



Bringing Software Productivity to the Next Level through Generalist Agents

Hongxin Li*, Jingran Su*, Yuntao Chen†, Qing Li†, and Zhaoxiang Zhang†

* Equal contribution. †Equally advising corresponding authors.



Accepted to
NeurIPS 2023

Content



- 1 Background**
- 2 SheetCopilot**

1

Background

1 Background



Large language models have shown impressive capabilities:

- Abundant world knowledge
- General intent understanding
- Powerful multi-turn interaction
- Logical reasoning
- Program synthesis



```
BR If you sum up 1 to 50 but exclude all primes, what is the result?
```

```
To find the sum of numbers from 1 to 50 excluding primes, you can start by listing the prime numbers between 1 and 50:
```

```
2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47
```

```
Now, you can calculate the sum of all the numbers from 1 to 50 and then subtract the sum of the prime numbers:
```

```
Sum of all numbers from 1 to 50 = (50 * 51) / 2 = 1275
```

```
Sum of prime numbers = 2 + 3 + 5 + 7 + 11 + 13 + 17 + 19 + 23 + 29 + 31 + 37 + 41 + 43 + 47 = 328
```

```
Now, subtract the sum of prime numbers from the sum of all numbers:
```

```
1275 - 328 = 947
```

```
So, the result of summing up all numbers from 1 to 50 but excluding prime numbers is 947.
```

```
BR Write Bubble Sort code in Python
```

```
Certainly! Below is the Python code for the Bubble Sort algorithm:
```

```
python Copy code
```

```
def bubble_sort(arr):  
    n = len(arr)  
  
    # Traverse through all array elements  
    for i in range(n):  
        # Flag to check if any swapping occurred  
        swapped = False  
  
        # Last i elements are already in place  
        for j in range(0, n-i-1):  
  
            # Traverse the array from 0 to n-i-1  
            # Swap if the element found is greater than the next element  
            if arr[j] > arr[j+1]:  
                arr[j], arr[j+1] = arr[j+1], arr[j]  
                swapped = True  
  
        # If no two elements were swapped in the inner loop, then the list is sorted  
        if not swapped:  
            break  
  
    return arr
```

1 Background



Can we build a generalist agent using these LLM capabilities?

world knowledge

Intent understanding

Multi-turn interaction

Logical reasoning

Program synthesis

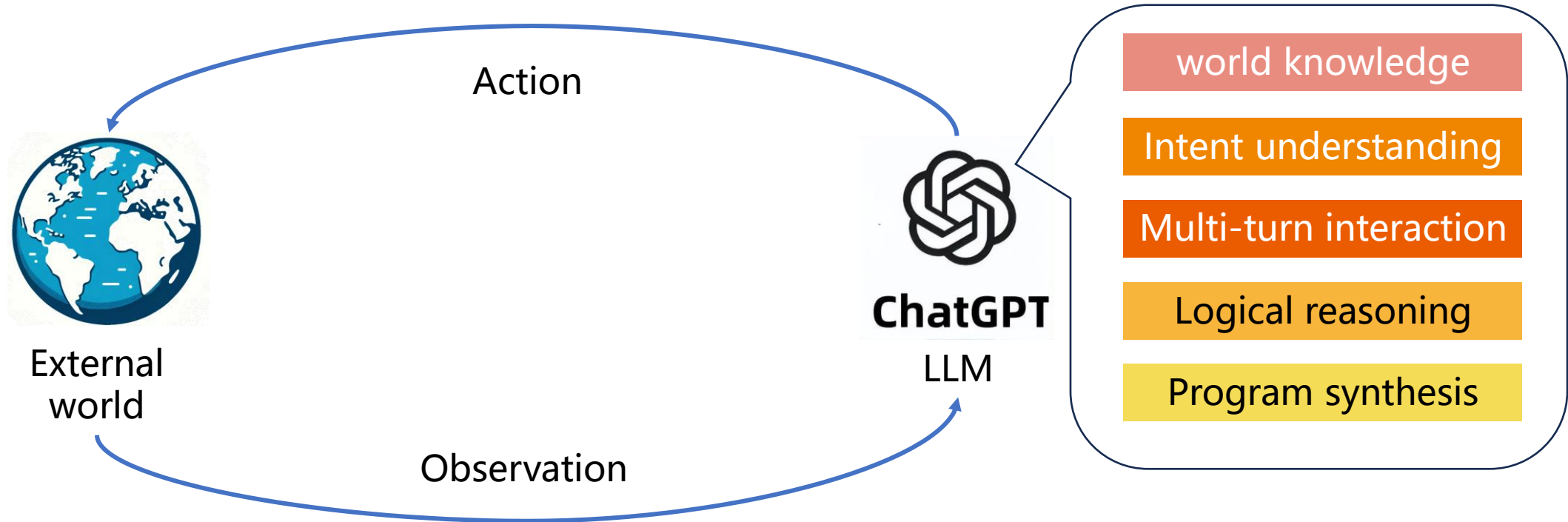


(Created with DALLE 3)

1 Background

Can we build a generalist agent using these LLM capabilities?

