

Rethinking Tabular Data Understanding with Large Language Models

NAACL 2024

Tianyang Liu, Fei Wang, Muhao Chen

UC San Diego, USC, UC Davis

Background

- LLMs confront challenge for reasoning tabular data**

- Linearizing tables
 - Suitable for the LLM paradigm
 - It can obscure the inherent structural and relational information
 - Underperform semantic reasoning capabilities
- Data Type
 - Interpretation process for header, column, row
 - Numerical reasoning and aggregation over tabular data

Table

Name	Height	Weight
Tom	5 feet 7	130 lbs
Jane	4 feet 9	90 lbs
Mike	6 feet 1	150 lbs
Lana	5 feet 6	123 lbs

Linearizing Table

```

<s> | Name | Height | Weight |<0x0A>|---|---|---|<0x0A>| Tom |
5 feet 7 | 130 lbs |<0x0A>| Jane | 4 feet 9 | 90 lbs |<0x0A>|
Mike | 6 feet 1 | 150 lbs |<0x0A>| Lana | 5 feet 6 | 123 lbs |

91, 4076, 765, 22147, 765, 16923, 9432, 91, 4521, 91, 4521, 91, 4521, 7511, 91,
8529, 765, 220, 20, 7693, 220, 22, 765, 220, 5894, 29160, 9432, 91, 22195,
765, 220, 19, 7693, 220, 24, 765, 220, 1954, 29160, 9432, 91, 11519, 765,
220, 21, 7693, 220, 16, 765, 220, 3965, 29160, 9432, 91, 95887, 765, 220,
20, 7693, 220, 21, 765, 220, 4513, 29160, 765
  
```

Background

- **LLMs confront challenge for reasoning tabular data**
 - Challenges faced by LLMs in comprehending table structures

Marek Plawgo

Year	1999	2000	...	2008	2012
Competition	European Junior Championships	World Junior Championships	...	Olympic Games	European Championships
Venue	Riga, Latvia	Santiago, Chile	...	Beijing, China	Helsinki, Finland
Position	4th	1st	...	7th	18th (sf)
Event	400 m hurdles	400 m hurdles	...	4x400 m relay	400 m hurdles
Notes	52.17	49.23	...	3:00.32	50.77

Table Header

LLMs

What are the headings of table?



The headings of the table are:
Year, Competition, Venue,
Position, Event, and Notes.

Are they located in first row
or first column?




The headings are located in
the **first row** of the table.

Codellama-34b-instruct

What's the headings of the table?

The headings of the table are:

1. Year
2. Competition
3. Venue
4. Position
5. Event
6. Notes

Column Headings

A	B	C	D

Row Headings

1			
2			
3			
4			
5			

Are they located in first row or first column?

The headings are located in the **first row** of the table.

Background

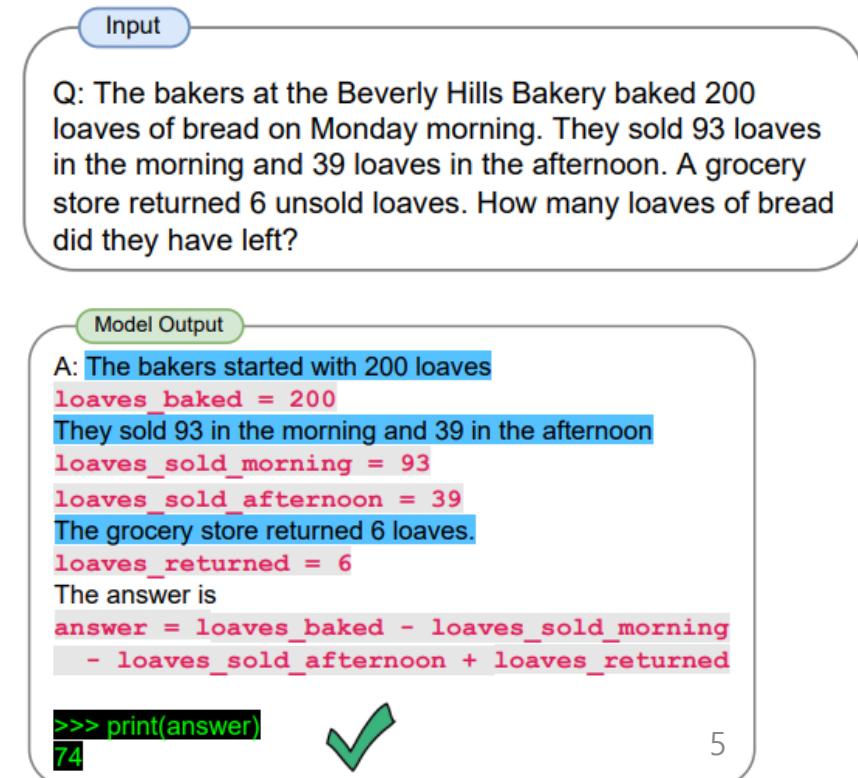
- **Pivotal Research Questions**
 - (1) How well do LLMs **perceive table structures** and how can we ensure robustness against structural variations?
 - (2) **Comparing textual and symbolic reasoning** for table data in LLMs, which prevails in effectiveness, and what advantages and challenges manifest in each strategy?
 - Textual Reasoning via Direct Prompting
 - Symbolic Reasoning via Python Shell Interactions
 - (3) Will the **aggregation of multiple reasoning pathways** enhance the accuracy and reliability of tabular data interpretation by LLMs?

Method

- **Each Reasoning Strategy**

- Textual/Generic Reasoning: Direct Prompting (DP)
 - Directly give the answer
 - COT(Chaing-of-thought): Let's think step by step and then give the answer
 - Better instructions for improvement
- Symbolic Reasoning: Python Shell Agent (PyAgent)
 - Program-aided LLM (e.g., Text-to-SQL)
 - Table QA: Use Python Library Pandas
 - ReAct Prompting (e.g., LangChain)
 - More complex framework

Program-aided Language models (this work)



Method

• Symbolic Reasoning: Python Shell Agent (PyAgent)

- Program-aided LLM
- Related Work for Table QA
 - Text-to-SQL

SQL generation from DIN-SQL(Pourreza et al, 2023)

Q: "Find the total budgets of the Marketing or Finance department."

Schema_links: [department.budget, department.dept_name, Marketing, Finance]

A: Let's think step by step. For creating the SQL for the given question, we need to join these tables = [].

First, create an intermediate representation, then use it to construct the SQL query.

Intermediate_representation: select sum(department.budget) from department where department.dept_name = "Marketing" or department.dept_name = "Finance"

SQL: `SELECT sum(budget) FROM department WHERE dept_name = 'Marketing' OR dept_name = 'Finance'`

DIN-SQL Generation Module

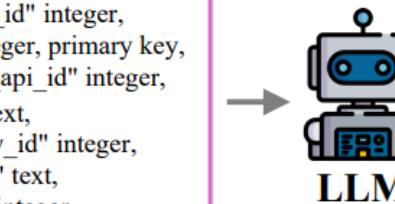
- Schema Linking
- Classification & Decomposition
- SQL Generation

User Question

Could you tell me the names of the 5 leagues with the highest matches of all time and how many matches were played in the said league?

Schema

```
TABLE Country
TABLE League
TABLE Match
{"league_id" integer,
 "id" integer, primary key,
 "match_api_id" integer,
 "date" text,
 "country_id" integer,
 "season" text,
 "stage" integer,
 "away_player_1" integer,
 "possession" text,
 "goal" text,
 primary key("id")}
```



Execution Results

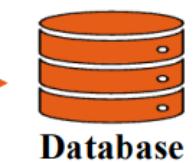
Match	League
3040	Spain LIGA BBVA
3040	France Ligue 1
3040	England Premier League
3017	Italy Serie A
2448	Netherlands Eredivisie

Generated SQL Query

```
SELECT League.name, count(Match.id) FROM Match INNER
JOIN League ON Match.league_id = league.id GROUP BY
League.name ORDER BY count(Match.id) DESC LIMIT 5
```



User



6

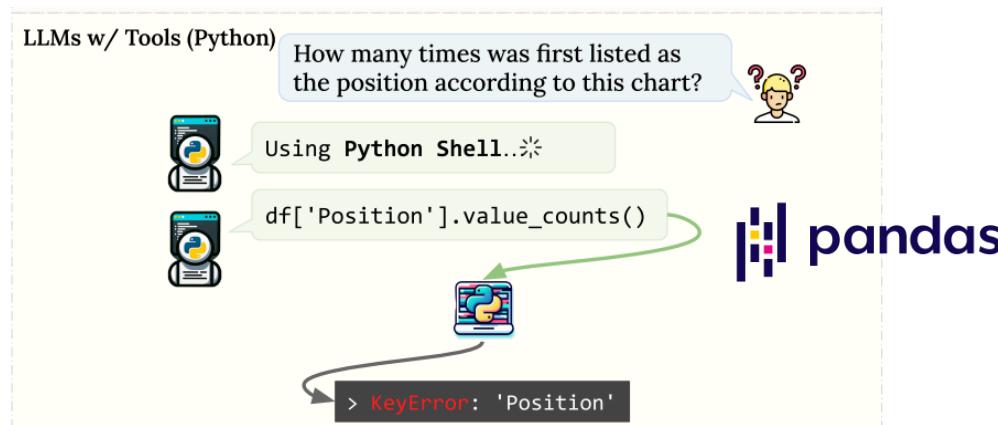
Method

- **Symbolic Reasoning: Python Shell Agent (PyAgent)**
 - Symbolic Reasoning

Marek Plawgo

Year	1999	2000	...	2008	2012
Competition	European Junior Championships	World Junior Championships	...	Olympic Games	European Championships
Venue	Riga, Latvia	Santiago, Chile	...	Beijing, China	Helsinki, Finland
Position	4th	1st	...	7th	18th (sf)
Event	400 m hurdles	400 m hurdles	...	4x400 m relay	400 m hurdles
Notes	52.17	49.23	...	3:00.32	50.77

Python Shell as an external tool



**Symbolic Reasoning
from Chain-of-thought prompt(Wei et al, 2022)**

Last Letter Concatenation

Q: Take the last letters of the words in “Lady Gaga” and concatenate them.

A: The last letter of “Lady” is “y”. The last letter of “Gaga” is “a”. Concatenating them is “ya”. So the answer is ya.

