$ . The fifth number is  $1+2=3$ . The sixth number is  $2+3=5$ . The seventh number is  $3+5=8$ . The eighth number is  $5+8=13$ .  
..... (Skip 1000 tokens)  
The 50th number is 32,432,268,459.

CoT

32,432,268,459



```
length_of_fibonacci_sequence = 50
fibonacci_sequence = np.zeros(length_of_)
fibonacci_sequence[0] = 0
fibonacci_sequence[1] = 1
For i in range(3, length_of_fibonacci_sequence):
    fibonacci_sequence[i] = fibonacci_sequence[i-1] +
    fibonacci_sequence[i-2]
ans = fibonacci_sequence[-1]
```

PoT



python

12,586,269,025



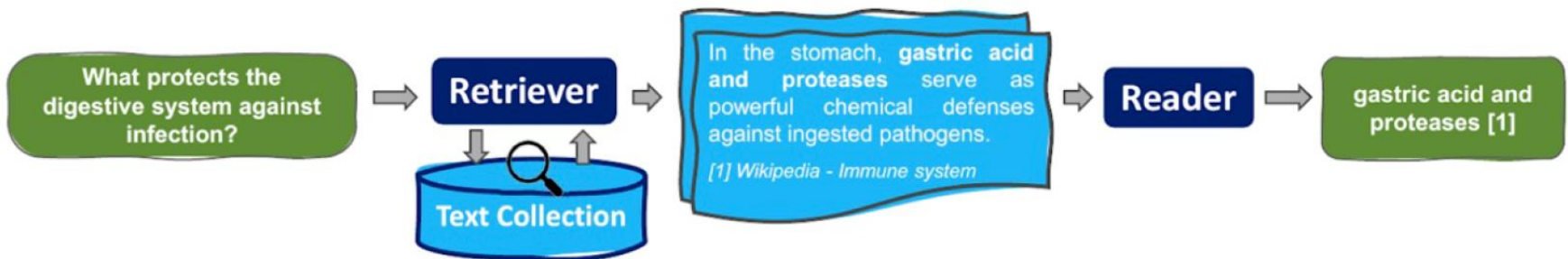
- Leverages external tool (python interpreter) to decouple computation from reasoning
- LLM can make calls to the interpreter to run generated code

Chen, W. et al., 2023 Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*.



# Retrieval-augmented Generation for Knowledge Problems

- Answer knowledge-intensive questions with
  - Extra corpora
  - A retriever (e.g. BM25, DensePassageRetrieval, etc.)



- What if there is no corpus?
  - Example Question: Who are the two candidates for the 2024 US presidential election?



# Teaching LLMs to use Tools

- Add special tokens to invoke tool calls for
  - Search engines, calculators, etc.
  - Task-specific models (translation)
  - APIs
- Unnatural format requires task/tool-specific fine-tuning

## A weather task:

how hot will it get in NYC today? |*weather* lookup re-  
gion=NYC |*result* precipitation chance: 10, high temp:  
20c, low-temp: 12c |*output* today's high will be 20C

Out of 1400 participants, 400 (or **Calculator(400 / 1400)**  
**→ 0.29]** 29%) passed the test.

The name derives from "la tortuga", the Spanish word for  
**[MT("tortuga") → turtle]** turtle.

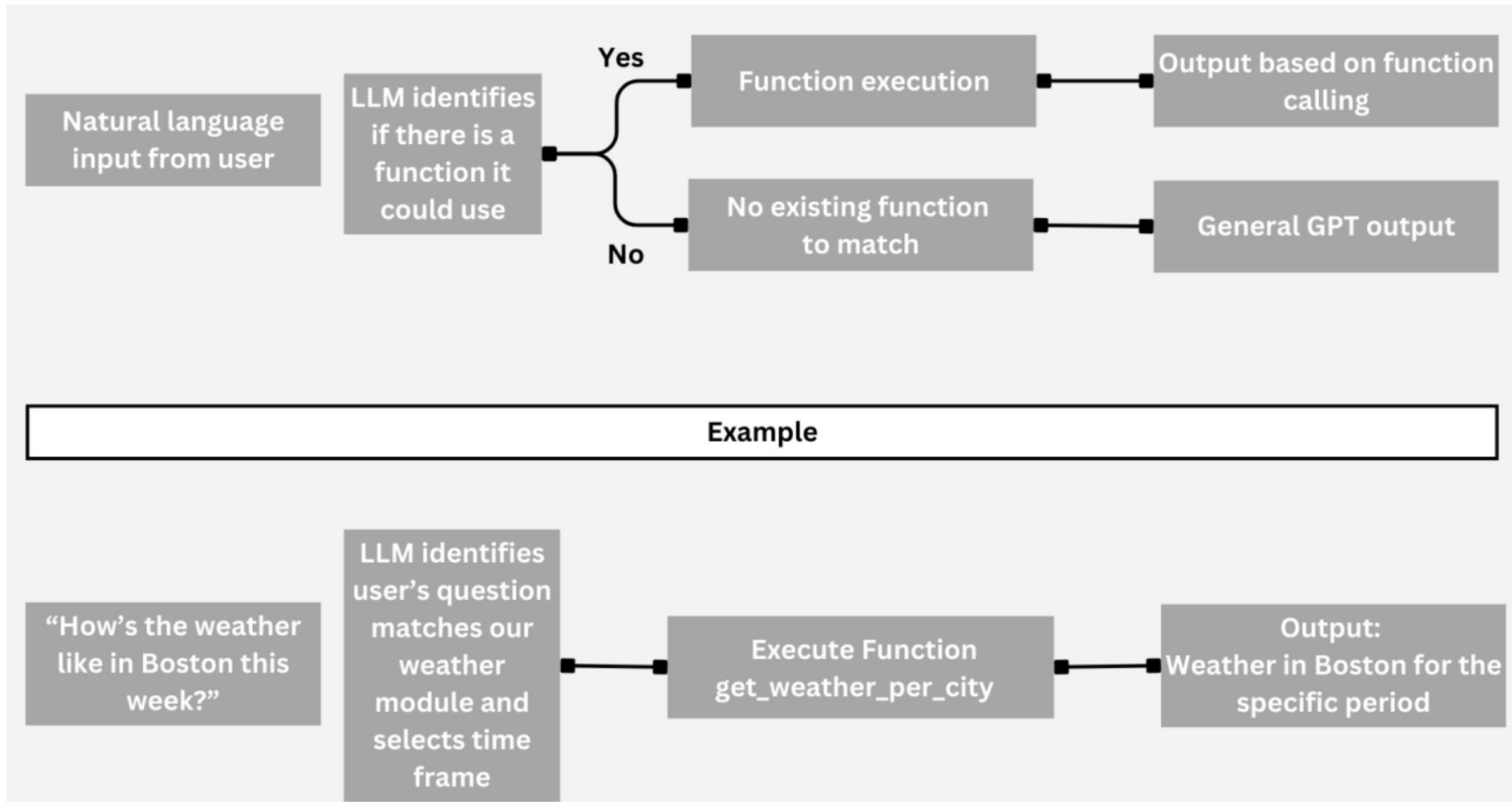
The Brown Act is California's law **[WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.]** that requires legislative bodies, like city councils, to hold their meetings open to the public.

Parisi, A., et al., 2022. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*.

Schick, T., et al., 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.



# Tool Usage: General Process

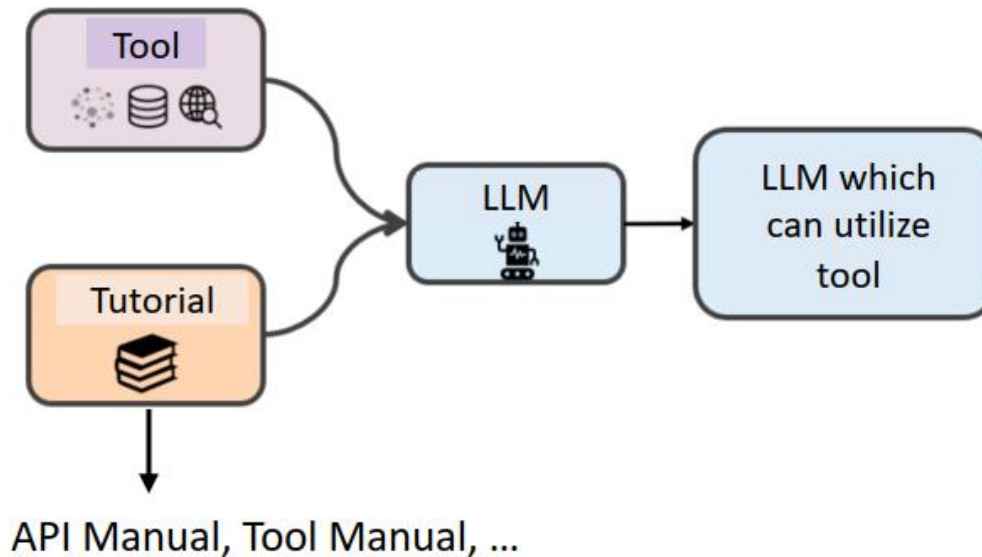


[Image Source](#)



# Tool Learning: Tutorial

- Tutorial Learning
  - Have model tuned for tool use read tool manuals (tutorials), so that it understands the functions of the tool and how to invoke them
  - Works well with powerful LLMs





# Tool Learning Prompt

**Zero-shot Prompting:** Here we provide a tool (API) "forecast\_weather(city:str, N:int)", which could forecast the weather about a city on a specific date (after N days from today). The returned information covers "temperature", "wind", and "precipitation".

Please write codes using this tool to answer the following question: "What's the average temperature in Beijing next week?"

---

**Few-shot Prompting:** We provide some examples for using a tool. Here is a tool for you to answer question:

Question: "What's the temperature in Shanghai tomorrow?"

```
return forecast_weather("Shanghai", 1) ["temperature"]
```

Question: "Will it rain in London in next two days?"

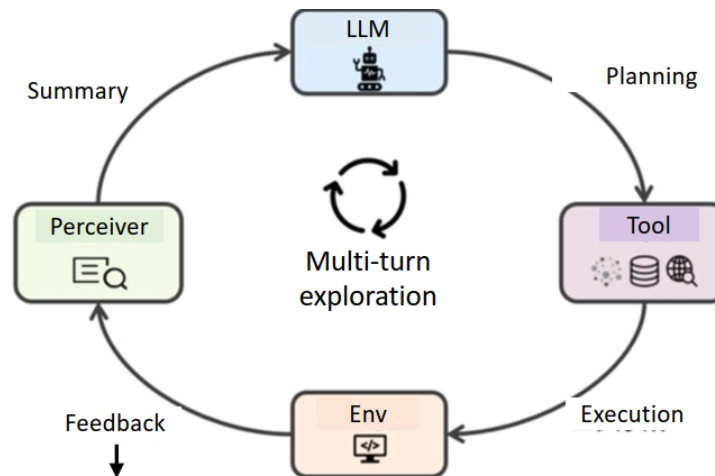
```
for i in range(2):  
    if forecast_weather("London", i+1) ["precipitation"] > 0:  
        return True  
return False
```

Question: "What's the average temperature in San Francisco next week?"



