Table meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study

Yuan Sui et al., (2024). Table Meets LLM: Can Large Language Models Understand Structured Table Data? A Benchmark and Empirical Study. 645-654. 10.1145/3616855.3635752. WSDM'24.

Introduction

- **Table** are widely used to manage data and facilitate data analysis.

2016 Sales

Metadata

Month	Forecast	Sales	Variation
Jan 17	42,000	38,532	-3,468
Feb 17	45,000	41,934	-3,066
Mar 17	45,000	42,163	-2,837
Apr 17	45,000	43,050	-1,950
May 17	45,000	45,145	145
Jun 17	48,000	47,745	-255
Jul 17	48,000	49,623	1,623
Aug 17	48,000	52,539	4,539
Sep 17	45,000	47,324	2,324
Oct 17	45,000	44,700	-300
Nov 17	42,000	44,923	
Dec 17	48,000	51,120	
	546,000	548,798	

James:
Forecast

Data

emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Metadata

Data

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
department_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee sti

Introduction

- **Table** are widely used to manage data and facilitate data analysis.
- *Table-based applications* requires **Structured Understanding Capabilities.**

Introduction

- **Table** are widely used to manage data and facilitate data analysis.
- *Table-based applications* requires **Structured Understanding Capabilities.**

Example from TabFact

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

TFV

Introduction

- **Table** are widely used to manage data and facilitate data analysis.
- *Table-based applications* requires **Structured Understanding Capabilities.**

Example from TabFact

United States House of Representatives Elections, 1972

District	Incumbent	Party	Result	Candidates
California 3	John E. Moss	democratic	re-elected	John E. Moss (d) 69.9% John Rakus (r) 30.1%
California 5	Phillip Burton	democratic	re-elected	Phillip Burton (d) 81.8% Edlo E. Powell (r) 18.2%
California 8	George Paul Miller	democratic	lost renomination democratic hold	Pete Stark (d) 52.9% Lew M. Warden , Jr. (r) 47.1%
California 14	Jerome R. Waldie	republican	re-elected	Jerome R. Waldie (d) 77.6% Floyd E. Sims (r) 22.4%
California 15	John J. Mcfall	republican	re-elected	John J. Mcfall (d) unopposed

Entailed Statement

1. John E. Moss and Phillip Burton are both re-elected in the house of representative election.
2. John J. Mcfall is unopposed during the re-election.
3. There are three different incumbents from democratic.

Refuted Statement

1. John E. Moss and George Paul Miller are both re-elected in the house of representative election.
2. John J. Mcfall failed to be re-elected though being unopposed.
3. There are five candidates in total, two of them are democrats and three of them are republicans.

TFV

TQA

Example from HybridQA

The 2016 Summer Olympics officially known as the Games of the XXXI Olympiad (Portuguese : Jogos da XXXI Olimpíada) and commonly known as **Rio** 2016 , was an international multi-sport event ……

Name	Year	Season	Flag bearer
XXXI	2016	Summer	Yan Naing Soe
XXX	2012	Summer	Zaw Win Thet
XXIX	2008	Summer	Phone Myint Tayzar
XXVIII	2004	Summer	Hla Win U
XXVII	2000	Summer	Maung Maung Nge
XX	1972	Summer	Win Maung

Yan Naing Soe (born **31 January 1979**) is a Burmese judoka . He competed at the 2016 Summer Olympics in the **men 's 100 kg event** , …… He was the flag bearer for Myanmar at the **Parade of Nations** .

Zaw Win Thet (born **1 March 1991** in Kyonpyaw , Pathein District , Ayeyarwady Division , Myanmar) is a Burmese runner who ……

Myint Tayzar Phone (Burmese : မြင့်တေဇာဖုန်း) born **July 2 , 1978**) is a sprint canoer from Myanmar who competed in the late 2000s .

……

Win Maung (born **12 May 1949**) is a Burmese footballer . He competed in the men 's tournament at the 1972 Summer Olympics …

Q: In which year did the judoka bearer participate in the Olympic opening ceremony?	A: 2016

Motivation

Motivation

date	result	score	brazil scorers	competition
may 11 , 1919	w	6 - 0	friedenreich (3) , neco (2) , harold	american championship
may 18 , 1919	w	3 - 1	heitor , amílcar , millon	american championship
may 26 , 1919	d	2 - 2	neco (2)	american championship
may 29 , 1919	w	1 - 0	friedenreich	american championship
june 1 , 1919	d	3 - 3	haroldo , arlindo (2)	taça roberto cherry

Brazilian football in 1919

How may goals have has Brazillian team player neco scored in 1919 south american championship?

Motivation

date	result	score	brazil scorers	competition
may 11 , 1919	w	6 - 0	friedenreich (3) , neco (2) , harold	american championship
may 18 , 1919	w	3 - 1	heitor , amílcar , millon	american championship
may 26 , 1919	d	2 - 2	neco (2)	american championship
may 29 , 1919	w	1 - 0	friedenreich	american championship
june 1 , 1919	d	3 - 3	haroldo , arlindo (2)	taça roberto cherry

Brazilian football in 1919

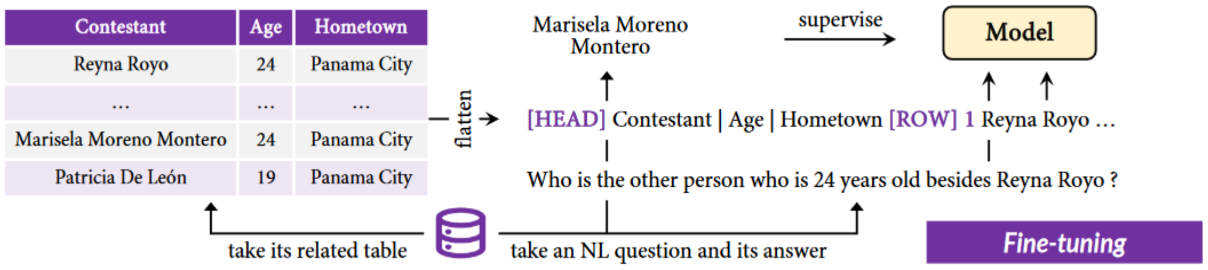
How may goals have has Brazillian team player neco scored in 1919 south american championship?