Coin Flip (state tracking)

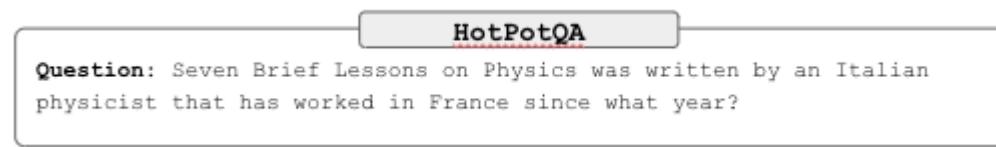
Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

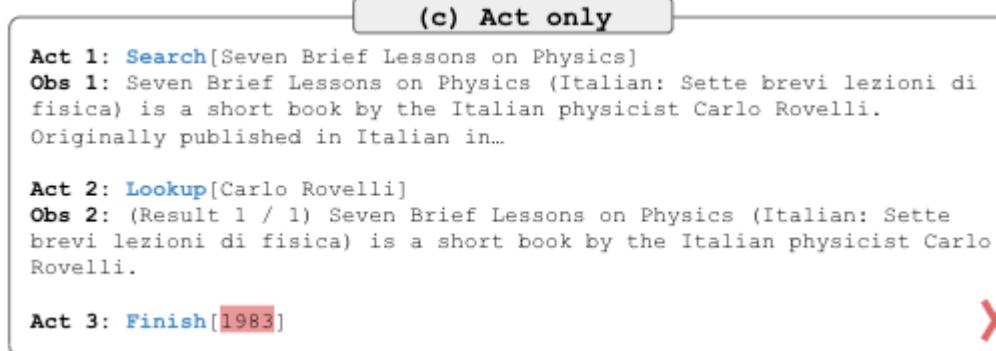
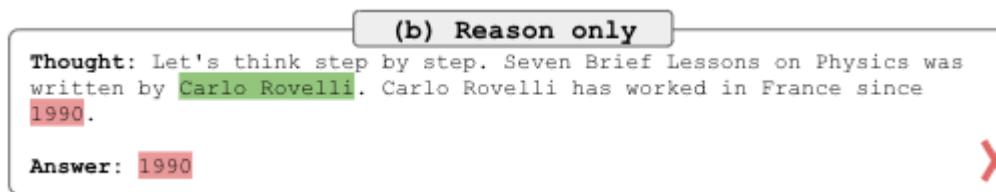
Method

• Symbolic Reasoning: Python Shell Agent (PyAgent)

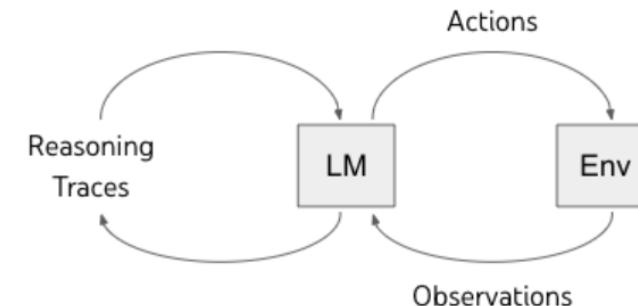
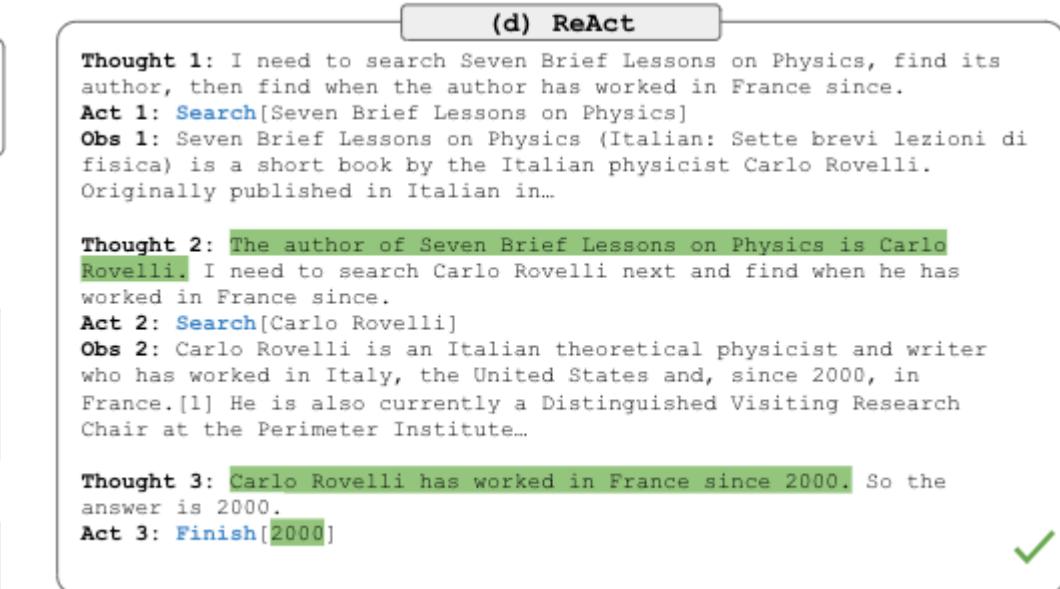
- ReAct Prompting (e.g., LangChain)



Chain-of-thought prompt(CoT)



ReAct: Synergizing Reasoning(CoT) + Acting



Background

- **Each Reasoning Strategy**
 - Reasoning examples sampled from the WikiTableQuestions dataset

World War II casualties of Poland							
Description Losses	Direct War Losses	Murdered	Deaths In Prisons & Camps	Deaths Outside of Prisons & Camps	Murdered in Eastern Regions	Deaths other countries	Total
1939/40	360,000	75,000	69,000	-	-	-	504,000
1940/41	-	100,000	210,000	42,000	-	-	352,000
...	-	...
1943/44	-	82,000	381,000	218,000	-	-	681,000
1944/45	183,000	-	-	-	100,000	-	270,000
Total	...	506,000	1,146,000	473,000	100,000	2,000	2,770,000

Textual Reasoning & Symbolic Reasoning

How many people were murdered in 1940/41?



To determine the number of people murdered in 1940/41, we need to look at the "Murdered" column in the table. From the table, we can see that in 1940/41, there were 210,000 people murdered.



```
df[df['Description Losses'] == '1940/41']['Murdered']
```



> 100,000

There were 100,000 people murdered in 1940/41.

During what year was the largest number of deaths recorded for ethnic poles outside of prisons & camps?



First, we need to identify

...

The row with the largest number in the "Deaths Outside of Prisons & Camps" column is row 5, which corresponds to the year 1943/44.



```
df[df['Deaths Outside of Prisons & Camps'] == df['Deaths Outside of Prisons & Camps'].max()]['Description Losses']
```



> Total

Therefore, the final answer is "Total".

Introduction

- **Our Proposed Method**
 - Table structure normalization
 - Enhance LLMs' resilience against structural table variations
 - Compare each Reasoning Strategies
 - Textual reasoning surpasses symbolic reasoning
 - Despite this, PyAgent can handle larger tables by processing partial table views
 - Each strategy possessing its unique advantages
 - Mix self-consistency mechanism
 - Aggregation of textual reasoning and symbolic reasoning in some proportion
 - Synergistic potential when reasoning strategies are aggregated
 - Model
 - GPT-3.5

Method

- **Problem Definition**

- Table Question Answering task with Table Variation

- Transposed Table

$$\mathcal{T}_{i,j}^\top = \mathcal{T}_{j,i} \quad \forall i \in [0, \mathcal{R} - 1], j \in [0, \mathcal{C} - 1].$$

- Row Shuffled Table

$$\mathcal{T}_{\Pi_{i,j}} = \mathcal{T}_{\pi(i),j} \quad \forall i \in [1, \mathcal{R} - 1], j \in [0, \mathcal{C} - 1]$$

- Row Shuffled and Transposed Table

$$\mathcal{T}_{\Pi_{i,j}}^\top = \mathcal{T}_{j,\pi(i)} \quad \forall i \in [1, \mathcal{R} - 1], j \in [0, \mathcal{C} - 1]$$

- Problem

$$f(\mathcal{T}', \tau, \mathcal{Q}) \rightarrow \mathcal{A}, \quad \forall \mathcal{T}' \in \{\mathcal{T}, \mathcal{T}^\top, \mathcal{T}_\Pi, \mathcal{T}_\Pi^\top\}$$

Method

- **Experimental Setup**

- Dataset: WikiTableQuestion (WikiTQ)
 - Answer a question using an HTML table as the knowledge source
 - For each question, we put one of the 36 generic prompts

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

- x_1 : “*Greece held its last Summer Olympics in which year?*”
 y_1 : {2004}
- x_2 : “*In which city’s the first time with at least 20 nations?*”
 y_2 : {Paris}
- x_3 : “*Which years have the most participating countries?*”
 y_3 : {2008, 2012}
- x_4 : “*How many events were in Athens, Greece?*”
 y_4 : {2}
- x_5 : “*How many more participants were there in 1900 than in the first year?*”
 y_5 : {10}

Method

- **Reasoning Methods**

- Textual Reasoning: Direct Prompting (DP)
 - Chain-of-Thought Prompting in a zero-shot manner
 - LLMs are instructed to reason step-by-step before concluding

Textual Reasoning

You are an advanced AI capable of analyzing and understanding information within tables. Read the table below regarding "[TITLE]".

[TABLE]

Based on the given table, answer the following question:

[QUESTION]

Let's think step by step, and then give the final answer. Ensure the final answer format is only "Final Answer: AnswerName1, AnswerName2..." form, no other form. And ensure the final answer is a number or entity names, as short as possible, without any explanation.

Method

• Reasoning Methods

- Symbolic Reasoning: Python Shell Agent (PyAgent)
 - ReAct Prompting: Thought, Action, Objection
 - Program-Aided: Pandas Library in Python Agent
 - Limited to a maximum of five iterative steps

Symbolic Reasoning

You are working with a pandas dataframe in Python. The name of the dataframe is `df`. Your task is to use `python_repl_ast` to answer the question posed to you.

Tool description:

- `python_repl_ast`: A Python shell. Use this to execute python commands. Input should be a valid python command. When using this tool, sometimes the output is abbreviated - ensure it does not appear abbreviated before using it in your answer.

Guidelines:

- **Aggregated Rows**: Be cautious of rows that aggregate data such as 'total', 'sum', or 'average'. Ensure these rows do not influence your results inappropriately.
- **Data Verification**: Before concluding the final answer, always verify that your observations align with the original table and question.

Strictly follow the given format to respond:

Question: the input question you must answer

Thought: you should always think about what to do to interact with `python_repl_ast`

Action: can **ONLY** be `python_repl_ast`

Action Input: the input code to the action

Observation: the result of the action

... (this Thought/Action/Action Input/Observation can repeat N times)

Thought: after verifying the table, observations, and the question, I am confident in the final answer

Final Answer: the final answer to the original input question (AnswerName1, AnswerName2...)

Notes for final answer:

- Ensure the final answer format is only "Final Answer: AnswerName1, AnswerName2..." form, no other form.
- Ensure the final answer is a number or entity names, as short as possible, without any explanation.
- Ensure to have a concluding thought that verifies the table, observations and the question before giving the final answer.

