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.

NEURAL INFORMATION PROCESSING SYSTEMS

Accepted to NeurIPS 2023

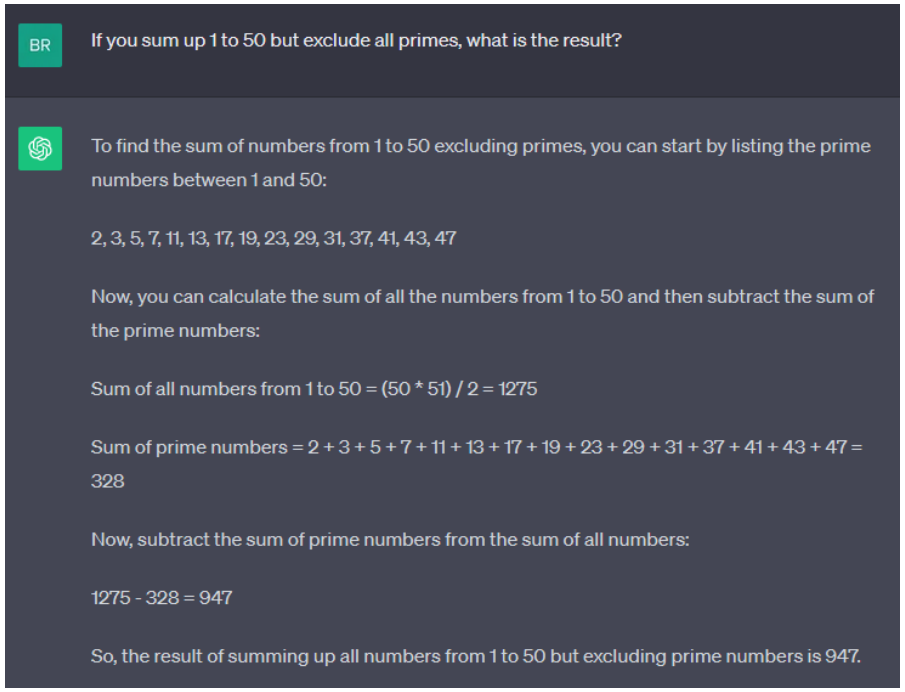
Content

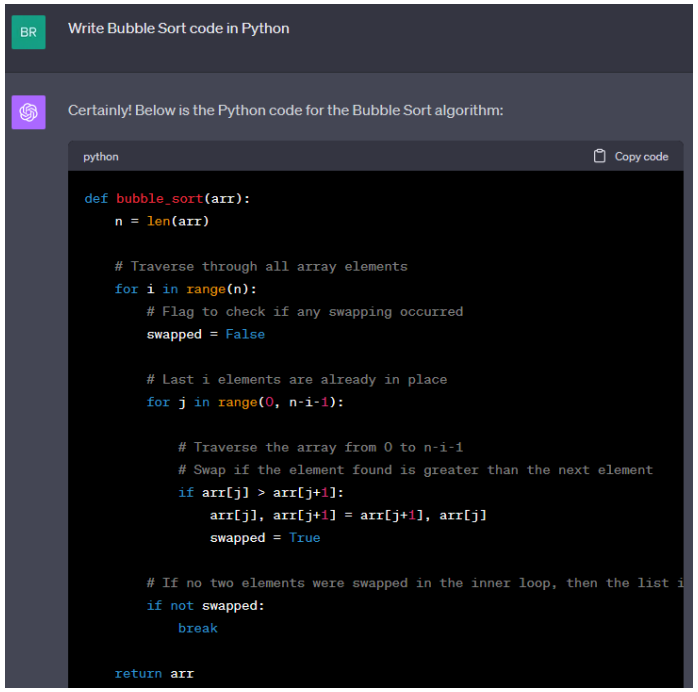
1 Background

1 Background

Large language models have shown impressive capabilities:

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





1 Background

Can we build a generalist agent using these LLM capabilities?

