Sheetpedia

A 300K-Spreadsheet Corpus for Spreadsheet Intelligence and LLM Fine-Tuning

Zailong Tian, Zhuoheng Han, Houfeng Wang, Lizi Liao

北京大学
PEKING UNIVERSITY

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Introduction and Motivation

- **The Gap:** A severe lack of large-scale, diverse, and formula-rich public datasets for spreadsheet research.
- **Limitation:** Existing corpora (EUSES, Enron, Fuse) are too small, domain-specific, or lack complex formulas, hindering AI/NLP model development.
- **Our Goal:** To create **Sheetpedia**, a high-quality corpus to fill this gap, enabling advanced spreadsheet intelligence and LLM fine-tuning.

Sheetpedia Overview

- **Scale:** 290,509 unique worksheets from 324,988 workbooks.
- **Sources:**

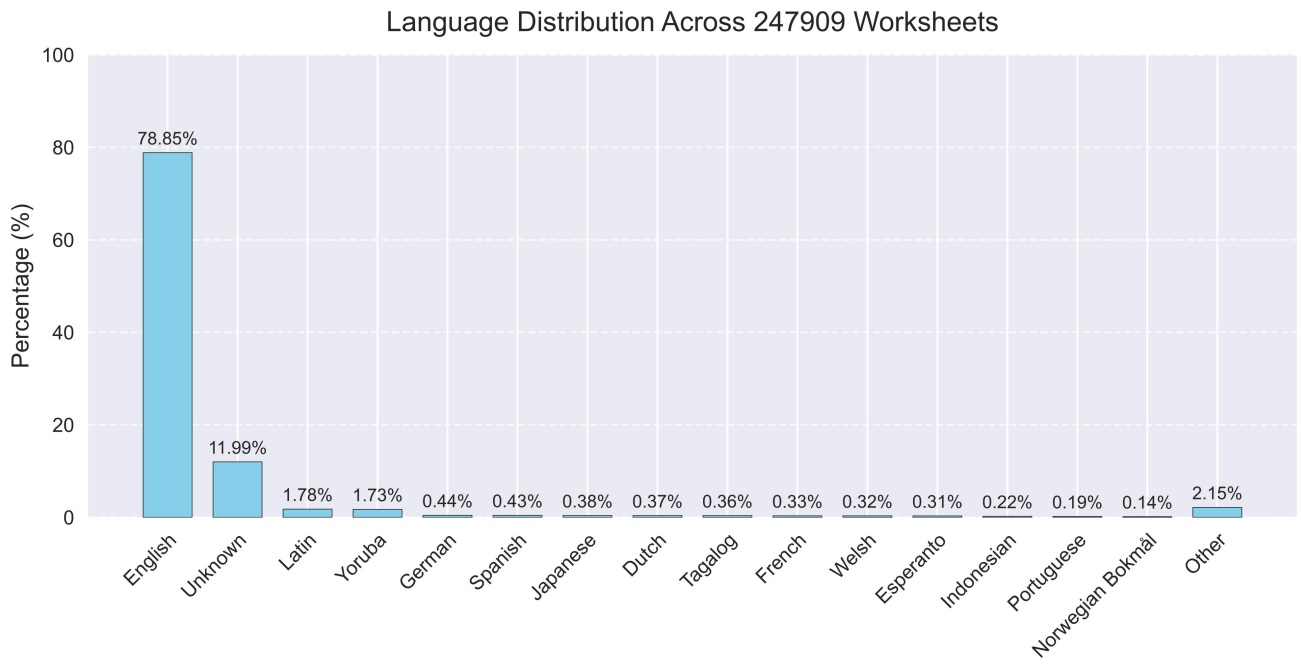
  Enron: 62,612 worksheets (enterprise).

  Fuse: 182,784 worksheets (web).

  ExcelForum: 320,489 worksheets (user-contributed).
- **Preprocessing:** Format standardization, language filtering (78%English), deduplication (48.7% reduction).

