EMNLP 2024

TAP4LLM: Table Provider on Sampling, Augmenting, and Packing Semi-structured Data for Large Language Model Reasoning

Yuan Sui[1*], Jiaru Zou[2*], Mengyu Zhou[3#], Xinyi He[4], Lun Du[5], Shi Han[3], Dongmei Zhang[3]

[1]National University of Singapore, [2]University of Illinois Urbana-Champaign, [3]Microsoft,

[4]Xi'an Jiaotong University, [5]Ant Research

yuansui@comp.nus.edu.sg, jiaruz2@illinois.edu, hxyhxy@stu.xjtu.edu.cn,
{mezho,shihan,dongmeiz}@microsoft.com, dulun.dl@antgroup.com

Overview

- Introduction & Demonstration
- Challenges when leveraging LLMs for table reasoning
- Framework of Tab4LLM (e.g., table sampling, table augmentation, table packing & serialization)
- Experiment Results & Findings

Introduction & Demonstration

Utterance: Which category achieves the most sales in 2016?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	…	…	…	…	…

Introduction & Demonstration

Utterance: Which category achieves the most sales in 2016?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	…	…	…	…	…

Introduction & Demonstration

How to leverage LLMs to solve table reasoning tasks?

Utterance: Which category achieves the most sales in 2016?

	Year	**Category**	**Product**	**Sales**	**Rating**
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	...	...	...	...	...

Introduction & Demonstration

How to leverage LLMs to solve table reasoning tasks?

Utterance: Which category achieves the most sales in 2016?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	...	...	...	...	

Which part of a table should be kept in the prompt?

Introduction & Demonstration

How to leverage LLMs to solve table reasoning tasks?

Utterance: Which category achieves the most sales in 2016?

	Year	**Category**	**Product**	**Sales**	**Rating**
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	…	…	…	…	

Which part of a table should be kept in the prompt?

What additional/external knowledge could help LLMs better understand a table? (e.g., Wikipedia, metadata, statistics, etc.)

Framework of TAP4LLM

Utterance: Which category achieves the most sales in 2016?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	...	...	...	...	...

Which part of a table should be kept in the prompt?

Framework of TAP4LLM

Utterance: Which category achieves the most sales in 2016?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	…	…	…	…	…

Which part of a table should be kept in the prompt?

(1) Table Sampling

- *Utterance (user query)*
- *Sampled column headers*
- *Sampled Rows {R1,R3,R4…}*

Sampled Table:

Year	Category	Sales
2016	Components	$20,000
2016	Clothing	$2,300
2016	Accessories	$3,400

Framework of TAP4LLM

Utterance: Which category achieves the most sales in 2016?

Which part of a table should be kept in the prompt?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	…	…	…	…	…

(1) Table Sampling

- *Utterance (user query)*
- *Sampled column headers*
- *Sampled Rows {R1,R3,R4…}*

Sampled Table:

Year	Category	Sales
2016	Components	$20,000
2016	Clothing	$2,300
2016	Accessories	$3,400

- **Table Sampling**: Decompose a large table T into a sub-table T' with specific rows and columns

Framework of TAP4LLM

What additional/external knowledge could help LLMs better understand a table

Utterance: Which category achieves the most sales in 2016?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	...	...	...	...	...

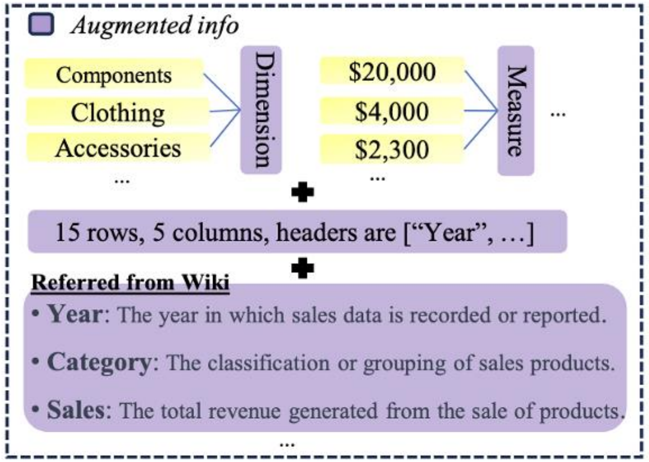
Framework of TAP4LLM

What additional/external knowledge could help LLMs better understand a table

Utterance: Which category achieves the most sales in 2016?

