

# SPREADSHEETBENCH

雷翔麟

华中科技大学计算机科学与技术学院

2024 年 10 月 11 日



- ① Ideas
- ② Details
- ③ Codes
- ④ Conclusion
- ⑤ Motivation

- ① Ideas
- ② Details
- ③ Codes
- ④ Conclusion
- ⑤ Motivation

# Five Key Ideas

- Collect high-quality **data** from real-world sources and select the questions by rigorous criteria
- Utilize GPT-4 to recreate a coherent **instruction**
- Categorize **answer positions** into sheet-level and cell-level
- Create multiple spreadsheets and develop multiple **test cases** for each instruction
- Use various methods to mitigate **data leakage**

## 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       | 33  | Male   | no           |
| 4 | June      | 22  | Female | yes          |
| 5 | Yang Ming | 18  | Male   | yes          |
| 6 | Jun Zhu   | 18  | Female | yes          |

## 5. OJ-Style Evaluation Pipeline

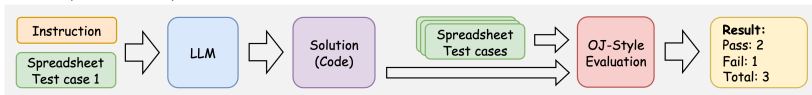
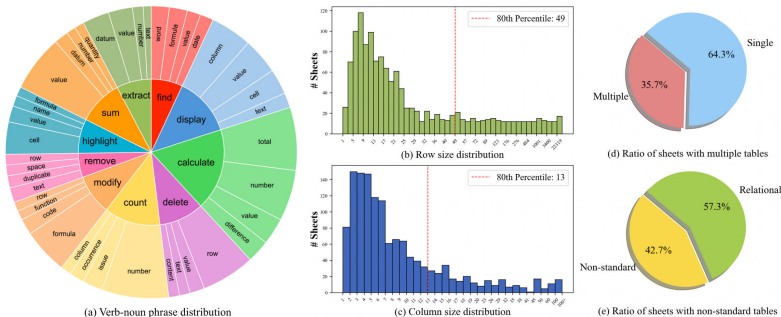


Figure 1: The benchmark construction pipeline and OJ-style evaluation.

- 1 Ideas
- 2 Details
- 3 Codes
- 4 Conclusion
- 5 Motivation

Figure 2: Key statistics of SPREADSHEETBENCH.



# Data Leakage

**Issue:** Datasets initially **obtained from online** forums may be susceptible to data leakage issues, given that many LLMs are pre-trained using a vast corpus of web text.

**Solutions:**

- **Revise the original questions** in the posts during the Instruction Generation process.
- **modifying the original provided spreadsheets** during the Spreadsheet Modification.
- **alter the position** of the tabular data in the original spreadsheets and the corresponding answer in the resulting spreadsheets during the Answer Position Changing



# Evaluation Metrics

- **Soft Restriction:**

$$S_{\text{soft}} = \frac{1}{|D|} \sum_{i=1}^{|D|} \left( \frac{1}{|T_i|} \sum_{j=1}^{|T_i|} 1_{r_i = \text{ACC}} \right)$$

- **Hard Restriction:**

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

# Inference Setting

Evaluate LLMs under two distinct settings:

- **Single-Round:** present the model with the initial few rows of spreadsheet files within the prompt, allowing for **only one inference**.
- **Multi-Round:** Building on the single-round prompt setting, furnish error feedback if the code fails to execute, enabling the model to refine its code in subsequent iterations.

- 1 Ideas
- 2 Details
- 3 Codes**
- 4 Conclusion
- 5 Motivation

# LLM Processing

<https://github.com/RUCKBReasoning/SpreadsheetBench>

<https://github.com/gersteinlab/MedAgents>

- ① Select datasets.
- ② Pass function parameters.
- ③ Based on the function parameters, choose:
  - Dataset
  - Model type
  - Processing method
  - Output location
- ④ Extract, clean, and format the data, then pass it to the large model.
- ⑤ Utilize the large model and the established pipeline for processing.
- ⑥ Output the processed results to files or other locations.
- ⑦ Use the evaluation module to assess and score the results.

# Enhancing LLM Problem-Solving

- **Prompt Engineering:** Designing precise prompts to guide LLMs towards generating higher-quality responses by framing the input effectively.
- **Fine-Tuning:** Adapting an LLM to specific domains by training it on domain-specific datasets, improving its performance in that area.
- **Chain-of-Thought (CoT):** Encouraging step-by-step reasoning in the model, simulating how humans break down complex tasks into simpler steps to enhance logical consistency.
- **Self-Consistency (SC):** Generating multiple independent reasoning chains for the same problem and choosing the most frequent answer to reduce randomness and errors.

# Enhancing LLM Problem-Solving

- **Multi-Round Interaction:** Enabling the model to refine its answers through multiple rounds of interaction, adjusting based on user feedback for improved accuracy.
- **Multi-Agent System:** Introducing multiple agents or models to collaboratively solve tasks, leveraging diverse skills and knowledge for more effective problem-solving.
- **Retrieval-Augmented Generation (RAG):** Combining generation with information retrieval, allowing the model to fetch relevant information from external knowledge sources to improve answer accuracy.
- **Classical Algorithms:** For example, using C4.5 Discretization for data handling and PCA for data analysis and preprocessing.

- ① Ideas
- ② Details
- ③ Codes
- ④ Conclusion
- ⑤ Motivation

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

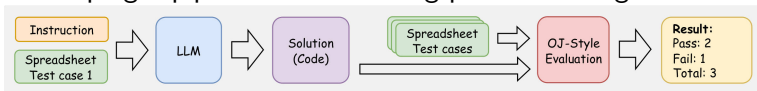
| Model                 | Soft Restriction (↑) |              |              | Hard Restriction (↑) |              |              |
|-----------------------|----------------------|--------------|--------------|----------------------|--------------|--------------|
|                       | Cell-Level           | 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                | <b>23.65</b> | 18.35        | <b>11.94</b>         | <b>19.94</b> | <b>15.02</b> |
| 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*     | <b>23.33</b>         | 15.00        | <b>20.00</b> | -                    | -            | -            |
| Human Performance     | 75.56                | 65.00        | 71.33        | 66.67                | 55.00        | 62.00        |

Figure 3: Performance of representative models on SPREADSHEETBENCH %.



- ① Ideas
- ② Details
- ③ Codes
- ④ Conclusion
- ⑤ Motivation

- The concept of constructing a benchmark:
  - Data quality
  - Data construction
  - Data diversity
- Methods to address data leakage issues
- Developing a pipeline for evaluating problems using LLMs



- Some methods to enable LLMs to handle problems more effectively

*Thanks!*