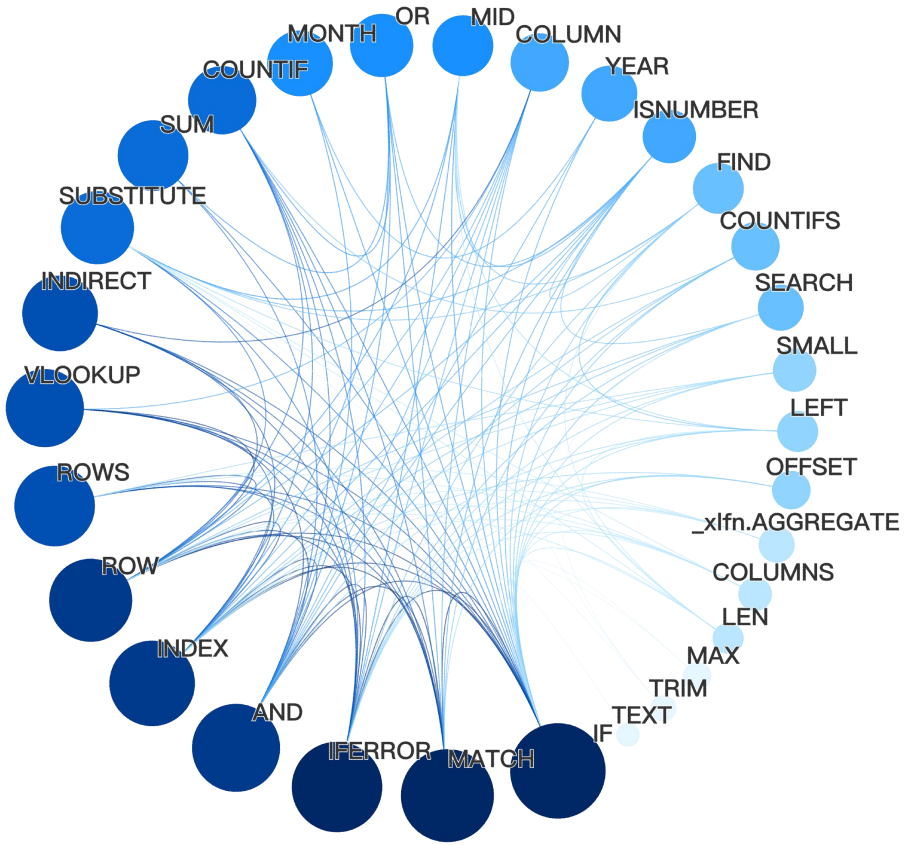
Date Collection and Preprocessing

- **Format Standardization:** Converted .xls to .xlsx, extracted cells/formulas to JSON.
- **Language Filtering:** Recognized by *lingua*
- **Deduplication:** Minhash-LSH, Jaccard similarity > 0.8.



Corpus Statistics

- **Formula Patterns:**
  Dominated by IF, SUM, VLOOKUP.
- **Workbook Level Statistics:**
  Median 270 cells, 1 sheet
- **Worksheet Level Statistics:**
  Median 300 cells, 48 rows, 10 columns.
  Max 60,000,000+ cells (Ultra Large Sheet).



Putting Sheetpedia to the Test

- **NL2SR:** Map natural language query to cell range (e.g., “Q1 sales” →A2:A4).
- **NL2Formula:** Generate Excel formula from query (e.g., “Sum Q1 sales” → =SUM(A2:A4)).
- **Why Important?**
  Simulate real-world spreadsheet interactions, requiring semantic and structural understanding.

Putting Sheetpedia to the Test

Query: Detailed Budget Situation of Cultural Festival Expenses

	A	B	C	D	E	F	G
1	Event Budget for Culture Festival						
2						Estimated	Actual
3	Total Expense					0	0
4							
5	Site	Estimated	Actual		Refreshments	Estimated	Actual
6	Room and hall fees	500	250		Food	250	130
7	Site staff	400	50		Drinks	50	220
8	Equipment	130	200		Linens	20	50
9	Tables and chairs	50	65		Staff and gratuiti	65	40
10	Total				Total		
11							
12	Decorations	Estimated	Actual		Program	Estimated	Actual
13	Flowers	200	50		Performers	70	65
14	Candles	130	130		Speakers	80	55
15	Lighting	160	160		Travel	90	35
16	Balloons	140	150		Hotel	120	100
17	Paper supplies	20	30		Other	150	140
18	Total				Total		
19							
20	Publicity	Estimated	Actual		Prizes	Estimated	Actual
21	Graphics work	45	50		Ribbons/Plaques	20	22
22	Photocopying/Print	70	77		Gifts	15	13
23	Postage	54	32		Total		
24	Total						

