CS11-711 Advanced NLP

LLM Tool Use and Agent Basics

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

https://phontron.com/class/anlp-fall2024/

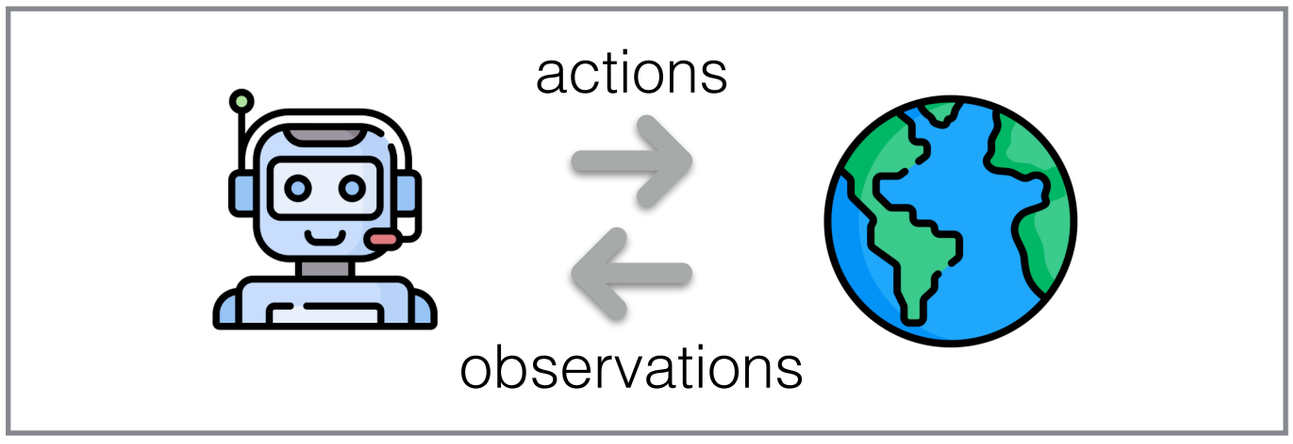
Some slides by Frank Xu and Zora Zhiruo Wang

From Words to Action

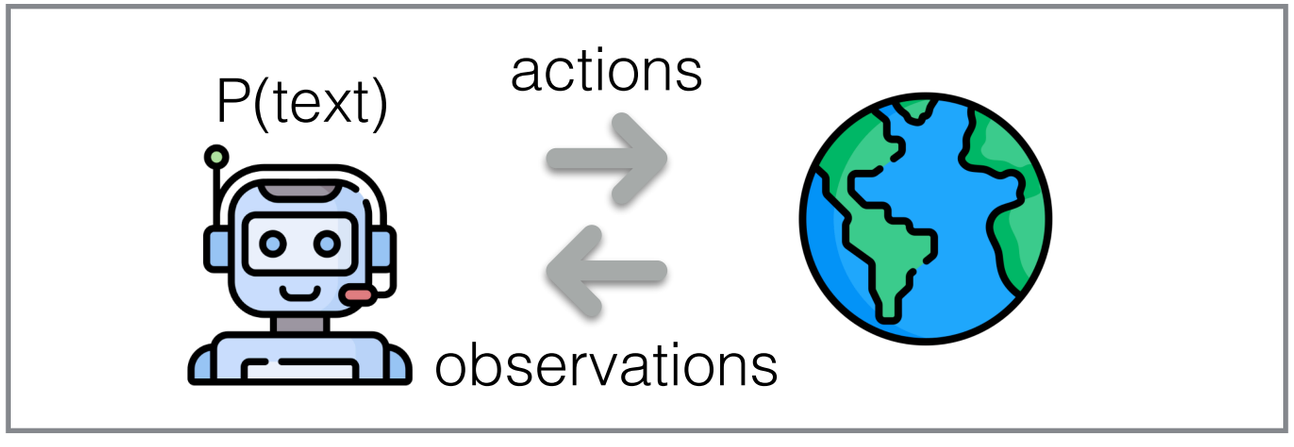
- Language models predict text

P(text)

- AI agents iteratively perform actions in the world



- LM agents are an agent with a an LM backbone



Agent Definition

- Disagreement on what “agent” or “agentic” means
- **Requirements:**
  - *Probably:* Proactive use of tools
  - *Probably:* An iterative, multi-step process
  - *Maybe:* Interaction with the outside world

What does an LM agent consist of?

- Underlying LLM
- Prompt
- Action/Observation Space

Is This an Agent?

- An LM system that browses the web - Yes
- An LM system that searches for files on your OS and processes them using code - Yes
- An LM system that retrieves then generates - Probably not (not incremental)
- An LM like o1 with complex CoT - No (no tools or outside world)

Requirements for Successful Agents

- Tool Use
- Environment Representation
- Environment Understanding
- Reasoning and Planning
- Interaction/ Communication

Agent Use Cases/ Environments

Chat Assistants

- e.g. ChatGPT plugins

ChatGPT plugins

We've implemented initial support for plugins in ChatGPT. Plugin
language model
help ChatGPT ac
computations, o

[ChatGPT plugins]

Expedia
Bring your trip plans to life—get there, stay there, find things to see and do.

FiscalNote
Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.

Instacart
Order from your favorite local grocery stores.

KAYAK
Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.

Klarna Shopping
Search and compare prices from thousands of online shops.

Milo Family AI
Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?

OpenTable
Provides restaurant recommendations, with a direct link to book.

Shop
Search for millions of products from the world's greatest brands.

Speak
Learn how to say anything in another language with Speak, your AI-powered language tutor.

Wolfram
Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language.

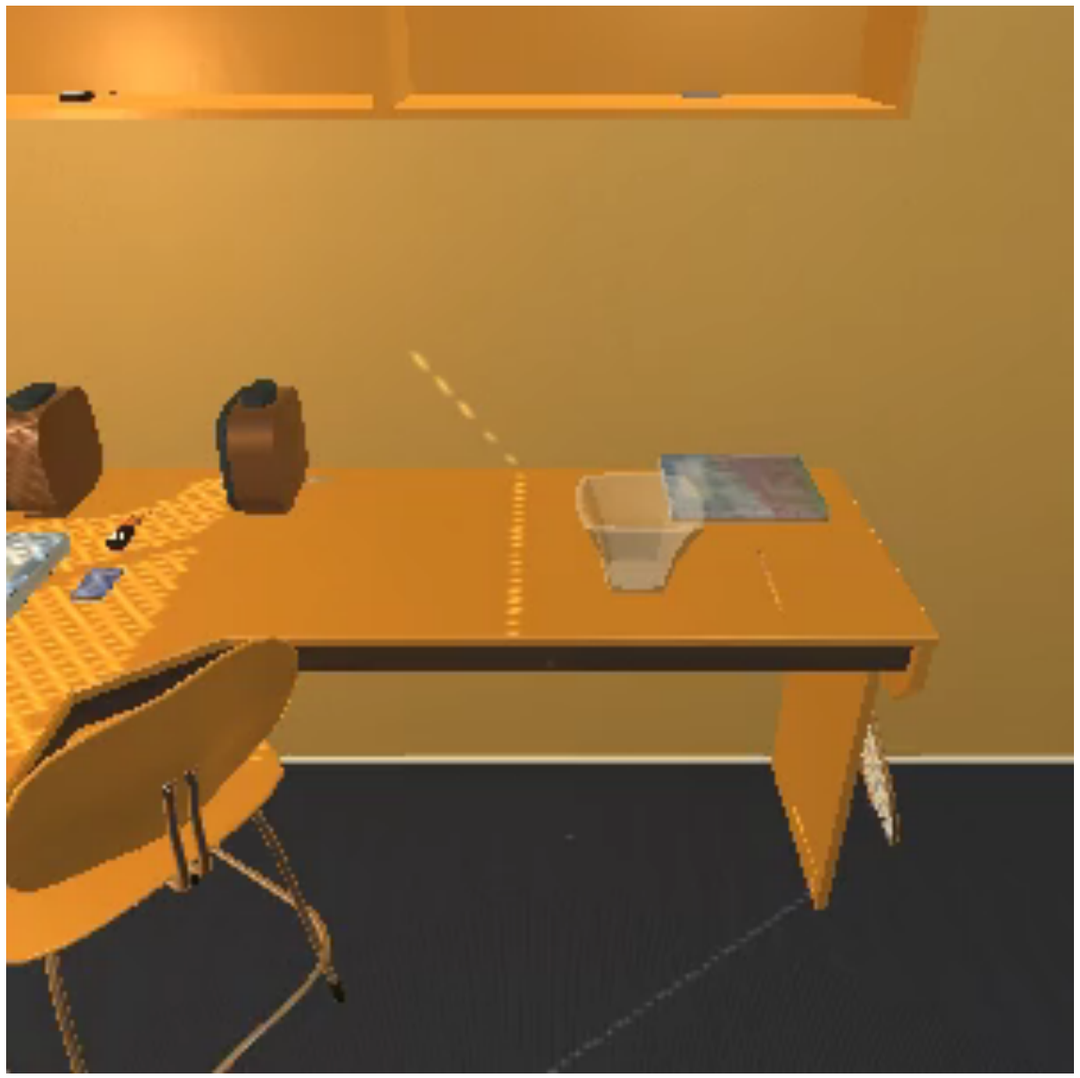
Zapier
Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

Robotics

Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Chen et al., 18’

Embodied Agents



You are in the middle of a room. Looking quickly around you, you see a safe 1, a shelf 4, a drawer 2, a bed 1, a drawer 1, a shelf 5, a shelf 2, a sidetable 2, a shelf 3, a drawer 3, a shelf 1, a sidetable 1, a desk 1, and a garbagecan 1.

