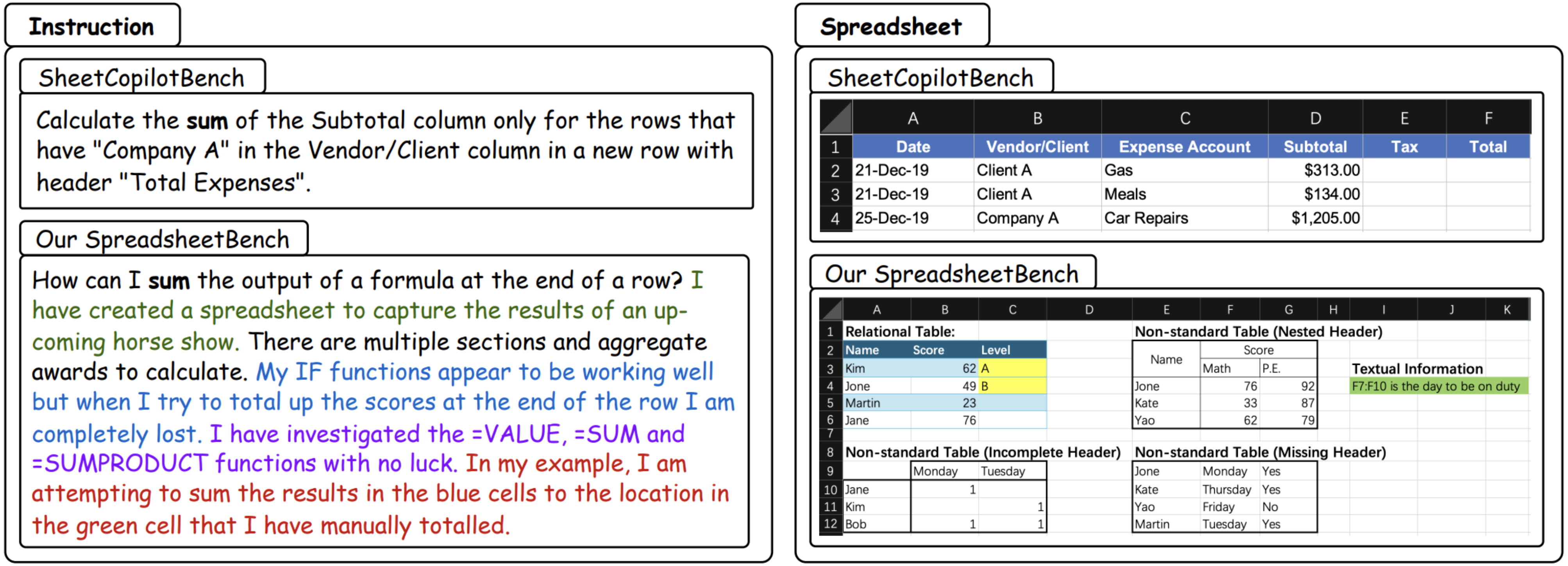
SpreadsheetBench: Towards Challenging Real World Spreadsheet Manipulation

Zeyao Ma, Bohan Zhang, Jing Zhang, Jifan Yu, Xiaokang Zhang,
Xiaohan Zhang, Sijia Luo, Xi Wang, Jie Tang

Spreadsheet Manipulation Task: Current Problem



Calculate the **sum** of the Subtotal column only for the rows that have "Company A" in the Vendor/Client column in a new row with header "Total Expenses".

How can I **sum** the output of a formula at the end of a row? I have created a spreadsheet to capture the results of an up-coming horse show. There are multiple sections and aggregate awards to calculate. My IF functions appear to be working well but when I try to total up the scores at the end of the row I am completely lost. I have investigated the =VALUE, =SUM and =SUMPRODUCT functions with no luck. In my example, I am attempting to sum the results in the blue cells to the location in the green cell that I have manually totalled.

	A	B	C	D	E	F
1	Date	Vendor/Client	Expense Account	Subtotal	Tax	Total
2	21-Dec-19	Client A	Gas	$313.00		
3	21-Dec-19	Client A	Meals	$134.00		
4	25-Dec-19	Company A	Car Repairs	$1,205.00		