You are provided with a table regarding "[TITLE]". This is the result of `print(df.to_markdown())`:

[TABLE]

Note: All cells in the table should be considered as `object` data type, regardless of their appearance.

Begin!

Question: [QUESTION]

Method

• Limitations of Table Transposition with LLMs

GPT-3.5 on the WTQ dataset

LLMs As	Task Description	Accuracy
Transposer	$f(\mathcal{T}) \rightarrow \mathcal{T}^\top$	53.68
	$f(\mathcal{T}^\top) \rightarrow \mathcal{T}$	51.07
Detector	$f(\mathcal{T}) \rightarrow 0$	93.35
	$f(\mathcal{T}^\top) \rightarrow 1$	32.54
Determinator	$f(\mathcal{T}, \mathcal{T}_{0,*}, \mathcal{T}_{*,0}) \rightarrow \mathcal{T}_{0,*}$	97.39
	$f(\mathcal{T}^\top, \mathcal{T}_{0,*}, \mathcal{T}_{*,0}) \rightarrow \mathcal{T}_{*,0}$	94.77

Prompt of LLMs as Table Transposer

You are given the following table:

[TABLE]

Please transpose this table. Maintain the format I give, with each row beginning with '|' and each cell separated by '|'. Do not change the content of any cell. Your response should solely consist of the transposed table, without any additional text.

Prompt of LLMs as Table Transposition Detector

Please examine the provided table:

[TABLE]

To enhance readability and facilitate efficient data analysis, it is often suggested that the table headings be horizontally located in the first/topmost row.

Please evaluate the table with this consideration in mind, and provide your response in the following format:

Table Headings: List the headings of the table, separated by commas.

Table Evaluation: Identify whether the headings listed are horizontally located in the first/topmost row. If not, describe the position.

Transpose Recommended: Indicate if transposing is recommended. Answer with only "YES" or "NO", without any additional explanation.

Prompt of Content-Aware Transposition Determination

You are an advanced AI capable of analyzing and understanding information within tables. Read the table below regarding "[TITLE]".

[TABLE]

Headings of a table are labels or titles given to rows or columns to provide a brief description of the data they contain.

Based on the given table, the headings of the table are more likely to be:

- (A) [FIRST_ROW]
- (B) [FIRST_COLUMN]
- (C) None of the above

Directly give your choice. Ensure the format is only "Choice: (A)/(B)/(C)" form, no other form, without any explanation.

• Table Structure Normalization

GPT-3.5 on the WTQ dataset

Perturbation	DP	PyAgent	Method	\mathcal{T}	\mathcal{T}_{Π}	\mathcal{T}^{\top}	\mathcal{T}_{Π}^{\top}
Original (\mathcal{T})	59.50	55.91	DP	59.50	52.21	51.14	37.51
+Shuffle (\mathcal{T}_{Π})	52.21	47.91	+NORM	58.66	58.66	58.30	57.71
+Transpose (\mathcal{T}^{\top})	51.14	12.45		-1.41%	+12.35%	+14.00%	+53.85%
+Transpose&Shuffle (\mathcal{T}_{Π}^{\top})	37.51	8.96	PyAgent	55.91	47.91	12.43	8.96
	-36.96%	-83.97%	+NORM	56.87	57.11	55.44	55.08
				+1.72%	+19.20%	+346.02%	+514.73%

Prompt of Resorting

You are an advanced AI capable of analyzing and understanding information within tables. Read the table below regarding "[TITLE]":

[TABLE]
Note: Only selected rows from the beginning and end of the table are displayed for brevity. Intermediate rows are omitted and represented by "..." for clarity.

The table column headings are provided below, separated by semicolons:

[HEADINGS]

In order to optimize the interpretability and readability of the data, follow these guidelines to determine the most suitable sorting method:

Sorting Guidelines:

1. Evaluate columns based on data types such as numerical, alphabetical, chronological, categorical, or other relevant sorting methods.
2. Identify any patterns or relationships in the data that would be highlighted by certain sorting methods.
3. Consider column position, as those on the left may sometimes have sorting priority.
4. If applicable, consider sorting by multiple columns in a prioritized sequence.

Provide your decision using one of the following statements:

- For sorting using a single column: "Sort by: [Name of Column]".
- For sorting using multiple columns: "Sort by: [Primary Column Name], [Secondary Column Name], ..." .
- If no specific sorting seems advantageous: "Sort by: N/A".

Your response should strictly follow the formats provided.

Prompt of Content-Aware Transposition Determination

You are an advanced AI capable of analyzing and understanding information within tables. Read the table below regarding "[TITLE]".

[TABLE]

Headings of a table are labels or titles given to rows or columns to provide a brief description of the data they contain.

Based on the given table, the headings of the table are more likely to be:

- (A) [FIRST_ROW]
 (B) [FIRST_COLUMN]
 (C) None of the above

$$\begin{cases} f(\mathcal{T}, \mathcal{T}_{0,*}, \mathcal{T}_{*,0}) \rightarrow \mathcal{T}_{0,*} \\ f(\mathcal{T}^{\top}, \mathcal{T}_{0,*}, \mathcal{T}_{*,0}) \rightarrow \mathcal{T}_{*,0} \end{cases}$$

Directly give your choice. Ensure the format is only "Choice: (A)/(B)/(C)" form, no other form, without any explanation.

Method

• Compare Prompting Model with Ablation Study

GPT-3.5 on the WTQ dataset

Method	Accuracy (%)
<i>Few-shot Prompting Methods</i>	
BINDER [★] (Cheng et al., 2023b)	63.61
BINDER [♣] (Cheng et al., 2023b)	55.07
DATER w/o SC [★] (Ye et al., 2023)	61.75
DATER w/ SC [★] (Ye et al., 2023)	68.99
<i>Zero-shot Prompting Methods</i>	
STRUCTGPT [♣] (Jiang et al., 2023a)	51.77
NORM+DP [♣]	58.66
NORM+PYAGENT [♣]	56.87
NORM+PYAGENT-OMITTED [♣]	52.45
NORM+DP&PYAGENT w/ EVAL [♣]	64.22
DP w/ SC [♣]	66.39
+NORM [♣]	64.10
+NORM w/o RESORT [♣]	66.99
PYAGENT w/ SC [♣]	61.39
+NORM [♣]	63.77
+NORM w/o RESORT [♣]	62.84
DP&PYAGENT w/ MIX-SC [♣]	73.06
+NORM [♣]	72.40
+NORM w/o RESORT [♣]	73.65

Error types of DP and PyAgent methods

Error Types	DP	PyAgent	Case Study
Table Misinterpretation	42%	- [†]	Counting Error / Locating Error
Coding Errors	-	38%	Attribute Noise Error / Special Row Misinterpretation Error / Incorrect Coding
Misalignment Issue	24%	28%	Answer Format Issue / Answer Deviation Error
Logical Inconsistency	20%	10%	Reasoning Conflict in DP / Reasoning Mistakes in PyAgent
Execution Issue	-	12%	Interaction Bound or Looping Error / Non-Observable Action Error
Resorting Issue	10%	8%	

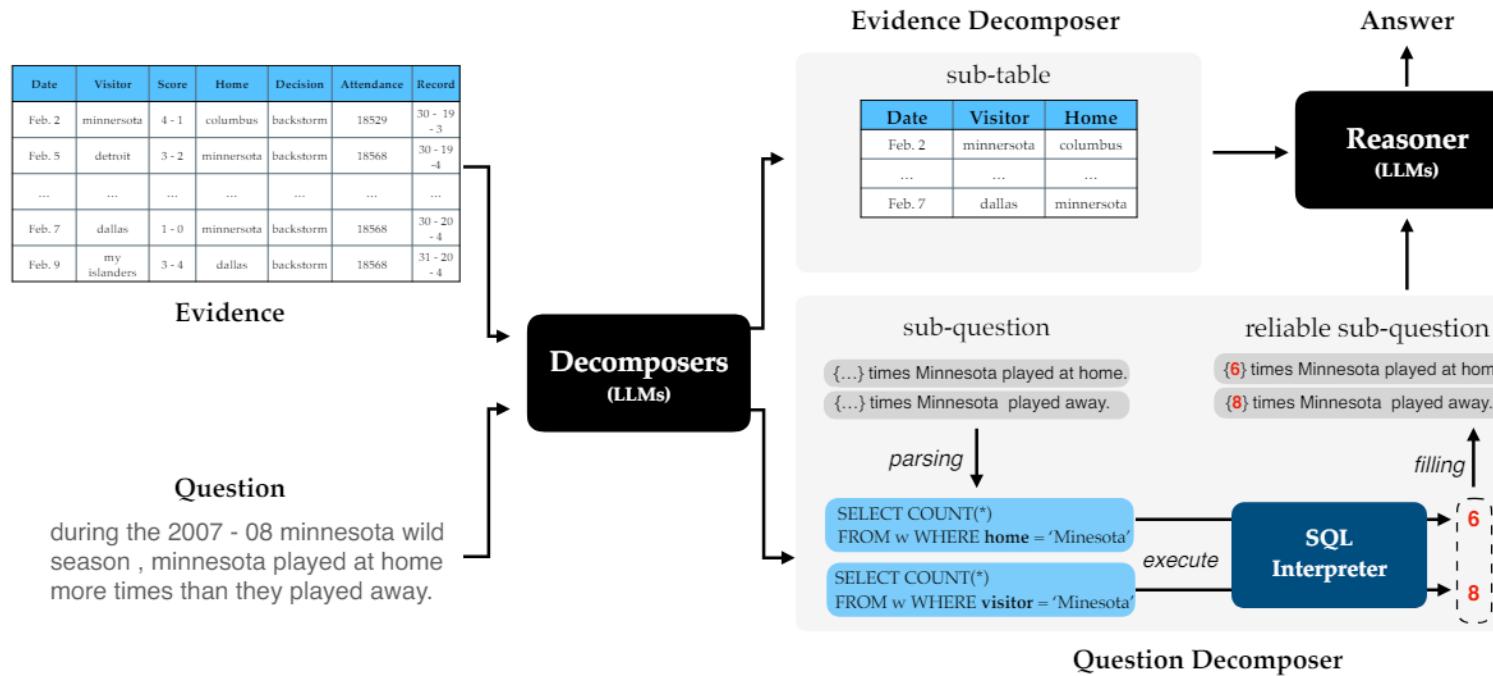
Ablation Study

- Reasoning Strategy: DP, PyAgent
- Table Structure Normalization: Norm
- Aggregate Multiple Reasoning: Self-consistency method(SC, Mix-SC)¹⁷

Method

- Comparable Model: DATER (Few-Shot Prompting)

- Extract sub-table by GPT-3 based in-context reasoning
 - Decompose a complex question into step-by-step sub-questions



during the 2007 - 08 minnesota wild season , minnesota played at home more times than they played away.