Motivation

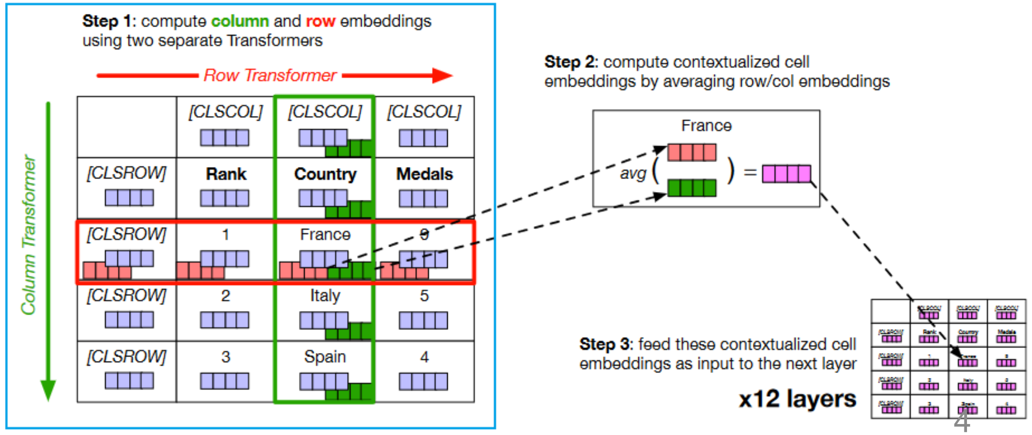
date	result	score	brazil scorers	competition
may 11 , 1919	w	6 - 0	friedenreich (3) , neco (2) , harold	american championship
may 18 , 1919	w	3 - 1	heitor , amílcar , millon	american championship
may 26 , 1919	d	2 - 2	neco (2)	american championship
may 29 , 1919	w	1 - 0	friedenreich	american championship
june 1 , 1919	d	3 - 3	haroldo , arlindo (2)	taça roberto cherry

Brazilian football in 1919

How may goals have has Brazillian team player neco scored in 1919 south american championship?



(Qian Liu et al., (2022), Tapex)



(Hiroshi Iida et al., (2021), Tabbie)

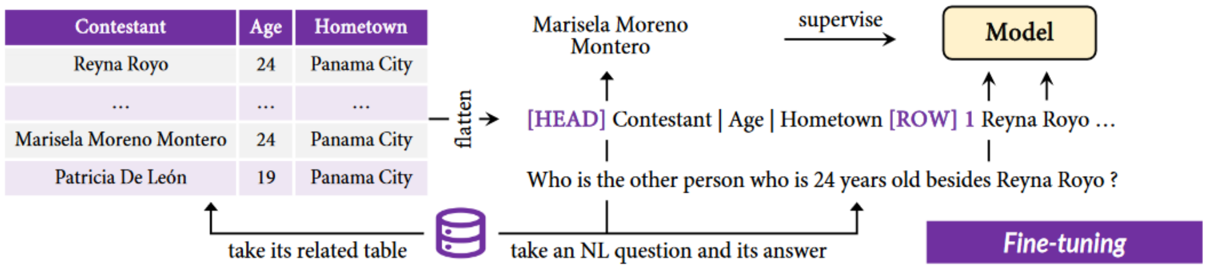
Motivation

date	result	score	brazil scorers	competition
may 11 , 1919	w	6 - 0	friedenreich (3) , neco (2) , harold	american championship
may 18 , 1919	w	3 - 1	heitor , amílcar , millon	american championship
may 26 , 1919	d	2 - 2	neco (2)	american championship
may 29 , 1919	w	1 - 0	friedenreich	american championship
june 1 , 1919	d	3 - 3	haroldo , arlindo (2)	taça roberto cherry

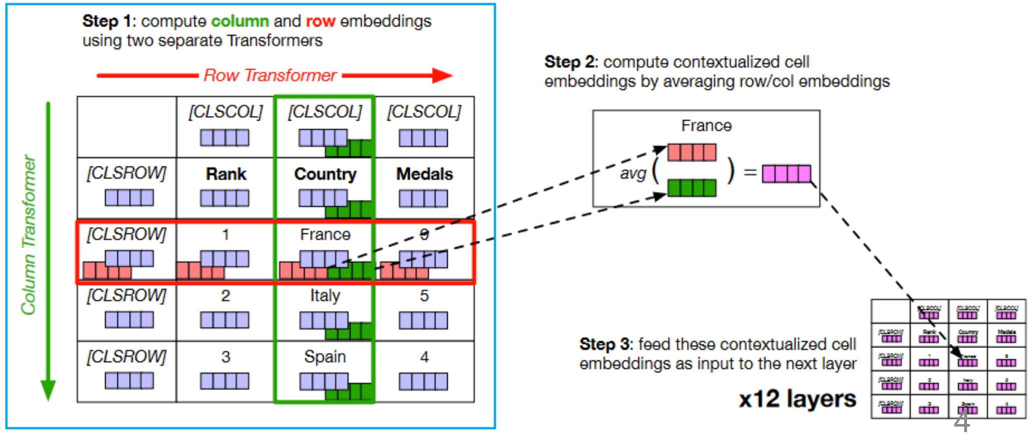
Brazilian football in 1919

How may goals have has Brazillian team player neco scored in 1919 south american championship?

- *Q1: What input designs and choices are most effective in enabling LLMs to understand tables?*



(Qian Liu et al., (2022), Tapex)



(Hiroshi Iida et al., (2021), Tabbie)

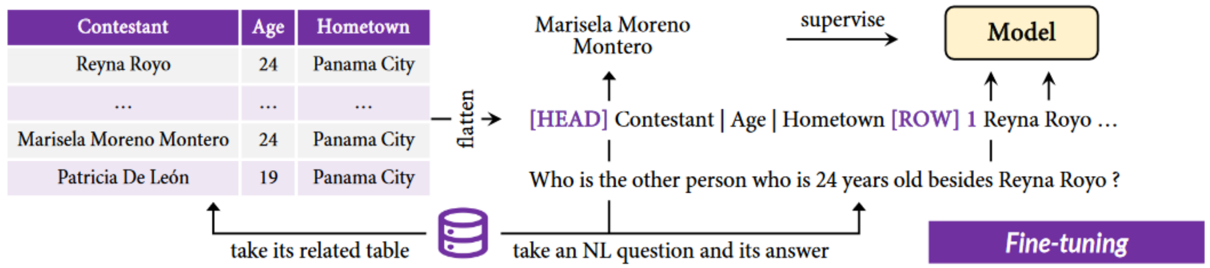
Motivation

date	result	score	brazil scorers	competition
may 11 , 1919	w	6 - 0	friedenreich (3) , neco (2) , harold	american championship
may 18 , 1919	w	3 - 1	heitor , amílcar , millon	american championship
may 26 , 1919	d	2 - 2	neco (2)	american championship
may 29 , 1919	w	1 - 0	friedenreich	american championship
june 1 , 1919	d	3 - 3	haroldo , arlindo (2)	taça roberto cherry

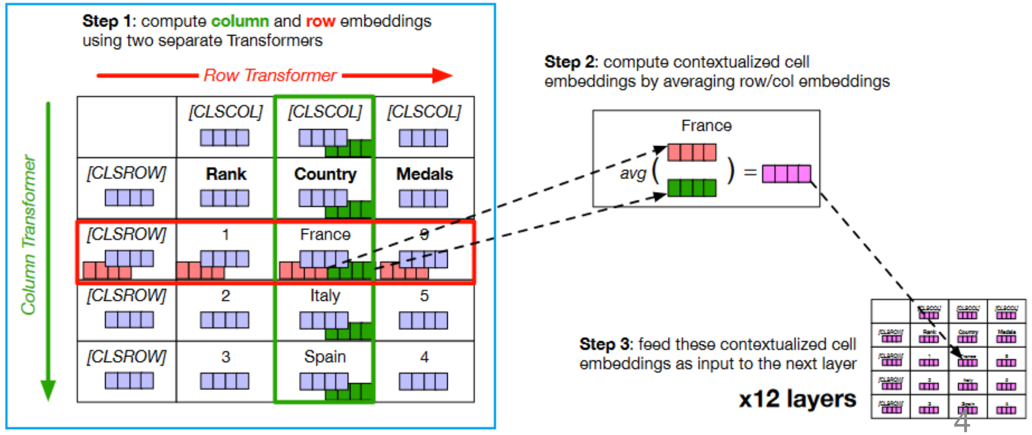
Brazilian football in 1919

How may goals have has Brazillian team player neco scored in 1919 south american championship?

