

# Language Models as Agents

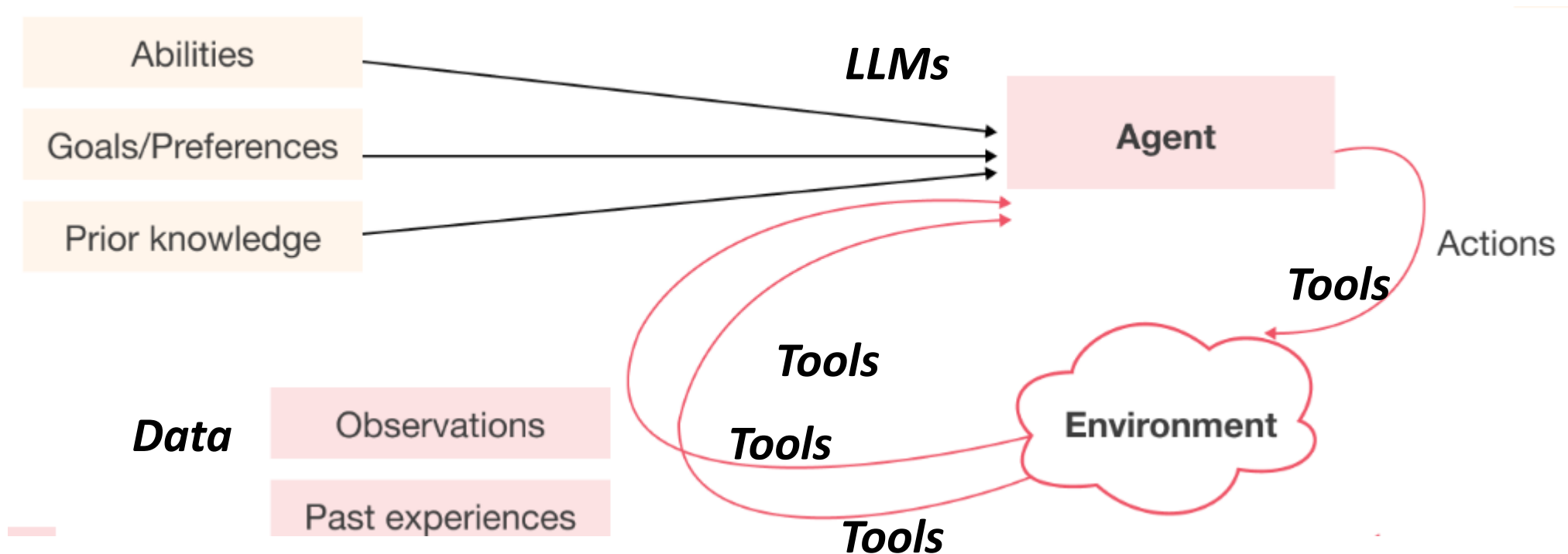
*Frank Xu*



**Carnegie Mellon University**  
Language Technologies Institute

# 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
- Training-free Methods for Building Agents
- Evaluation Environment and Benchmark
- Training Methods for Improving Agents

# Tasks and Applications For LLM Agents

# Why do we want agents?

Imagine if things get done by just talking...

# How People Interact with Computers?

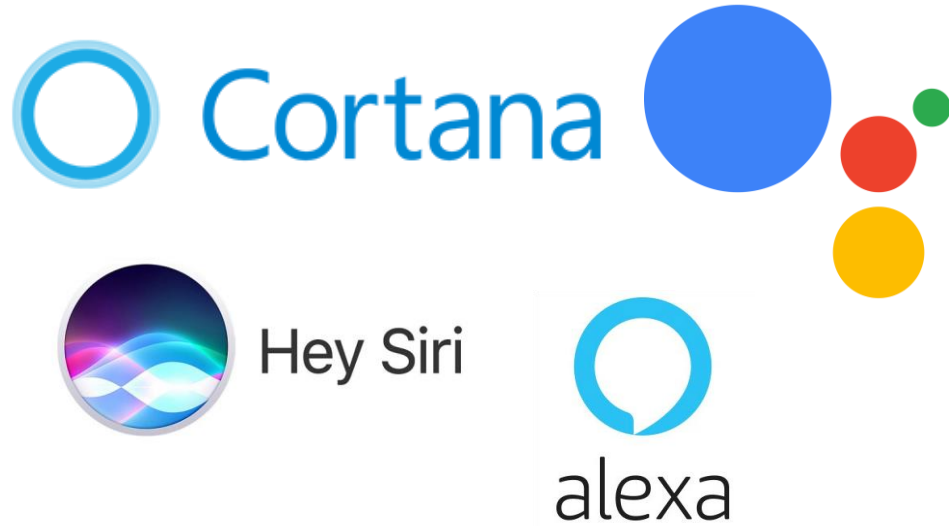
- Traditionally ...






- What about using natural language?
- Save time, natural, accessible, no need to browse, no programming learning curve, etc.






# Natural Language Interfaces to Computers

A screenshot of a Python IDE window titled "Untitled-1". The window contains a Python script with five lines of code. Line 1: `my_list = [3, 5, 1]`. Line 2: `sort in descending order →`. Line 3: `sorted(my_list, reverse=True)`, which is highlighted in green. Line 4: . Line 5: . The IDE interface includes a toolbar at the top with a Python logo, a file icon, and a window icon. The status bar at the bottom shows "master\*", a refresh icon, "Python 3.6.5 64-bit", and a close icon.

## Virtual Assistants

-  *Set an alarm at 7 AM*
-  *Remind me for the meeting at 5pm*
-  *Play Jay Chou's latest album*

## Natural Language Programming

-  *Sort my\_list in descending order*
-  *Copy my\_file to home folder*
-  *Dump my\_dict as a csv file output.csv*

# Tool Integrations into Chatbots

## ChatGPT plugins

We've implemented initial support for plugins in

ChatGPT. Plugins help ChatGPT access external data sources, perform computations, or

[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.



# Robots



⋮



*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.*

# Robots



AlfWorld, Shridhar et al. 21'

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.