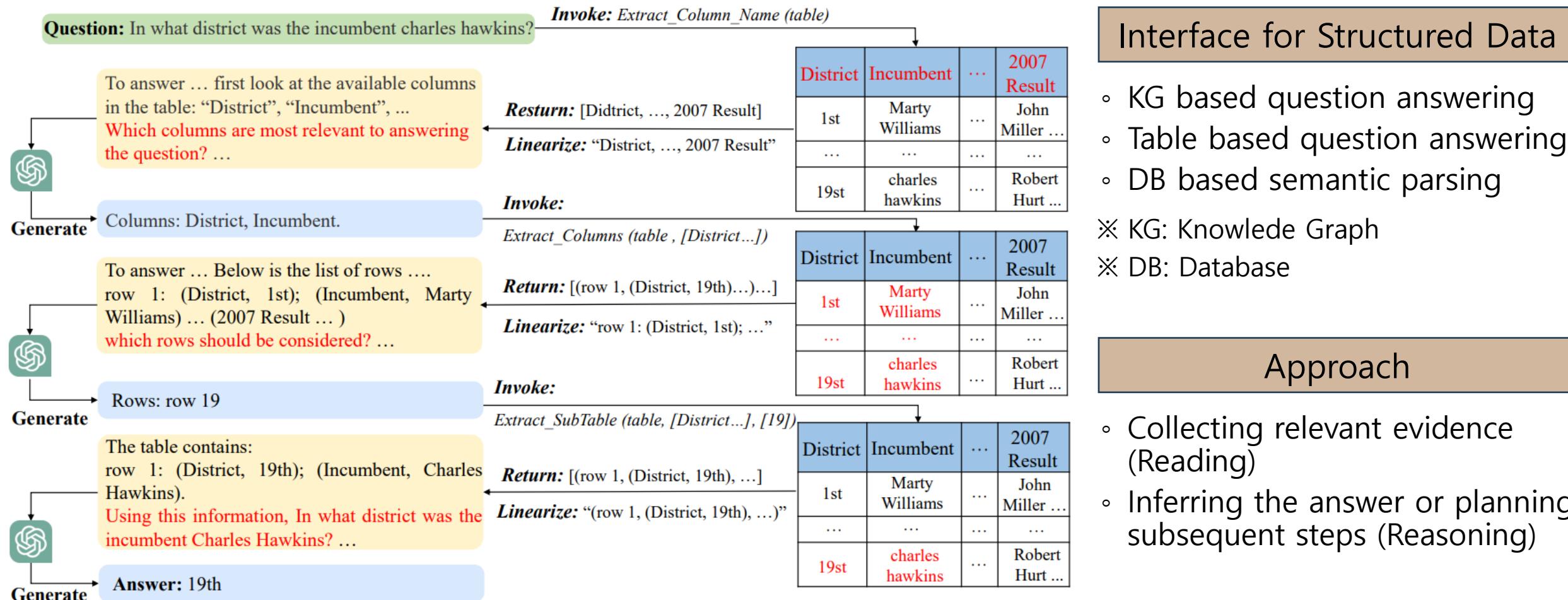
Hierarchical semantic parsing method

Evidence Decomposer

Instruction:	Select relevant columns / rows in the given table that support or oppose the statement.
Table:	<p>table_caption : south wales derby</p> <p>col: competition total matches cardiff win draw swansea win</p> <p>row 1: league 55 6/19 16 20</p> <p>row 2: fa cup 2 6/0 27 2</p> <p>row 3: league cup 5 6/2 0 3</p> <p>...</p>
Question:	there are no cardiff wins that have a draw greater than 26 or smaller than 1.
Sub-table Column, Row:	col(cardiff win, draw), row(2, 3, 17, 23, 28)
Table:	<p>table_caption : dan hardy</p> <p>col: res opponent event round</p> <p>row 1: win amir sadollah ufc on fuel tv 4</p> <p>row 2: loss duane ludwig ufc 146 1</p> <p>row 3: win chris lytle ufc live 3</p> <p>...</p>
Question:	Dan hardy have a game win in turn 4 and he also play against chris lytle in turn 3.
Sub-table Column, Row:	col(res, round, opponent), row(1, 3)
Table:	<p>table_caption : 2008 - 09 season</p> <p>col: nat name moving to type cost</p> <p>row 1: sco chris burke motherwell end of contract free</p> <p>row 2: ivory coast lacine cherif kilmarnock loan 7.8 m</p> <p>row 3: sco alistair park hibernian loan free</p> <p>...</p>
Question:	chris burke , be the only player to have a free transfer fee
Sub-table Column, Row:	col(name, cost), row(1, 3, 13)

Method

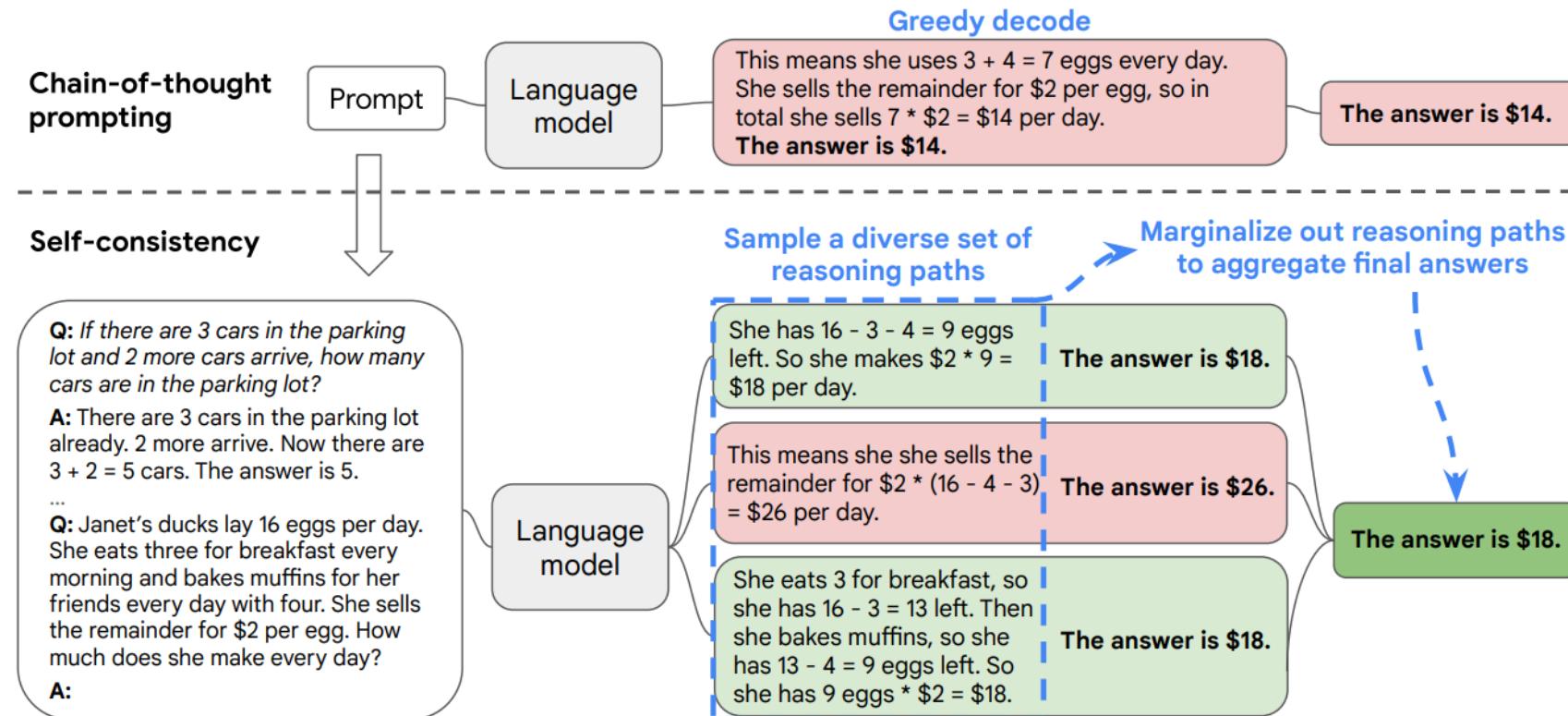
• Comparable Model: StructGPT (Zero-Shot Prompting)



Method

- **Self-consistency method(SC)**

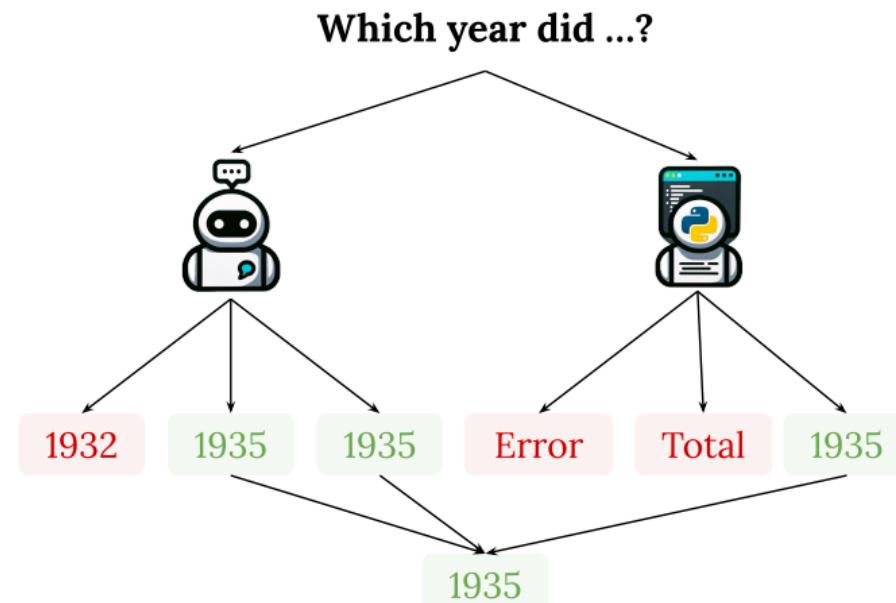
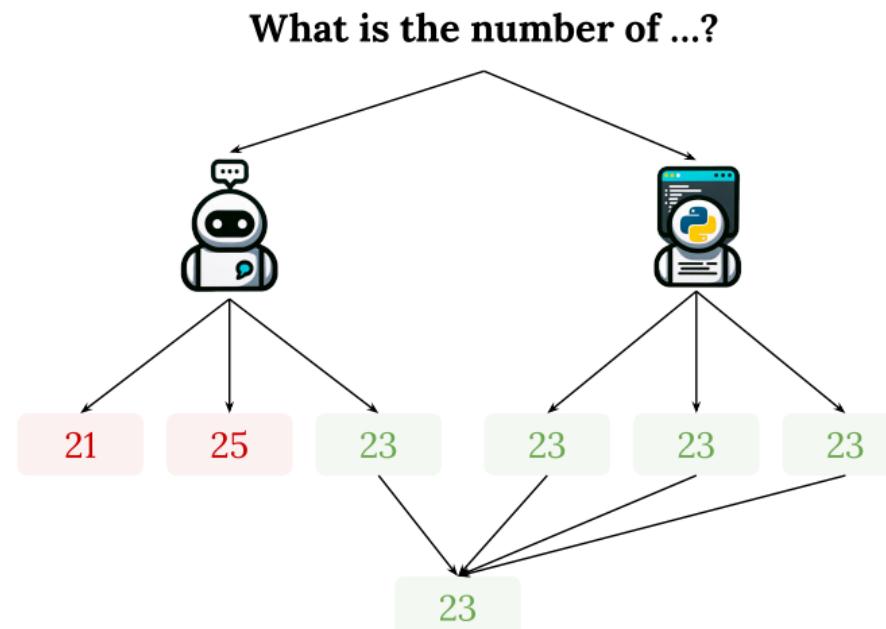
- Prompt a language model using chain-of-thought (CoT) prompting
- Replace “greedy decode” in CoT prompting to generate a diverse set of reasoning paths
- Marginalize out reasoning paths and aggregate by choosing consistent answer