- *Q1: What input designs and choices are most effective in enabling LLMs to understand tables?*
- *Q2: Do LLMs have the* ***structural understanding capabilities*** *and what extent do LLMs already have achieved in understanding structured data?*



(Qian Liu et al., (2022), Tapex)



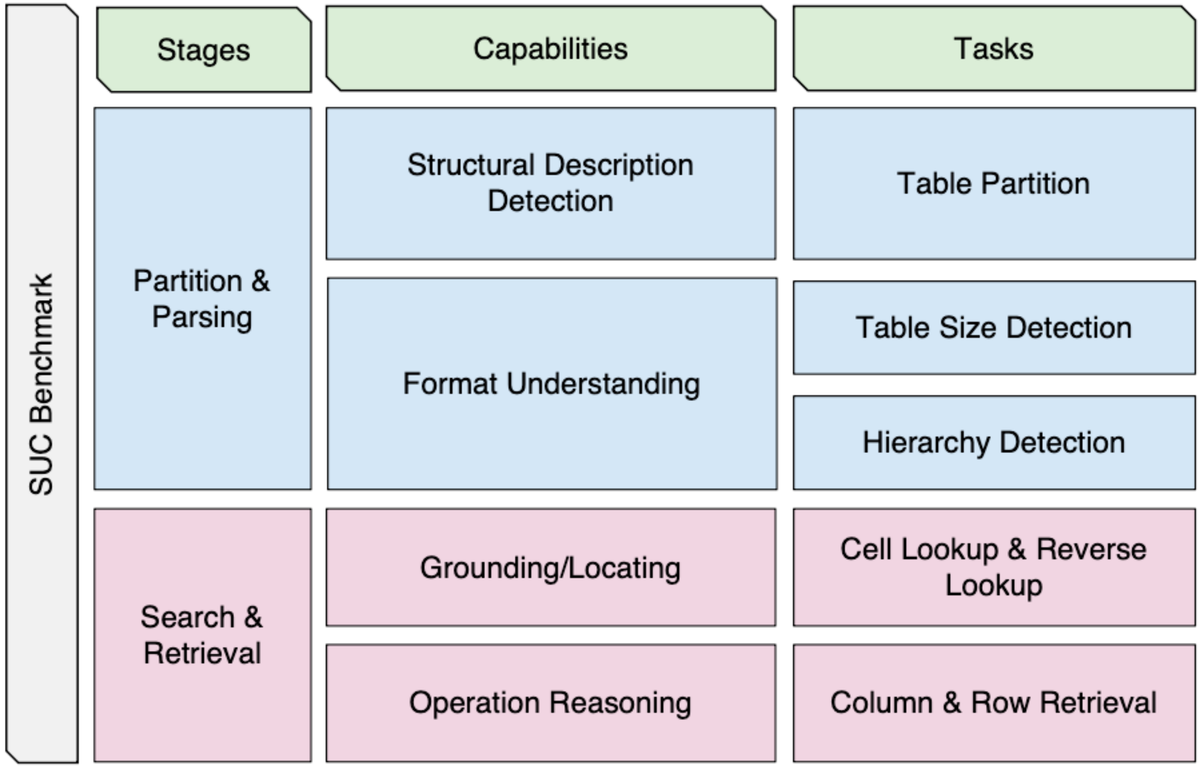
(Hiroshi Iida et al., (2021), Tabbie)

Benchmark: SUC

- *Q1: Do LLMs have the **structural understanding capabilities** and what extent do LLMs already have achieved in understanding structured data?*

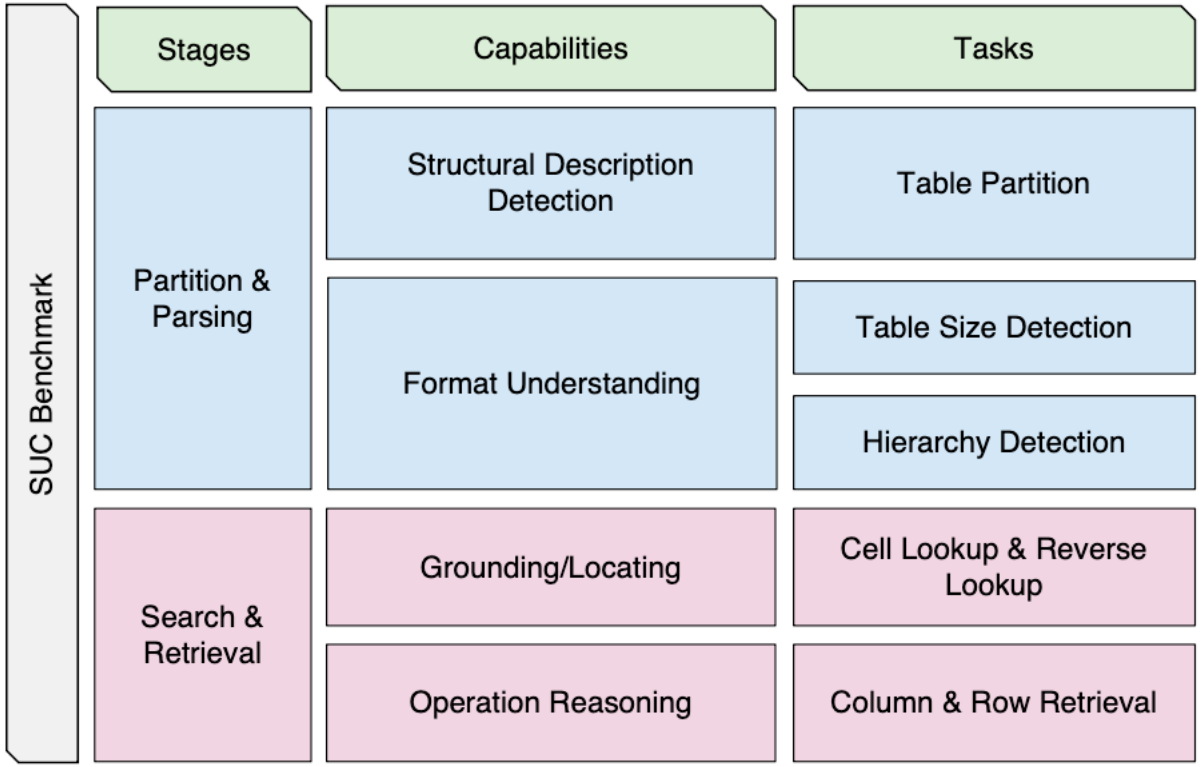
Benchmark: SUC

- *Q1: Do LLMs have the* ***structural understanding capabilities*** *and what extent do LLMs already have achieved in understanding structured data?*



