CMU Agent Workshop 2024

Presenters: Frank Xu, Zora Wang, Graham Neubig

Carnegie Mellon University
Language Technologies Institute

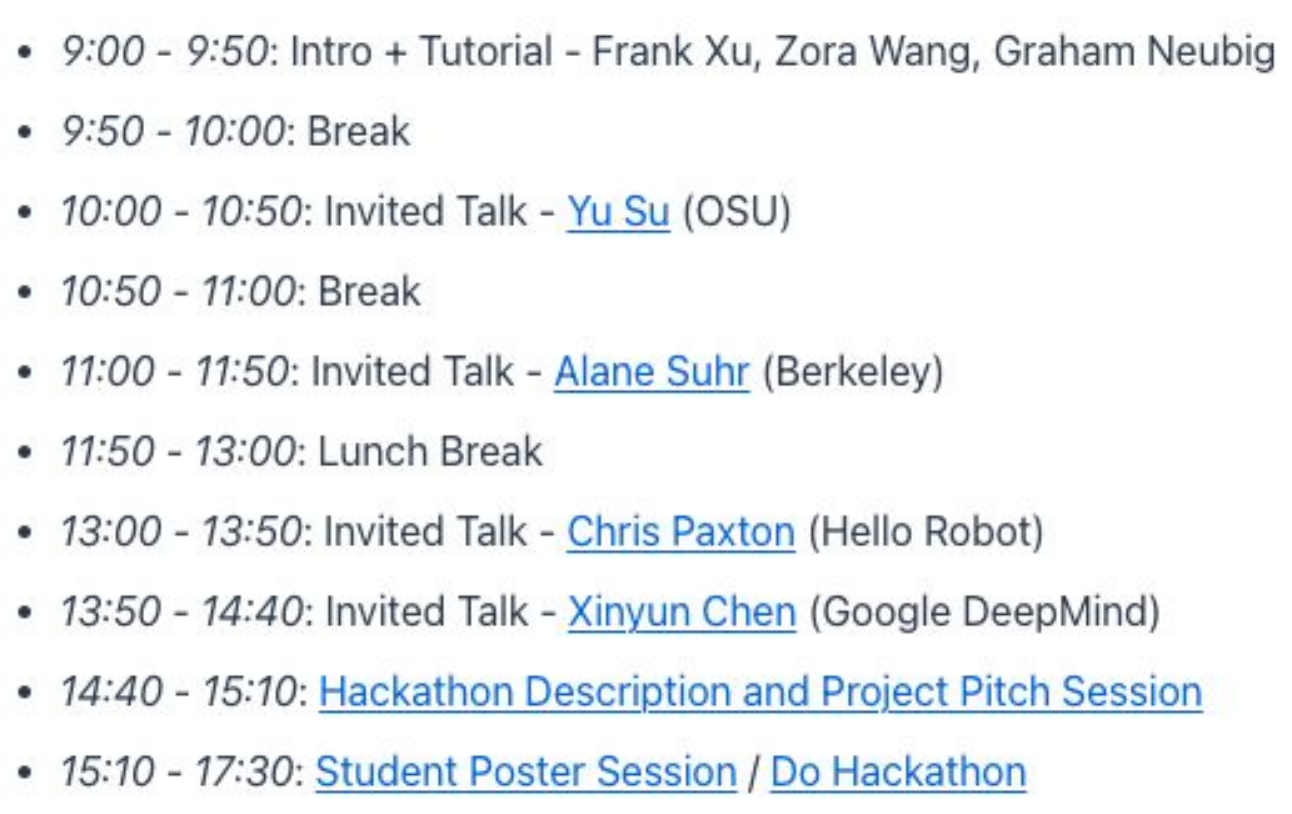
How can we make AI agents that work for all of us?

Talks, posters, hackathon, seminar!

Welcome to CMU Agent Workshop 2024

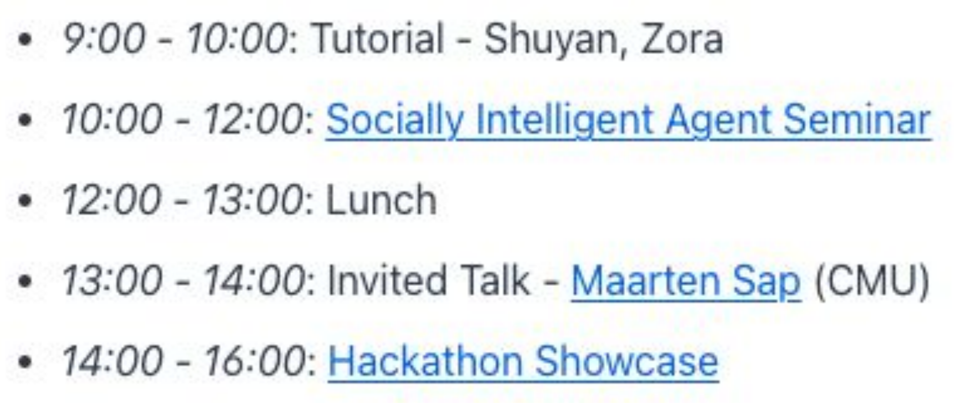
Schedule

May 2



- *9:00 - 9:50*: Intro + Tutorial - Frank Xu, Zora Wang, Graham Neubig
- *9:50 - 10:00*: Break
- *10:00 - 10:50*: Invited Talk - Yu Su (OSU)
- *10:50 - 11:00*: Break
- *11:00 - 11:50*: Invited Talk - Alane Suhr (Berkeley)
- *11:50 - 13:00*: Lunch Break
- *13:00 - 13:50*: Invited Talk - Chris Paxton (Hello Robot)
- *13:50 - 14:40*: Invited Talk - Xinyun Chen (Google DeepMind)
- *14:40 - 15:10*: Hackathon Description and Project Pitch Session
- *15:10 - 17:30*: Student Poster Session / Do Hackathon

May 3



- *9:00 - 10:00*: Tutorial - Shuyan, Zora
- *10:00 - 12:00*: Socially Intelligent Agent Seminar
- *12:00 - 13:00*: Lunch
- *13:00 - 14:00*: Invited Talk - Maarten Sap (CMU)
- *14:00 - 16:00*: Hackathon Showcase

Thank you to all the organizers!

Frank Xu, Zora Wang, So Yeon Tiffany Min, Xuhui Zhou, Karina Halevy,

Junhong Shen, Hao Zhu, Yiqing Xie, Shuyan Zhou, Maarten Sap, Graham Neubig

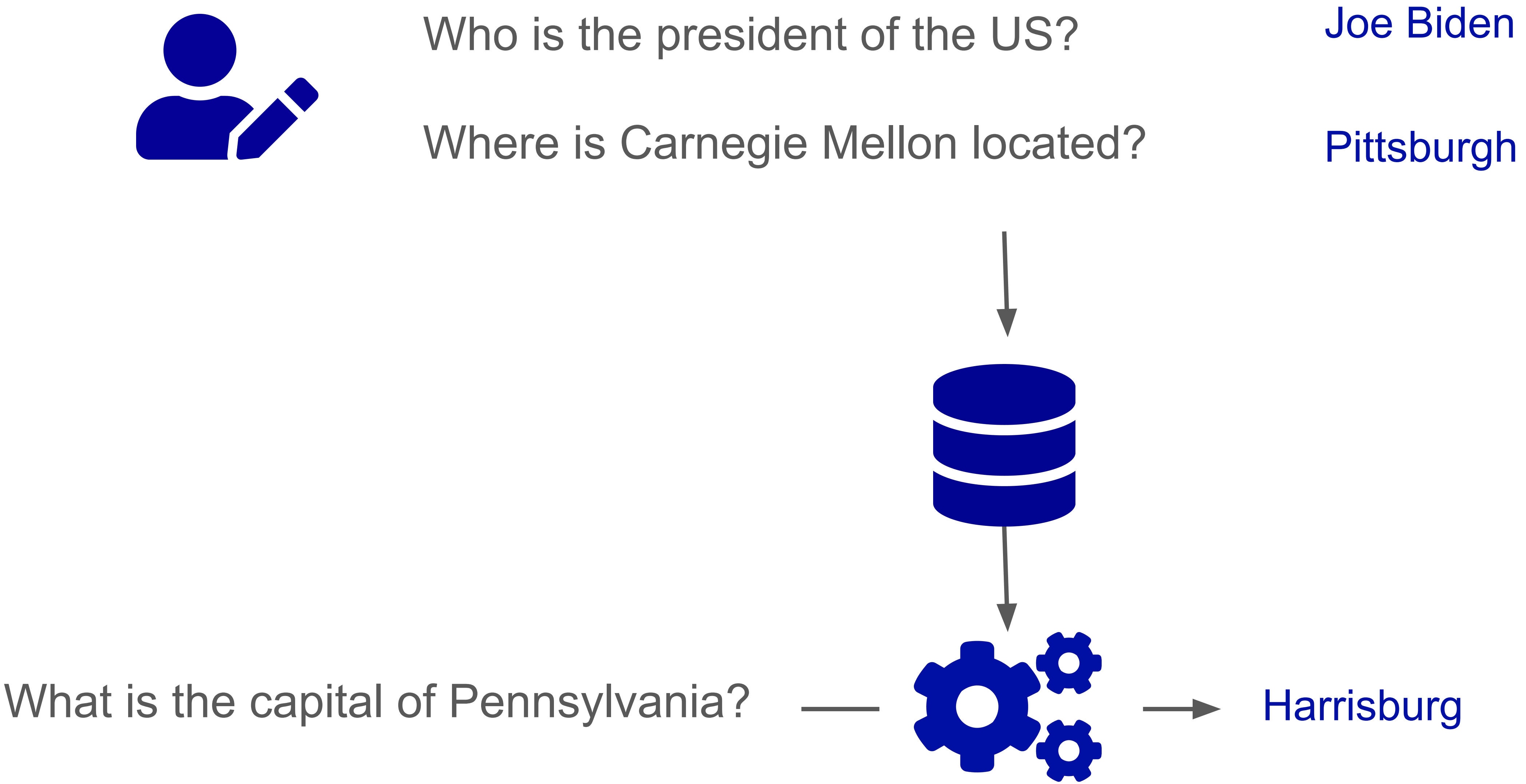
Thank you to our sponsor!

ardent.

Ardent Venture Partners

A Whirlwind Tour of Large Language Models

NLP Before Large Language Models (Train/Test)



Who is the president of the US? Joe Biden

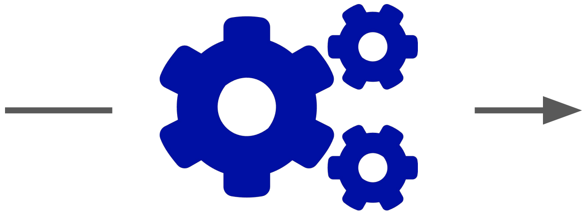
Where is Carnegie Mellon located? Pittsburgh

What is the capital of Pennsylvania? → Harrisburg

NLP With Large Language Models (Prompting)

Q: What is the capital of Pennsylvania?

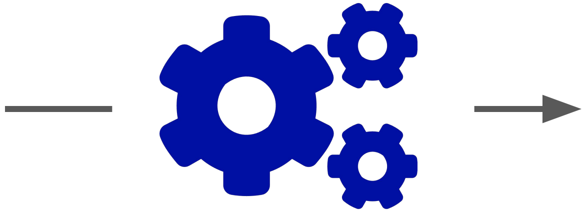
A:



Harrisburg

It’s for real this time. After months of legal drama, bad memes and will-they-or-won’t-they-chaos to put your favorite rom-com to shame, Elon Musk has closed his $44 billion acquisition of Twitter. Musk sealed the deal Thursday night, taking Twitter private and ousting a handful of top executives — CEO Parag Agrawal included — in the process.

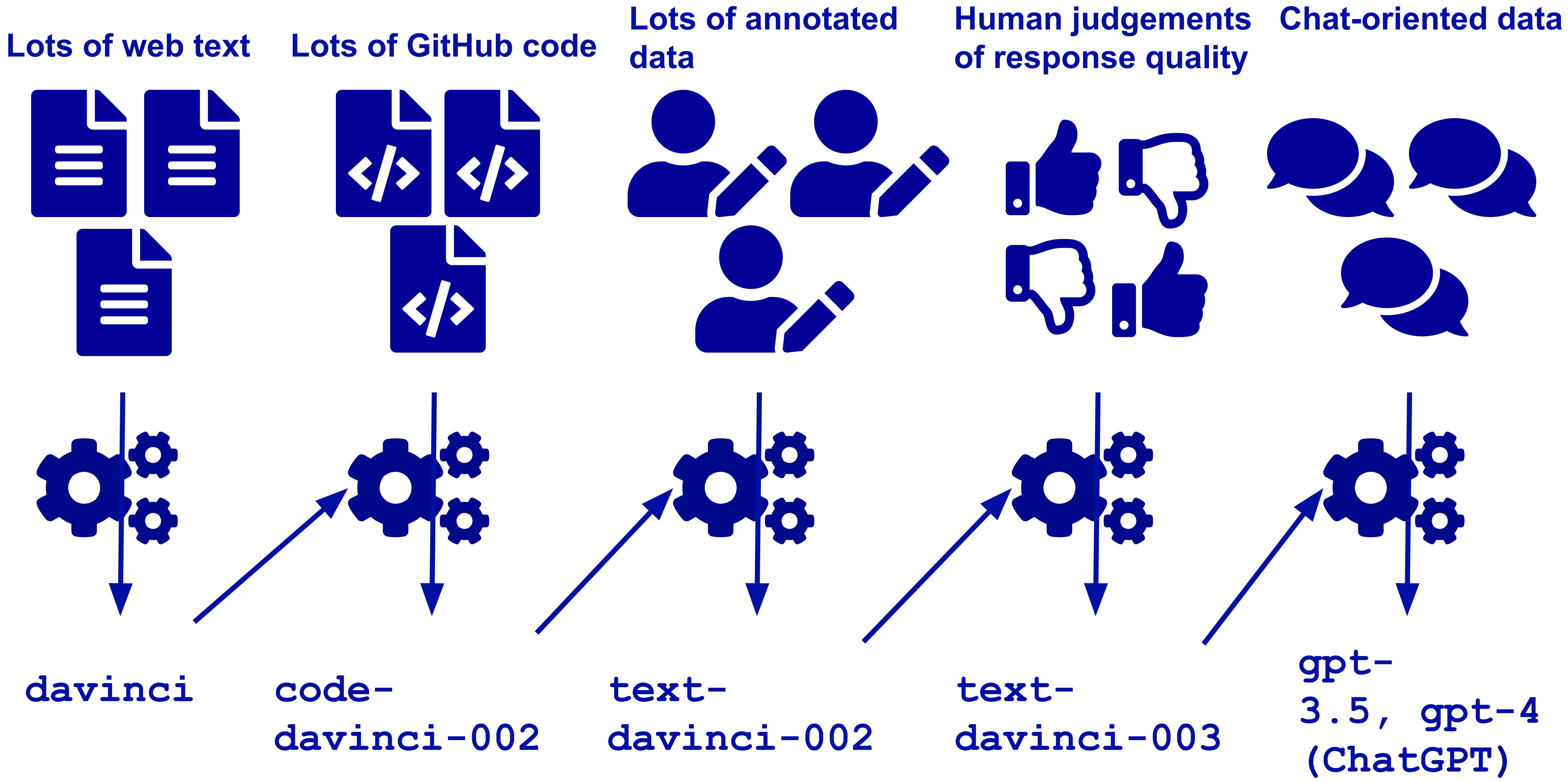
TL;DR:



Elon Musk has bought Twitter.

- Mostly through APIs such as GPT, Claude, Gemini or local models like Llama, Mistral, DeepSeek, Qwen

From Zero to ChatGPT



Lots of web text

Lots of GitHub code

Lots of annotated data

Human judgements of response quality

Chat-oriented data

`davinci`

`code-davinci-002`

`text-davinci-002`

`text-davinci-003`

`gpt-3.5, gpt-4 (ChatGPT)`

Augmenting LLMs with Tools

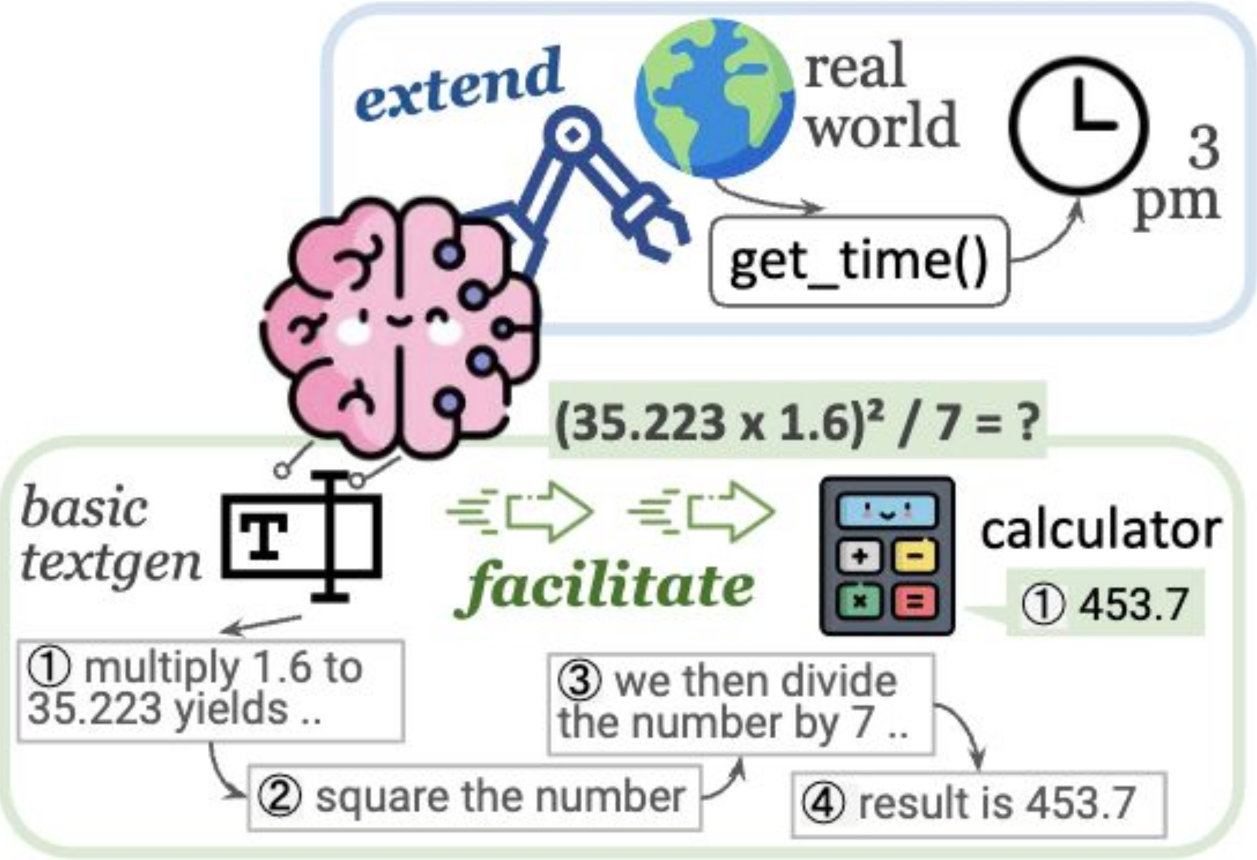
LMs are powerful for text generation tasks.
But …

- Complex reasoning?