Yes! We can integrate **perception, reasoning, and planning** into one LLM and **augment it** with external tools.



2

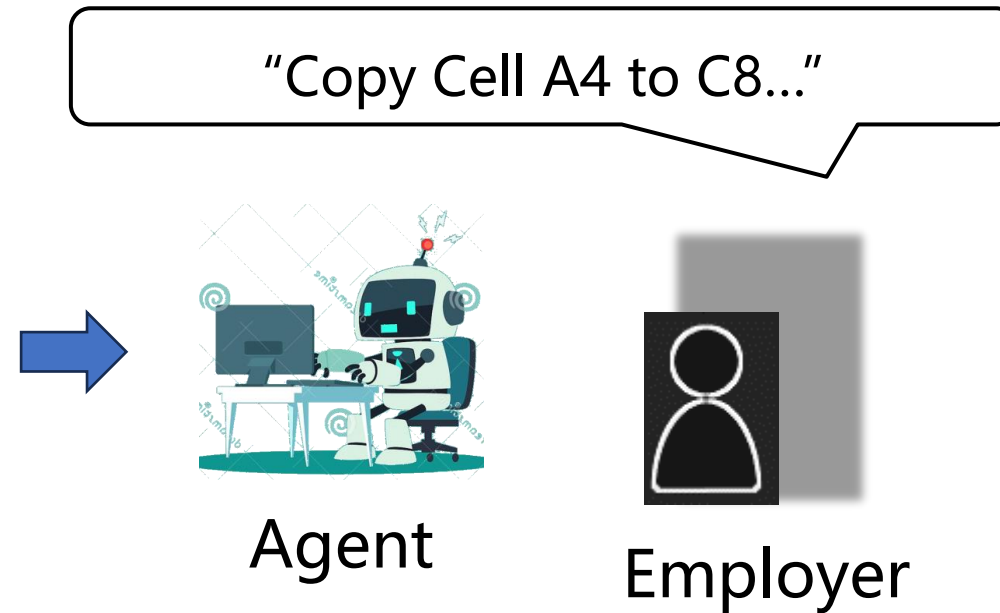
SheetCopilot

Motivation

What about building an agent to free us from boring routines?



Now



Future

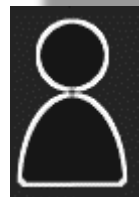
Motivation

What about building an agent to **free us from boring routines**?

"Copy Cell A4 to C8..."



Agent



Employer

Future

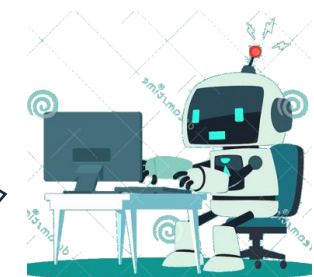
Some issues to address:

1. Can **average users use complex software** without acquiring expertise in advance?
2. How to build a generalist agent that **masters various software**?
3. How to **evaluate** such an agent?

Research content

Build a generalist agent that **follow instructions** to solve complex spreadsheet tasks