Your task is to: examine an alarmclock with the desklamp.

> go to desk 1

You arrive at loc 8. On the desk 1, you see a pen 1, a bowl 1, a alarmclock 2, a pencil 2, a pencil 3, a creditcard 3, a book 1, a alarmclock 3, a keychain 3, and a book 2.

> take alarmclock 2 from desk 1

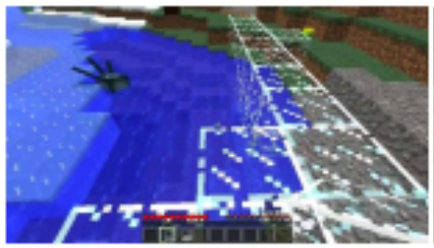
You pick up the alarmclock 2 from the desk 1.

AlfWorld, Shridhar et al. 21’

Games

Open-ended Environments

Craft Glass Bridge



Build Oak House

Make Ice Igloo

Combat Zombie

Fish Squid

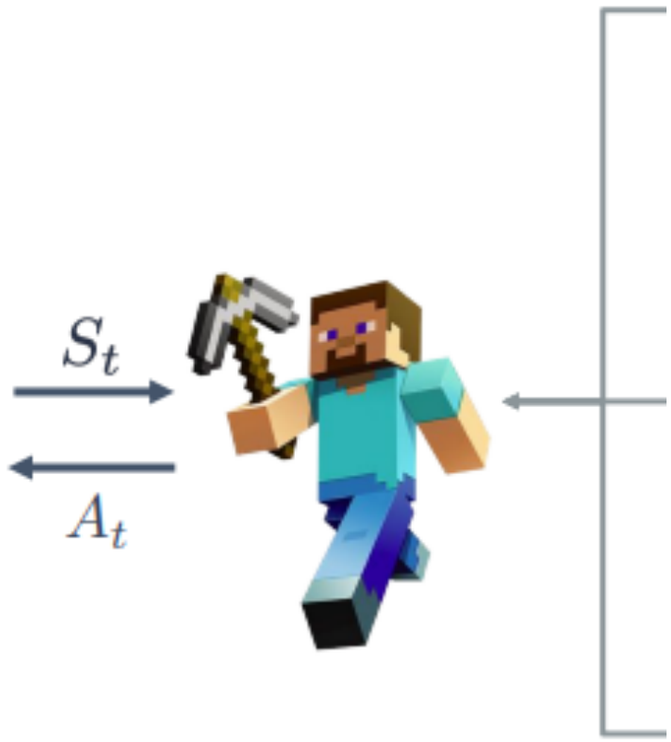
Farm Sugar Cane

Find Ocean Monument

Explore Desert Temple

Treasure Hunt in End City

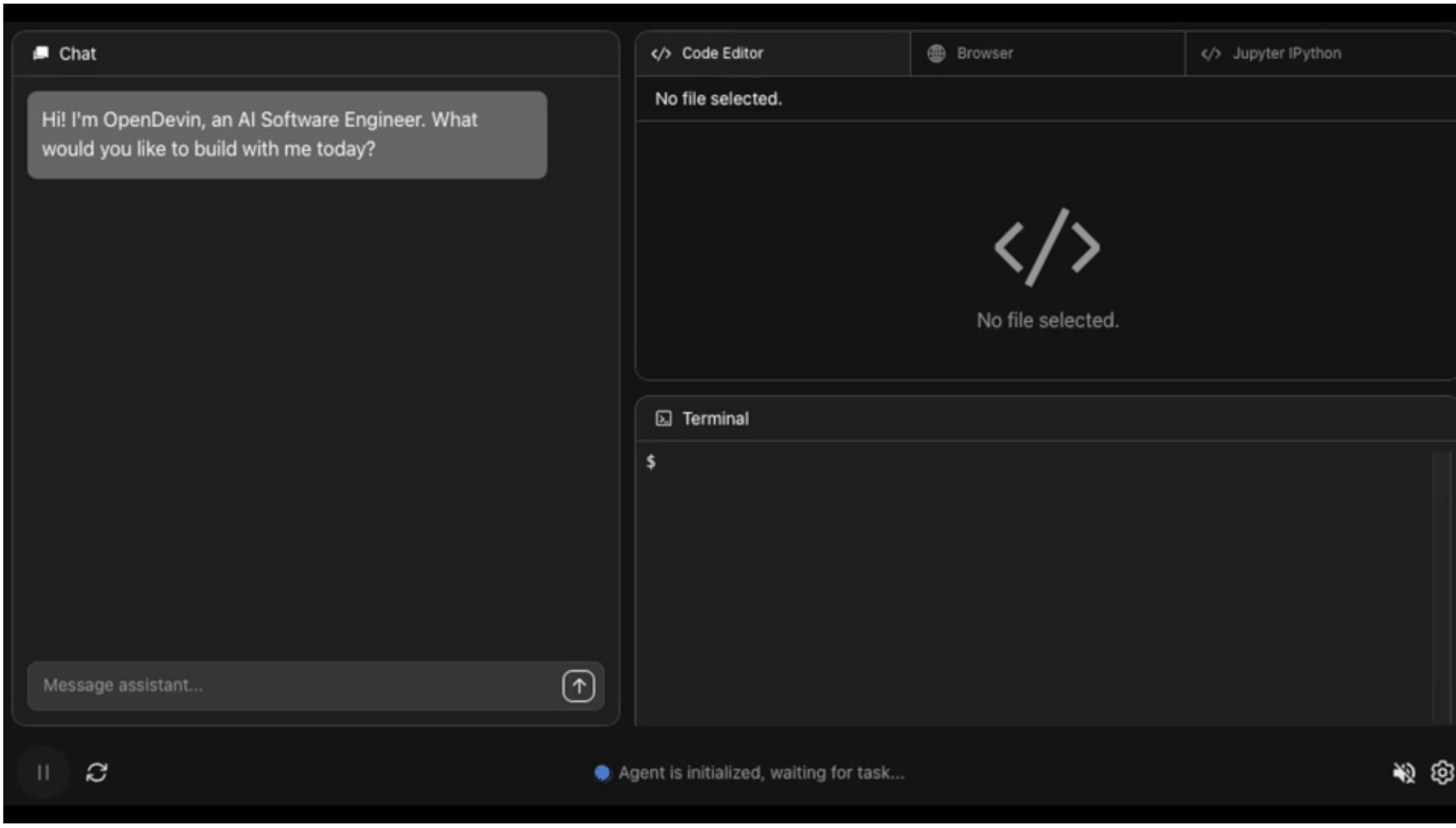
Generalist Agent



MineDojo, Fan et.al 22’

