Inducing Functions to Improve LLM Agents

Daniel Fried

Language Technologies Institute

Carnegie Mellon University

Language Interfaces

Science Fiction

Her, 2013

Let’s start with your emails. You have several thousand emails regarding LA Weekly, but it looks like you haven’t worked there in many years.

Oh yeah, I guess I was saving those because in some of them I thought I might have written some funny stuff.

Yeah, there are some funny ones. I’d say there are about 86 that we should save. We can delete the rest.

Today

ChatGPT, 2025

Please help me organize my emails.

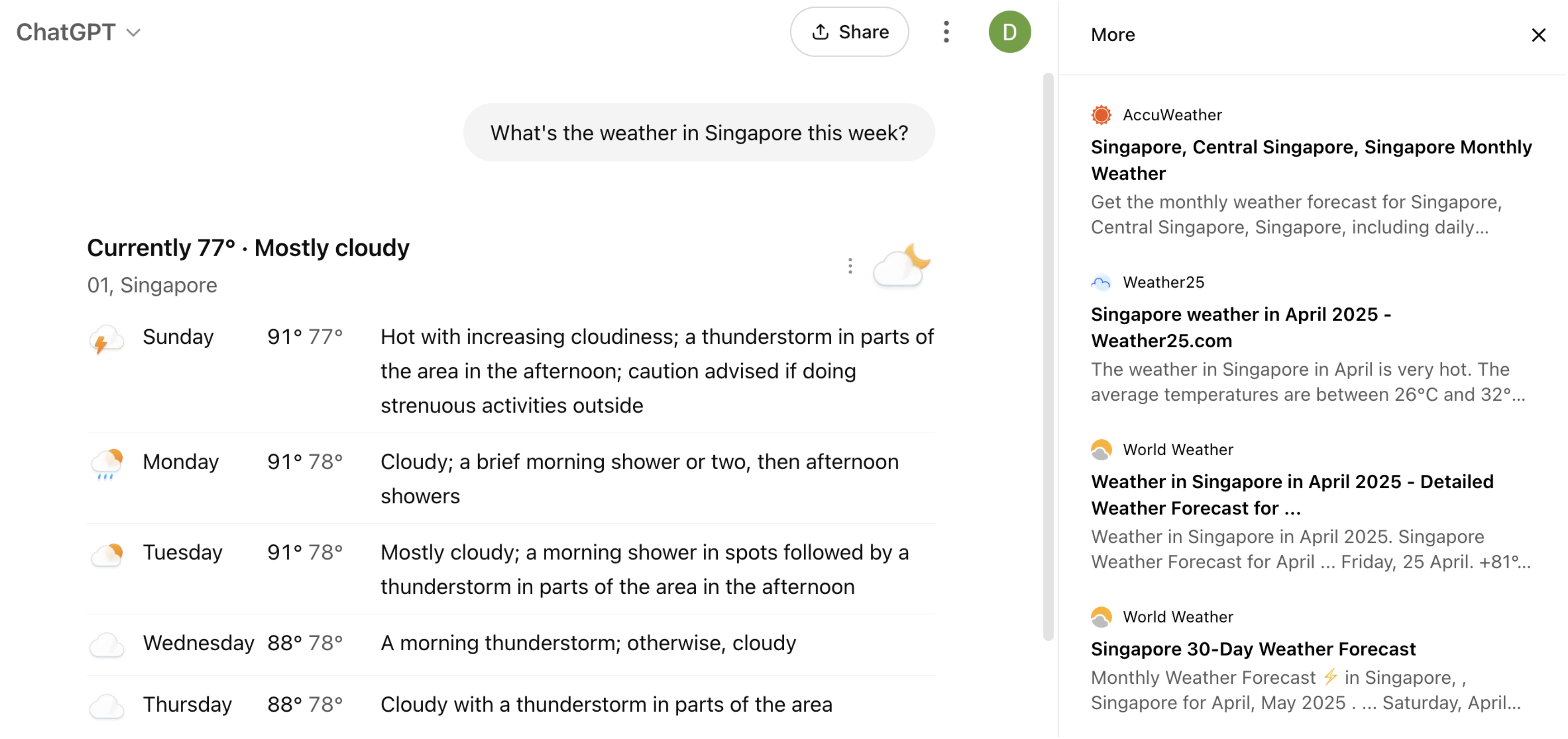
Absolutely! I can help you get your inbox in shape. To tailor the advice a bit, could you tell me:

1. What email provider are you using? (e.g., Gmail, Outlook, Apple Mail)
2. Do you use any email apps (on desktop or phone)?

Thanks, but I want you to do it!

I wish I could jump in and drag those emails into folders for you—but I can't directly access your inbox. 😅 I’ll guide you step-by-step, and we’ll get your inbox whipped into shape fast.

Tool Use



What's the weather in Singapore this week?

Currently 77° · Mostly cloudy

01, Singapore

Sunday	91° 77°	Hot with increasing cloudiness; a thunderstorm in parts of the area in the afternoon; caution advised if doing strenuous activities outside
Monday	91° 78°	Cloudy; a brief morning shower or two, then afternoon showers
Tuesday	91° 78°	Mostly cloudy; a morning shower in spots followed by a thunderstorm in parts of the area in the afternoon
Wednesday	88° 78°	A morning thunderstorm; otherwise, cloudy
Thursday	88° 78°	Cloudy with a thunderstorm in parts of the area

AccuWeather

Singapore, Central Singapore, Singapore Monthly Weather

Get the monthly weather forecast for Singapore, Central Singapore, Singapore, including daily...

Weather25

Singapore weather in April 2025 - Weather25.com

The weather in Singapore in April is very hot. The average temperatures are between 26°C and 32°...

World Weather

Weather in Singapore in April 2025 - Detailed Weather Forecast for ...

Weather in Singapore in April 2025. Singapore Weather Forecast for April ... Friday, 25 April. +81°...

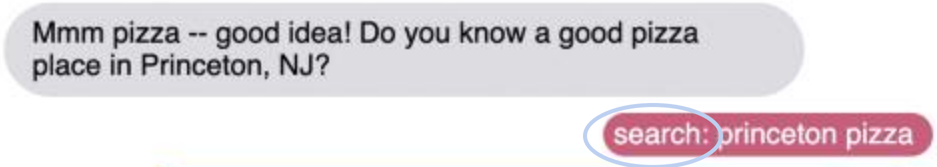
World Weather

Singapore 30-Day Weather Forecast

Monthly Weather Forecast ⚡ in Singapore, , Singapore for April, May 2025 . ... Saturday, April...

Sources of Tools

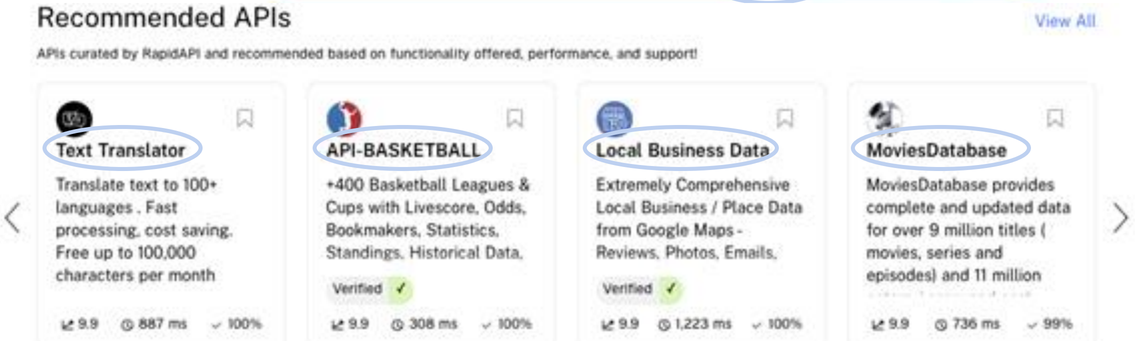
Application-specific software



Mmm pizza -- good idea! Do you know a good pizza place in Princeton, NJ?

search: princeton pizza

Real-world APIs



Recommended APIs

APIs curated by RapidAPI and recommended based on functionality offered, performance, and support!

Text Translator

Translate text to 100+ languages . Fast processing, cost saving. Free up to 100,000 characters per month

9.9 887 ms 100%

API-BASKETBALL

+400 Basketball Leagues & Cups with Livescore, Odds, Bookmakers, Statistics, Standings, Historical Data,

Verified ✓

9.9 308 ms 100%

Local Business Data

Extremely Comprehensive Local Business / Place Data from Google Maps - Reviews, Photos, Emails,

Verified ✓

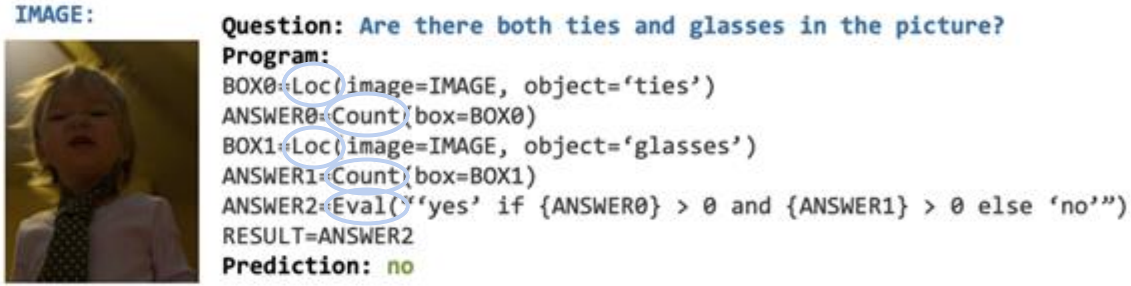
9.9 1,223 ms 100%

MoviesDatabase

MoviesDatabase provides complete and updated data for over 9 million titles (movies, series and episodes) and 11 million

9.9 736 ms 99%

Expert-designed functions



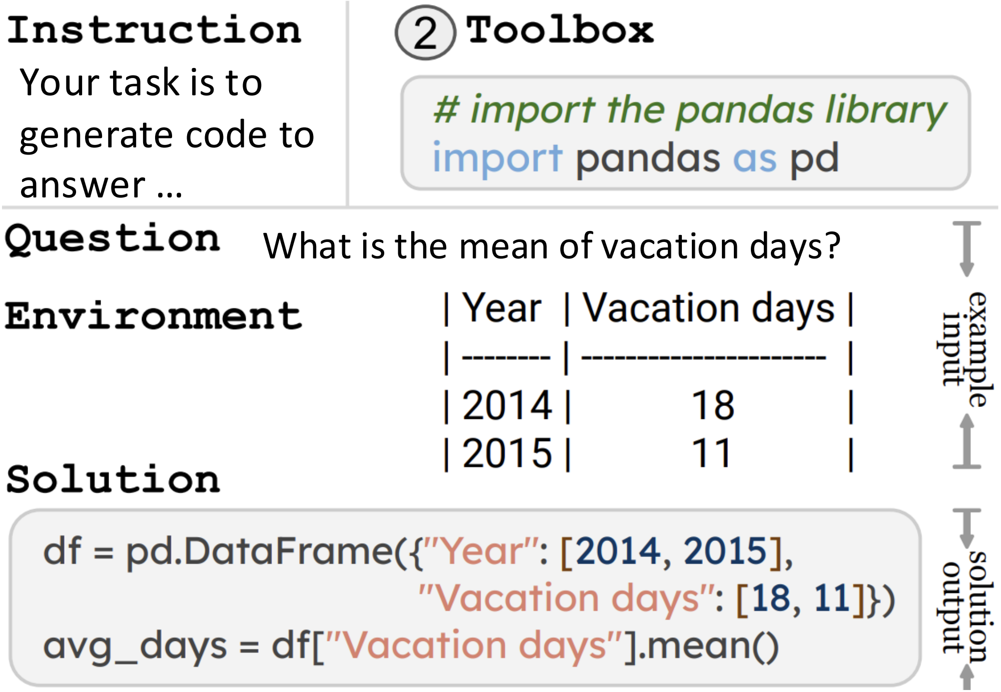
```
IMAGE:
Question: Are there both ties and glasses in the picture?
Program:
BOX0=Loc(image=IMAGE, object=‘ties’)
ANSWER0=Count(box=BOX0)
BOX1=Loc(image=IMAGE, object=‘glasses’)
ANSWER1=Count(box=BOX1)
ANSWER2=Eval(“‘yes’ if {ANSWER0} > 0 and {ANSWER1} > 0 else ‘no’”)
RESULT=ANSWER2
Prediction: no
```

Zora Wang

Creating Tools on-the-Fly

Zora Zhiruo Wang, Graham Neubig, and Daniel Fried. “TroVE: Inducing Verifiable and Efficient Toolboxes for Solving Programmatic Tasks.” ICML 2024

Answering Questions via Code Generation



Instruction
Your task is to generate code to answer …

② **Toolbox**

```
# import the pandas library
import pandas as pd
```

Question What is the mean of vacation days?

