NAACL 2025 Main Conference

NAACL

H-STAR: LLM-driven Hybrid SQL-Text Adaptive Reasoning on Tables

Nikhil Abhyankar [1], Vivek Gupta [2], Dan Roth [3], Chandan K. Reddy [1]

[1] Virginia Tech, [2] Arizona State University, [3] University of Pennsylvania




VIRGINIA TECH

Arizona State University

Penn UNIVERSITY of PENNSYLVANIA

Table Reasoning

- Tabular Reasoning involves **reasoning over unstructured text and structured data**.
- It combines natural language understanding with structured data analytics.
- Table Reasoning involves:
  - Fact Verification
    - Tabfact
  - Question Answering (QA)
    - WikiTQ (Short form QA)
    - FetaQA (Long form QA)

row_id	Year	Division	Playoffs	National Cup
0	1935/36	1	Champion	?
1	1936/37	1	DNQ	Champion
...	...	...	...	...
18	1953/54	1	Champion	Champion
19	1954/55	1	No playoff	?

Q: NY Americans did not qualify for playoffs in 1936/37
Evidence: Columns: [year, national cup]; Rows: [1]
A: False
Fact Verification

Q: When did NY Americans win the cup after 1936?
Evidence: Columns: [year, playoffs]; Rows: [1,18]
A: 1953/54
Short-form QA

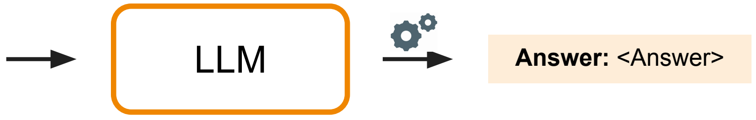
Q: How was the cup performance in 1936/37 and 1953/54?
Evidence: Columns: [year, national cup]; Rows: [1, 18]
A: NY Americans won the national cup in 1936/37 and 1953/54
Long-form QA

Traditional Approaches for Tabular Reasoning

- Traditional methods leveraging Large Language Models (LLMs) use either **semantic reasoning** or **symbolic reasoning** approaches for tabular reasoning.
- Semantic Reasoning

row_id	Year	Division	Playoffs	National Cup
0	1935/36	1	Champion	?
1	1936/37	1	DNQ	Champion
...	...	...	...	...
18	1953/54	1	Champion	Champion
19	1954/55	1	No playoff	?

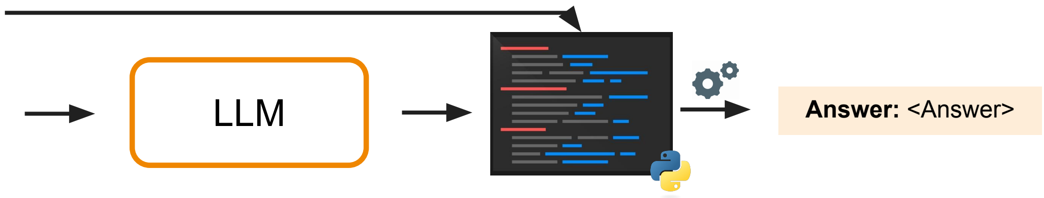
Question: <Question>



- Symbolic Reasoning

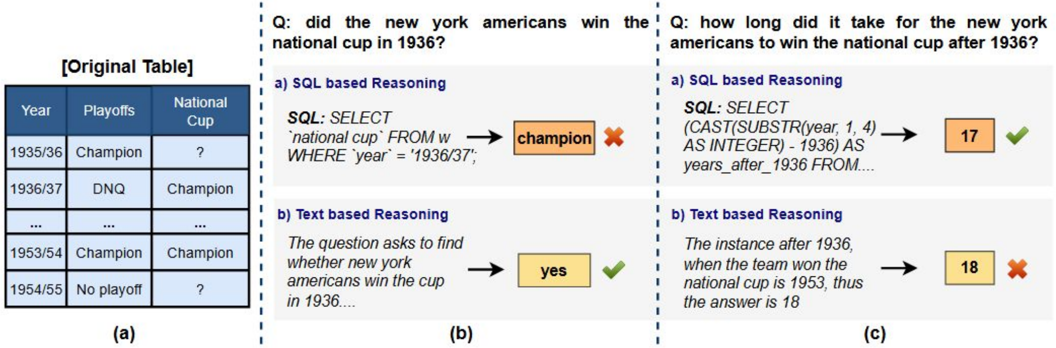
row_id	Year	Division	Playoffs	National Cup
0	1935/36	1	Champion	?
1	1936/37	1	DNQ	Champion
...	...	...	...	...
18	1953/54	1	Champion	Champion
19	1954/55	1	No playoff	?