world knowledge

Intent understanding

Multi-turn interaction

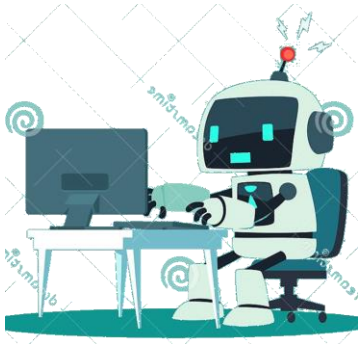
Logical reasoning

Program synthesis

ChatGPT

LLaMA

PaLM 2

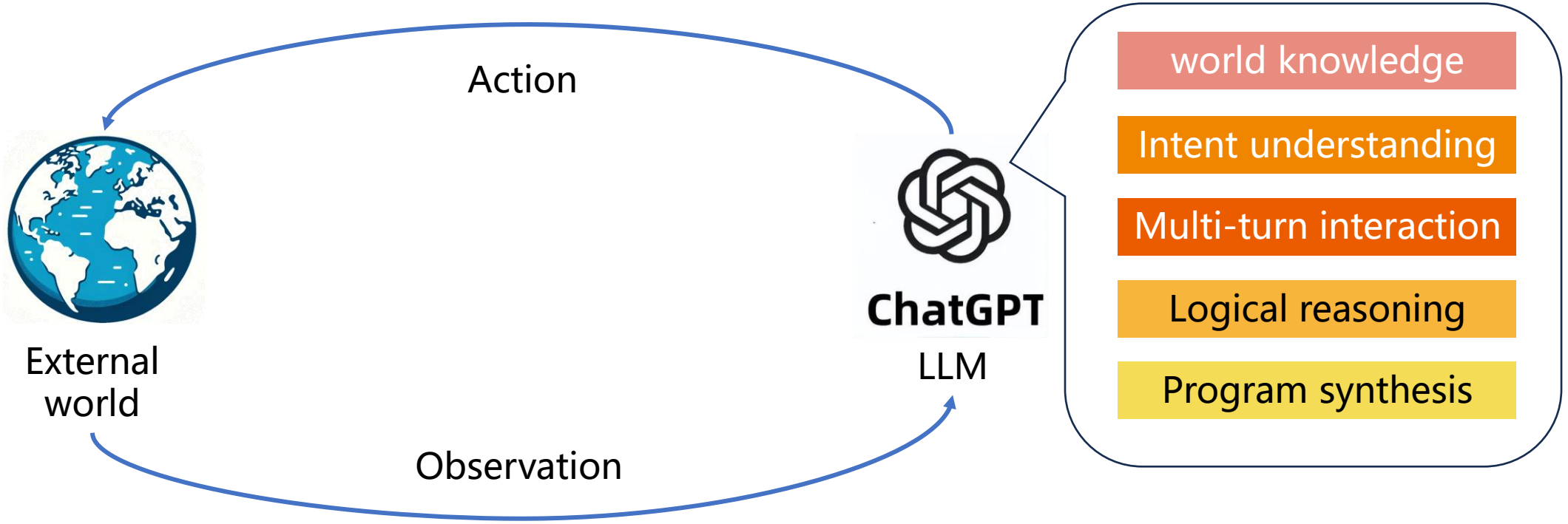


(Created with DALLE 3)

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.



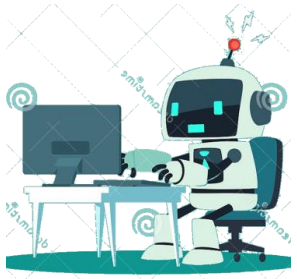
2 SheetCopilot

Motivation

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

Now

"Copy Cell A4 to C8..."



Agent

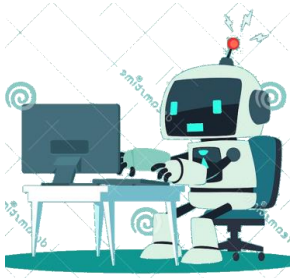
Employer

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:

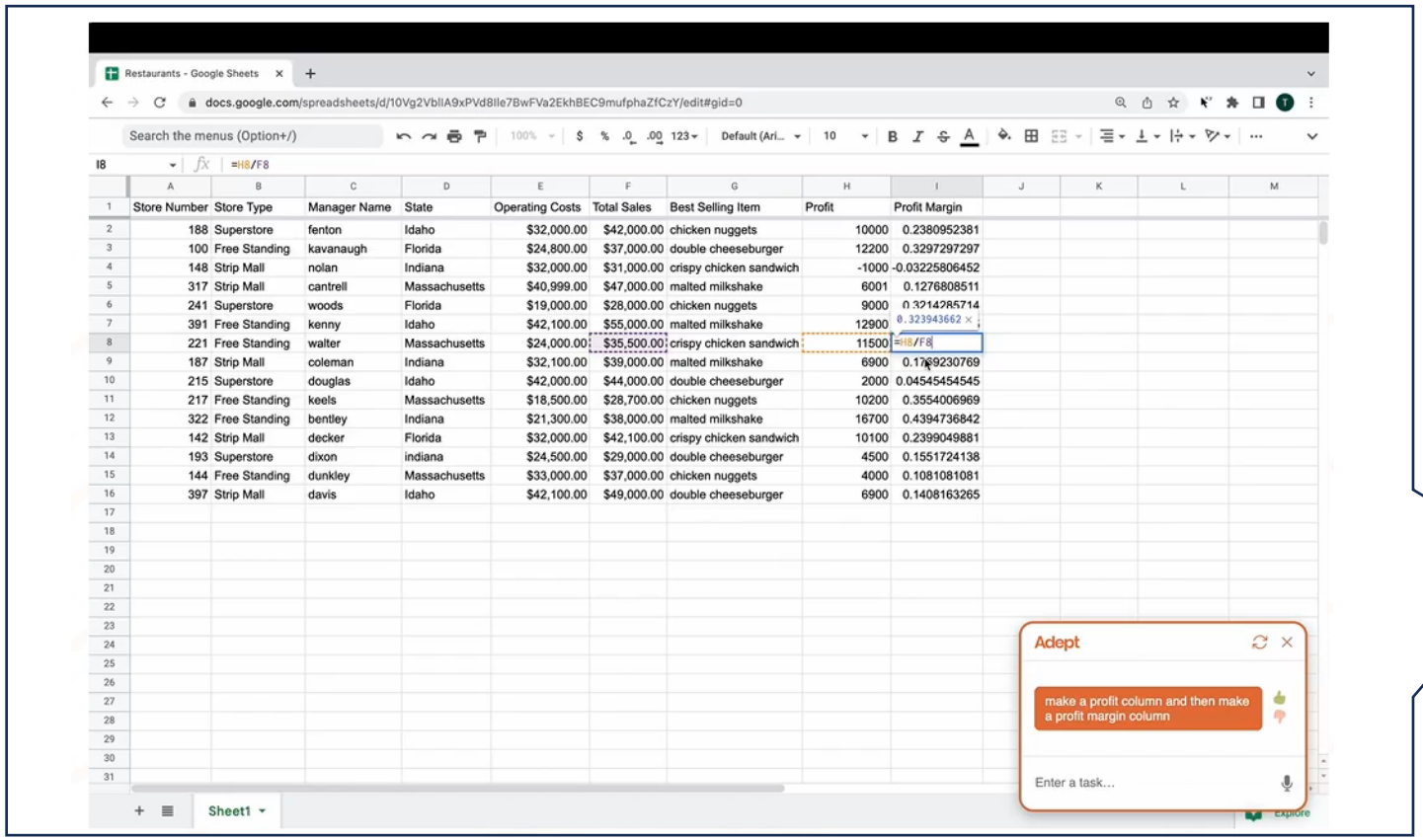
1. Can average users use complex software without acquiring expertise in advance?