Environment

Year	Vacation days
2014	18
2015	11

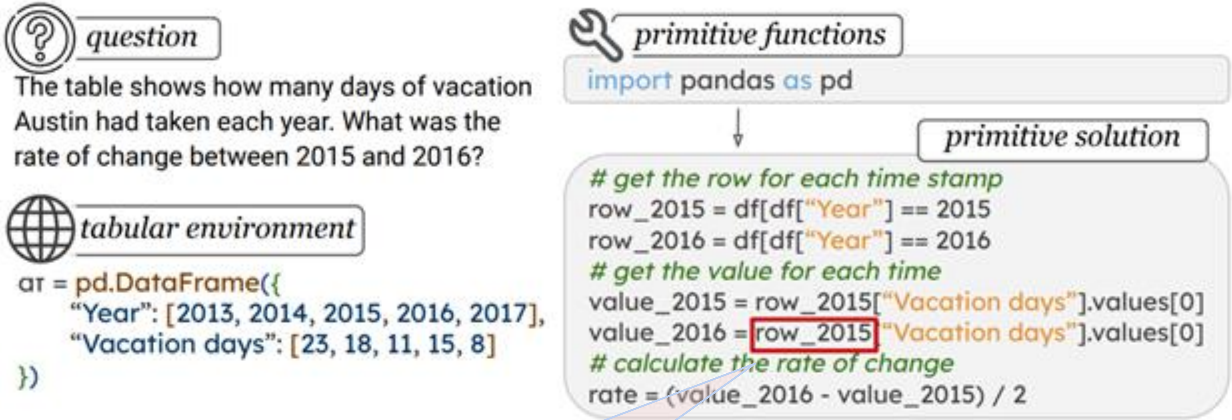
Solution

```
df = pd.DataFrame({"Year": [2014, 2015],
                   "Vacation days": [18, 11]})
avg_days = df["Vacation days"].mean()
```

example input
solution output

What If No Useful Tools Exist Yet?

- Can only solve problems using primitive actions, e.g., Python built-in functions



question

The table shows how many days of vacation Austin had taken each year. What was the rate of change between 2015 and 2016?

tabular environment

```
df = pd.DataFrame({
      "Year": [2013, 2014, 2015, 2016, 2017],
      "Vacation days": [23, 18, 11, 15, 8]
})
```

primitive functions

```
import pandas as pd
```

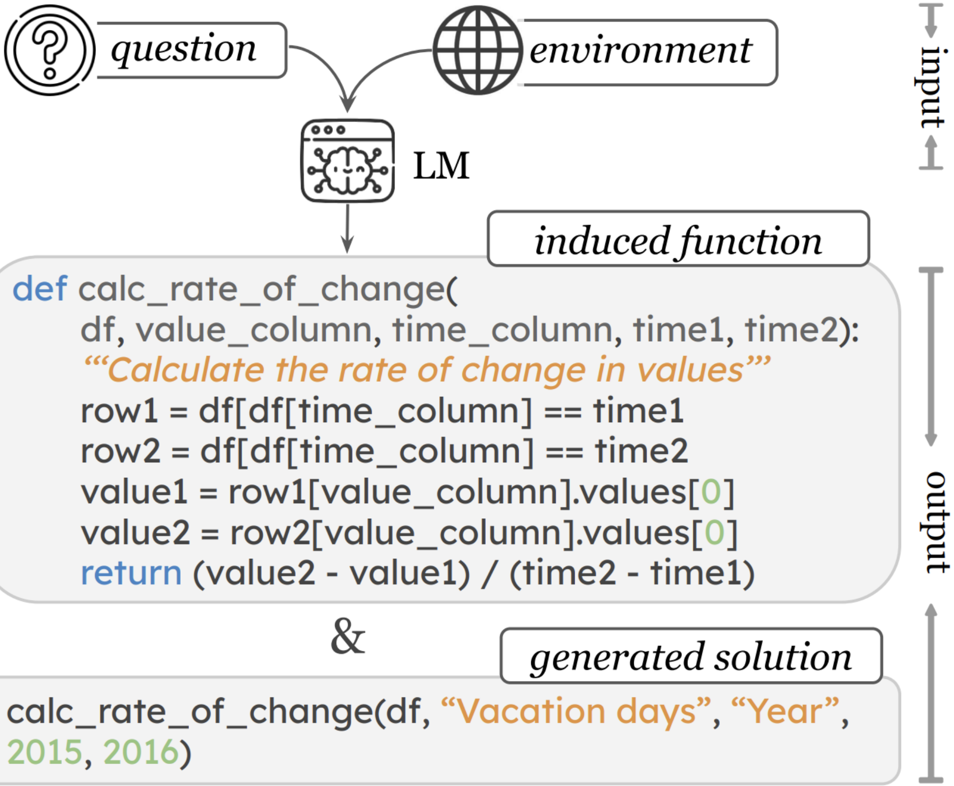
primitive solution

```
# get the row for each time stamp
row_2015 = df[df["Year"] == 2015
row_2016 = df[df["Year"] == 2016
# get the value for each time
value_2015 = row_2015["Vacation days"].values[0]
value_2016 = row_2015["Vacation days"].values[0]
# calculate the rate of change
rate = (value_2016 - value_2015) / 2
```

Prone to errors :(

Or let the agents make new tools!

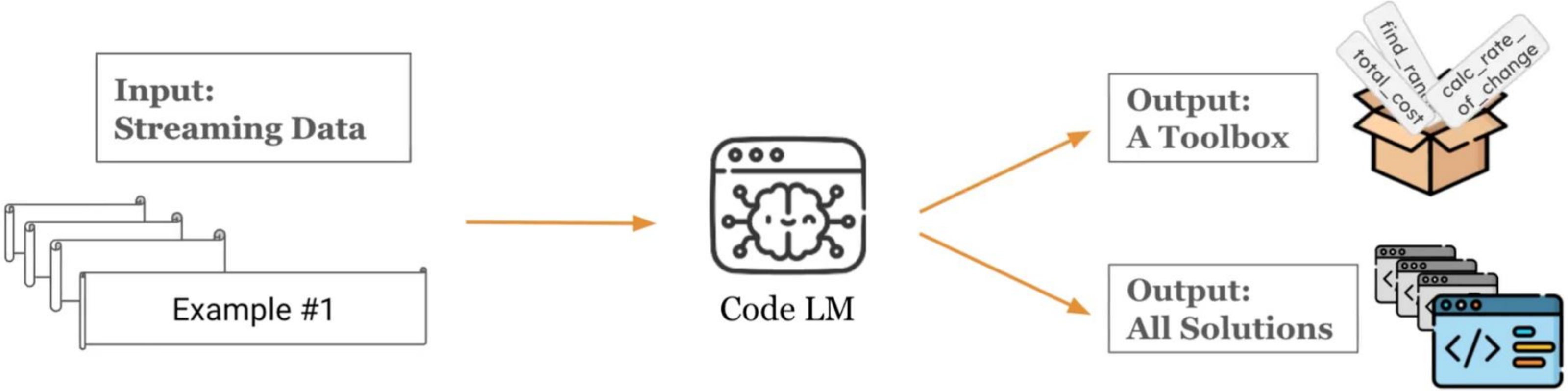
Generating Reusable Tools



```
def calc_rate_of_change(
        df, value_column, time_column, time1, time2):
        ‘“Calculate the rate of change in values”’
        row1 = df[df[time_column] == time1
        row2 = df[df[time_column] == time2
        value1 = row1[value_column].values[0]
        value2 = row2[value_column].values[0]
        return (value2 - value1) / (time2 - time1)
```

```
calc_rate_of_change(df, “Vacation days”, “Year”,
2015, 2016)
```

How does TroVE make tools?



Input:
Streaming Data

Example #1

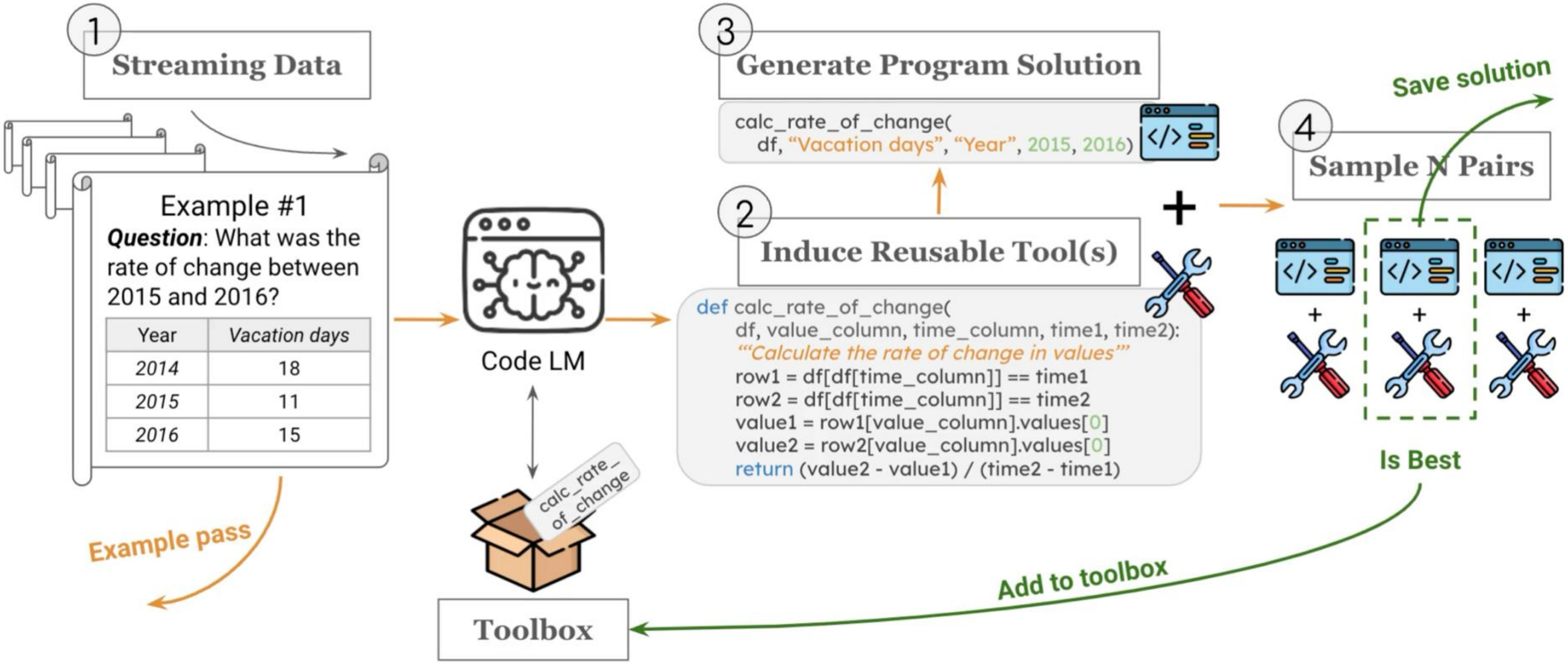
Code LM

Output:
A Toolbox

Output:
All Solutions

How does TroVE make tools?