Question: <Question>



Challenges in Traditional Approaches

Either methods fall short due to the complexities of data and intricate table structures.



Semantic Reasoning

- ✓ Excels in natural language understanding and common-sense queries
- ✓ Handles noisy/ unstructured data
- ✗ Misinterprets table structure for long tables
- ✗ Struggles with quantitative problem-solving

Symbolic Reasoning

- ✗ Struggles with noisy/ unstructured inputs
- ✗ Struggles with complex lexical queries
- ✓ Excels in quantitative reasoning and mathematical reasoning
- ✓ Handles longer table data

H-STAR: A Hybrid Approach

Can we **efficiently integrate both symbolic and textual approaches** into a hybrid method **to leverage their complementary benefits** to enhance tabular reasoning?

	Semantic Reasoning	Symbolic Reasoning	Hybrid
Common-sense/ lexical queries	✓	✗	✓
Noisy/ unstructured data	✓	✗	✓
Long table data	✗	✓	✓
Quantitative problem-solving	✗	✓	✓

H-STAR: A Hybrid Approach

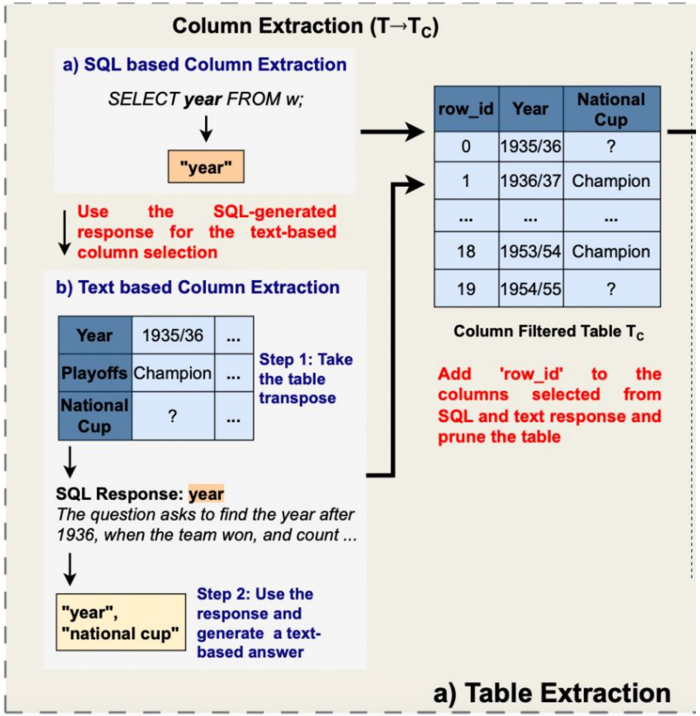
Integrates **symbolic** and **semantic** reasoning to get the **best of both worlds**.

H-STAR consists of:

1. **Table Extraction:** LLMs struggle on reasoning for longer tables.
    - Only **few cells are relevant,** the rest acting as noise leading to hallucinations
    - Use multi-view approach (table transpose) for column extraction followed by row extraction.