Benchmark: SUC

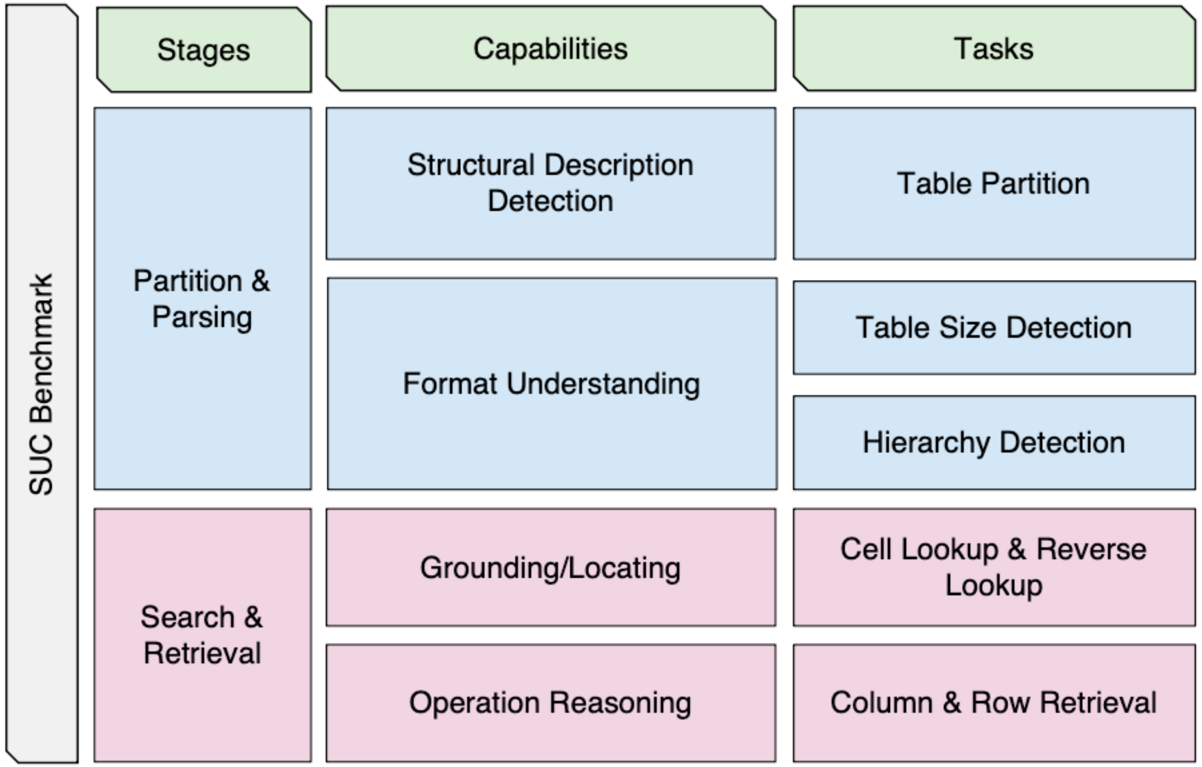
- *Q1: Do LLMs have the **structural understanding capabilities** and what extent do LLMs already have achieved in understanding structured data?*



(1) tabular dataset are always paried with knowledge from other sources to provide more context (e.g., passage, images, human annotations, etc.)

Benchmark: SUC

- *Q1: Do LLMs have the **structural understanding capabilities** and what extent do LLMs already have achieved in understanding structured data?*

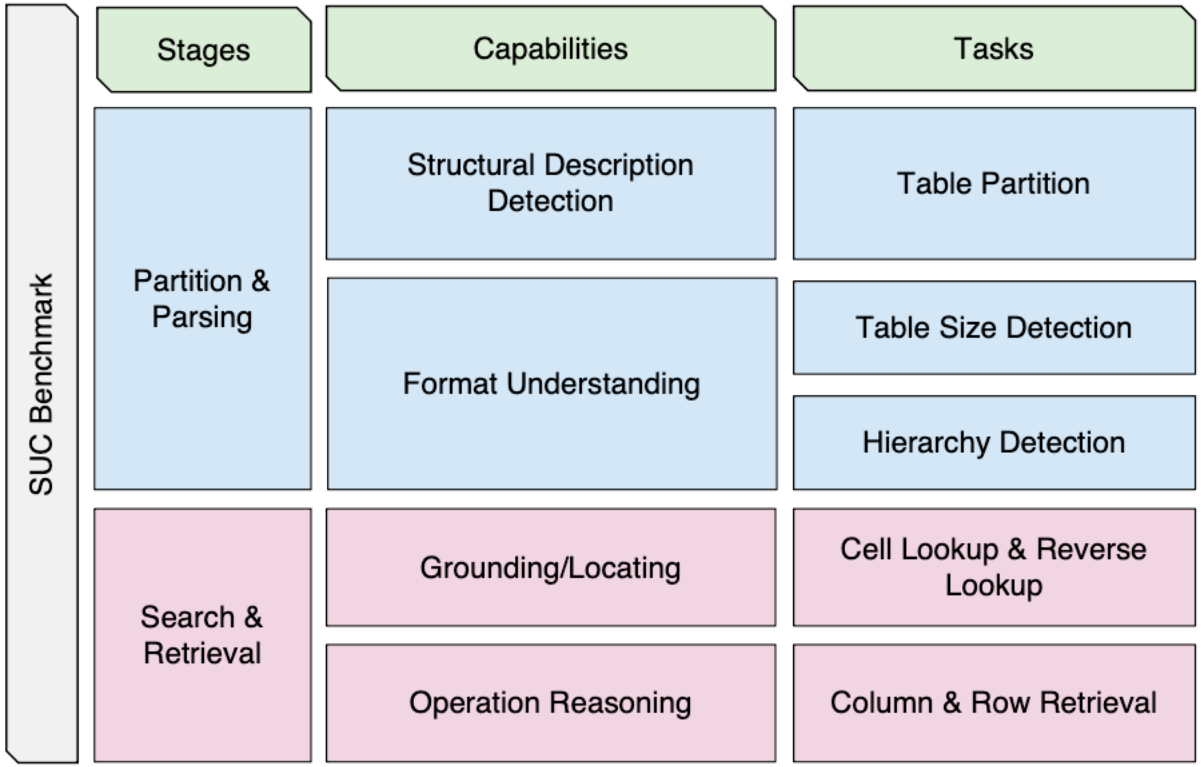


(1) tabular dataset are always paried with knowledge from other sources to provide more context (e.g., passage, images, human annotations, etc.)