	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	...	...	...	...	...

(2) Table Augmentation



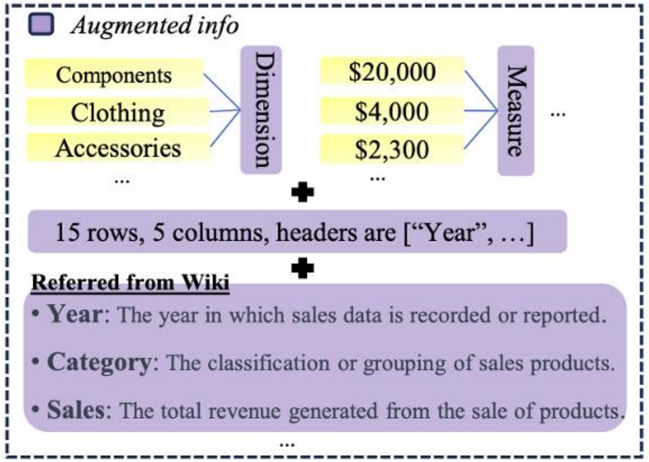
Framework of TAP4LLM

What additional/external knowledge could help LLMs better understand a table

Utterance: Which category achieves the most sales in 2016?

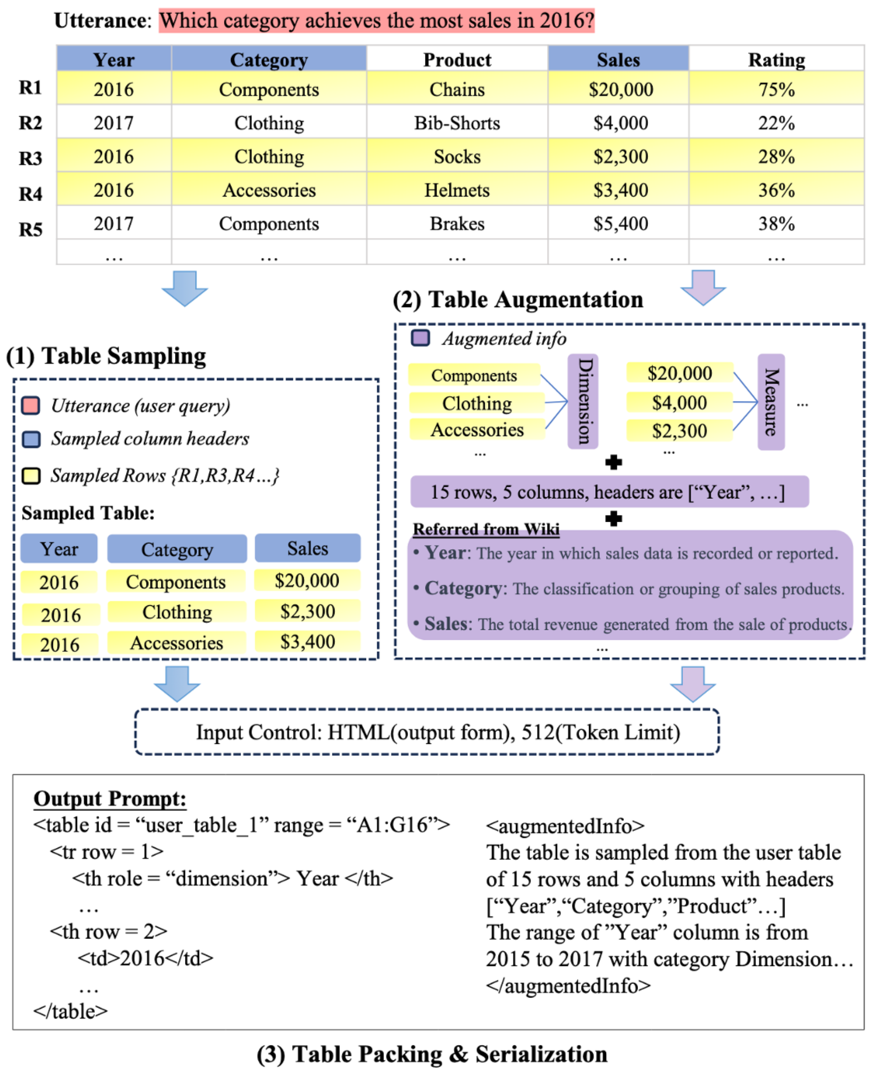
	Year	Category	Product	Sales	Rating
R1	2016	Components	Chains	$20,000	75%
R2	2017	Clothing	Bib-Shorts	$4,000	22%
R3	2016	Clothing	Socks	$2,300	28%
R4	2016	Accessories	Helmets	$3,400	36%
R5	2017	Components	Brakes	$5,400	38%
	…	…	…	…	…