# Games

## Open-ended Environments

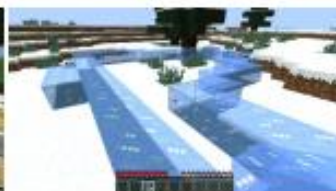
Craft Glass Bridge



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





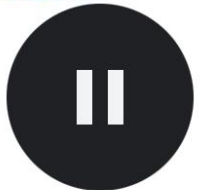
# Games



No Man's Sky: Go to spaceship



No Man's Sky: Shoot Asteroid



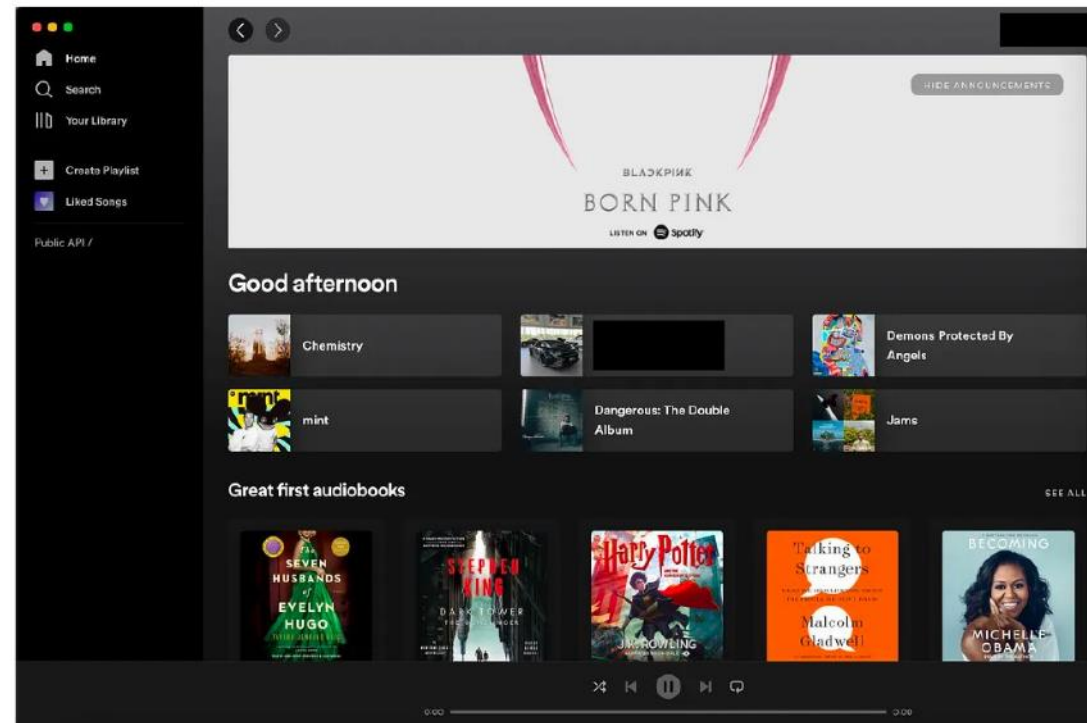
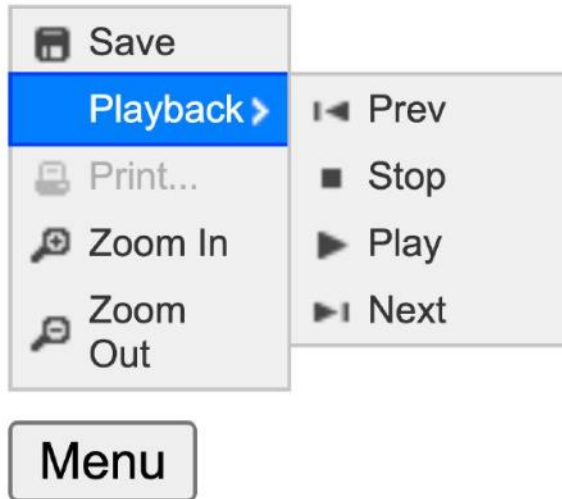
# Software Development

Devin (https://www.cognition-labs.com/introducing-devin)

The screenshot displays the 'Devin's Workspace' interface. On the left, a chat window shows a message from 'Devin' to 'Scott' regarding updates to a benchmarking script. Below the chat, a status indicator shows 'Devin is currently debugging the script to resolve API response errors and the KeyError 'text' issue.' The main workspace is divided into two panes. The top pane, titled 'default', contains a terminal window with HTML code for a script and its output, which shows 'None' for token counts. The bottom pane, titled 'benchmark.py', shows a code editor with Python code for a benchmark script, including API keys and endpoints. A browser window in the bottom-left pane displays the Perplexity AI 'Chat Completions' reference page. The interface includes a sidebar with navigation icons and a 'Following' toggle in the top right.

# UI Automation

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



“Play some synthwave songs”

# Training-free Methods for Building Agents



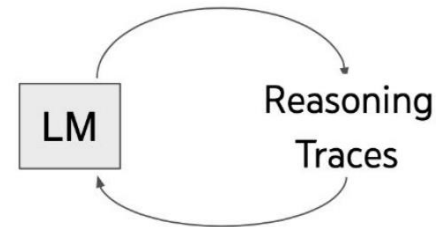


# How to let LLM become an agent

- 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:

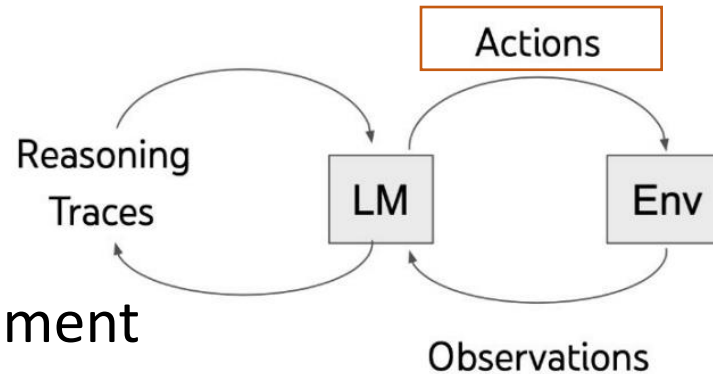
First I need to find a pepper shaker ... more likely to appear in cabinets (1-6), countertops (1-3) ...

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 in prompt



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

- Planning and reasoning ability

Chain-of-thoughts (CoT)

"Let's think step by step ...."

- Tool-use ability

Generate API calls with arguments

Execute the API calls

Get API call results as new observation

- What if lots of APIs?

Supposing you have the following APIs, answer the given question.