Method

- **Mix Self-consistency method(MIX-SC)**

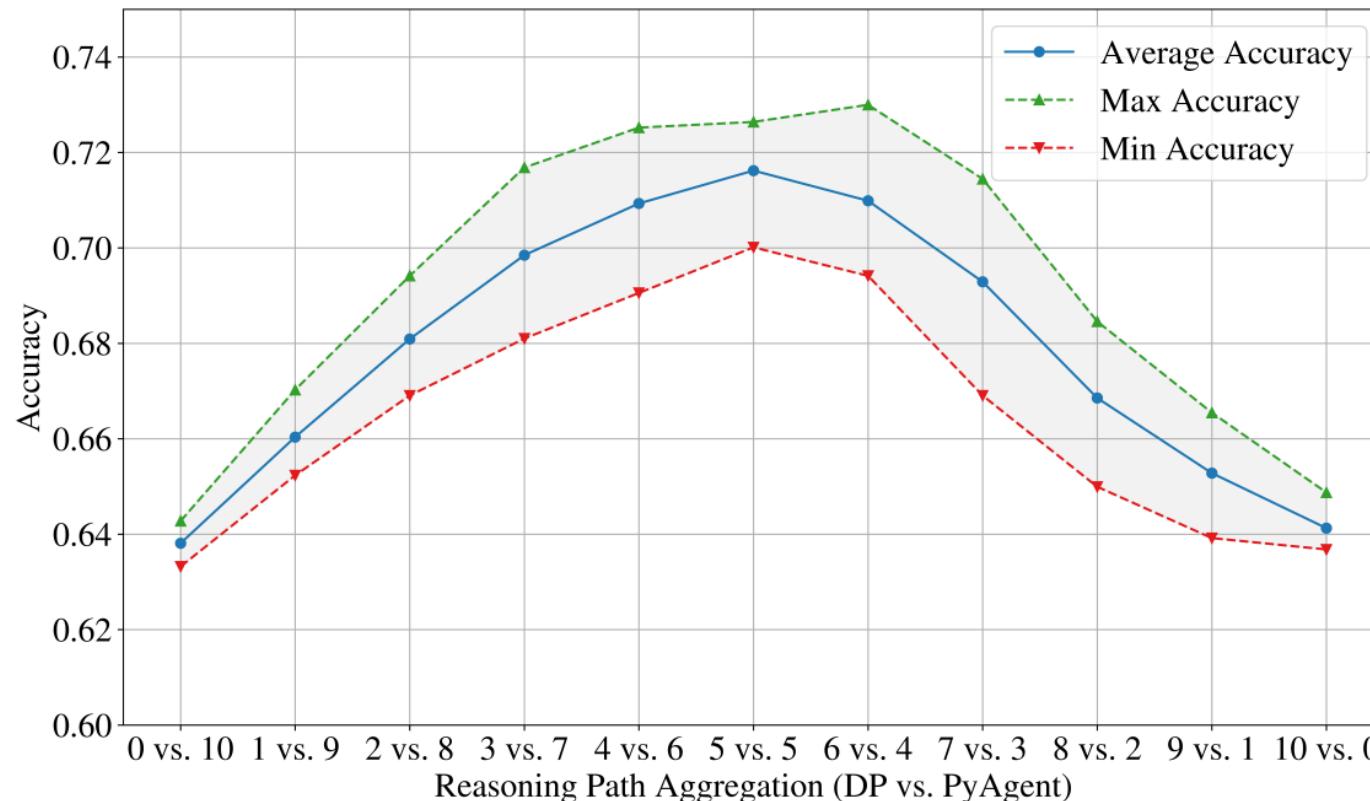
- Confidence Level
 - More diverse results, Lower level of confidence
- Aggregating Multiple outputs from each reasoning strategy
 - Applying majority voting for outputs
 - Refine variations into an accurate prediction



Method

- **Mix Self-consistency method(MIX-SC)**
 - Combinations range from **10 DP vs. 0 PyAgent** to **0 DP vs. 10 PyAgent**
 - Accuraciy results obtained from 100 tests, using random sampling

Accuracy for Mix Self-Consistency method on WTQ dataset



Method

• Self-Evaluation

GPT-3.5 on the WTQ dataset

Method	Accuracy (%)
<i>Few-shot Prompting Methods</i>	
BINDER [★] (Cheng et al., 2023b)	63.61
BINDER [♣] (Cheng et al., 2023b)	55.07
DATER w/o SC [★] (Ye et al., 2023)	61.75
DATER w/ SC [★] (Ye et al., 2023)	68.99
<i>Zero-shot Prompting Methods</i>	
STRUCTGPT [♣] (Jiang et al., 2023a)	51.77
NORM+DP [♣]	58.66
NORM+PYAGENT [♣]	56.87
NORM+PYAGENT- OMITTED [♣]	52.45
NORM+DP&PYAGENT w/ EVAL^{★★}	64.22
DP w/ SC [♣]	66.39
+NORM [♣]	64.10
+NORM w/o RESORT [♣]	66.99
PYAGENT w/ SC [♣]	61.39
+NORM [♣]	63.77
+NORM w/o RESORT [♣]	62.84
DP&PYAGENT w/ Mix-SC [♣]	73.06
+NORM [♣]	72.40
+NORM w/o RESORT [♣]	73.65

Prompt of Self-Evaluation

Below is a markdown table regarding "[TITLE]":

[TABLE]

You're tasked with answering the following question:

[QUESTION]

You have 2 answers derived by two different methods. Answer A was derived by prompting the AI to think step-by-step. Answer B was derived by interacting with a Python Shell.

Answer A is [COT_ANSWER].

Answer B is [AGENT_ANSWER].

Your task is to determine which is the correct answer. It is crucial that you strictly adhere to the following evaluation process:

- **Preliminary Evaluation**: Begin by evaluating which of the two answers directly addresses the question in a straightforward and unambiguous manner. A direct answer provides a clear response that aligns closely with the query without introducing additional or extraneous details. If one of the answers is not a direct response to the question, simply disregard it.
- **Nature of the Question**: If both answers appear to be direct answers, then evaluate the nature of the question. For tasks involving computation, counting, and column-locating, especially when for extensive table, the Python Shell (Answer B) might be more precise. However, always remain cautious if the Python Shell's output appears off (e.g., error messages, success notifications, etc.). Such outputs may not be trustworthy for a correct answer.
- **Final Verdict**: Finally, after thorough evaluation and explanation, provide your verdict strictly following the given format:
 - Use "[[A]]" if Answer A is correct.
 - Use "[[B]]" if Answer B is correct.

Note:

- Each method has its own strengths and weaknesses. Evaluate them with an unbiased perspective. When in doubt, consider the nature of the question and lean towards the method that is most suited for such queries.
- Ensure that your verdict is provided after evaluation, at the end.

Method

- Overall Evaluation

GPT-3.5 with Comparable Model on the WTQ dataset

Method	Accuracy (%)
<i>Fine-tuning Based Models</i>	
TAPAS (Herzig et al., 2020)	48.8
T5-3B (Xie et al., 2022)	49.3
TAPAX (Liu et al., 2022)	57.5
REASTAP (Zhao et al., 2022)	58.7
OMNITAB (Jiang et al., 2022)	63.3
<i>LLMs Based Methods</i>	
STRUCTGPT★ (Jiang et al., 2023a)	48.4
BINDER★ (Cheng et al., 2023b)	55.5
BINDER♠ (Cheng et al., 2023b)	64.6
LEVER♠ (Ni et al., 2023)	65.8
DATER♠ (Ye et al., 2023)	65.9
Ours★	73.6

Conclusion

- **Our Proposed Method**

- Table Structure Normalization
 - LLMs are sensitive to structural variance of tables
 - Application of a normalization strategy enhance resilience against table variation
- Each Reasoning Strategy
 - Textual reasoning surpasses symbolic reasoning
 - Despite this, PyAgent can handle larger tables by processing partial table views
- Mix Self-Consistency (Mix-SC)
 - Integrating multiple reasoning strategy is proved beneficial for overall interpretation accuracy
- Limitation
 - Exclusive utilization of GPT-3.5 may limit the generalizability of our findings
 - GPT-4 might offer enhanced outcomes
 - All table data are sourced from Wikipedia
 - Certain answers might be available within data, thus potentially biasing results

Appendix

• Table Misinterpretation: Counting Error

Speed skating at the 2006 Winter Olympics - Men's 1000 metres

	Rank	Pair	Name	Country	Time	Time behind
...
9	10	6	Dmitry Dorofeyev	Russia	1:09.74	+0.85
10	11	19	Jeremy Wotherspoon	Canada	1:09.76	+0.87
11	12	9	Beorn Nijenhuis	Netherlands	1:09.85	+0.96
...
39	—	2	Ermanno Ioriatti	Italy	DQ	—
40	—	6	Erik Zachrisson	Sweden	DQ	—

What is the name of the last manufacturer on this chart?

To determine the number of finishers Canada had in the Men's 1000 metres speed skating event at the 2006 Winter Olympics, we need to identify the rows in the table that correspond to Canadian athletes who completed the race.

From the table, we can see that the Canadian athletes are listed in rows 10, 15, 18, 23, 24, 27, 28, and 36. These rows correspond to the following athletes: Jeremy Wotherspoon, François-Olivier Roberge, Denny Morrison, Mun Jun, Janne Hänninen, Takahiro Ushiyama, Steven Elm, and Risto Rosendahl.