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



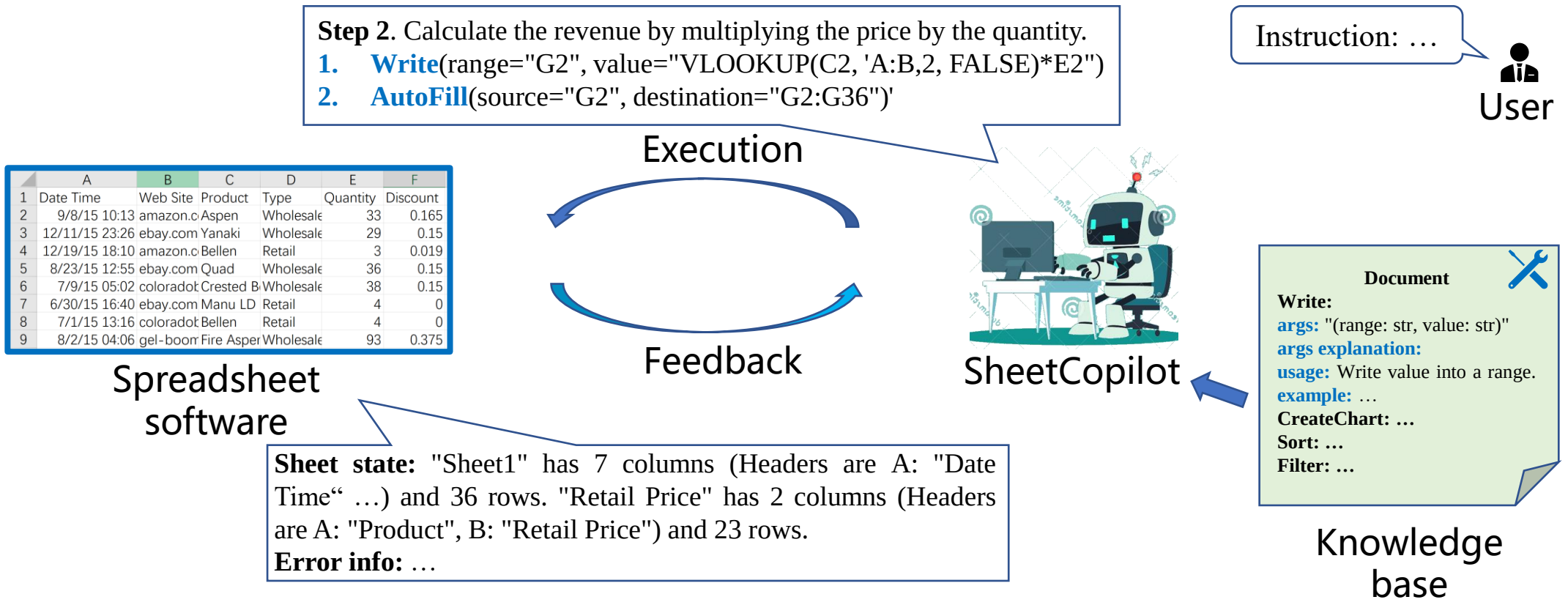
SheetCopilot

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.



	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-boom	Fire Asper	Wholesale	93	0.375

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	coloradol	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	coloradol	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	Sunspot	Wholesale	64	0.375