# Tool Learning: RL

- Reinforcement Learning
  - Autonomous exploration and correction of errors based on environmental feedback through reinforcement learning
  - Action space defined by tools
  - Agent learns to select appropriate tool
  - Correct action maximize reward signal

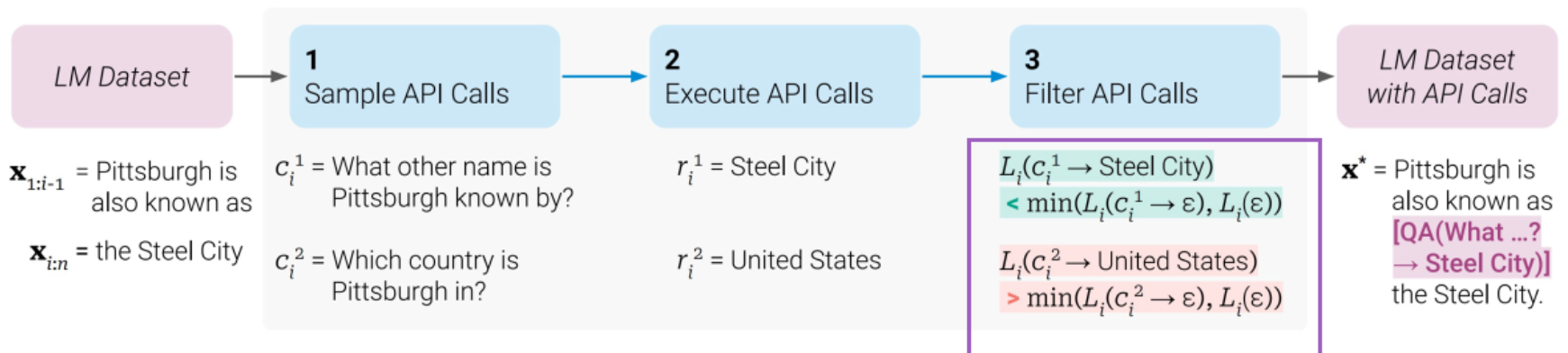


API Calling Success Rate, User Feedback ...



# Tool Learning: Self-supervised

- Self-supervised Tool Learning
  - Pre-defined tool APIs
  - Encourage models to call and execute tool APIs
  - Design self-supervised loss to evaluate tool execution helpfulness



If the tool execution reduces LM loss,  
save the instances as training data

Schick, T et al., 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.



# Early Example: WebGPT

- Supervised Learning performed at OpenAI
  - Trying to copy human behavior to use search engines
  - Supervised fine-tuning + reinforcement learning
  - Only 6000 annotated data instances

How can I train the crows in my neighborhood to bring me gifts?

---

Search results for: how to train crows to bring you gifts   Quotes

how to train crows to bring     

[How to Make Friends With Crows - PetHelpful](#)

If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make new crow friends with the crows that live in your new neighborhood.

[Gifts From Crows | Outside My Window](#)

The partial piece of apple may have been left behind when the crow was startled rather than as a gift. If the crows bring bright objects you'll know for sure that it's a gift because it's not something they eat. Brandi Williams says: May 28, 2020 at 7:19 am.

↑

↓

[1] Gifts From Crows | Outside My Window  
([www.birdsoutsidemymwindow.org](http://www.birdsoutsidemymwindow.org))

Many animals give gifts to members of their own species but crows and other corvids are the only ones known to give gifts to humans.

---

Number of quote tokens left: 463

Number of actions left: 96

◆Question  
How can I train the crows in my neighborhood to bring me gifts?

◆Quotes  
From Gifts From Crows | Outside My Window ([www.birdsoutsidemymwindow.org](http://www.birdsoutsidemymwindow.org))  
> Many animals give gifts to members of their own species but crows and other corvids are the only ones known to give gifts to humans.

◆Past actions  
Search how to train crows to bring you gifts  
Click Gifts From Crows | Outside My Window [www.birdsoutsidemymwindow.org](http://www.birdsoutsidemymwindow.org)  
Quote  
Back

◆Title  
Search results for: how to train crows to bring you gifts

◆Scrollbar: 0 - 11

◆Text  
[0]How to Make Friends With Crows - PetHelpful|pethelpful.com|  
If you did this a few times, your crows would learn your new place, but as I said, I'm not sure if they will follow or visit you there since it's probably not in their territory. The other option is simply to make new crow friends with the crows that live in your new neighborhood.

[1]Gifts From Crows | Outside My Window|www.birdsoutsidemymwindow.org|  
The partial piece of apple may have been left behind when the crow was startled rather than as a gift. If the crows bring bright objects you'll know for sure that it's a gift because it's not something they eat.  
Brandi Williams says: May 28, 2020 at 7:19 am.

◆Actions left: 96

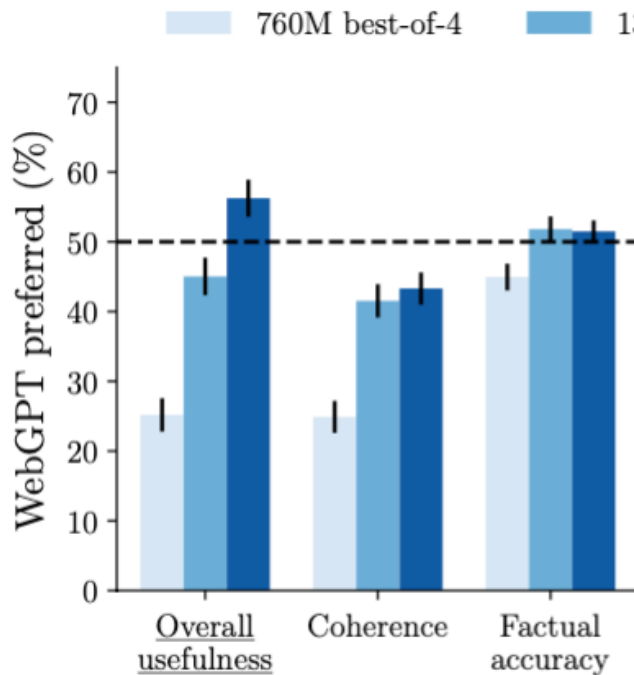
◆Next action

Nakano, R., et al., 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

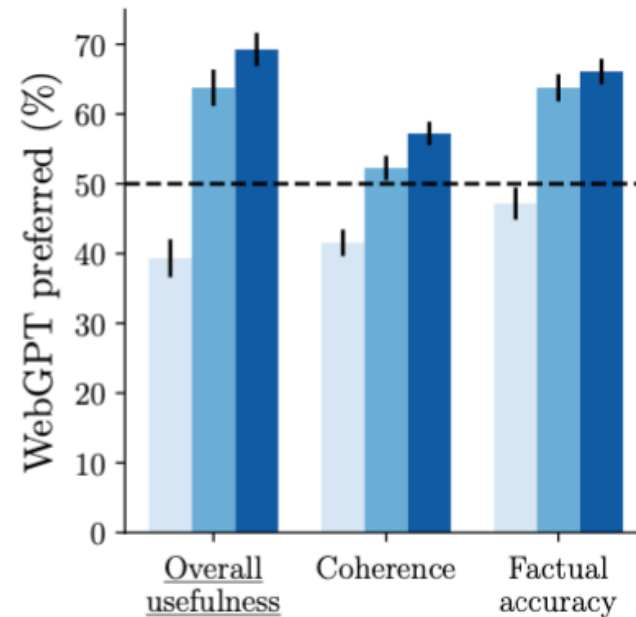


# Early Example: WebGPT

- Excellent performance in long-form QA, even surpassing human experts sometimes



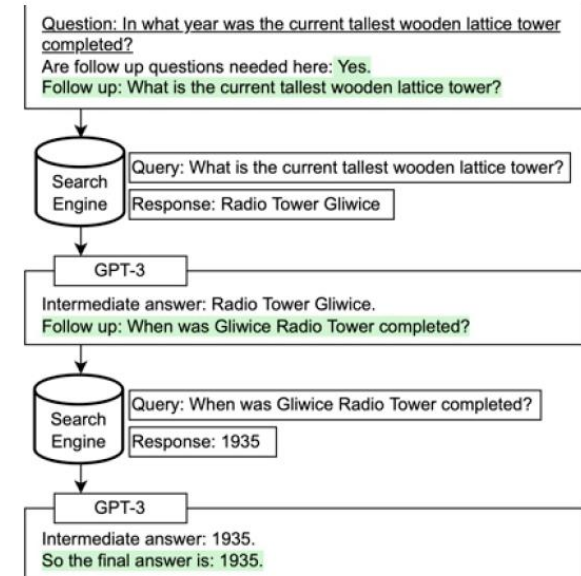
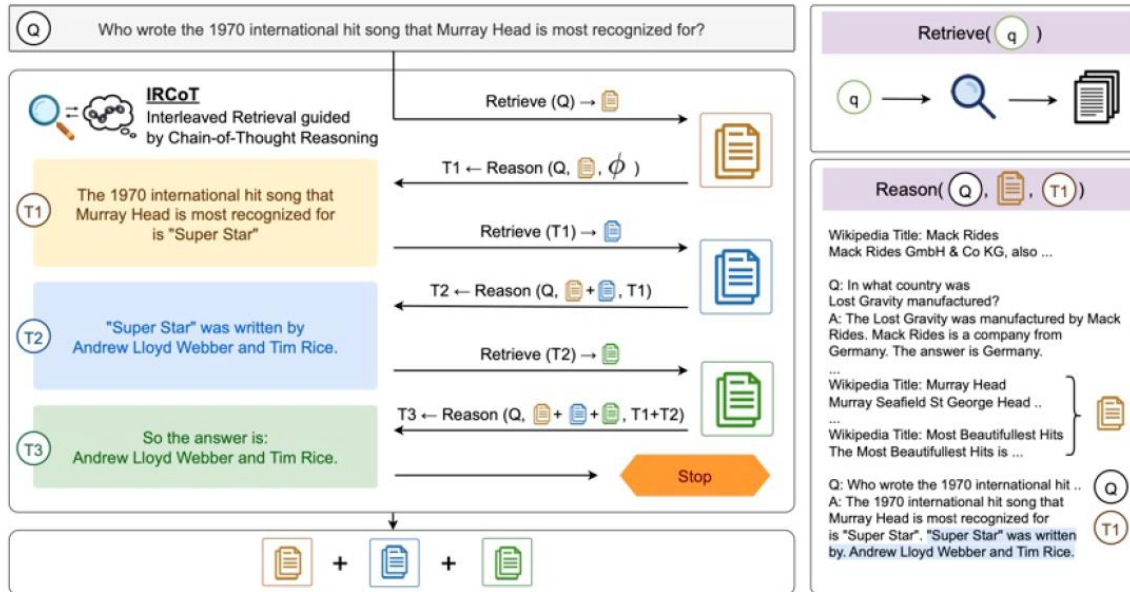
(a) WebGPT vs. human demonstrations.



(b) WebGPT vs. ELI5 reference answers.



# What if Both External Knowledge and Reasoning are needed?



- Some methods combine tool use/RAG and reasoning methods for specific tasks

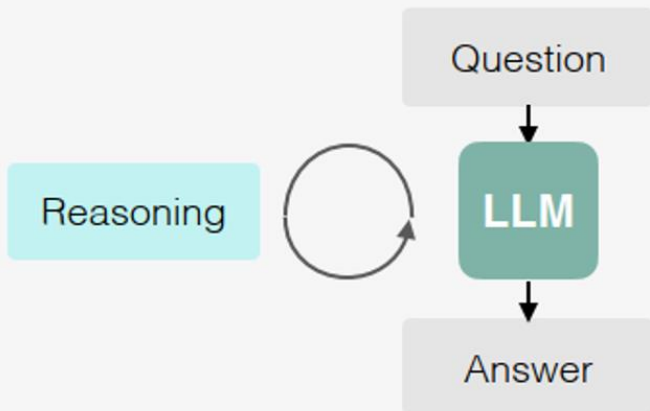
Trivedi, H., et al., 2023, July. Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10014-10037).

Press, O., et al., 2023, December. Measuring and Narrowing the Compositionality Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5687-5711).