Creating Tools



Example #1

Question: What was the rate of change between 2015 and 2016?

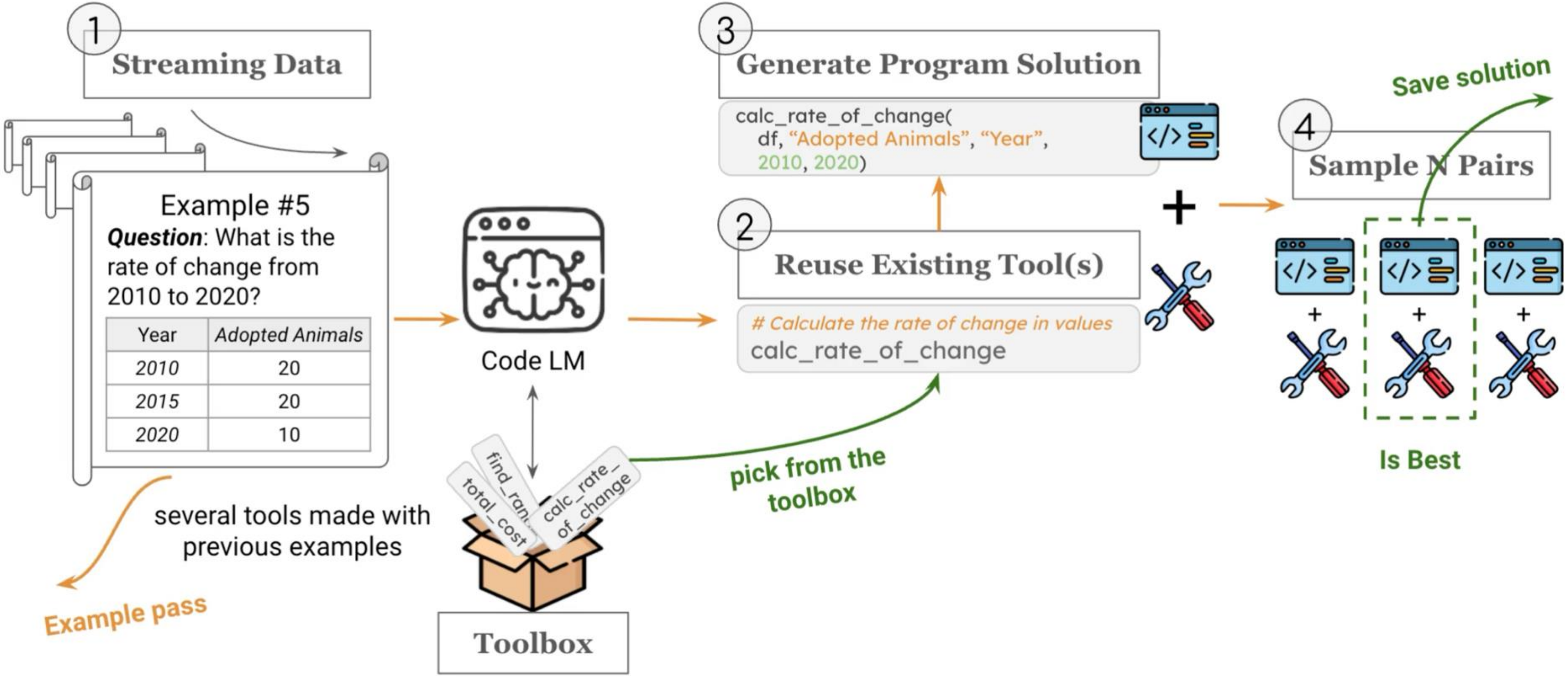
Year	*Vacation days*
2014	18
2015	11
2016	15

```
calc_rate_of_change(
   df, "Vacation days", "Year", 2015, 2016)
```

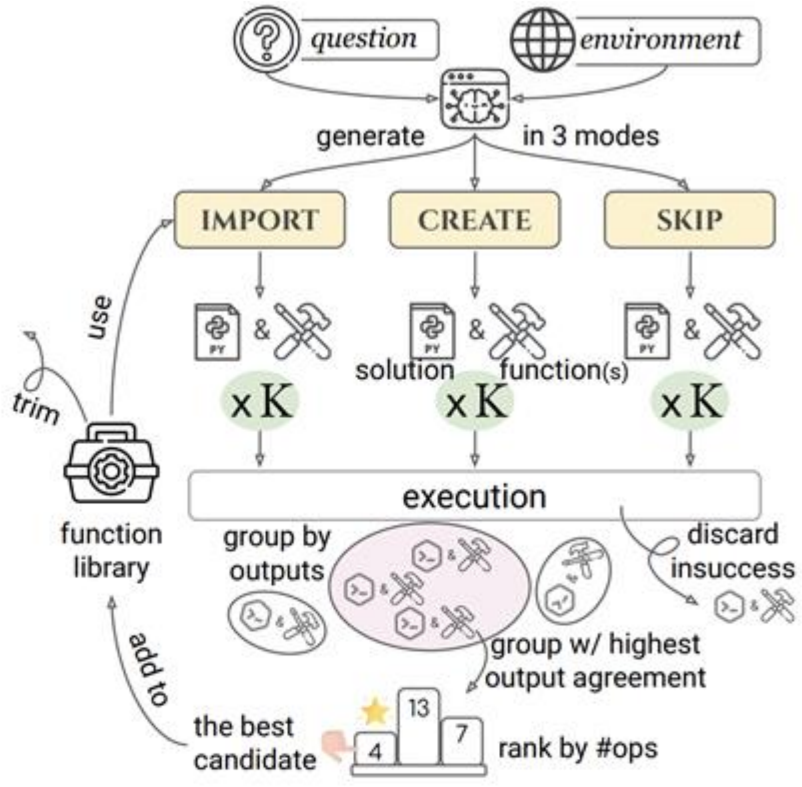
```
def calc_rate_of_change(
    df, value_column, time_column, time1, time2):
    '''Calculate the rate of change in values'''
    row1 = df[df[time_column]] == time1
    row2 = df[df[time_column]] == time2
    value1 = row1[value_column].values[0]
    value2 = row2[value_column].values[0]
    return (value2 - value1) / (time2 - time1)
```

How does TroVE make tools?

Importing Tools

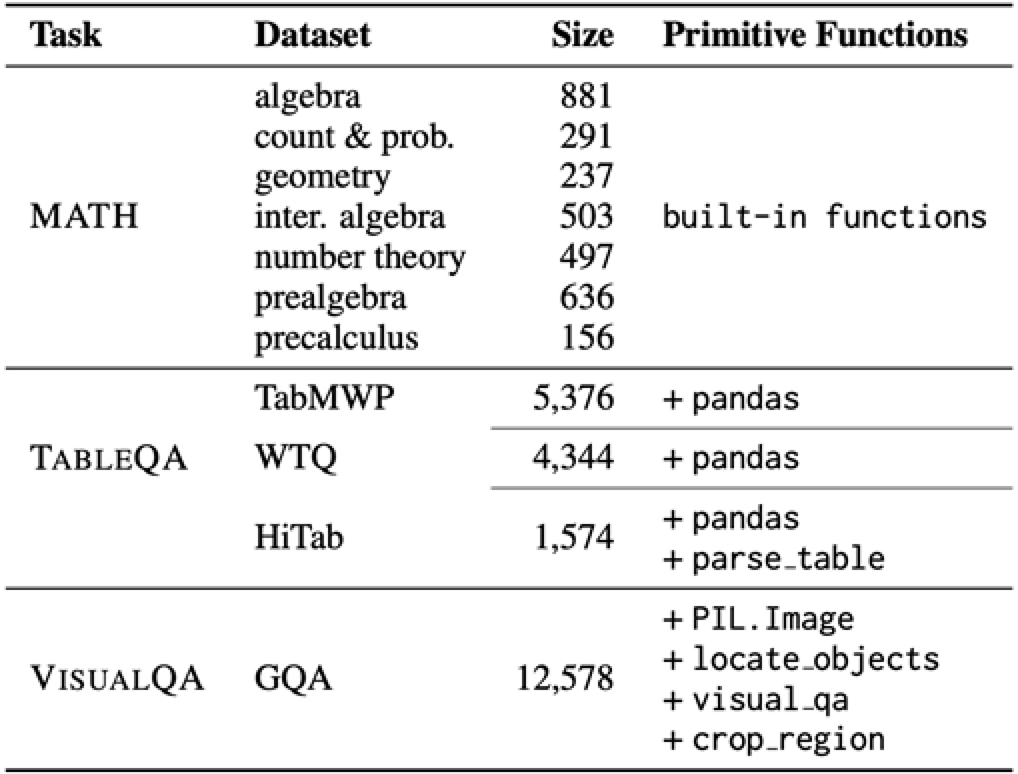


How does TroVE make tools?



- Using and growing the toolbox
- Agreement-based selection
- Periodic toolbox trimming

Evaluation on Programmatic Tasks

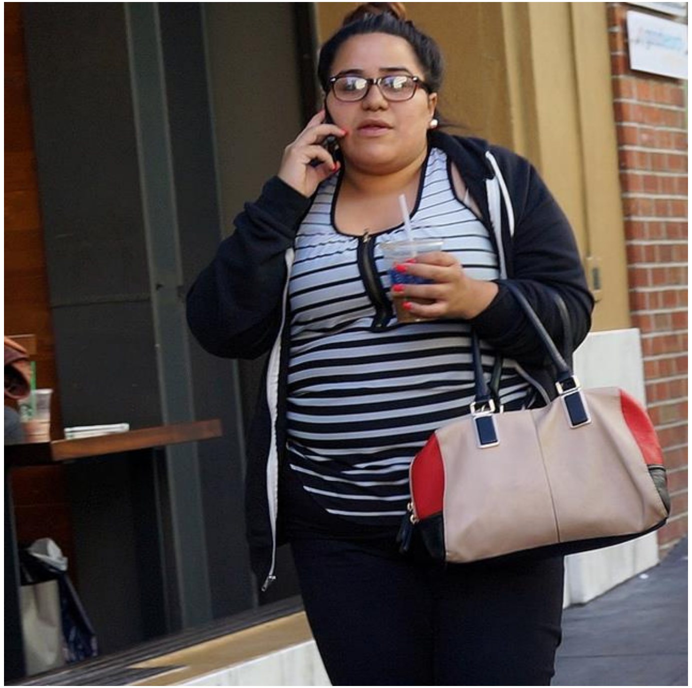


Task	Dataset	Size	Primitive Functions
MATH	algebra	881	`built-in functions`
	count & prob.	291	
	geometry	237	
	inter. algebra	503	
	number theory	497	
	prealgebra	636	
	precalculus	156	
TABLEQA	TabMWP	5,376	+ `pandas`
	WTQ	4,344	+ `pandas`
	HiTab	1,574	+ `pandas` + `parse_table`
VISUALQA	GQA	12,578	+ `PIL.Image` + `locate_objects` + `visual_qa` + `crop_region`

Table 1. Statistics and primitives for three tasks.

Example from GQA

On which side of the picture is the leather bag?