1	Store Number	Store Type	Manager Name	State	Operating Costs	Total Sales	Best Selling Item	Profit	Profit Margin
2	188	Superstore	fenton	Idaho	\$32,000.00	\$42,000.00	chicken nuggets	10000	0.2380952381
3	100	Free Standing	kavanaugh	Florida	\$24,800.00	\$37,000.00	double cheeseburger	12200	0.3297297297
4	148	Strip Mall	nolan	Indiana	\$32,000.00	\$31,000.00	crispy chicken sandwich	-1000	-0.03225806452
5	317	Strip Mall	cantrell	Massachusetts	\$40,999.00	\$47,000.00	malted milkshake	6001	0.1276808511
6	241	Superstore	woods	Florida	\$19,000.00	\$28,000.00	chicken nuggets	9000	0.3214285714
7	391	Free Standing	kenny	Idaho	\$42,100.00	\$55,000.00	malted milkshake	12900	0.323943662
8	221	Free Standing	walter	Massachusetts	\$24,000.00	\$35,500.00	crispy chicken sandwich	11500	=H8/F8
9	187	Strip Mall	coleman	Indiana	\$32,100.00	\$39,000.00	malted milkshake	6900	0.179230769
10	215	Superstore	douglas	Idaho	\$42,000.00	\$44,000.00	double cheeseburger	2000	0.04545454545
11	217	Free Standing	keels	Massachusetts	\$18,500.00	\$28,700.00	chicken nuggets	10200	0.3554006969
12	322	Free Standing	bentley	Indiana	\$21,300.00	\$38,000.00	malted milkshake	16700	0.4394736842
13	142	Strip Mall	decker	Florida	\$32,000.00	\$42,100.00	crispy chicken sandwich	10100	0.2399049881
14	193	Superstore	dixon	indiana	\$24,500.00	\$29,000.00	double cheeseburger	4500	0.1551724138
15	144	Free Standing	dunkley	Massachusetts	\$33,000.00	\$37,000.00	chicken nuggets	4000	0.1081081081
16	397	Strip Mall	davis	Idaho	\$42,100.00	\$49,000.00	double cheeseburger	6900	0.1408163265



SheetCopilot

Methodology

- 1. LLM planning:** Prompt a **large language model** (GPT-4) to generate efficient multi-step solutions.
- 2. Closed-loop control:** Provide **software state feedback** for the LLM to revise solutions.
- 3. Knowledge retrieval:** Utilize **external knowledge base** to optimize solutions.

Step 2. Calculate the revenue by multiplying the price by the quantity.
1. Write(range="G2", value="VLOOKUP(C2, 'A:B,2, FALSE)*E2")
2. AutoFill(source="G2", destination="G2:G36")

Instruction: ...



User

Execution



Feedback



SheetCopilot

	A	B	C	D	E	F
1	Date Time	Web Site	Product	Type	Quantity	Discount
2	9/8/15 10:13	amazon.c	Aspen	Wholesale	33	0.165
3	12/11/15 23:26	ebay.com	Yanaki	Wholesale	29	0.15
4	12/19/15 18:10	amazon.c	Bellen	Retail	3	0.019
5	8/23/15 12:55	ebay.com	Quad	Wholesale	36	0.15
6	7/9/15 05:02	coloradot	Crested B	Wholesale	38	0.15
7	6/30/15 16:40	ebay.com	Manu LD	Retail	4	0
8	7/1/15 13:16	coloradot	Bellen	Retail	4	0
9	8/2/15 04:06	gel-boon	Fire Asper	Wholesale	93	0.375

Spreadsheet software

Sheet state: "Sheet1" has 7 columns (Headers are A: "Date Time" ...) and 36 rows. "Retail Price" has 2 columns (Headers are A: "Product", B: "Retail Price") and 23 rows.
Error info: ...

Document

Write:
args: "(range: str, value: str)"
args explanation:
usage: Write value into a range.
example: ...
CreateChart: ...
Sort: ...
Filter: ...

Knowledge base

Methodology

1. LLM planning: Prompt an LLM (GPT-4) to generate efficient multi-step solutions.

① You are a **spreadsheet agent**

② Here is the **API document**:

Write # Args: (range: str, value: str)

CopyPaste # Args: (source: str, destination: str)

...

③ Requirements:

1. The user will show you the headers and row numbers of sheets.

2. Please provide **step-by-step solutions with explanations**.

...

④ An exemplar

⑤ Workbook context: ...

⑥ Instruction: I want to **calculate the revenue** for each transaction

⑦ Initial sheet state: ...

Chain-of-Thought + In-context learning

Methodology

2. **Closed-loop control:** Provide **software state feedback** to revise solutions.