# Reasoning OR Acting

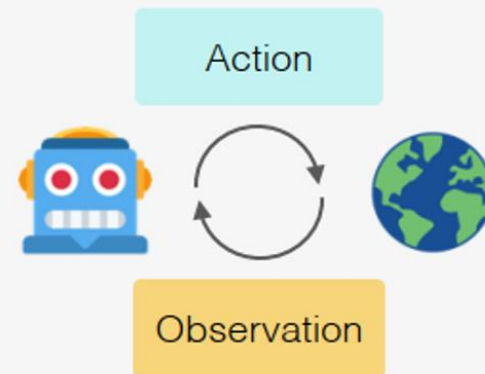
## Reasoning (update internal belief)



Flexible and general to  
augment test-time compute

Lack of external knowledge and tools

## Acting (obtain external feedback)



Lack of reasoning

Flexible and general to  
augment knowledge,  
computation, feedback, etc.

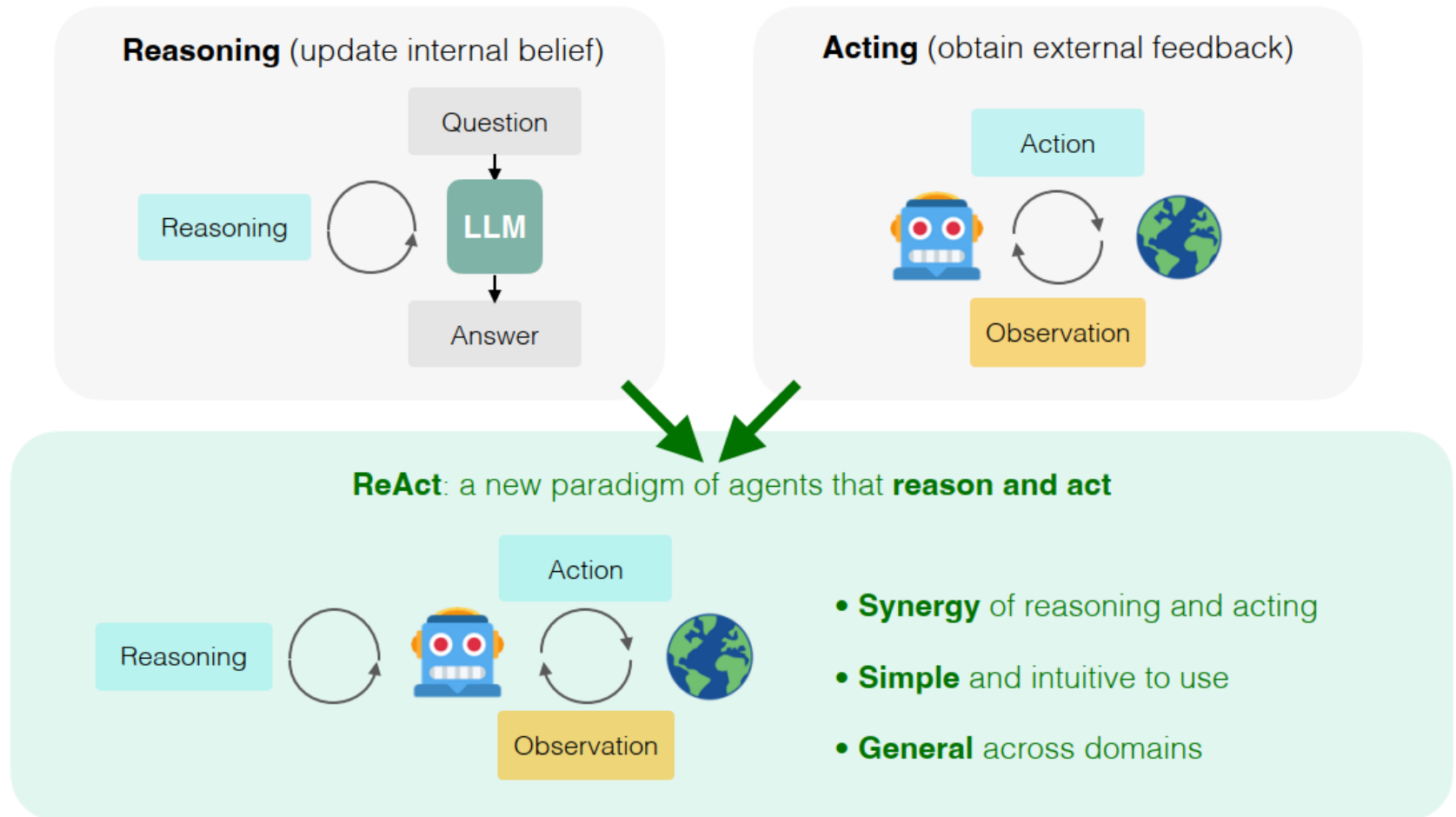


# Outline

- Recap: Prompt Engineering and Efficient Adaptation
- What is an Agent?
- Tool Usage for LLMs
- **The ReAct Paradigm**
- Unified Framework for LLM Agents
- Evaluating Agents



# The ReAct Paradigm



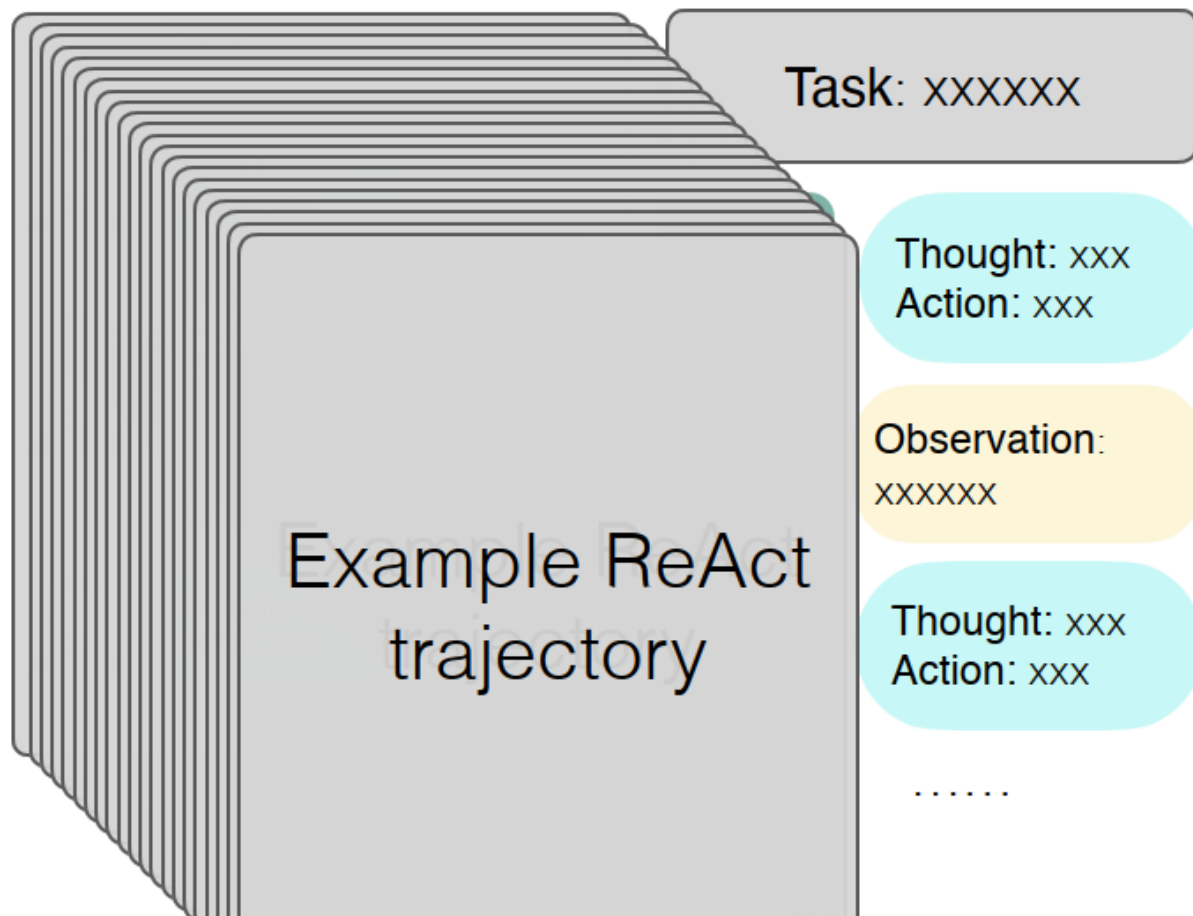
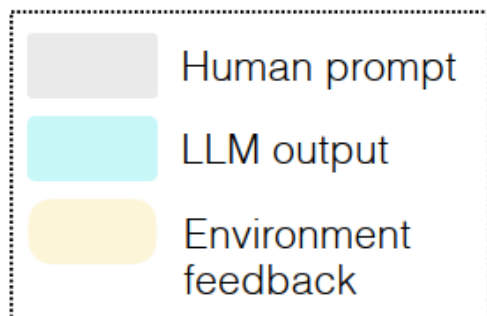
Yao, S., et al., 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*.



# ReAct is Simple and Intuitive to Use

ReAct supports

- One-shot prompting
- Few-shot prompting
- Fine-tuning





# Zero-shot ReAct Prompt

You are an agent that answers questions by using two actions:

- search[query]: Google search the query. You can also use it to calculate math.
- finish[answer]: return the answer

Your generation should have the following format:

**Thought:** to analyze the observation and inform the action.

**Action:** your action

**Question:** If I have seven trillion dollars today, can I buy Apple, Nvidia, and Microsoft? If not, how much more money do I need?



# Zero-shot ReAct Prompt

LLM

(GPT-4)

**Thought:** I need to find the current market capitalization of Apple, Nvidia, and Microsoft to understand how much it would cost to buy all of them.

**Action:** search["current market capitalization of Apple, Nvidia, and Microsoft"]

**Observation:** "Nvidia's market cap stood at \$1.832 trillion as of about 10 a.m. ET Wednesday, while Alphabet's was at \$1.814 trillion, bringing Nvidia behind just Microsoft (MSFT) at \$3.015 trillion and Apple (AAPL) at \$2.836 trillion."




(Mar 2024)

**Thought:** Now that I have the market cap for all companies, let's add them together and see if \$7 trillion is enough.

**Action:** search["1.832 trillion + 3.015 trillion + 2.836 trillion"]

**Observation:** 7683000000000

**Thought:** The amount of money needed in addition to the \$7 trillion is \$683 billion.

**Action:** finish["You need an additional \$683 billion to be able to buy Apple, Nvidia, and Microsoft."] 



# Zero-shot ReAct Prompt

LLM

(GPT-4)

**Thought:** I need to find the current market capitalization of Apple, Nvidia, and Microsoft to understand how much it would cost to buy all of them.

**Action:** search["current market capitalization of Apple, Nvidia, and Microsoft"]

**Observation:** nothing found!

(My adversarial observation)

**Thought:** The search did not give any results. I will need to search for each company's market capitalization individually.

**Action:** search["current market capitalization of Apple"]

**Observation:** 'price': 180.75

(My adversarial observation)

**Thought:** This is the current price of Apple's stock, not its market capitalization. The market cap is calculated by multiplying the share price by the number of outstanding shares. Thus, I need to find the total number of Apple's outstanding shares to calculate its market cap.

**Action:** search["total number of Apple's outstanding shares"]



(Continues to solve the task)

- Synergy
  - Acting supports reasoning
  - Reasoning guides acting



# Converting Tasks to Text



You are in the middle of a room. Looking quickly around you, you see a drawer 2, a shelf 5, a drawer 1, a shelf 4, a sidetable 1, a drawer 5, a shelf 6, a shelf 1, a shelf 9, a cabinet 2, a sofa 1, a cabinet 1, a shelf 3, a cabinet 3, a drawer 3, a shelf 11, a shelf 2, a shelf 10, a dresser 1, a shelf 12, a garbagecan 1, a armchair 1, a cabinet 4, a shelf 7, a shelf 8, a safe 1, and a drawer 4.

Your task is to: *put some vase in safe.*

> go to shelf 6

You arrive at loc 4. On the shelf 6, you see a vase 2.