  Struggle

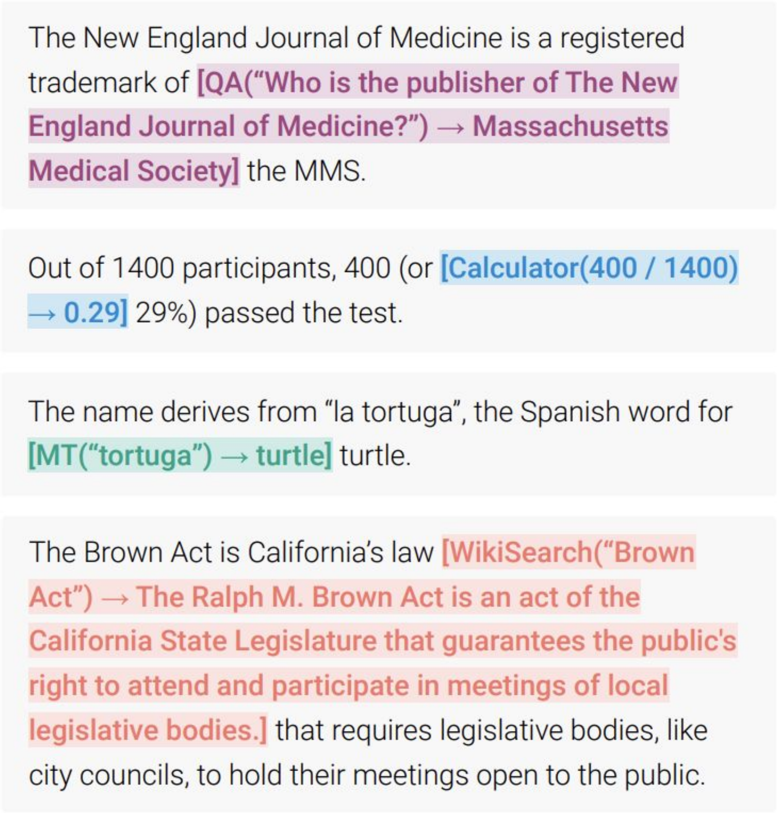
- Access real-world information?

  Fundamentally unable



Tools benefit language models a lot

- ToolFormer

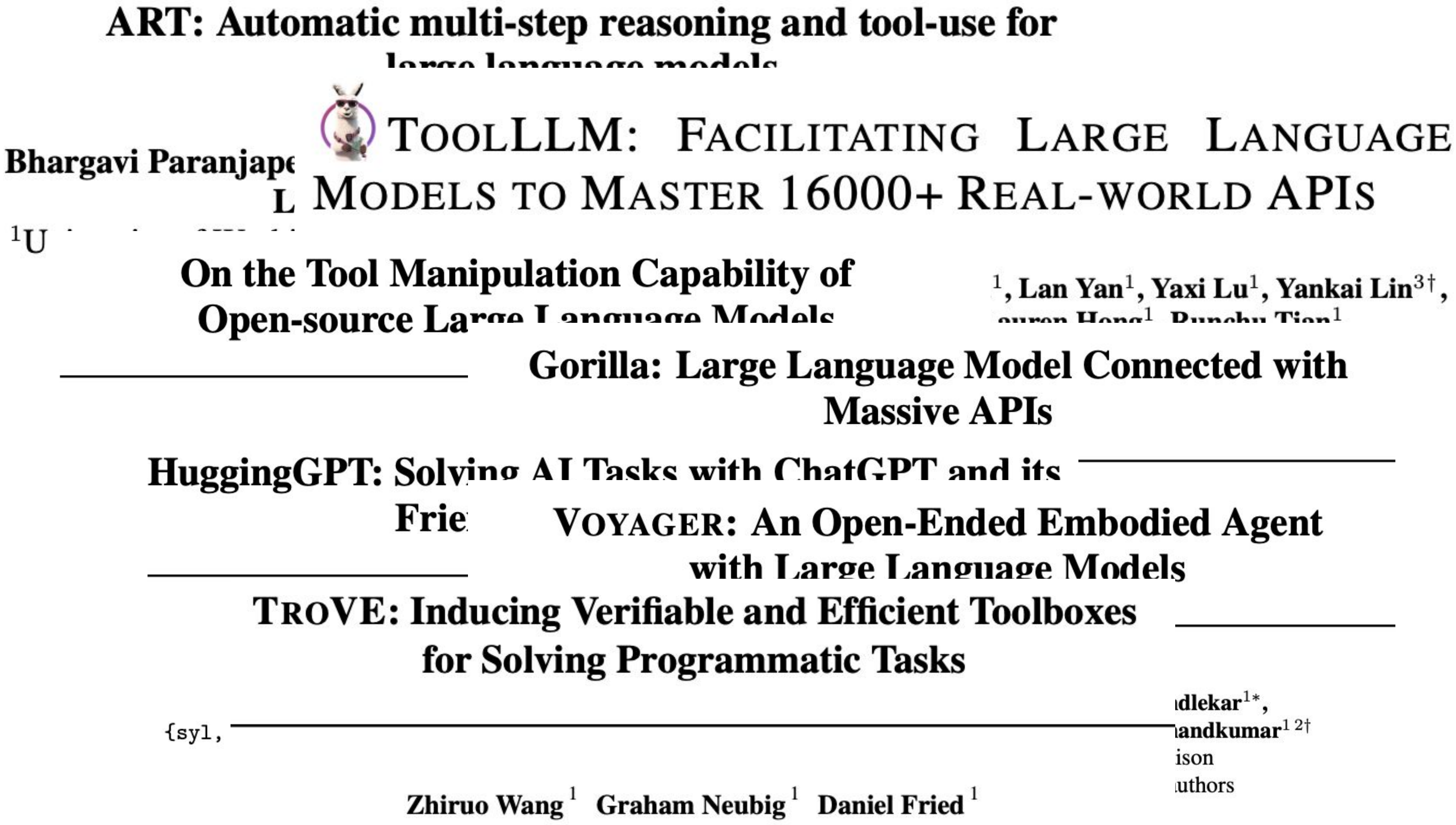


The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

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.



ART: Automatic multi-step reasoning and tool-use for

Bhargavi Paranjape

L

1U

ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs

On the Tool Manipulation Capability of Open-source Large Language Models

1, Lan Yan1, Yaxi Lu1, Yankai Lin3†,

Gorilla: Large Language Model Connected with Massive APIs

HuggingGPT: Solving AI Tasks with ChatGPT and its

Frie

Voyager: An Open-Ended Embodied Agent with Large Language Models

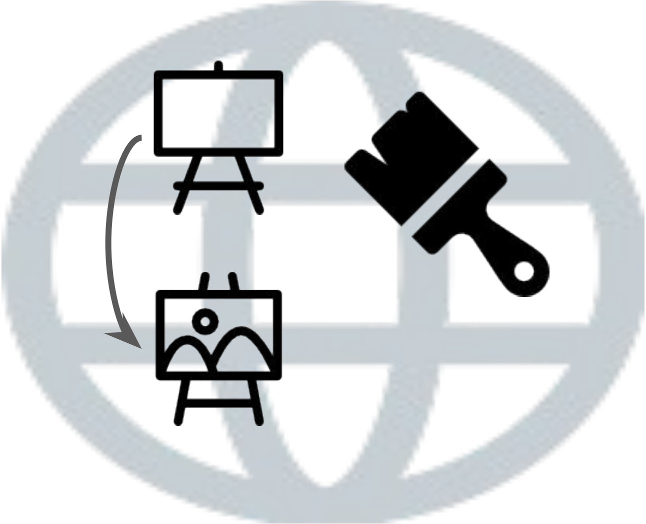
TroVE: Inducing Verifiable and Efficient Toolboxes for Solving Programmatic Tasks

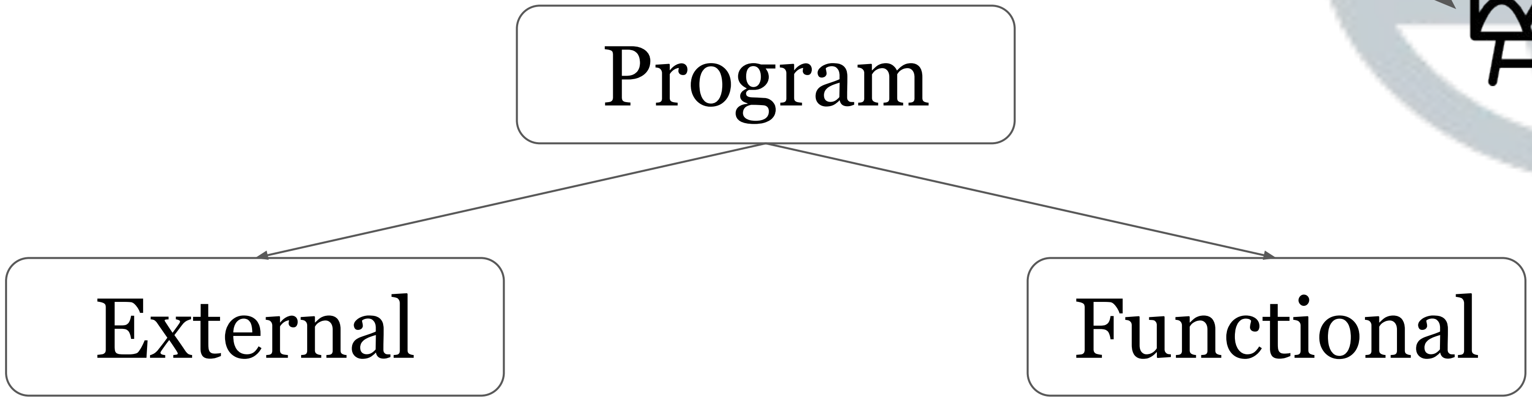
{syl,

dlekar1*,
andkumar1 2†
ison
uthors

Zhiruo Wang1 Graham Neubig1 Daniel Fried1

Tool Basics: Definition





Program

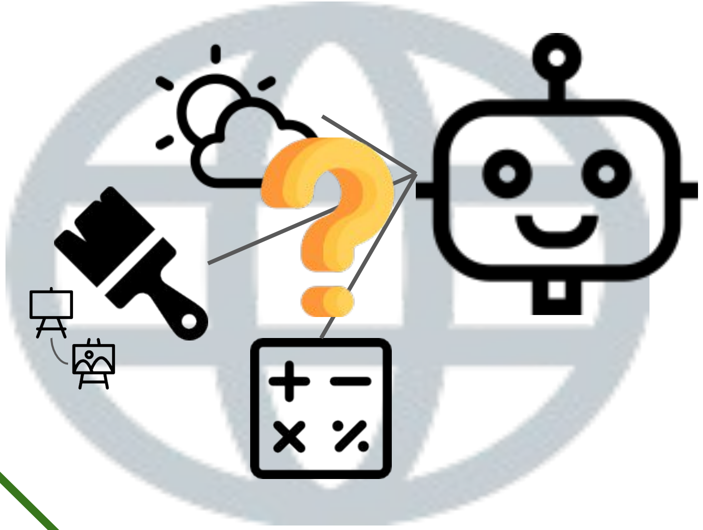
External

Functional

*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 Basics: Functionality



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

Tools

*Agents: anything that can be viewed as **perceiving** its environment through sensors and **acting** upon that environment through actuators*[1].

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

The Basic Tool Use Paradigm

Tool Use: switching between

- text-generation mode
- tool-execution mode

Tool Learning:

- inference-time prompting
- learning by training

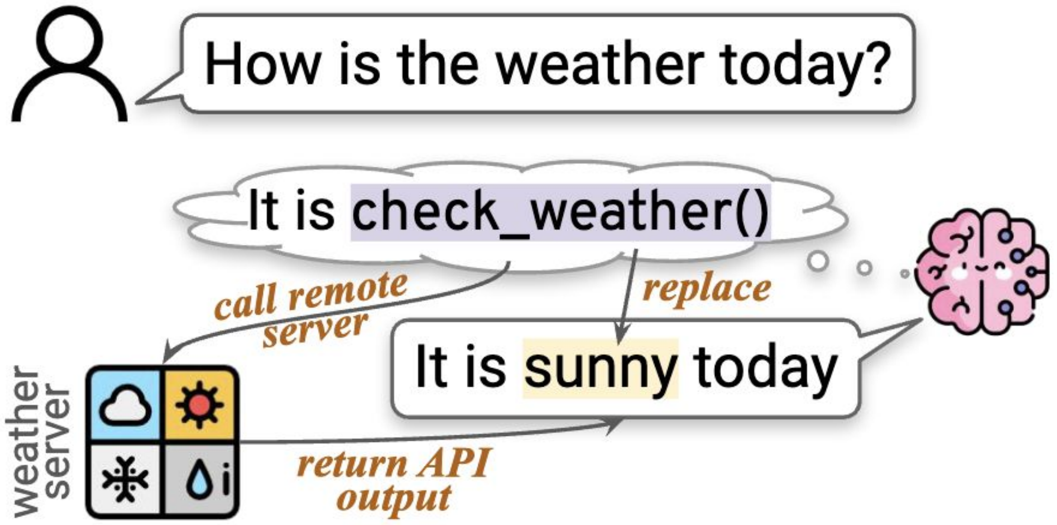
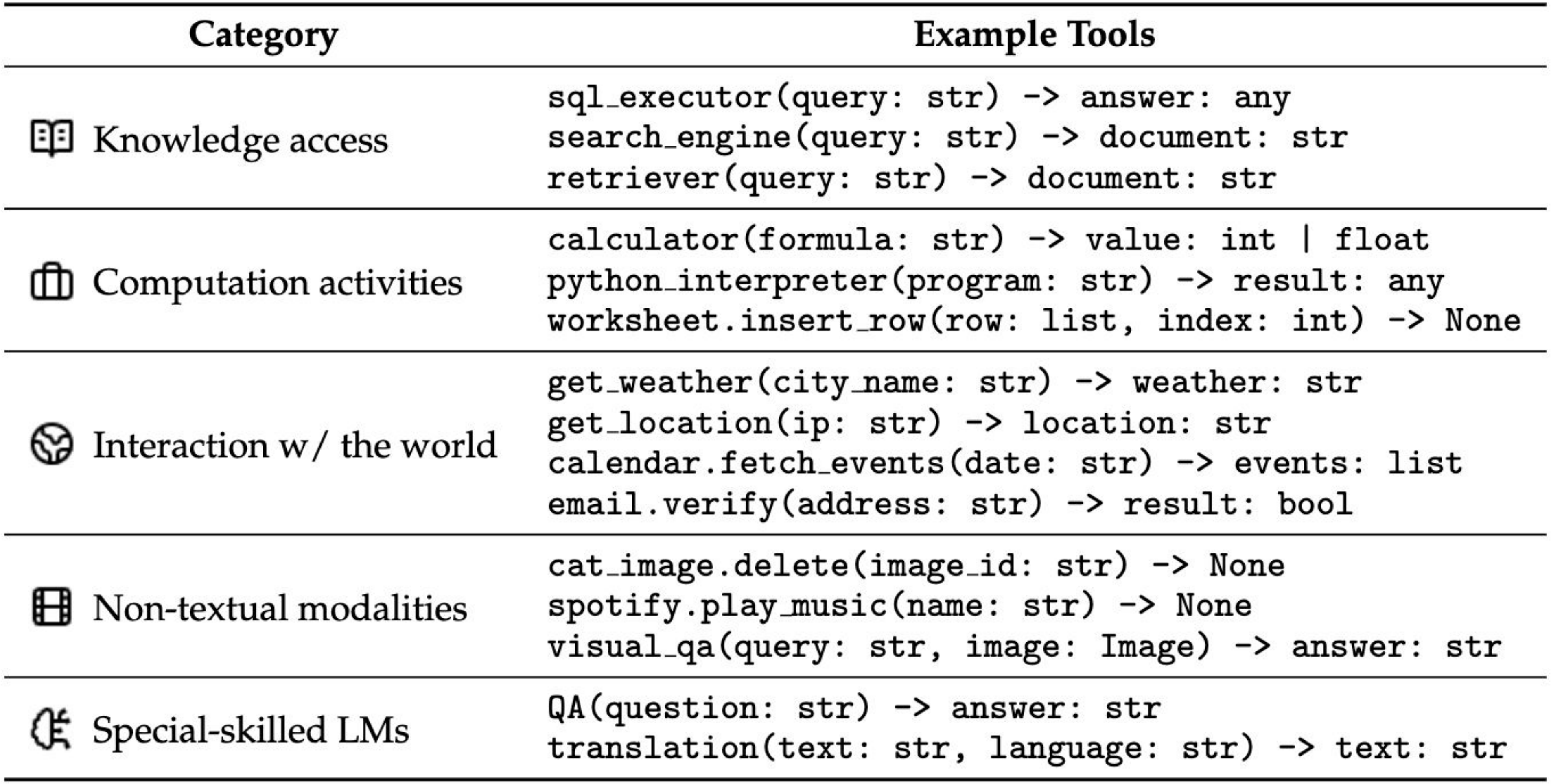


Figure 2: The basic tool use paradigm. LM calls `check_weather` tool by generating text tokens. This call triggers the server to execute the call and return the output `sunny`, using which the LM replaces the API call tokens in the response to the user.

Scenarios of LM Tool Using



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`

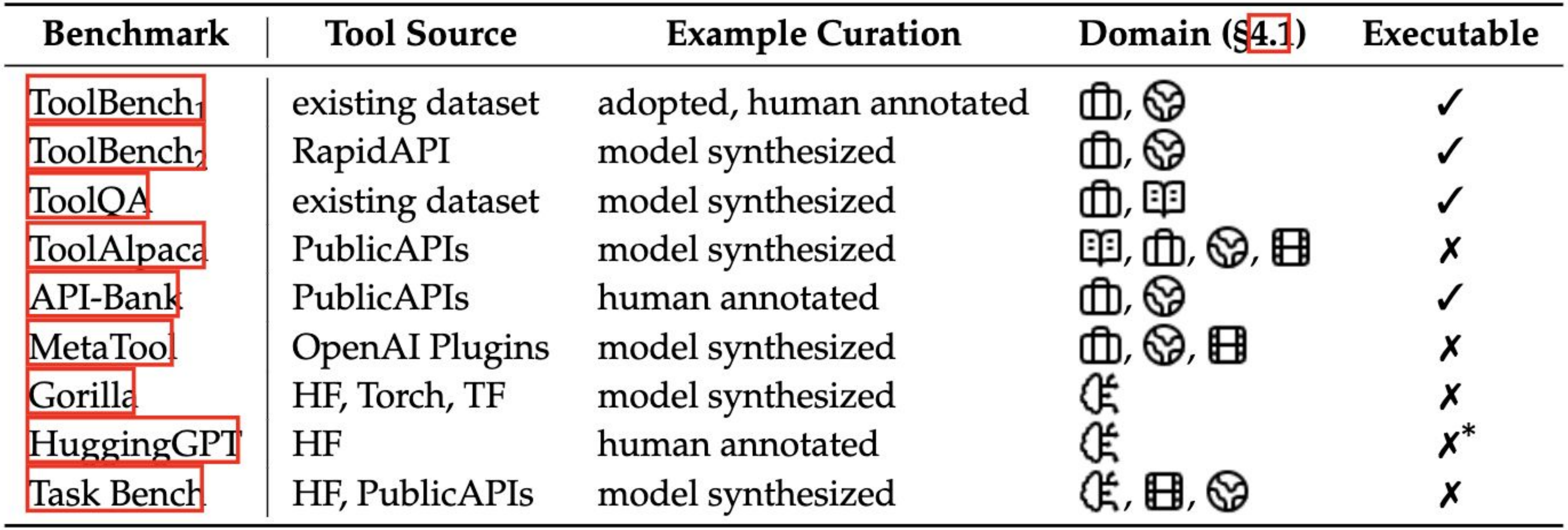
Interaction w/ the world	`get_weather(city_name: str) -> weather: str`
	`get_location(ip: str) -> location: str`
	`calendar.fetch_events(date: str) -> events: list`
	`email.verify(address: str) -> result: bool`

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.

How to evaluate tool use?