Signature of Created Function

```
def get_object_side(image: str | Image.Image,
object_name: str) -> str:
  """Get the side of the image where the
object is located.""”
  ...
```

Solution

```
from PIL import Image

image_file = "data/gqa/testdev_images/n313060.jpg"
image = Image.open(image_file).convert('RGB')

side = get_object_side(image, "leather bag")

print(side)
```

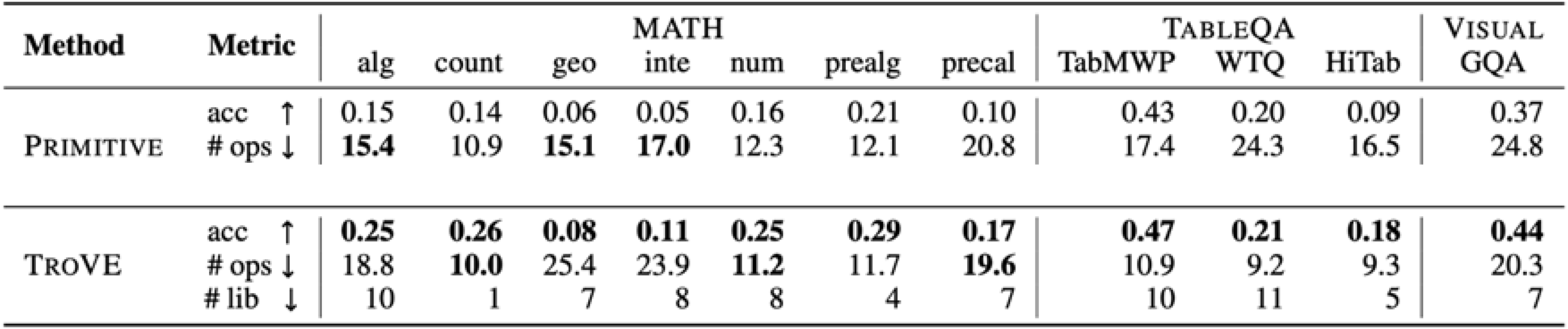
Example from GQA

Created Function:

```python
from PIL import Image
from toolbox import locate_objects

def get_object_side(image: str | Image.Image, object_name: str) -> str:
  """Get the side of the image where the object is located.""”

  boxes = locate_objects(image=image, object_name=object_name)

  if isinstance(image, str):
    image = Image.open(image).convert('RGB’)

  image_width = image.width
  object_center = (boxes[0][0] + boxes[0][2]) / 2

  if object_center < image_width / 2:
    return "left"
  else:
    return "right"
```

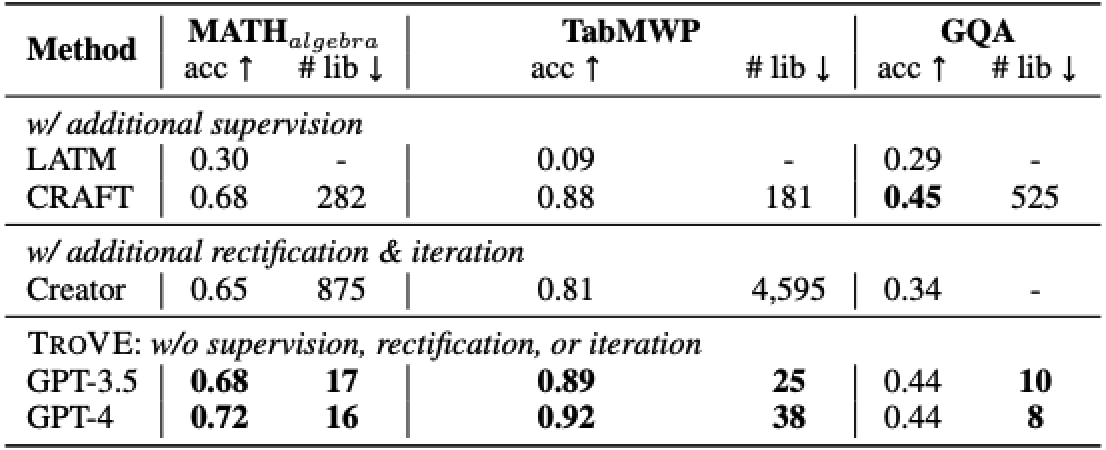
CodeLLaMa Performance w/ TroVE

- Answer correctness (acc ↑)
- Solution complexity (#ops ↓)



Method	**Metric**	MATH							TABLEQA			VISUAL
		alg	count	geo	inte	num	prealg	precal	TabMWP	WTQ	HiTab	GQA
PRIMITIVE	acc ↑	0.15	0.14	0.06	0.05	0.16	0.21	0.10	0.43	0.20	0.09	0.37
	# ops ↓	**15.4**	10.9	**15.1**	**17.0**	12.3	12.1	20.8	17.4	24.3	16.5	24.8
TROVE	acc ↑	**0.25**	**0.26**	**0.08**	**0.11**	**0.25**	**0.29**	**0.17**	**0.47**	**0.21**	**0.18**	**0.44**
	# ops ↓	18.8	**10.0**	25.4	23.9	**11.2**	11.7	**19.6**	10.9	9.2	9.3	20.3
	# lib ↓	10	1	7	8	8	4	7	10	11	5	7

GPT-4 Performance w/ TroVE



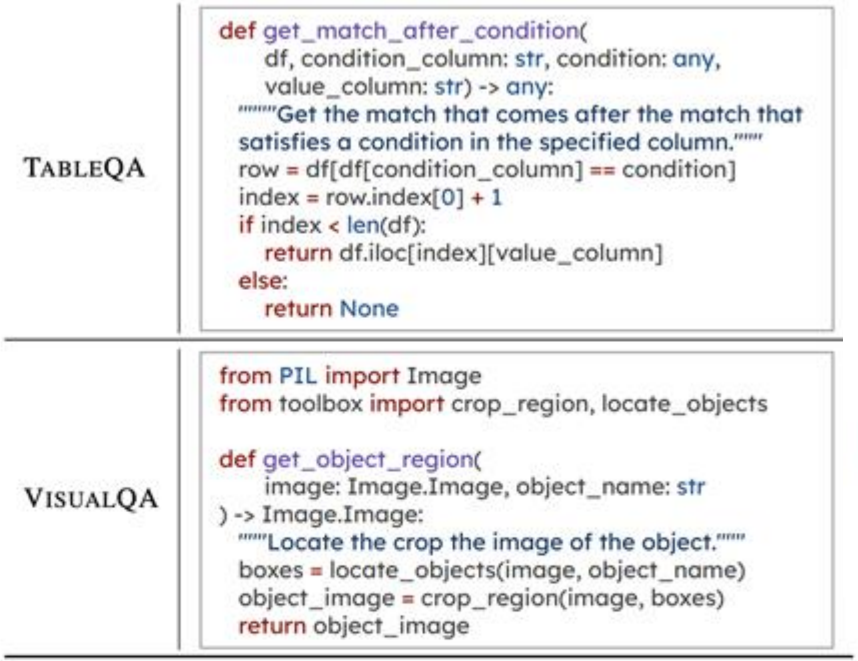
Method	**MATH$_{algebra}$** acc ↑	# lib ↓	**TabMWP** acc ↑	# lib ↓	**GQA** acc ↑	# lib ↓
w/ additional supervision						
LATM	0.30	-	0.09	-	0.29	-
CRAFT	0.68	282	0.88	181	**0.45**	525
w/ additional rectification & iteration						
Creator	0.65	875	0.81	4,595	0.34	-
TROVE: *w/o supervision, rectification, or iteration*						
GPT-3.5	**0.68**	**17**	**0.89**	**25**	0.44	**10**
GPT-4	**0.72**	**16**	**0.92**	**38**	0.44	**8**

- Higher accuracy
- Much smaller toolbox

- GPT-4 performs comparably to CodeLLaMa-7B on GQA

Diverse Tools Across Domains

- Varied function types across tasks



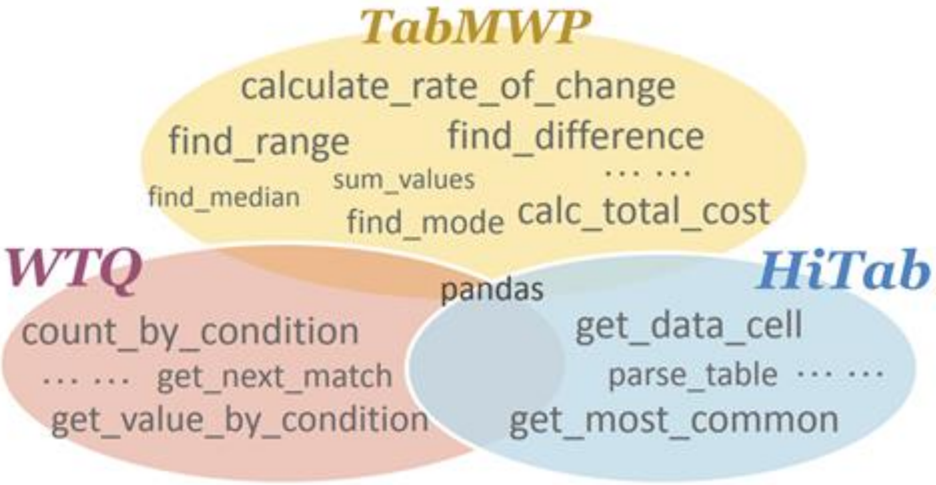
TABLEQA

```python
def get_match_after_condition(
        df, condition_column: str, condition: any,
        value_column: str) -> any:
    """Get the match that comes after the match that
    satisfies a condition in the specified column."""
    row = df[df[condition_column] == condition]
    index = row.index[0] + 1
    if index < len(df):
        return df.iloc[index][value_column]
    else:
        return None
```

VISUALQA

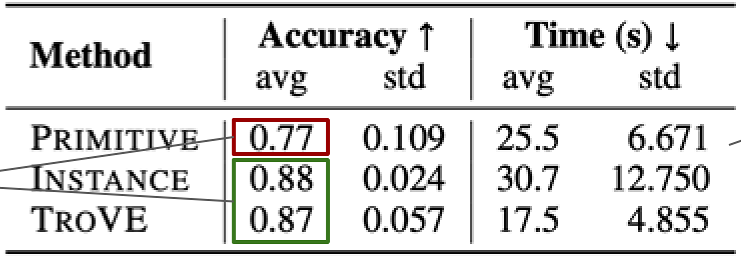
```python
from PIL import Image
from toolbox import crop_region, locate_objects

def get_object_region(
        image: Image.Image, object_name: str
) -> Image.Image:
    """Locate the crop the image of the object."""
    boxes = locate_objects(image, object_name)
    object_image = crop_region(image, boxes)
    return object_image
```

Varied functionalities across datasets



Accurate, Efficient Human Verification

10% more accurate



Method	Accuracy ↑		Time (s) ↓	
	avg	std	avg	std
PRIMITIVE	0.77	0.109	25.5	6.671
INSTANCE	0.88	0.024	30.7	12.750
TROVE	0.87	0.057	17.5	4.855

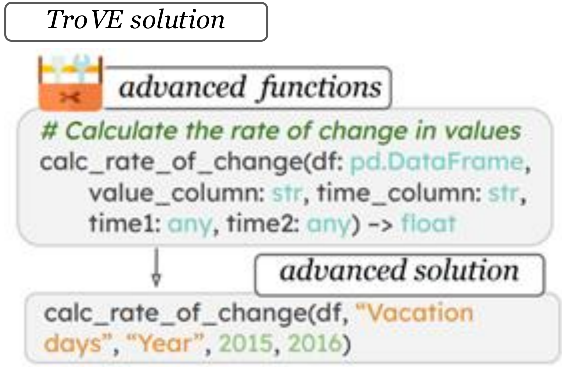
31.4 – 43.0% faster

Table 5. Human accuracy and time in verifying model-produced solutions with three methods experimented.

primitive solution

```
# get the row for each time stamp
row_2015 = df[df["Year"] == 2015
row_2016 = df[df["Year"] == 2016
# get the value for each time
value_2015 = row_2015["Vacation days"].values[0]
value_2016 = row_2015["Vacation days"].values[0]
# calculate the rate of change
rate = (value_2016 - value_2015) / 1
```