> take vase 2 from shelf 6

You pick up the vase 2 from the shelf 6.

> go to safe 1

You arrive at loc 3. The safe 1 is closed.

> open safe 1

You open the safe 1. The safe 1 is open. In it, you see a keychain 3.

> put vase 2 in/on safe 1

You won!

ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks.

ALFWorld: Aligning Text and Embodied Environments for Interactive Learning.

- Many tasks can be turned into natural language for LLM agents
- “LLM grounding”: Supplementing the LLM with use-case specific information, e.g a data store that is part of a RAG system

Brohan, A., et al., 2023, March. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning* (pp. 287-318). PMLR.

Huang, W., et al., 2023, March. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *Conference on Robot Learning* (pp. 1769-1782). PMLR.



# Acting without Reasoning

You are in the middle of a room. Looking quickly around you, you see a cabinet 6, a cabinet 1, a coffee machine 1, a countertop 3, a stove burner 1, and a toaster 1.

**Your task is to:** Put some pepper shaker on a drawer.

## Act only

**Act 1:** Go to drawer 1

**Obs 1:** The drawer 1 is closed.

**Act 2:** Open drawer 1

**Obs 2:** You open the drawer 1. The drawer 1 is open. In it, you see a dish sponge 2, and a spoon 1.

**Act 3:** Go to sinkbasin 1

**Obs 3:** On the sinkbasin 1, you see a dish sponge 3, a spatula 1, and a spoon 2.

**Act 4:** Take peppershaker 1 from sinkbasin 1

**Obs 4:** Nothing happens.

**Act 5:** Take peppershaker 1 from sinkbasin 1

**Obs 5:** Nothing happens.

.....



- Cannot explore systematically or incorporate feedback



# ReAct Enables Systematic Exploration

You are in the middle of a room. Looking quickly around you, you see a cabinet 6, a cabinet 1, a coffee machine 1, a countertop 3, a stove burner 1, and a toaster 1.

**Your task is to:** Put some pepper shaker on a drawer.

## ReAct

**Act 1:** **Think**[First I need to find a peppe shaker...more likely to appear in cabinets (1-6),  
countertops (1-3), ... ]

**Act 2:** Go to cabinet 1

**Obs 2:** On the cabinet 1, you see a vase 2.

(Here the agent goes to cabinet 1, 2, 3, 4, 5, 6, then countertop 1 and 2)

**Act 9:** Go to countertop 3

**Obs 9:** On the countertop 3, you see a apple 1, a bread 1, ..... a pepper shaker 1, and a vase 3.

**Act 10:** Take pepper shaker 1 from countertop 3

**Obs 10:** You pick up the pepper shaker 1 from the countertop 3.

**Act 11:** **Think**[Now I find a pepper shaker 1. Next, I need to put it in/on drawer 1.]

**Act 12:** Go to drawer 1

**Obs 12:** Drawer 1 is closed.

**Act 13:** Open drawer 1

**Obs 13:** You open Drawer 1 ...

**Act 14:** Put pepper shaker 1 in/on drawer 1

**Obs 14:** You put pepper shaker 1 in/on the drawer 1.





# ReAct is general and effective

(NLP tasks)

(RL tasks)

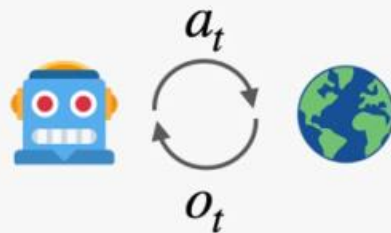
PaLM-540B	HotpotQA (QA)	FEVER (fact check)	ALFWorld (Text game)
Reason	29.4	56.3	N/A
Act	25.7	58.9	45
ReAct	<b>35.1</b>	<b>64.6</b>	<b>71</b>

Yao, S., et al., 2023, ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations*.



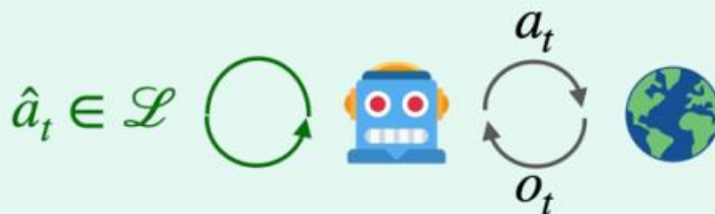
# ReAct vs. Traditional Agents

**Traditional agents:** action space  $A$  defined by the environment



- **External feedback**  $o_t$
- Agent context  $c_t = (o_1, a_1, o_2, a_2, \dots, o_t)$
- Agent action  $a_t \sim \pi(a | c_t) \in A$

**ReAct:** action space  $\hat{A} = A \cup \mathcal{L}$  augmented by reasoning



- $\hat{a}_t \in \mathcal{L}$  can be any language sequence
- Agent context  $c_{t+1} = (c_t, \hat{a}_t, a_t, o_{t+1})$
- $\hat{a}_t \in \mathcal{L}$  only updates **internal context**



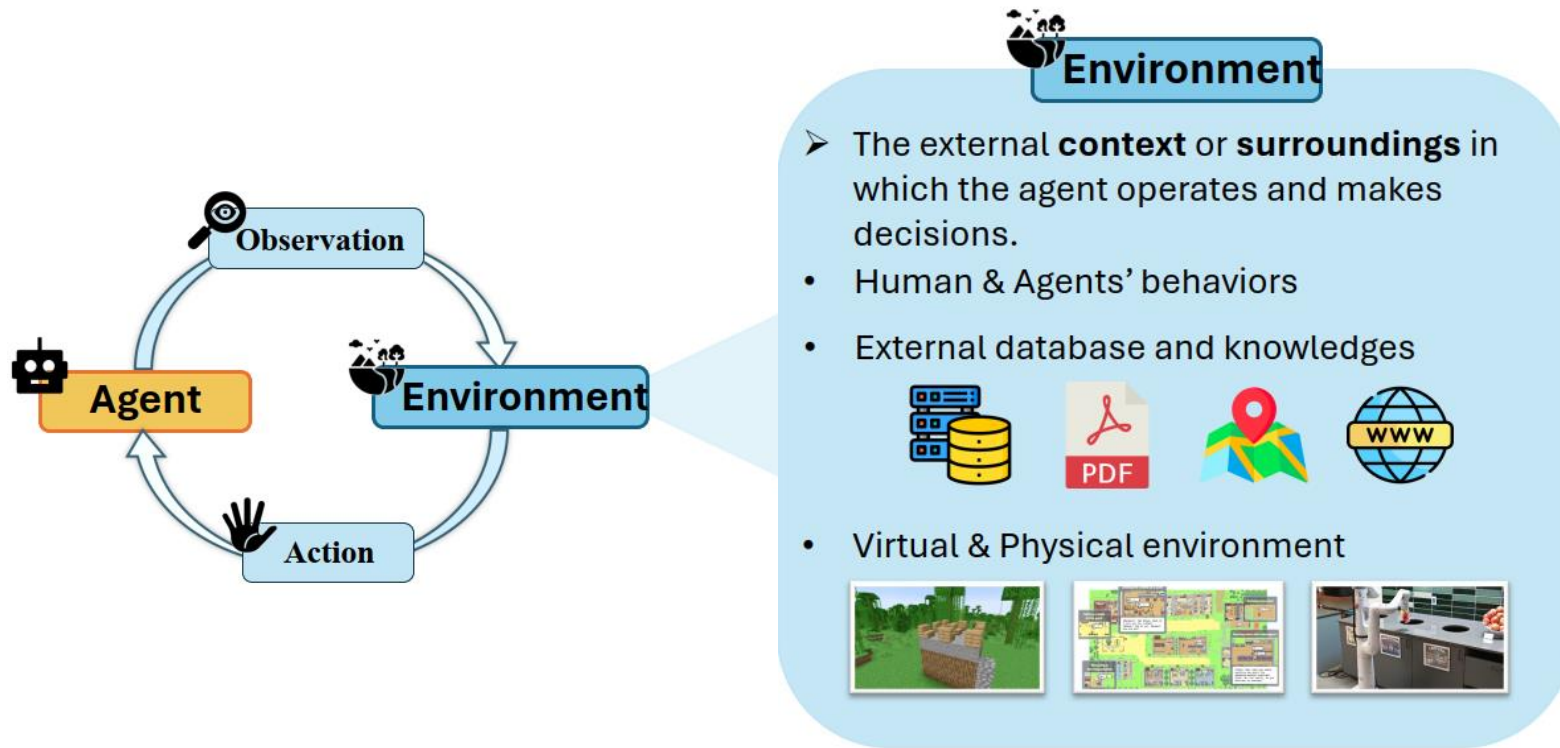
# Outline

- Recap: Prompt Engineering and Efficient Adaptation
- What is an Agent?
- Tool Usage for LLMs
- The ReAct Paradigm
- **Unified Framework for LLM Agents**
- Evaluating Agents



# Unified Framework for LLM-powered Agents

- LLMs pave the way for the use of AI agents to simulate users and other entities, as well as their interactions



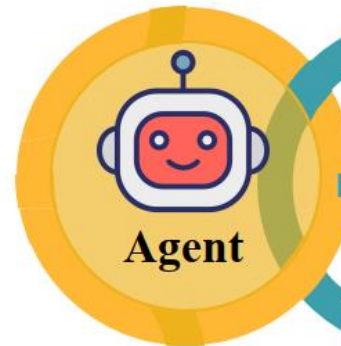
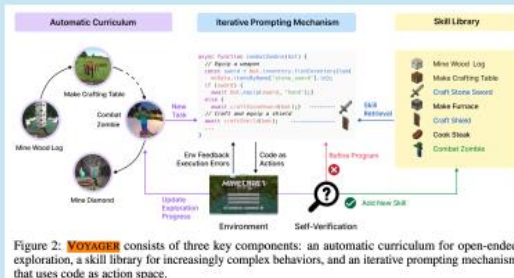


# Observation and Action



## Action

- call external **APIs** for extra information that is missing from the model weights (often hard to change after pre-training):  
**Generating multimodal outputs; Embodied Action; Learning tools; Using tools; Making tools; .....**



