

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.



Content





Background SheetCopilot





>> 1 Background

EXCEPTION OF A CARDINAL AND A CARDIN

Large language models have shown impressive capabilities:

- Abundant world knowledge
- General intent understanding
- Powerful multi-turn interaction



• Program synthesis













Can we build a generalist agent using these LLM capabilities?



(Created with DALLE 3)

1 Background

RECEPCIÓN DE LA CONTRACTIÓN DE LA CONTRACTICACIÓN DE LA CONTRACTIC

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.











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?



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

÷	-> C	iocs.google.com	/spreadsheets/d/10	0Vg2VblIA9xPVd	8lle7BwFVa2EkhBE	C9mufphaZf0	22Y/edit#gid=0							Q	۵	☆	е ж			
	Search the me	enus (Option+/)		~~~ • •	100% - \$	% .000	123 - Default (Ari	10 👻	E	3 <i>I</i> \$ <u>A</u>	è .	H 3	E -	≡.	<u>+</u> +	÷ -	81-		~	
18		=H8/F8																		
	A	В	С	D	E	F	G	н		1		J		К		L		м		
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	10	000	0.2380952381										
3	100	Free Standing	kavanaugh	Florida	\$24,800.00	\$37,000.00	double cheeseburger	12	200	0.3297297297										
4	148	Strip Mall	nolan	Indiana	\$32,000.00	\$31,000.00	crispy chicken sandwich	-1	000 -	-0.03225806452										
5	317	Strip Mall	cantrell	Massachusetts	\$40,999.00	\$47,000.00	malted milkshake	6	001	0.1276808511										
6	241	Superstore	woods	Florida	\$19,000.00	\$28,000.00	chicken nuggets	9	000	0.3214285714										
7	391	Free Standing	kenny	Idaho	\$42,100.00	\$55,000.00	malted milkshake	12	900	0.323943662 ×										
8	221	Free Standing	walter	Massachusetts	\$24,000.00	\$35,500.00	crispy chicken sandwich	11	500	=H8/F8										
9	187	Strip Mall	coleman	Indiana	\$32,100.00	\$39,000.00	malted milkshake	6	900	0.1769230769										
10	215	Superstore	douglas	Idaho	\$42,000.00	\$44,000.00	double cheeseburger	2	000	0.04545454545										
11	217	Free Standing	keels	Massachusetts	\$18,500.00	\$28,700.00	chicken nuggets	10	200	0.3554006969										
12	322	Free Standing	bentley	Indiana	\$21,300.00	\$38,000.00	malted milkshake	16	700	0.4394736842										
13	142	Strip Mall	decker	Florida	\$32,000.00	\$42,100.00	crispy chicken sandwich	10	100	0.2399049881										
14	193	Superstore	dixon	indiana	\$24,500.00	\$29,000.00	double cheeseburger	4	500	0.1551724138										
15	144	Free Standing	dunkley	Massachusetts	\$33,000.00	\$37,000.00	chicken nuggets	4	000	0.1081081081										
16	397	Strip Mall	davis	Idaho	\$42,100.00	\$49,000.00	double cheeseburger	6	900	0.1408163265										
17																				
18																				
19																				
20																				
21																				
23												_							-	
24												Ade	ept					C	×)	
25												-						~	-	
26																				
27												m	ake a	profit c	olumn	and th	ien mal			
28												a	profit r	nargin	column			1		
29																			-	
30												-								
01																				





EXCEPCE Methode Metho

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.