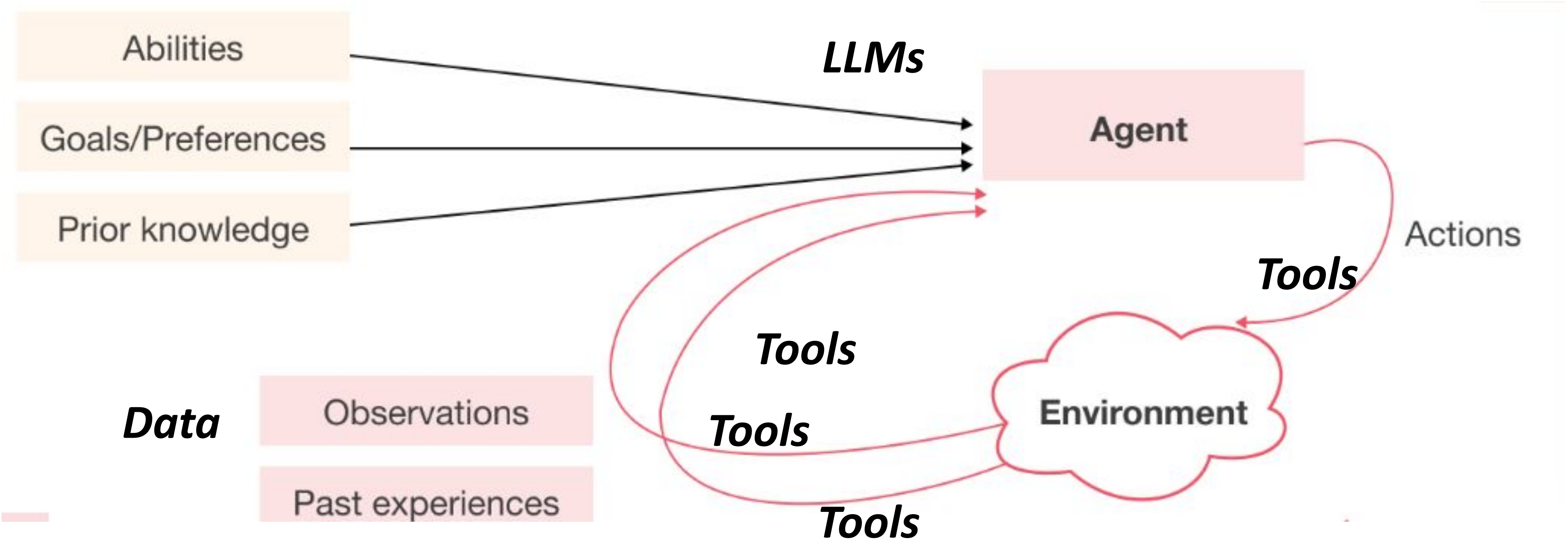
Benchmark	Tool Source	Example Curation	Domain (§4.1)	Executable
ToolBench$_1$	existing dataset	adopted, human annotated	🧰, 🌐	✓
ToolBench$_2$	RapidAPI	model synthesized	🧰, 🌐	✓
ToolQA	existing dataset	model synthesized	🧰, 📖	✓
ToolAlpaca	PublicAPIs	model synthesized	📖, 🧰, 🌐, 🎞	✗
API-Bank	PublicAPIs	human annotated	🧰, 🌐	✓
MetaTool	OpenAI Plugins	model synthesized	🧰, 🌐, 🎞	✗
Gorilla	HF, Torch, TF	model synthesized	🧠	✗
HuggingGPT	HF	human annotated	🧠	✗*
Task Bench	HF, PublicAPIs	model synthesized	🧠, 🎞, 🌐	✗

- ools
  - Tasks that necessitate tools
  - Issues ☹
    - Naturalness
    - Executability
  - Reproducible testing
  - Safe usage

From Tool Use to Agents

What are agents?

Anything that can be viewed as **perceiving** its environment through sensors and **acting** upon that environment through actuators.



How to get started in LLM Agents

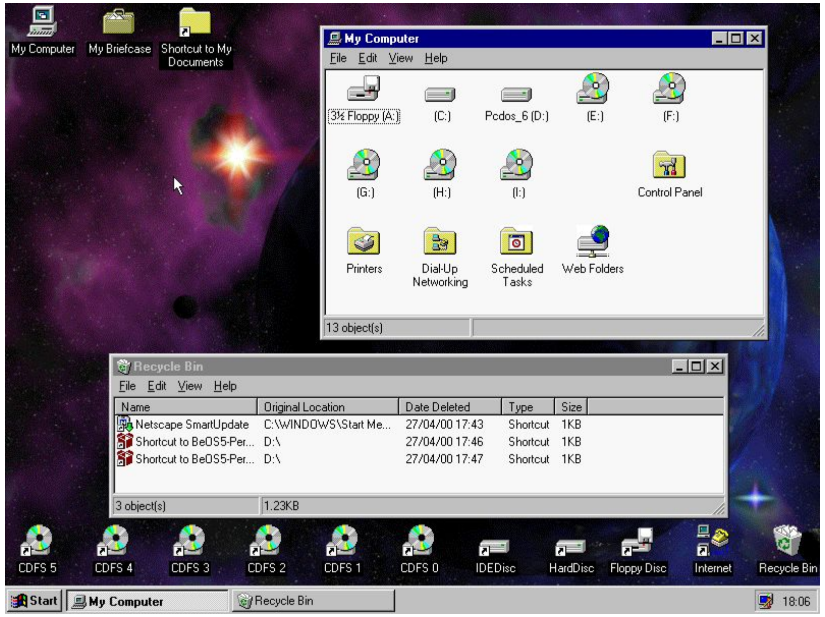
- Tasks and Applications
- Methods for Building Agents

Tasks and Applications For LLM Agents

Why do we want agents?

Imagine if things get done by just talking...

How Do People Interact with Computers?

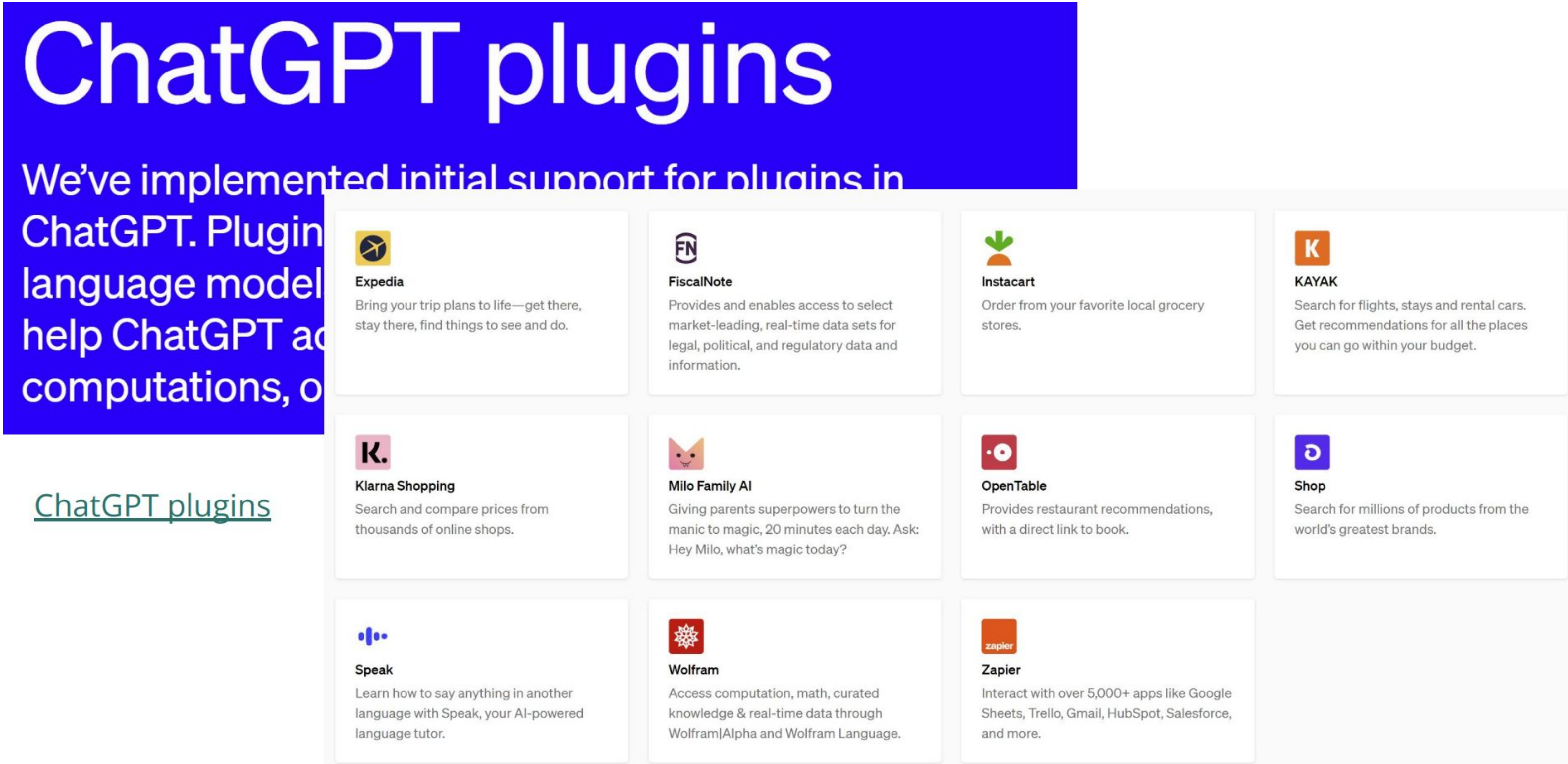


Name	Original Location	Date Deleted	Type	Size
Netscape SmartUpdate	C:\WINDOWS\Start Me...	27/04/00 17:43	Shortcut	1KB
Shortcut to BeOS5-Per...	D:\	27/04/00 17:46	Shortcut	1KB
Shortcut to BeOS5-Per...	D:\	27/04/00 17:47	Shortcut	1KB

- What about using natural language?

- Save time, natural, accessible, no need to browse, no programming learning curve, etc.

Tool Integrations into Chatbots



ChatGPT plugins

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

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.

ChatGPT plugins

Embodied and Robots

Alfred and AlfWorld

You are in the middle of a room. Looking quickly around you, you see a safe 1, a shelf 4, … a garbage can 1.

Your task is to: examine an alarm clock

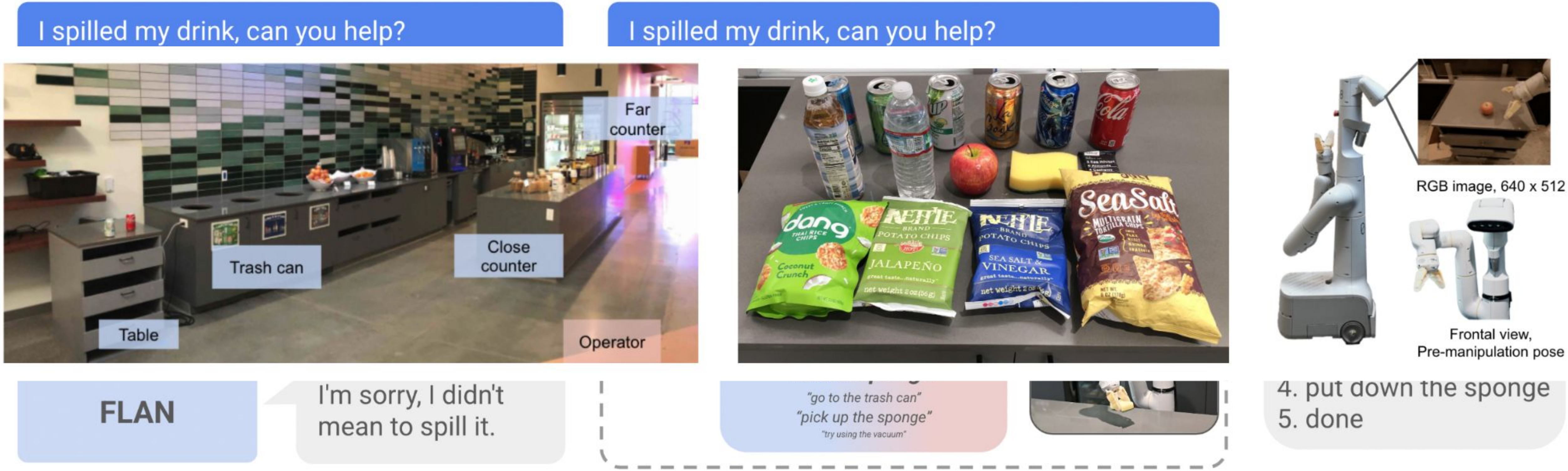
> go to desk 1

You arrive at loc 8. On the desk 1, you see a pen 1, a bowl 1, a alarm clock 2 …

> take alarmclock 2 from desk 1

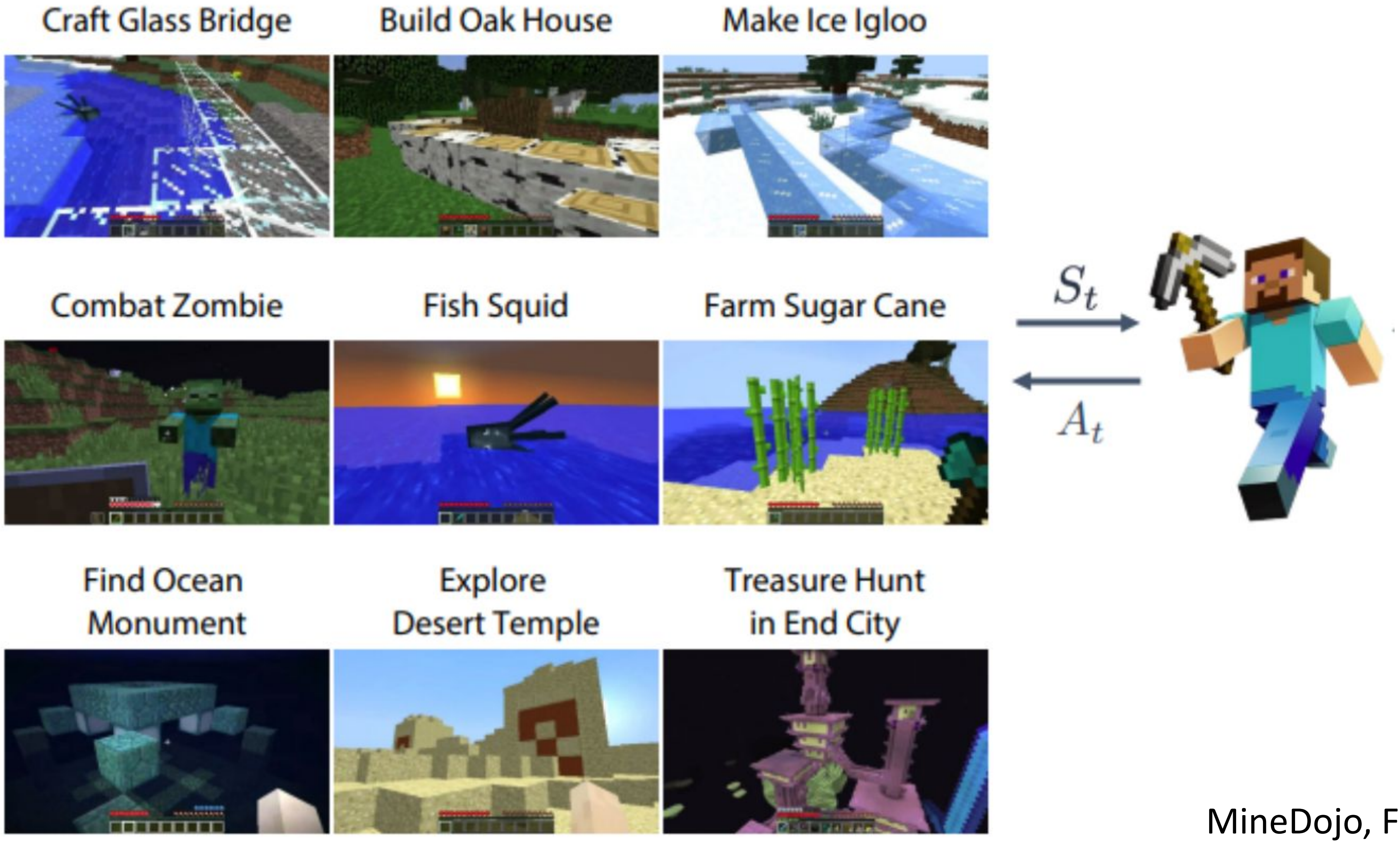
You pick up the alarm clock 2 from the desk 1.

Embodied and Robots



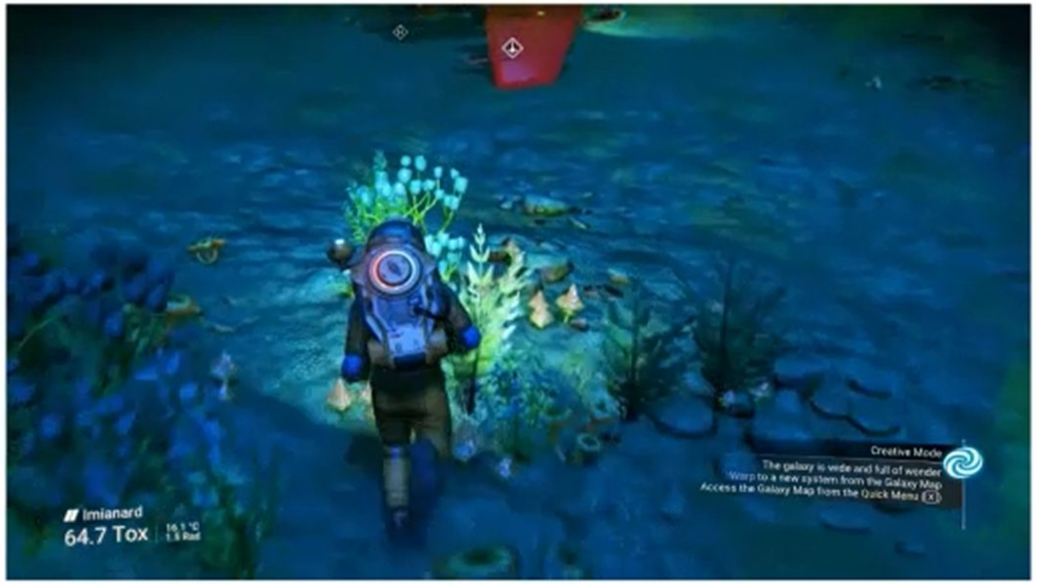
SayCan, Google et al. 22'

Games

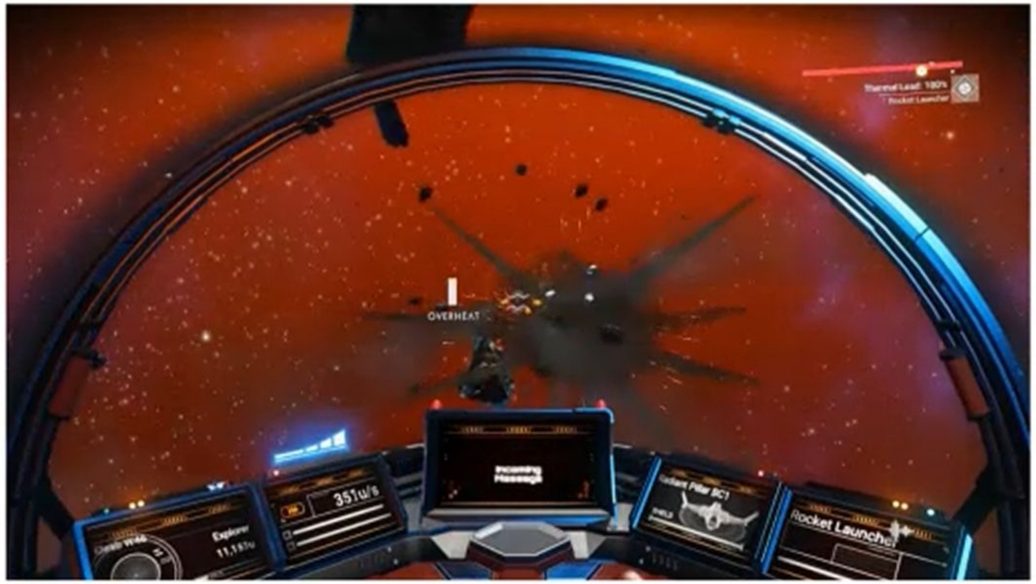


MineDojo, Fan et.al 22’

Games


Creative Mode
The galaxy is wide and full of wonder
Warp to a new system from the Galaxy Map
Access the Galaxy Map from the Quick Menu
Imianard
64.7 Tox