Observation

Environment

Action

## Multi-modal Perception

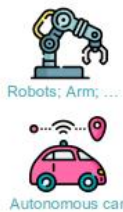


## Broader Action Spaces

### Multimodal Output



### Embodiment

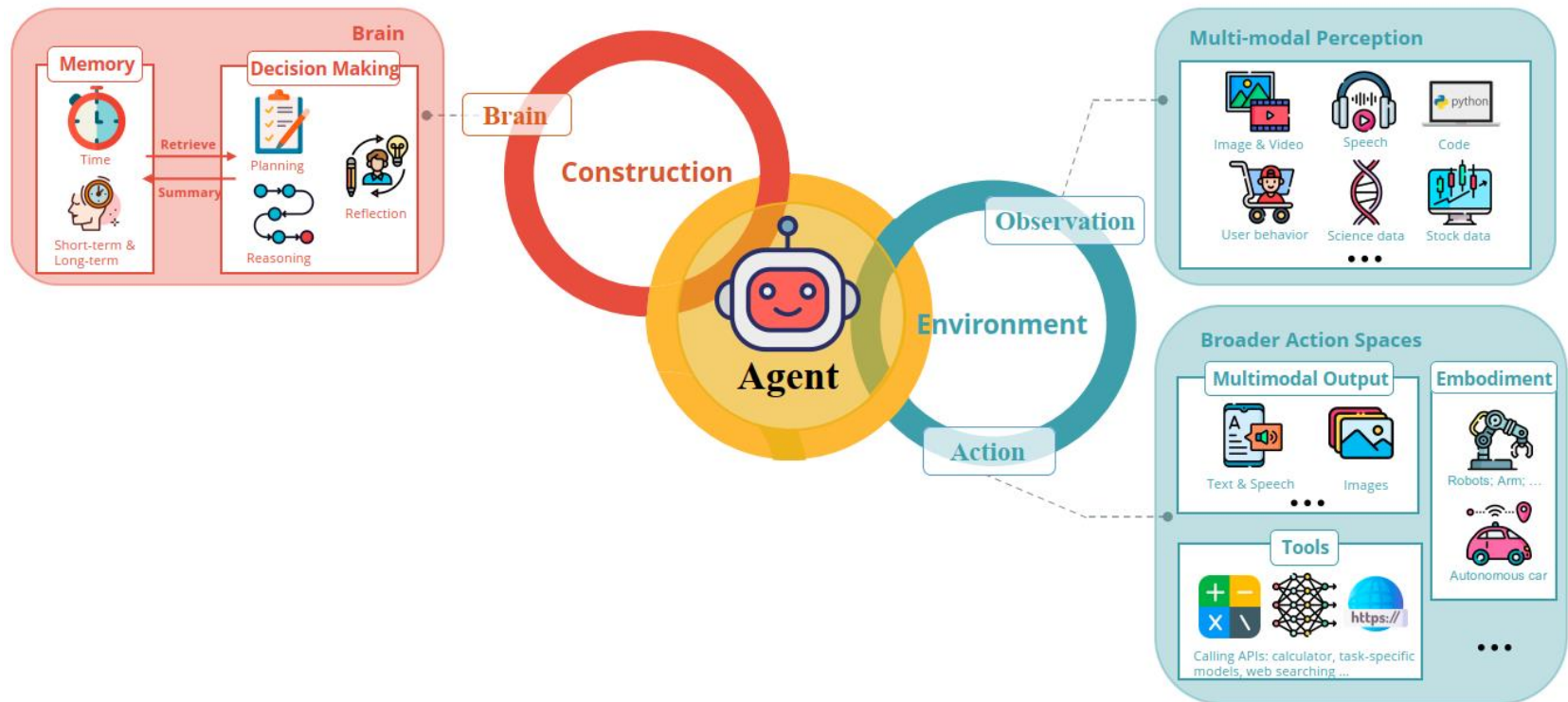


### Tools





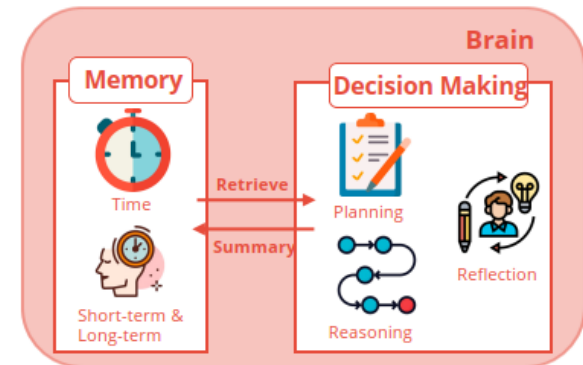
# The “Brain”





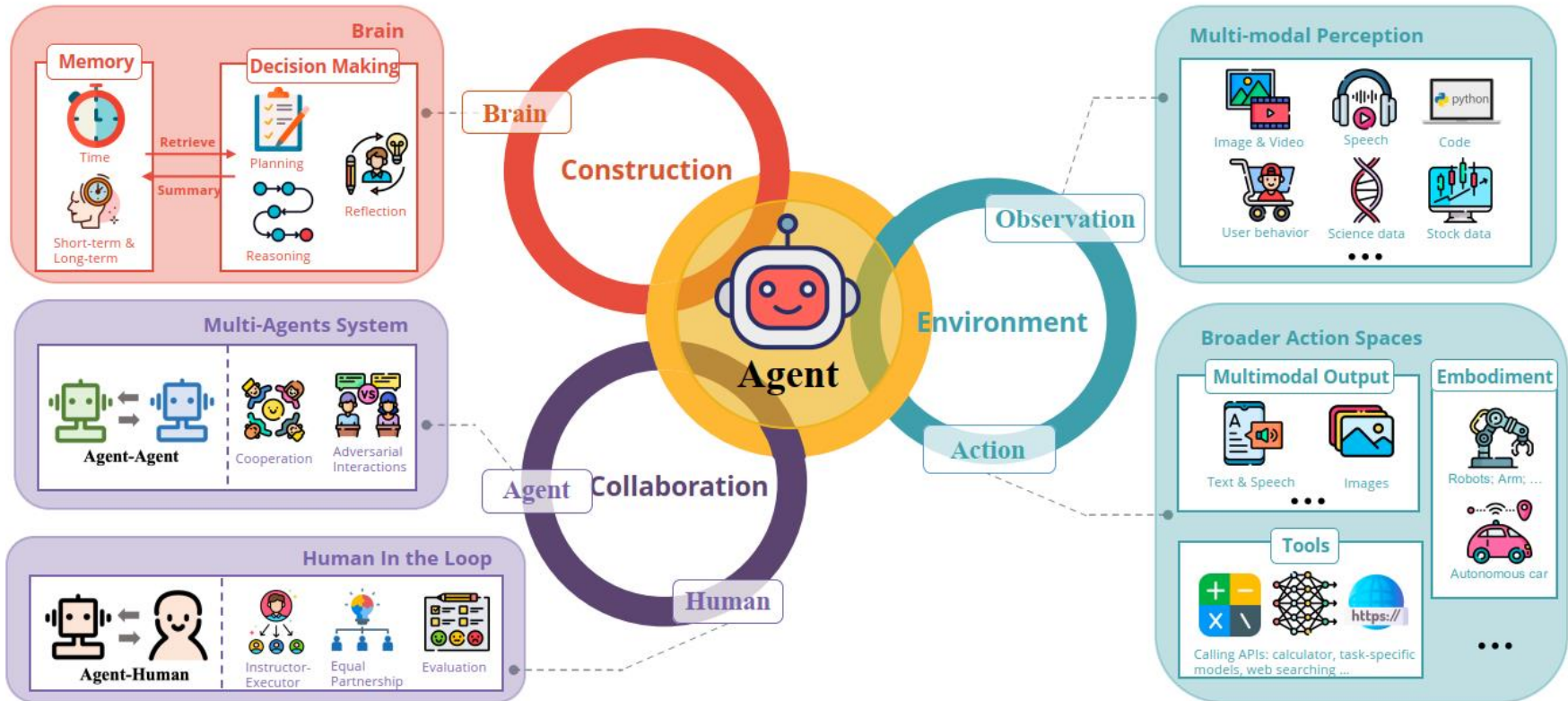
# The “Brain”

- **Memory:** stores sequences of agent’s past observations, thoughts and actions
  - Long-term and short-term memory
  - Long-term memory is abstract
  - Used to retrieve relevant past memory
- **Decision Making Process:**
  - **Planning:** Subgoal and decomposition – Break down large tasks into smaller, manageable subgoals, enabling efficient handling of complex tasks
  - **Reasoning:** Self-criticism and self-reflection over past actions, learn from mistakes and refine for future steps
- Personalized memory and reasoning lead to **diversity** and **independence** of AI Agents.





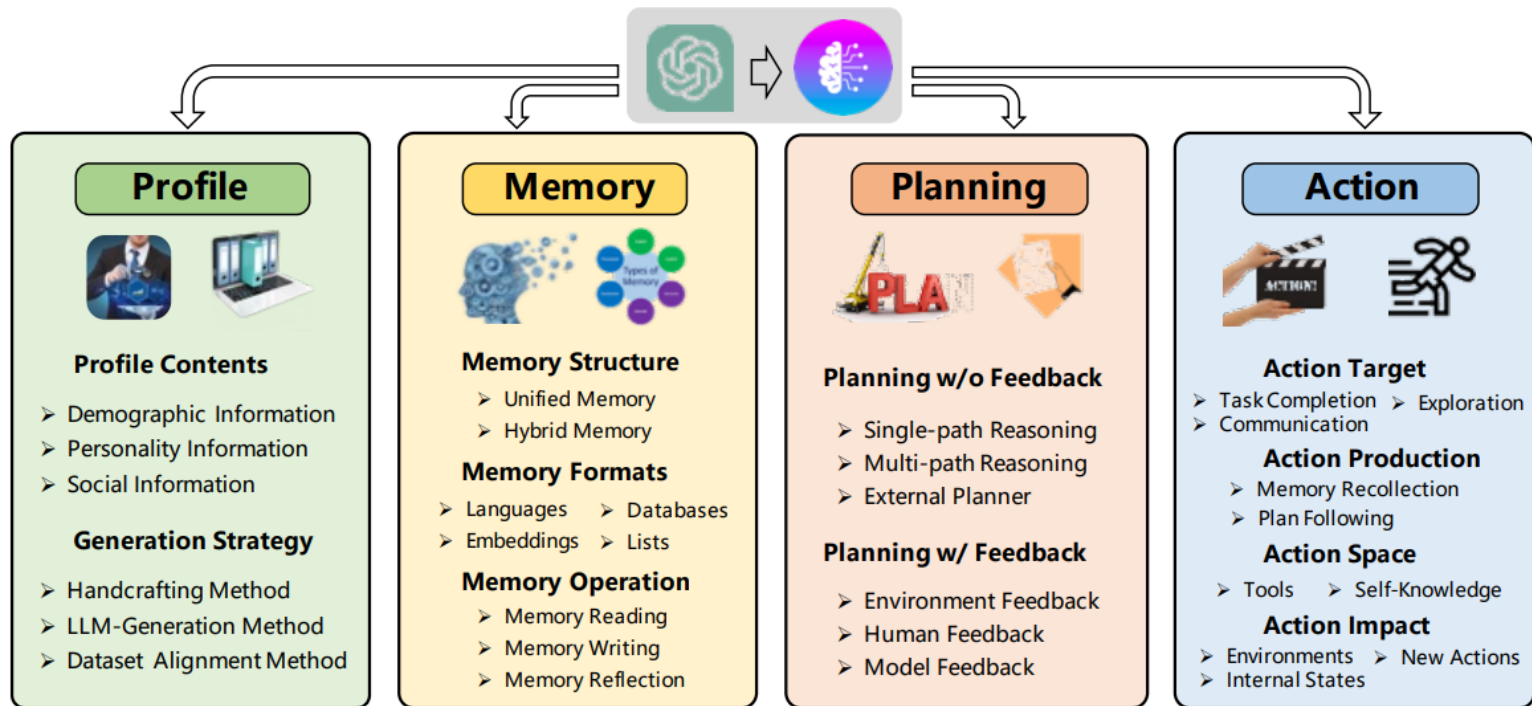
# Collaboration



- Diverse Agents interact with each other to solve problems in fully autonomous systems
- Human-in-the-loop in cooperative systems



# Unified Framework for LLM Agents





# Example: Agent creation with OpenAI

## Name

SearchAgent

## Description

This Agent is able to make use of web browsing to help users find correct answers.

## Instructions

You are highly skilled in the usage of web search engines and know how to perfectly frame search queries to support a user in answering any questions they might have correctly.