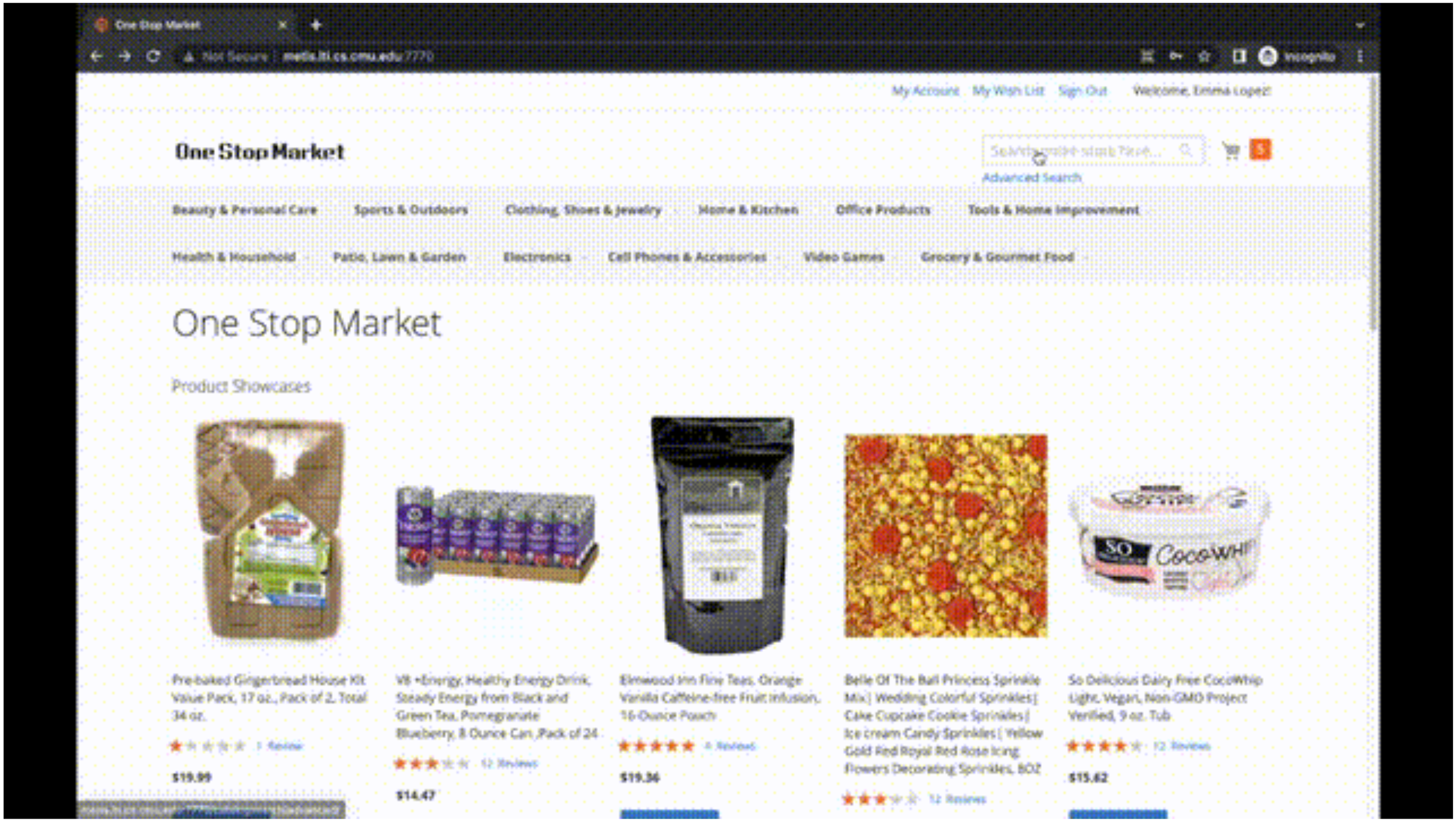
Software Development

e.g. OpenHands (Wang et al. 2024)



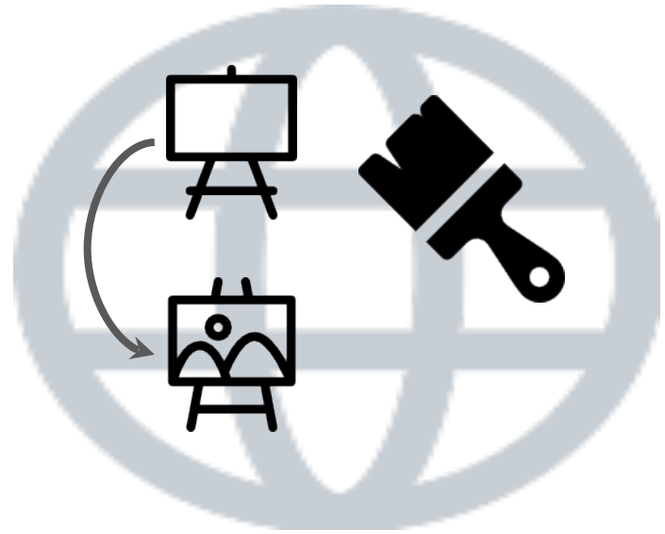
UI automation

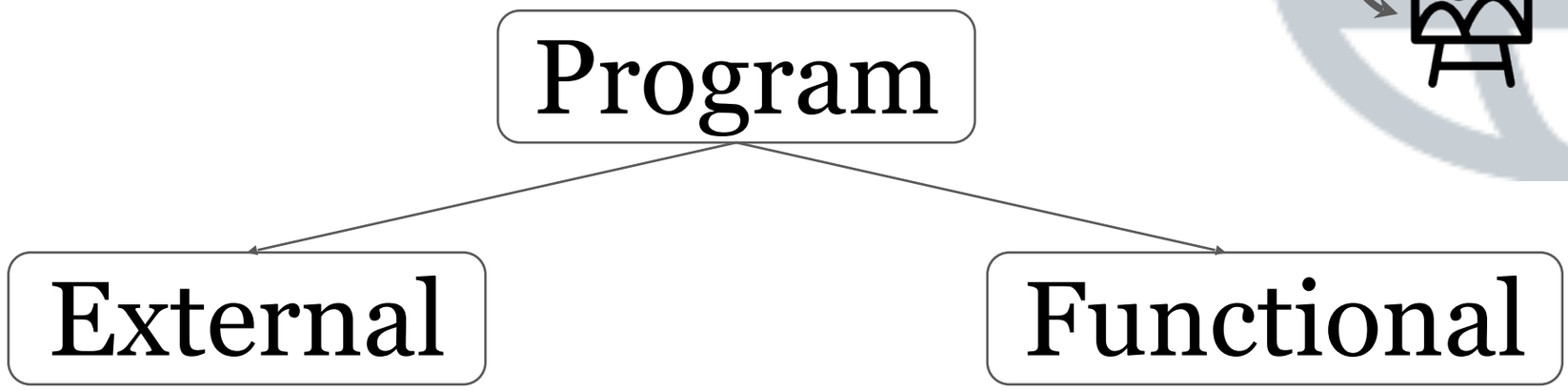
e.g. WebArena (Zhou+Xu et al. 2023)



Tool Use in LMs

What is a Tool (for LLM Agents)?

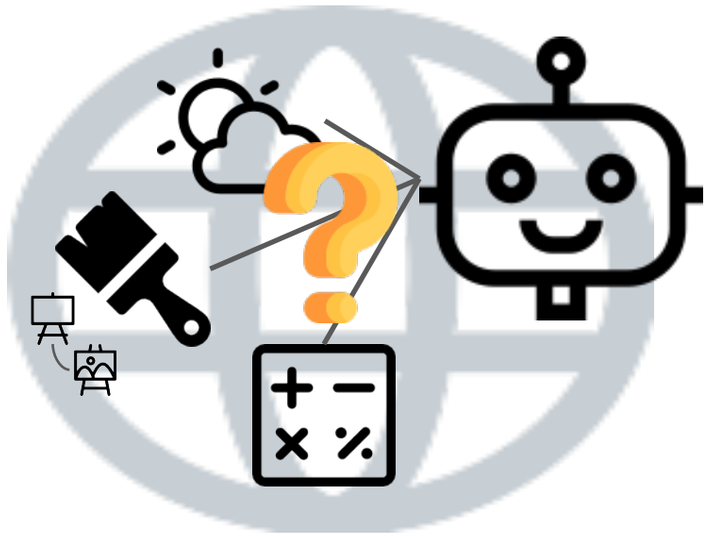




*An LM-used tool is a **function** interface to a computer **program** that runs **external** to the LM, where the LM generates the function calls and input arguments in order to use the tool.*

[1] Shumaker et al. *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press, 2011.

Tool Functionality



- 👀 Perception: collect data from the env
- 🦾 Action: exert actions, change env state
- 🔢 Computation: general acts of computing

→ ***Tools***

[1] Russell, Stuart J., and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

Tool Use Scenarios

Category	Example Tools

Knowledge access	`sql_executor(query: str) -> answer: any`
	`search_engine(query: str) -> document: str`
	`retriever(query: str) -> document: str`

Computation activities	`calculator(formula: str) -> value: int | float`
	`python_interpreter(program: str) -> result: any`
	`worksheet.insert_row(row: list, index: int) -> None`

Interaction w/ the world	`get_weather(city_name: str) -> weather: str`
	`get_location(ip: str) -> location: str`
	`calendar.fetch_events(date: str) -> events: list`
	`email.verify(address: str) -> result: bool`

Non-textual modalities	`cat_image.delete(image_id: str) -> None`
	`spotify.play_music(name: str) -> None`
	`visual_qa(query: str, image: Image) -> answer: str`

Special-skilled LMs	`QA(question: str) -> answer: str` `translation(text: str, language: str) -> text: str`

Table 1: Exemplar tools for each category.