Available APIs:

1. weather(location): for getting the weather information given a location
2. location(): for getting the location of the user.
3. bus\_route(start, end): for finding the current bus route from the start position to the end position.
4. count\_characters(s): for counting the number of characters in a string s.

Question: Is it okay to go hiking today?

To know the answer, we first need to know the current location of the user.

API call: location() = Seattle.

Next, we need to know the current weather in Seattle.

API call: weather(Seattle) = Cloudy with a chance of rain.

Based on this information, it is not recommended to go hiking today in Seattle.

Available APIs:

1. weather(location): for getting the weather information given a location
2. location(): for getting the location of the user.
3. bus\_route(start, end): for finding the current bus route from the start position to the end position.
4. count\_characters(s): for counting the number of characters in a string s.

Question: how long does it take to take a bus to Vancouver?

To answer this question, we first need to know the current location of the user.

API call: location() = Seattle.

Next, we need to find the bus route from Seattle to Vancouver.

API call: bus\_route(Seattle, Vancouver) = 4 hours.

Therefore, it takes 4 hours to take a bus from Seattle to Vancouver.

# How to let LLM become an agent

- Generate code to perform the task  
Reasoning + planning + action unified  
PAL (Gao et al. 23')

YI

Assuming you can use Python, you have installed all the common APIs and you have completed 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 up a meeting or send invitations. However, I can guide you on how to use Python to interact with calendar APIs to set up a meeting.

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

```
python Copy code

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.UTC),
        'timeZone': 'UTC',
    },
    'end': {
        'dateTime': datetime.datetime(2023, 2, 22, 11, 0, 0, tzinfo=pytz.UTC),
        'timeZone': 'UTC',
    },
    'reminders': {
        'useDefault': True,
    },
}

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

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

```

# Evaluation Environment and Benchmark

# Evaluation of LLM Agents

- Simplified environments and basic tasks
- Performance is saturating.

1.Stateless, non interactive environment, e.g.

*Mind2Web (Deng et al. 2023)* has only dumped pages.

2.Checking action sequence accuracy (step-wise, surface form only)

3.Simple interactive environment, short horizon, e.g.

*WebShop (Yao et al. 2023), MiniWoB++ (Humphreys et al. 2022)*

## Task

Follow one of the team leaders of one of the NHL teams from the Atlantic Division

## Actions

1. HOVER ➤ [link]  
*ESPN NHL Home Page*
2. CLICK ➤ [link]  
*Teams*
3. CLICK ➤ [heading]  
*Boston Bruins*
4. CLICK ➤ [link]  
*Full Team Statistics*
5. CLICK ➤ [heading]  
*David Pastrnak RW*
6. CLICK ➤ [button]



Move the cube around so that "5" is the active side facing the user.



Submit

Set the sliders to the combination [13,20,13] and submit.

Three sliders with values 14, 19, and 17. A mouse cursor is on the second slider.

Submit

Draw the number "2" in the checkboxes using the example on the right and press Submit when finished.

A 4x4 grid of checkboxes and a 4x4 grid of squares. The number 2 is drawn in the checkboxes.

Submit

Drag Ree to the 4th position.

A list of names: Jade, Ree, Karlen, Millie, Noelyn. A mouse cursor is over 'Ree'.

Keep your mouse inside the circle as it moves around.

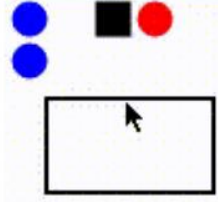


Enter the value of Country into the text field and press Submit.

A form with fields for Gender (Male), First name (Annecorinne), Country (Guam), Year of Birth (1934), and Religion (Hinduism).

Submit

Drag all triangles into the black box.



Submit

Select 09/23/2016 as the date and hit submit.

A date picker showing September 2016 with the 23rd selected.

Sort the numbers in increasing order, starting with the lowest number at the top of the list.

A list of numbers: 9, -12, 49, -28. A mouse cursor is over -28.

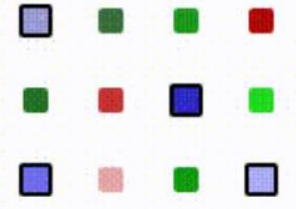
Submit

Copy the text from the 1st text area below and paste it into the text input.

Text areas containing 'Blandit quisque.', 'Amet ac odio aliquam.', and 'Ultrices ornare'. A mouse cursor is over the first text area.

Submit

Select all the shades of blue and press Submit.



Submit

Find the 4th word in the paragraph, type that into the textbox and press "Submit".

Non arcu ut ultrices est. Gravidam gravida. Porta erat nulla eget condimentum posuere a

A text input field with the cursor at the end.

Submit

Click the button in the dialog box labeled "Cancel".

A dialog box with the text 'Massa. Facilisi luctus. Est vel.' and buttons for 'Cancel' and 'OK'. A mouse cursor is over 'Cancel'.

Highlight the text in the paragraph below and click submit.

A 'Submit' button and a text field containing 'Ultrices. Sagittis in' which is highlighted in blue.

Highlight the text in the paragraph below and click submit.

Tempor posuere nibh. Vel nisl, faucibus. Feugiat condimentum

Submit

Find the 11th word in the paragraph, type that into the textbox and press "Submit".

Ullamcorper aliquet amet ullamcorper. Elit. Mattis luctus diam. Lobortis nulla fermentum ornare faucibus

A text input field.

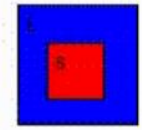
Submit

Move the cube around so that "2" is the active side facing the user.



Submit

Drag the smaller box so that it is completely inside the larger box.