(2) Table Augmentation

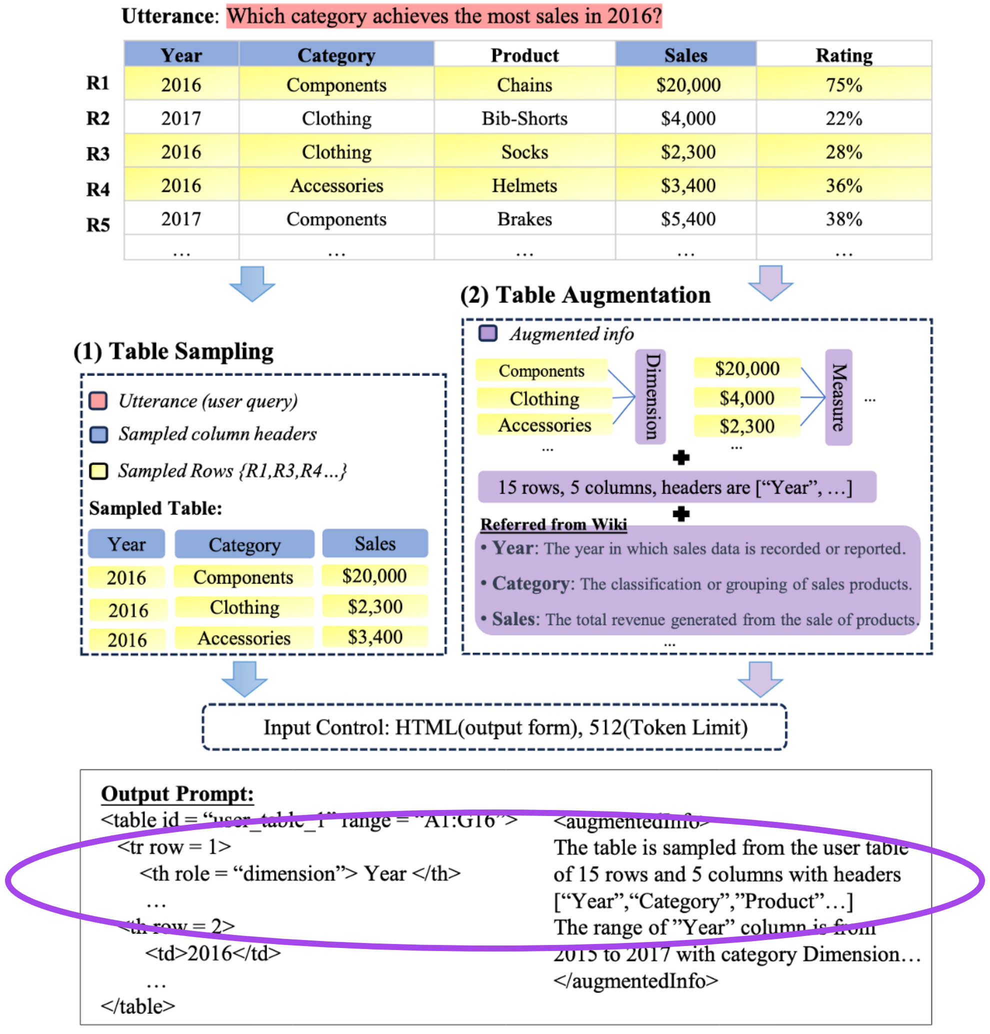


- **Table Augmentation**: Incorporate relevant external knowledge, metadata, and attributes about the original table T explicitly.

Framework of TAP4LLM



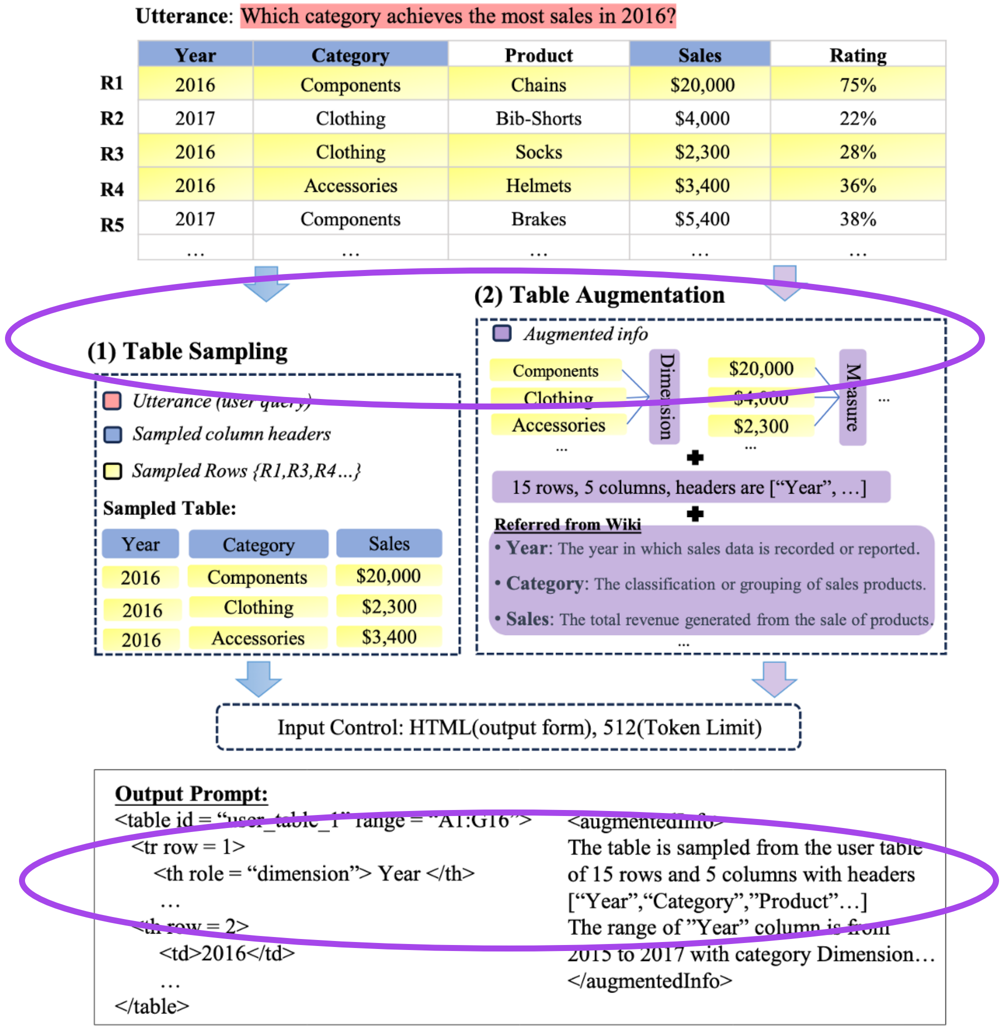
Framework of TAP4LLM



(3) Table Packing & Serialization

Table Packing & Serialization: convert table(s) into various formats suitable for LLMs’ understanding while control the token allocation for table sampling and augmentation.