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





Evaluation

Environment

Methodology

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

① Sheet state description

		А	В	С	D	Е	F		А	В
	1	Date Time	Web Site	Product	Туре	Quantity	Discount	1	Product	Retail Price
	2	9/8/15 10:13	amazon.c	Aspen	Wholesale	33	0.165	2	Alpine	21.95
	3	12/11/15 23:26	ebav.com	Yanaki	Wholesale	29	0.15	3	Aspen	21.95
Cierre e d	4	12/19/15 18.10	amazon c	Bellen	Retail	3	0 019	4	Carlota	19.95
Spread	5	8/23/15 12:55	ebay com	Quad	Wholesale	36	0.15	5	Crested Beaut	17.95
sheet	6	7/9/15 05:02	coloradok	Crested B	Wholesale	38	0.15	6	Majestic Beaut	26.95
		6/20/15 16:40			Detail	1	0.15	7	Sunshine	20
	1	0/30/15 10.40	ebay.com		Retail	4	0	8	Bellen	25
	8	//1/15 13:16	colorado	Bellen	Retail	4	0	9	Eagle	21.95
	9	8/2/15 04:06	gel-boom	Fire Aspe	Wholesale	93	0.375	10	Yanaki	23.95
	10	11/16/15 10·15 Sheet1 Retail Price ⊕	abaycom	Cupapat	Wholesole	61	<u> </u>		Sheet1 Retail Price +	00.05

Sheet state:

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

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

rows





Methodology

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

② Error feedback







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

				Succ	ess	Effici	ency
	Analysis			/			λ]
1.	SheetCopilot (GPT-4) demonstrates	Data	Models	Exec@1↑	Pass@1↑	A50↓	A90↓
	the strongest planning capability	10% 10%	GPT-3.5-Turbo GPT-4	85.0% 65.0%	45.0% 55.0%	2.00 1.33	4.50 2.00
2.	SheetCopilot (GPT-3.5-Turbo)	10%	Claude	80.0%	40.0%	1.50	4.40
	outperforms the VBA-based method	100%	GPT-3.5-Turbo	87.3%	44.3%	1.50	3.00
	(Exec@1 + 7.2% and Pass@1 + 7.9%)	100%	VBA	77.8%	37.1%	-	-

Exec@1[†]: Execution success rate Pass@1[†]: Functional correctness A50/A90[‡]: Median/90th percentile action number







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

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

1	▼ fx 0			e wie	10105										
	A	В	c a	🔰 Ap	ps Script			G	Н	1	J	К	L	М	N
	Date	Web Site	Product												
	9/8/15 10:13	amazon.com	Aspen	7 Ap	pSheet		►								
3	12/11/15 23:26	ebay.com	Yanaki												
1	12/19/15 18:10	amazon.com	Bellen E	E sh	eetConilot										
5	8/23/15 12:55	ebay.com	Quad	-+ 011	ceroophor										
6	7/9/15 05:02	coloradoboomerangs.com	Crested Be	eaut	Wholesale	38	0.15								
7	6/30/15 16:40	ebay.com	Manu LD		Retail	4	0								
В	7/1/15 13:16	coloradoboomerangs.com	Bellen		Retail	4	0								
9	8/2/15 04:06	gel-boomerang.com	Fire Aspen	i (Wholesale	93	0.375								
0	11/16/15 12:15	ebay.com	Sunspot		Wholesale	64	0.375								
1	12/9/14 19:48	coloradoboomerangs.com	Bower Aus	sie R	Retail	3	0								
2	12/8/15 08:23	amazon.com	Quad		Wholesale	124	0.594								
3	12/13/14 16:30	ebay.com	Fun Fly		Wholesale	30	0.15								
4	8/23/15 03:07	amazon.com	Aspen		Retail	8	0.072								
15	11/20/15 11:00	amazon.com	Carlota Do	ubler	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/