Submit

Instruction: i am looking for x-large, red color women faux fur lined winter warm jacket coat, and price lower than 70.00 dollars

Current Query: women fur jacket coat

## Results

Page 1 (1-10) of 50 total results

[Back to Search](#)

[Next >](#)



[B09KP78G37](#)

Women Faux Fur Lined Jacket Coat  
Winter Warm Thick Fleece Outwear  
Trench Zipper Plus Size Long  
Sleeve Plush Overcoat



[B07ZXBGDXF](#)

Women's Coat, FORUU Winter Faux  
Fur Fleece Outwear Warm Lapel  
Biker Motor Aviator Jacket



[B098XT346Y](#)

Fjackets Real Lambskin Sherpa  
Jacket - Mens Leather Jacket

★★★★★ 4.7

**Current Action: click[Fjackets Real Lambskin...]**



# Key to Agent Benchmarks

## Environment:

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

## Tasks:

- Long horizon tasks
- Enough difficulty
- Involves multiple websites

## Evaluation:

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

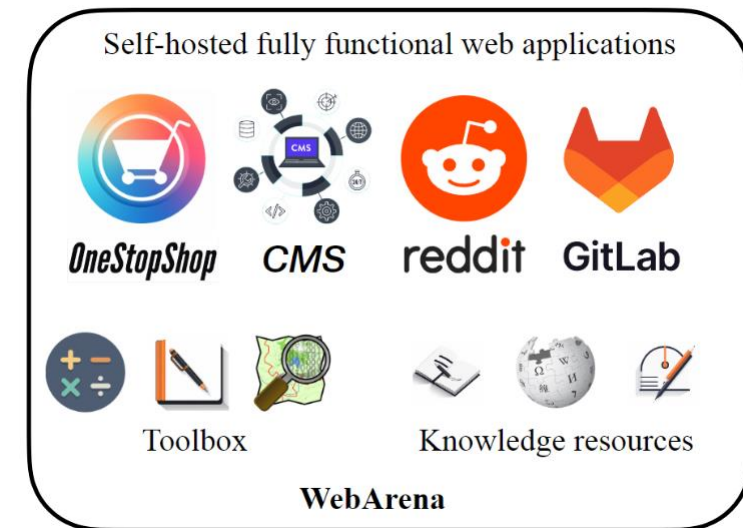
# WebArena Environment Design

## Environment:

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

## A sandbox Internet:

- Open source, production-ready implementation of the websites
- Data populated from real-world websites
- Easily distributable – Docker, AWS images, etc.



WebArena (Zhou et al. 23')

# Collecting Realistic Intents

We checked our own web browsing history!

- Information seeking
  - *“When was the last time I bought shampoo?”*
- Site navigation
  - *“Checkout merge requests assigned to me.”*
- Content & configuration operation
  - *“Post my question, “is a car necessary in NYC”, in a subreddit where I’m likely to get an answer.”*

# Example Tasks 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”

webarena.wikipedia.com

Wikipedia Pittsburgh museums

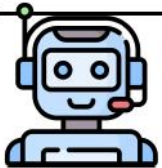
## List of museums in Pittsburgh

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

webarena.openstreetmap.com

OpenStreetMap Edit History Export

Schenley Park, Pittsburgh, Allegheny County

The Andy Warhol Museum, 117, Sandusky Str

Car (OSRM) Go

Reverse Directions

## 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

webarena.gitlab.com

Update README.md

README.md 158 B Edit Replace

## 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

- “When was the last time I bought shampoo?”
- **Directly compare with the annotated answer:** Answer is “Dec 15th, 2022”

# Outcome/Execution-based Evaluation

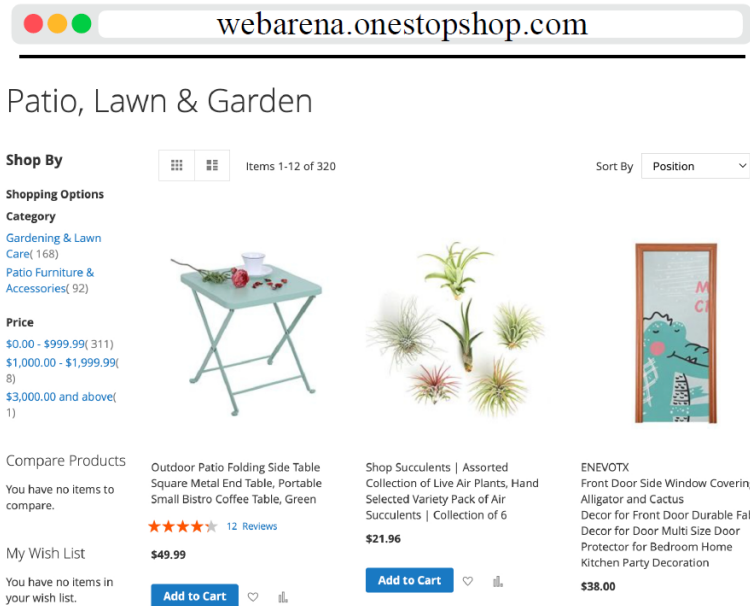
*Post my question, "is car necessary in NYC", in a subreddit where I'm likely to get an answer*

The screenshot shows a browser window with the URL `webarena.redd.it.com/f/nyc/28/need-your-`. The browser's navigation bar is red and contains links for Postmill, Forums, Wiki, a search bar, and an email icon. Below the navigation bar, the page title is `/f/nyc`. The main content shows a post titled "Need your answer" submitted by `convexeggtarxxx` 0 seconds ago in the `nyc` subreddit. The post content is "is car necessary in NYC?". Below the post, there are options for "No comments", "Edit", and "Delete". At the bottom, the browser's developer tools are open, showing the DOM tree. The selected element is `<div class="submission_inner">`, which contains a `<div class="submission_content flow-slim">` element. This element contains a `<div class="submission_body break-text text-flow">` element, which contains a `<p lang="en" dir="ltr">is car necessary in NYC?</p>` element.

"f/nyc" in `page.url`

"Is car necessary in NYC?" in `document.querySelector(".submission_inner").outText`

# Observation & Action Space



Screenshot

Keyboard: type

```

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

Text

Mouse: click, hover, scroll

```

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

Browser: New tab, go back

# Prompting LLM as Agent

- ***Few-shot in-context learning***: General guideline + two examples

You are an autonomous intelligent agent tasked with navigating a web browser. You will be given web-based tasks. These tasks will be accomplished through the use of specific actions you can issue.

*You can observe the following information:*

...

*You can do the following actions:*

...

...