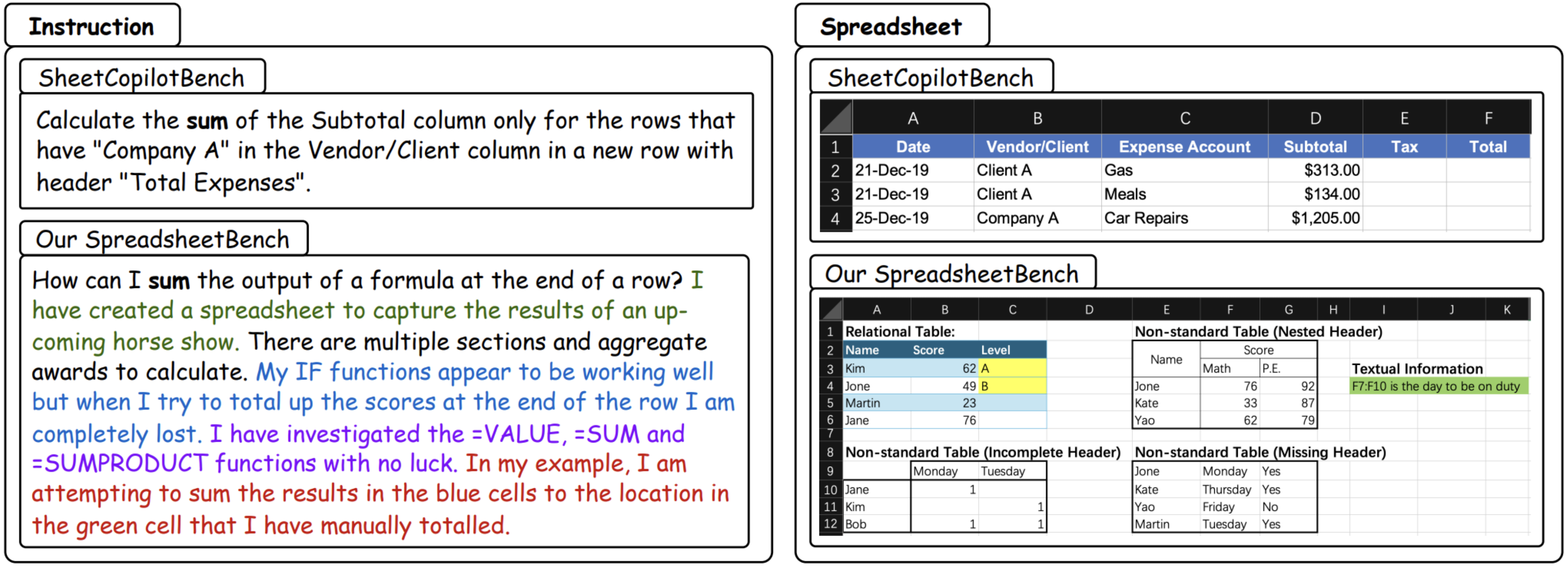
	A	B	C	D	E	F	G	H	I	J	K
1	Relational Table:				Non-standard Table (Nested Header)						
2	Name	Score	Level		Name	Score					
3	Kim	62	A			Math	P.E.		Textual Information		
4	Jone	49	B		Jone	76	92		F7:F10 is the day to be on duty		
5	Martin	23			Kate	33	87				
6	Jane	76			Yao	62	79				
7											
8	Non-standard Table (Incomplete Header)				Non-standard Table (Missing Header)						
9		Monday	Tuesday		Jone	Monday	Yes				
10	Jane	1			Kate	Thursday	Yes				
11	Kim		1		Yao	Friday	No				
12	Bob	1	1		Martin	Tuesday	Yes				

(1) **Short and Simple Instructions:** Self-instruct or manual writing based on a few given examples.

(2) **Over Simplified Spreadsheets:** Contain only one regular relational table.

(3) **Lack of Test Cases:** Involve only one single test case for each instruction.

Spreadsheet Manipulation Task: Our Solution



Instruction

SheetCopilotBench

Calculate the **sum** of the Subtotal column only for the rows that have "Company A" in the Vendor/Client column in a new row with header "Total Expenses".

Our SpreadsheetBench

How can I **sum** the output of a formula at the end of a row? I have created a spreadsheet to capture the results of an up-coming horse show. There are multiple sections and aggregate awards to calculate. My IF functions appear to be working well but when I try to total up the scores at the end of the row I am completely lost. I have investigated the =VALUE, =SUM and =SUMPRODUCT functions with no luck. In my example, I am attempting to sum the results in the blue cells to the location in the green cell that I have manually totalled.

Spreadsheet

SheetCopilotBench

	A	B	C	D	E	F
1	Date	Vendor/Client	Expense Account	Subtotal	Tax	Total
2	21-Dec-19	Client A	Gas	$313.00		
3	21-Dec-19	Client A	Meals	$134.00		
4	25-Dec-19	Company A	Car Repairs	$1,205.00		

Our SpreadsheetBench

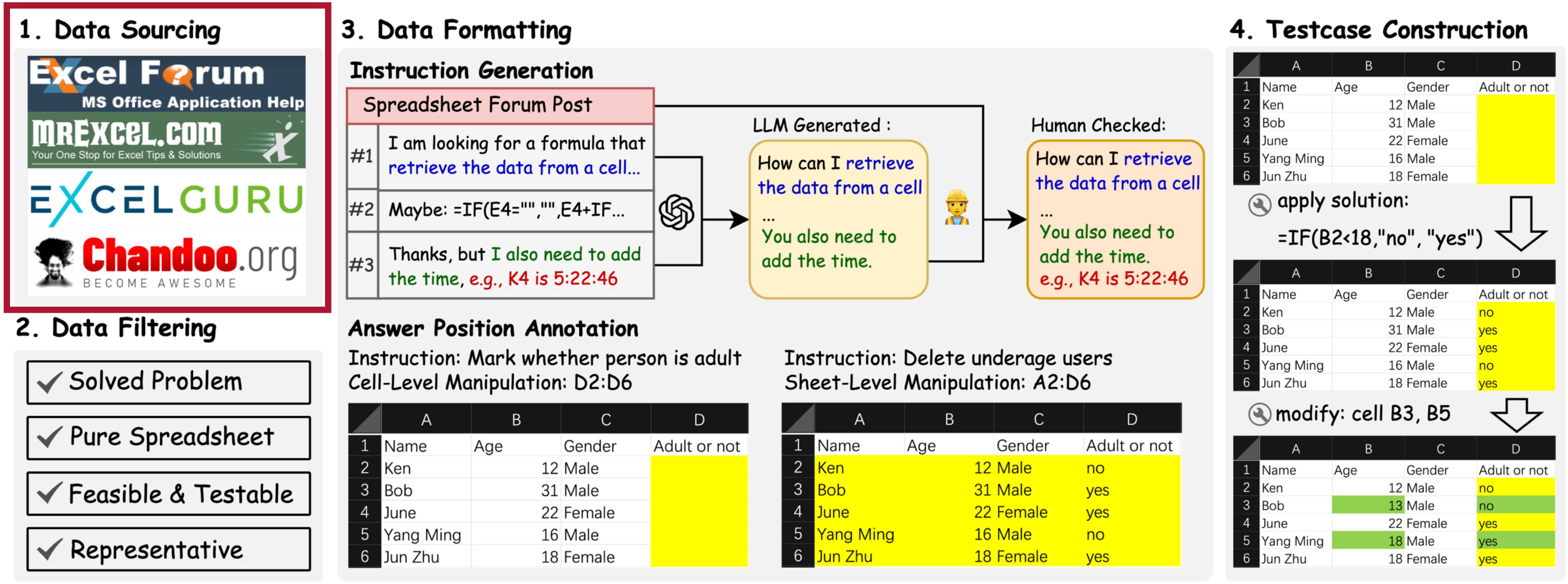
	A	B	C	D	E	F	G	H	I	J	K
1	Relational Table:				Non-standard Table (Nested Header)						
2	Name	Score	Level		Name	Score					
3	Kim	62	A			Math	P.E.		Textual Information		
4	Jone	49	B		Jone	76	92		F7:F10 is the day to be on duty		
5	Martin	23			Kate	33	87				
6	Jane	76			Yao	62	79				
7											
8	Non-standard Table (Incomplete Header)				Non-standard Table (Missing Header)						
9		Monday	Tuesday		Jone	Monday	Yes				
10	Jane	1			Kate	Thursday	Yes				
11	Kim		1		Yao	Friday	No				
12	Bob	1	1		Martin	Tuesday	Yes				