Tool Use Paradigms

Tool Use: switching between

- text-generation mode
- tool-execution mode

How to induce tool use

- Inference-time prompting
- Training

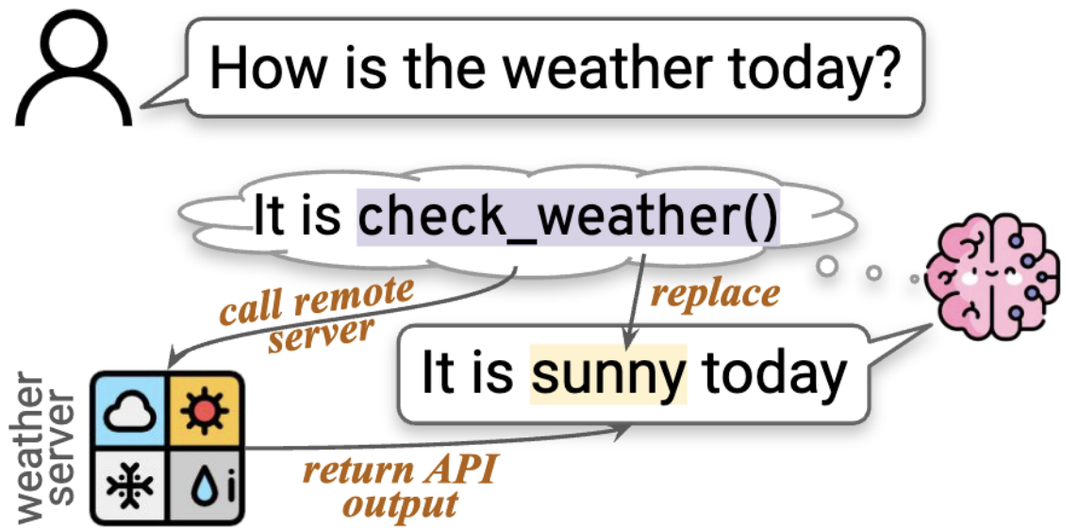
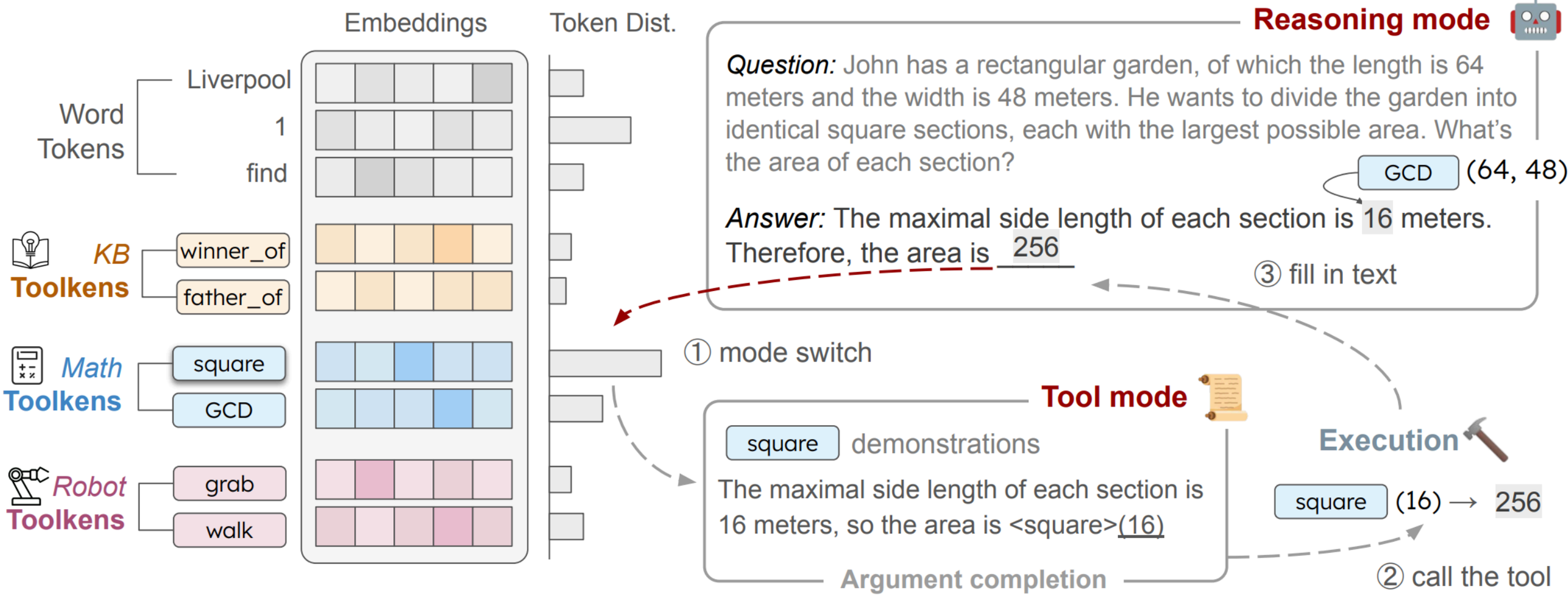


Figure 2: The basic tool use paradigm. LM calls `check_weather` tool by generating text tokens. This call triggers the server to execute the call and return the output `sunny`, using which the LM replaces the API call tokens in the response to the user.

Tool Execution: Tool Tokens

- e.g. ToolkenGPT (Hao et al. 2023)



Tool Execution: Code Tags

- e.g. CodeAct (Wang et al. 2024)

I should calculate the phone price in USD for each country, then find the most cost-effective country.
<execute_python>

```
countries = ['USA', 'Japan', 'Germany', 'India']
final_prices = {}

for country in countries:
    exchange_rate, tax_rate = lookup_rates(country)
    local_price = lookup_phone_price("xAct 1", country)
    converted_price = convert_and_tax(
        local_price, exchange_rate, tax_rate
    )
    shipping_cost = estimate_shipping_cost(country)
    final_price = estimate_final_price(converted_price, shipping_cost)
    final_prices[country] = final_price

most_cost_effective_country = min(final_prices, key=final_prices.get)
most_cost_effective_price = final_prices[most_cost_effective_country]
print(most_cost_effective_country, most_cost_effective_price)
```

<execute_python>

Prompting for Tool Use

- e.g. DocPrompting (Zhou et al. 2022) retrieves library documentation

Input

Potential document 0: w displays information about the users currently on the machine, and their processes. The header shows, in this order …
Potential document 1: -s, –short Use the short format. Don't print the login time, JCPU or PCPU times.

display information without including the login, jcpu and pcpu columns

Output

w --short

Learning for Tool Use

- e.g. ToolFormer (Schick et al. 2023)

Prompt Unsupervised

Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[QA(question)]" where "question" is the question you want to ask. Here are some examples of API calls:

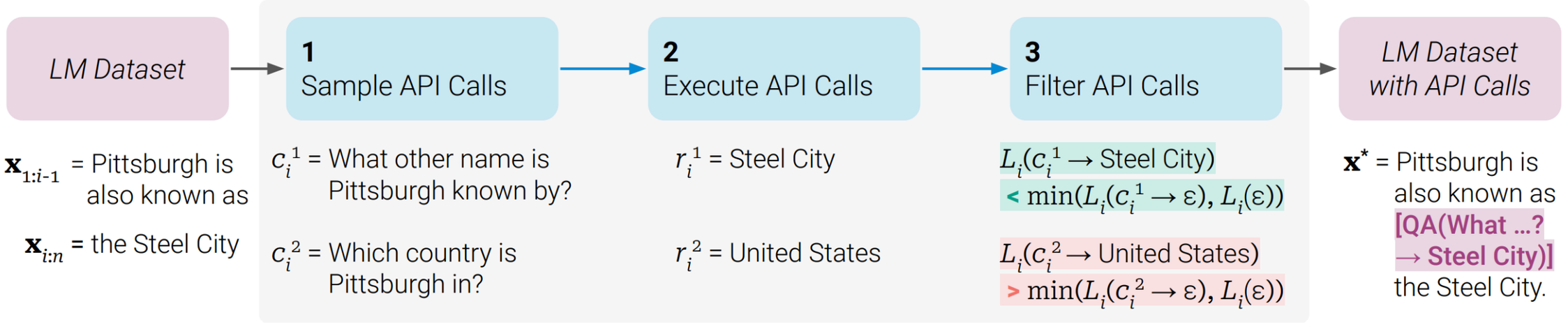
Input: Coca-Cola, or Coke, is a carbonated soft drink manufactured by the Coca-Cola Company.

Output: Coca-Cola, or **[QA("What other name is Coca-Cola known by?")]** Coke, is a carbonated soft drink manufactured by **[QA("Who manufactures Coca-Cola?")]** the Coca-Cola Company.

Input: x

Output:

Filter for Success and Train



OpenAI Function Calling Standard

(OpenAI 2024)

- Define a function signature in a Python dictionary

```
tools = [
  {
    "name": "get_delivery_date",
    "description": "Get the delivery date for a customer's order. Call this
whenever you need to know the delivery date, for example when a customer asks
'Where is my package'",
    "parameters": {
        "type": "object",
        "properties": {
            "order_id": {
                "type": "string",
                "description": "The customer's order ID."
            }
        },
        "required": ["order_id"],
        "additionalProperties": false
    }
  }
]
```

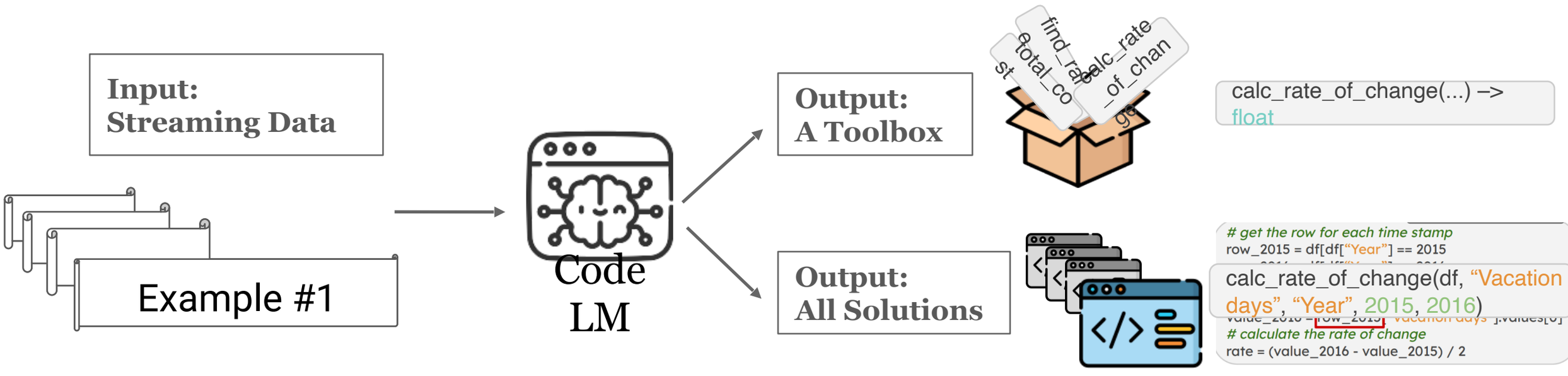
- Send it together with your prompt

```
response = openai.chat.completions.create(
    model="gpt-4o",
    messages=messages,
    tools=tools,
)
```