VS



TroVE solution

advanced functions

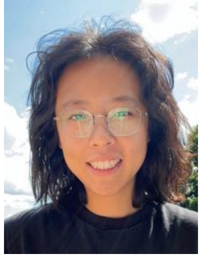
```
# Calculate the rate of change in values
calc_rate_of_change(df: pd.DataFrame,
    value_column: str, time_column: str,
    time1: any, time2: any) -> float
```

advanced solution

```
calc_rate_of_change(df, "Vacation
days", "Year", 2015, 2016)
```

TroVE: Takeaways

- Programmatic tools enable accurate, concise solutions
- Facilitates faster human verification
- Naturally adaptive to various tasks/domains



Shuyan
Zhou

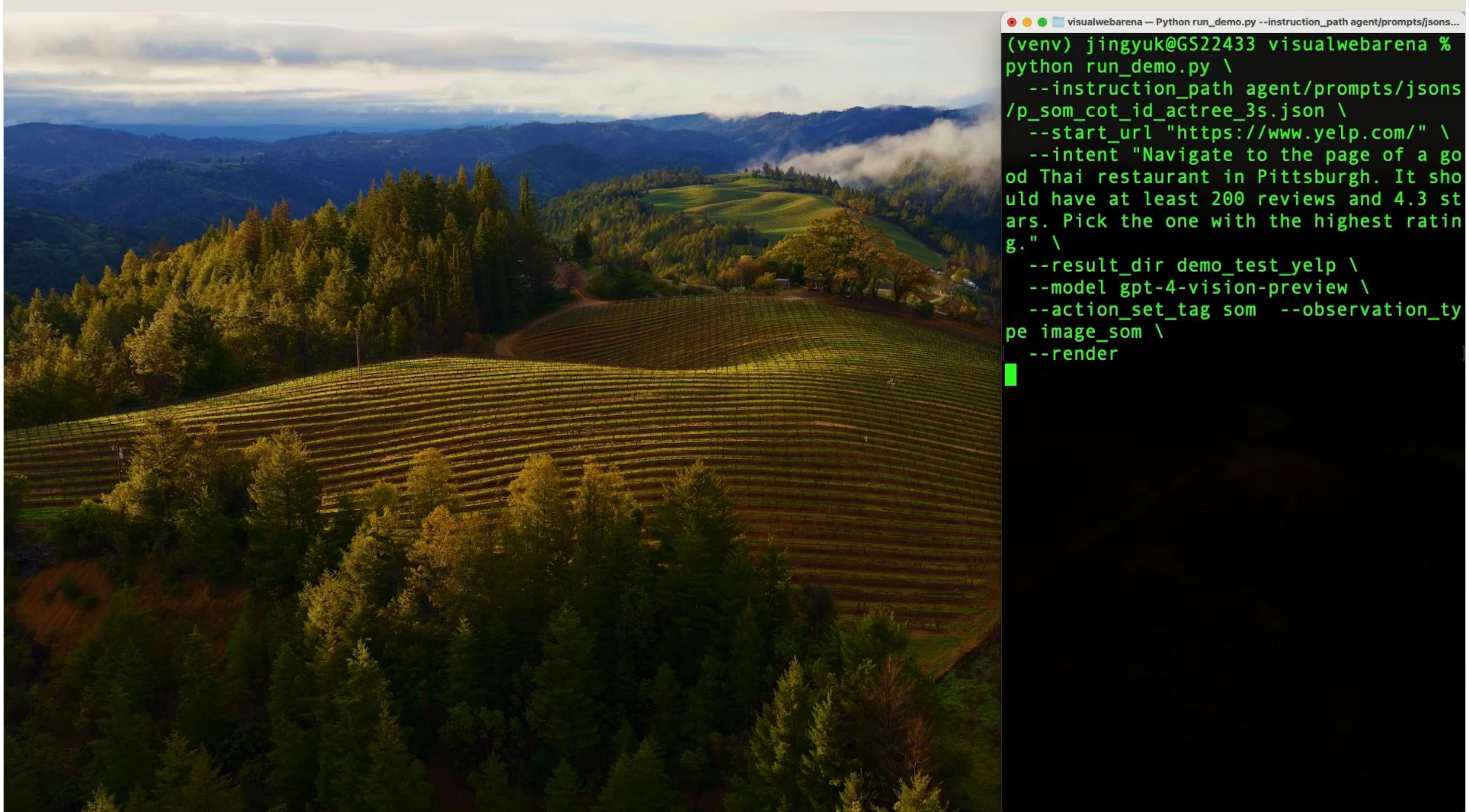
Frank Xu

Jing Yu
Koh

Acting in an Environment

Shuyan Zhou and Frank Xu et al. WebArena. ICLR 2024
Jing Yu Koh et al. VisualWebArena. ACL 2024

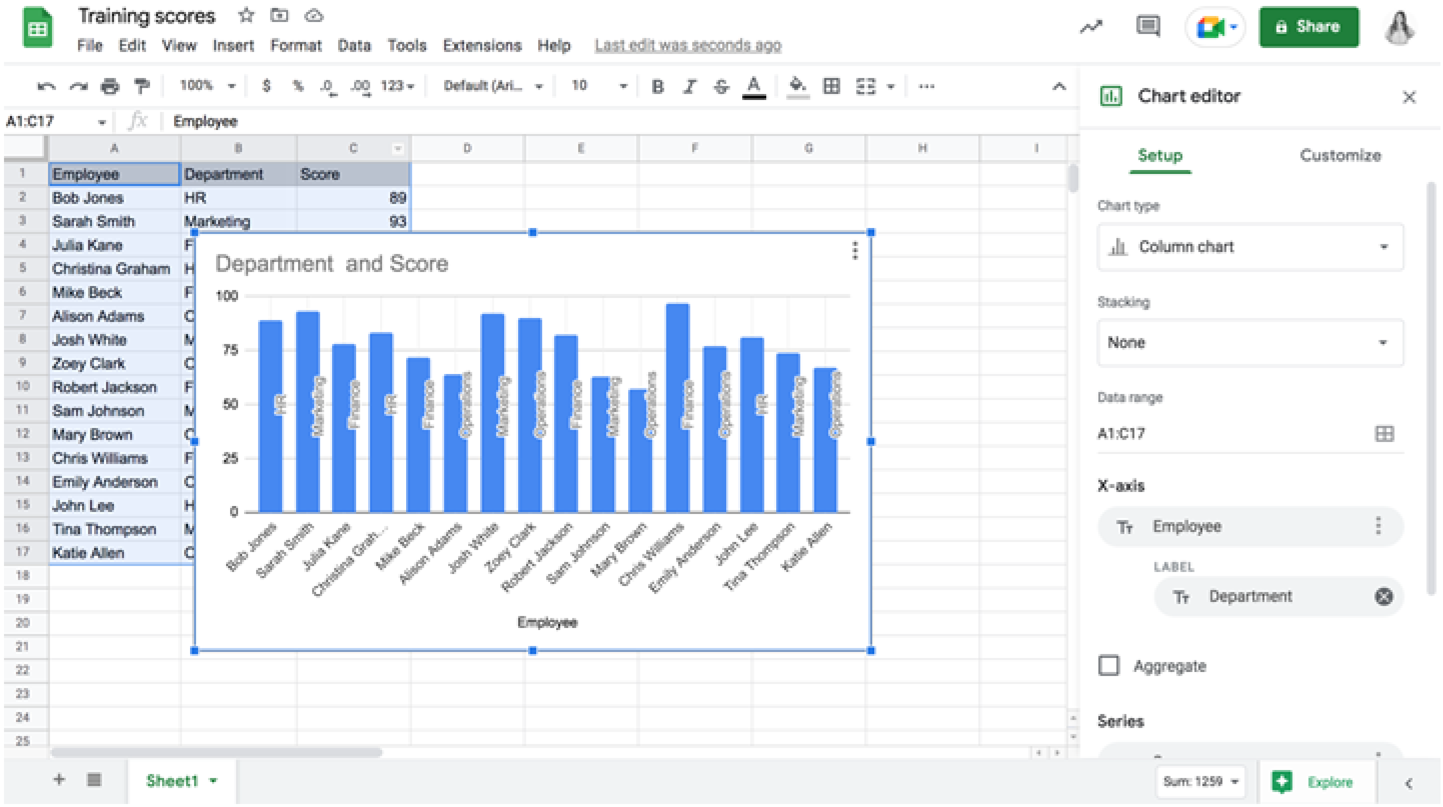
Task: Navigate to the page of a good Thai restaurant in Pittsburgh. It should have at least 200 reviews and 4.3 stars. Pick the one with the highest rating.



visualwebarena — Python run_demo.py --instruction_path agent/prompts/jsons...

```
(venv) jingyuk@GS22433 visualwebarena %
python run_demo.py \
  --instruction_path agent/prompts/jsons
/p_som_cot_id_actree_3s.json \
  --start_url "https://www.yelp.com/" \
  --intent "Navigate to the page of a go
od Thai restaurant in Pittsburgh. It sho
uld have at least 200 reviews and 4.3 st
ars. Pick the one with the highest ratin
g." \
  --result_dir demo_test_yelp \
  --model gpt-4-vision-preview \
  --action_set_tag som  --observation_ty
pe image_som \
  --render
```

Why Web Agents?



	A	B	C
1	Employee	Department	Score
2	Bob Jones	HR	89
3	Sarah Smith	Marketing	93
4	Julia Kane	F	
5	Christina Graham	H	
6	Mike Beck	F	
7	Alison Adams	C	
8	Josh White	M	
9	Zoey Clark	C	
10	Robert Jackson	F	
11	Sam Johnson	M	
12	Mary Brown	C	
13	Chris Williams	F	
14	Emily Anderson	C	
15	John Lee	H	
16	Tina Thompson	M	
17	Katie Allen	C	

Chart type
Column chart

Stacking
None

Data range
A1:C17

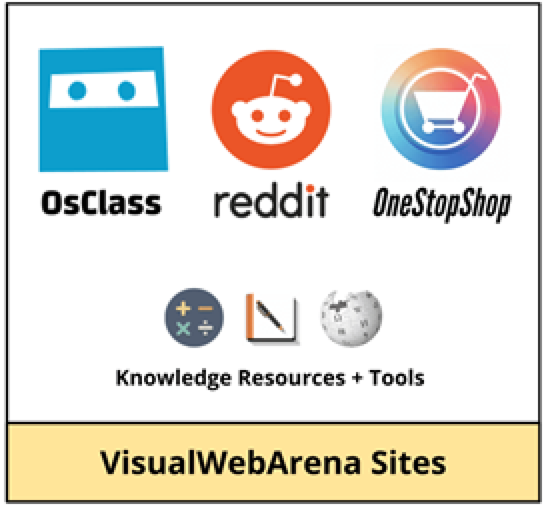
X-axis
Employee

LABEL
Department

☐ Aggregate

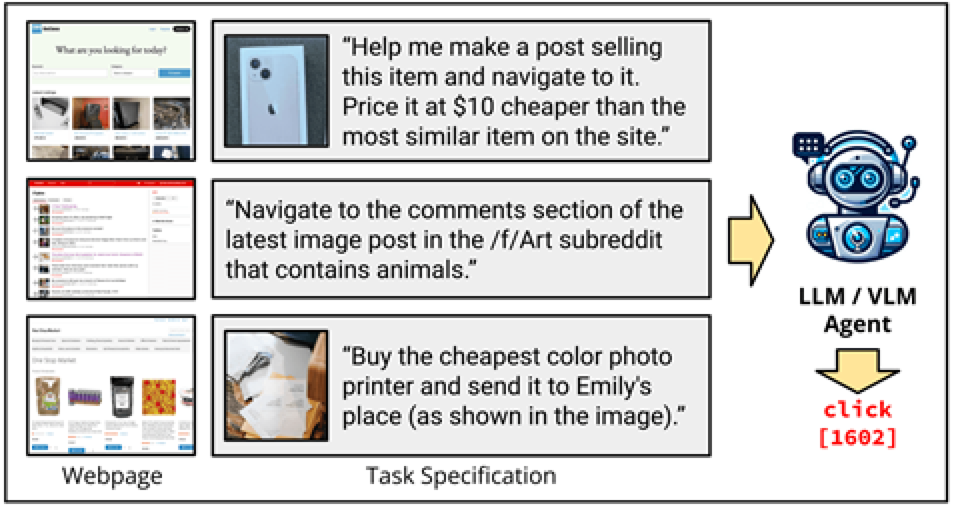
Series

Simulators with Real-World Sites



WebArena (Zhou*, Xu* et al., ICLR 2024)

Standalone, self-hostable web environments


"Help me make a post selling this item and navigate to it. Price it at $10 cheaper than the most similar item on the site."