(1) **Complex Instructions from Real World:** Gather real user queries from popular Excel forums.

(2) **Spreadsheet in Diverse Formats:** Feature non-standard tables, multiple tables and styles

(3) **OJ-style Evaluation Metric:** Three test cases per instruction and comprehensive evaluation.

Construction Pipeline: Data Sourcing



1. Data Sourcing

Excel Forum
MS Office Application Help
MrExcel.com
Your One Stop for Excel Tips & Solutions
EXCELGURU
Chandoo.org
BECOME AWESOME

2. Data Filtering

- ✓ Solved Problem
- ✓ Pure Spreadsheet
- ✓ Feasible & Testable
- ✓ Representative

3. Data Formatting

Instruction Generation

Spreadsheet Forum Post

#1	I am looking for a formula that retrieve the data from a cell...
#2	Maybe: =IF(E4="","",E4+IF...
#3	Thanks, but I also need to add the time, e.g., K4 is 5:22:46

LLM Generated :

How can I retrieve the data from a cell
...
You also need to add the time.

Human Checked:

How can I retrieve the data from a cell
...
You also need to add the time.
e.g., K4 is 5:22:46

Answer Position Annotation

Instruction: Mark whether person is adult
Cell-Level Manipulation: D2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

Instruction: Delete underage users
Sheet-Level Manipulation: A2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

4. Testcase Construction

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

apply solution:
=IF(B2<18,"no", "yes")

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

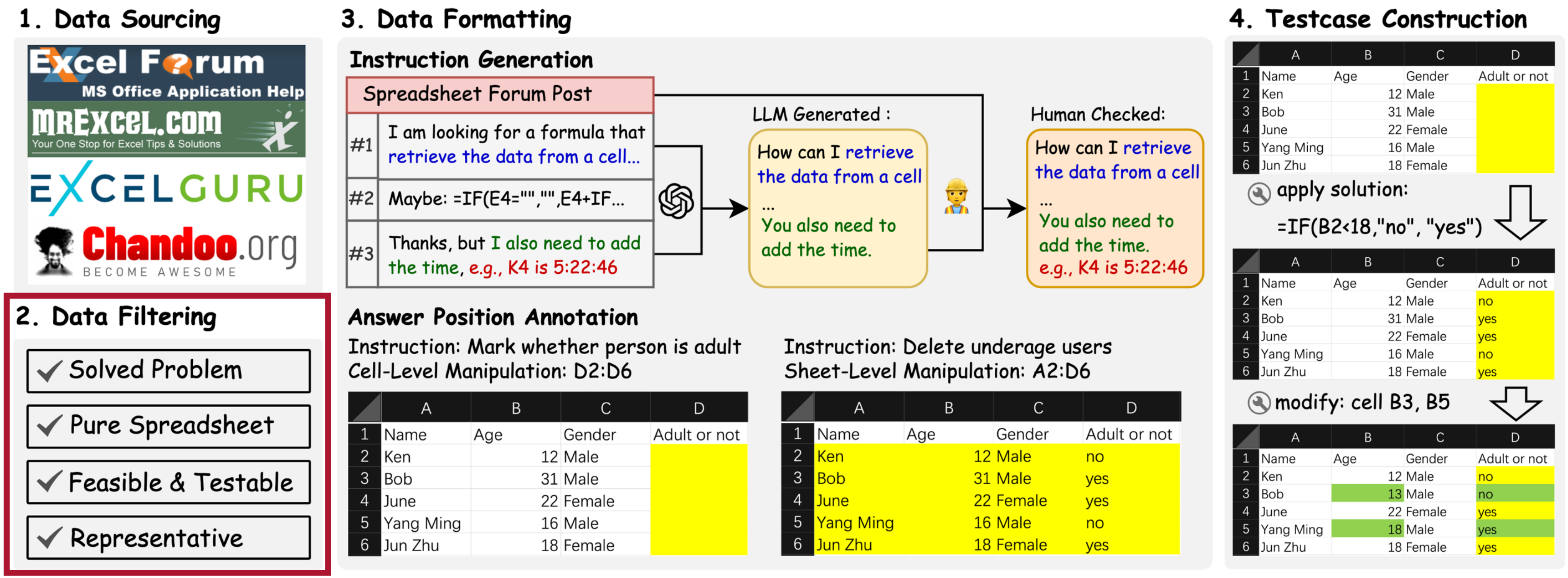
modify: cell B3, B5

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	13	Male	no
4	June	22	Female	yes
5	Yang Ming	18	Male	yes
6	Jun Zhu	18	Female	yes

1. Data Sourcing:

- Our source data is collected on four frequently updated Excel forums.
- We target posts that fall under categories *Formula*, *VBA & Marco*, etc, to ensure relevance to spreadsheet manipulation.