Tool Induction

TROVE: Inducing Verifiable and Efficient Toolboxes for Solving Programmatic Tasks

Zhiruo Wang [1] **Graham Neubig** [1] **Daniel Fried** [1]



Environment Representation

Environment Understanding

- For an agent to understand its environment it needs
- **Tools** to access the environment (last section)
- A **representation** of the environment
- Methods for **holistic understanding/exploration**

Environment Representation: Text

- e.g. ALFWorld (Shridhar et al. 2021)

```
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!
```

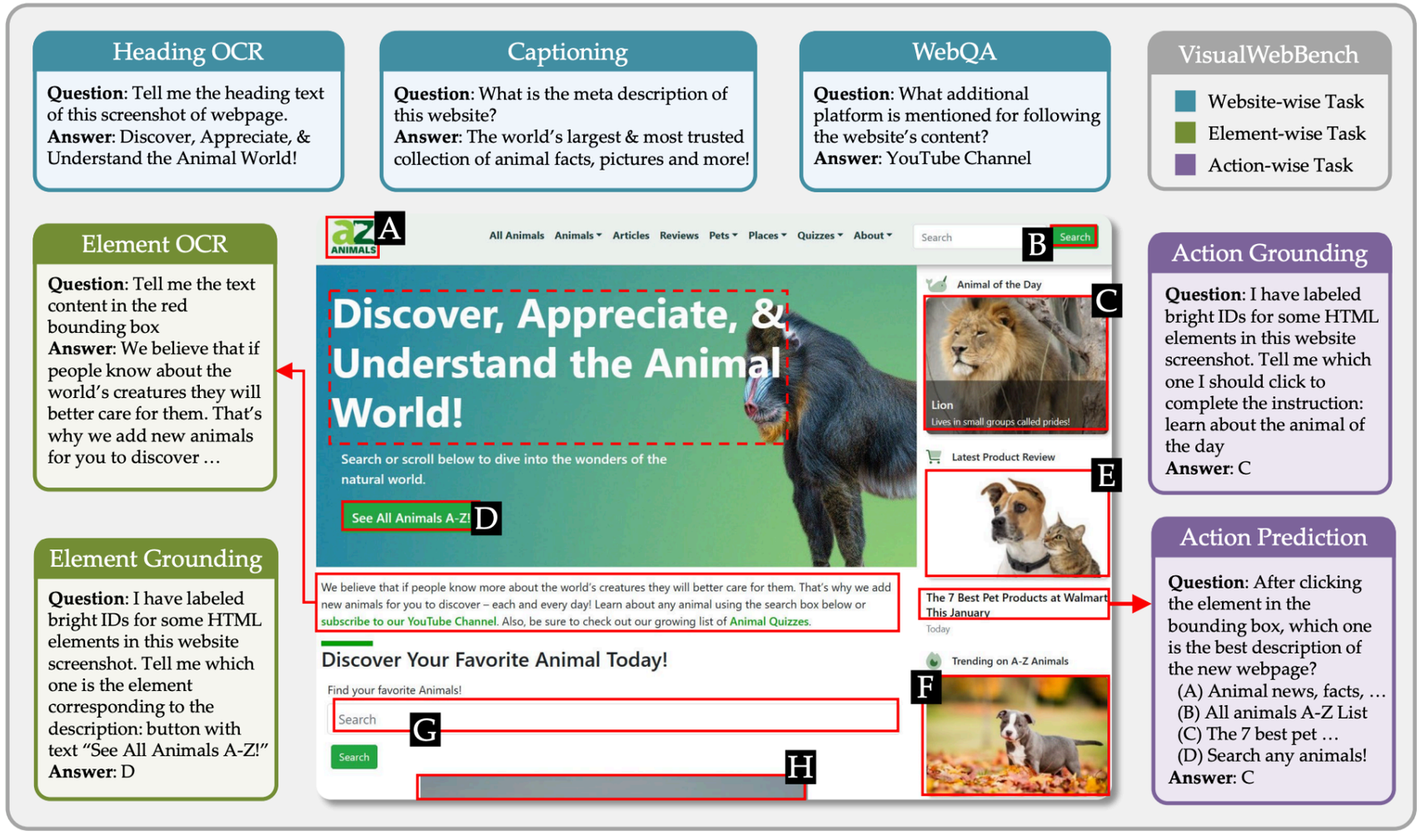
Environment Representation: Images

- e.g. Touchdown (Chen et al. 2018)

Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

Problems w/ Image Representations

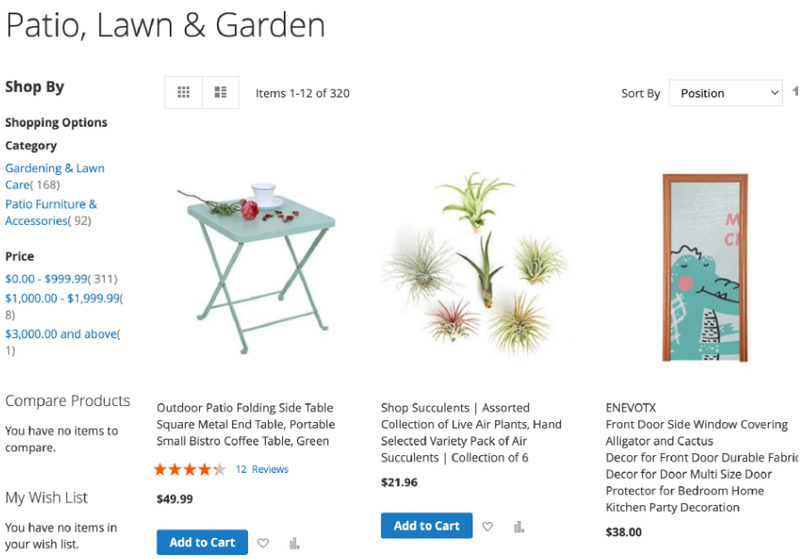
- In order to perform well as an agent details matter!
- For instance, OCR, grounding over complex layouts
- Many models fail at these (Liu et al. VisualWebBench), but this can be remedied somewhat with training (Liu et al. MultiUI)



Environment Representation: Textual Web Representations

- e.g. WebArena (Zhou+Xu et al. 2024)

webarena.onestopshop.com



Screenshot

webarena.onestopshop.com

```
<li>
  <div>
    <a href="..."><img src="..."></a>
    <div class>
      <a href="...">Outdoor Patio …
</a>
      <div>
          <span>Rating:</span>
          <div>
            <span>82%</span>
          </div>
          <a href="…#reviews">12
<span>Reviews</span></a>
```

Text

webarena.onestopshop.com

```
RootWebArea ‘Patio, Lawn ..’
  link 'Image'
    img 'Image'
link 'Outdoor Patio..’
LayoutTable ''
    StaticText 'Rating:'
    generic '82%'
    link '12 Reviews'
StaticText ‘$49.99'
button 'Add to Cart’ focusable: True
button 'Wish List’ focusable: …
button 'Compare’ focusable: …
```

Accessibility tree

Environment Representation: Set of Marks (Yang et al. 2023)

- Mark each item with a number and ask the LLM to identify by number

Environment Understanding

How can we Understand Complex Environments?