SheetCopilot vs. Humans



A	. 8	С	D	E	F	6	. Я	. I	1	К	L	м	N.
Canada	Gross Doe	1.11=12	1.11-12	1.21-12	1.21-12	1.25=12	1.25+12	1.25+12	1.25+12	1.38+12	1.38+12	1.36+12	
Reputito	+Gross Dor	8.2E+11	8.6E+11	9E+11	9.4E+11	15+12	1E+12	16+12	1.16+12	1.1E+12	1.2E+12	1.26+12	
Spen	Gross Dor	1.1E+12	1.1E+12	1.2E+12	1.25-12	1.35-12	1.35+12	1.26+12	1.26+12	1.2E+12	1.26+12	1.2E+12	
Brezh	Gross Dox	8.16=11	8.0E+11	8.8E=11	9.2E+11	9.7E=11	1E+12	16+12	1.16+12	1.16+12	1.16+12	1.26+12	
Mexico	Gross Dor	8.1E+11	8.4E+11	8.6E+11	9.15+11	9.42+11	9.5E+11	£+11	9.55+11	9.96+11	1E+12	16+12	
Russian	F Gross Dor	6.7E+11	7.2E+11	7.								9.9E+11	
Australia	Cross Dor	7.21=11	7.4E=11	7			Top 1	0 GDP			- 3	9.5E+11	
Notheria	r Gross Dox	6.5E+11	6.6f=11	6. 15.1								7.26+11	
Turkey.	Gross Dor	4.1E+11	4.5E=11	4				-	 Republik 	out Konea Gra	- 3	6.56+11	
Sauch Ar	a Gross Dor	2.81+11	3.15+11	3 3.56+1	1	-			Detest	s Product \$25	17) 7	5.25+11	
Saitzers	er Gross Dor	3.8E+11	4E+11	4 38+3		-		-	# Cerate	Gross Dorwood	i 5	4.8E+11	
Indones	a Gross Dor	2.6E+11	2.7E-11	2					Product	(409)	S	4.5E+11	
Sanden	Gross Dor	3.6E+11	3.8E+11	3	1				atoba de	os Datesta	Product 3	4.4E+11	
Refriger	Gross Dor	3.76+11	3.85=11	3 28+3					01041			4.25+11	
Poland	Gross Dor	2.8F+11	295+11	1.56+5	3				· Day feet	a bandar i	hadar 1	4.25+11	
Austria	Gross Dor	31+11	315+11	2					(00041			3.55+11	
Norway	Gross Dos	2.01+11	31+11	3	·				· france i	and Diverse	. 3	3.45+11	
Argentin	aGross Dor	1.9E=11	2E+11	2 58+3	2				Product	3(2397)		3.3E+11	
South A	h Gross Dee	2.25+11	245+11	2					attended to	Same Carry		326+11	
Materia	Gross Dra	1.55+11	1.76+11	1	8 3 1	6 8 8	888	1 3 3	Correct	e Product \$12	in 1	2.06+11	
Depthad	Gross Dor	2.5E+11	261-11	5	2 2 1	2 2 2	8 2 2	222				2 7E+11	
	Groce Devi	1.75-11	1.02-11	1.05-11	25+11	215+11	215+11	215+11	2 25+11	2 26 + 11	2/5+11	255+11	
Ten /kis	e Cross Dor	101-11	11=10	215-11	2.05+11	2.45+11	2.45+11	255+11	265+11	275+11	265+11	266+11	
Chate h		1.66-11	1.76+11	1.01-11	100-11	215-11	216+11	215+11	2 26 + 11	2.06+11	216+11	246+11	
I britest 2		1.65+11	176-11	1.05-11	26+11	25+11	216+11	26+11	26.111	215+11	2 20411	216+11	
hadanad	Course Days	1 105 - 22	77-22	237-31	1 10 - 11	1 10 - 11	10.000 - 1.1	110.11	210.11	1.10.11	1.101.00	1.10.11	
fieland	Cross Dor	1.01-11	31-11	11-11	215-11	2.02-11	2.05711	210-11	210+11	2.25*11	2.20733	2.25*11	
Contactor.	Cart Share	1		aca.	A.44.744			A.A.A.C.A.A.			A.416.7.4.8		
-		0											
		0	6			E	5		C.	-		6	1
P		Michiel		······································									