# Prompting LLM as Agent

- ***Few-shot in-context learning***: General guideline + two examples

Example input:

**OBSERVATION:**

[1744] link 'HP CB782A#ABA 640 Inkjet Fax Machine (Renewed)'

[1757] button 'Add to Cart'

**URL:** <http://onestopmarket.com/office-products/office-electronics.html>

**OBJECTIVE:** What is the price of HP Inkjet Fax Machine

**PREVIOUS ACTION:** None

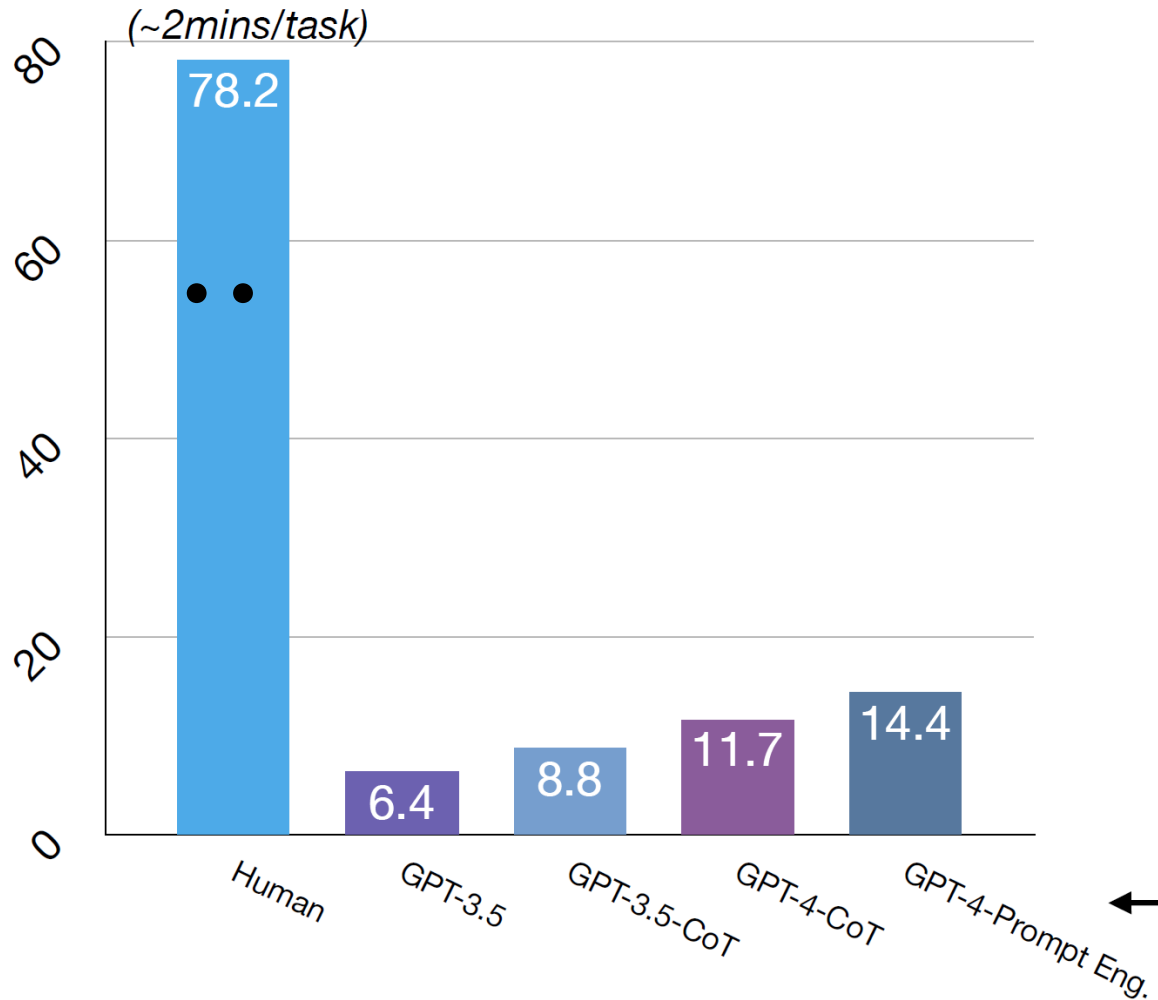
Example output:

Let's think step-by-step. This page lists ... the next action I will perform is `click [1744]`

(Optional chain-of-thought reasoning)

(Issued action)


# WebArena is Challenging





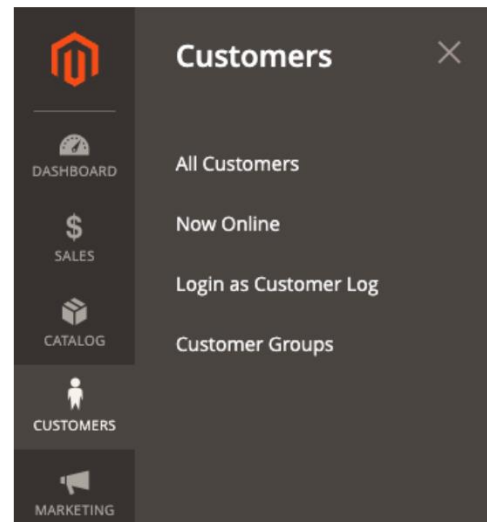
- Chain-of-thought prompting provides *limited* benefits.
- GPT-4 remains *significantly behind* human performance.
- Prompt engineering emphasizes LLMs' *sensitivity* to subtle instruction changes that typically don't affect humans.

# Failures: Not Knowing How

 Show me the customers who have expressed dissatisfaction with Olivia zip jacket

 Either going to the **catalog (product)** section or the **marketing (review)** section

  Decided to go to **customers** section which is not easy to select and filter reviews



...d of your business' performance, using our dynamic product, order, and customer reports

# Failures: Not being Accurate

“... and set the due date to 2023 / 12 / 23”

Due date



“... and set the due date to 2023-12-13”

Due date



Due date

December 2023						
sun	mon	tue	wed	thu	fri	sat
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31						

# Failures: Trivial Errors



## Search

Search query

DMV area

Search

50 results for *DMV area*:

[2430] searchbox 'Search query'  
[5172] StaticText 'DMV area'