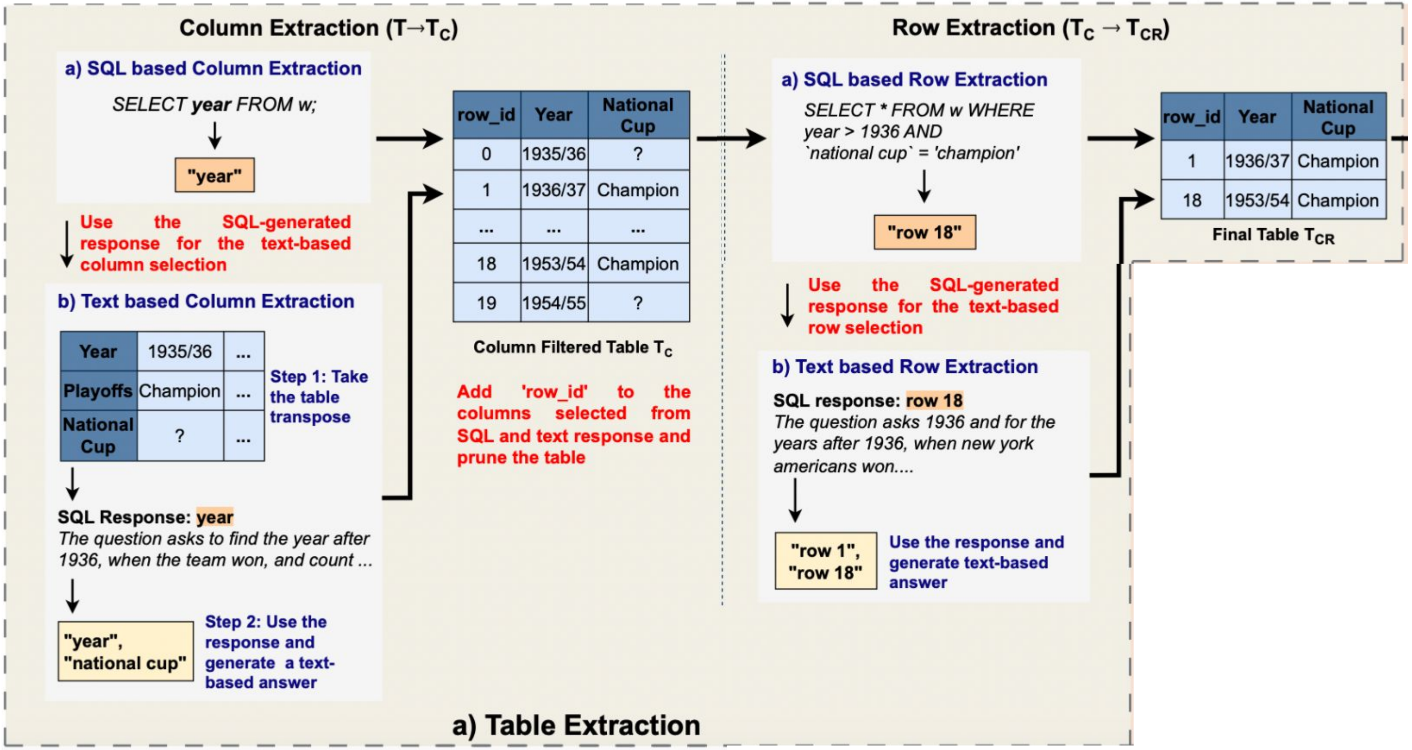
Year	Playoffs	National Cup
1935/36	Champion	?
1936/37	DNQ	Champion
...	...	...
1953/54	Champion	Champion
1954/55	No playoff	?

How many years after 1936 did NY Americans win the national cup?



Year	Playoffs	National Cup
1935/36	Champion	?
1936/37	DNQ	Champion
...	...	...
1953/54	Champion	Champion
1954/55	No playoff	?

How many years after 1936 did NY Americans win the national cup?