Construction Pipeline: Data Filtering


Excel Forum
MS Office Application Help
MREXCEL.COM
Your One Stop for Excel Tips & Solutions
EXCELGURU
Chandoo.org
BECOME AWESOME

1. Data Sourcing

2. Data Filtering

- ✓ Solved Problem
- ✓ Pure Spreadsheet
- ✓ Feasible & Testable
- ✓ Representative

3. Data Formatting

Instruction Generation

Spreadsheet Forum Post

#1 I am looking for a formula that retrieve the data from a cell...

#2 Maybe: =IF(E4="","",E4+IF...

#3 Thanks, but I also need to add the time, e.g., K4 is 5:22:46

LLM Generated :

How can I retrieve the data from a cell
...
You also need to add the time.

Human Checked:

How can I retrieve the data from a cell
...
You also need to add the time.
e.g., K4 is 5:22:46

Answer Position Annotation

Instruction: Mark whether person is adult
Cell-Level Manipulation: D2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

Instruction: Delete underage users
Sheet-Level Manipulation: A2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

4. Testcase Construction

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

apply solution:
=IF(B2<18,"no", "yes")

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

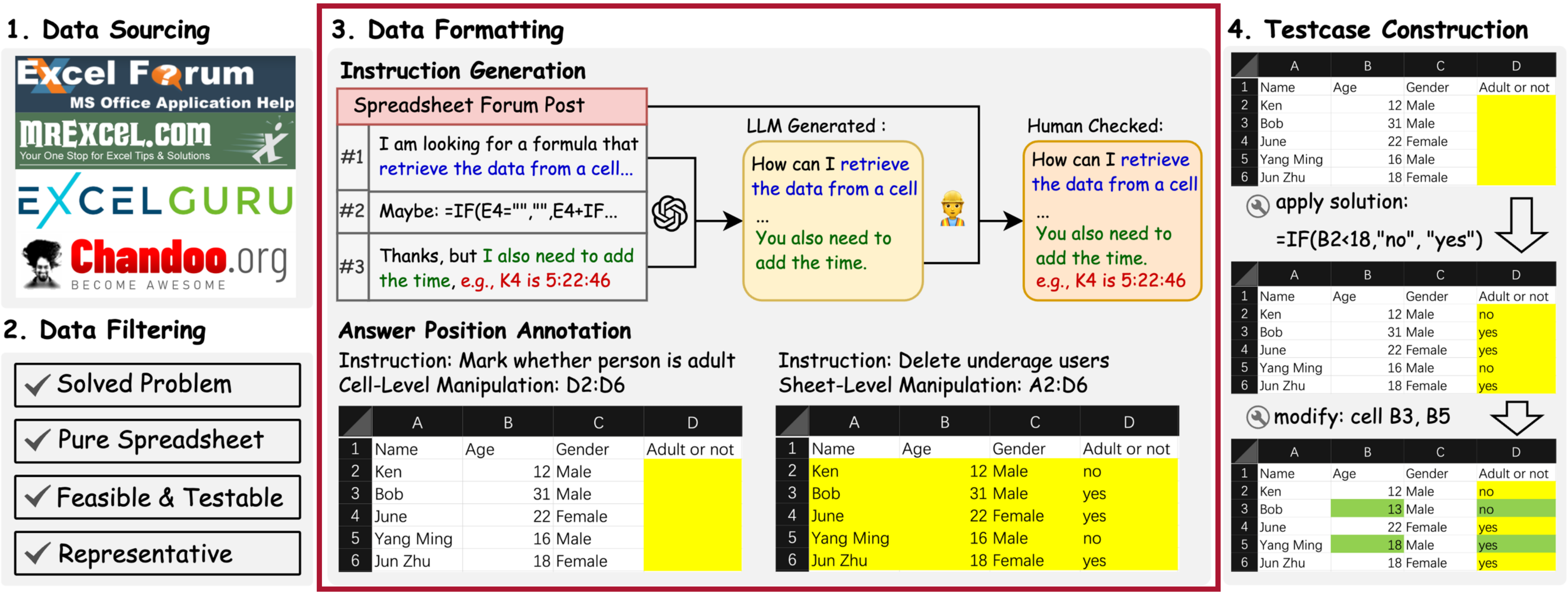
modify: cell B3, B5

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	13	Male	no
4	June	22	Female	yes
5	Yang Ming	18	Male	yes
6	Jun Zhu	18	Female	yes

2. Data Filtering:

- Identify solved problem based on the tag provided by the forum or the judgment of GPT-4.
- Discard all irrelevant software-specific questions using keywords such as “input box”, “forms”, etc.
- Exclude posts without spreadsheet attachments or ambiguous presentations.
- Filter those with high view counts and ask the annotators to retain both simple and difficult questions.

Construction Pipeline: Data Formatting


Excel Forum
MS Office Application Help
MREXCEL.COM
Your One Stop for Excel Tips & Solutions
EXCELGURU
Chandoo.org
BECOME AWESOME

1. Data Sourcing

2. Data Filtering

✓ Solved Problem

✓ Pure Spreadsheet

✓ Feasible & Testable

✓ Representative

3. Data Formatting

Instruction Generation

Spreadsheet Forum Post

#1	I am looking for a formula that retrieve the data from a cell...
#2	Maybe: =IF(E4="","",E4+IF...
#3	Thanks, but I also need to add the time, e.g., K4 is 5:22:46

LLM Generated :

How can I retrieve the data from a cell
...
You also need to add the time.