- 15	Date	Month	Week	Promotion	Sales Channe Product	Revenue		
2	2020/9/16	Sep	3	8 None	In Store Sales Doublers	160.04		
3	2020/7/3	Jul	2	7 None	In Store Sales Sunshine	40.76		
4	2020/6/2	Jun	2	3 Festival of Fligh	E-mail Coupe Crested 8	3 75.62		
5	2020/5/15	May	2	0 None	Web Site Sale Quad	99.29		
6	2020/5/17	May	2	1 None	In Store Sales Carlota	46.25		
	2020/6/10	Jun	2	4 Festival of Fligh	In Store Sales Majectic	1 28.28		
8	2020/5/3	May	1	9 None	In Store Sales Sunset	71.86		
9	2020/5/13	May	2	0 None	Web Site Sale Carlota	46.1		
	2020/9/2	Sep	3	6 None	Web Site Sale Bellen	23.25		
	2020/7/3	Jul	2	7 None	Web Site Sale Sunbell	50.9		
	2020/6/4	Jun	Z	3 Festival of Fligh	In Store Sales Quad	66.85		
	2020/9/3	Sep	3	6 None	Web Site Sale Sunset	24.68		
	2020/9/10	Sep	3	7 None	Web Site Sale Carlota	89		
	2020/5/21	May	2	L None	Web Site Sale Quad	68.25		
	2020/7/26	Jul	3	L None	In Store Sales Sunshine	20.88		
	2020/5/12	May	2	0 None	Web Site Sale Sunbell	24.75		
	2020/5/12	May	2	0 None	Web Site Sale Doublers	80.95		
	2020/7/29	Jul	3	L None	Web Site Sale Quad	306.04		
	2020/5/18	May	2	1 None	Web Site Sale Sunshine	20 01		
	2028/6/30	Jun	2	7 None	Web Site Sale Carlota	69.85		
	2020/5/29	May	2	2 None	In Store Sales Quad	99		
	2020/6/13	Jun	2	4 Festival of Fligh	In Store Sales Quad	68.39		
	2020/7/16	Jul	2	9 None	Web Site Sale FlatTop	28.92		
	2020/7/29	Dul	3	1 None	In Store Sales Carlota	46.85		
26	2020/9/11	Sep.	3	7 None	Web Site Sale Aspen	479.13		
	2020/6/10	Dun	2	E Festival of Fligh	Web Site Sale Sunset	47.35		
	2020/5/18	May	2	1 None	Web Site Sale Sunbell	75.91		
29	2020/6/28	Jun	2	7 None	Web Site Sale Sunshine	60.4		
	2020/6/14	Jun	2	5 Festival of Flip	Web Site Sale Sunbell	48.59		
	2020/7/8	Dul	2	8 None	Web Site Sale Sunbell	23.48		
	2020/5/7	May	1	9 None	In Store Sales Aspen	20.72		
	2020/7/28	Jul	3	1 None	Web Site Sale Sunshine	60.73		
	2020/5/10	May	2	0 None	Web Site Sale Majectic	I 413.97		