Sheet1 Retail Price

	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

Sheet1 Retail Price

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

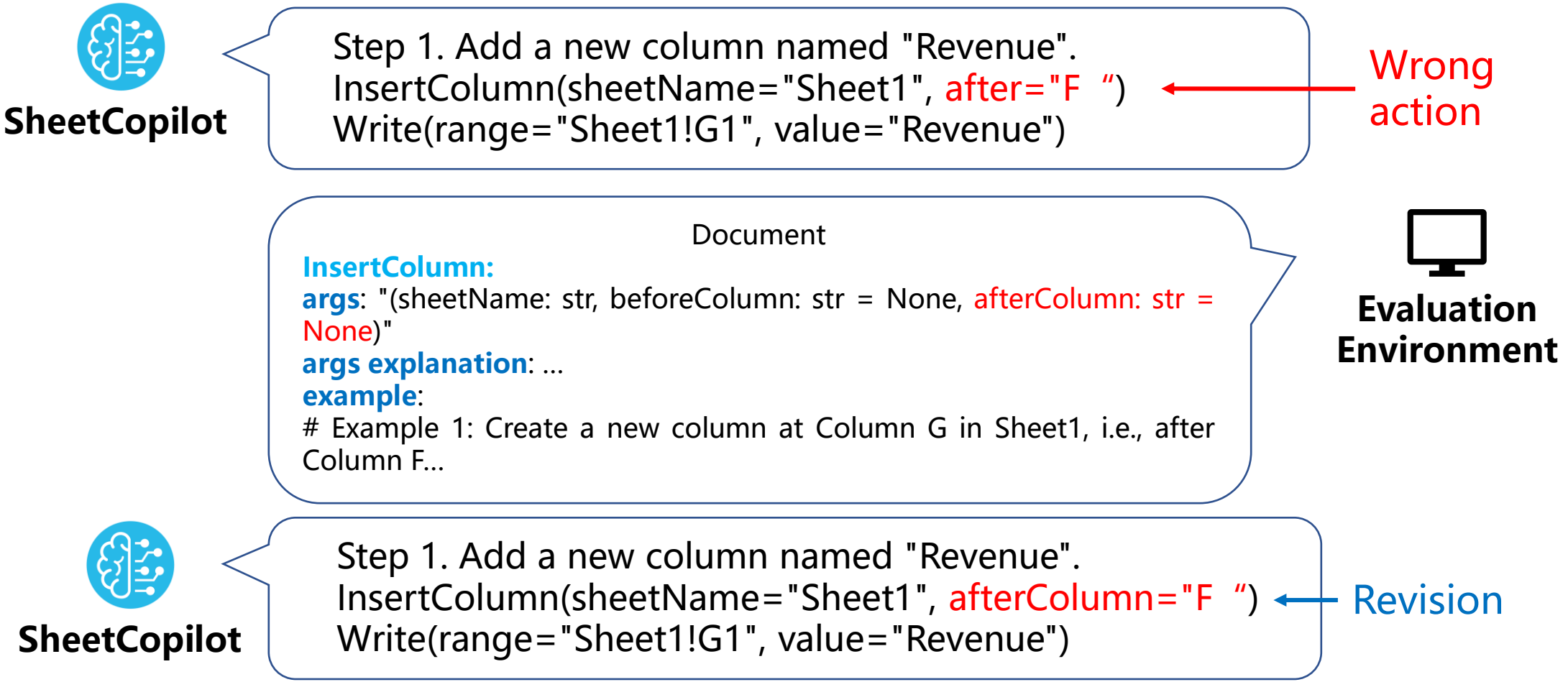
SheetCopilot

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.



Dataset

Benchmark creation pipeline:

- Scraping raw data
- Data cleaning
- Clustering
- Adaptation
- Simplification
- Human Verification

→ **221 tasks with Ground truths**

Word cloud of the instructions

Word cloud of the atomic actions

Experiments

Analysis

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

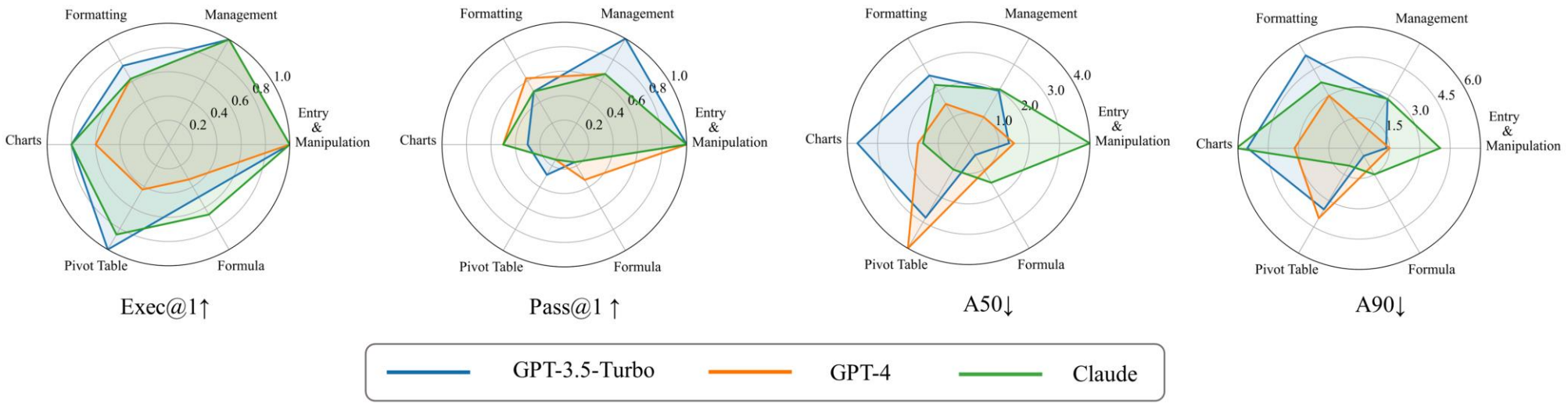
		Success		Efficiency	
Data	Models	Exec@1↑	Pass@1↑	A50↓	A90↓
10%	GPT-3.5-Turbo	**85.0%**	45.0%	2.00	4.50
10%	GPT-4	65.0%	**55.0%**	**1.33**	**2.00**
10%	Claude	80.0%	40.0%	1.50	4.40
100%	GPT-3.5-Turbo	87.3%	44.3%	1.50	3.00
100%	VBA	77.8%	37.1%	-	-