Therefore, the number of finishers Canada had is 8.

Final Answer: 8 ✗

Correct Answer: 4

Appendix

• Table Misinterpretation: Locating Error

2005-06 FIS Ski Jumping World Cup

	Rank	Name	Nationality	1st (m)	2nd (m)	Points
0	11	Jakub Janda	CZE	123.5	128.5	252.6
1	5	Andreas Kofler	AUT	126.5	132	261.3
2	4	Adam Małysz	POL	127	130.5	261.5
3	3	Janne Ahonen	FIN	127	130.5	264
4	2	Tami Kiuru	FIN	128.5	130.5	264.7
5	1	Matti Hautamäki	FIN	132	131.5	274.3

Question: Who has the top 2nd score?

To determine who has the top 2nd score, we need to find the highest value in the "2nd (m)" column of the table.

By analyzing the table, we can see that the highest value in the "2nd (m)" column is 132, achieved by both Andreas Kofler and Matti Hautamäki.

Therefore, the top 2nd score is 132 achieved by Andreas Kofler and Matti Hautamäki.

Final Answer: Andreas Kofler, Matti Hautamäki X

Correct Answer: Andreas Kofler

Appendix

• Coding Error: Attribute Noise Error

List of The Citadel Bulldogs football seasons

	Season	Team	Record	Head Coach	Quarterback	...	Runner Up
0	1970	Dallas Cowboys	10–4	Tom Landry*	Craig Morton	...	San Francisco 49ers
1	1971	Dallas Cowboys†	11–3	Tom Landry*	Roger Staubach*	...	San Francisco 49ers
2	1972	Washington Redskins	11–3	George Allen*	Billy Kilmer	...	Dallas Cowboys
3	1973	Minnesota Vikings	12–2	Bud Grant*	Fran Tarkenton*	...	Dallas Cowboys
...
43	2013	Seattle Seahawks†	13–3	Pete Carroll	Russell Wilson	...	San Francisco 49ers

How many times did the cowboys win the nfc championship in the 1970s?

🤔Thought: To answer this question, I need to count the number of times the Cowboys won the NFC championship in the 1970s. I can do this by filtering the dataframe based on the team and season columns.

🛠Action: Use `python_repl_ast` to filter the dataframe and count the number of rows.

💡Action Input: `df[(df['Team'] == 'Dallas Cowboys') & (df['Season'] >= 1970) & (df['Season'] < 1980)].shape[0]`

💻Observation:

3

💡Final Answer: 3 ✗

Correct Answer: 5

Appendix

• Coding Error: Special Row Misinterpretation Error

Nicolás Terol

	Season	Class	Moto	Races	Win	Podiums	Pole	Pts	Position
0	2004	125cc	Aprilia	1	0	0	0	0	NC
1	2005	125cc	Derbi	13	0	0	0	1	36th
2	2006	125cc	Derbi	16	0	0	0	53	14th
3	2007	125cc	Derbi	17	0	0	0	19	22nd
...
11	Total	nan	nan	147	16	39	9	1213.5	nan

How many times did the cowboys win the nfc championship in the 1970s?

🤔Thought: To answer this question, I need to count the number of times the Cowboys won the NFC championship in the 1970s. I can do this by filtering the dataframe based on the team and season columns.

🛠Action: Use `python_repl_ast` to filter the dataframe and count the number of rows.

💡Action Input: `df[(df['Team'] == 'Dallas Cowboys') & (df['Season'] >= 1970) & (df['Season'] < 1980)].shape[0]`

💻Observation:

3

💡Final Answer: 3 ✗

Correct Answer: 5

Appendix

• Coding Error: Incorrect Coding

Stoewer

	Type	Construction period	Cylinder	Capacity	Power	Vmax
0	10 PS (7 kW; 10 hp)	1901–1902	straight-2	1.527 cc	18 PS (13,2 kW)	50 km/h (31 mph)
1	8/14 PS	1902–1905	straight-2	1.527 cc	14 PS (10,3 kW)	50 km/h (31 mph)
2	20 PS (15 kW; 20 hp)	1904–1905	straight-4	7.946 cc	45 PS (33 kW)	85 km/h (53 mph)
...
40	Greif Junior	1936–1939	flat-4	1.484 cc	34 PS (25 kW)	100 km/h (62 mph)

How many cars had a maximum velocity of at least 100 km/h?

💡Thought: To answer this question, I need to count the number of cars that had a maximum velocity of at least 100 km/h. I can do this by filtering the 'Vmax' column and counting the number of rows that meet the condition.

✖Action: I will filter the 'Vmax' column for cars that have a maximum velocity of at least 100 km/h and then count the number of rows.

💻Action Input: `df[df['Vmax'].str.contains('100 km/h')].shape[0]`

💻Observation:

7

💡Thought: After filtering the 'Vmax' column for cars with a maximum velocity of at least 100 km/h, I found that there are 7 cars that meet this condition.

💡Final Answer: 7 ✗

Correct Answer: 17

Appendix

• Misalignment Issue: Answer Format Issue

1984 IMSA GT Championship season

Rnd	Circuit	GTP Winning TeamGTP Winning Drivers	GTO Winning TeamGTO Winning Drivers	GTU Winning TeamGTU Winning Drivers	Results
-----	---------	-------------------------------------	-------------------------------------	-------------------------------------	---------

What number did Brooks Racing have?

💡Thought: To find the number of Brooks Racing, I need to filter the dataframe based on the team name and then extract the number.

✗Action: Use pandas filtering to find the row with the team name "Brooks Racing" and extract the number.

💡Action Input: `df[df['GTP Winning TeamGTP Winning Drivers'].str.contains('Brooks Racing')]['GTP Winning TeamGTP Winning Drivers']`

─Observation:

```
Series([], Name: GTP Winning TeamGTP Winning Drivers, dtype: object)
```

💡Thought: It seems that there is no row in the dataframe where the GTP Winning TeamGTP Winning Drivers column contains the name "Brooks Racing". I should double-check the spelling and formatting of the team name.

✗Action: Print the unique values in the GTP Winning TeamGTP Winning Drivers column to verify the spelling and formatting.

💡Action Input: `df['GTP Winning TeamGTP Winning Drivers'].unique()`

─Observation:

```
['#00 Kreepy Krauly Racing'
 'Sarel van der Merwe\n Graham Duxbury\n Tony Martin'
 'Doc Bundy\n Brian Redman' '#04 Group 44'
 'Mauricio DeNarvaez\n Hans Heyer\n Stefan Johansson'
 '#48 DeNarvaez Enterprises' '#16 Marty Hinze Racing' 'Don Whittington'
 'Don Whittington\n Randy Lanier' '#56 Blue Thunder Racing'
 'Randy Lanier' 'Bill Whittington\n Randy Lanier' 'Sarel van der Merwe'
 'Al Holbert\n Derek Bell' '#14 Holbert Racing'
 'Al Holbert\n Jim Adams\n Derek Bell' 'Bill Whittington'
 '#57 Blue Thunder Racing' 'Dale Whittington\n Randy Lanier']
```

💡Thought: The unique values in the GTP Winning TeamGTP Winning Drivers column do not include "Brooks Racing". It seems that there is no row in the dataframe where the team name is "Brooks Racing". I should double-check the column name to ensure I am searching in the correct column.

✗Action: Print the column names to verify the column name.

💡Action Input: `df.columns`

─Observation:

```
Index(['Rnd', 'Circuit', 'GTP Winning TeamGTP Winning Drivers',
       'GTO Winning TeamGTO Winning Drivers',
       'GTU Winning TeamGTU Winning Drivers', 'Results'],
      dtype='object')
```

💡Thought: The column name is 'GTP Winning TeamGTP Winning Drivers', not 'GTO Winning TeamGTO Winning Drivers'. I should modify the filtering condition to search for 'Brooks Racing' in the correct column.

✗Action: Use pandas filtering to find the row with the team name "Brooks Racing" and extract the number.

💡Action Input: `df[df['GTO Winning TeamGTO Winning Drivers'].str.contains('Brooks Racing')]['GTO Winning TeamGTO Winning Drivers']`

─Observation:

```
10    #77 Brooks Racing
22    #77 Brooks Racing
Name: GTO Winning TeamGTO Winning Drivers, dtype: object
```

💡Final Answer: Brooks Racing had the number 77. ✓

Correct Answer: 77

Appendix

• Misalignment Issue: Answer Deviation Error

2007–08 Washington Huskies men's basketball team

	#	Name	Height	Weight (lbs.)	Position	Class	Hometown	Previous Team(s)
0	21	Artem Wallace	6'8"	250	C	Jr.	Toledo, WA, U.S.	Toledo HS
1	32	Joe Wolfinger	7'0"	255	C	RS So.	Portland, OR, U.S.	Northfield Mount Hermon School
2	22	Justin Holiday	6'6"	170	F	Fr.	Chatsworth, CA, U.S.	Campbell Hall School
3	24	Quincy Pondexter	6'6"	210	F	So.	Fresno, CA, U.S.	San Joaquin Memorial HS
...
11	4	Tim Morris	6'4"	210	G	Sr.	Spokane Wa, U.S.	Central Valley HS

Are there at least 13 names on the chart?

To determine if there are at least 13 names on the chart, we need to count the number of rows in the table.

Step 1: Count the number of rows in the table.

There are 12 rows in the table.

Step 2: Compare the count to 13.

The count of 12 is less than 13.

Final Answer: 12 X

Correct Answer: no

Appendix

• Logical Inconsistency: Reasoning Conflict in DP

Chief of the National Guard Bureau

	Number	Name	Service	From	To
13	13	MG Kenneth F. Cramer	USA	September 30, 1947	September 4, 1950
14	14	MG Raymond H. Fleming	USA	August 14, 1951	February 15, 1953
15	15	MG Edgar C. Erickson	USA	June 22, 1953	May 31, 1959
...
39	(Acting)	MG John R. D'Araujo, Jr.	USA	August 1, 1994	September 30, 1994

Question: Was george e. leach or kenneth f. cramer the chief of the national guard bureau in 1934?

To determine whether George E. Leach or Kenneth F. Cramer was the Chief of the National Guard Bureau in 1934, we need to find the individuals who held the position during that time period.

From the table, we can see that George E. Leach served as the Chief of the National Guard Bureau from December 1, 1931, to November 30, 1935. Kenneth F. Cramer, on the other hand, served as the Chief of the National Guard Bureau from September 30, 1947, to September 4, 1950.

Since neither George E. Leach nor Kenneth F. Cramer held the position in 1934, the answer is neither.