Search query

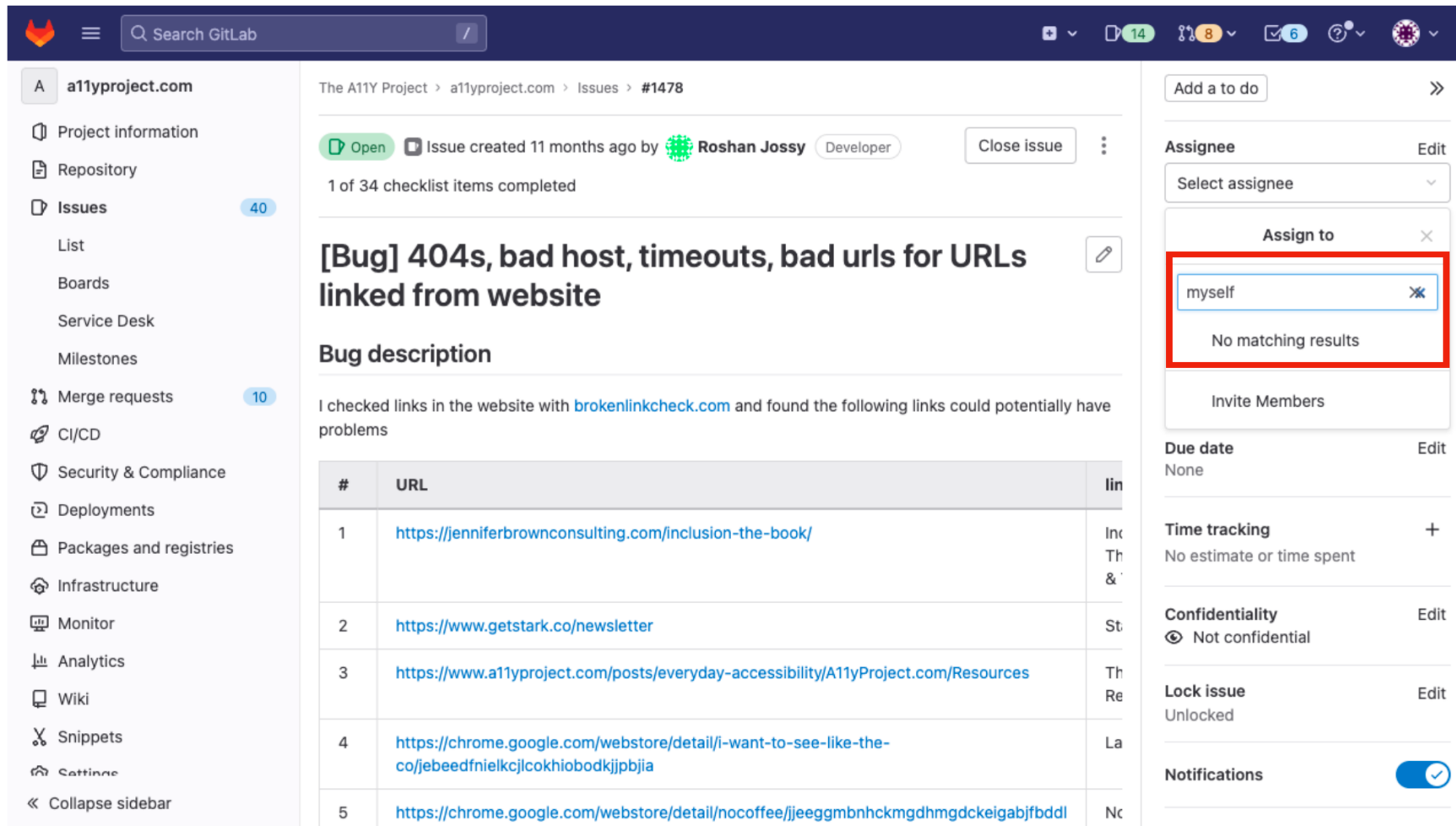
DMV areaDMV areaDMV areaDMV area

Search

- GPT-4 : 21% examples failed due to repeated typing.
- May be related to hallucination effect, generates repeated actions
- Irrelevant content in a webpage hurts!

# Failures: (Not So) Trivial Errors

- “Assign this issue to myself.”



The screenshot shows a GitLab issue page for a11yproject.com. The issue is titled "[Bug] 404s, bad host, timeouts, bad urls for URLs linked from website" and was created 11 months ago by Roshan Jossy. The issue description states: "I checked links in the website with brokenlinkcheck.com and found the following links could potentially have problems". A table lists five broken links. On the right side, the "Assign to" dropdown menu is open, showing "myself" as the selected option, which is highlighted with a red box. Below it, it says "No matching results".

**Issue Details:**

- Project: a11yproject.com
- Issue #1478
- Status: Open
- Created: 11 months ago by Roshan Jossy (Developer)
- Checklist: 1 of 34 items completed

**Bug description:**

I checked links in the website with [brokenlinkcheck.com](https://brokenlinkcheck.com) and found the following links could potentially have problems

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

**Assign to dropdown:**

- myself (selected)
- No matching results
- Invite Members

# Training Methods for Improving Agents

# Learning of LLM Agents

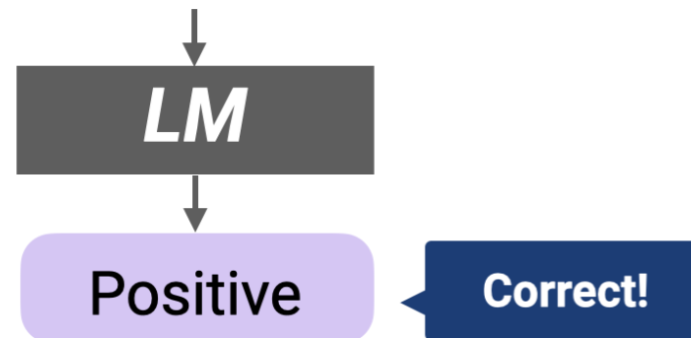
- In-Context Learning – Learning from few-shot exemplars
- Supervised Finetuning – Learning From *Experts*
- Reinforcement Learning – Learning from *Environment*



# In-context Learning

- LM performs a task just by ***conditioning on input-output examples***, without optimizing any parameters.