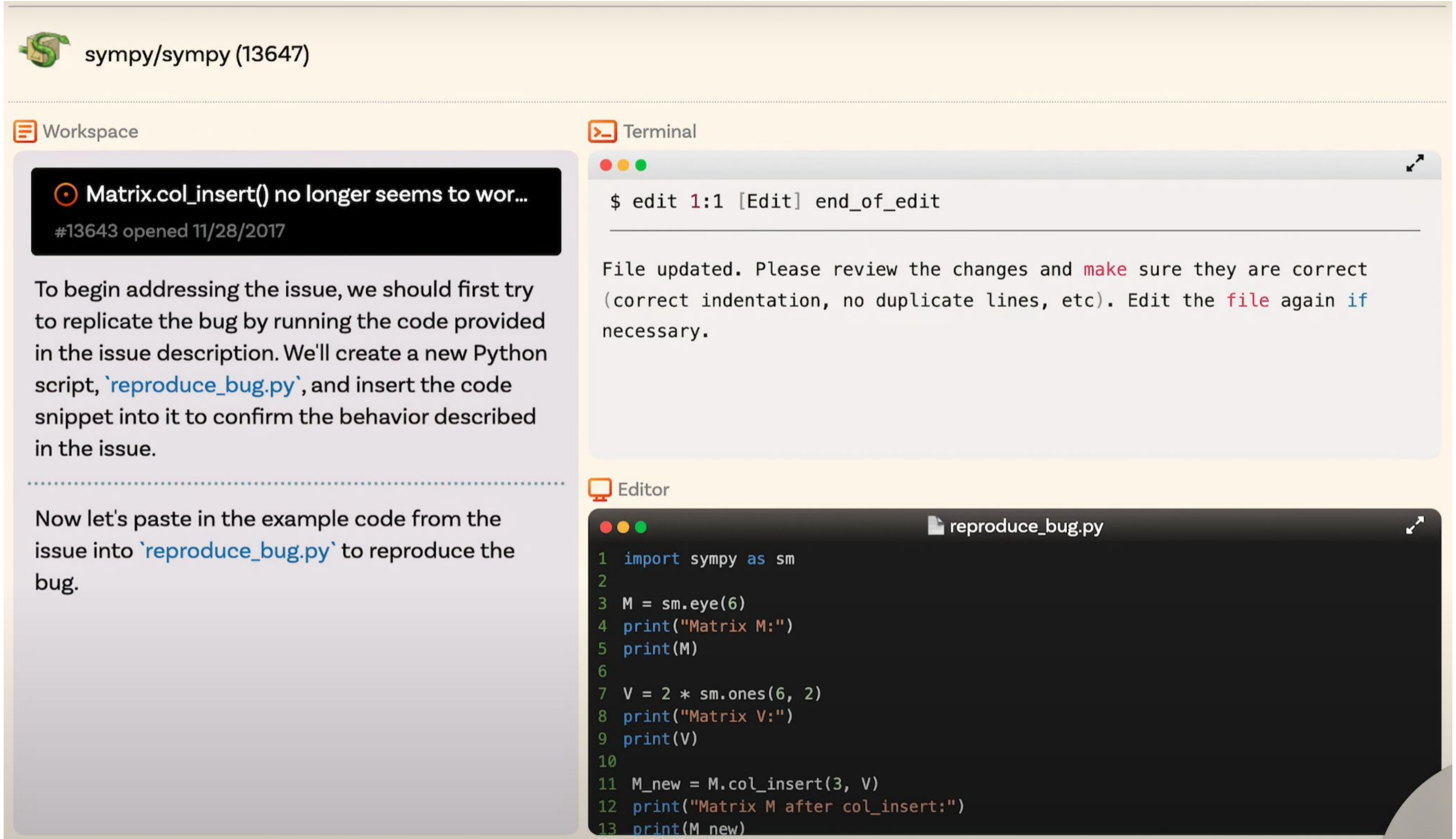
No Man’s Sky: Go to spaceship



No Man’s Sky: Shoot Asteroid

SIMA, DeepMind SIMA Team 24’

Software Development



Matrix.col_insert() no longer seems to wor...

#13643 opened 11/28/2017

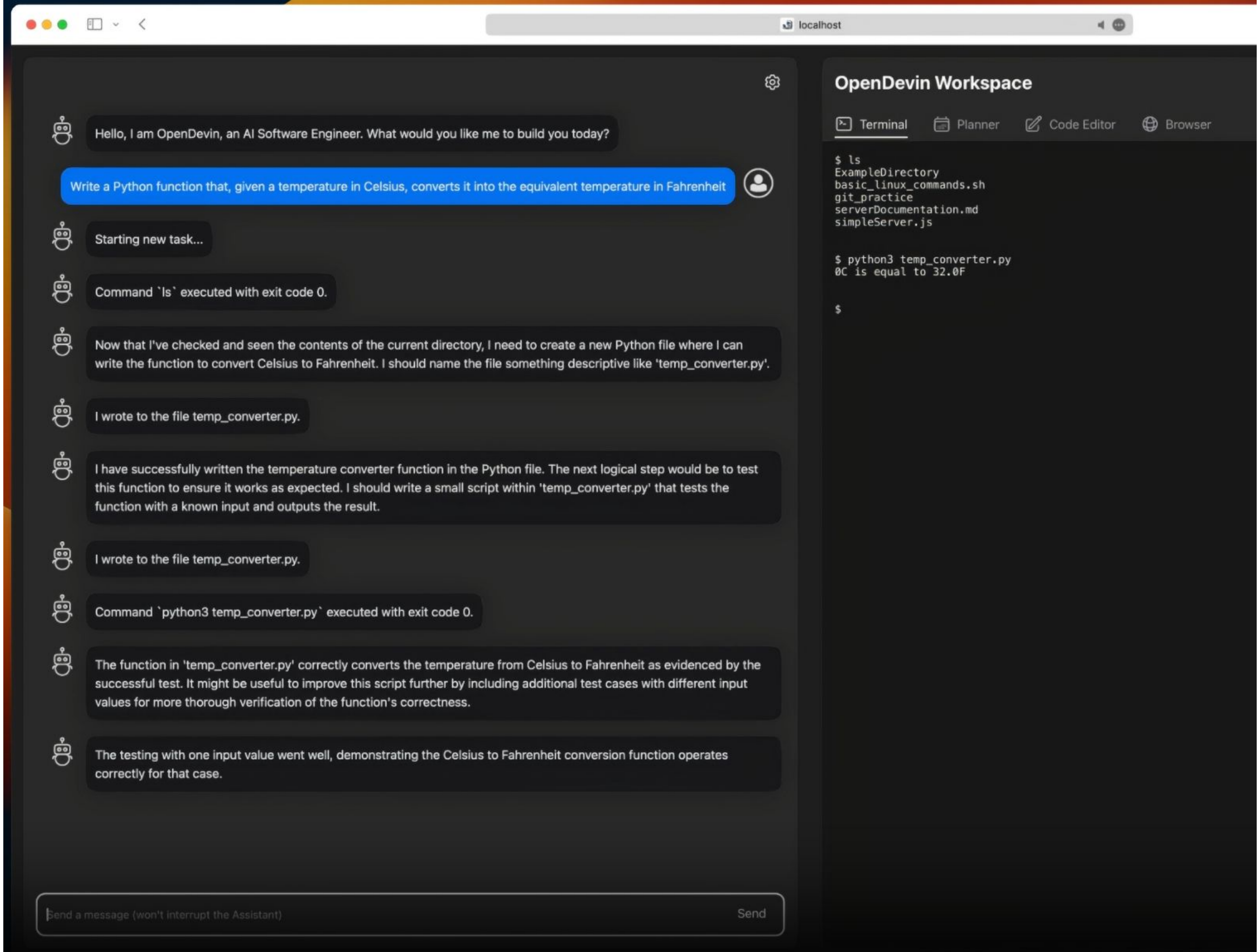
To begin addressing the issue, we should first try to replicate the bug by running the code provided in the issue description. We'll create a new Python script, `reproduce_bug.py`, and insert the code snippet into it to confirm the behavior described in the issue.

Now let's paste in the example code from the issue into `reproduce_bug.py` to reproduce the bug.

```
$ edit 1:1 [Edit] end_of_edit

File updated. Please review the changes and make sure they are correct
(correct indentation, no duplicate lines, etc). Edit the file again if
necessary.
```

```python
import sympy as sm

M = sm.eye(6)
print("Matrix M:")
print(M)

V = 2 * sm.ones(6, 2)
print("Matrix V:")
print(V)

 M_new = M.col_insert(3, V)
 print("Matrix M after col_insert:")
 print(M_new)
```

SWE-Agent, Princeton NLP 24’



OpenDevin Workspace

Hello, I am OpenDevin, an AI Software Engineer. What would you like me to build you today?

Write a Python function that, given a temperature in Celsius, converts it into the equivalent temperature in Fahrenheit

Starting new task...

Command `ls` executed with exit code 0.

Now that I've checked and seen the contents of the current directory, I need to create a new Python file where I can write the function to convert Celsius to Fahrenheit. I should name the file something descriptive like 'temp_converter.py'.

I wrote to the file temp_converter.py.

I have successfully written the temperature converter function in the Python file. The next logical step would be to test this function to ensure it works as expected. I should write a small script within 'temp_converter.py' that tests the function with a known input and outputs the result.

I wrote to the file temp_converter.py.

Command `python3 temp_converter.py` executed with exit code 0.

The function in 'temp_converter.py' correctly converts the temperature from Celsius to Fahrenheit as evidenced by the successful test. It might be useful to improve this script further by including additional test cases with different input values for more thorough verification of the function's correctness.

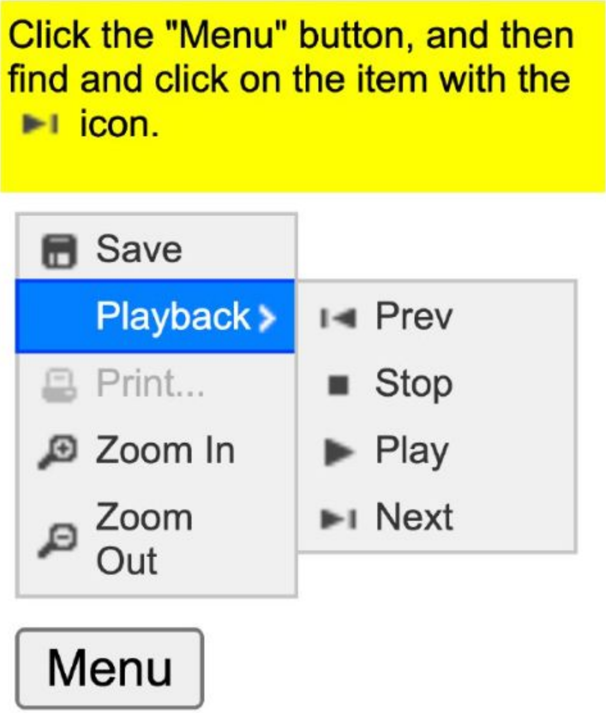
The testing with one input value went well, demonstrating the Celsius to Fahrenheit conversion function operates correctly for that case.

Send a message (won't interrupt the Assistant)

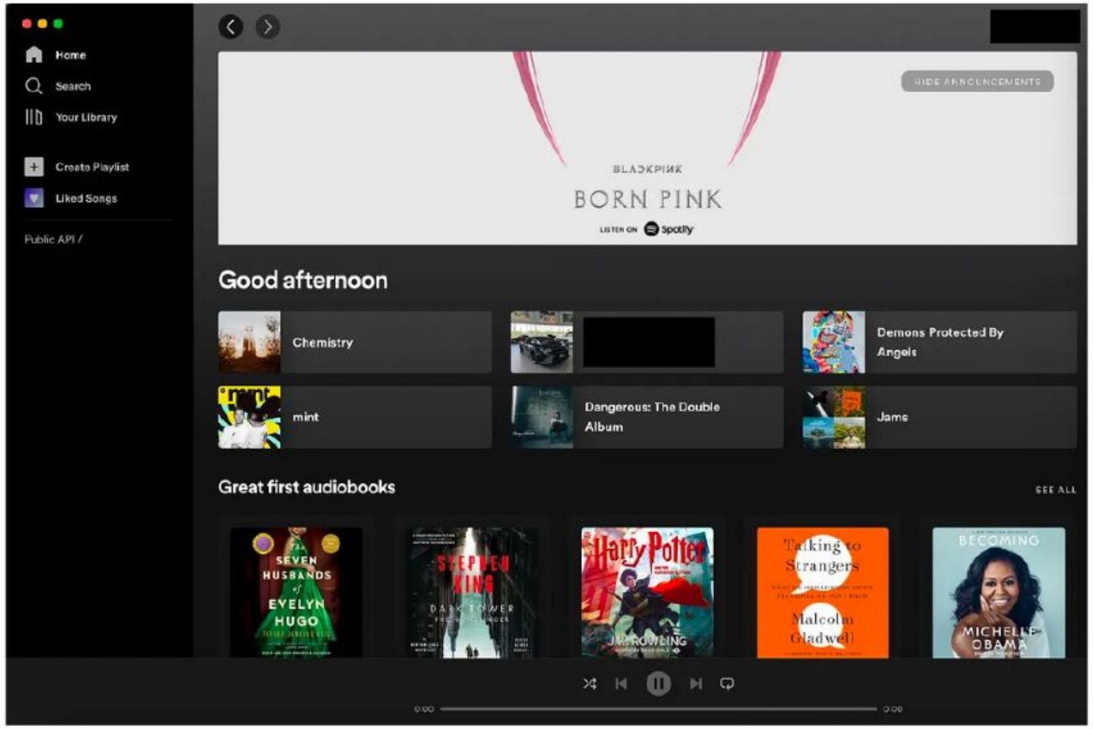
```
$ ls
ExampleDirectory
basic_linux_commands.sh
git_practice
serverDocumentation.md
simpleServer.js


$ python3 temp_converter.py
0C is equal to 32.0F


$
```

Check out OpenDevin in Hackathon!

UI Automation



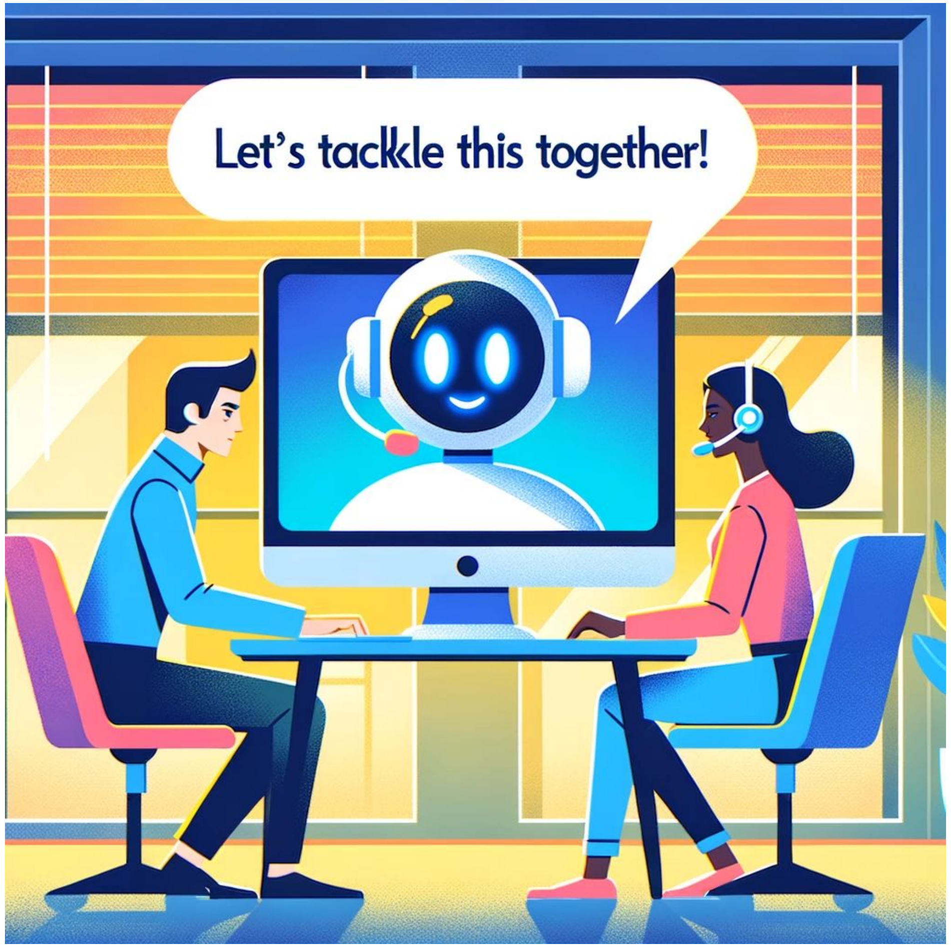
Click the "Menu" button, and then find and click on the item with the ⏭ icon.



"Play some synthwave songs"

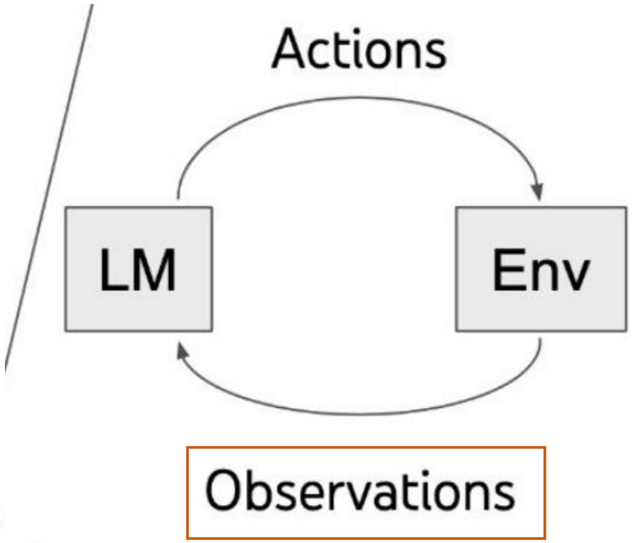
Why Web Agents?

- Many productive tasks we perform today are done on the computer
  - And many of these are on the web
- Many opportunities to automate menial tasks
- Augment human capabilities



Methods for Building Agents

How to let LLM become an agent

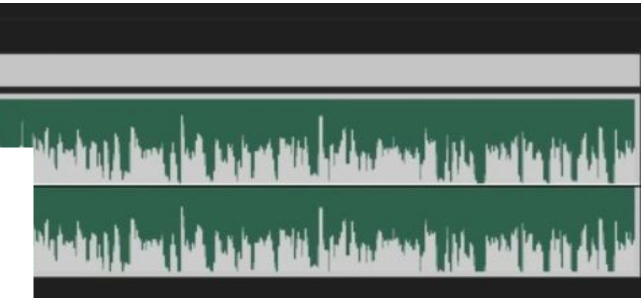


- Observation
  - Text input

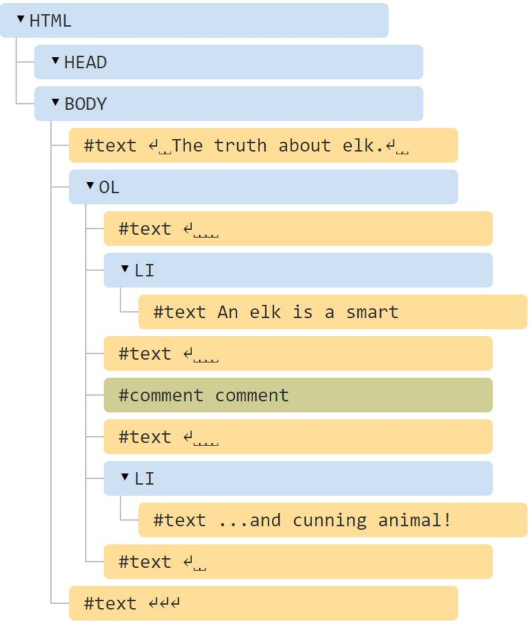
You are in the middle of a room. Looking quickly around you, you see a cabinet 6, a cabinet 1, a coffee machine 1, a countertop 3, a stove burner 1, and a toaster 1.