H-STAR: A Hybrid Approach

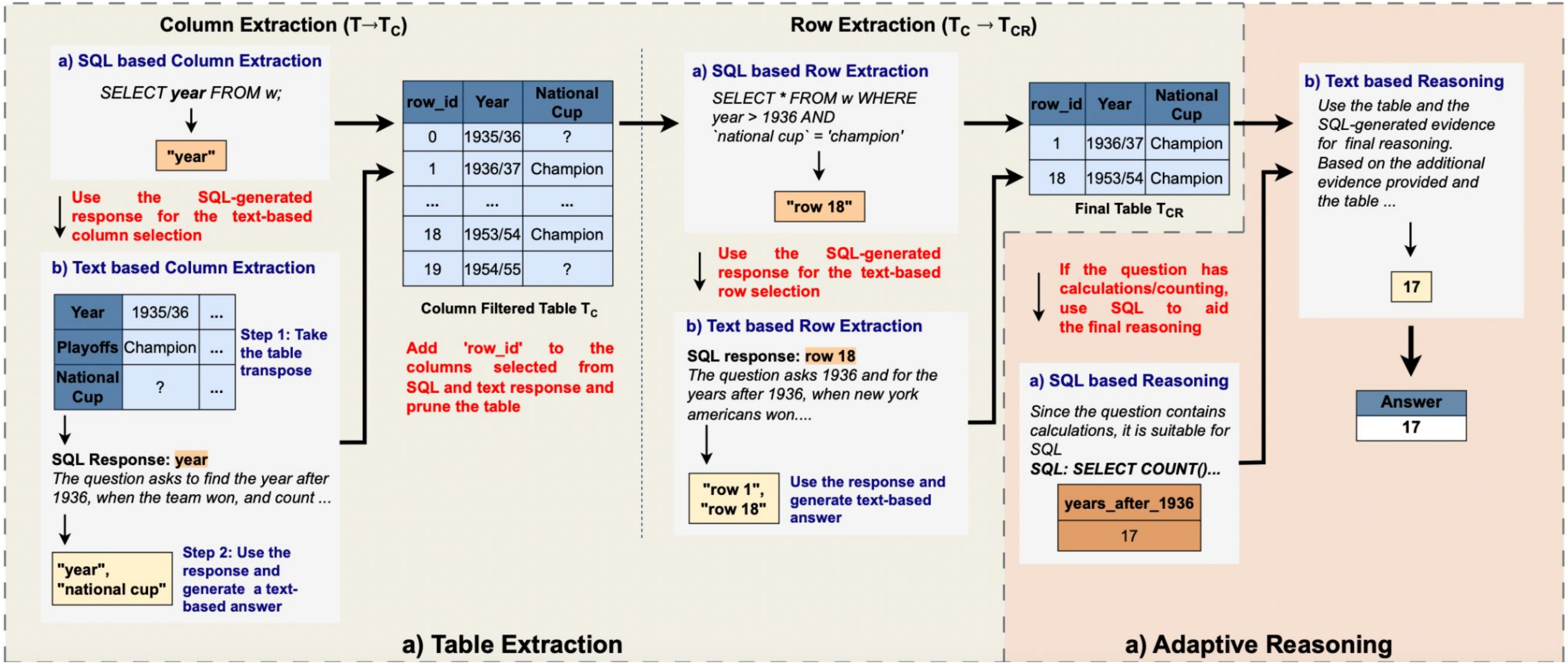
Integrates **symbolic** and **semantic** reasoning to get the **best of both worlds**.

H-STAR consists of:

1. **Table Extraction:** LLMs struggle on reasoning for longer tables.
    - Only **few cells are relevant,** the rest acting as noise leading to hallucinations
    - Use multi-view approach (table transpose) for column extraction followed by row extraction.
2. **Adaptive Reasoning:** LLM chooses between symbolic and semantic methods.
    - Uses **symbolic reasoning for quantitative, mathematical, and logical** tasks.
    - **Semantic reasoning for direct lookup, common-sense, and lexical** queries.

Year	Playoffs	National Cup
1935/36	Champion	?
1936/37	DNQ	Champion
...	...	...
1953/54	Champion	Champion
1954/55	No playoff	?

How many years after 1936 did NY Americans win the national cup?



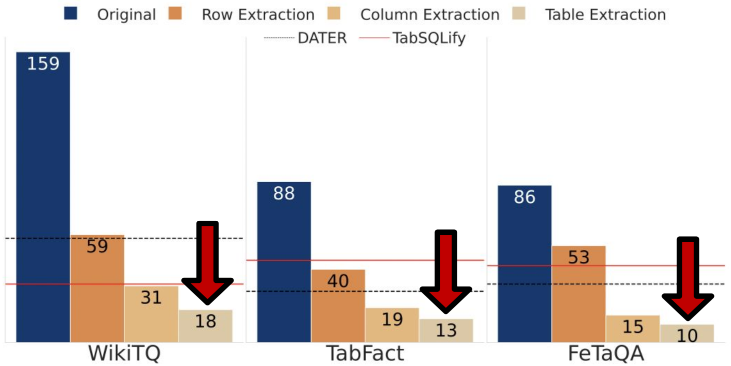
Main Results

	GPT-3.5-Turbo		PaLM-2	
	TabFact	WikiTQ	TabFact	WikiTQ
Generic Reasoning				
End-to-End QA	70.45	51.84	77.92	60.59
Few-shot QA	71.54	52.56	78.06	60.33
CoT	65.37	53.48	79.05	60.43
Table Manipulation				
BINDER	79.17	56.74	76.98	54.88
DATER	78.01	52.90	84.63	61.48
Chain-of-Table*	80.20	59.94	**86.61**	67.31
TabSQLify	79.50	64.70	79.78	55.78
H-STAR	**85.03**	**69.56**	**86.51**	**68.62**