① Sheet state description

Spread
sheet

	A	B	C	D	E	F
1	Date Time	Web Site	Product	Type	Quantity	Discount
2	9/8/15 10:13	amazon.c	Aspen	Wholesale	33	0.165
3	12/11/15 23:26	ebay.com	Yanaki	Wholesale	29	0.15
4	12/19/15 18:10	amazon.c	Bellen	Retail	3	0.019
5	8/23/15 12:55	ebay.com	Quad	Wholesale	36	0.15
6	7/9/15 05:02	colorado	Crested B	Wholesale	38	0.15
7	6/30/15 16:40	ebay.com	Manu LD	Retail	4	0
8	7/1/15 13:16	colorado	Bellen	Retail	4	0
9	8/2/15 04:06	gel-boom	Fire Asper	Wholesale	93	0.375
10	11/16/15 12:15	ebay.com	Sunnet	Wholesale	64	0.275

	A	B
1	Product	Retail Price
2	Alpine	21.95
3	Aspen	21.95
4	Carlota	19.95
5	Crested Beaut	17.95
6	Majestic Beaut	26.95
7	Sunshine	20
8	Bellen	25
9	Eagle	21.95
10	Yanaki	23.95
11		22.95

Sheet state:

Sheet **"Sheet1"** has 6 columns (Headers are ...) and 36 rows

Sheet **"Retail Price"** has 2 columns (Headers are ...) and 23 rows



Evaluation Environment

Methodology

2. **Closed-loop control:** Provide **software state feedback** to revise solutions.

② Error feedback



SheetCopilot

Step 2. Calculate the revenue for each transaction.
`Write(range="Sheet1!G2", value="=VLOOKUP(C2, "Retail Price"!A:B,2, FALSE)*E2*(1-F2)")`
`AutoFill(source="Sheet1!G2", destination="G2:G36 ")` ←

Wrong
action

Execution error: **Failed to execute**
`AutoFill(source="Sheet1!G2", destination="G2:G36").`
Error: The **range must contain a sheet name.**



Evaluation
Environment



SheetCopilot

Step 2. Calculate the revenue for each transaction.
`Write(range="Sheet1!G2", value="=E2*VLOOKUP(C2,"Retail Price"!A:B,2,FALSE)*(1-F2)")`
`AutoFill(source="Sheet1!G2", destination="Sheet1!G2:G36 ")` ←

Revision

Methodology

3. **Knowledge retrieval:** Utilize **external knowledge base** to optimize solutions.



SheetCopilot

Step 1. Add a new column named "Revenue".
InsertColumn(sheetName="Sheet1", after="F ")
Write(range="Sheet1!G1", value="Revenue")

Wrong
action

Document

InsertColumn:

args: "(sheetName: str, beforeColumn: str = None, afterColumn: str = None)"

args explanation: ...

example:

Example 1: Create a new column at Column G in Sheet1, i.e., after Column F...



Evaluation
Environment



SheetCopilot

Step 1. Add a new column named "Revenue".
InsertColumn(sheetName="Sheet1", afterColumn="F ")
Write(range="Sheet1!G1", value="Revenue")