Framework of TAP4LLM



(3) Table Packing & Serialization

How to encode the table into a prompt, balancing table augmentation and table sampling?

Table Packing & Serialization: convert table(s) into various formats suitable for LLMs' understanding while control the token allocation for table sampling and augmentation.

Trade-off between Table Sampling & Augmentation

Trade-off between Table Sampling & Augmentation

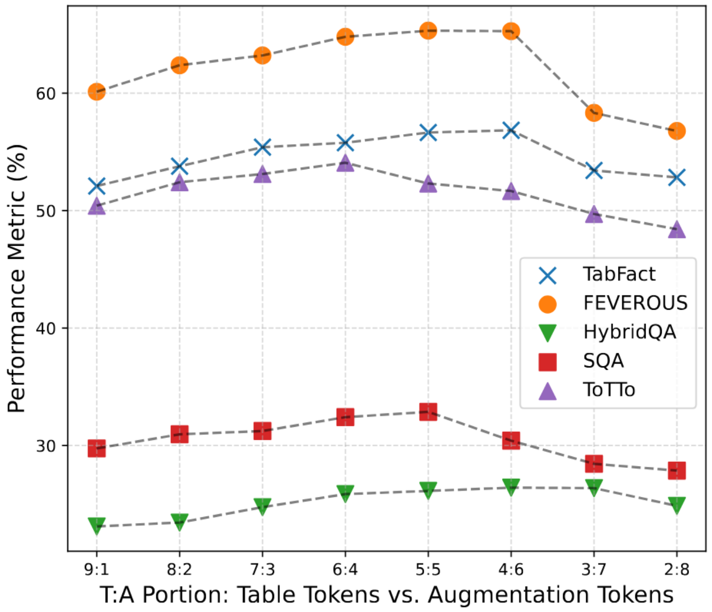
- **Table augmentation** prevents LLMs from partially understanding table(s) after table sampling, which may remove essential rows/columns.

Trade-off between Table Sampling & Augmentation

- **Table augmentation** prevents LLMs from partially understanding table(s) after table sampling, which may remove essential rows/columns.
- It leverages the summarization, statistics, and metadata derived from the entire table to **compromise the trade-off and reduce information loss**.

Trade-off between Table Sampling & Augmentation

- **Table augmentation** prevents LLMs from partially understanding table(s) after table sampling, which may remove essential rows/columns.
- It leverages the summarization, statistics, and metadata derived from the entire table to **compromise the trade-off and reduce information loss**.



A balanced token distribution between the table and augmentation (approximately 5:5 or 4:6, referred to as the balanced T:A ratio)

Experiment Results & Findings

- Table sampling: Focusing on **key** rows/columns can improve LLMs' comprehension of tables

Sampling Type	Table Sampling Methods	SQA	FEVEROUS	TabFact	HybridQA	ToTTo
Rule-based Sampling	Random Sampling	27.30%	60.30%	55.17%	23.60%	40.12%
	Evenly Sampling	26.72%	61.87%	54.63%	5.32%	29.41%
	Content Snapshot (Yin et al., 2020)	28.24%	63.10%	56.92%	23.40%	47.51%
Embedding-based Sampling	Centroid-based Sampling	28.10%	63.50%	55.40%	24.03%	48.30%
	Semantic-based Sampling	28.32%	63.32%	59.80%	24.32%	49.14%
	w/ Column Grounding	**29.12%**	64.74%	60.23%	**25.14%**	**53.42%**
	Hybrid Sampling	28.79%	**65.34%**	**61.37%**	24.71%	51.63%
LLM-based Sampling	LLM-Decomposer (Ye et al., 2023b)	27.98%	62.34%	58.74%	24.98%	48.13%
-	No sampling (GPT-3.5)	27.60%	60.12%	56.20%	14.10%	47.42%
	No sampling (GPT-3.5, truncated)	23.54%	43.54%	52.12%	23.12%	30.42%