(2) various table storage formats, (including csv, json, xml, html, xlsx, etc.) represent *different levels of challenges* for LLMs to understand the table content.

Benchmark: SUC

- *Q1: Do LLMs have the **structural understanding capabilities** and what extent do LLMs already have achieved in understanding structured data?*



(1) tabular dataset are always paried with knowledge from other sources to provide more context (e.g., passage, images, human annotations, etc.)

(2) various table storage formats, (including csv, json, xml, html, xlsx, etc.) represent *different levels of challenges* for LLMs to understand the table content.

(3) the ability to accurately search and retrieve information from specific positions within structured data is crucial for LLMs. The capability is highly relevant to the downstream tasks.

Benchmark: SUC

- *Q2: What input designs and choices are most effective in enabling LLMs to understand tables?*

Benchmark: SUC

- *Q2: What input designs and choices are most effective in enabling LLMs to understand tables?*

PROMPT DESIGN

Role Prompting

You are a brilliant table executor with the capabilities of table partition, table parsing, table search/retrieval, and table operation/manipulation. You can solve any tasks related to table.

Partition Mark

<title> United States House of Representatives Elections, 1972
<context> NA
<caption> District, Incumbent and Candidates Collection
<HTML grammar>

Format Explanation

Each table cell is defined by a <td> and a </td> tag. Each table row starts with a <tr> and ends with a </tr> tag. th stands for table header.

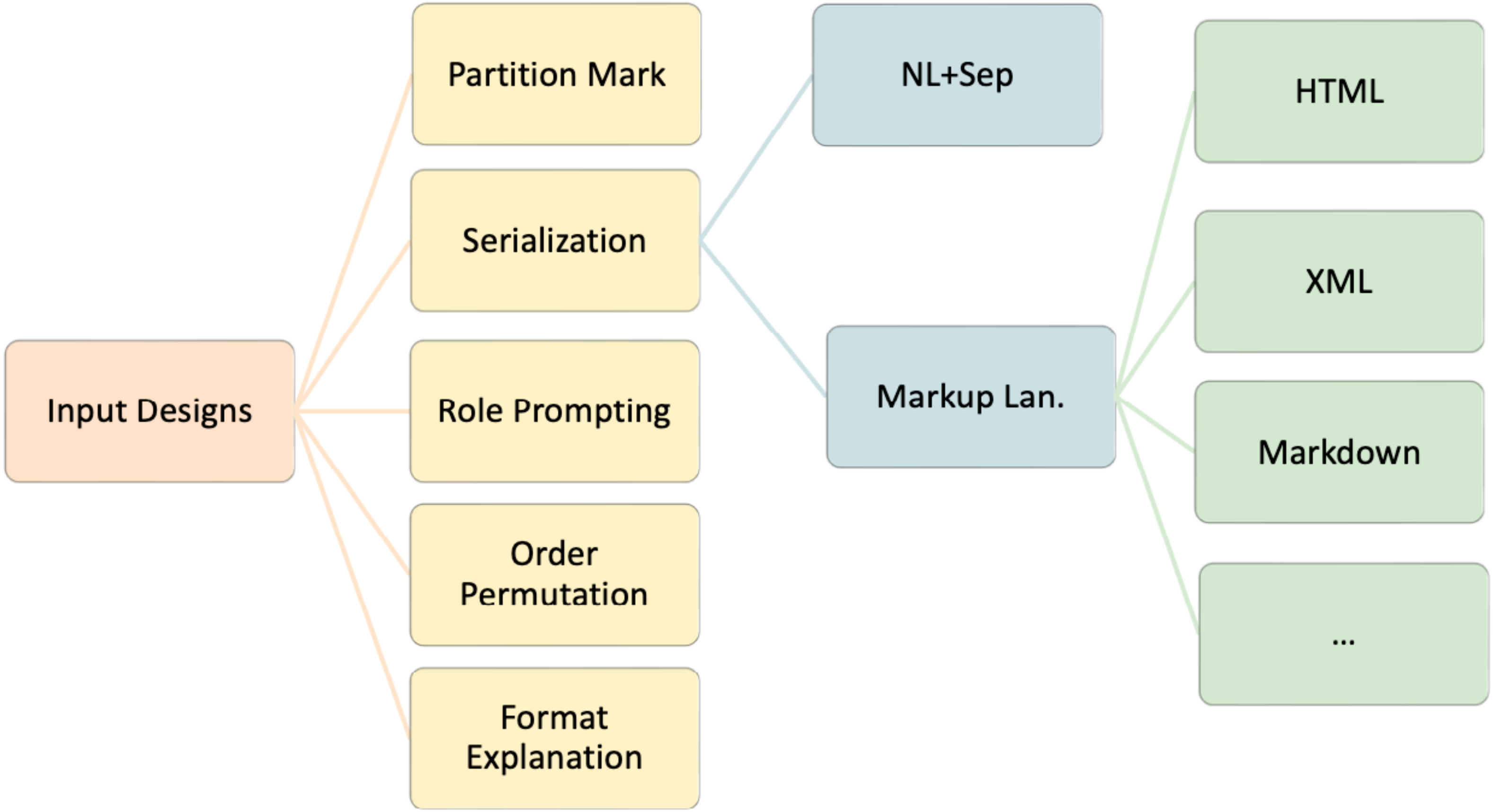
<table border="1" class="dataframe">
<thead> <tr style="text-align: right;"><th></th><th>Atlantic Division</th><th>Home</th><th>Road</th><th>Div</th></tr></thead>
<tbody><tr><th>0</th> <td>y-Philadelphia 76ers</td><td>29–12</td><td>27–14</td><td>18–6</td></tr></tr></tbody>
</table>

Order Permutation: put external text ahead of tables.

<request>
What is the position of the cell value 30? Use row index and column index to answer, e.g., 2 | 3)
The answer is

Benchmark: SUC

- *Q2: What input designs and choices are most effective in enabling LLMs to understand tables?*



Input Designs for Structural Understanding Capabilities Evaluation

Experiment Settings

- **Models.** GPT-3.5 and GPT-4. (close-sourced model); Llama-7b, 13b, PaLM-2 (open-sourced models, TBD)
- **Downstream Tasks and Datasets.** In addition to evaluate LLMs' capabilities towards understanding structured data through our benchmark. We also conduct experiments on five typical tabular downstream tasks: *SQA*, *HybridQA*, *ToTTo*, *Feverous, TabFact*.
- **Data Collection and Reformatting of SUC.**
    - only consider the structural portions of the original datasets, which are labeled with “table”, “rows”, or “headers”, exclude other parts, “ID”, “Answer”.
    - to identify a specific value within the structured data, we append each parsed sample with a unique question. e.g., “How many rows in the table? How many columns in the table?” Each question is accompanied by a set of reference answers ("groundtruth") sourced from the original datasets.
    - we evaluate these questions using Text-Davinci-03 and manually eliminate any question that the model consistently answers correctly when multiple random samples are generated at a nonzero temperature.