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

Experiments

Success

Efficiency



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

Alation 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

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

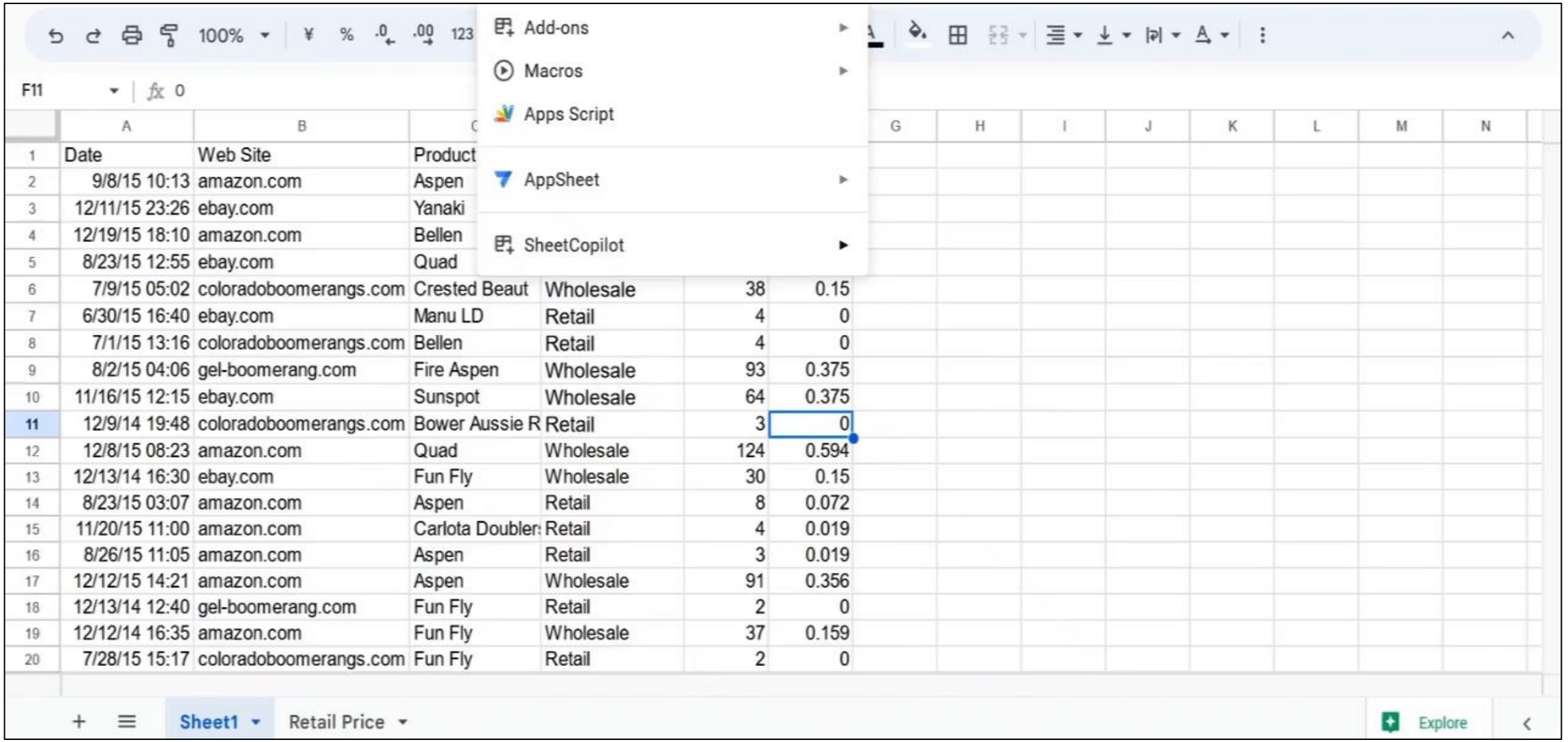
An interesting insight

- 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.