Sheet1 Sheet4

A.	A 8	C	0	E	F		G			8
1 0	ustorn Seatt	le Mèner	al Bernie	g Oakia	n:Min Distance	Shipping Center			Total Shipping	Charge
	handk1447	1830	1686	765	=MIN(B2.E2)	=IF(F2=82."Seattle	JEFE2=C2. Weiner	isee" ##2+D2."B	er= F(F2<3.11<7	5.75.F2+3.11)
	endal 3405	1870	1/15	744	=MIN(83±3)	=P(F3=K3, Seattle	JHP3=C3 Mines	According to DALL	# =1F(F3<3.11<7	5/5/3-311
6.0	offert 1635	1818	1205	254	-58N(85.65)	-BUES-IS "Seattle	URS-CS Man	AND 1015-05 10	er = 1605x3 11e2	57515(312)
6 0	ttle R: 2277	743	373	1955	=M9N(B5:55)	=#076=86."Seattle	HURSTCR MANA	Ace" JFIT0 : D6 TB	er #1F(F6-3.11 </td <td>57576+3113</td>	57576+3113
TA	ttioch 795	2145	2299	36	=MN(87.27)	=8(97=87."Seattle	JIGT=C7, Minut	Ace" ##7=07.78	er =1F(F7+3.11<7	5.75J7+3.11)
8 54	orkolo:795	2165	2319	4	=MRN(B8:28)	=#(F8=\$8,"Seattle	JHHE=CR. MANN	Aver JF/FE=DE/B	er =1F(F8+3.11<7	5.7538-3.11)
9 0	1ufa V 1258	2124	1996	499	=NRN(89:E9)	=#(F9=89,"Seattle	"JEF9=CR."Milmai	1400" (F)[F9=D9,"B)	er = IF(F9+3.11<7	5.7539-311)
	20000 785	2155	2319	25	=MIN(B10E10)	=#(#10=010."Seat	tie", #(F10=C10,"N	Ryankies, http://www.	D =IF(F10+3.11<	75.75,710-3.11)
	average 200	2183	2329	41	-14N/(011211)	-16/011-011 Seat	No. SPECIAL COLUMN	Auguster HULL-	0-0012-011-0	
IR H	htmp: 515	2185	2306	20	=MIN(R12F13)	=IFIF13::R13 Seat	the" IFIF13=C13"N	HAWGERS" IFIFTE	D =1FIF13+3 11<	75 75 F13-3 111
14 14	unting 1162	2063	2034	404	=MIN(B)4E14)	=IF(F)4=B)4."Seat	tlo".IFIF14=C14."N	Avaulase" JF(F14=	D=IFIF14+3.11<	75.75.F14-3.111
15 14	ing Bk 1116	2071	2042	398	=MIN(825-825)	=IF(F15=015,"Seat	the".#EF15=C15."N	Hwaukee* JF(F15=	D=IF(F15+3.11<	75,75,F15+3.11)
16 M	odest 822	2158	2248	82	=MIN(B16.E16)	=IF(F16=B16."Seat	tle", #[F16=C16,"N	Dwackee" JF(F16=	D =IF(F16=3.11<	75.75,916+3.11)
17 S	irita C834	2204	2290	-44	=NRN(B17E17)	=IF(F17=B17,"Seat	the", #EF17=C17."W	Dwackee" JF(F17=	D =IF(F17+3.11<	75,75,F17+3.11)
10 X	annya 635	2005	2003	505	-hen(d18118)	-stirig-Bill Seat	De .mp1d=U18.W	Auguster (10112-	0 - IF(F18-3.11<	75,75,748+311
20	0010 1140	Acos	1.010	5455	-midinary	- epita-tata seat	in septembers in	commence of these	a sub as sure	-2.129 10-2.11
23										
	Liberit's	÷						1.(4)		
1	1	A			В	С	D	E	F	G
1	Invoid	e No	. D.	ate		Sales Rep	Product	Price	Units	Sales
2	_	1050	00		2011/5/25	Joe	Majestic	\$30.00	25	\$750.0
3		1050)1		2011/5/25	Moe	Majestic	\$30.00	g	\$270.0
4		1050	01		2011/5/25	Moe	Quad	\$32.00	21	\$672.0
5		1050)1		2011/5/25	Moe	Alpine	\$22.00	7	\$154.0
6		1050)1		2011/5/25	Moe	Carlota	\$25.00	11	\$275.0
7		1050)2		2011/5/27	Moe	Majestic	\$30.00	5	\$150.0
8		1050)2		2011/5/27	Moe	Carlota	\$25.00	25	\$625.0
9	_	1050)3		2011/5/28	Chin	Carlota	\$25.00	21	\$525.0
10		1050)3		2011/5/28	Chin	Alpine	\$22.00	16	\$352.0
11		1050	03		2011/5/28	Chin	Quad	\$32.00	4	\$128.0
12		1050)3		2011/5/28	Chin	Majestic	\$30.00	18	\$540.0
13		1050)4		2011/5/28	Moe	Bellen	\$23.00	17	\$391.0
14		1050)4		2011/5/28	Moe	Quad	\$32.00	8	\$256.0
15		1050)5		2011/5/28	Joe	Bellen	\$23.00	21	\$483.0
16		1050)5		2011/5/28	Joe	Carlota	\$25.00	8	\$200.0
4.72		1050)5		2011/5/28	Joe	Quad	\$32.00	17	\$544.0
17										
17		1050)5		2011/5/28	Joe	Majestic	\$30.00	12	\$360.0
17 18 19		1050)5)5		2011/5/28 2011/5/28	Joe Joe	Alpine	\$30.00	12	\$360.0

VD .

5 \$150.00 25 \$625.00 21 \$525.00 16 \$352.00 4 \$128.00 18 \$540.00 17 \$391.00 8 \$256.00 21 \$483.00 21 \$483.00

21 \$483.00 8 \$200.00 17 \$544.00 12 \$360.00





SheetCopilot





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/