Task	Input
Table Partition	What is the first token (cell value instead of separator |) of the given table? What is the end token (cell value instead of separator |) of the given table? Answer questions one by one and use | to split the answer.
Cell Lookup	What is the position of the cell value cell_value? Use row index and column index to answer
Reverse Lookup	What is the cell value of row index, column index ? Only output the cell value without other information
Column Retrieval	What is the column name with the index column_idx of the following table? Only give the column name without any explanation
Row Retrieval	What are the cell values of the row_idx row in following table? Only list the cell values one by one using | to split the answers
Size Detection	How many rows in the table? How many columns in the table. Answer the questions one by one and use | to split the answer
Merged Cell Detection	What is the column index of the cell which span is over 1. use | to split the answer (e.g., 3 | 4), the column index starts from 0. If there's no answer, return None

Experiments: Benchmark

Table 1: Micro results of the benchmark (See full results from Table 6). Change order [37] refers to put external text (like questions, statement) ahead of tables. Noted that "GPT-4" refers to the evaluation outcomes utilizing the GPT-4 model. Given the resource-intensive nature of GPT-4 calls, we only conducting the GPT-4 inference test on a subset of 300 samples (randomly sampled) from each task set. Each column follows the roles of graded color scale, *i.e.*, the deeper color refers to better perf.

Format	Table Partition		Cell Lookup		Reverse Lookup		Column Retrieval		Row Retrieval		Size Detection		Merged Cell Detection	
	Acc	GPT-4	Acc	GPT-4	Acc	GPT-4	Acc	GPT-4	Acc	GPT-4	Acc	GPT-4	Acc	GPT-4
NL + Sep	93.00%	96.78%	39.67%	72.48%	52.00%	59.12%	60.67%	66.32%	31.00%	48.67%	42.00%	73.12%	71.33%	74.98%
Markdown	92.33%	**98.32%**	43.33%	71.93%	51.00%	57.32%	35.33%	60.12%	**42.33%**	49.98%	40.67%	82.12%	**78.00%**	**82.64%**
JSON	94.00%	97.12%	42.67%	68.32%	54.33%	58.12%	54.33%	64.32%	29.00%	48.32%	42.67%	76.43%	73.33%	78.98%
XML	96.00%	97.64%	43.33%	72.28%	**55.00%**	**60.32%**	41.33%	68.28%	41.00%	**50.28%**	43.67%	80.21%	75.00%	80.32%
HTML	**96.67%**	**98.32%**	**44.00%**	**73.34%**	47.33%	59.45%	**63.33%**	**69.32%**	42.00%	50.19%	**67.00%**	**83.43%**	76.67%	81.28%

Highlights:

- LLMs achieves the highest overall accuracy 65.43% overall seven tasks when using ***HTML***, indicating that LLM has significant potential for understanding the structural information of tables in this specific format.
- Compared to the commonly used format "NL+Sep", ***hierarchy structure*** is essential. This may be due to the language models being fine-tuned on code and the training data containing substantial amounts of web data, while most tabular data is sourced from web pages such as Wikipedia.

Experiments: Benchmark

Table 2: Micro ablation results of the input designs over benchmark. Find more detailed ablation results from Table 6

Input Design	Table Partition		Cell Lookup		Reverse Lookup		Column Retrieval		Row Retrieval		Size Detection		Merged Cell Detection	
	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
Markup Lan. HTML	**96.67%**	0.00%	44.00%	0.00%	47.33%	0.00%	63.33%	0.00%	42.00%	0.00%	67.00%	0.00%	76.67%	0.00%
w/o format explanation	92.00%	-4.67%	52.00%	8.00%	52.33%	5.00%	64.33%	1.00%	36.00%	-6.00%	**78.00%**	11.00%	**77.67%**	1.00%
w/o partition mark	**98.00%**	1.33%	**59.00%**	15.00%	53.00%	5.67%	66.00%	2.67%	39.67%	-2.33%	**72.00%**	5.00%	70.33%	-6.33%
w/o role prompting	95.00%	3.00%	40.67%	-11.33%	44.67%	-7.67%	59.00%	-5.33%	39.33%	3.33%	**69.00%**	-9.00%	76.00%	-1.67%
w/o change order	**96.67%**	0.00%	52.33%	8.33%	40.67%	-6.67%	55.67%	-7.67%	31.67%	-10.33%	52.67%	-14.33%	65.67%	-11.00%
w/o 1-shot	63.00%	-33.67%	9.33%	-34.67%	17.33%	-30.00%	50.00%	-13.33%	30.00%	-12.00%	16.67%	-50.33%	38.00%	-38.67%
GPT-4 w/ Lan. HTML	**98.32%**	1.65%	73.34%	29.34%	59.45%	12.12%	69.32%	5.99%	50.19%	8.19%	83.43%	16.43%	81.28%	4.61%

Experiments: Benchmark

Table 2: Micro ablation results of the input designs over benchmark. Find more detailed ablation results from Table 6

Input Design	Table Partition		Cell Lookup		Reverse Lookup		Column Retrieval		Row Retrieval		Size Detection		Merged Cell Detection	
	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
Markup Lan. HTML	**96.67%**	0.00%	44.00%	0.00%	47.33%	0.00%	63.33%	0.00%	42.00%	0.00%	67.00%	0.00%	76.67%	0.00%
w/o format explanation	92.00%	-4.67%	52.00%	8.00%	52.33%	5.00%	64.33%	1.00%	36.00%	-6.00%	**78.00%**	11.00%	**77.67%**	1.00%
w/o partition mark	**98.00%**	1.33%	**59.00%**	15.00%	53.00%	5.67%	66.00%	2.67%	39.67%	-2.33%	**72.00%**	5.00%	70.33%	-6.33%
w/o role prompting	95.00%	3.00%	40.67%	-11.33%	44.67%	-7.67%	59.00%	-5.33%	39.33%	3.33%	**69.00%**	-9.00%	76.00%	-1.67%
w/o change order	**96.67%**	0.00%	52.33%	8.33%	40.67%	-6.67%	55.67%	-7.67%	31.67%	-10.33%	52.67%	-14.33%	65.67%	-11.00%
w/o 1-shot	63.00%	-33.67%	9.33%	-34.67%	17.33%	-30.00%	50.00%	-13.33%	30.00%	-12.00%	16.67%	-50.33%	38.00%	-38.67%
GPT-4 w/ Lan. HTML	**98.32%**	1.65%	73.34%	29.34%	59.45%	12.12%	69.32%	5.99%	50.19%	8.19%	83.43%	16.43%	81.28%	4.61%

Highlights:

- External information should appear ahead of tables.
- Partition mark. & format explanation may undermine Search & Retrieval capability.