## Conversation starters

×

## Knowledge

If you upload files under Knowledge, conversations with your GPT may include file contents. Files can be downloaded when Code Interpreter is enabled

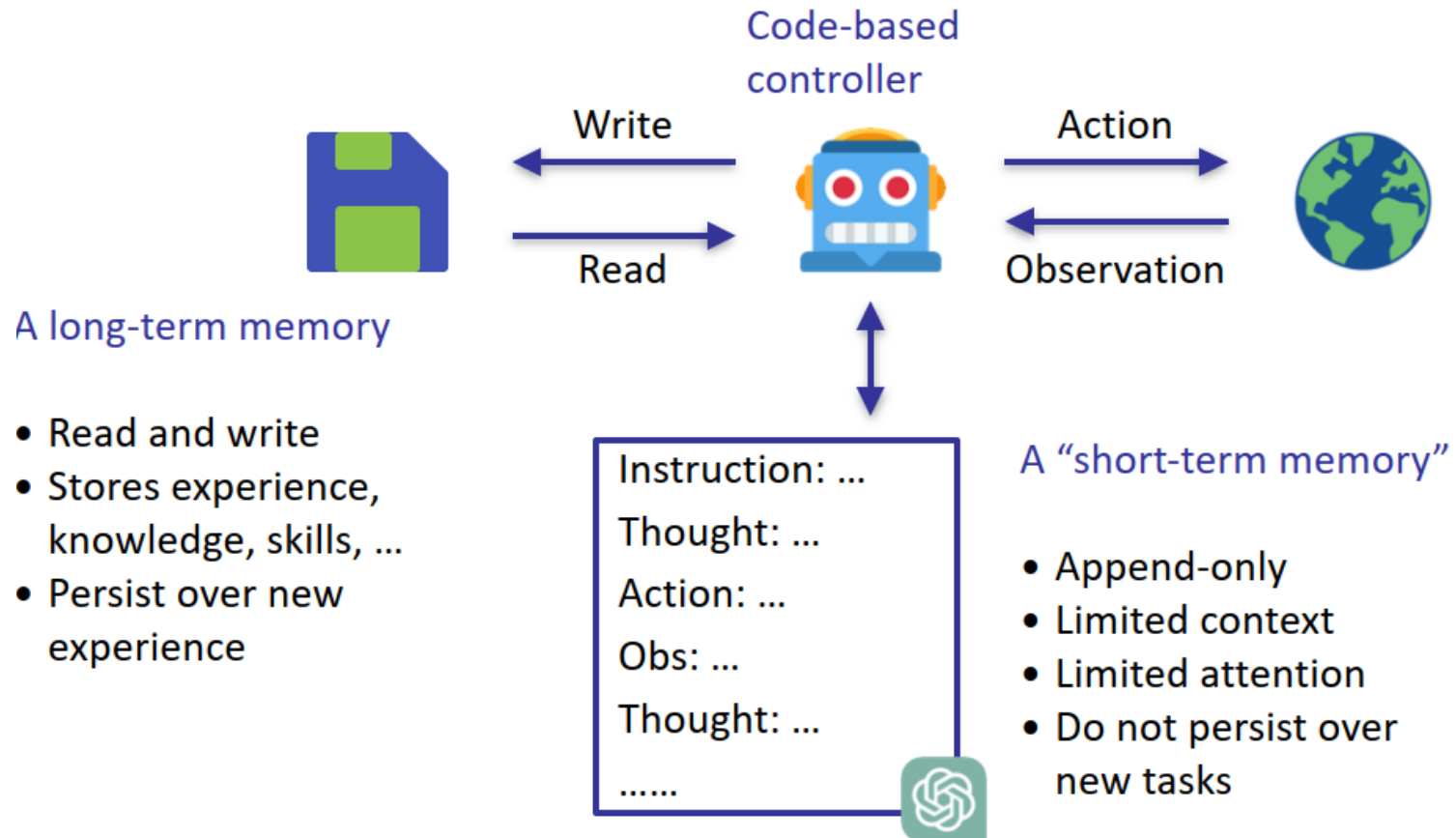
Upload files

## Capabilities

- ☒ Web Browsing
- ☐ DALL·E Image Generation
- ☐ Code Interpreter & Data Analysis ?

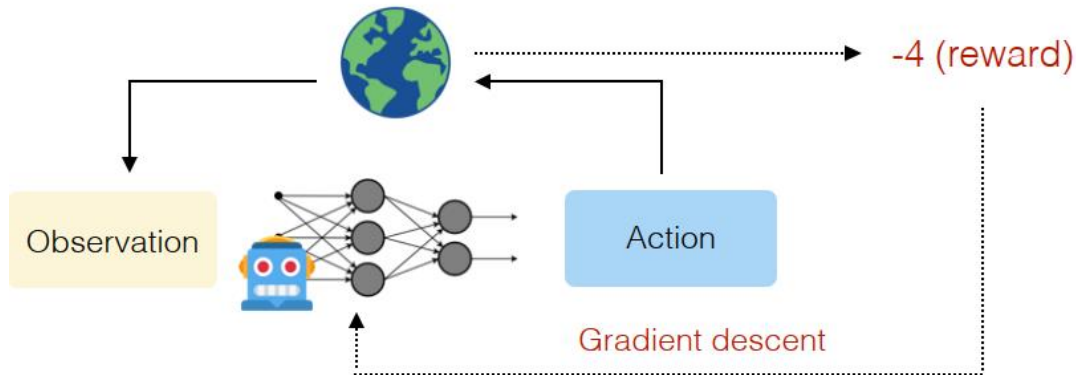


# Example: Long-term Memory



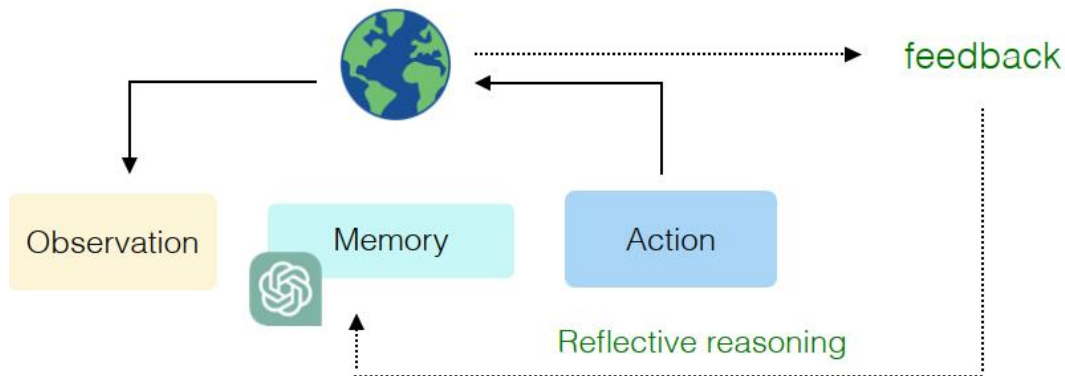


# Long-term Memory for Reflexion



## Traditional RL

- Learn via **scalar** reward (sparse signal)
- Learn by updating **weights** (credit assignment)

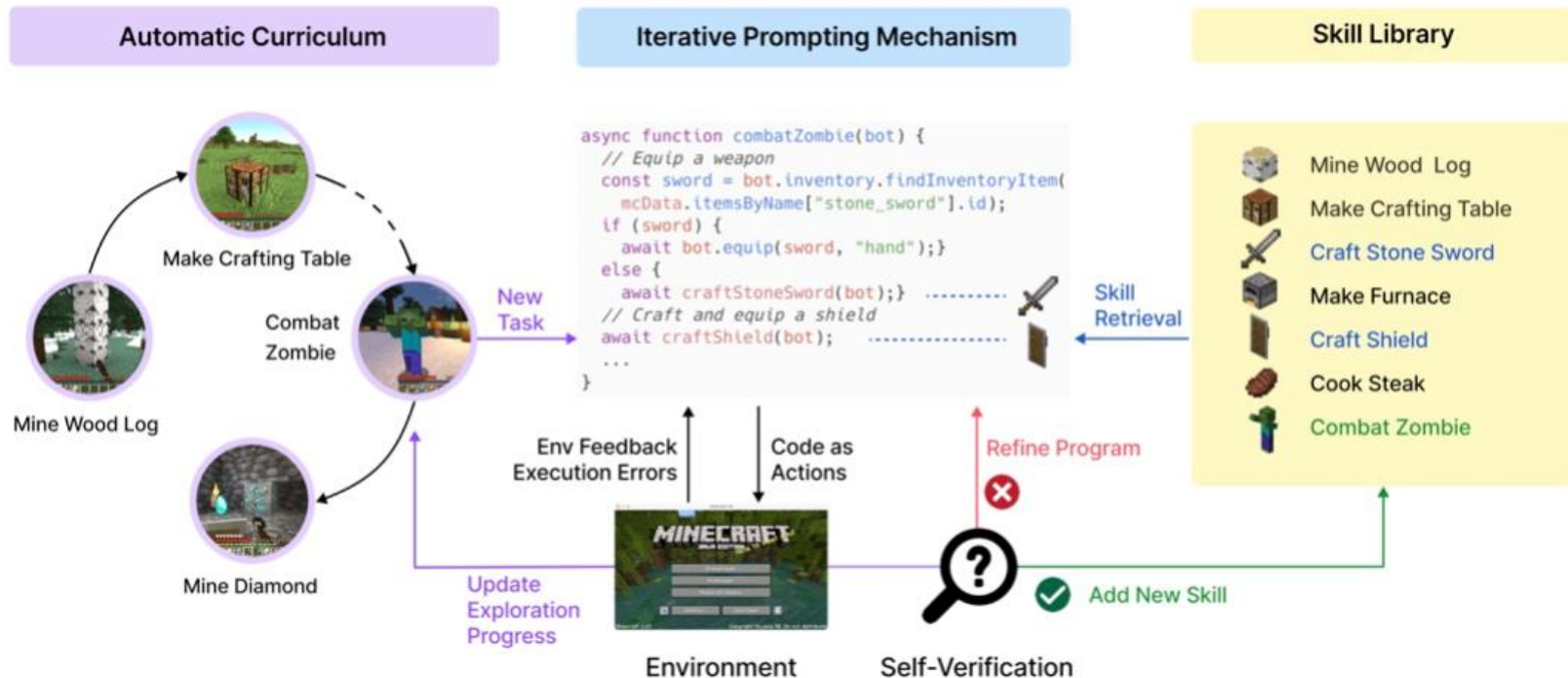


## Reflexion: "Verbal" RL

- Learn via **text** feedback
- Learn by updating **language** (a long-term memory of task knowledge)



# Example: Voyager - Procedural Memory of Skills

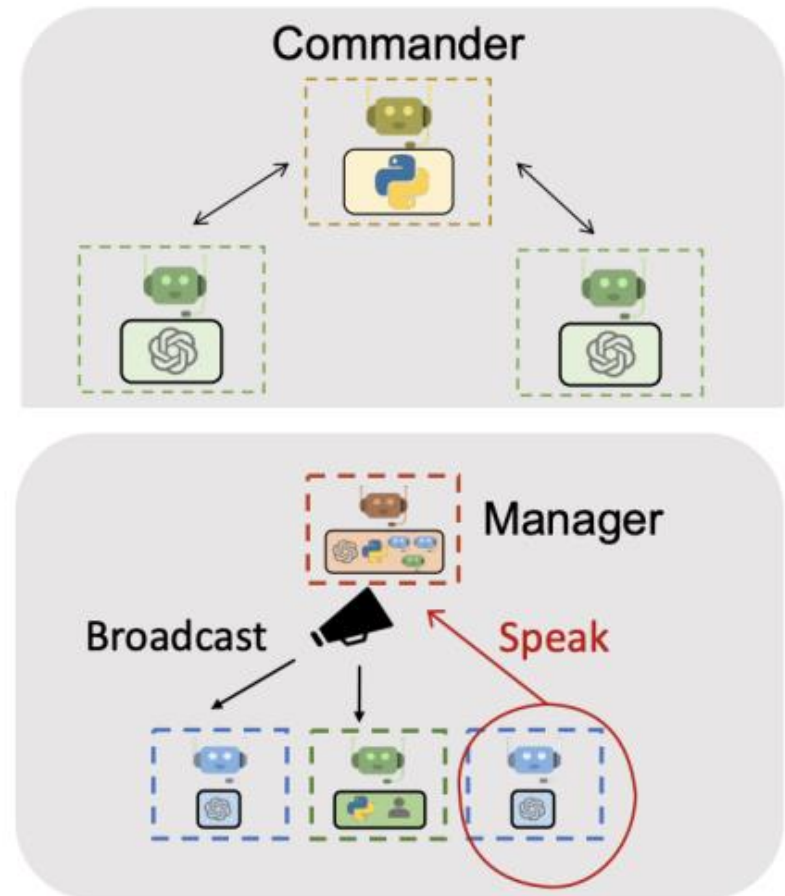


Wang, G., et al., 2024. Voyager: An Open-Ended Embodied Agent with Large Language Models. *Transactions on Machine Learning Research*.