Human Checked:

How can I retrieve the data from a cell
...
You also need to add the time.
e.g., K4 is 5:22:46

Answer Position Annotation

Instruction: Mark whether person is adult
Cell-Level Manipulation: D2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

Instruction: Delete underage users
Sheet-Level Manipulation: A2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

4. Testcase Construction

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

apply solution:

=IF(B2<18,"no", "yes")

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

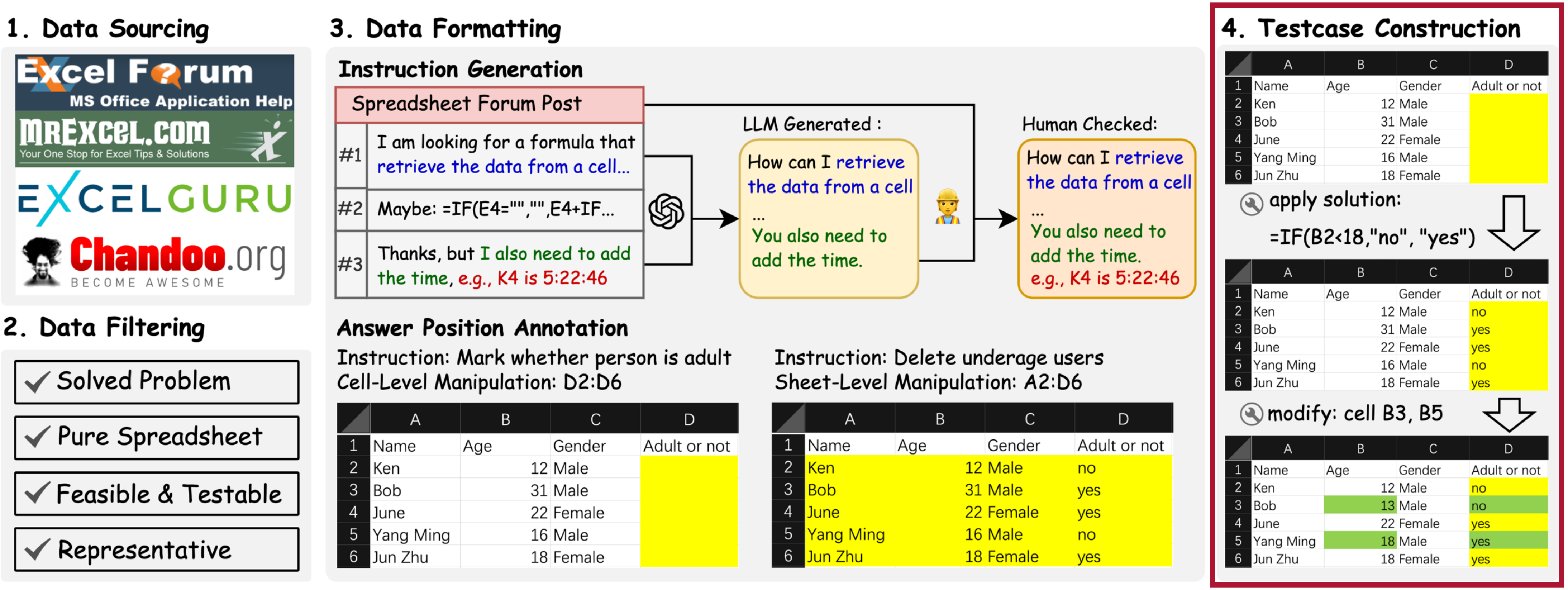
modify: cell B3, B5

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	13	Male	no
4	June	22	Female	yes
5	Yang Ming	18	Male	yes
6	Jun Zhu	18	Female	yes

3. Data Formatting:

- Instruction Generation: First, we utilize GPT-4 to recreate a coherent instruction from the original post. Then, annotators verify this instruction to resolve any issues arising from incomplete context extraction.
- Answer Position Annotation: Through this annotation, we restrict some “open-ended” question to “fill in the blanks” question.

Construction Pipeline: Test Case Construction



Spreadsheet Forum Post

#1	I am looking for a formula that retrieve the data from a cell...
#2	Maybe: =IF(E4="","",E4+IF...
#3	Thanks, but I also need to add the time, e.g., K4 is 5:22:46

LLM Generated : How can I retrieve the data from a cell ... You also need to add the time.

Human Checked: How can I retrieve the data from a cell ... You also need to add the time. e.g., K4 is 5:22:46

Instruction: Mark whether person is adult
Cell-Level Manipulation: D2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

Instruction: Delete underage users
Sheet-Level Manipulation: A2:D6

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

4. Testcase Construction

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	
3	Bob	31	Male	
4	June	22	Female	
5	Yang Ming	16	Male	
6	Jun Zhu	18	Female	

apply solution: =IF(B2<18,"no", "yes")

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	31	Male	yes
4	June	22	Female	yes
5	Yang Ming	16	Male	no
6	Jun Zhu	18	Female	yes

modify: cell B3, B5

	A	B	C	D
1	Name	Age	Gender	Adult or not
2	Ken	12	Male	no
3	Bob	13	Male	no
4	June	22	Female	yes
5	Yang Ming	18	Male	yes
6	Jun Zhu	18	Female	yes

4. Test Case Construction:

- Some solutions provided by forum users may be applicable only to the specific example spreadsheet.
- We modify the spreadsheet files to construct more test cases and corner cases.

Against Data Leakage

We utilize the following perturbations to minimize data leakage:

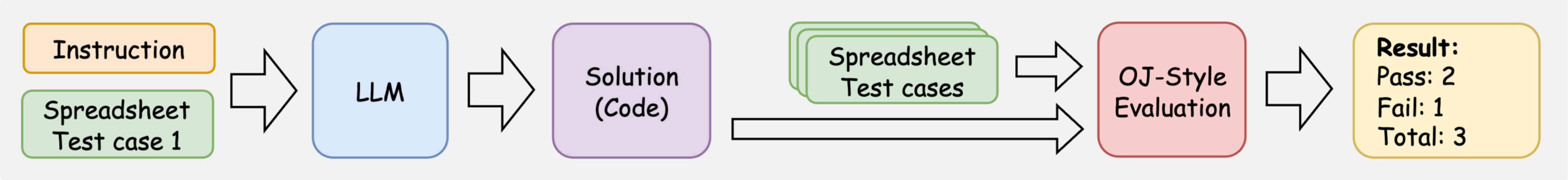
- Instruction Generation: The aforementioned instruction generation strategy, carried out by GPT-4 and human annotators, involves revising the original questions in the posts, thereby preventing LLMs from memorizing the original questions.