Experiments: Benchmark

Table 2: Micro ablation results of the input designs over benchmark. Find more detailed ablation results from Table 6

Input Design	Table Partition		Cell Lookup		Reverse Lookup		Column Retrieval		Row Retrieval		Size Detection		Merged Cell Detection	
	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ	Acc	Δ
Markup Lan. HTML	**96.67%**	0.00%	44.00%	0.00%	47.33%	0.00%	63.33%	0.00%	42.00%	0.00%	67.00%	0.00%	76.67%	0.00%
w/o format explanation	92.00%	-4.67%	52.00%	8.00%	52.33%	5.00%	64.33%	1.00%	36.00%	-6.00%	**78.00%**	11.00%	**77.67%**	1.00%
w/o partition mark	**98.00%**	1.33%	**59.00%**	15.00%	53.00%	5.67%	66.00%	2.67%	39.67%	-2.33%	**72.00%**	5.00%	70.33%	-6.33%
w/o role prompting	95.00%	3.00%	40.67%	-11.33%	44.67%	-7.67%	59.00%	-5.33%	39.33%	3.33%	**69.00%**	-9.00%	76.00%	-1.67%
w/o change order	**96.67%**	0.00%	52.33%	8.33%	40.67%	-6.67%	55.67%	-7.67%	31.67%	-10.33%	52.67%	-14.33%	65.67%	-11.00%
w/o 1-shot	63.00%	-33.67%	9.33%	-34.67%	17.33%	-30.00%	50.00%	-13.33%	30.00%	-12.00%	16.67%	-50.33%	38.00%	-38.67%
GPT-4 w/ Lan. HTML	**98.32%**	1.65%	73.34%	29.34%	59.45%	12.12%	69.32%	5.99%	50.19%	8.19%	83.43%	16.43%	81.28%	4.61%

Highlights:

- External information should appear ahead of tables.
- Partition mark. & format explanation may undermine Search & Retrieval capability.

Based on the highlights, ***guidelines*** are proposed to answer the questions:

- LLMs have basic structural understanding capabilities, but far from perfect, even for some trivial tasks, e.g., table size detection;
- Correctly choosing the combination of input designs is a potential factor in improving the performance of LLMs over structured data.

Experiments: Downstream Tasks

Table 3: Main results of the downstream tasks ablation study

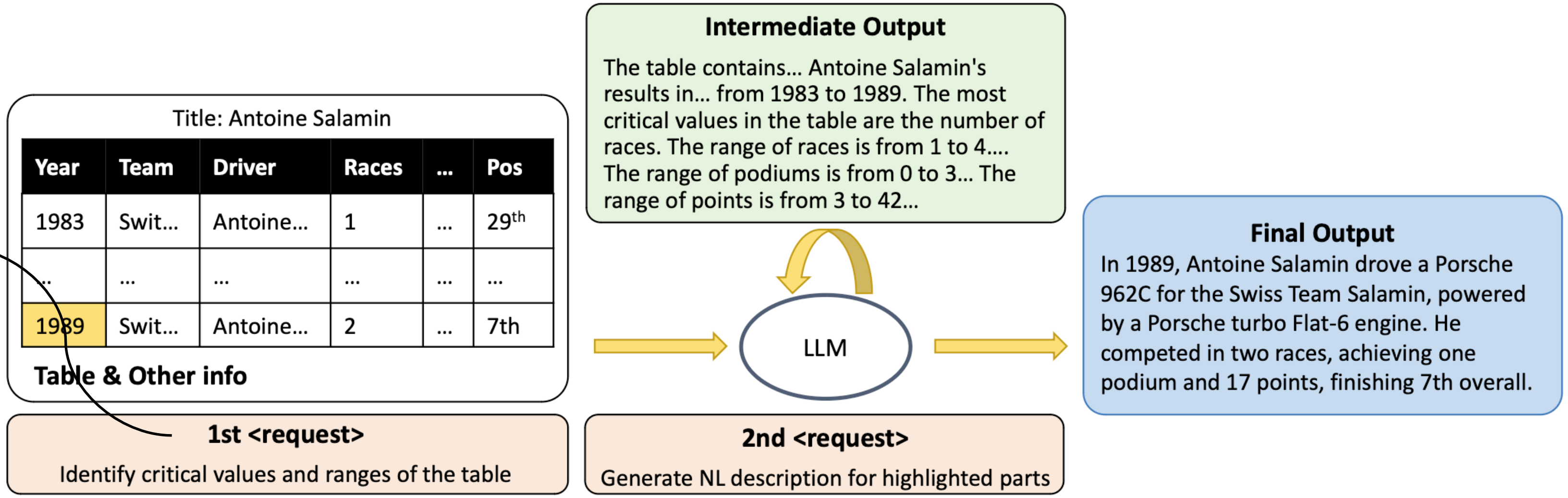
Format	TabFact	HybridQA	SQA	Feverous	ToTTo
	Acc	Acc	Acc	Acc	BLEU-4
NL + Sep	70.26%	45.02%	70.41%	75.15%	**12.70%**
Markdown	68.40%	45.88%	66.59%	71.88%	8.57%
JSON	68.04%	42.40%	70.39%	73.84%	8.82%
XML	70.00%	47.20%	70.74%	73.14%	8.82%
HTML	**71.33%**	**47.29%**	**71.31%**	**75.20%**	12.30%
GPT-4 w/ HTML	78.40%	56.68%	75.35%	83.21%	20.12%

Observations on downstream tasks are the same!

Type	Choice	TabFact	HybridQA	SQA	Feverous	ToTTo			
		Acc	Acc	Acc	Acc	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1-shot	1-shot	72.04%	46.07%	73.81%	75.56%	72.43%	44.36%	27.01%	17.24%
1-shot	w/o table size	71.33%	45.52%	72.91%	74.66%	72.30%	44.23%	27.14%	17.25%
1-shot	w/o partition mark	71.25%	45.48%	73.09%	75.11%	71.18%	43.17%	26.36%	16.34%
1-shot	w/o format explanation	70.87%	45.39%	71.69%	75.97%	70.54%	43.59%	26.52%	16.74%
1-shot	w/o role prompting	71.35%	46.05%	73.39%	75.52%	70.61%	43.10%	26.02%	16.15%