- Models don't know everything about the environments they interact with
- Some knowledge is included in LLM parameters (coding, navigating popular web sites, etc.)
- Other knowledge must be discovered on the fly

Environment-specific Prompts

- Manually craft prompts that give directions about the environment
- e.g. SteP (Sochi et al. 2023) prompts for web navigation

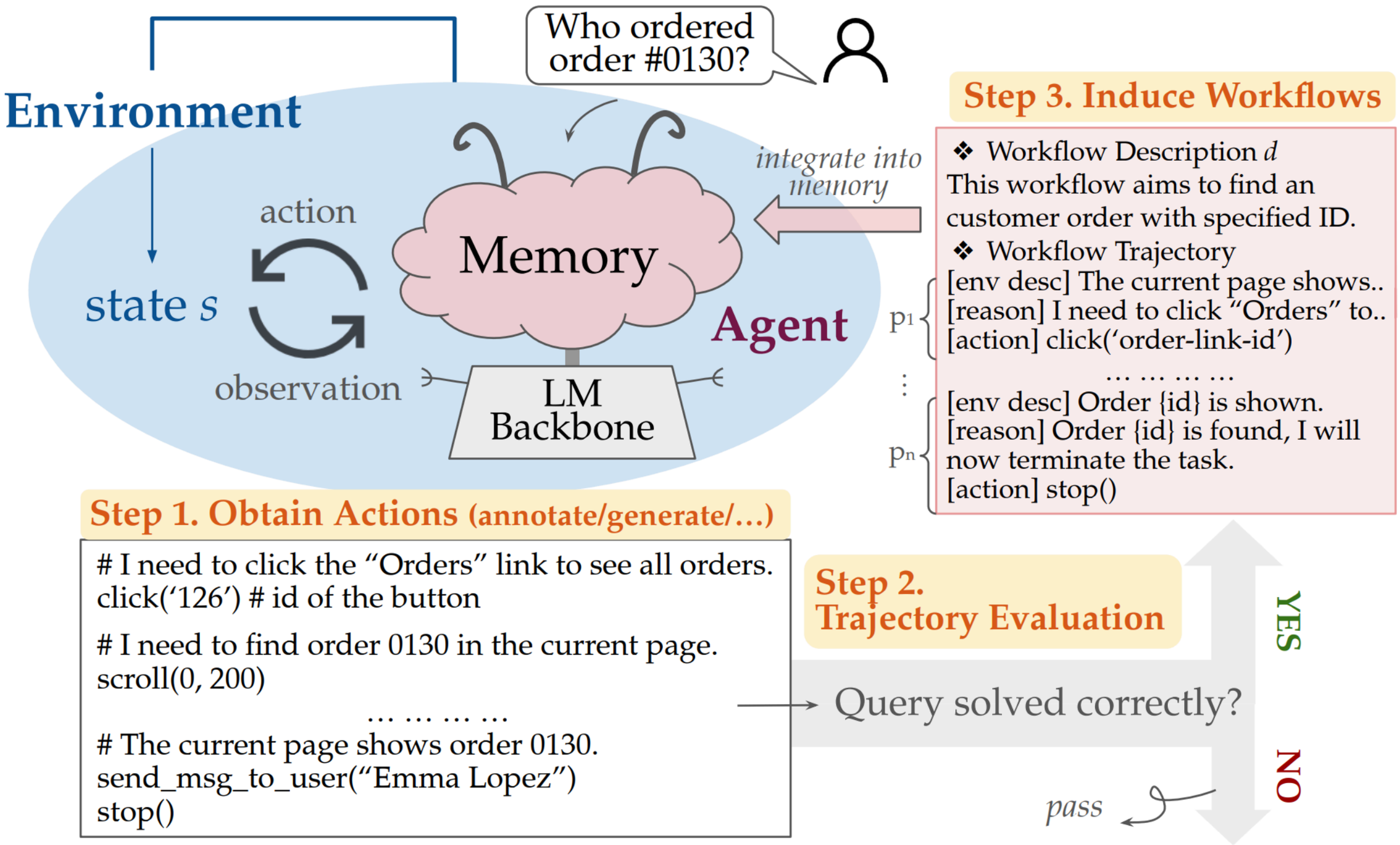
```
search_issues = {
 "instruction": """
 {general_instruction_template}

 Please follow these general instructions:
 * First navigate the Issues page
 * Once you are in the Issues page, you MUST first navigate to all issues so
that you see both open and closed issues for solving the objective
 * You may not see all issues listed at once, use the search bar to search for
appropriate keywords and filter down to relevant set of issues
 * If the objective says to "Open ... issue, check if it is X", you must first
open the specific issue page by clicking it. Do not stop [] until you have
navigated to the specific issue page.
 * Once you are on the issue page, return the appropriate status
 * In your status, if the objective is to check if an issue is open or clossed,
respond as though you are answering a question, e.g. "No, it is open", "Yes, it
is closed"
 """
}
```

- Issue: generalization!

Unsupervised Induction of Prompts: Agent Workflow Memory (Wang et al. 2024)

- remembers successful workflows and prompts the model with them

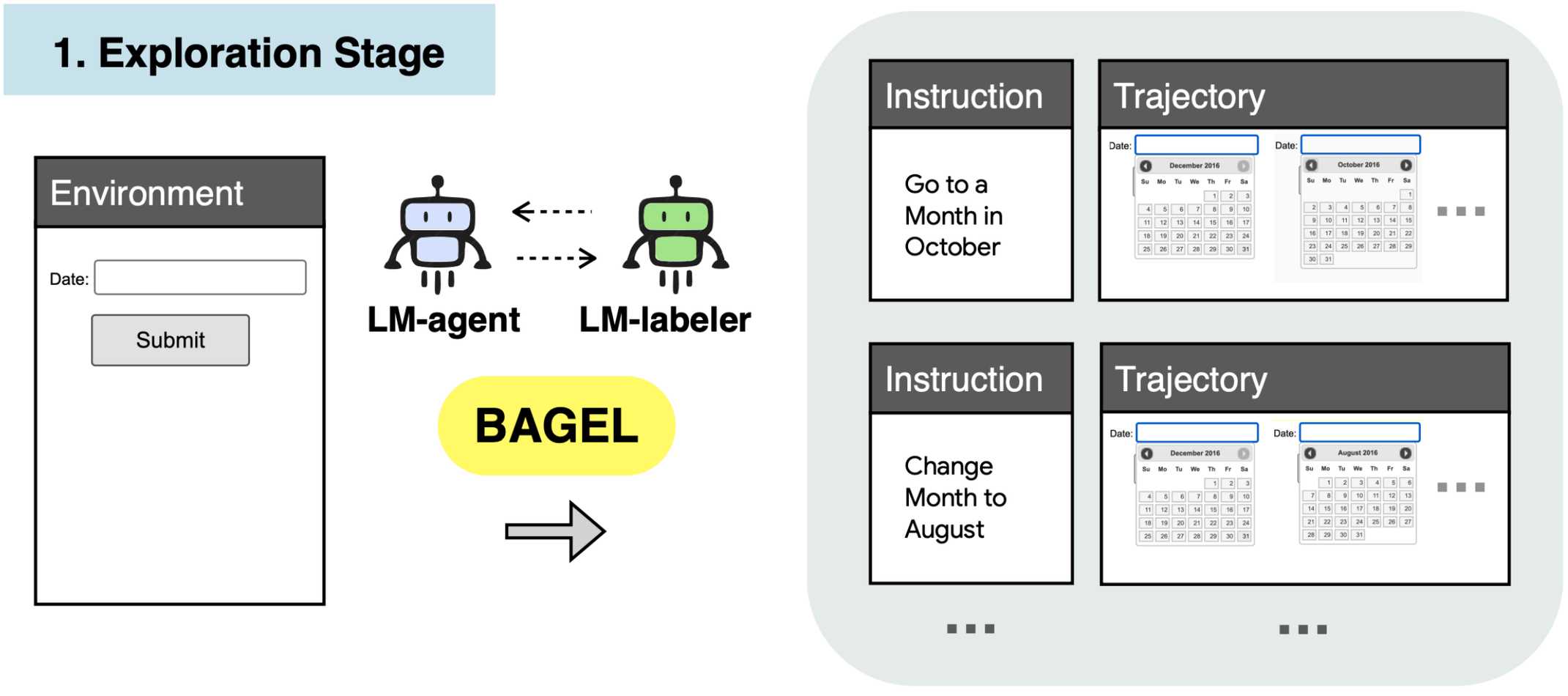


Environment Exploration

- Have the model be rewarded for exploring the environment: “curiosity”
- e.g. use RL and increase reward when the model enters a part of the state space that is not predictable (Pathak et al. 2017)

Exploration-based Trajectory Memorization: BAGEL (Murty et al. 2024)

- Sample instructions and follow them, then re-label trajectory with new more accurate instructions



Reasoning and Planning

Reasoning within Agents

- ReAct (Yao et al. 2022) performs local CoT reasoning before acting

(1d) ReAct (Reason + Act)

Thought 1: I need to search Apple Remote and find the program it was originally designed to interact with.
Act 1: Search[Apple Remote]
Obs 1: The Apple Remote is a remote control introduced in October 2005 by Apple … originally designed to control the Front Row media center program …