	GPT-4o-mini		Gemini-1.5		Llama-3	
	TF	WTQ	TF	WTQ	TF	WTQ
Generic Reasoning						
End-to-End QA	73.22	59.43	81.12	58.47	78.41	57.89
CoT	75.99	64.31	79.99	64.11	75.34	65.49
Table Manipulation						
TabSQLify	78.30	68.74	79.50	63.92	60.70	66.85
Chain-of-Table	85.09	68.53	86.95	70.05	85.86	70.76
H-STAR	**89.42**	**74.93**	**89.08**	**73.14**	**89.23**	**75.76**

H-STAR outperforms state-of-the-art methods such as **Chain-of-Table, TabSQLify, BINDER, and DATER** across diverse models and datasets!

Effective Table Extraction



Succinct Table Extraction
(# of cells reduce in final extraction)

Method	Small	Medium	Large
BINDER	56.54	25.13	6.41
DATER	62.50	42.34	34.62
Chain-of-Table	68.13	52.25	44.87
TabSQLify	68.15	57.91	52.34
H-STAR	**71.64**	**65.20**	**64.84**

Effective on Longer Tables

H-STAR **efficiently reduces the table size** leading to a **better overall performance**, particularly **over longer tables** (> 4000 tokens).

Ablation Analysis

Method	TabFact	WikiTQ
H-STAR	**86.51**	**68.62**
w/o row extraction	86.17	66.30
w/o column extraction	84.04	67.03
w/o table extraction	83.79	63.58
w/o adaptive reasoning	79.35	61.47

All steps are essential

- Table extraction is essential
- Adaptive reasoning is essential

Method	TabFact	WikiTQ
H-STAR	**86.51**	**68.62**
w/o SQL extraction	85.22	64.39
w/o text extraction	83.74	60.31
w/o SQL reasoning	84.48	64.76
w/o text reasoning	58.70	54.35

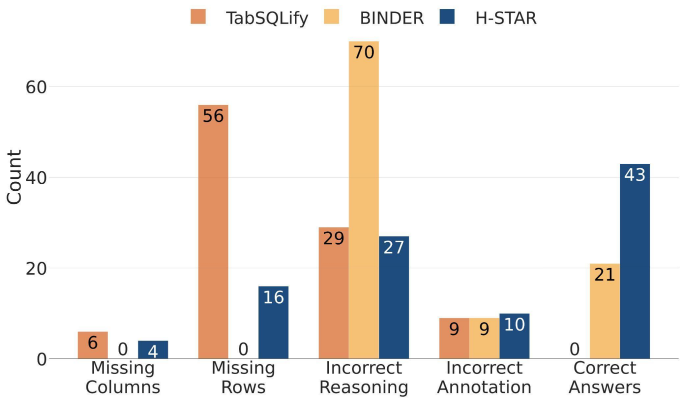
Hybrid approach is essential

- SQL and Text extraction helps
- SQL and Text reasoning helps

Analysis

Error Analysis

- H-STAR is better at table extraction
- Adaptive reasoning works better (27/80 incorrect)



H-STAR outperforms on **FetaQA** long-form Question Answering (Human Evaluation)

Method	Fluent	Correct	Adequate	Faithful
T5-large	94.6	54.8	50.4	50.4
Human	95	92.4	95.6	95.6
TableCoT	96	82	75	87
Tabsqlify	97	88	84	93
H-STAR	96.6	87.6	89.6	94

Summary

- **Integrating symbolic & textual reasoning**, **H-STAR achieves the best of both worlds,** outperforming state-of-the-art approaches for table reasoning
- **Decomposing the task into two modular steps is very effective**.
- **Table Extraction** provides the LLM with the **right context for right reasoning.**
- **Adaptive Reasoning** i.e. augmenting semantic reasoning with symbolic reasoning **is effective.**