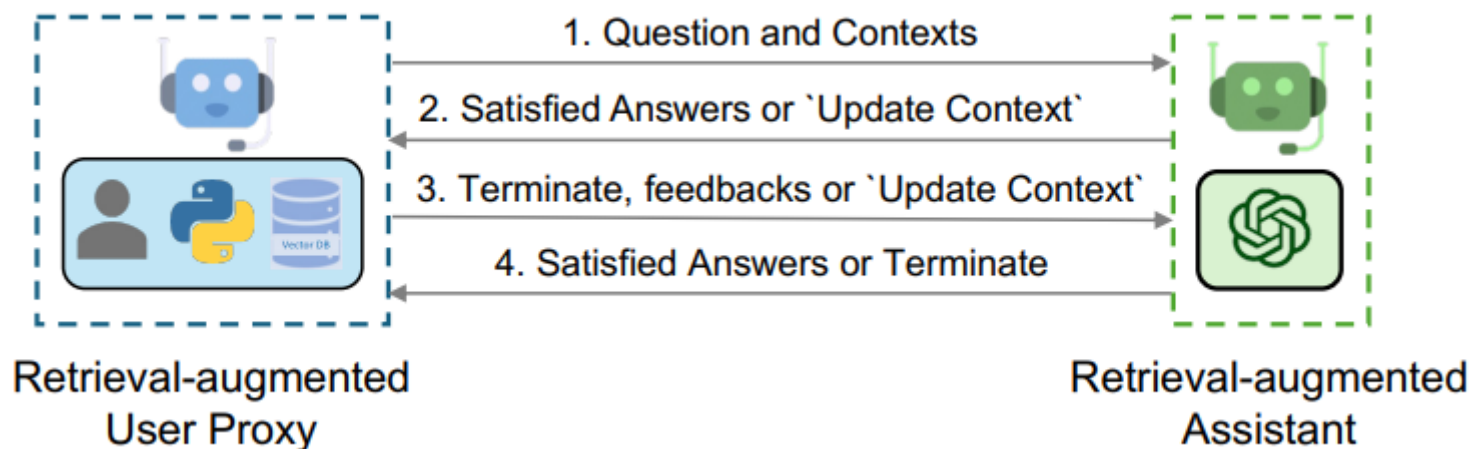
# Multi-Agent Orchestration

- Usually a “Manager” or “Commander” for orchestrating many agents
- Context may be shared or isolated
- Cooperative vs. competitive environments
- Centralized vs. decentralized communication
- Human intervention vs. full automation





# Example: Retrieval-Augmented QA

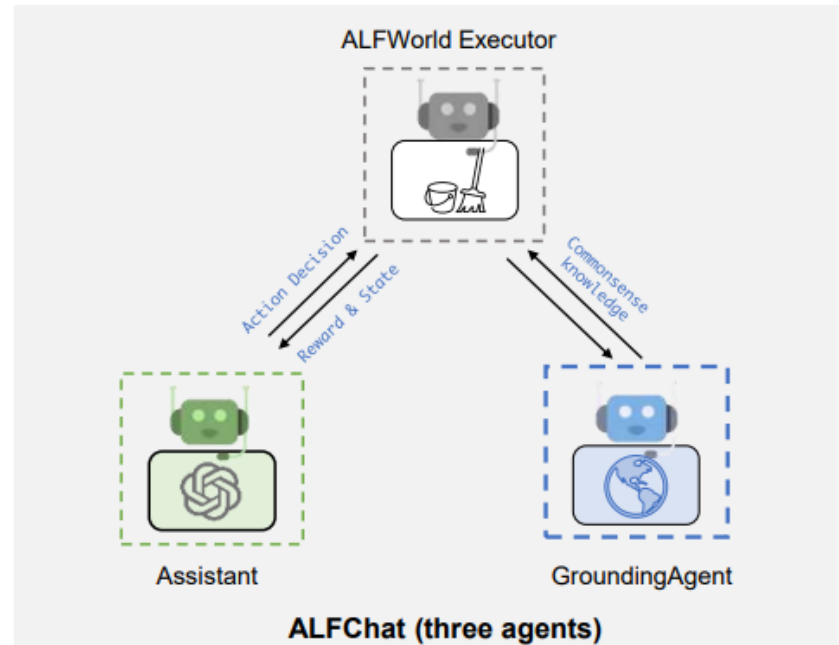
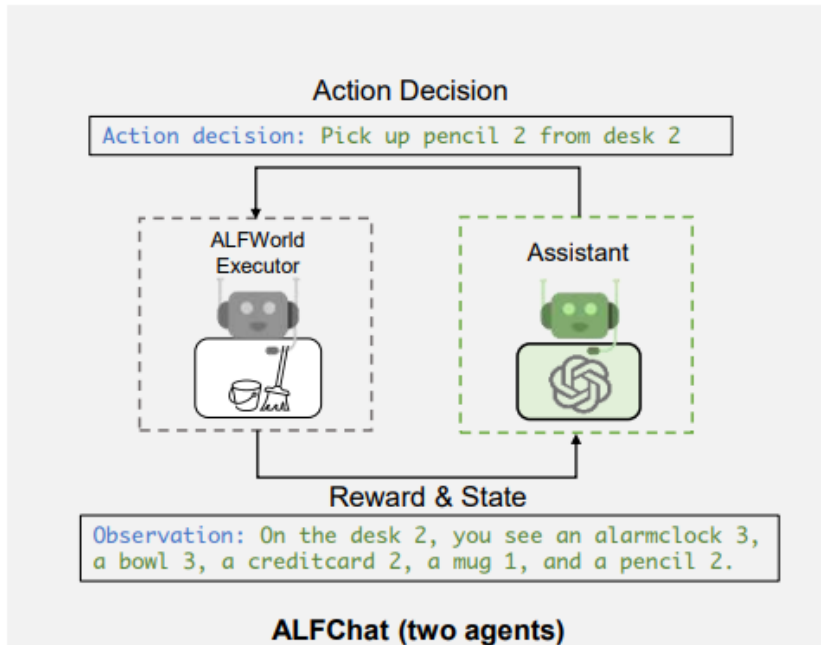


- Two agents
- User Proxy processes documents into vectorstore
- User Question and relevant context passed to assistant that generates answer
- Conversation continues until satisfactory answer

Wu, Q., et al., 2024, AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.



# Example: Decision Making

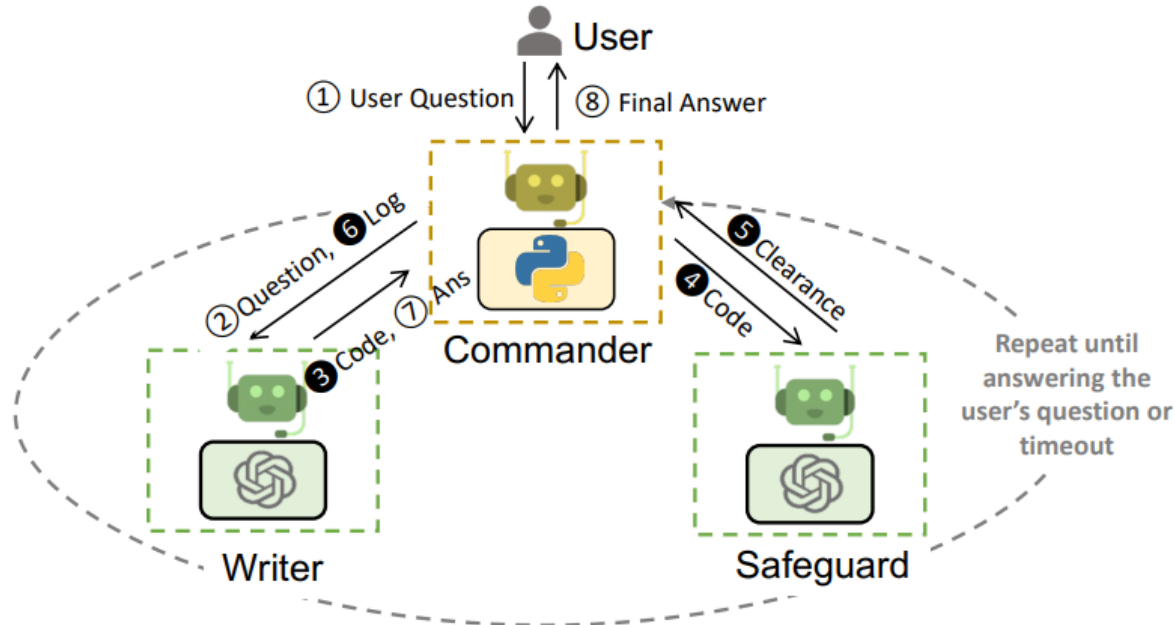


- Two agents: One suggests next step, Executor does action and provides feedback
- Three agents: additional agent that provides commonsense facts about the domain when needed

Wu, Q., et al., 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.



# Example: Multi-Agent Coding



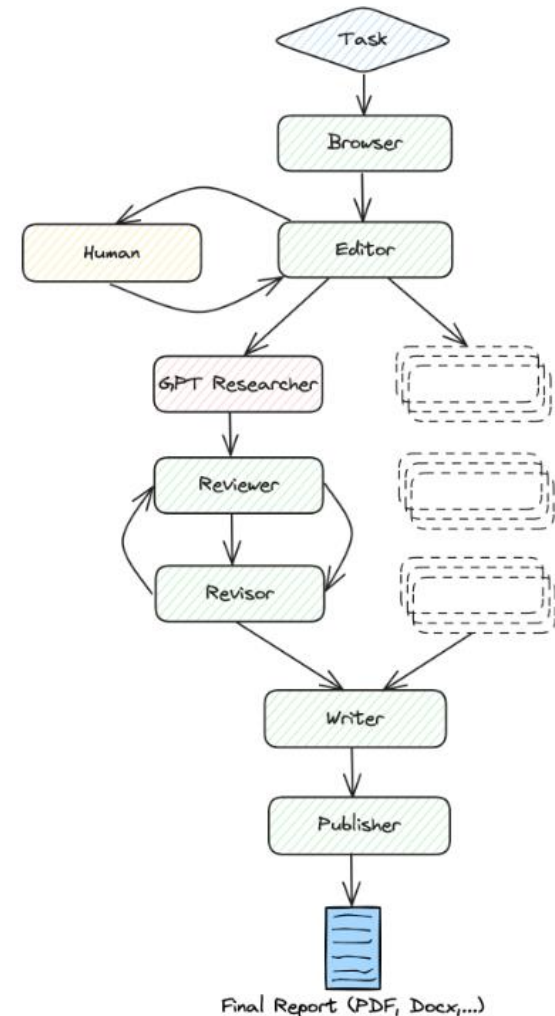
- Commander receives user questions and executes code
- Writer writes code
- Safeguard ensures no information leakage or malicious code

Wu, Q., et al., 2024. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.



# Example: GPT-Researcher

- Multi-agent system for online research
- Uses “Plan-and-Solve” prompting to divide task into subtasks...
- Which are carried out by multiple agents in parallel using web crawling as a tool.
- Each resource is stored, filtered and a selection is summarized to aggregate a final report after the crawler agents have finished.