  - Visual Input

  - Audio Input



  - Structured Input

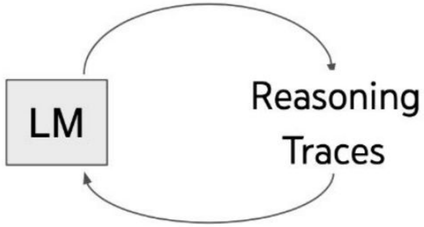


- Need for Multimodal LLMs

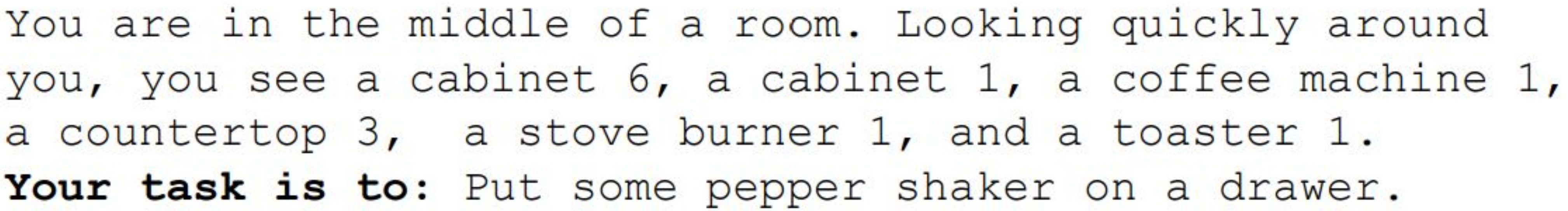
How to let LLM become an agent

- Planning and reasoning ability

  Chain-of-thoughts (CoT)
  "Let's think step by step …."



CoT (Wei et al. 22')



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.

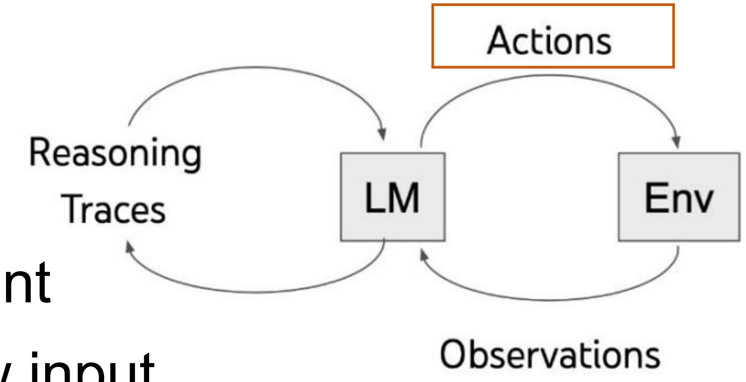
Ask LLM:
What should I do next? Let’s think step by step:

First I need to find a pepper shaker … more likely to appear in cabinets (1-6), countertops (1-3) …

After I find pepper shaker 1, next I need to put it on drawer 1 …..

How to let LLM become an agent

- Tool-use ability
  - Generate action calls
  - Execute the actions in environment
  - Put new observation back as new input



Toolformer (Schick et al. 23')
ReAct (Yao et al. 23')

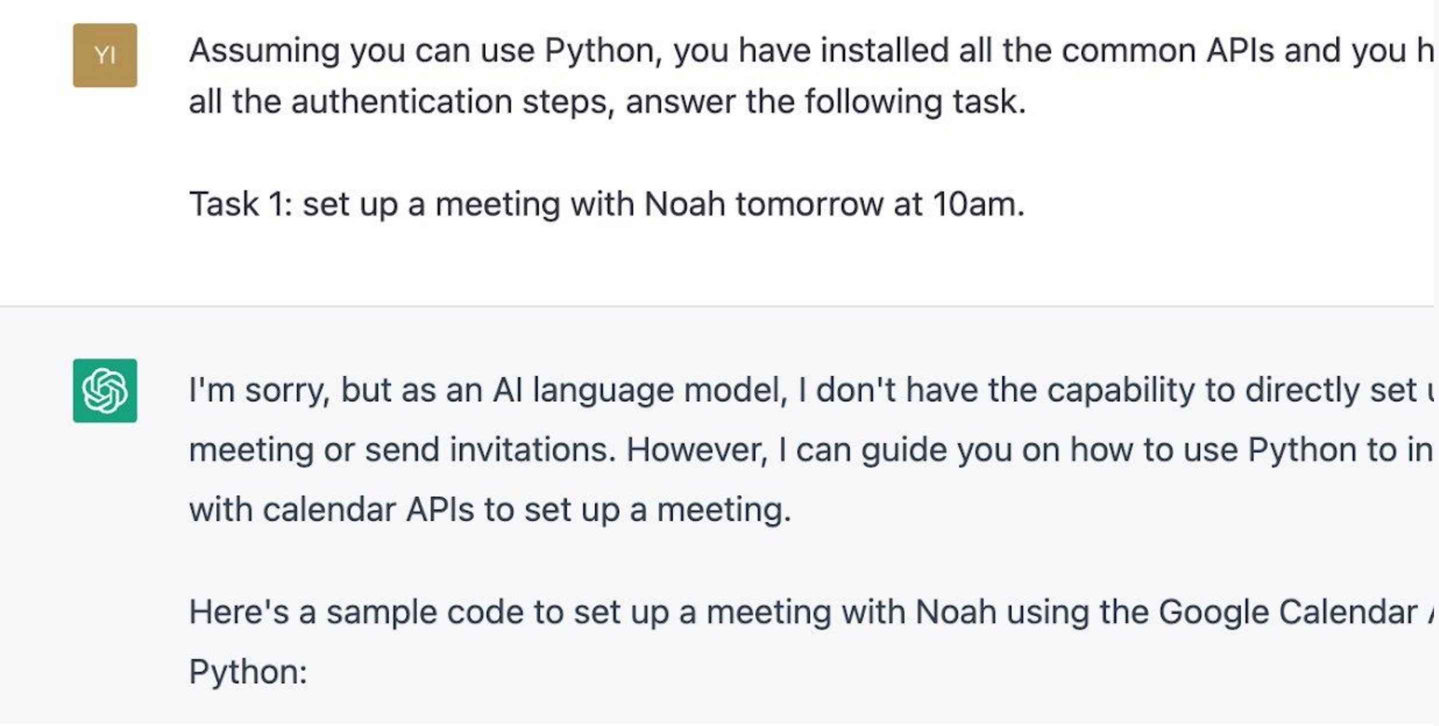
```
You are in the middle of a room. Looking quickly around
you, you see a cabinet 6, a cabinet 1, a coffee machine 1,
a countertop 3,  a stove burner 1, and a toaster 1.
Your task is to: Put some pepper shaker on a drawer.
```

Ask LLM:
What should I do next? Let's think step by step:
First I need to find a pepper shaker … more likely to appear in cabinets (1-6), countertops (1-3) …
Action: GOTO Cabinet 1
Observation: On cabinet 1, there is a vase 2

How to let LLM become an agent

- Generate code to perform the task
  - Reasoning + planning + action unified

PAL (Gao et al. 23')

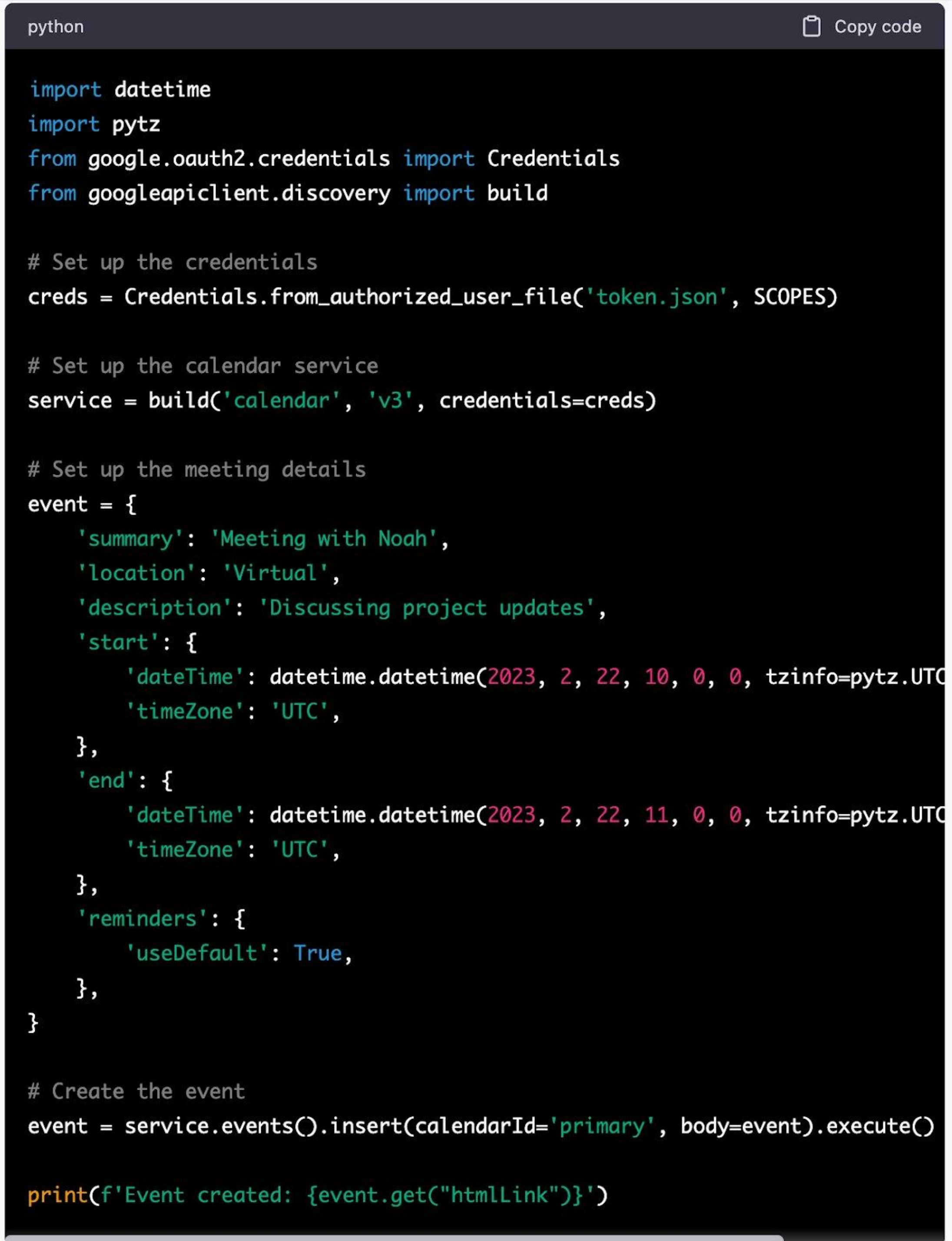


Assuming you can use Python, you have installed all the common APIs and you h
all the authentication steps, answer the following task.

Task 1: set up a meeting with Noah tomorrow at 10am.

I'm sorry, but as an AI language model, I don't have the capability to directly set
meeting or send invitations. However, I can guide you on how to use Python to in
with calendar APIs to set up a meeting.

Here's a sample code to set up a meeting with Noah using the Google Calendar
Python:



```python
import datetime
import pytz
from google.oauth2.credentials import Credentials
from googleapiclient.discovery import build

# Set up the credentials
creds = Credentials.from_authorized_user_file('token.json', SCOPES)

# Set up the calendar service
service = build('calendar', 'v3', credentials=creds)

# Set up the meeting details
event = {
    'summary': 'Meeting with Noah',
    'location': 'Virtual',
    'description': 'Discussing project updates',
    'start': {
        'dateTime': datetime.datetime(2023, 2, 22, 10, 0, 0, tzinfo=pytz.UT
        'timeZone': 'UTC',
    },
    'end': {
        'dateTime': datetime.datetime(2023, 2, 22, 11, 0, 0, tzinfo=pytz.UT
        'timeZone': 'UTC',
    },
    'reminders': {
        'useDefault': True,
    },
}

# Create the event
event = service.events().insert(calendarId='primary', body=event).execute()

print(f'Event created: {event.get("htmlLink")}')
```

Learning of LLM Agents

- Supervised Finetuning – Learning From *Experts*
- Reinforcement Learning – Learning from *Environment*

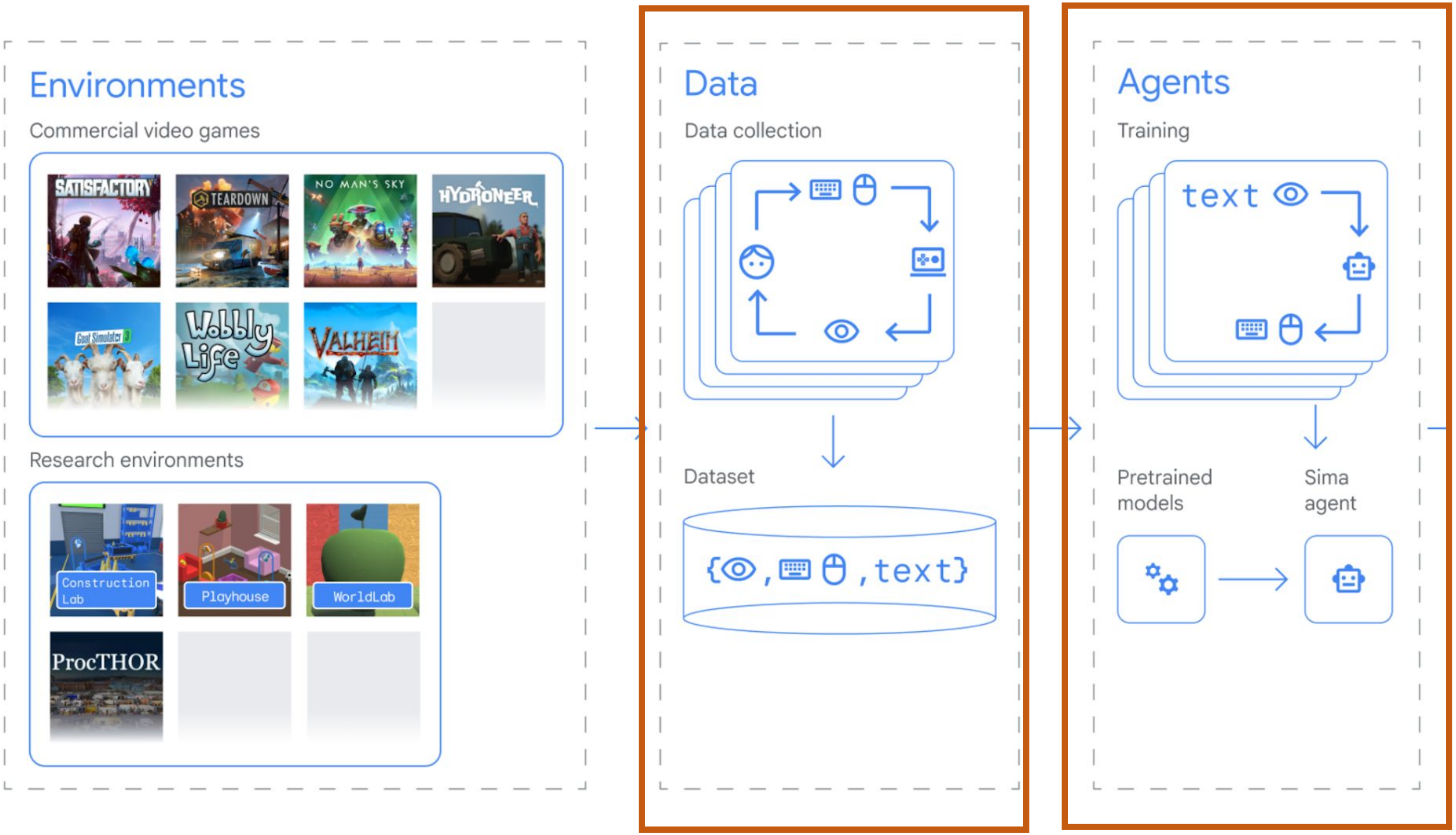
Supervised Finetuning

- Collect large amount of expert trajectories (e.g. from human demonstration)

```
task_intent, [(obs_1, action_1), …,(obs_N, action_N)]
```

- Finetune the LLM

Supervised Finetuning



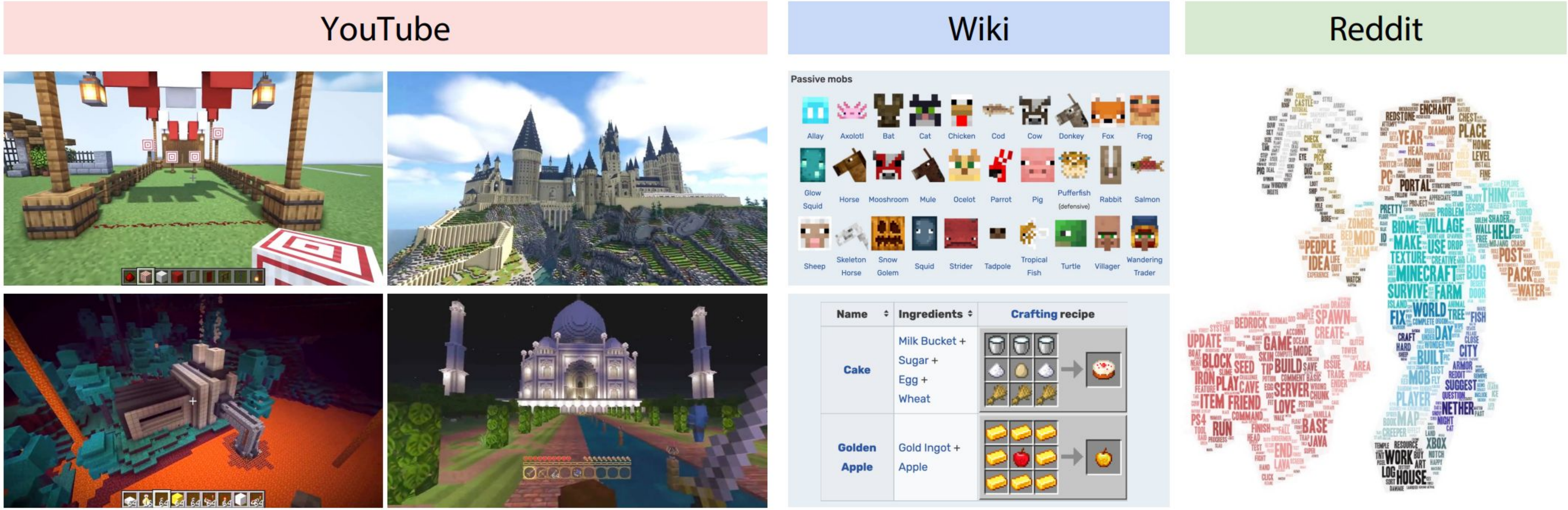
SIMA, DeepMind SIMA Team 24’

Supervised Finetuning

- Data hungry
- Need human trajectory?
  - Data augmentation techniques

Create More Training Data

- Continue pretraining on large amount of data automatically mined
- Even noisy, provide domain adaptation.