Final Answer: Neither X

Correct Answer: George E. Leach

Appendix

• Logical Inconsistency: Reasoning Mistakes in PyAgent

Wiseman hypothesis

	Tablet	Genealogy	Narrative	Colophon
...
8	9	Abraham to Isaac 25:19	25:20 to 35:29	"This is the account of Esau." 36:1 (eldest son)
9	10	Descendants of Esau 36:2 - 5	36:6 - 8	"This is the account of Esau." 36:9
10	11	Descendants of Esau 36:10 to 37:1	no narrative	"This is the account of Jacob." 37:2
11	nan	no genealogy	37:2 to 50:26	no colophon

Tablets 9 and 10 are each accounted by who?

🤔 Thought: To answer this question, I need to find the rows corresponding to Tablets 9 and 10 and check the values in the "Genealogy" column.

✖ Action: Filter the dataframe to include only the rows with Tablets 9 and 10.

💻 Action Input: `df[(df['Tablet'] == 9) | (df['Tablet'] == 10)]`

💻 Observation:

Tablet	Genealogy	Narrative	Colophon
8 9.0	Abraham to Isaac 25:19	25:20 to 35:29	"This is the account of Esau." 36:1 (eldest son)
9 10.0	Descendants of Esau 36:2 - 5	36:6 - 8	"This is the account of Esau." 36:9

🤔 Thought: The rows corresponding to Tablets 9 and 10 are accounted by Abraham and the descendants of Esau, respectively.

💡 Final Answer: Abraham, Descendants of Esau ✗

Correct Answer: Esau

Appendix

• Execution Issue: Interaction Bound or Looping Error

Comparison of Linux distributions

#	Date	Venue	Opponent	Score	Result	Competition
---	------	-------	----------	-------	--------	-------------

Which team competed for the Euro 2000 qualifying the most consecutive years?

🤔 Thought: To find the team that competed for the Euro 2000 qualifying the most consecutive years, I need to count the number of consecutive years each team participated in the Euro 2000 qualifying.

✗ Action: I will group the dataframe by the "Opponent" column and count the number of unique years in the "Date" column for each opponent.

💻 Action Input: `df.groupby('Opponent')['Date'].apply(lambda x: x.dt.year.nunique())`

💻 Observation:

`AttributeError: Can only use .dt accessor with datetimelike values`

🤔 Thought: The error message suggests that the "Date" column is not in a datetime format. I need to convert the "Date" column to a datetime format before using the `.dt` accessor.

✗ Action: I will convert the "Date" column to a datetime format using the `pd.to_datetime()` function.

💻 Action Input: `df['Date'] = pd.to_datetime(df['Date'])`

💻 Observation:

`success!`

🤔 Thought: Now that the "Date" column is in a datetime format, I can proceed with grouping the dataframe and counting the number of unique years for each opponent.

✗ Action: I will group the dataframe by the "Opponent" column and count the number of unique years in the "Date" column for each opponent.

💻 Action Input: `df.groupby('Opponent')['Date'].apply(lambda x: x.dt.year.nunique())`

💻 Observation:

`Opponent`

`Austria 1`

`Cyprus 1`

`Poland 1`

`San Marino 1`

`Name: Date, dtype: int64`

Appendix

- Execution Issue: Interaction Bound or Looping Error

Comparison of Linux distributions

#	Date	Venue	Opponent	Score	Result	Competition
---	------	-------	----------	-------	--------	-------------

Which team competed for the Euro 2000 qualifying the most consecutive years?

💡Thought: The output shows the number of unique years each opponent participated in the Euro 2000 qualifying. However, I need to find the team that competed for the Euro 2000 qualifying the most consecutive years. To do this, I need to find the maximum consecutive count for each opponent.

✗Action: I will modify the lambda function to calculate the maximum consecutive count for each opponent.

💡Action Input: `df.groupby('Opponent')['Date'].apply(lambda x: x.diff().dt.days.ne(1).cumsum().value_counts().max())`

■Observation:

```
Opponent
Austria      1
Cyprus        1
Poland        1
San Marino   1
Name: Date, dtype: int64
```

💡Thought: The output is not what I expected. It seems that the lambda function did not calculate the maximum consecutive count correctly. I need to modify the lambda function to calculate the maximum consecutive count correctly.

✗Action: I will modify the lambda function to calculate the maximum consecutive count correctly.

💡Action Input: `df.groupby('Opponent')['Date'].apply(lambda x: x.diff().dt.days.ne(1).cumsum().value_counts().max() + 1)`

■Observation:

```
Opponent
Austria      2
Cyprus        2
Poland        2
San Marino   2
Name: Date, dtype: int64
```

Appendix

• Execution Issue: Non-Observable Action Error

Stoewer

	Game	Date	Team	Score	...	LocationAttendance	Record
--	------	------	------	-------	-----	--------------------	--------

How many consecutive games were played in the ARCO Arena from January 2nd to January 24th?

💡Thought: To answer this question, I need to find the consecutive games played in the ARCO Arena from January 2nd to January 24th. I can do this by filtering the dataframe based on the date and location, and then counting the number of rows.

🛠Action: Filter the dataframe based on the date and location.

💻Action Input: `df_filtered = df[(df['Date'] >= 'January 2') & (df['Date'] <= 'January 24') & (df['LocationAttendance'].str.contains('ARCO Arena'))]`

💻Observation:

success!

💡Thought: Now that I have filtered the dataframe, I can count the number of rows to determine the number of consecutive games played in the ARCO Arena from January 2nd to January 24th.

🛠Action: Count the number of rows in the filtered dataframe.

💻Action Input: `consecutive_games_count = len(df_filtered)`

💻Observation:

success!

💡Thought: Now that I have counted the number of rows in the filtered dataframe, I have the number of consecutive games played in the ARCO Arena from January 2nd to January 24th.

💡Final Answer: 4 ✗

Correct Answer: 3

Appendix

- Resorting Issue

Table Before NORM

	Year	Manufacturer	Model	Length (feet)	Quantity	Fleet Series	Fuel Propulsion	Powertrain
0	1996	New Flyer	D60	60 (articulated)	30	1901-1930*	Diesel	Detroit Diesel Series 50\nAllison B400R
1	1998	NABI	416	40	133	3001-3067, 3101-3166*	Diesel	Cummins M11E\nAllison B400R
...
19	2013	Gillig	Low-floor Advantage	40	65	1301-1365	Diesel	Cummins ISL 280 HP \nAllison B400 6-speed
20	2013	New Flyer	Xcelsior D60	60	23	2201-2223	Diesel	Cummins ISL 330 HP\nAllison B400 6-speed
21	2013	Gillig	Low-floor Advantage	40	55	6101-6155	Diesel	Cummins ISL 280 HP\nAllison B400 6-speed

What is the name of the last manufacturer on this chart?

Answer before NORM (Ground Truth): Gillig

Answer after NORM: New Flyer

	Year	Manufacturer	Model	Length (feet)	Quantity	Fleet Series	Fuel Propulsion	Powertrain
0	1996	New Flyer	D60	60 (articulated)	30	1901-1930*	Diesel	Detroit Diesel Series 50\nAllison B400R
1	1998	NABI	416	40	133	3001-3067, 3101-3166*	Diesel	Cummins M11E\nAllison B400R
...
19	2013	Gillig	Low-floor Advantage	40	65	1301-1365	Diesel	Cummins ISL 280 HP \nAllison B400 6-speed
20	2013	Gillig	Low-floor Advantage	40	55	6101-6155	Diesel	Cummins ISL 280 HP\nAllison B400 6-speed
21	2013	New Flyer	Xcelsior D60	60	23	2201-2223	Diesel	Cummins ISL 330 HP\nAllison B400 6-speed

Appendix

- **ReAct Prompting (e.g., LangChain)**

- Dataset: Hotpot QA

ReAct

Question	What is the elevation range for the area that the eastern sector of the Colorado orogeny extends into?
Thought 1	I need to search Colorado orogeny, find the area that the eastern sector of the Colorado orogeny extends into, then find the elevation range of the area.
Action 1	Search[Colorado orogeny]
Observation 1	The Colorado orogeny was an episode of mountain building (an orogeny) in Colorado and surrounding areas.
Thought 2	It does not mention the eastern sector. So I need to look up eastern sector.
Action 2	Lookup[eastern sector]
Observation 2	(Result 1 / 1) The eastern sector extends into the High Plains and is called the Central Plains orogeny.
Thought 3	The eastern sector of Colorado orogeny extends into the High Plains. So I need to search High Plains and find its elevation range.
Action 3	Search[High Plains]
Observation 3	High Plains refers to one of two distinct land regions
Thought 4	I need to instead search High Plains (United States).
Action 4	Search[High Plains (United States)]
Observation 4	The High Plains are a subregion of the Great Plains. From east to west, the High Plains rise in elevation from around 1,800 to 7,000 ft (550 to 2,130 m). [3]
Thought 5	High Plains rise in elevation from around 1,800 to 7,000 ft, so the answer is 1,800 to 7,000 ft.
Action 5	Finish[1,800 to 7,000 ft]

Appendix

- **WikiTableQuestion (WikiTQ)**

- Answer a question using an HTML table as the knowledge source
- For each question, we put one of the 36 generic prompts

Year	City	Country	Nations
1896	Athens	Greece	14
1900	Paris	France	24
1904	St. Louis	USA	12
...
2004	Athens	Greece	201
2008	Beijing	China	204
2012	London	UK	204

- x_1 : “*Greece held its last Summer Olympics in which year?*”
 y_1 : {2004}
- x_2 : “*In which city's the first time with at least 20 nations?*”
 y_2 : {Paris}
- x_3 : “*Which years have the most participating countries?*”
 y_3 : {2008, 2012}
- x_4 : “*How many events were in Athens, Greece?*”
 y_4 : {2}
- x_5 : “*How many more participants were there in 1900 than in the first year?*”
 y_5 : {10}

Appendix

- **TABFACT**

- Table-based fact verification dataset with logical inferences in the annotated statements
- Linguistic Reasoning
 - **Correctly classify the entailment relation**
- Symbolic Reasoning
 - **Symbolic execution on the table structure (e.g., "There are three Democrats incumbents")**

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.

Appendix

- **TABFACT**

- Higher-order Operations

1. Aggregation: the aggregation operation refers to sentences like “the averaged age of all”, “the total amount of scores obtained in ...”, etc.
2. Negation: the negation operation refers to sentences like “xxx did not get the best score”, “xxx has never obtained a score higher than 5”.
3. Superlative: the superlative operation refers to sentences like “xxx achieves the highest score in”, “xxx is the lowest player in the team”.
4. Comparative: the comparative operation refers to sentences like “xxx has a higher score than yyy”.
5. Ordinal: the ordinal operation refers to sentences like “the first country to achieve xxx is xxx”, “xxx is the second oldest person in the country”.
6. Unique: the unique operation refers to sentences like “there are 5 different nations in the tournament, ”, “there are no two different players from U.S”
7. All: the for all operation refers to sentences like “all of the trains are departing in the morning”, “none of the people are older than 25.”
8. None: the sentences which do not involve higher-order operations like “xxx achieves 2 points in xxx game”, “xxx player is from xxx country”.