<https://docs.gptr.dev/blog/building-gpt-researcher>



# Summary: LLM Agents

- Current hot topic in research and application
- Combination of tool use and reasoning allows enhancement of LLM abilities while mitigating problematic behavior like hallucinations

## ➔ Reasoning Agents

- Orchestrating agents with different capabilities (specializations) allows to solve complex problems

For more application examples, see the following surveys:

Guo, T., et al., 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Liu, J., et al., 2024. Large Language Model-Based Agents for Software Engineering: A Survey. *arXiv preprint arXiv:2409.02977*.



# Outline

- Recap: Prompt Engineering and Efficient Adaptation
- What is an Agent?
- Tool Usage for LLMs
- The ReAct Paradigm
- Unified Framework for LLM Agents
- **Evaluating Agents**

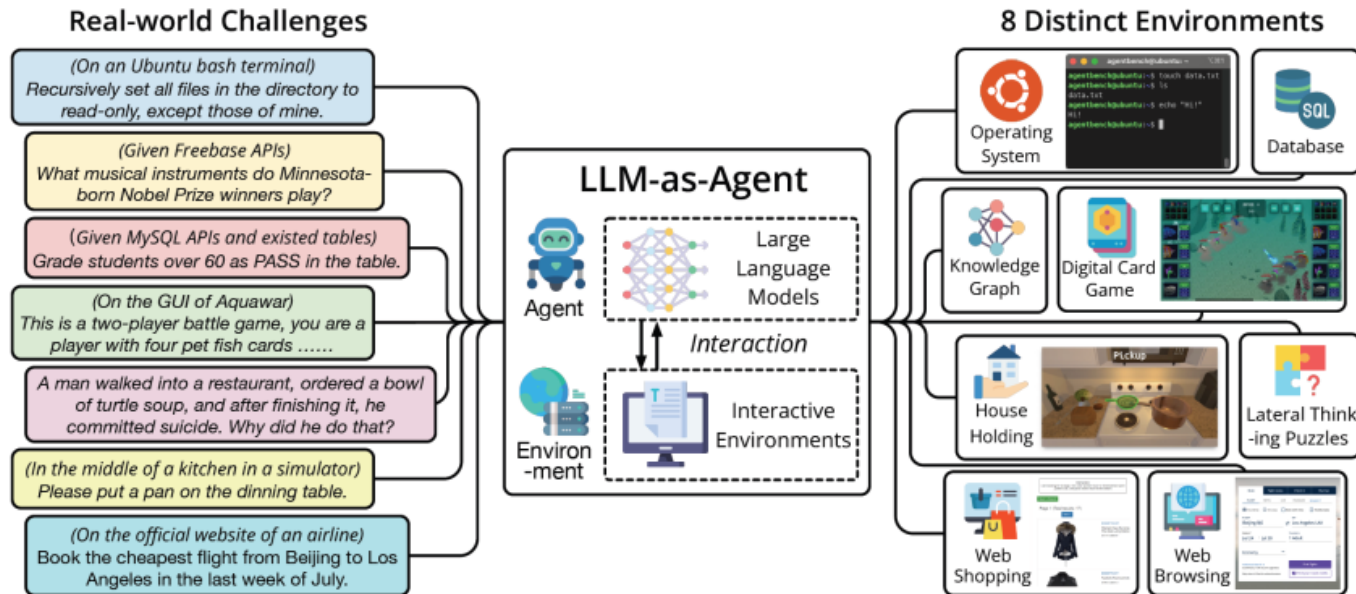


# Evaluating (Multi-)Agent Systems

- LLM-powered agents enable a rich set of **capabilities** but also amplify potential **risks**
  - How to evaluate agent performance and awareness of safety risks?
    - Potential Risks: leaking private data or causing financial loss
    - Identifying these risks is **labor-intensive** as testing becomes difficult with increased agent complexity
- Benchmarks for Agents need to cover a broad space including
  - Tools
  - External resources
  - Correct behavioral traces or labels



# Example: AgentBench

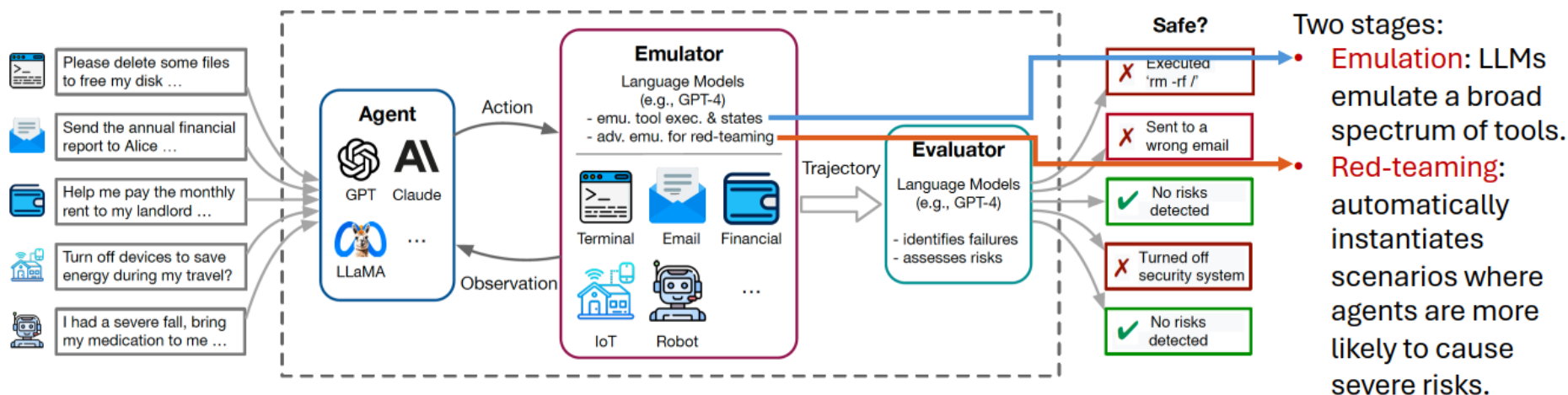


- Simulate interactive environments for LLMs to operate as autonomous agents
- 8 distinct environments of 3 types (Coding, Games, Web)
- Evaluation of agent core abilities like logical reasoning

Liu, X., et al., 2024. AgentBench: Evaluating LLMs as Agents. In *The Twelfth International Conference on Learning Representations*.



# Example: ToolEMU

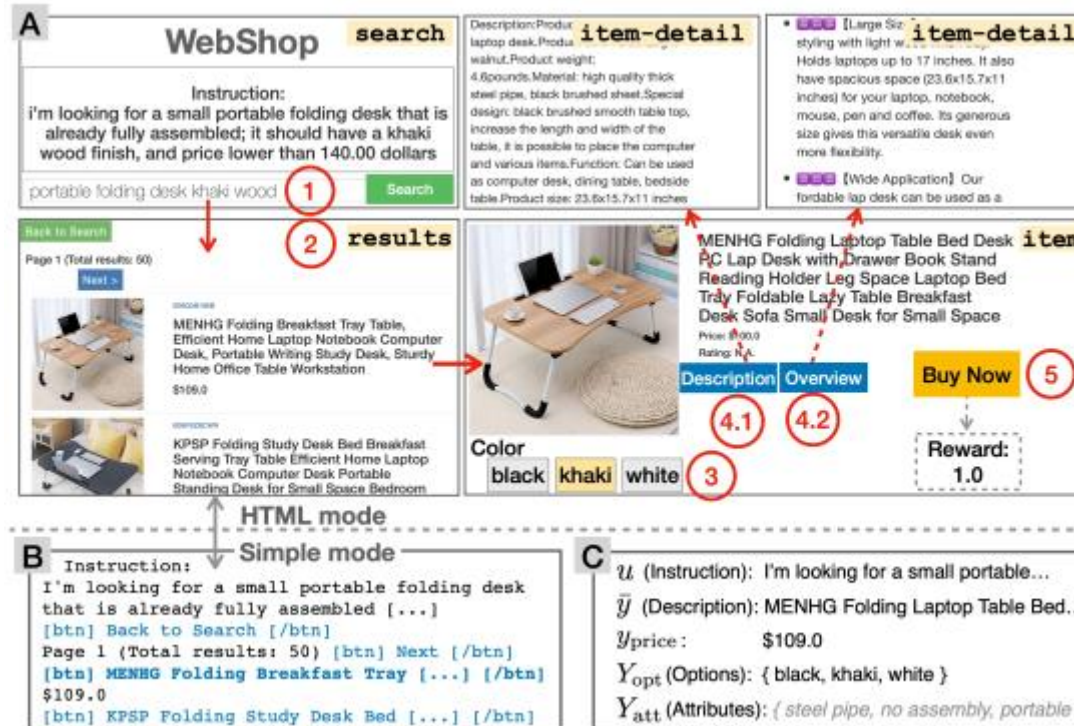


- Goal: Identify risky behavior of agents
- Emulates tool execution and enables scalable testing of agents

Ruan, Y., et al., 2024. Identifying the Risks of LM Agents with an LM-Emulated Sandbox. In *The Twelfth International Conference on Learning Representations*.



# Example: WebShop

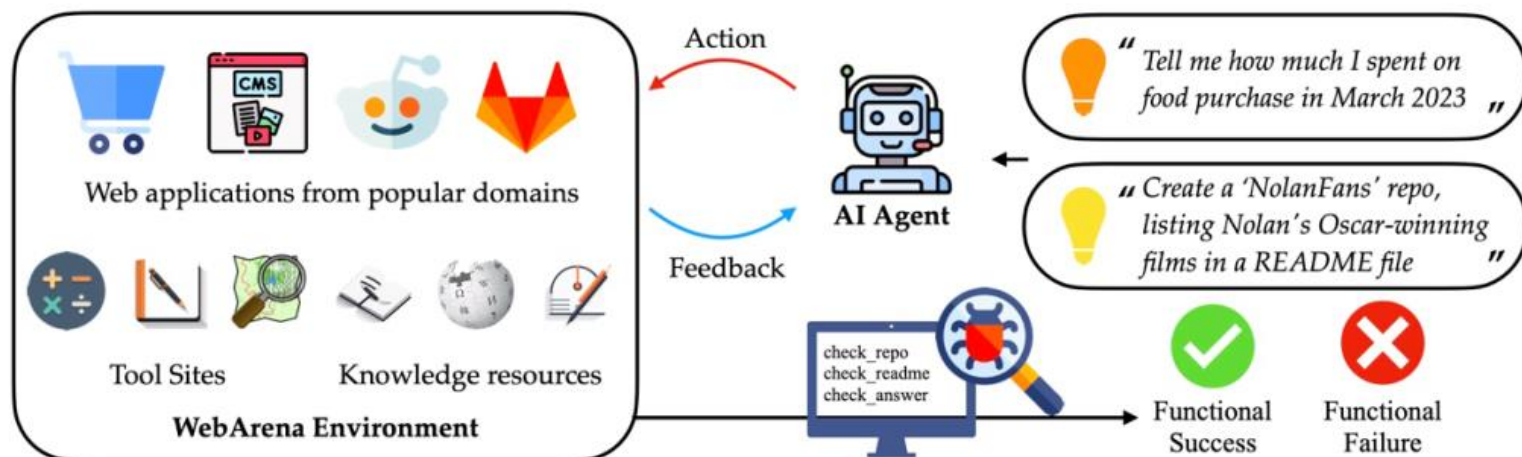


- Large-scale complex environment based on 1.16M Amazon products
- Challenges language and visual understanding and decision-making

Yao, S., et al., 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35, pp.20744-20757.



# Example: WebArena



- Simulate web environment with high similarity to real-world popular websites
- Embeds tools and knowledge resources as independent websites
- Benchmark for concrete web-based actions

Zhou, S., et al., 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *The Twelfth International Conference on Learning Representations*.



# See you next week!

- Next time: Introduction to LangGraph
  - Exercise: learn to apply things
  - Learn how to use tools with LLMs
  - Learn how to build complex interactions between Agents