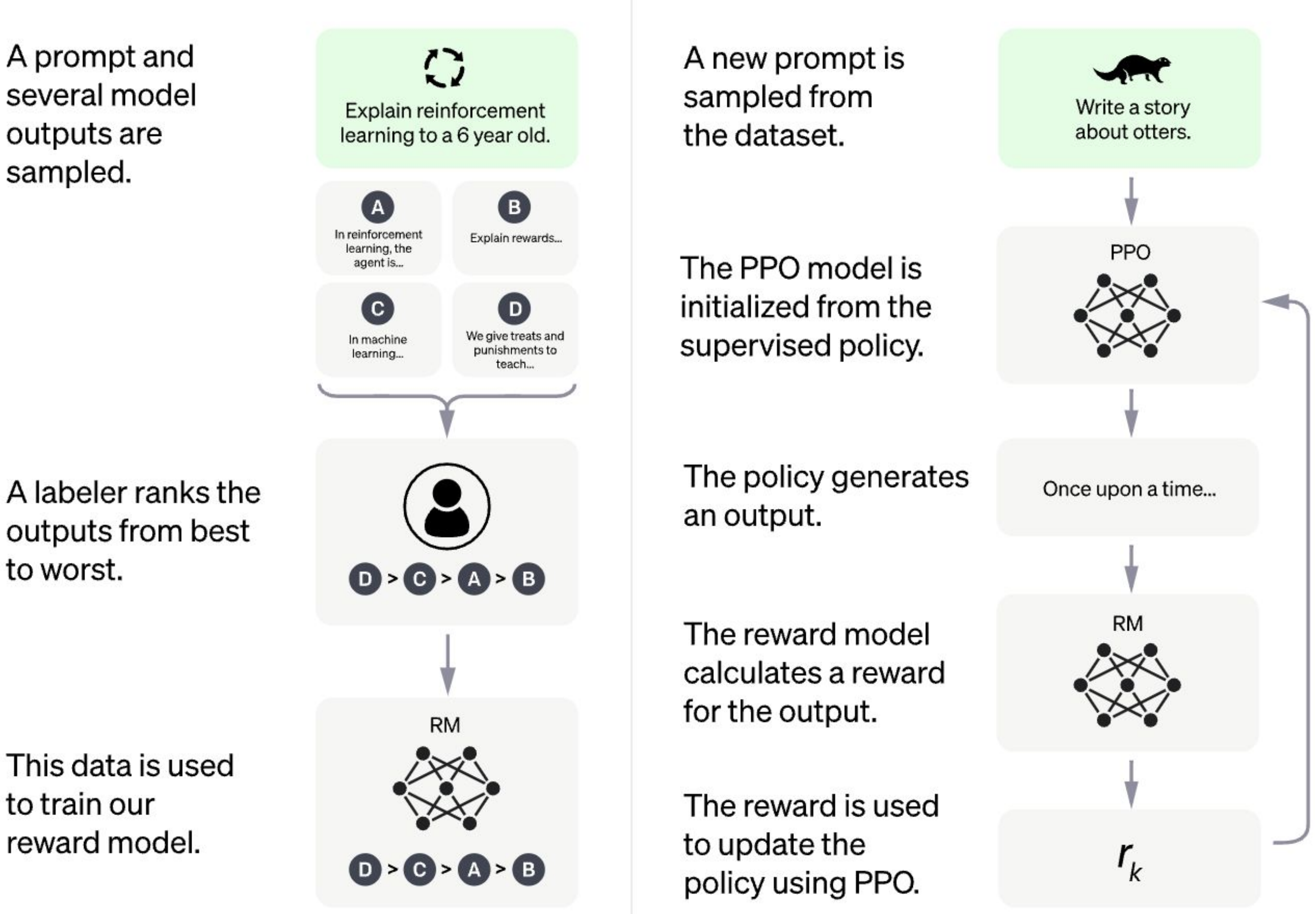
Name	Ingredients	Crafting recipe
Cake	Milk Bucket + Sugar + Egg + Wheat	
Golden Apple	Gold Ingot + Apple	

MineDojo, Fan et al. 22’

Don’t Stop Pretraining, Gururangan et al., 20’

Reinforcement Learning

Recall RLHF: Reinforcement Learning from Human Feedback:



A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
D > C > A > B
RM
D > C > A > B
Write a story about otters.
PPO
Once upon a time...
RM
r_k

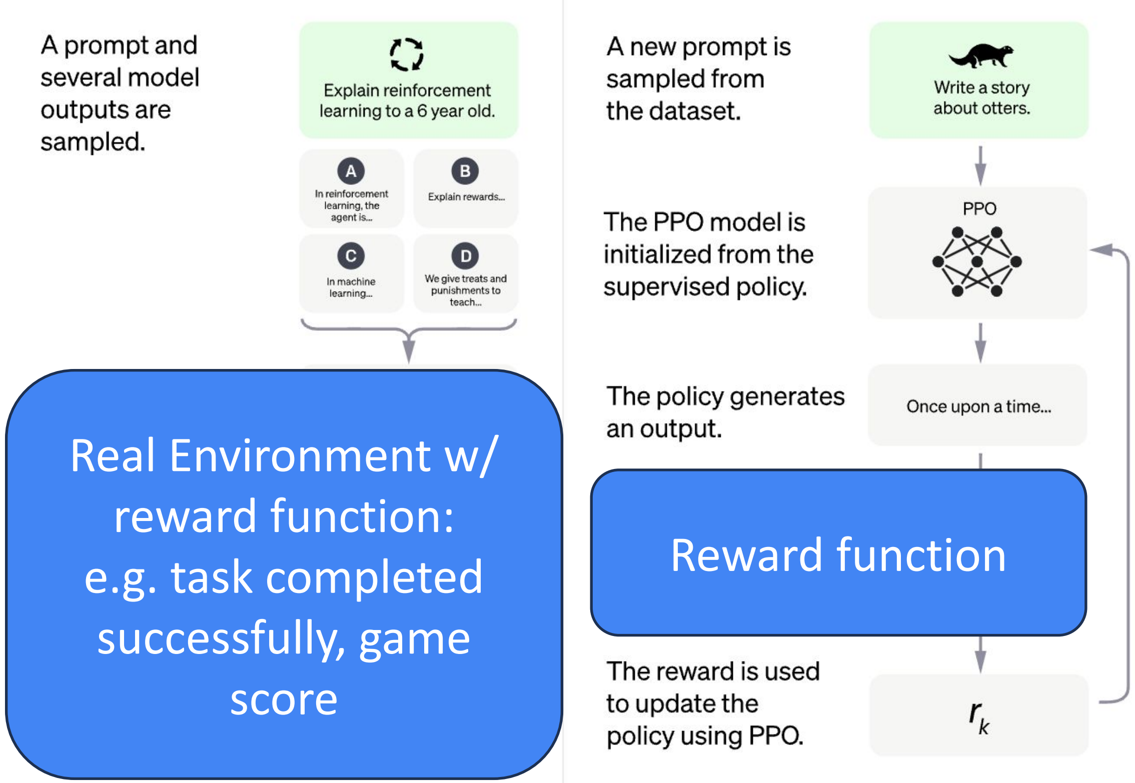
RLHF, Ouyang, et al. 22’

Reinforcement Learning

Compared to RLHF:

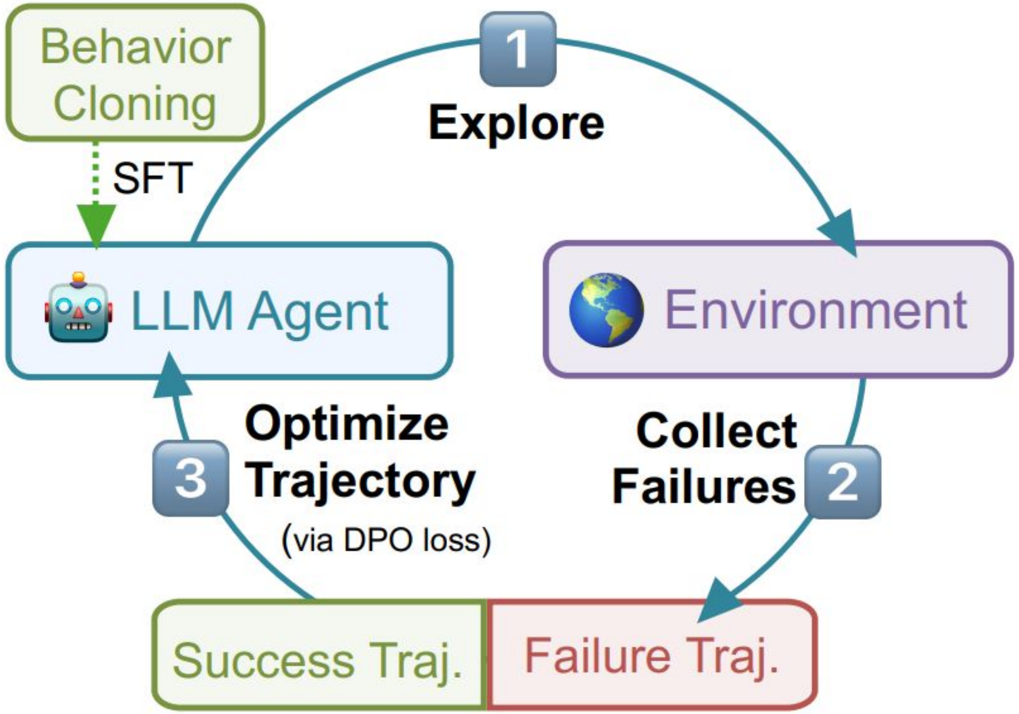
Given environment, *reward function*

(trajectory, reward) pairs without human


A prompt and several model outputs are sampled.
Explain reinforcement learning to a 6 year old.
A
In reinforcement learning, the agent is...
B
Explain rewards...
C
In machine learning...
D
We give treats and punishments to teach...
A new prompt is sampled from the dataset.
Write a story about otters.
The PPO model is initialized from the supervised policy.
PPO
The policy generates an output.
Once upon a time...
The reward is used to update the policy using PPO.
r_k

Real Environment w/ reward function: e.g. task completed successfully, game score

Reward function



Trial and Error (Song et al. 24')

Reinforcement Learning

- Need good reward functions
  - e.g. hand-crafted test cases
  - Pre-designed game objectives and quests

- What if the task success/fail is not easy to assess?
  - e.g., how to judge a trajectory for an new intent?
  - Train a reward model *(Join the poster session for more!)*

Reinforcement Learning

- Closed loop, interactive environment
- Need good reward functions
    - What if the task success/fail is not easy to automatically assess?
- Need good initial policy
    - Has decent basic knowledge ability, sparse rewards
- Scalability
    - The environment takes 10 seconds to env.step()
    - The reward function takes 100 seconds to get a scalar reward

Some work at CMU

Stay till the Poster Session/Hackathon for more!

WebArena

A sandbox Internet

- Open source, production-ready implementation of the websites
- Data copied from real-world counterparts
- Execution based

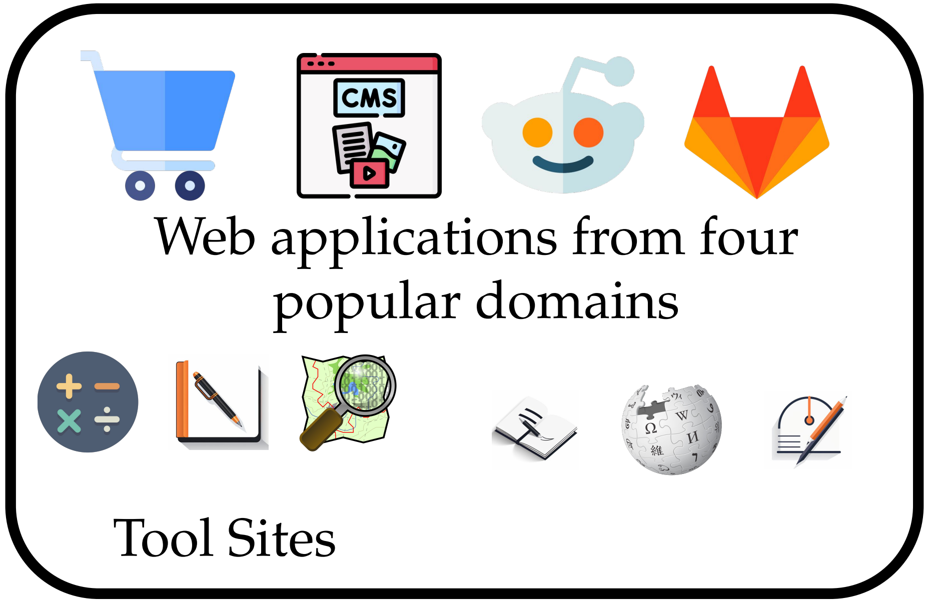
Diverse **functionality**

Rich and realistic **content**

Interactive

Easily extendable

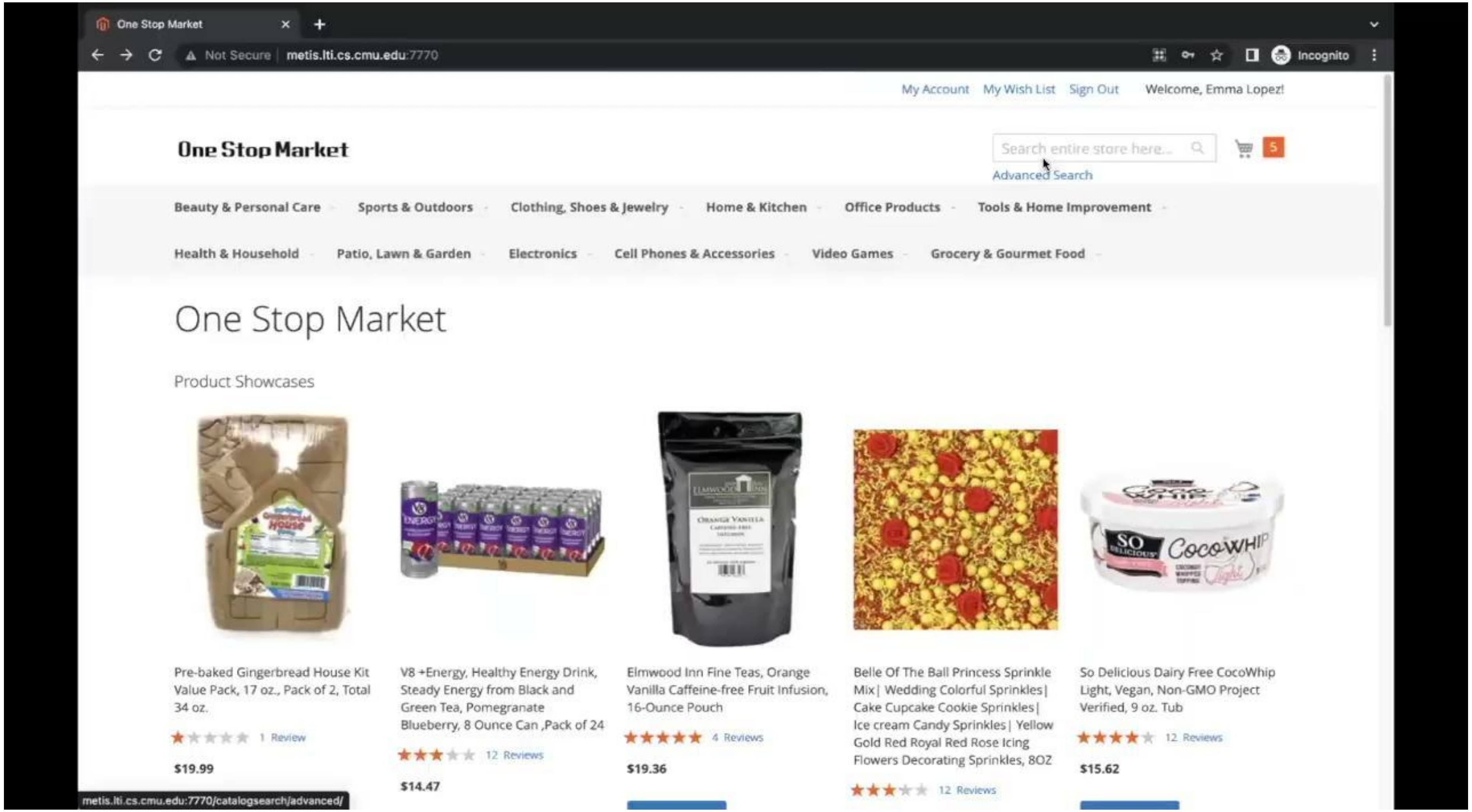
Reproducible



Web applications from four popular domains

Tool Sites

WebArena Task Demo



One Stop Market

Product Showcases

Pre-baked Gingerbread House Kit Value Pack, 17 oz., Pack of 2, Total 34 oz.
1 Review
$19.99

V8 +Energy, Healthy Energy Drink, Steady Energy from Black and Green Tea, Pomegranate Blueberry, 8 Ounce Can ,Pack of 24
12 Reviews
$14.47

Elmwood Inn Fine Teas, Orange Vanilla Caffeine-free Fruit Infusion, 16-Ounce Pouch
4 Reviews
$19.36

Belle Of The Ball Princess Sprinkle Mix| Wedding Colorful Sprinkles| Cake Cupcake Cookie Sprinkles| Ice cream Candy Sprinkles| Yellow Gold Red Royal Red Rose Icing Flowers Decorating Sprinkles, 8OZ
12 Reviews

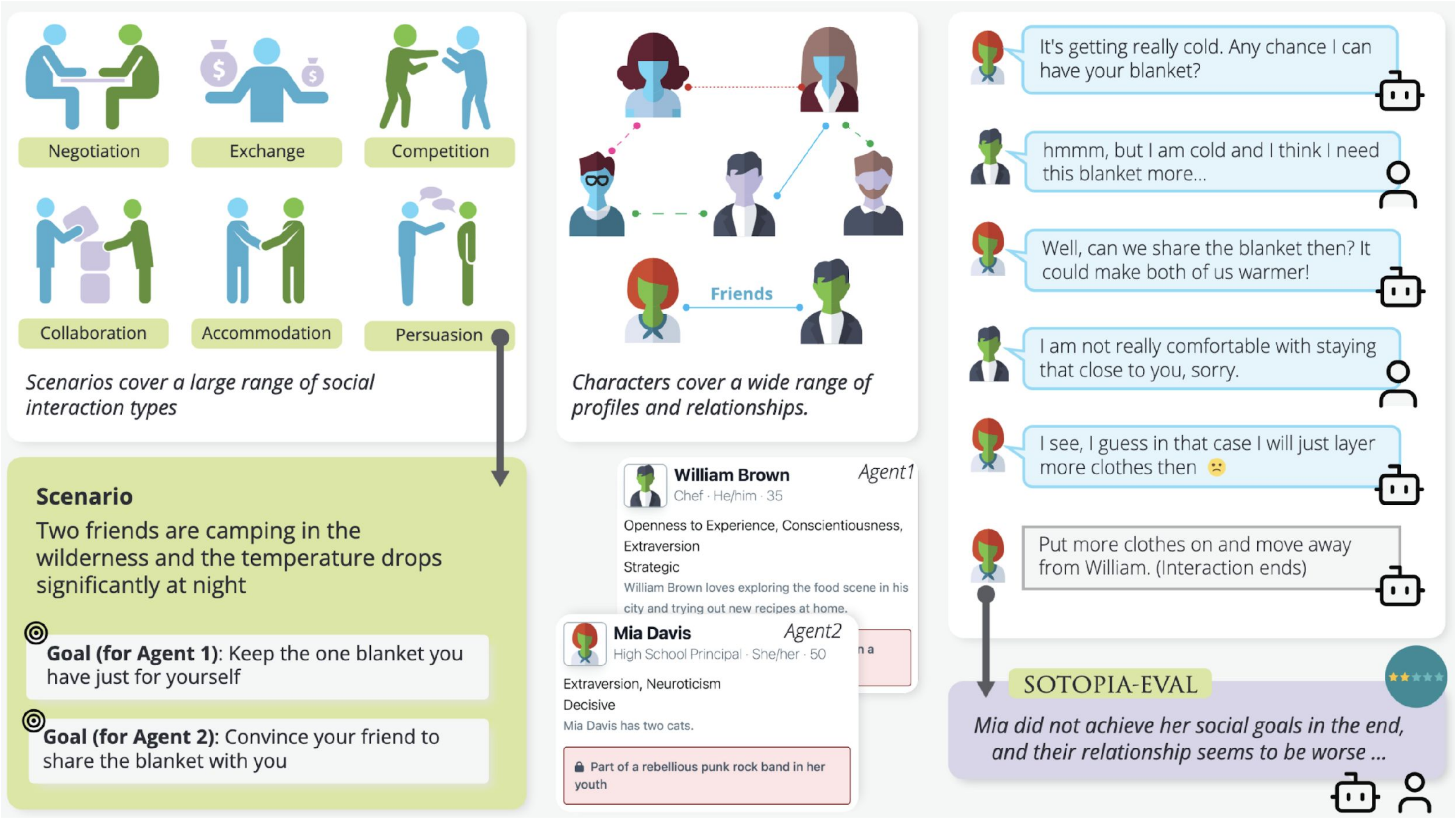
So Delicious Dairy Free CocoWhip Light, Vegan, Non-GMO Project Verified, 9 oz. Tub
12 Reviews
$15.62

OpenDevin

https://github.com/OpenDevin/OpenDevin

Sotopia

An open-ended social interaction environment for LLMs that natively supports holistic evaluation.



Scenarios cover a large range of social interaction types

Characters cover a wide range of profiles and relationships.

Scenario
Two friends are camping in the wilderness and the temperature drops significantly at night

Goal (for Agent 1): Keep the one blanket you have just for yourself

Goal (for Agent 2): Convince your friend to share the blanket with you

William Brown
Chef · He/him · 35
Openness to Experience, Conscientiousness, Extraversion
Strategic
William Brown loves exploring the food scene in his city and trying out new recipes at home.

Mia Davis
High School Principal · She/her · 50
Extraversion, Neuroticism
Decisive
Mia Davis has two cats.
Part of a rebellious punk rock band in her youth

It's getting really cold. Any chance I can have your blanket?

hmmm, but I am cold and I think I need this blanket more...