"Navigate to the comments section of the latest image post in the /f/Art subreddit that contains animals."

"Buy the cheapest color photo printer and send it to Emily's place (as shown in the image)."

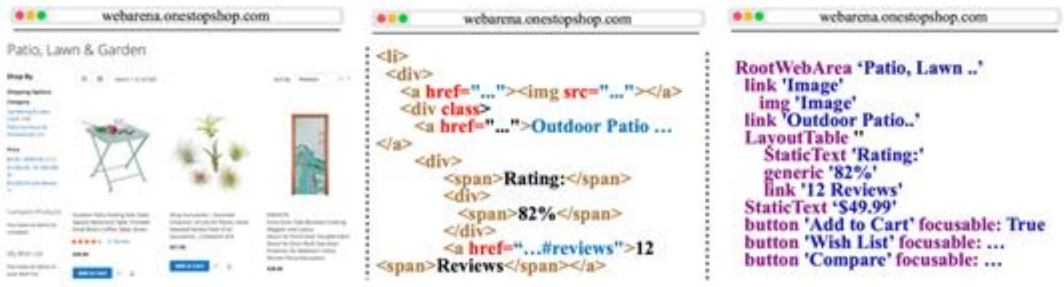
VisualWebArena (Koh et al., ACL 2024)

Benchmark for *multimodal* web agents

Simulators with Real-World Sites

POMDP environment: $\mathcal{E} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T} \rangle$

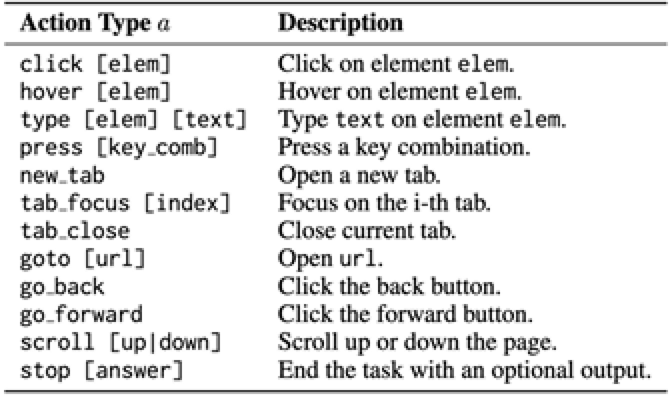
Observations $\mathcal{O}$



```
<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>
```

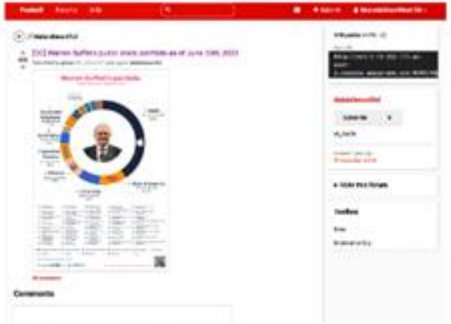
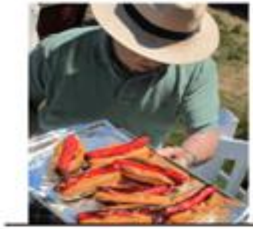
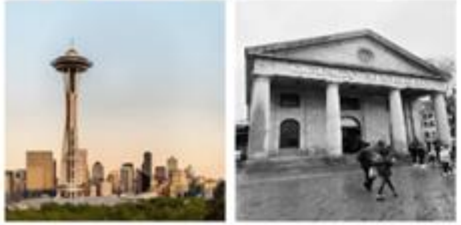
```
RootWebArea 'Patio, Lawn ..'
  link 'Image'
    img 'Image'
  link 'Outdoor Patio..'
  LayoutTable ''
     StaticText 'Rating:'
     generic '82%'
     link '12 Reviews'
  StaticText '$49.99'
  button 'Add to Cart' focusable: True
  button 'Wish List' focusable: ...
  button 'Compare' focusable: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

Actions $\mathcal{A}$

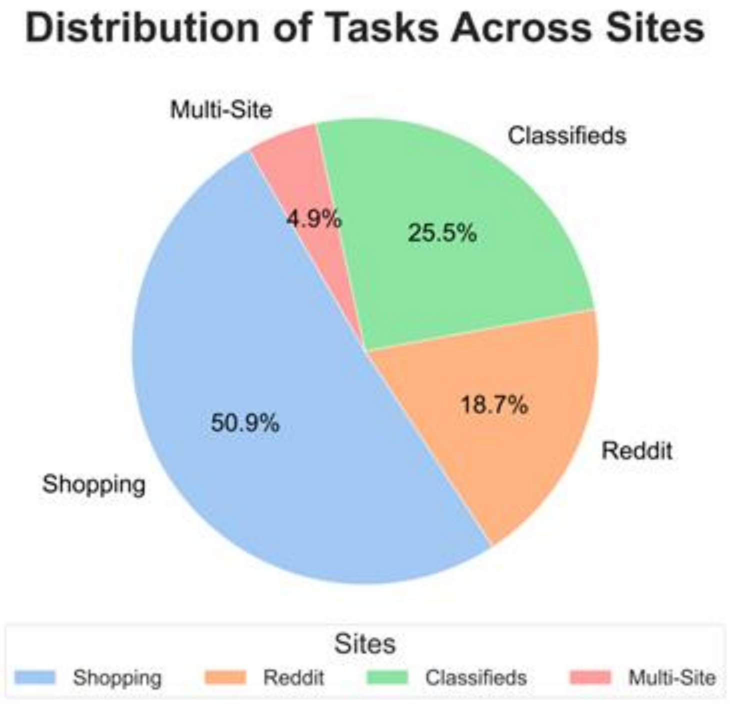


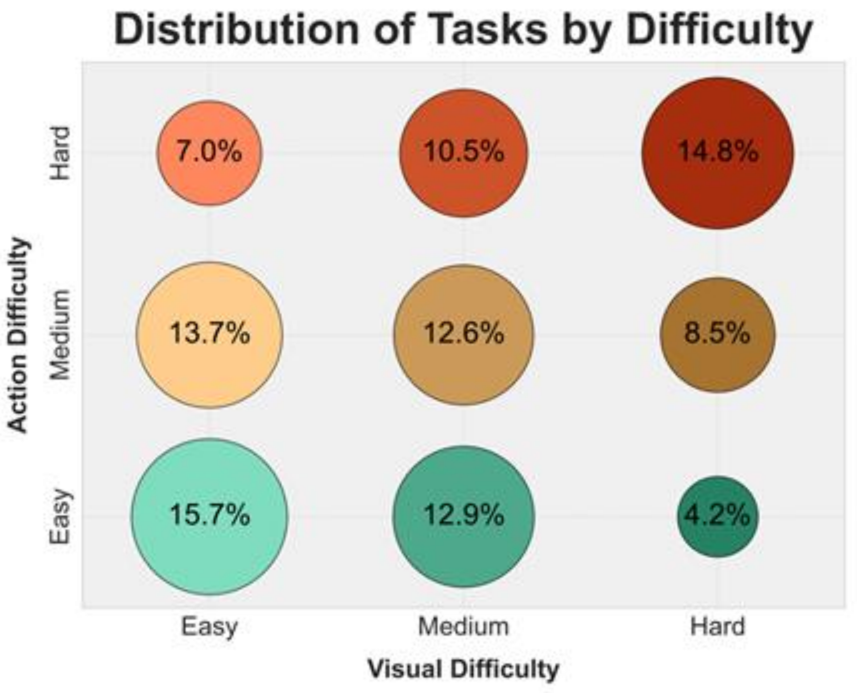
Action Type a	Description
`click [elem]`	Click on element `elem`.
`hover [elem]`	Hover on element `elem`.
`type [elem] [text]`	Type `text` on element `elem`.
`press [key_comb]`	Press a key combination.
`new_tab`	Open a new tab.
`tab_focus [index]`	Focus on the i-th tab.
`tab_close`	Close current tab.
`goto [url]`	Open `url`.
`go_back`	Click the back button.
`go_forward`	Click the forward button.
`scroll [up|down]`	Scroll up or down the page.
`stop [answer]`	End the task with an optional output.

Example Tasks

Webpage / Input Image(s)	Example Intent
	What is the ISIN of the company that occupies the largest portion in Warren Buffet's portfolio? Answer using the information from the Wikipedia site in the second tab.
	Add something like what the man is wearing to my wish list.
	Create a post for each of the following images in the most related forums.

VisualWebArena: Task Distribution


Distribution of Tasks Across Sites


Distribution of Tasks by Difficulty

Long-Horizon Web Tasks Are Hard

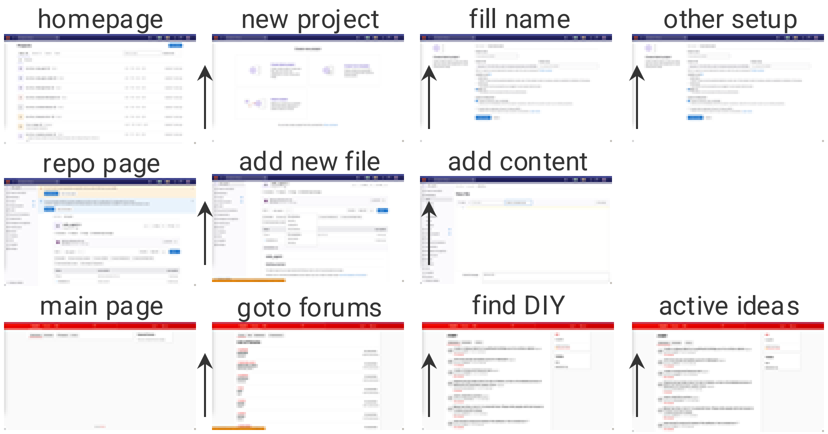
- Example Long-Horizon Task

`create a repository named Great_DIY_ideas`

`that includes a README file`

`with the links to the most active 6 DIY ideas on DIY subreddit?`

Each part requires many steps



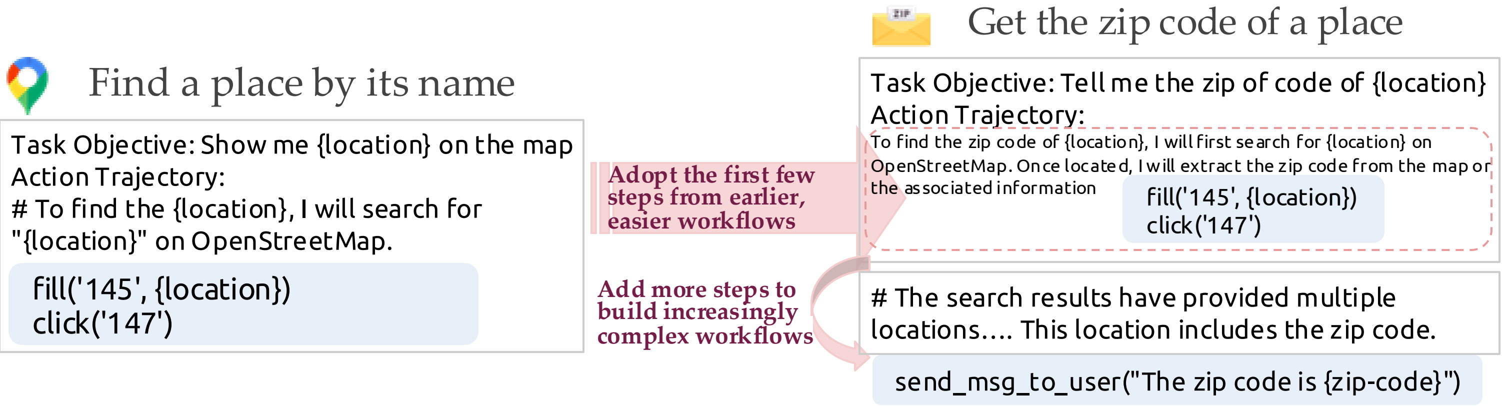
skills ↕ Know how to do these **atomic tasks** a priori

Zora Wang

Inducing Agentive Skills via Workflows and Tools

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. ”Agent Workflow Memory."
Zora Zhiruo Wang, Apurva Gandhi, Graham Neubig, and Daniel Fried. "Inducing Programmatic Skills for Agentic Tasks."

Skills: Reusable Subcomponents



Find a place by its name

Task Objective: Show me {location} on the map
Action Trajectory:
# To find the {location}, I will search for "{location}" on OpenStreetMap.