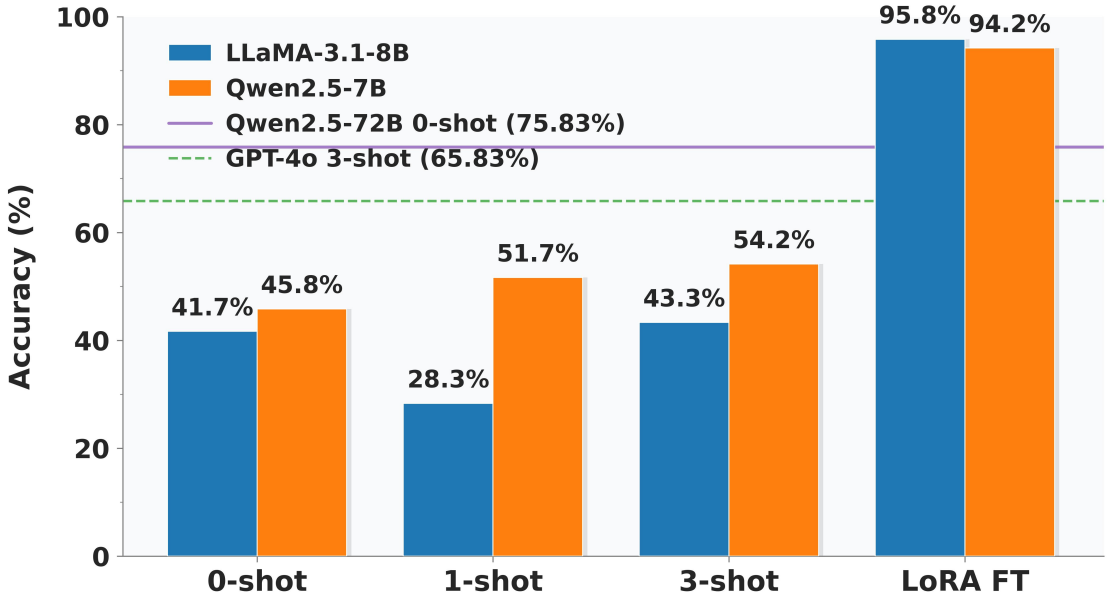
Label: B12, F12, B20, F20, B26, F25
(Yellow Regions)

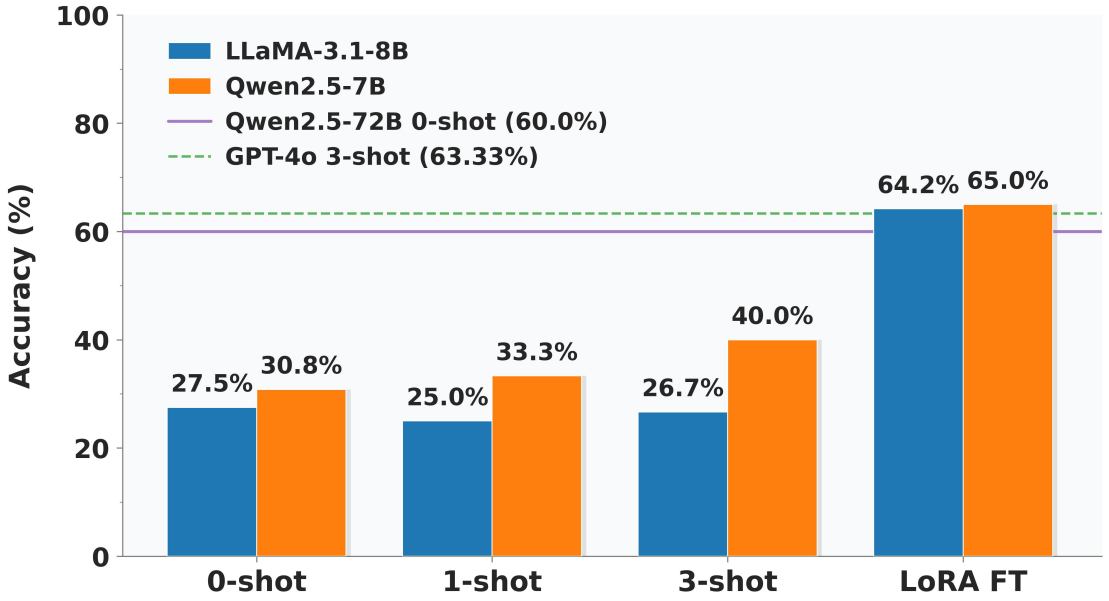
Query: Calculate the Credit enrollment for Spring 2009

	A	B	C	D	E	F	G
1	FTES = Full Time Equiv. Student	FTES Total by Subject					
3	Academic Year	AY07-08			AY08-09		
4	Credit	1,073			1,130		
5	Non Credit	10			13		
6	Total	1,083			1,144		
7	% Change vs Year Ago				6%		
8							
9	Terms	Summer07	Fall07	Spring08	Summer08	Fall08	Spring09
10	Credit	98	495	480	98	525	507
11	Non Credit	2	3	5	1	6	6
12	Total	99	498	485	99	531	514
13	% Change vs Year Ago				0%	7%	6%
14							
15	FTES = Full Time Equiv. Student	CREDIT FTES Total by Subject					
17		AY07-08			AY08-09		
18	Subject	Summer07	Fall07	Spring08	Summer08	Fall08	Spring09
19	Accounting		16.7	14.9		19.5	18.4
20	Business Administration		14.0	19.3		17.4	15.7
21	Computer Applications	3.6	12.3	10.9	0.4	15.0	17.4
22	Economics	7.3	29.5	31.7	7.0	32.3	33.2
23	Mathematics	82.1	366.1	350.4	82.7	378.6	353.2
24	Sociology	4.9	56.1	52.9	7.7	62.4	69.6