Appendix

- TABFACT

- Latent Program Algorithm (LPA)

Algorithm 1 Latent Program Search with Comments

```

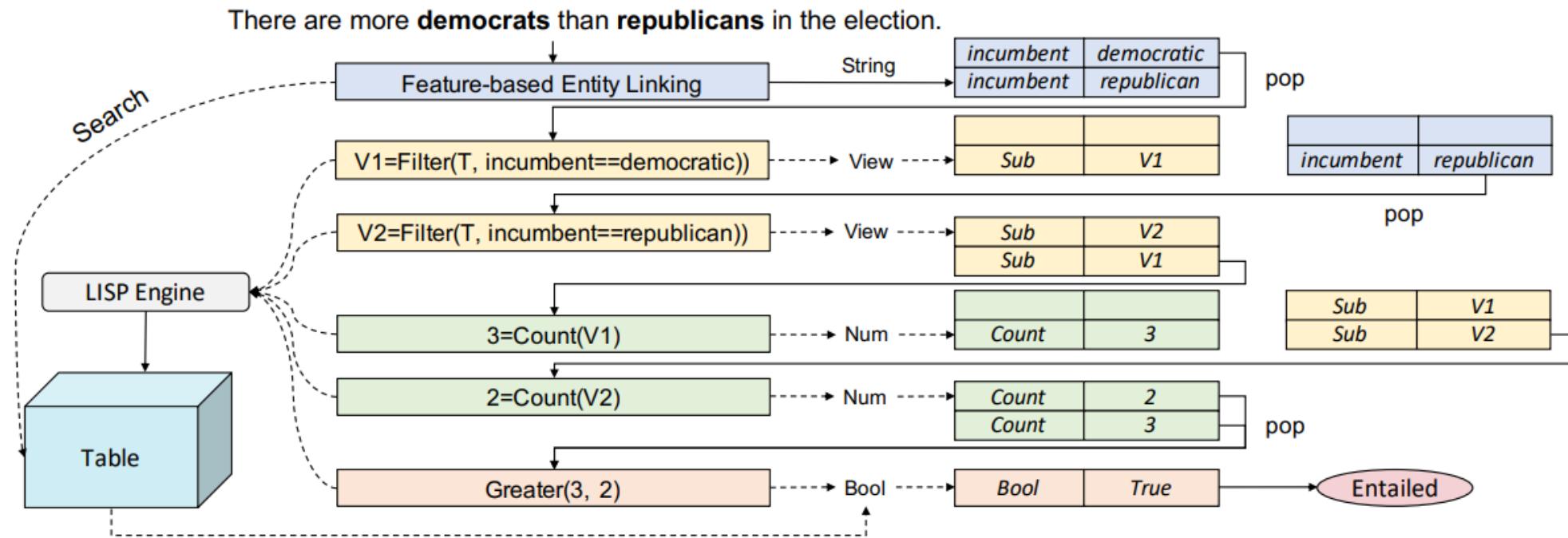
1: Initialize Number Cache  $\mathcal{N}$ , String Cache  $\mathcal{R}$ , Bool Cache  $\mathcal{B}$ , View Cache  $\mathcal{V} \rightarrow \emptyset$ 
2: Push linked numbers, strings from the given statement  $S$  into  $\mathcal{N}, \mathcal{R}$ , and push  $\mathbf{T}$  into  $\mathcal{V}$ 
3: Initialize the result collector  $\mathcal{P} \rightarrow \emptyset$  and an empty program trace  $P = \emptyset$ 
4: Initialize the Queue  $\mathcal{Q} = [(P, \mathcal{N}, \mathcal{R}, \mathcal{B}, \mathcal{V})]$ , we use  $\mathcal{Q}$  to store the intermediate states
5: Use trigger words to find plausible function set  $\mathcal{F}$ , for example, more will trigger Greater function.
6: while loop over time  $t = 1 \rightarrow \text{MAXSTEP}$  do:
7:   while  $(P, \mathcal{N}, \mathcal{R}, \mathcal{B}, \mathcal{V}) = \mathcal{Q}.pop()$  do:
8:     while loop over function set  $f \in \mathcal{F}$  do:
9:       if arguments of  $f$  are in the caches then
10:        Pop out the required arguments  $arg_1, arg_2, \dots, arg_n$  for different cachess.
11:        Execute  $A = f(arg_1, \dots, arg_n)$  and concatenate the program trace  $P$ .
12:        if Type( $A$ )=Bool then
13:          if  $\mathcal{N} = \mathcal{S} = \mathcal{B} = \emptyset$  then
14:             $\mathcal{P}.push((P, A))$  # The program  $P$  is valid since it consumes all the variables.
15:             $P = \emptyset$  # Collect the valid program  $P$  into set  $\mathcal{P}$  and reset  $P$ 
16:          else
17:             $\mathcal{B}.push(A)$  # The intermediate boolean value is added to the bool cache
18:             $\mathcal{Q}.push((P, \mathcal{N}, \mathcal{R}, \mathcal{B}, \mathcal{V}))$  # Add the refreshed state to the queue again
19:        if Type( $A$ )  $\in \{\text{Num, Str, View}\}$  then
20:          if  $\mathcal{N} = \mathcal{S} = \mathcal{B} = \emptyset$  then
21:             $P = \emptyset$ ; break # The program ends without consuming the cache, throw it.
22:          else
23:            push  $A$  into  $\mathcal{N}$  or  $\mathcal{S}$  or  $\mathcal{V}$  # Add the refreshed state to the queue for further search
24:             $\mathcal{Q}.push((P, \mathcal{N}, \mathcal{R}, \mathcal{B}, \mathcal{V}))$ 
25: Return the triple  $(\mathbf{T}, S, \mathcal{P})$  # Return (Table, Statement, Program Set)

```

Appendix

- TABFACT

- Latent Program Algorithm (LPA)
 - Entity Linking



Appendix

- TABFACT

- Latent Program Algorithm (LPA)
 - Tirgger Words

Trigger	Function
'average'	average
'difference', 'gap', 'than', 'separate'	diff
'sum', 'summation', 'combine', 'combined', 'total', 'add', 'all', 'there are'	ddd, sum
'not', 'no', 'never', "didn't", "won't", "wasn't", "isn't", "haven't", "weren't", "won't", 'neither', 'none', 'unable', 'fail', 'different', 'outside', 'unable', 'fail'	not_eq, not_within, Filter_not_eq, none
'not', 'no', 'none'	none
'first', 'top', 'latest', 'most'	first
'last', 'bottom', 'latest', 'most'	last
'RBR', 'JJR', 'more', 'than', 'above', 'after'	filter_greater, greater
'RBR', 'JJR', 'less', 'than', 'below', 'under'	filter_less, less
'all', 'every', 'each'	all_eq, all_less, all_greater,
['all', 'every', 'each'], ['not', 'no', 'never', "didn't", "won't", "wasn't"]	all_not_eq
'at most', 'than'	all_less_eq, all_greater_eq
'RBR', 'RBS', 'JJR', 'JJS'	max, min
'JJR', 'JJS', 'RBR', 'RBS', 'top', 'first'	argmax, argmin
'within', 'one', 'of', 'among'	within
'follow', 'following', 'followed', 'after', 'before', 'above', 'precede'	before
'follow', 'following', 'followed', 'after', 'before', 'above', 'precede'	after
'most'	most_freq
ordinal	First, second, third, fourth

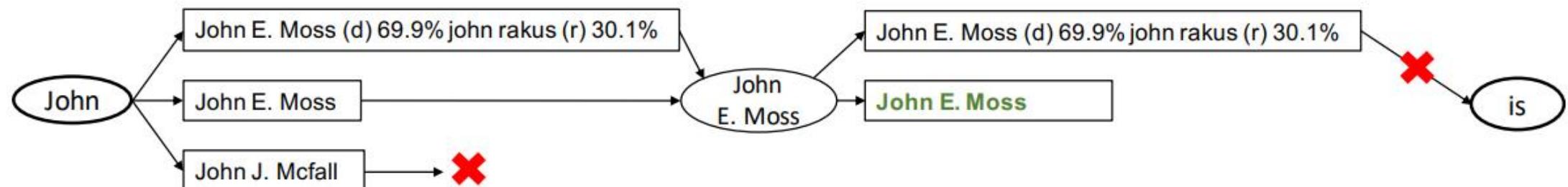
Appendix

- TABFACT

- Entity Linking

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

Statement: John E. Moss is a democratic who is from California 3 district



Appendix

- **Free-form Table Question Answering**

- Both questions and answers is natural and grounded in the context of the entire table
- Retrieving and reasoning over relations of multiple entities

Page Title: Hawaii demographics - ancestry				
Racial composition	1970	1990	2000	2010
White	38.80%	33.40%	24.30%	24.70%
Asian			41.60%	38.60%
Native Hawaiian and other Pacific Islander	57.70%	61.80%	9.40%	10.00%
Black	1.00%	2.50%	1.80%	1.60%
Native American and Alaskan native	0.10%	0.50%	0.30%	0.30%
Q: What ethnic groups are the majorities back in 1970?	A: In 1970, Hawaii's population mainly consists of 38.8% white and 57.7% asian, native hawaiian and other pacific islander.			

Dataset	Answer Format	Avg # Words in Answer
SQuAD (Rajpurkar et al., 2016)	Text-span	3.2
HotpotQA (Yang et al., 2018)	Short-form entity	2.2
NarrativeQA (Kočiský et al., 2018)	Free-form text	4.7
ELI5 (Fan et al., 2019)	Free-form text	130.6
WikiTableQuestions (Pasupat and Liang, 2015)	Short-form entity	1.7
SequenceQA (Saha et al., 2018)	Short-form entity	1.2
HybridQA (Chen et al., 2020e)	Short-form entity	2.1
FeTaQA	Free-form text	18.9

Appendix

- **WikiSQL**

- Inputs consist of a table and a question
- Outputs consist of a SQL query and the result from execution

Table: CFLDraft

Pick #	CFL Team	Player	Position	College
27	Hamilton Tiger-Cats	Connor Healy	DB	Wilfrid Laurier
28	Calgary Stampeders	Anthony Forgone	OL	York
29	Ottawa Renegades	L.P. Ladouceur	DT	California
30	Toronto Argonauts	Frank Hoffman	DL	York
...

Question:

How many CFL teams are from York College?

SQL:

```
SELECT COUNT CFL Team FROM  
CFLDraft WHERE College = "York"
```

Result:

2