```
fill('145', {location})
click('147')
```

Adopt the first few steps from earlier, easier workflows

Add more steps to build increasingly complex workflows

Get the zip code of a place

Task Objective: Tell me the zip of code of {location}
Action Trajectory:
To find the zip code of {location}, I will first search for {location} on OpenStreetMap. Once located, I will extract the zip code from the map or the associated information

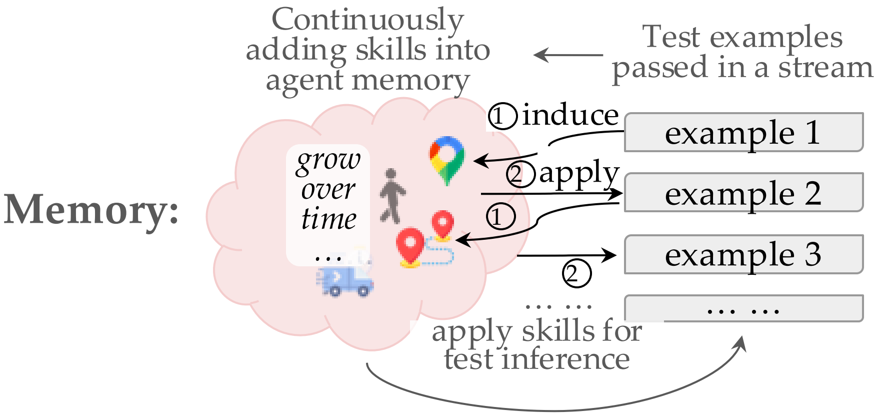
```
fill('145', {location})
click('147')
```

# The search results have provided multiple locations.... This location includes the zip code.

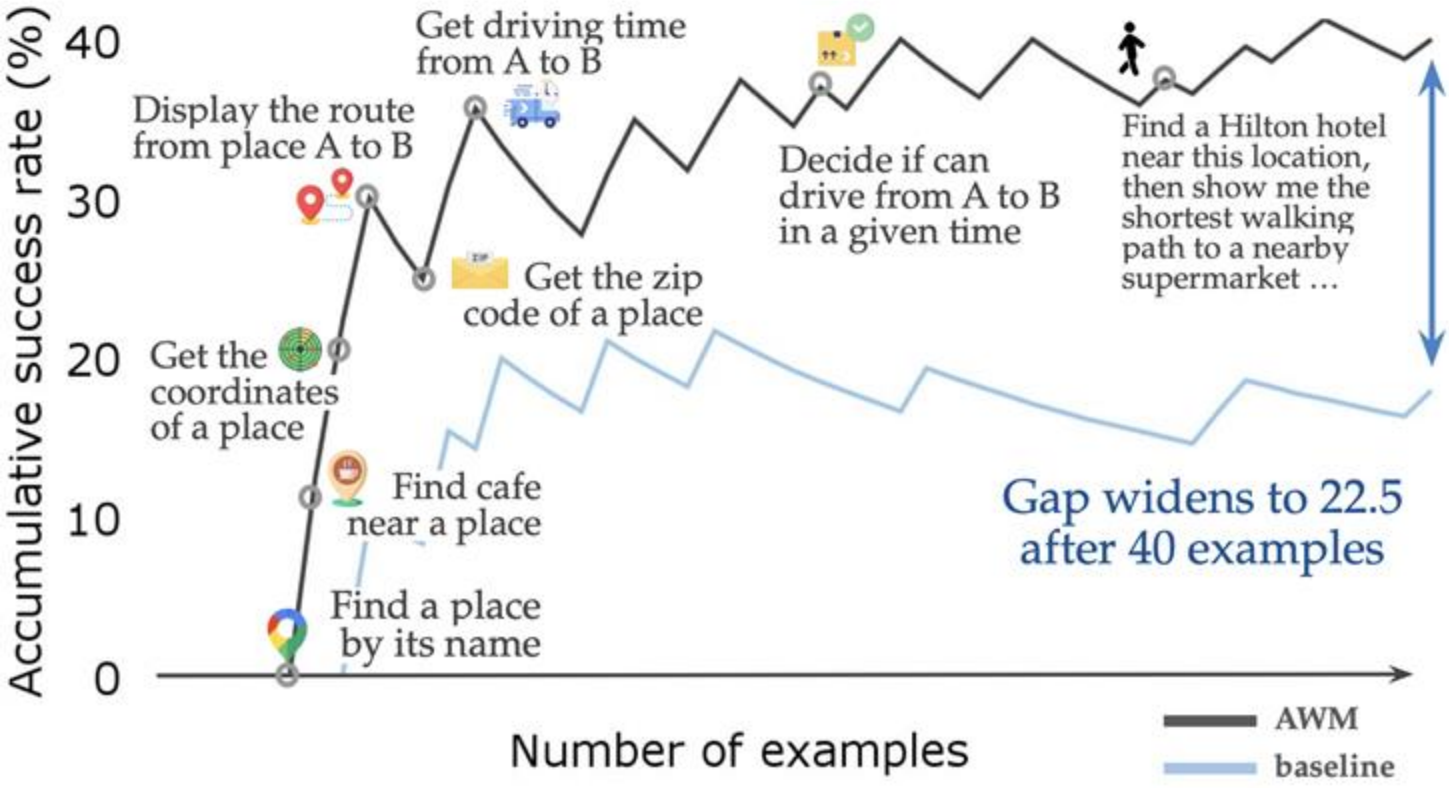
```
send_msg_to_user("The zip code is {zip-code}")
```

Learning Skills Online

- Like our previous work, maintain a memory of skills induced so far
- Induce skills from solved examples, add to the memory
- Retrieve from the memory to use in solving new examples

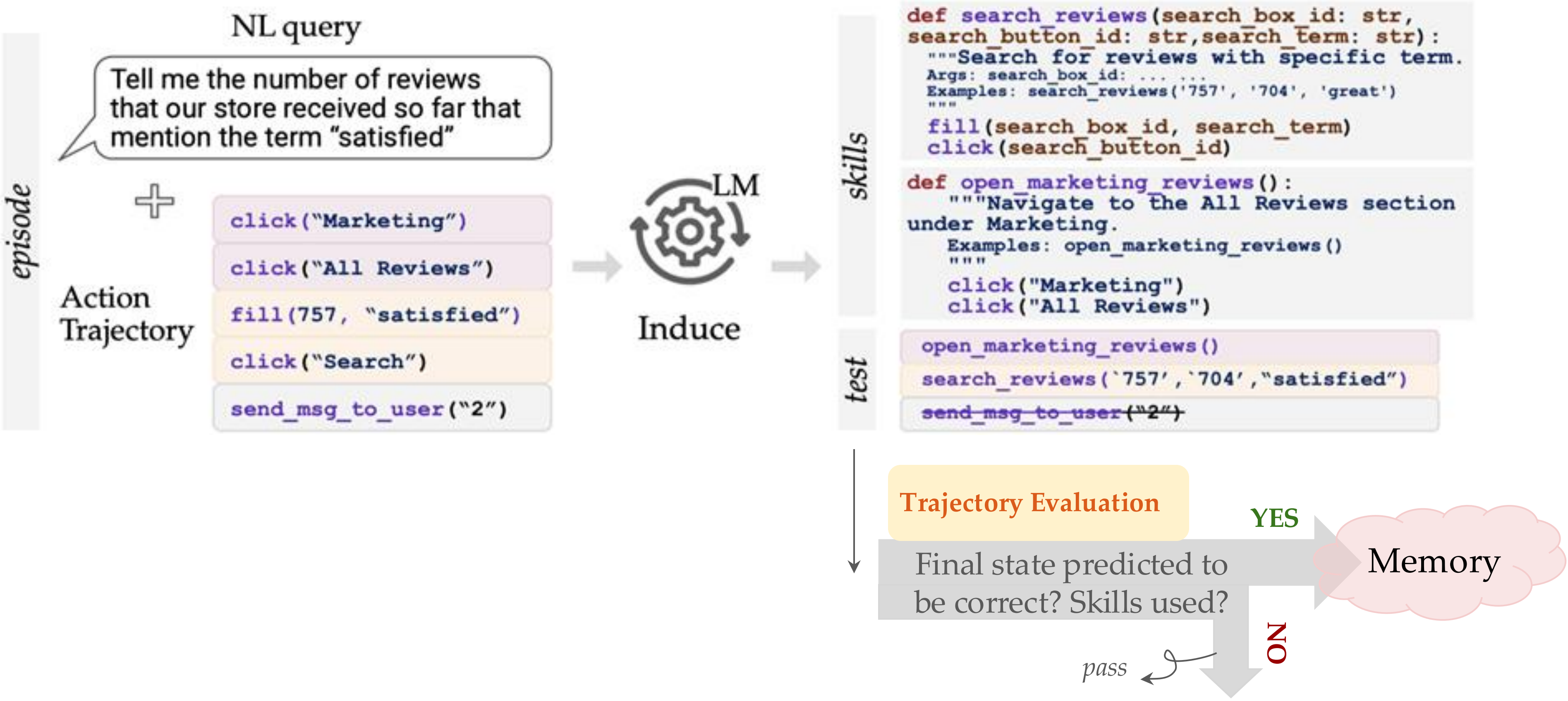

Memory:
Continuously adding skills into agent memory
Test examples passed in a stream
grow over time
…
① induce
② apply
①
②
example 1
example 2
example 3
… …
… …
apply skills for test inference

Online Learning of Skills Improves Performance



Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. ”Agent Workflow Memory."

Inducing and Verifying Programmatic Tools



Tell me the number of reviews that our store received so far that mention the term “satisfied”

```
click(“Marketing”)
click(“All Reviews”)
fill(757, “satisfied”)
click(“Search”)
send_msg_to_user(“2”)
```

```
def search_reviews(search_box_id: str,
search_button_id: str,search_term: str):
  """Search for reviews with specific term.
  Args: search_box_id: ... ...
  Examples: search_reviews('757', '704', 'great')
  """
  fill(search_box_id, search_term)
  click(search_button_id)
```

```
def open_marketing_reviews():
    """Navigate to the All Reviews section
under Marketing.
    Examples: open_marketing_reviews()
    """
    click("Marketing")
    click("All Reviews")
```