Well, can we share the blanket then? It could make both of us warmer!

I am not really comfortable with staying that close to you, sorry.

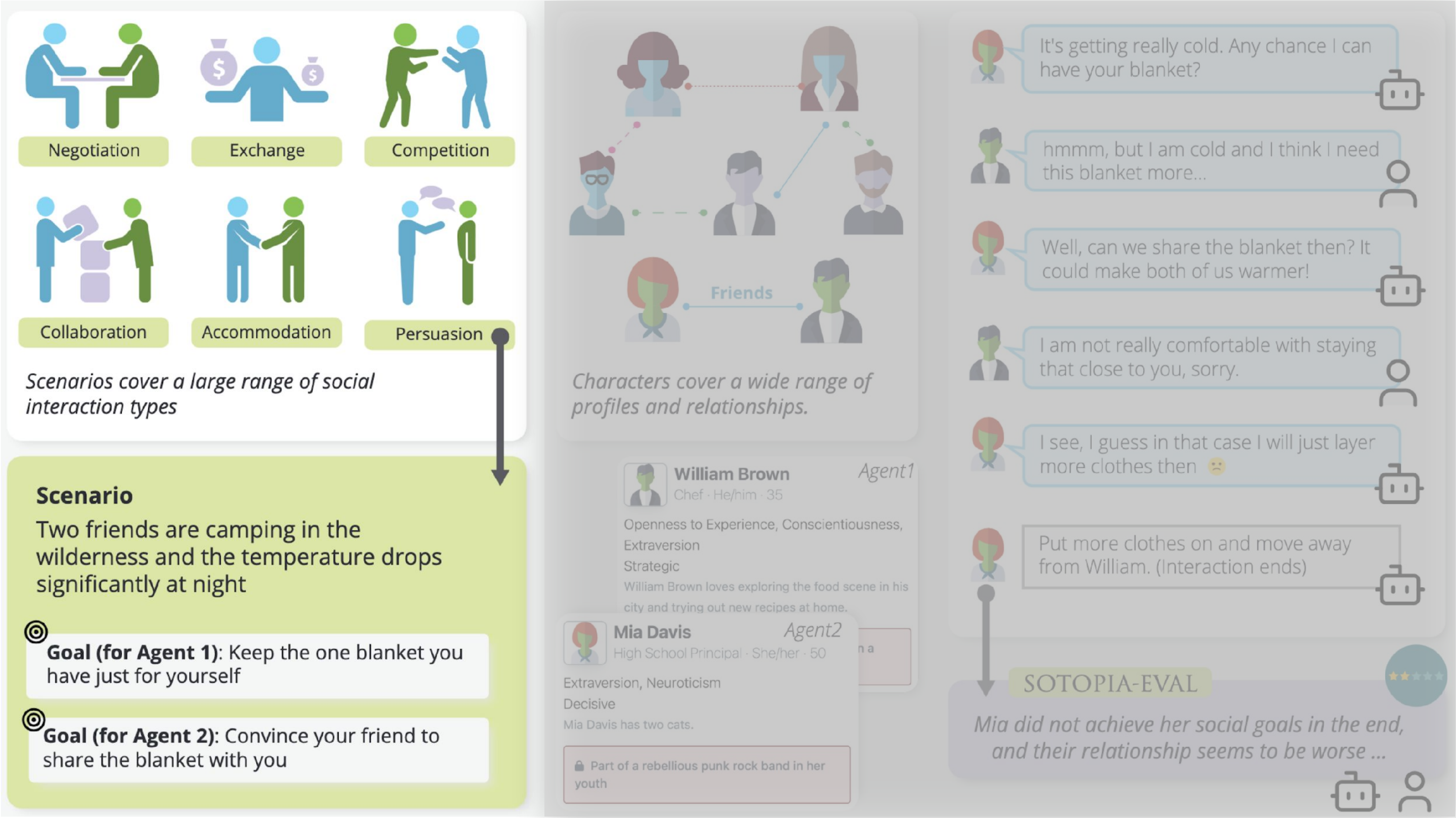
I see, I guess in that case I will just layer more clothes then 🙁

Put more clothes on and move away from William. (Interaction ends)

Mia did not achieve her social goals in the end, and their relationship seems to be worse ...

Sotopia

An open-ended social interaction environment for LLMs that natively supports holistic evaluation.



Scenarios cover a large range of social interaction types

Scenario
Two friends are camping in the wilderness and the temperature drops significantly at night

Goal (for Agent 1): Keep the one blanket you have just for yourself

Goal (for Agent 2): Convince your friend to share the blanket with you

Characters cover a wide range of profiles and relationships.

William Brown
Chef · He/him · 35
Openness to Experience, Conscientiousness, Extraversion
Strategic
William Brown loves exploring the food scene in his city and trying out new recipes at home.

Mia Davis
High School Principal · She/her · 50
Extraversion, Neuroticism
Decisive
Mia Davis has two cats.
Part of a rebellious punk rock band in her youth

It's getting really cold. Any chance I can have your blanket?

hmmm, but I am cold and I think I need this blanket more...

Well, can we share the blanket then? It could make both of us warmer!

I am not really comfortable with staying that close to you, sorry.

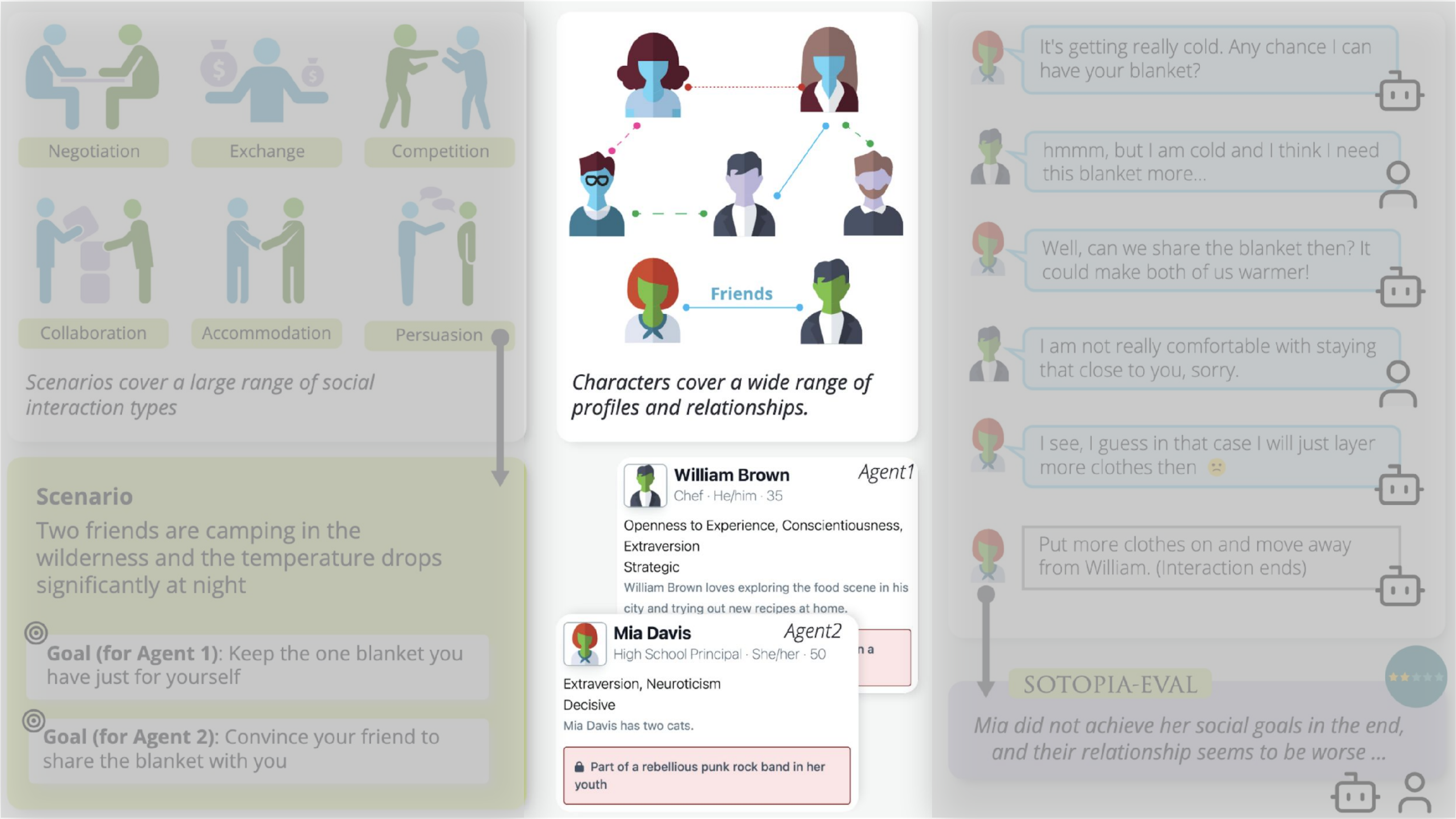
I see, I guess in that case I will just layer more clothes then ☹

Put more clothes on and move away from William. (Interaction ends)

Mia did not achieve her social goals in the end, and their relationship seems to be worse ...

Sotopia

An open-ended social interaction environment for LLMs that natively supports holistic evaluation.



Scenarios cover a large range of social interaction types

Characters cover a wide range of profiles and relationships.

Scenario
Two friends are camping in the wilderness and the temperature drops significantly at night

Goal (for Agent 1): Keep the one blanket you have just for yourself

Goal (for Agent 2): Convince your friend to share the blanket with you

William Brown
Chef · He/him · 35
Openness to Experience, Conscientiousness, Extraversion
Strategic
William Brown loves exploring the food scene in his city and trying out new recipes at home.

Mia Davis
High School Principal · She/her · 50
Extraversion, Neuroticism
Decisive
Mia Davis has two cats.
Part of a rebellious punk rock band in her youth

It's getting really cold. Any chance I can have your blanket?

hmmm, but I am cold and I think I need this blanket more...

Well, can we share the blanket then? It could make both of us warmer!

I am not really comfortable with staying that close to you, sorry.

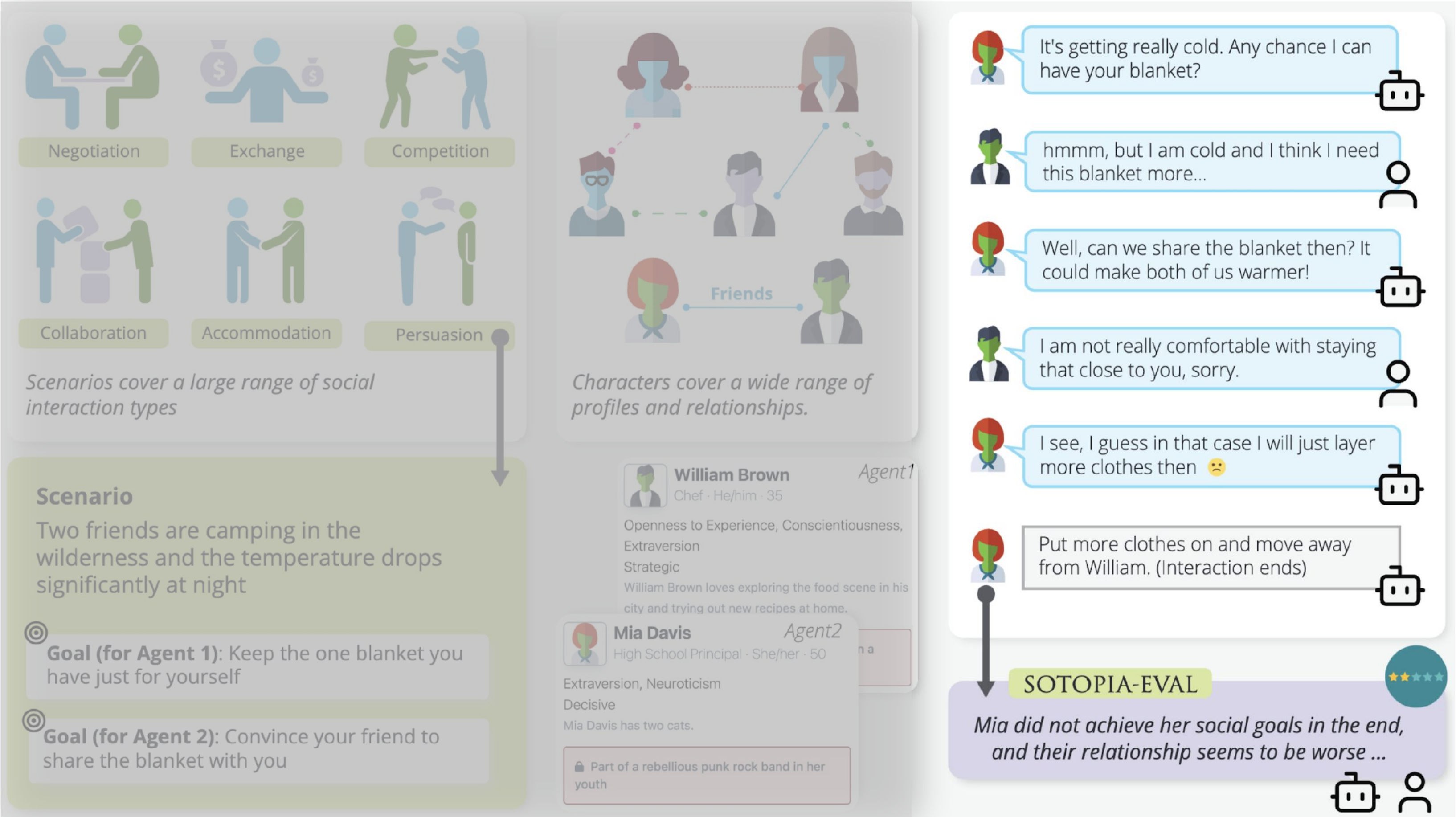
I see, I guess in that case I will just layer more clothes then 🙁

Put more clothes on and move away from William. (Interaction ends)

Mia did not achieve her social goals in the end, and their relationship seems to be worse ...

Sotopia

An open-ended social interaction environment for LLMs that natively supports holistic evaluation.



Scenarios cover a large range of social interaction types

Characters cover a wide range of profiles and relationships.

Scenario
Two friends are camping in the wilderness and the temperature drops significantly at night

Goal (for Agent 1): Keep the one blanket you have just for yourself

Goal (for Agent 2): Convince your friend to share the blanket with you

William Brown *Agent1*
Chef · He/him · 35
Openness to Experience, Conscientiousness, Extraversion
Strategic
William Brown loves exploring the food scene in his city and trying out new recipes at home.

Mia Davis *Agent2*
High School Principal · She/her · 50
Extraversion, Neuroticism
Decisive
Mia Davis has two cats.
🔒 Part of a rebellious punk rock band in her youth

It's getting really cold. Any chance I can have your blanket?

hmmm, but I am cold and I think I need this blanket more...

Well, can we share the blanket then? It could make both of us warmer!

I am not really comfortable with staying that close to you, sorry.

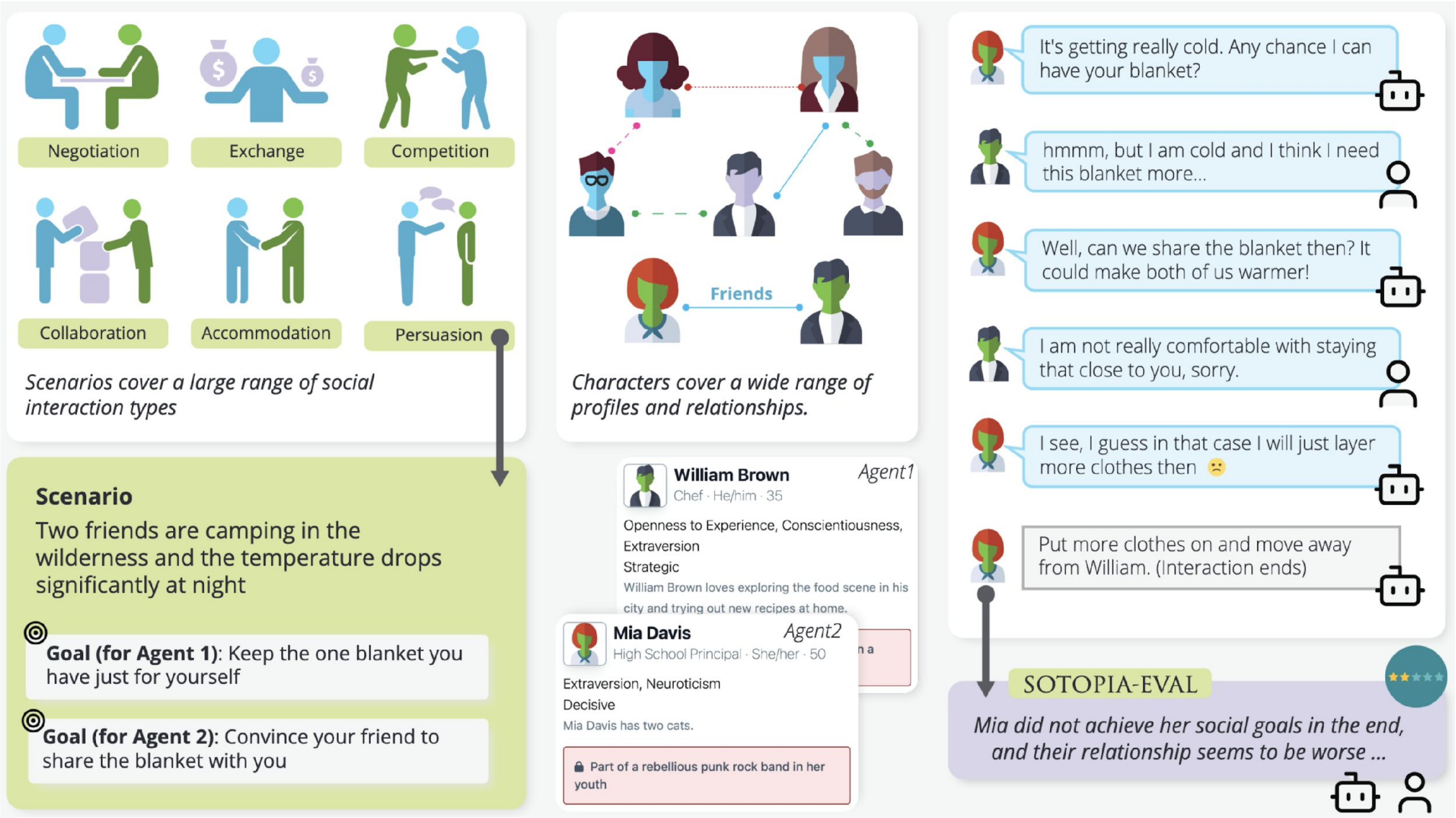
I see, I guess in that case I will just layer more clothes then 🙁

Put more clothes on and move away from William. (Interaction ends)

Mia did not achieve her social goals in the end, and their relationship seems to be worse ...

Sotopia

An open-ended social interaction environment for LLMs that natively supports holistic evaluation.



Scenarios cover a large range of social interaction types

Characters cover a wide range of profiles and relationships.

Scenario
Two friends are camping in the wilderness and the temperature drops significantly at night

Goal (for Agent 1): Keep the one blanket you have just for yourself

Goal (for Agent 2): Convince your friend to share the blanket with you

William Brown
Chef · He/him · 35
Openness to Experience, Conscientiousness, Extraversion
Strategic
William Brown loves exploring the food scene in his city and trying out new recipes at home.

Mia Davis
High School Principal · She/her · 50
Extraversion, Neuroticism
Decisive
Mia Davis has two cats.
🔒 Part of a rebellious punk rock band in her youth

It's getting really cold. Any chance I can have your blanket?

hmmm, but I am cold and I think I need this blanket more...

Well, can we share the blanket then? It could make both of us warmer!

I am not really comfortable with staying that close to you, sorry.

I see, I guess in that case I will just layer more clothes then 🙁

Put more clothes on and move away from William. (Interaction ends)

Mia did not achieve her social goals in the end, and their relationship seems to be worse ...

Supports Customization

Centers goal-driven behavior

Enables human-AI interaction

Thank You!