Revision

Experiments

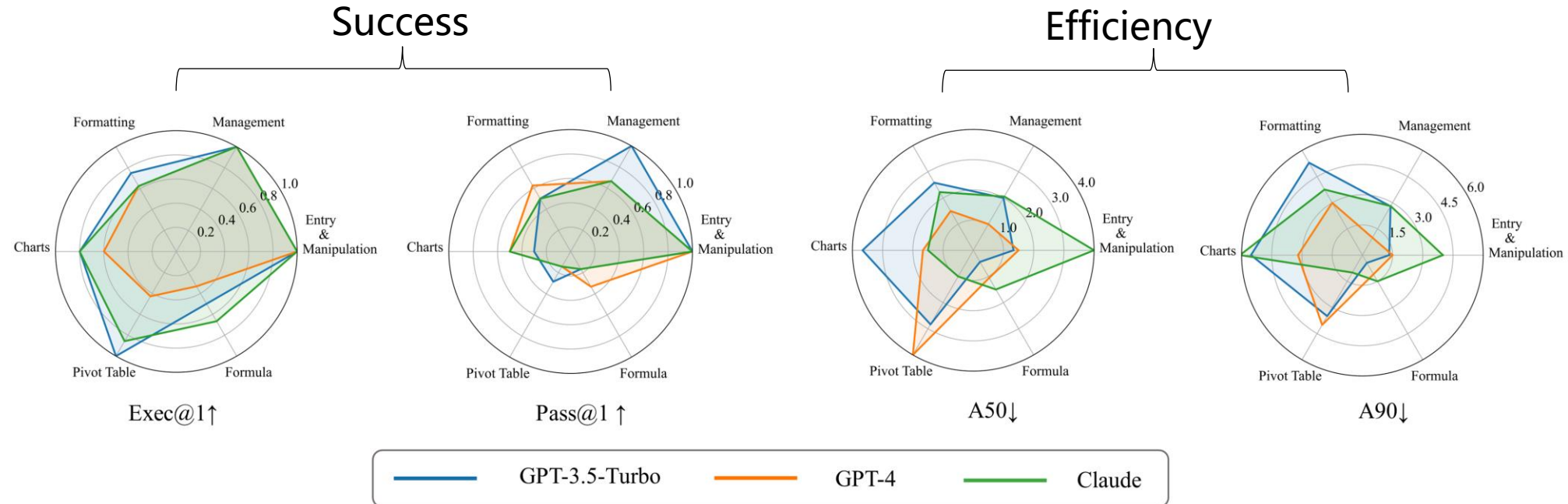
Analysis

- SheetCopilot (GPT-4) demonstrates **the strongest planning capability**
- SheetCopilot (GPT-3.5-Turbo) **outperforms the VBA-based method** (Exec@1 + 7.2% and Pass@1 + 7.9%)

Data	Models	Success		Efficiency	
		Exec@1↑	Pass@1↑	A50↓	A90↓
10%	GPT-3.5-Turbo	85.0%	45.0%	2.00	4.50
10%	GPT-4	65.0%	55.0%	1.33	2.00
10%	Claude	80.0%	40.0%	1.50	4.40
100%	GPT-3.5-Turbo	87.3%	44.3%	1.50	3.00
100%	VBA	77.8%	37.1%	-	-

Exec@1↑: Execution success rate
 Pass@1 ↑ : Functional correctness
 A50/A90↓: Median/90th percentile action number

Experiments



Analysis

1. The 2 GPT models both achieve **100% success** in the Management and Entry & manipulation tasks
2. It is **difficult** for the 3 models to excel in all task categories.

Exec@1↑: Execution success rate
 Pass@1 ↑ : Functional correctness
 A50/A90↓: Median/90th percentile action number

Ablation study on atomic action names

The knowledge in the LLMs may be confused with our API definitions, so we use the synonyms **far away from** the official names:

Write → RangeInputValue

SetConditionalFormat → FormatWithRules

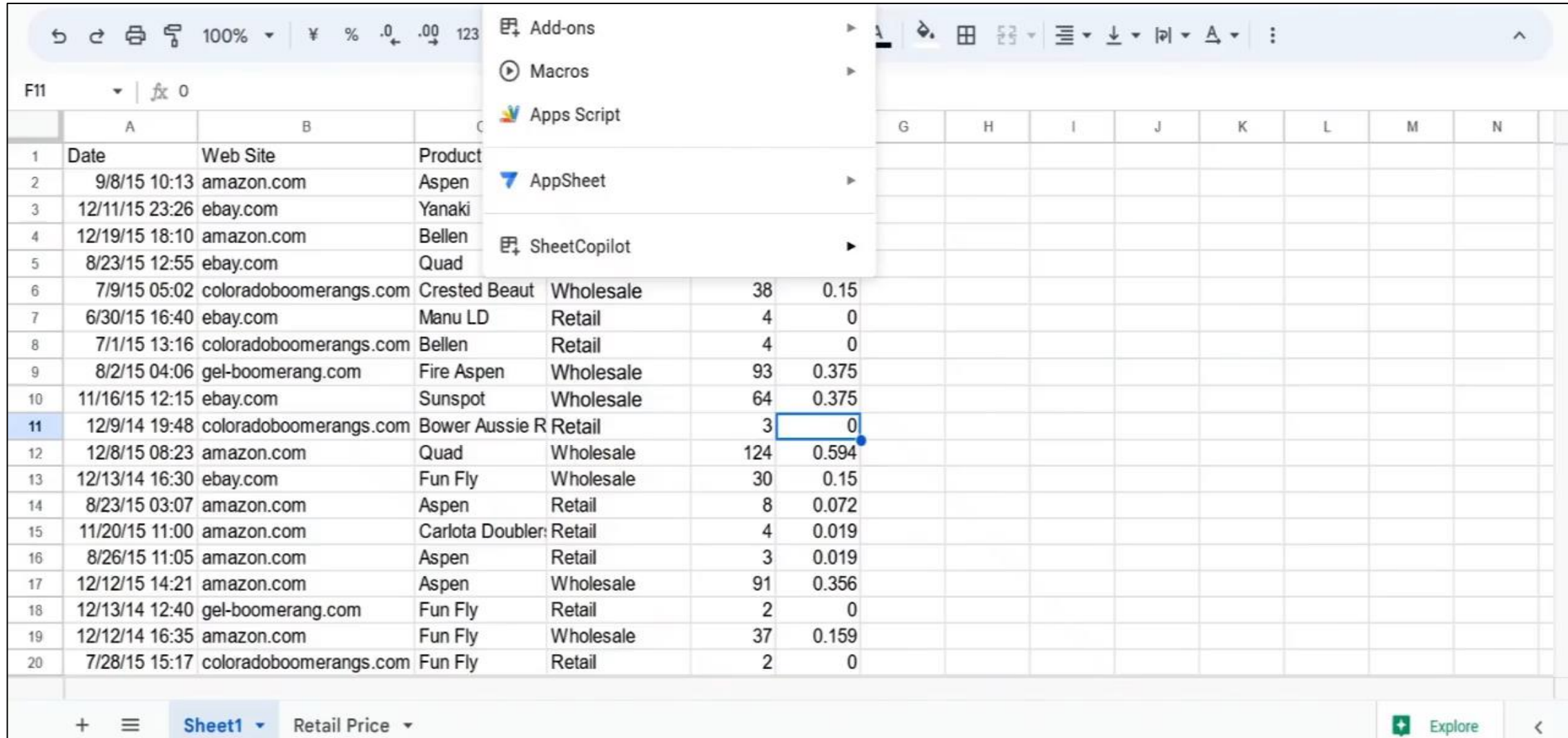
An interesting insight

- **Better** Pass@1 and efficiency.
- This result demonstrates **the flexibility of our method**: users can define their own atomic actions and prompt LLMs to use them.

Models	Exec@1↑	Pass@1↑	A50↓	A90↓
Official names	87.3%	44.3%	1.50	3.00
Synonyms	86.9%	46.2%	1.33	2.78

Exec@1↑: Execution success rate
Pass@1 ↑ : Functional correctness
A50/A90↓: Median/90th percentile action number

Interactive demo on Google Sheets



The screenshot displays a Google Sheets interface with a data table and the SheetCopilot add-on menu open. The table contains 20 rows of data with columns for Date, Web Site, Product, and numerical values. The SheetCopilot menu is visible, listing options like Add-ons, Macros, Apps Script, AppSheet, and SheetCopilot.

	A	B	C			
1	Date	Web Site	Product			
2	9/8/15 10:13	amazon.com	Aspen			
3	12/11/15 23:26	ebay.com	Yanaki			
4	12/19/15 18:10	amazon.com	Bellen			
5	8/23/15 12:55	ebay.com	Quad			
6	7/9/15 05:02	coloradoboomerangs.com	Crested Beaut	Wholesale	38	0.15
7	6/30/15 16:40	ebay.com	Manu LD	Retail	4	0
8	7/1/15 13:16	coloradoboomerangs.com	Bellen	Retail	4	0
9	8/2/15 04:06	gel-boomerang.com	Fire Aspen	Wholesale	93	0.375
10	11/16/15 12:15	ebay.com	Sunspot	Wholesale	64	0.375
11	12/9/14 19:48	coloradoboomerangs.com	Bower Aussie R	Retail	3	0
12	12/8/15 08:23	amazon.com	Quad	Wholesale	124	0.594
13	12/13/14 16:30	ebay.com	Fun Fly	Wholesale	30	0.15
14	8/23/15 03:07	amazon.com	Aspen	Retail	8	0.072
15	11/20/15 11:00	amazon.com	Carlota Doubler	Retail	4	0.019
16	8/26/15 11:05	amazon.com	Aspen	Retail	3	0.019
17	12/12/15 14:21	amazon.com	Aspen	Wholesale	91	0.356
18	12/13/14 12:40	gel-boomerang.com	Fun Fly	Retail	2	0
19	12/12/14 16:35	amazon.com	Fun Fly	Wholesale	37	0.159
20	7/28/15 15:17	coloradoboomerangs.com	Fun Fly	Retail	2	0

See more on our website: <https://sheetcopilot.github.io/>

Summary

1. We build a software agent by integrating perception, reasoning , and planning into large language models
2. We provide a comprehensive benchmark for testing spreadsheet agents
3. We create add-ons for **Excel and Google Sheets**
4. SheetCopilot improves office productivity and brings smooth interaction experience to users.

Out paper: <http://arxiv.org/abs/2305.19308>

Our website: <https://sheetcopilot.github.io/>