Circulation revenue has increased by 5% in Finland. \n Positive  
Panostaja did not disclose the purchase price. \n Neutral  
Paying off the national debt will be extremely painful. \n Negative  
The company anticipated its operating profit to improve. \n \_\_\_\_\_



# In-context Learning

## example\_user

OBSERVATION:

[1744] link 'HP CB782A#ABA 640 Inkjet Fax Machine (Renewed)'

[1749] StaticText '\$279.49'

[1757] button 'Add to Cart'

[1760] button 'Add to Wish List'

[1761] button 'Add to Compare'

URL: <http://onestopmarket.com/office-products/office-electronics.html>

OBJECTIVE: What is the price of HP Inkjet Fax Machine

PREVIOUS ACTION: None

## example\_assistant

Let's think step-by-step. This page lists the information of HP Inkjet Fax Machine, which is the product identified in the objective. Its price is \$279.49. I think I have achieved the objective. I will issue the stop action with the answer. In summary, the next action I will perform is `` ` stop [\$279.49]` ``

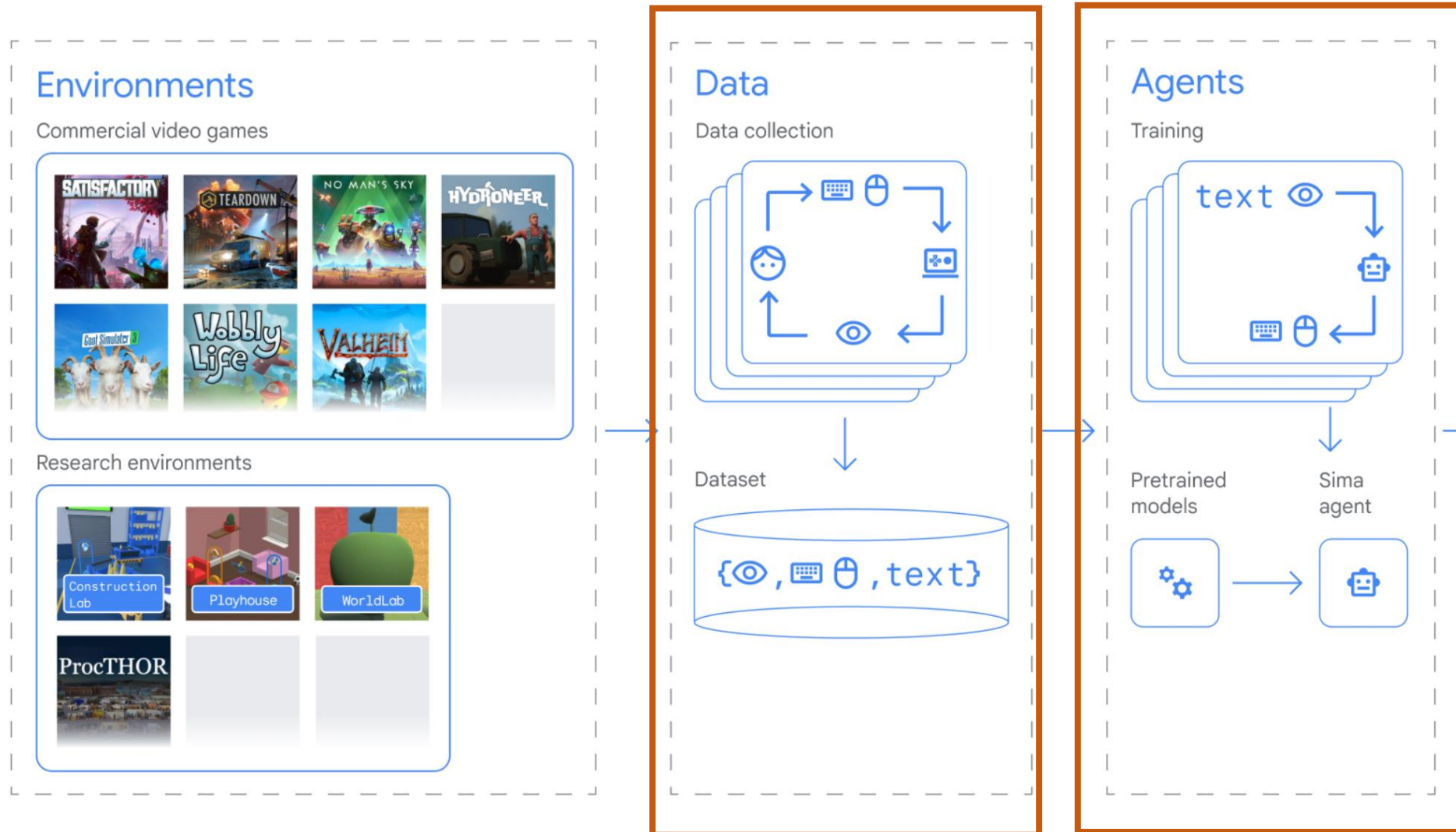
# Supervised Finetuning

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

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

- Finetune the LLM with standard cross-entropy loss.

# Supervised Finetuning



# Supervised Finetuning

- Data hungry
- Cannot learn much from failed trajectories
  - $a_1, a_2, a_3, \dots, a_{10}$  - Success
  - $a_1, a_2, a_3, \dots, a_{10}$  - Fail (Wasted)
- Need human trajectory?
  - Data augmentation techniques

# Create More Training Data

## Open-ended Environments

Craft Glass Bridge



Build Oak House



Make Ice Igloo



Combat Zombie



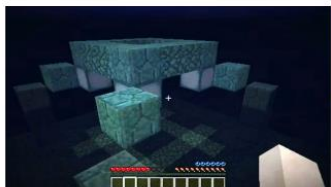
Fish Squid



Farm Sugar Cane



Find Ocean Monument



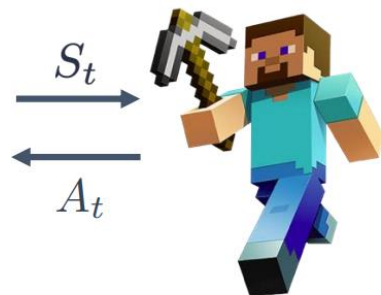
Explore Desert Temple



Treasure Hunt in End City



## Generalist Agent



## Internet-scale Knowledge Base

**YouTube**

**Wiki**

Features	Description	Screenshot	[hide]
<p>River</p> <p>Water, Sand, Clay, Sugar Cane, Seagrass, Salmon, Squid, Drowned</p>	<p><b>Temperature: 0.5. Rainfall: 0.5.</b> A biome that consists of water blocks in an elongated, curving shape similar to a real river. Rivers are a reliable source of clay. They are good for fishing, but drowned can spawn at night.</p>	<p>River</p>	

**Reddit**

r/Minecraft · Posted by u/Anime-ghostGirl 6 days ago

190 I present to you me struggling to get up stairs in the end city

i dig a staircase in the wall ^^

Or just use enderpearl.