Label: =SUM(G19:G25)
(Yellow Regions)

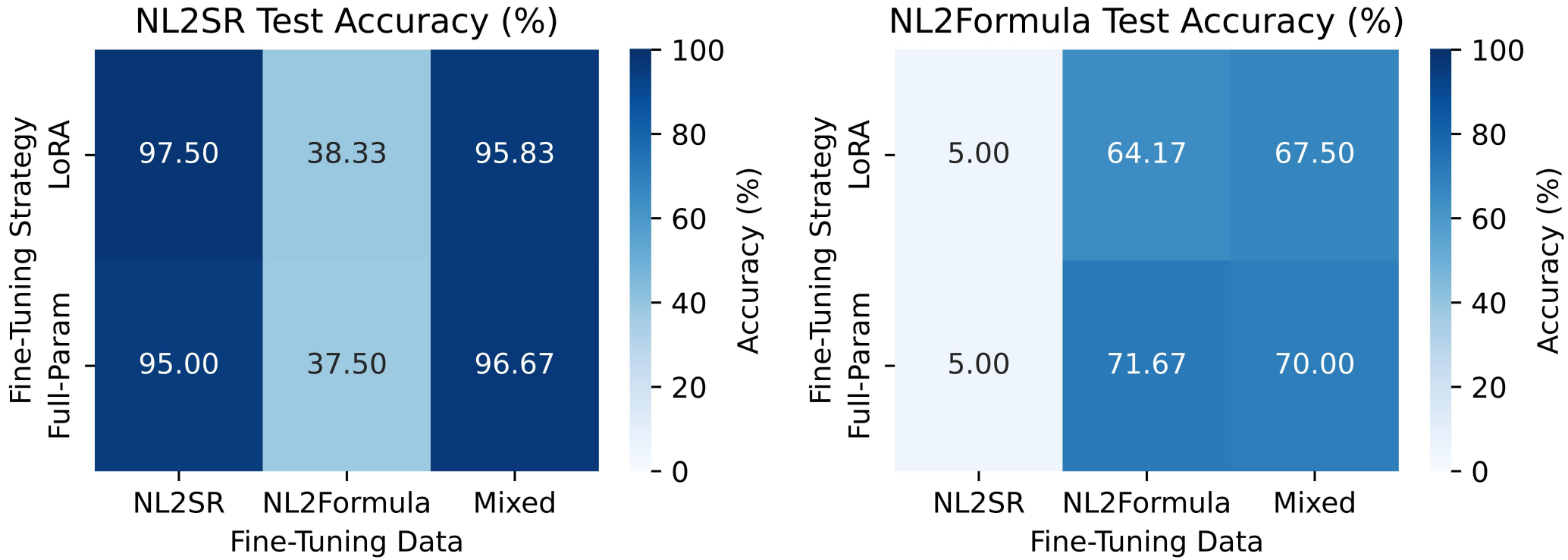
Putting Sheetpedia to the Test





- For both NL2SR and NL2Formula tasks, LoRA fine-tuning provides a substantial accuracy boost, outperforming all few-shot settings and surpassing strong baselines from larger models.

Putting Sheetpedia to the Test



- **LoRA:** Efficient, excels in single-task (97.5% NL2SR).
- **Full-Param:** Robust for multi-task (96.67% NL2SR, 70% NL2Formula).

Conclusion and Impact

- Sheetpedia provides critical infrastructure for the future of spreadsheet AI.
- It Enables:
- Next-Generation Tools: Powers spreadsheet automation, natural language interaction, and intelligent data analysis.

Advanced Research Frontiers:

- Multimodal Analysis (charts, layouts)
- Ultra-Large Spreadsheets Processing
- Retrieval-Augmented Generation (RAG) for context-aware assistance

Accessing Sheetpedia

- Sheetpedia Page:
  - https://tttiantt.github.io/Sheetpedia/
- Code:
  - https://github.com/TTtianTT/Sheetpedia
- Huggingface Dataset:
  - https://huggingface.co/datasets/tianzl66/Sheetpedia_xlsx