- Spreadsheet Modification: The test case construction strategy outlined above involves modifying the original provided spreadsheets, preventing LLMs from memorizing the original spreadsheets.

- Answer Position Changing: We also alter the position of the tabular data in the original spreadsheets and the corresponding answer in the resulting spreadsheets. By doing so, the originally provided solution from the posts cannot be directly used to derive answers with changed positions.

Evaluation Metrics

5. OJ-Style Evaluation Pipeline



Instruction

Spreadsheet Test case 1

LLM

Solution (Code)

Spreadsheet Test cases

OJ-Style Evaluation

Result:
Pass: 2
Fail: 1
Total: 3

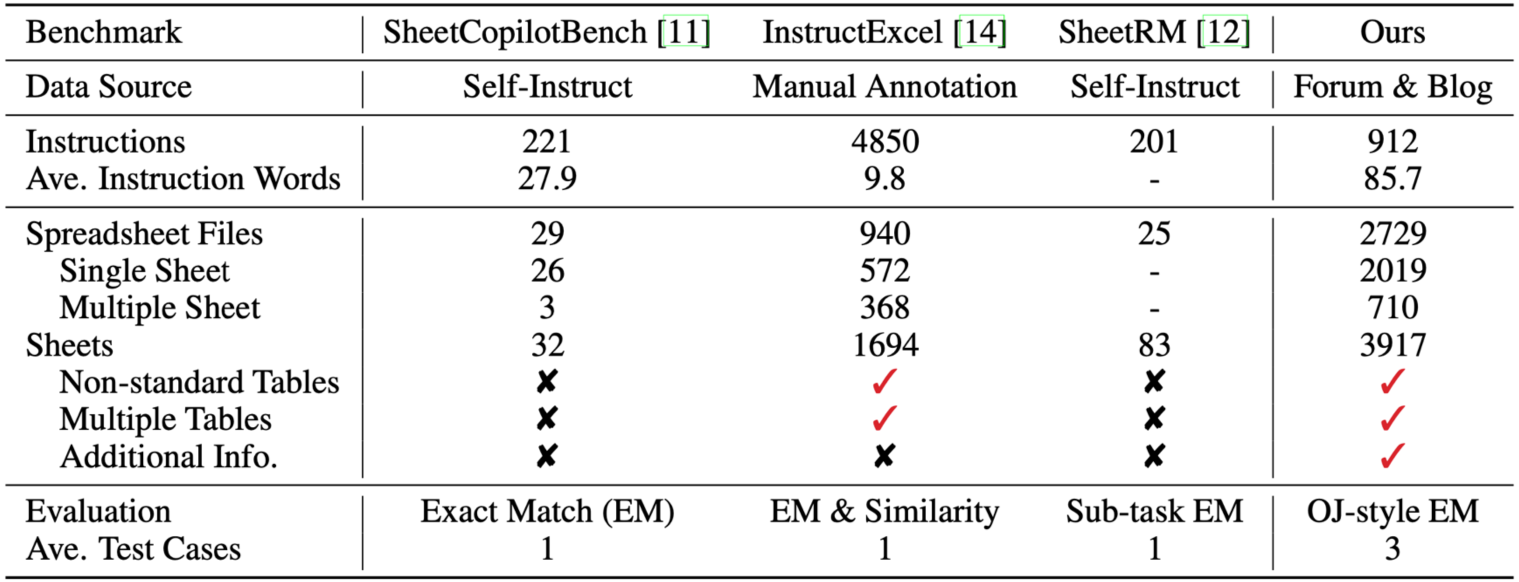
Soft restriction adheres to the scoring principles of the OJ system from the IOI, granting partial credit when a solution only passes some test cases:

$$S_{soft} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\left(\frac{1}{|T_i|}\sum_{j=1}^{|T_i|}\mathbb{1}_{r_{ij}=ACC}\right).$$

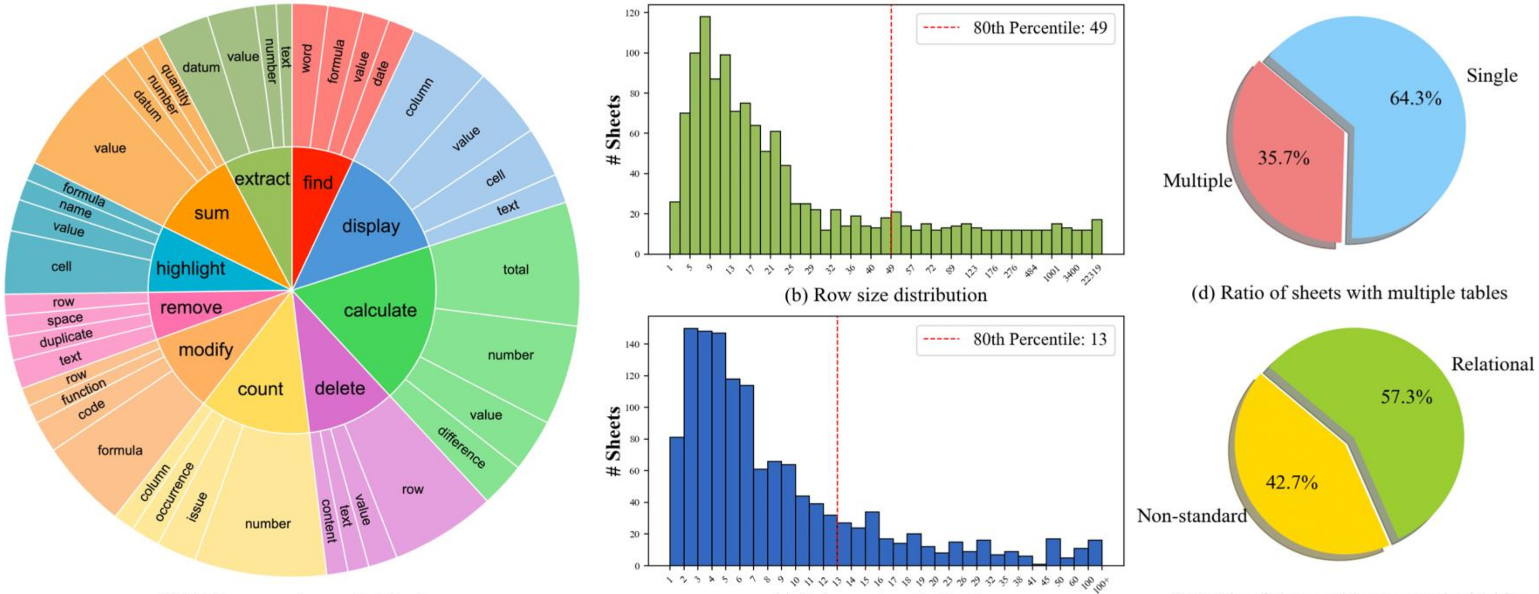
Hard restriction follows the ICPC scoring rules of the OJ system, where no partial credit is awarded:

$$S_{hard} = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\mathbb{1}_{r_{ij}=ACC,\forall j=1,2,\dots,|T_i|}.$$

Benchmark Statistic



Benchmark	SheetCopilotBench [11]	InstructExcel [14]	SheetRM [12]	Ours
Data Source	Self-Instruct	Manual Annotation	Self-Instruct	Forum & Blog
Instructions	221	4850	201	912
Ave. Instruction Words	27.9	9.8	-	85.7
Spreadsheet Files	29	940	25	2729
Single Sheet	26	572	-	2019
Multiple Sheet	3	368	-	710
Sheets	32	1694	83	3917
Non-standard Tables	✘	✓	✘	✓
Multiple Tables	✘	✓	✘	✓
Additional Info.	✘	✘	✘	✓
Evaluation	Exact Match (EM)	EM & Similarity	Sub-task EM	OJ-style EM
Ave. Test Cases	1	1	1	3



(a) Verb-noun phrase distribution

(b) Row size distribution

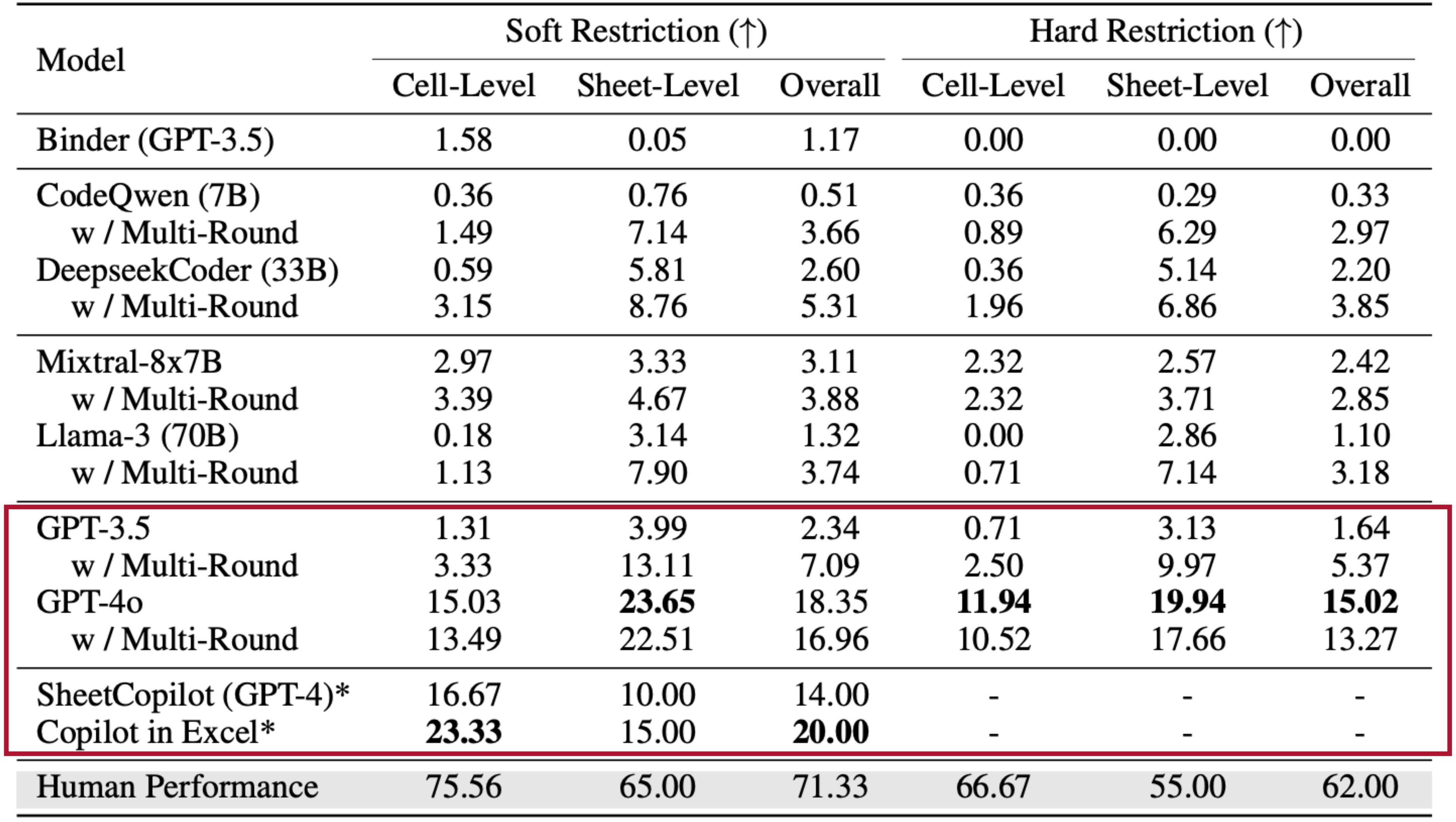
(c) Column size distribution

(d) Ratio of sheets with multiple tables

(e) Ratio of sheets with non-standard tables

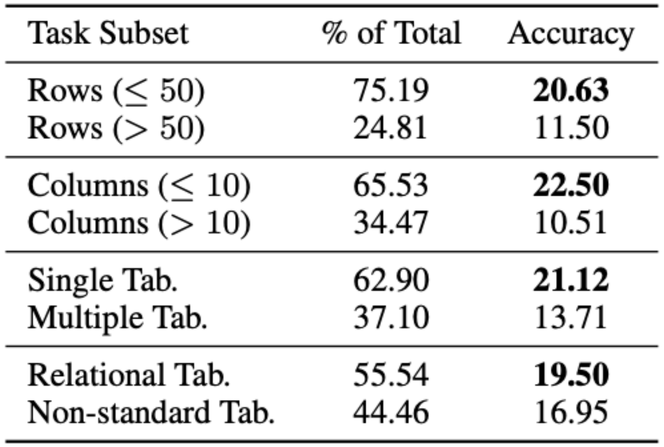
Experiments: Main Result

Table 2: Performance of representative models on SpreadsheetBench (%).



Model	Soft Restriction (↑)			Hard Restriction (↑)		
	Cell-Level	Sheet-Level	Overall	Cell-Level	Sheet-Level	Overall
Binder (GPT-3.5)	1.58	0.05	1.17	0.00	0.00	0.00
CodeQwen (7B)	0.36	0.76	0.51	0.36	0.29	0.33
w / Multi-Round	1.49	7.14	3.66	0.89	6.29	2.97
DeepseekCoder (33B)	0.59	5.81	2.60	0.36	5.14	2.20
w / Multi-Round	3.15	8.76	5.31	1.96	6.86	3.85
Mixtral-8x7B	2.97	3.33	3.11	2.32	2.57	2.42
w / Multi-Round	3.39	4.67	3.88	2.32	3.71	2.85
Llama-3 (70B)	0.18	3.14	1.32	0.00	2.86	1.10
