# 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

# 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**

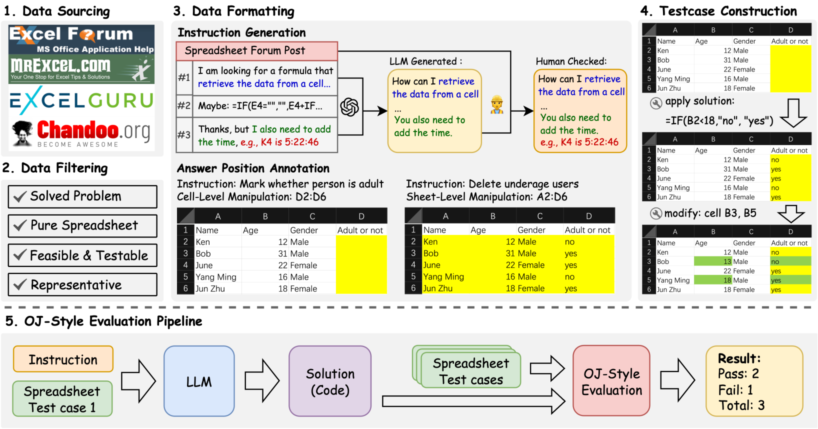



Figure 1: The benchmark construction pipeline and OJ-style evaluation.

# Data Info

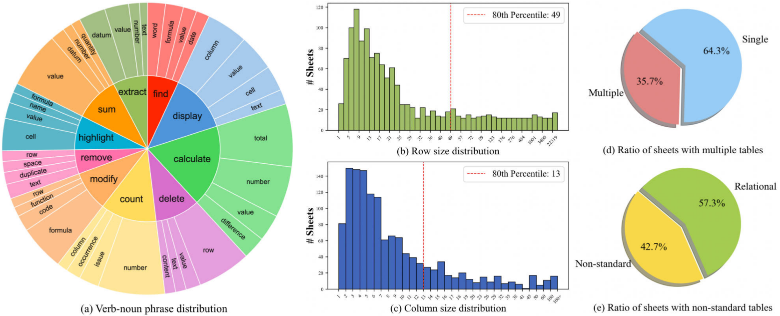



Figure 2: Key statistics of SPREADSHEETBENCH.

# 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:**

$$S_{hard} = \frac{1}{|D|} \sum_{i=1}^{|D|} 1_{rij} = ACC, \forall j = 1, 2, \ldots, |T_i|$$

# 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.

## 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.

# 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.

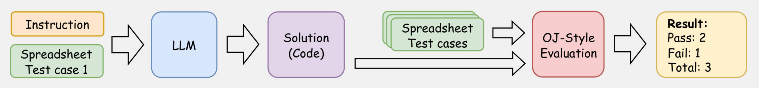
1. Ideas
2. Details
3. Codes
4. **Conclusion**
5. Motivation

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 | **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 |

Figure 3: Performance of representative models on SPREADSHEETBENCH %.

- 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!*