Water is useful in a lot of situations. Early game, and late game



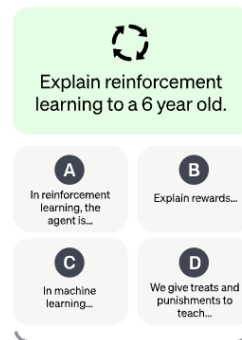


# Reinforcement Learning

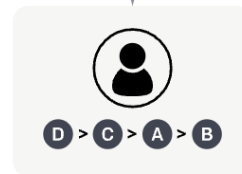
Lots of on-going research in this area!

Recall RLHF: Reinforcement Learning from Human Feedback:

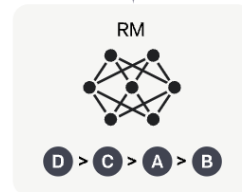
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



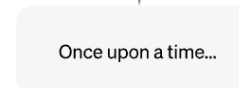
A new prompt is sampled from the dataset.



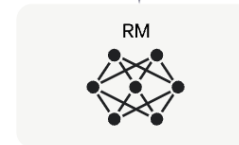
The PPO model is initialized from the supervised policy.



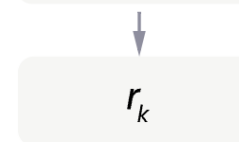
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

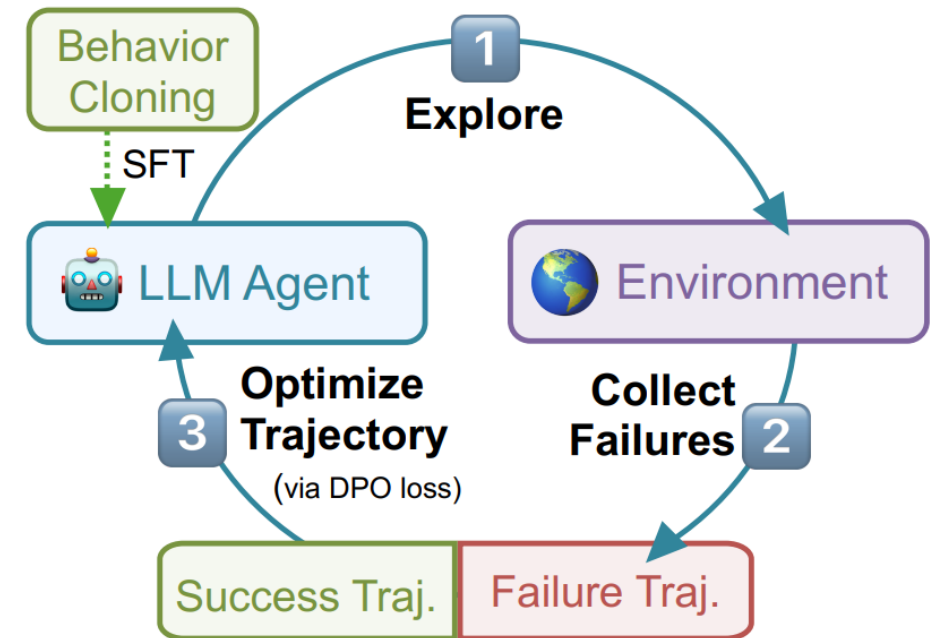
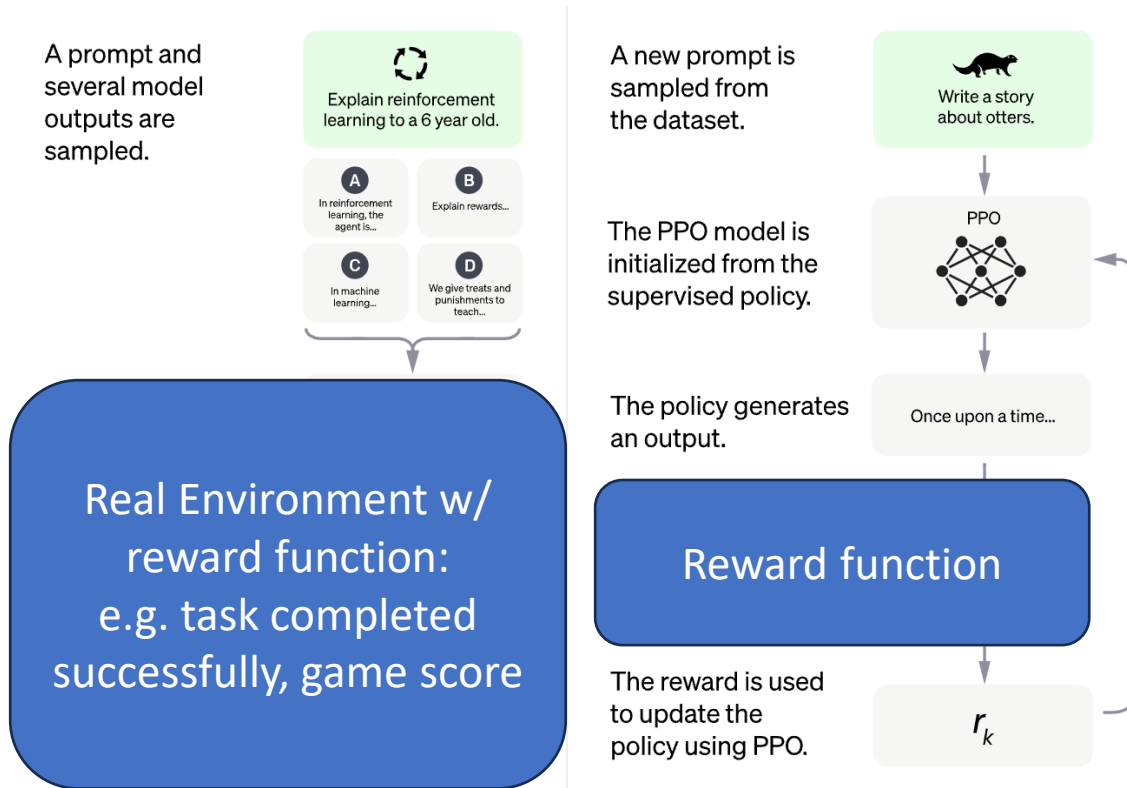




# Reinforcement Learning

Compared to RLHF:

Given environment, *reward function*  
(*trajectory, reward*) pairs without human



# 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 models
  - 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

# Reinforcement Learning

- No access to policy model parameters? (e.g. GPT-4)
- We can train with environment reward a second LLM (smaller open source models), that generates additional “reflection” prompts.