```
open_marketing_reviews()
search_reviews(`757′,`704′,“satisfied”)
```

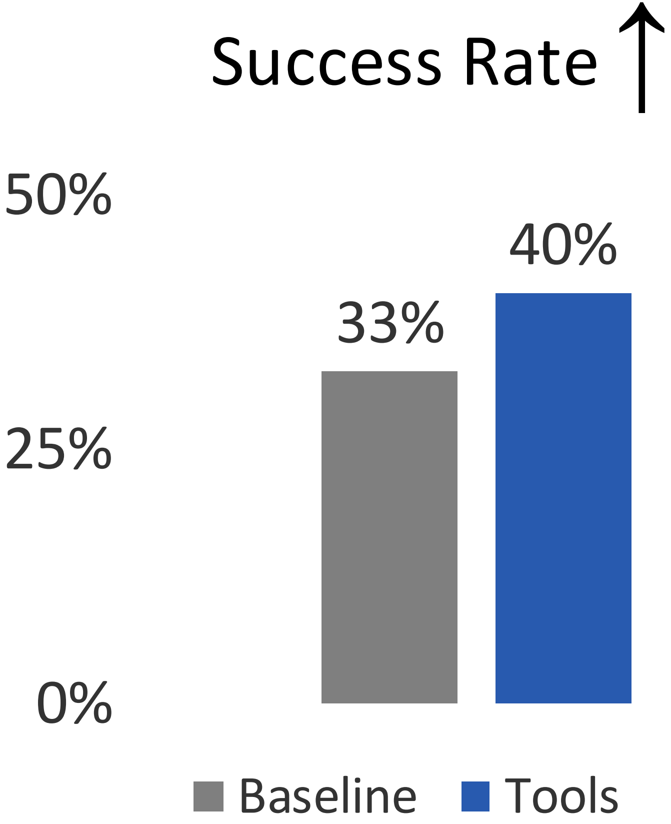
~~send_msg_to_user(“2”)~~

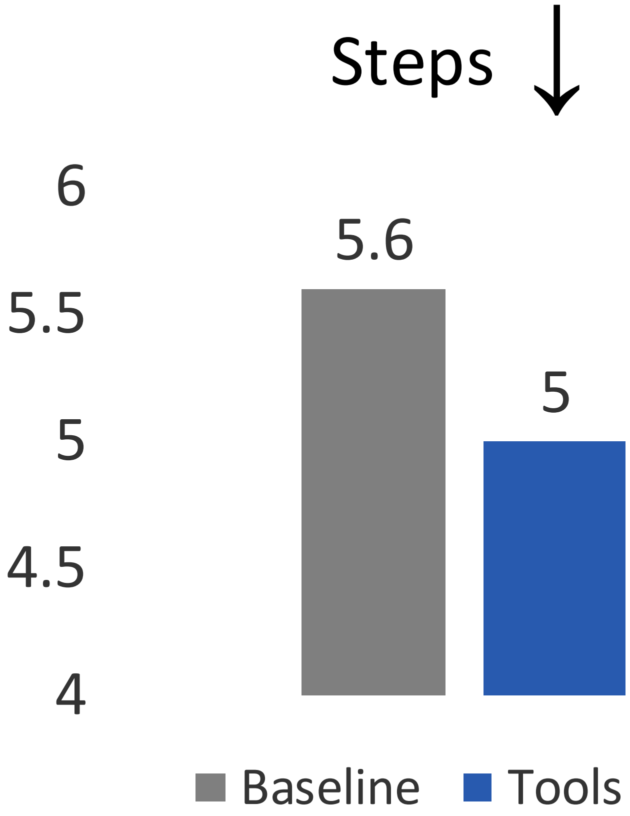
Trajectory Evaluation

Final state predicted to be correct? Skills used?

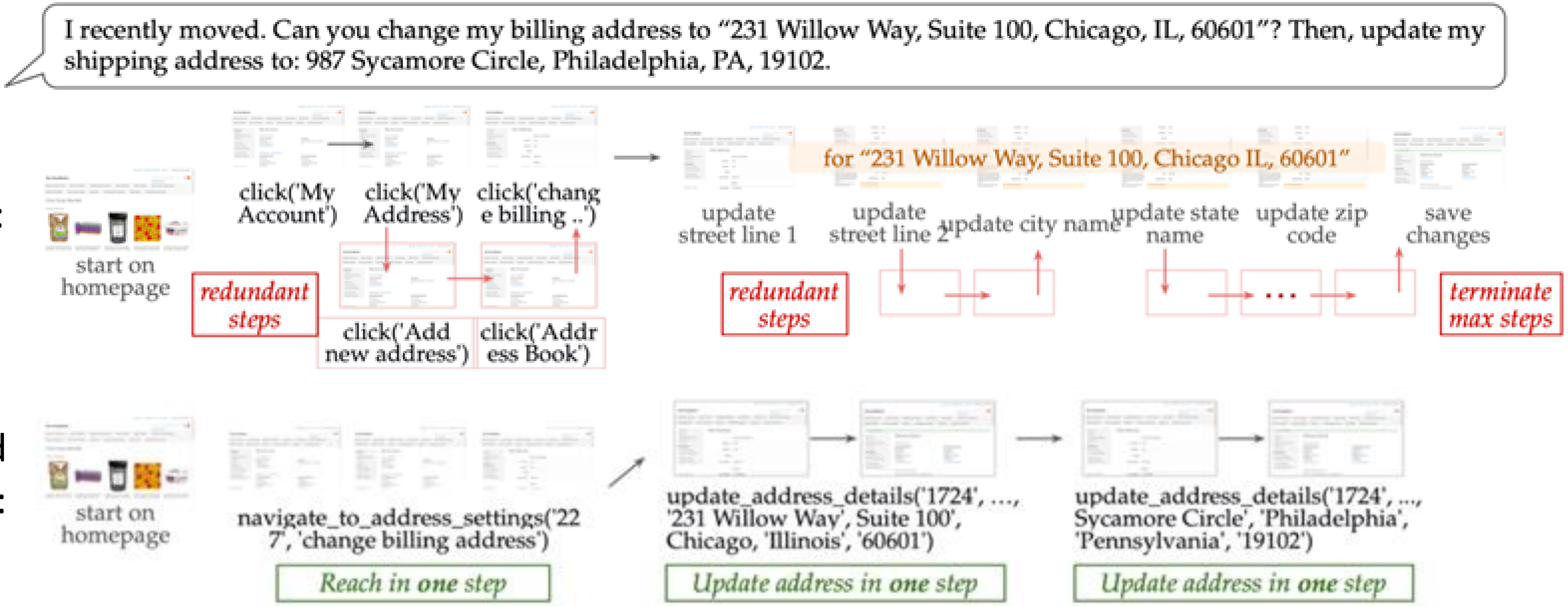
Zora Zhiruo Wang, Apurva Gandhi, Graham Neubig, and Daniel Fried. "Inducing Programmatic Skills for Agentic Tasks."

Results on WebArena





Scaled-Up Web Activities



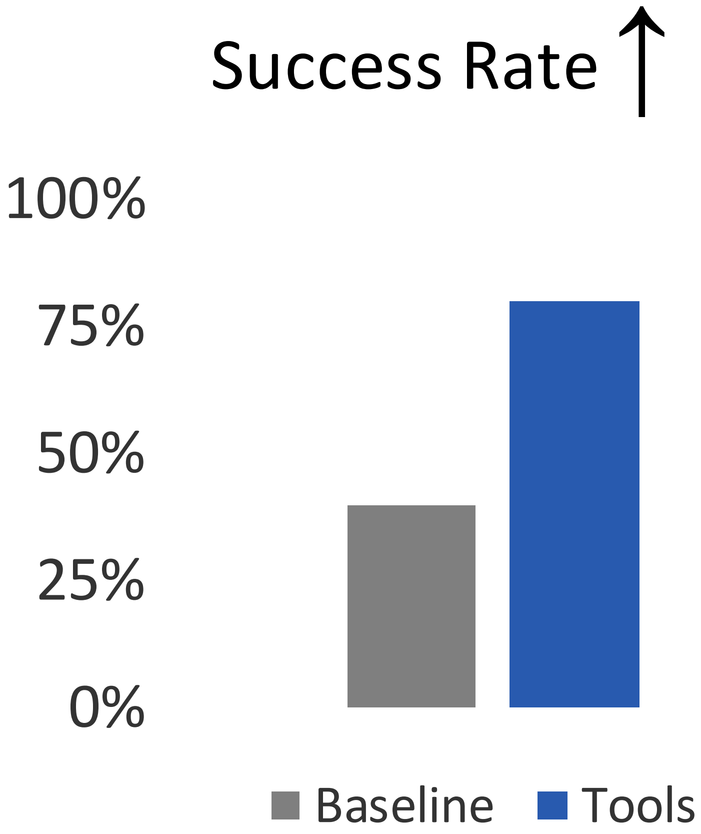
I recently moved. Can you change my billing address to "231 Willow Way, Suite 100, Chicago, IL, 60601"? Then, update my shipping address to: 987 Sycamore Circle, Philadelphia, PA, 19102.

start on homepage
click('My Account')
click('My Address')
click('change billing ..')
click('Add new address')
click('Address Book')
redundant steps
for "231 Willow Way, Suite 100, Chicago IL, 60601"
update street line 1
update street line 2
update city name
update state name
update zip code
save changes
redundant steps
. . .
terminate max steps
start on homepage
navigate_to_address_settings('227', 'change billing address')
Reach in one step
update_address_details('1724', ..., '231 Willow Way', Suite 100', Chicago, 'Illinois', '60601')
Update address in one step
update_address_details('1724', ..., Sycamore Circle', 'Philadelphia', 'Pennsylvania', '19102')
Update address in one step

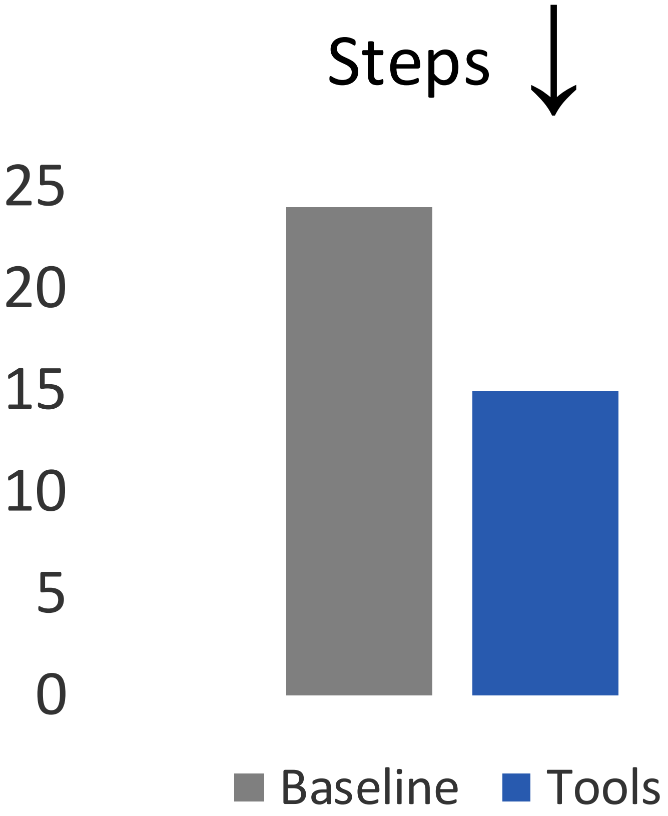
Baseline:

w/ Induced Tools:

Results on Scaled-Up Web Activities



Success Rate ↑



Steps ↓

Takeaways

- Tools / skills / functions can be induced to capture shared sub-structure
- Often more successful and more efficient!
- But how to learn over very long time horizons (e.g. years of interaction)?
- Toward human interaction and editing of learned skills

Thanks!

dfried@cs.cmu.edu
http://dpfried.github.io

TroVE: https://github.com/zorazrw/trove
Agent Skill Induction: https://github.com/zorazrw/agent-skill-induction