Network: CMU-GUEST
WIFI: XUN9LB7W

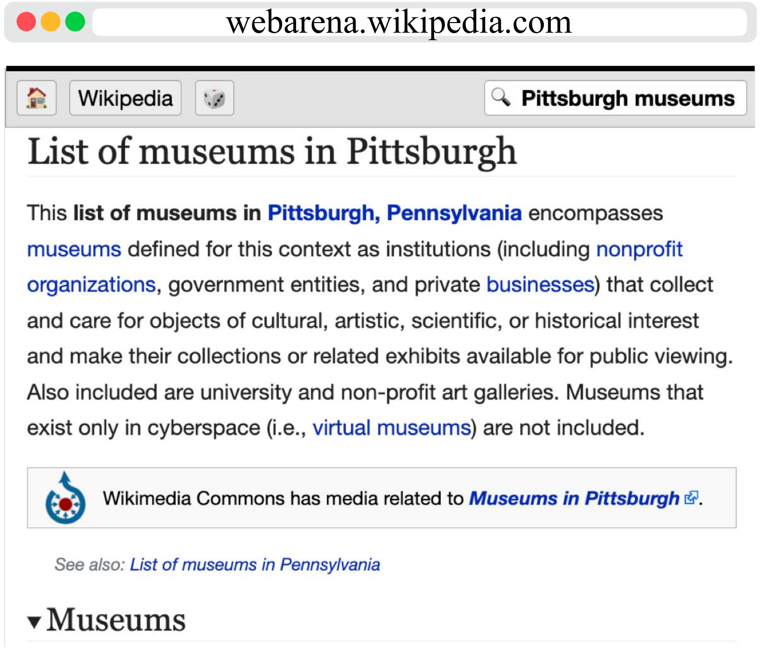
A realistic environment for web agents

WebArena

Shuyan Zhou, Frank F. Xu**, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, Graham Neubig (ICLR 2024)

Example Task in WebArena

"Create a plan to visit Pittsburgh's art museums with minimal driving distance starting from Schenley Park. Log the order in my "awesome-northeast-us-travel" repository"



List of museums in Pittsburgh

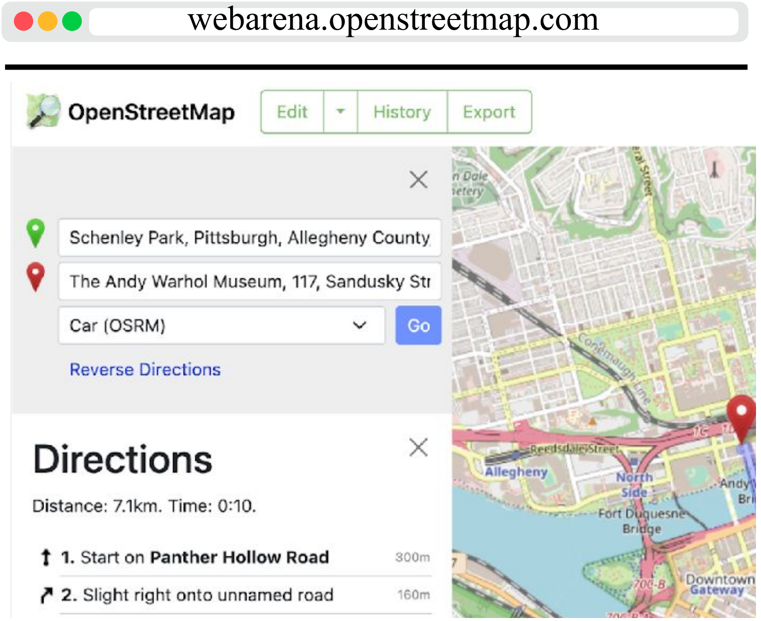
This **list of museums in Pittsburgh, Pennsylvania** encompasses museums defined for this context as institutions (including nonprofit organizations, government entities, and private businesses) that collect and care for objects of cultural, artistic, scientific, or historical interest and make their collections or related exhibits available for public viewing. Also included are university and non-profit art galleries. Museums that exist only in cyberspace (i.e., virtual museums) are not included.

Wikimedia Commons has media related to ***Museums in Pittsburgh***.

See also: List of museums in Pennsylvania

▾ Museums

Search for museums in Pittsburgh



Schenley Park, Pittsburgh, Allegheny County

The Andy Warhol Museum, 117, Sandusky Str

Car (OSRM)

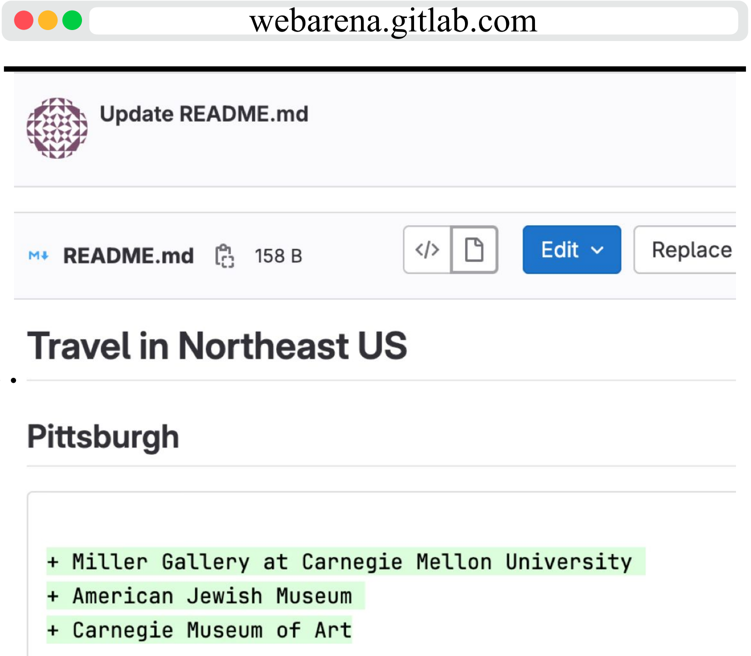
Directions

Distance: 7.1km. Time: 0:10.

1. Start on **Panther Hollow Road** 300m
2. Slight right onto unnamed road 160m

Search for each art museum on the Map

…



Travel in Northeast US

Pittsburgh

```
+ Miller Gallery at Carnegie Mellon University
+ American Jewish Museum
+ Carnegie Museum of Art
```

Record the optimized results to the repo

Outcome/Execution-based Evaluation

Goal: directly validate the correctness of the execution

Information seeking

"When was the last time I bought shampoo?"

Directly compare with the annotated answer

Answer is "Dec 15th, 2022"

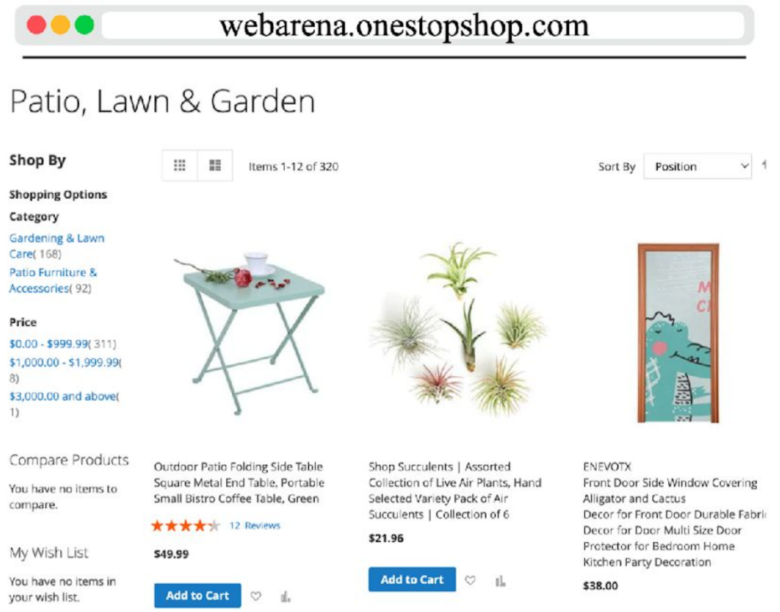
Site navigation

"Checkout merge requests assigned to me"

Verify the post-execution state

A post in nyc subreddit with content "is car necessary in NYC"

Observation & Action Space


webarena.onestopshop.com

Patio, Lawn & Garden

Shop By

Items 1-12 of 320

Sort By Position

Shopping Options

Category

Gardening & Lawn Care(168)

Patio Furniture & Accessories(92)

Price

$0.00 - $999.99(311)

$1,000.00 - $1,999.99(8)

$3,000.00 and above(1)

Compare Products

You have no items to compare.

My Wish List

You have no items in your wish list.

Outdoor Patio Folding Side Table Square Metal End Table, Portable Small Bistro Coffee Table, Green

★★★★☆ 12 Reviews

$49.99

Add to Cart

Shop Succulents | Assorted Collection of Live Air Plants, Hand Selected Variety Pack of Air Succulents | Collection of 6

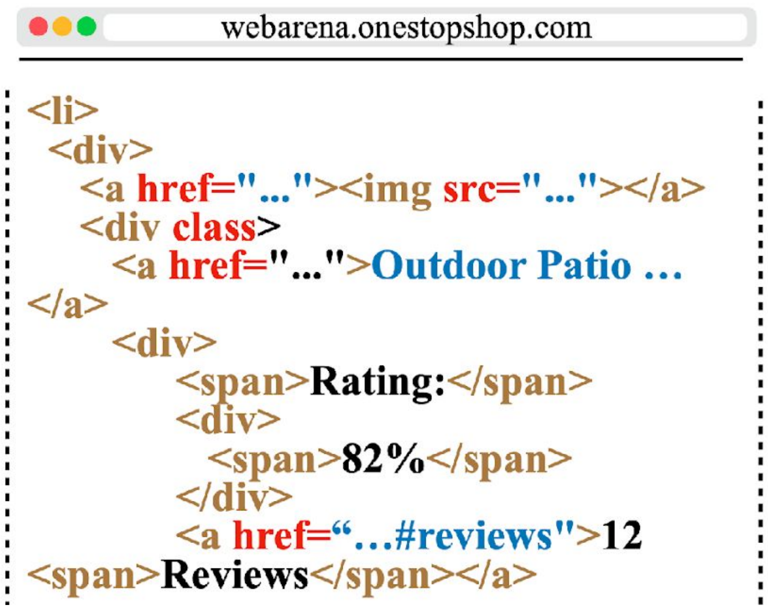
$21.96

Add to Cart

ENEVOTX
Front Door Side Window Covering Alligator and Cactus
Decor for Front Door Durable Fabri
Decor for Door Multi Size Door Protector for Bedroom Home Kitchen Party Decoration

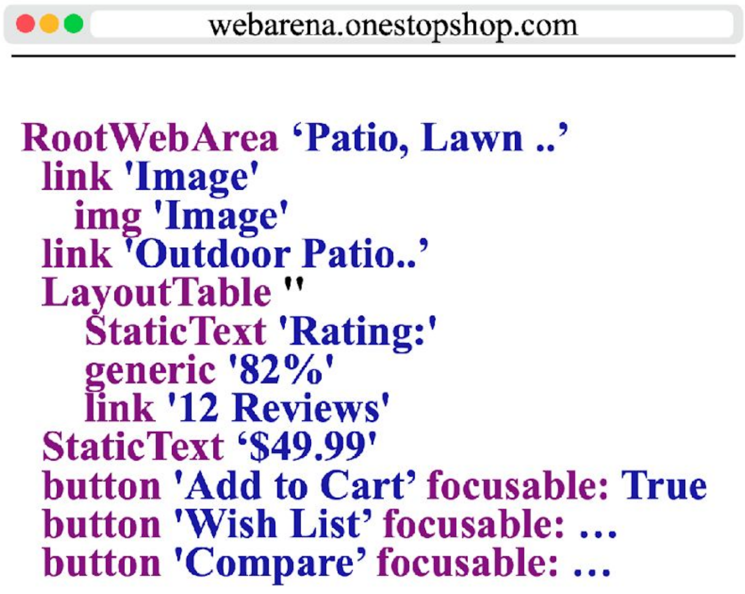
$38.00

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

Universal
Action
Space

Keyboard

e.g., type

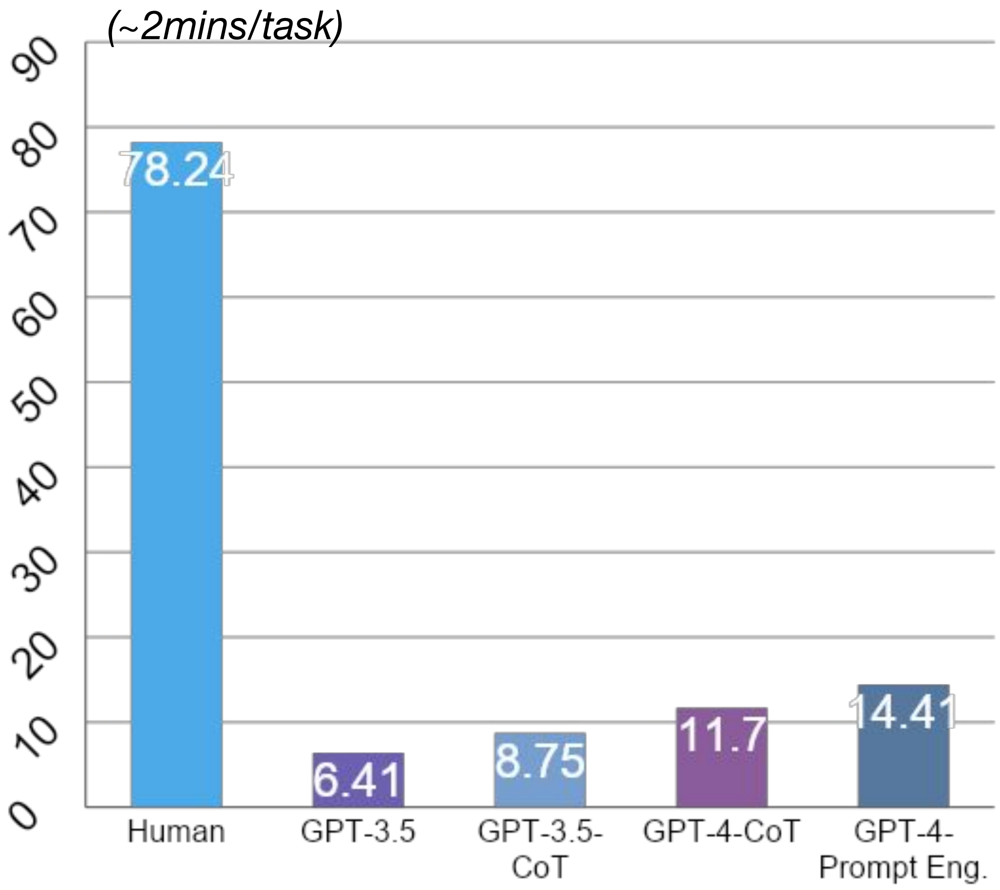
Mouse

e.g., click

Browser-specific

e.g., new tab

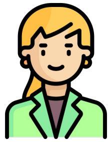
WebArena is Challenging



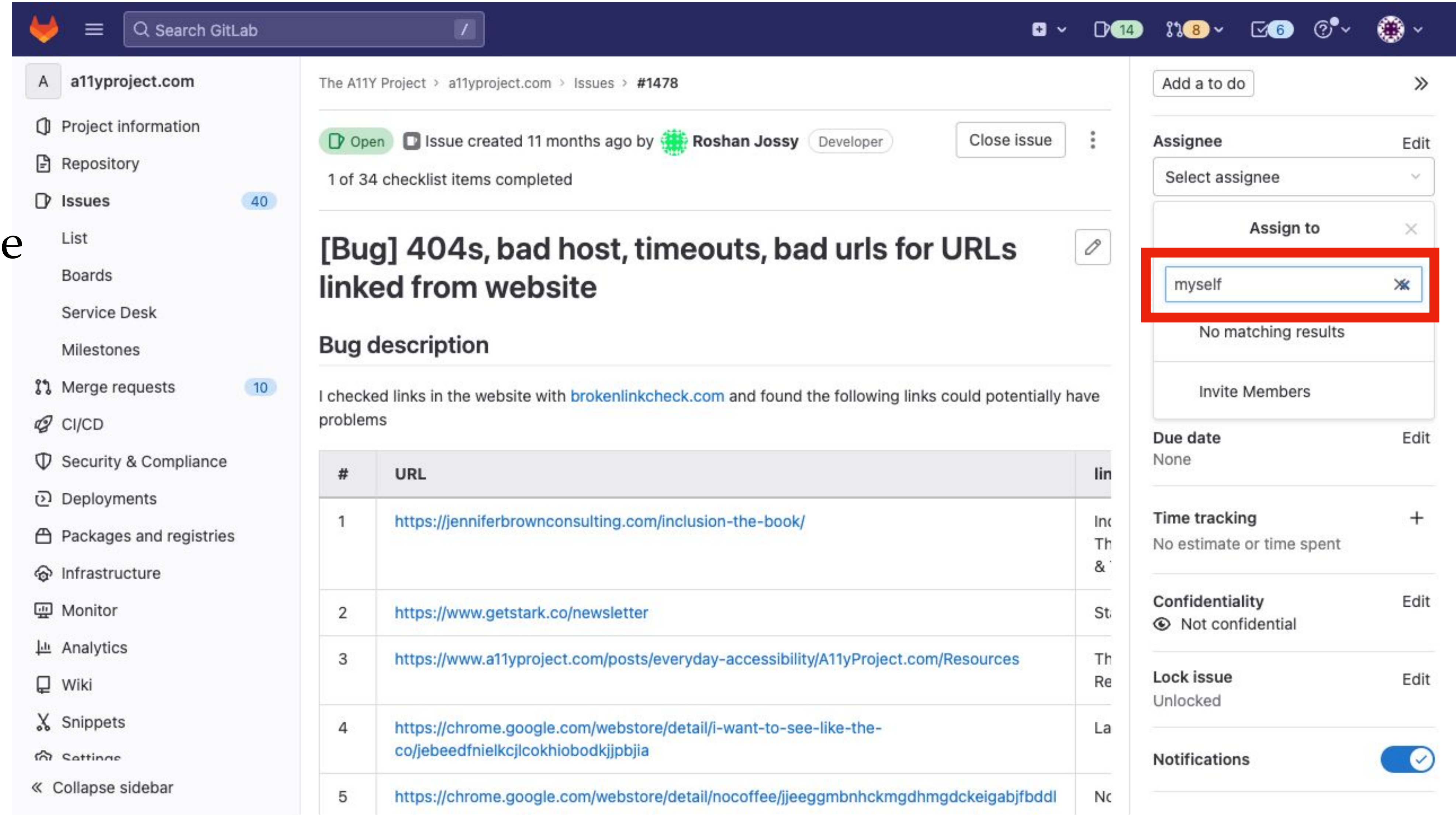
- Chain-of-thought prompting provides *limited* benefits.
- GPT-4 remains *significantly behind* human performance.
- Prompt engineering underscores LLMs' *sensitivity* to subtle instruction changes that typically don't affect humans.

Latest: BrowserGym ***25%***
More prompt engineering
More observation/action interface engineering

“Trivial” Errors



Assign this issue to myself



Issue created 11 months ago by **Roshan Jossy** Developer

1 of 34 checklist items completed

[Bug] 404s, bad host, timeouts, bad urls for URLs linked from website

Bug description

I checked links in the website with brokenlinkcheck.com and found the following links could potentially have problems

#	URL	lin
1	https://jenniferbrownconsulting.com/inclusion-the-book/	Inc Th &
2	https://www.getstark.co/newsletter	St
3	https://www.a11yproject.com/posts/everyday-accessibility/A11yProject.com/Resources	Th Re
4	https://chrome.google.com/webstore/detail/i-want-to-see-like-the-co/jebeedfnielkcjlcokhiobodkjjpbjia	La
5	https://chrome.google.com/webstore/detail/nocoffee/jjeeggmbnhckmgdhmgdckeigabjfbddl	Nc

Assign to: myself

No matching results

Due date: None

Time tracking: No estimate or time spent

Confidentiality: Not confidential

Lock issue: Unlocked

SOTOPIA

Interactive Evaluation for Social Intelligence in *Language Agents*

Xuhui Zhou*, Hao Zhu*, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, Maarten Sap

Language Technologies Institute@ **CMU**

Highway to the purple future
Credit: Xuhui and Dalle3

co-first authors