w / Multi-Round	1.13	7.90	3.74	0.71	7.14	3.18
GPT-3.5	1.31	3.99	2.34	0.71	3.13	1.64
w / Multi-Round	3.33	13.11	7.09	2.50	9.97	5.37
GPT-4o	15.03	**23.65**	18.35	**11.94**	**19.94**	**15.02**
w / Multi-Round	13.49	22.51	16.96	10.52	17.66	13.27
SheetCopilot (GPT-4)*	16.67	10.00	14.00	-	-	-
Copilot in Excel*	**23.33**	15.00	**20.00**	-	-	-
Human Performance	75.56	65.00	71.33	66.67	55.00	62.00

Experiments: Analysis



Task Subset	% of Total	Accuracy
Rows (≤ 50)	75.19	**20.63**
Rows (> 50)	24.81	11.50
Columns (≤ 10)	65.53	**22.50**
Columns (> 10)	34.47	10.51
Single Tab.	62.90	**21.12**
Multiple Tab.	37.10	13.71
Relational Tab.	55.54	**19.50**
Non-standard Tab.	44.46	16.95

Table 3: Overall soft restriction of GPT-4o on different subsets (%).

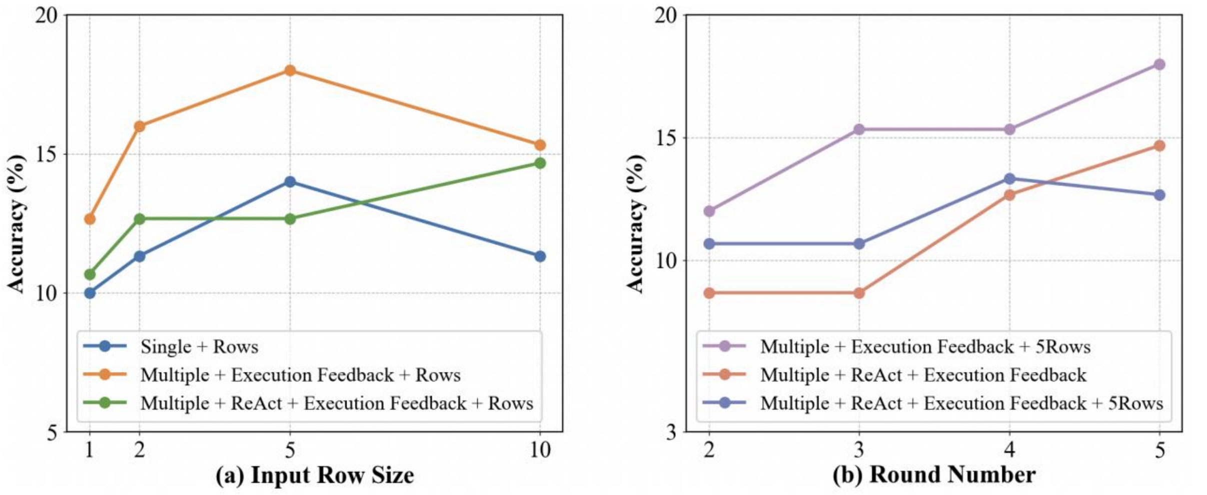


Figure 5: The impact of input row size and round number on GPT-4o across different inference settings. Accuracy represents the value of the overall soft restriction.

- Spreadsheets with more rows, more columns, multiple tables and non-standard tables is more difficult for current LLMs.
- Including more rows in the prompt can enhance the performance of LLM but two much rows may lead to the performance degradation.
- LLMs benefit from multi-round setting for spreadsheet manipulation. More interaction rounds will improve model performance.

Thank you!

[Paper] https://arxiv.org/pdf/2406.14991

[Code] https://github.com/RUCKBReasoning/SpreadsheetBench

[Sample Data] https://github.com/RUCKBReasoning/SpreadsheetBench