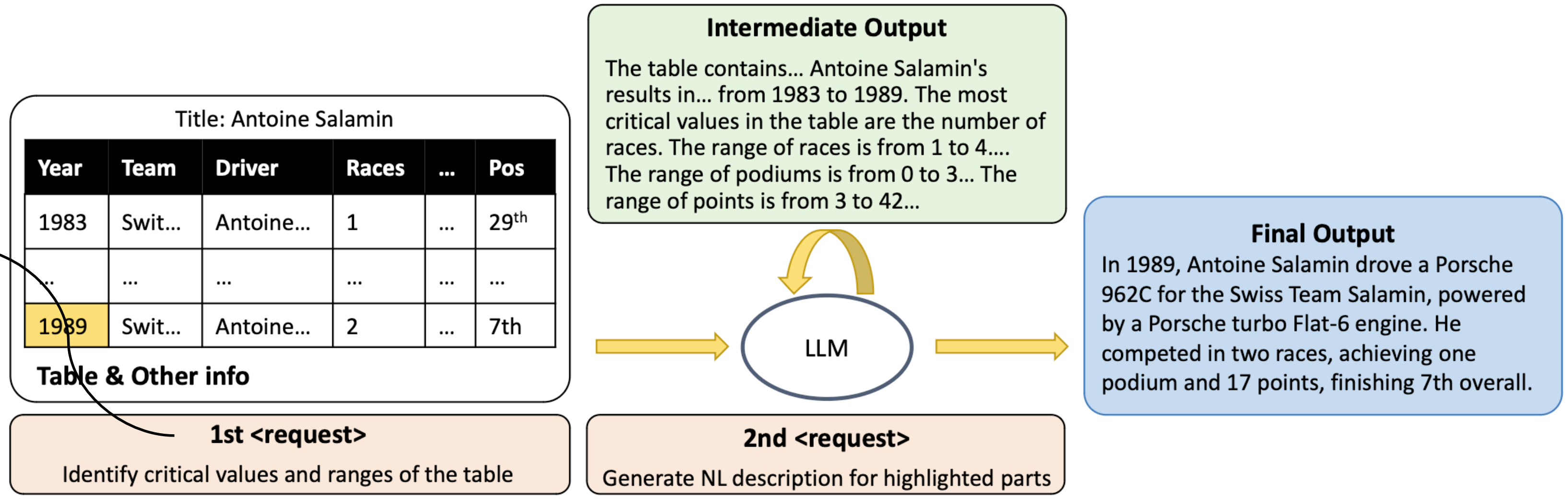
Method: Self-augmented Prompting

Choice	Prompt Design
self format explanation	Generate short format specification and description of the last {data_type} within five sentences.
self key range and values identification	Identify critical values and ranges of the last {data_type} related to the {context_type} within five sentences
self structural information description	Describe structural information, patterns and statistics of the last {data_type} related to the {context_type} within five sentences.



Method: Self-augmented Prompting

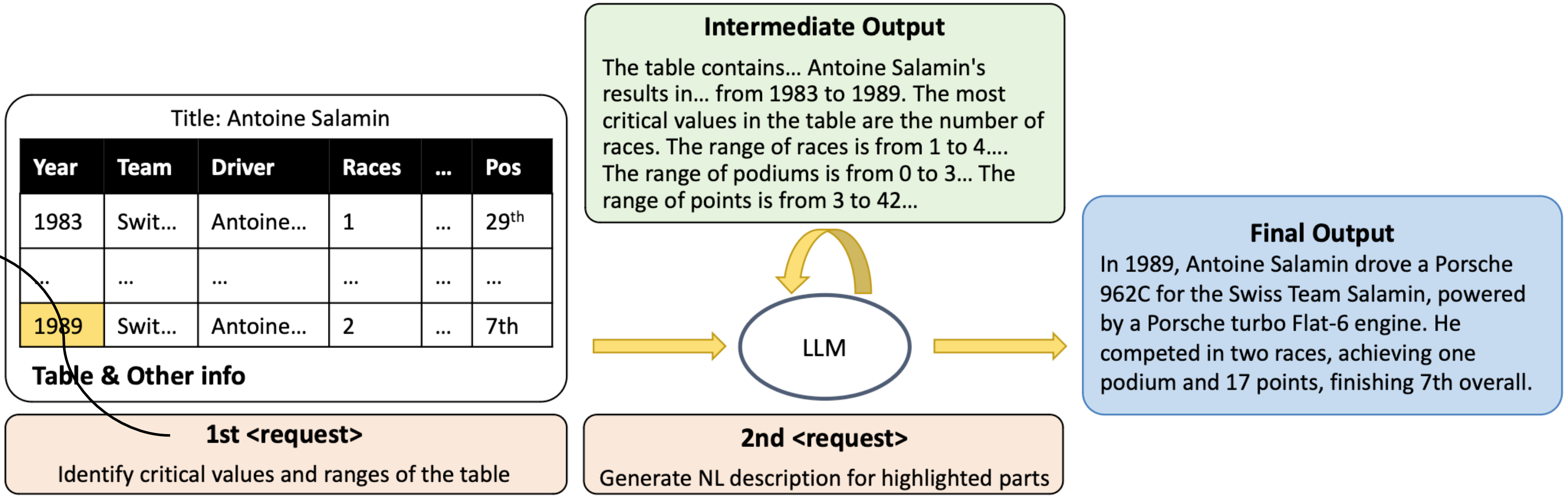
Choice	Prompt Design
self format explanation	Generate short format specification and description of the last {data_type} within five sentences.
self key range and values identification	Identify critical values and ranges of the last {data_type} related to the {context_type} within five sentences
self structural information description	Describe structural information, patterns and statistics of the last {data_type} related to the {context_type} within five sentences.



- use self-augmented prompt to ask LLM to generate additional knowledge (intermediate output) about this table;
- add the self-augmented response to form the second prompt to ask for final answer of a downstream task.
- the LLM can tell some important values in the table which help itself generate a better answer for the downstream task.

Method: Self-augmented Prompting

Choice	Prompt Design
self format explanation	Generate short format specification and description of the last {data_type} within five sentences.
self key range and values identification	Identify critical values and ranges of the last {data_type} related to the {context_type} within five sentences
self structural information description	Describe structural information, patterns and statistics of the last {data_type} related to the {context_type} within five sentences.



- use self-augmented prompt to ask LLM to generate additional knowledge (intermediate output) about this table;
- add the self-augmented response to form the second prompt to ask for final answer of a downstream task.
- the LLM can tell some important values in the table which help itself generate a better answer for the downstream task.

Type	Choice	TabFact	HybridQA	SQA	Feverous	ToTTo			
		Acc	Acc	Acc	Acc	BLEU-1	BLEU-2	BLEU-3	BLEU-4
SA	self format explanation	72.23%	46.12%	73.91%	76.15%	74.18%	45.25%	27.32%	18.34%
SA	self critical values and ranges identification	74.35%	48.20%	76.53%	76.32%	80.83%	47.96%	30.68%	22.92%
SA	self structural information description	73.42%	46.97%	75.97%	77.28%	78.93%	46.91%	28.94%	19.32%

Wrapup

- In this work, we propose a benchmark to compare various input designs in order to study the structural understanding capabilities of LLMs on tables.
- Suprisingly, we obtain some insights of the input designs and the comparison reveal that LLMs have the basic capabilities towards understanding structural information of tables.
- We also give some guidance on how to apply our benchmark insights on downstream tasks and propose a simple, generic but effective method, i.e., self-augmented prompting, by generating additional knowledge with LLMs self-knowledge.
- We believe this study will be beneficial for table-based, even structured data based research, or serve as a auxiliary tool to help better understand the table(s) from structural perspectives.