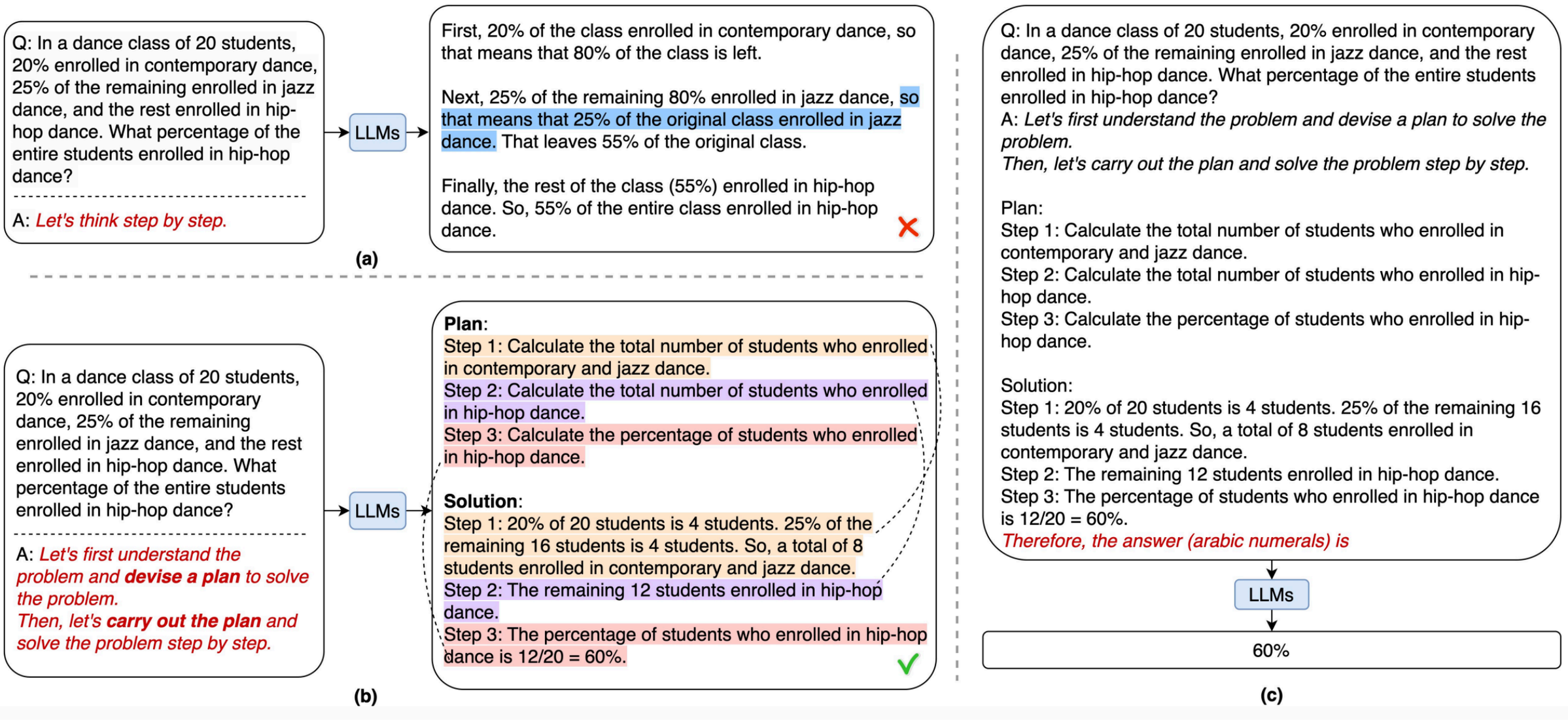
Thought 2: Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.
Act 2: Search[Front Row]
Obs 2: Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports',' Front Row (software) ', …]

Thought 3: Front Row is not found. I need to search Front Row (software) .
Act 3: Search[Front Row (software)]
Obs 3: Front Row is a discontinued media center software …

Thought 4: Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.
Act 4: Finish[keyboard function keys]

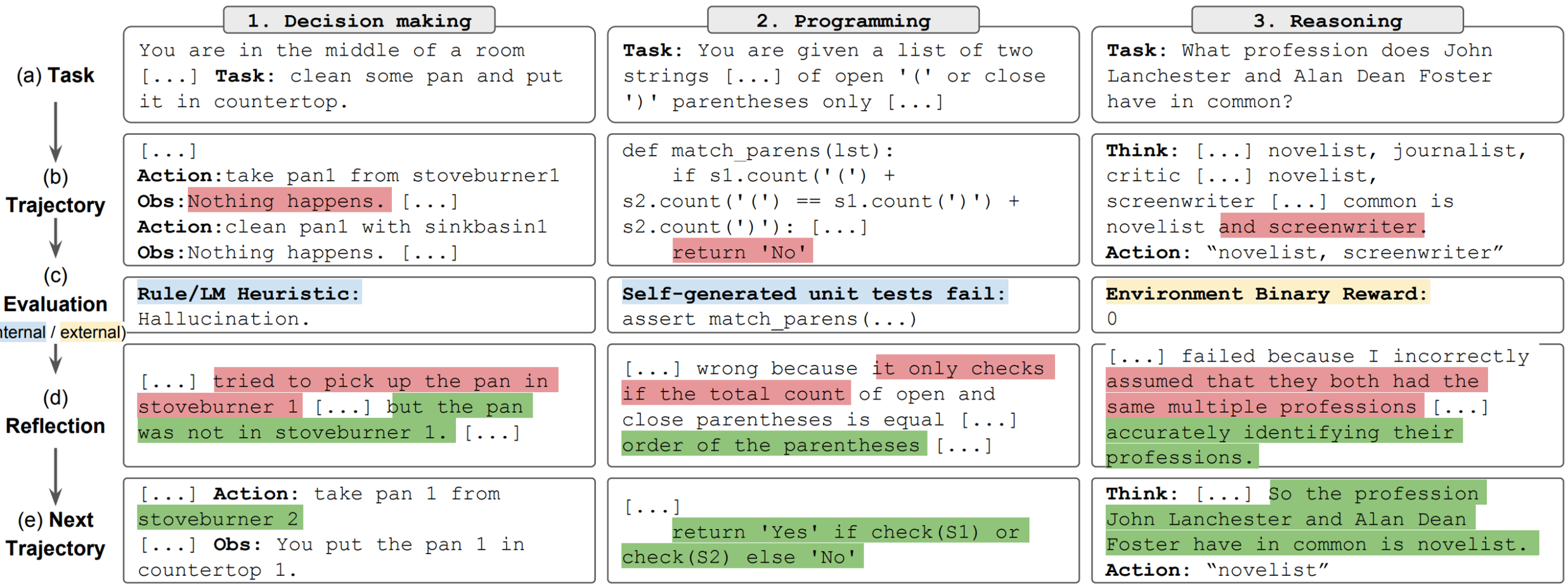
Global Planning

- First, devise a global plan and solve for the plan
- e.g. Plan-and-solve Prompting (Wang et al. 2023)



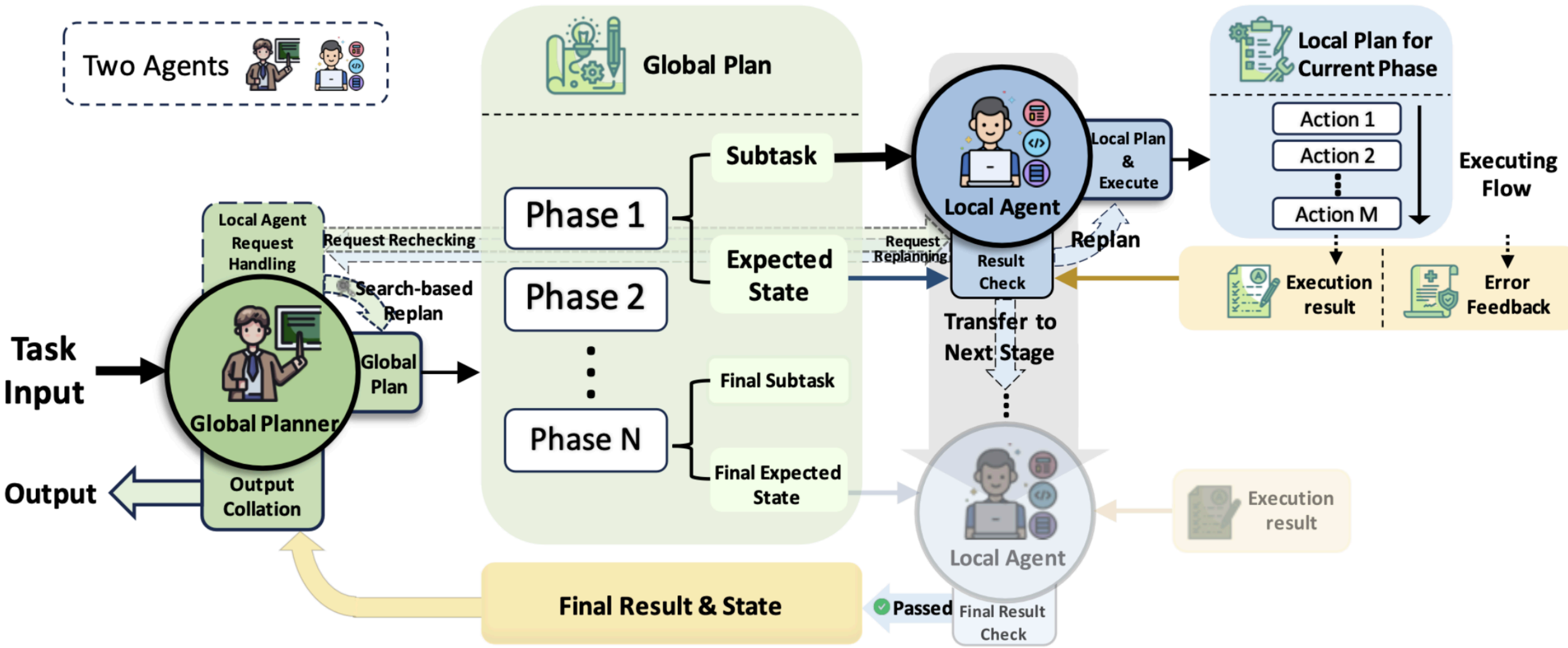
Error Identification and Recovery

- Need to have a way to recover from mistakes!
- e.g. Reflexion (Shinn et al 2023)