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

2 SheetCopilot

Interactive demo on Google Sheets

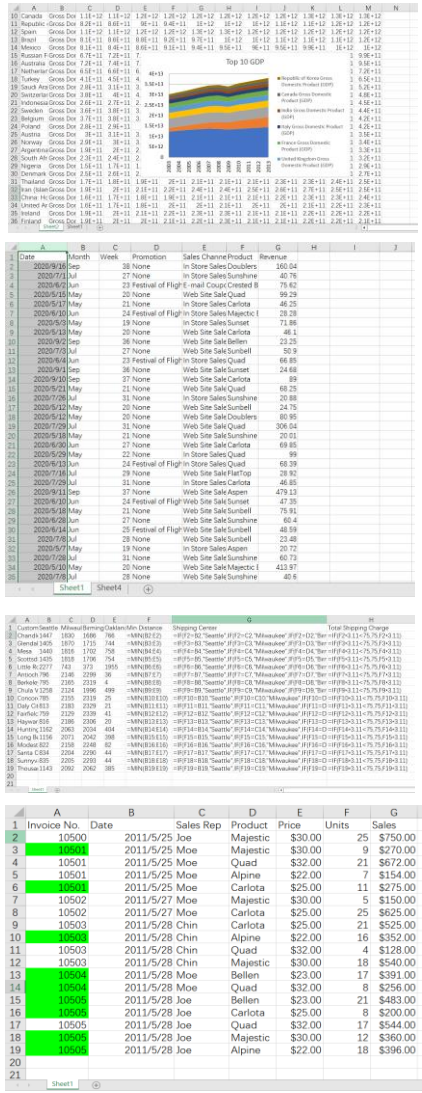


	A	B	C	D	E	F
1	Date	Web Site	Product			
2	9/8/15 10:13	amazon.com	Aspen			
3	12/11/15 23:26	ebay.com	Yanaki			
4	12/19/15 18:10	amazon.com	Bellen			
5	8/23/15 12:55	ebay.com	Quad			
6	7/9/15 05:02	coloradoboomerangs.com	Crested Beaut	Wholesale	38	0.15
7	6/30/15 16:40	ebay.com	Manu LD	Retail	4	0
8	7/1/15 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/

2 SheetCopilot

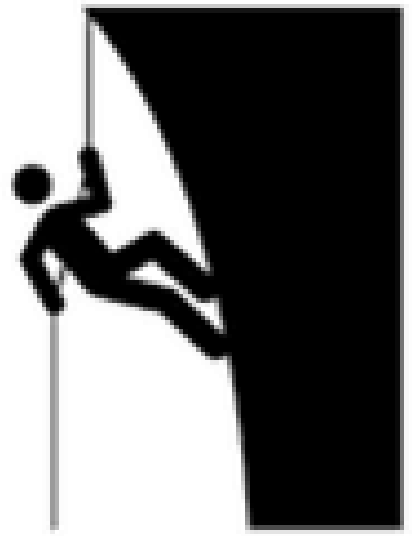
SheetCopilot vs. Humans



Invoice No.	Date	Sales Rep	Product	Price	Units	Sales
10500	2011/5/25	Joe	Majestic	$30.00	25	$750.00
10501	2011/5/25	Moe	Majestic	$30.00	9	$270.00
10501	2011/5/25	Moe	Quad	$32.00	21	$672.00
10501	2011/5/25	Moe	Alpine	$22.00	7	$154.00
10501	2011/5/25	Moe	Carlota	$25.00	11	$275.00
10502	2011/5/27	Moe	Majestic	$30.00	5	$150.00
10502	2011/5/27	Moe	Carlota	$25.00	25	$625.00
10503	2011/5/28	Chin	Carlota	$25.00	21	$525.00
10503	2011/5/28	Chin	Alpine	$22.00	16	$352.00
10503	2011/5/28	Chin	Quad	$32.00	4	$128.00
10503	2011/5/28	Chin	Majestic	$30.00	18	$540.00
10504	2011/5/28	Moe	Bellen	$23.00	17	$391.00
10504	2011/5/28	Moe	Quad	$32.00	8	$256.00
10505	2011/5/28	Joe	Bellen	$23.00	21	$483.00
10505	2011/5/28	Joe	Carlota	$25.00	8	$200.00
10505	2011/5/28	Joe	Quad	$32.00	17	$544.00
10505	2011/5/28	Joe	Majestic	$30.00	12	$360.00
10505	2011/5/28	Joe	Alpine	$22.00	18	$396.00

SheetCopilot

VS.



Do it by hand

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/