Revisiting Plans

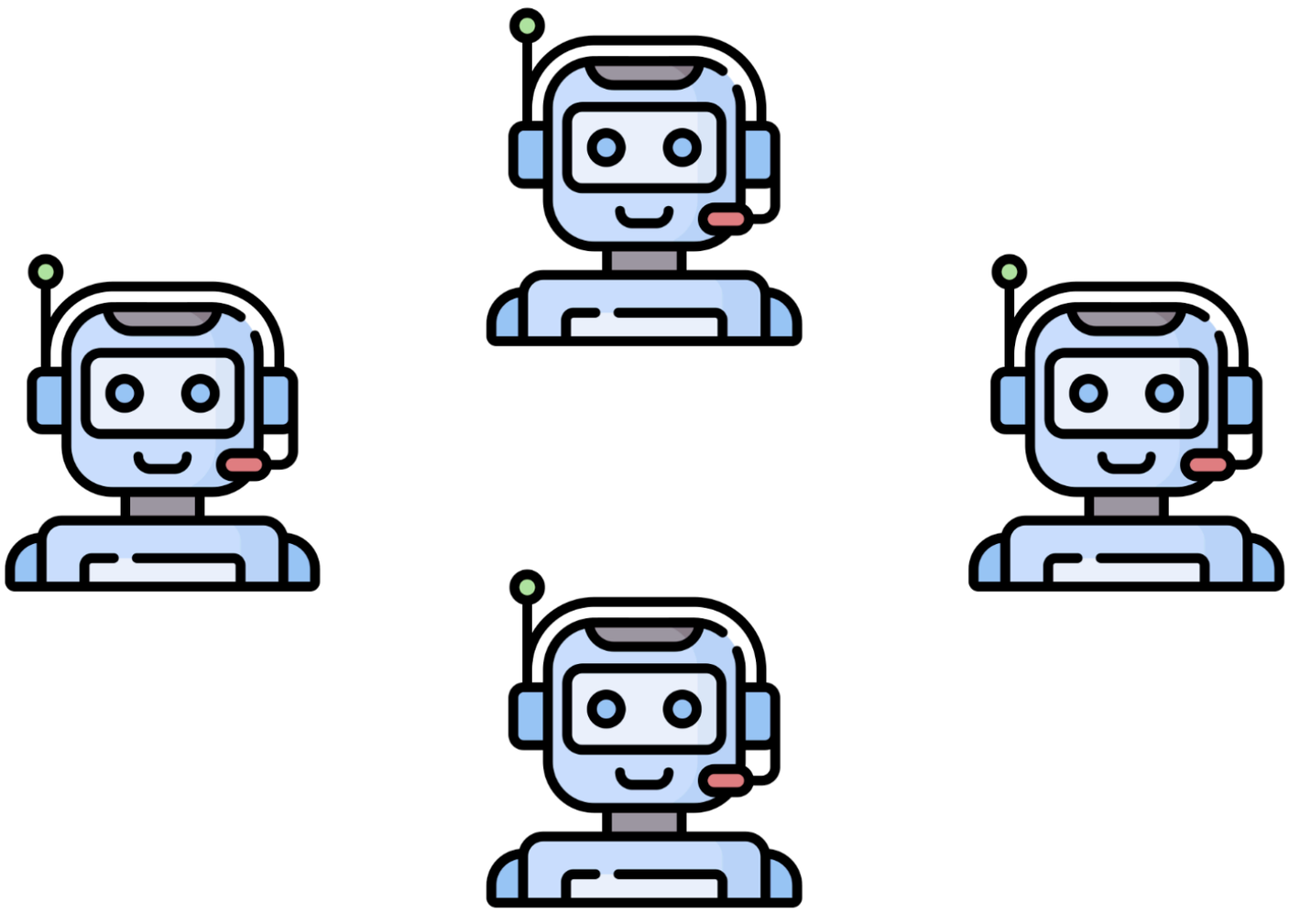
- CoAct goes back and fixes plans (Hou et al. 2024)



Multi-agent Systems

Multi-agent Systems

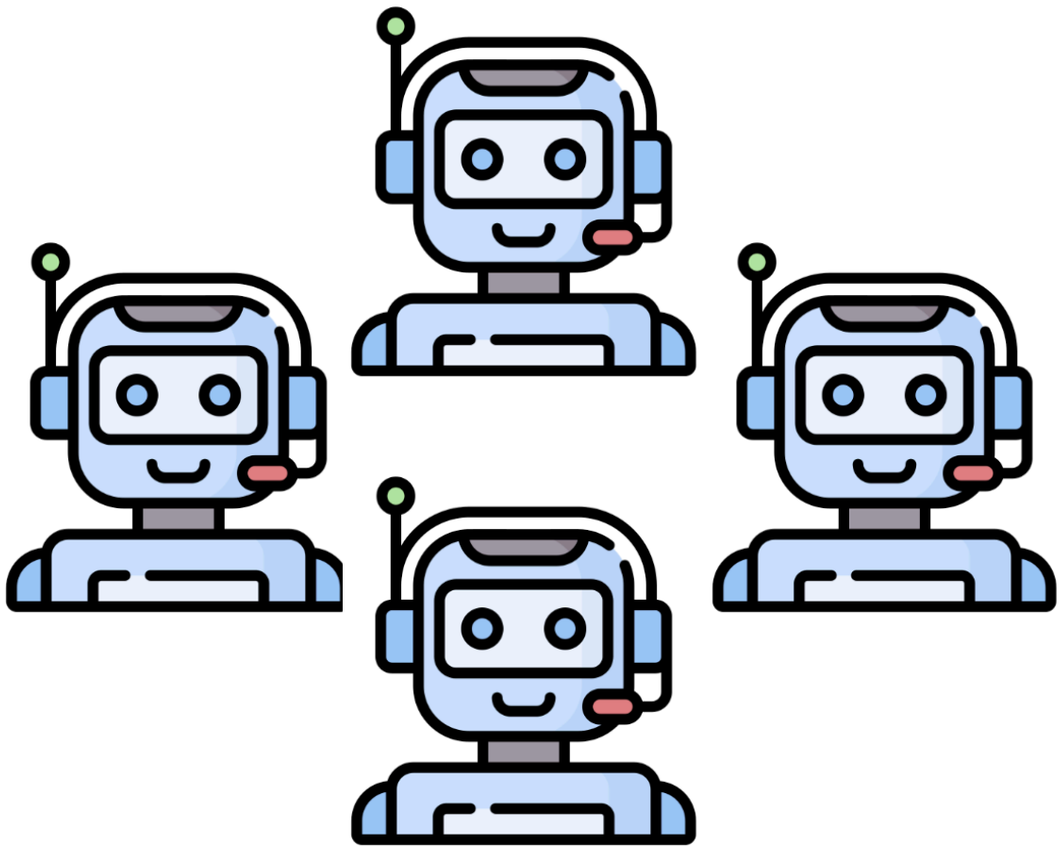
- When we have multiple agents interacting with each-other

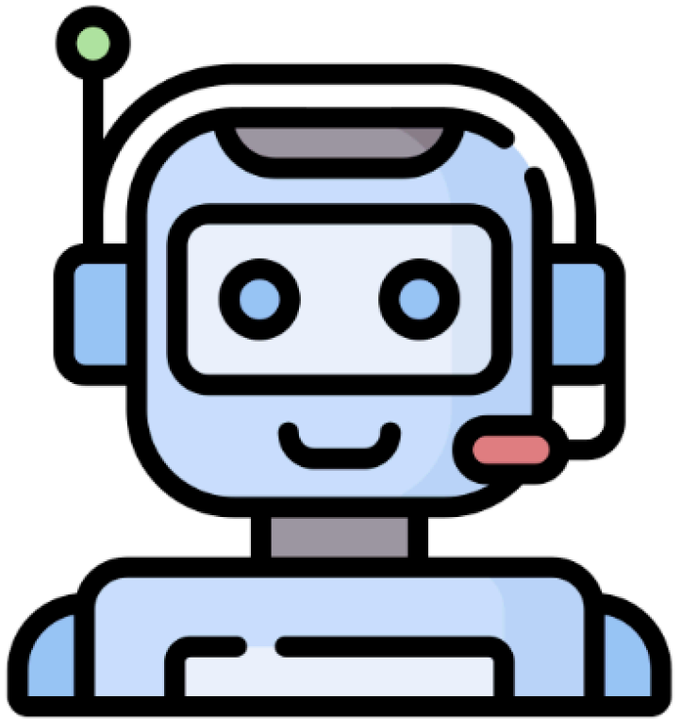


Why Multi-agent?

- Define problem structure
- Provide the right knowledge at the right time
- Switch between LLMs
- Provide security/safety
- Simulate human interactions

Advantages/Disadvantages





- Explicitly define structure, but can be overly rigid
- Works well on typical settings but not eventualities
- Allows for more flexible credentialing, but not a silver bullet

Single Agent Alternatives
(Neubig 2024)

- **Problem structure:** use a descriptive prompt
- **Provide the right knowledge:** use retrieval
- **Switch between LLMs:** use model routing
- **Provide security/safety:** credential the whole system, not individual parts
- **Simulate human interactions:** none, really

Questions?