Experiment Results & Findings

- Integrating **metadata** or **statistics features** of tables can consistently reduce factual inaccuracies in LLMs and improve overall reasoning performance
- Explaining **unusual terms** in table(s) or adding supplemental relevant web pages as the **references** could further enhance LLMs' understanding of table(s)

Augmentation Aspect	SQA		FEVEROUS		TabFact		HybridQA		ToTTo	
	Acc	Delta	Acc	Delta	Acc	Delta	Acc	Delta	BLEU-4	Delta
baseline	28.32%	0.00%	63.32%	0.00%	59.80%	0.00%	24.32%	0.00%	49.14%	0.00%
D/M + SF	30.12%	1.80%	65.72%	2.40%	**62.67%**	2.87%	26.12%	1.80%	51.25%	2.11%
Table Size	28.85%	0.53%	63.40%	0.08%	60.30%	0.50%	24.94%	0.62%	49.03%	-0.11%
Statistics Feature	31.22%	2.90%	**66.51%**	3.19%	62.33%	2.53%	26.13%	1.81%	50.57%	1.43%
Header Hierarchy	-	-	-	-	-	-	-	-	48.64%	-0.50%
Docs References	**33.45%**	5.13%	63.13%	-0.19%	61.32%	1.52%	25.12%	0.80%	52.74%	3.60%
Term Explanations										
- LLM-based	31.59%	3.27%	64.12%	0.80%	62.32%	2.52%	26.24%	1.92%	**53.21%**	4.07%
- Heuristics-based	29.59%	1.27%	63.72%	0.40%	61.58%	1.78%	25.24%	0.92%	51.21%	2.07%
Self Prompting	30.45%	2.13%	65.24%	1.92%	62.32%	2.52%	**26.64%**	2.32%	52.36%	3.22%

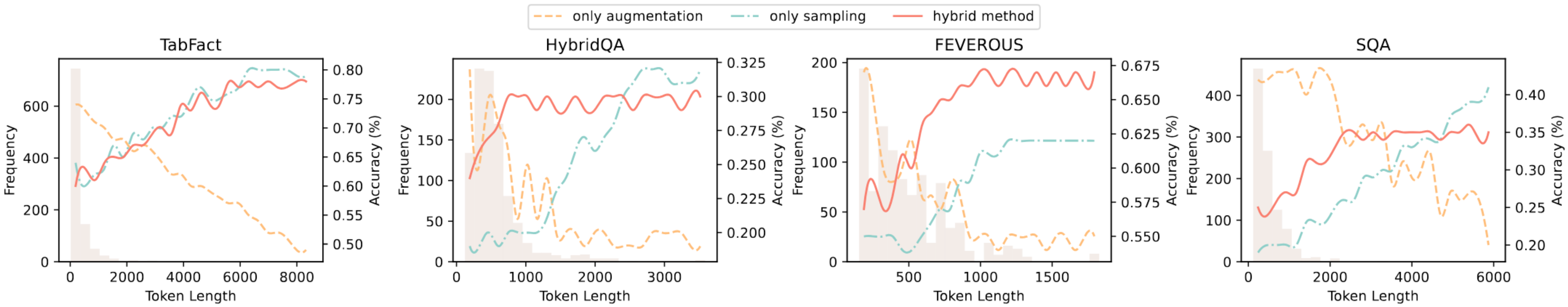
Ablation Study

- All components of TAP4LLM contribute to its performance, with table sampling and augmentation being particularly critical.

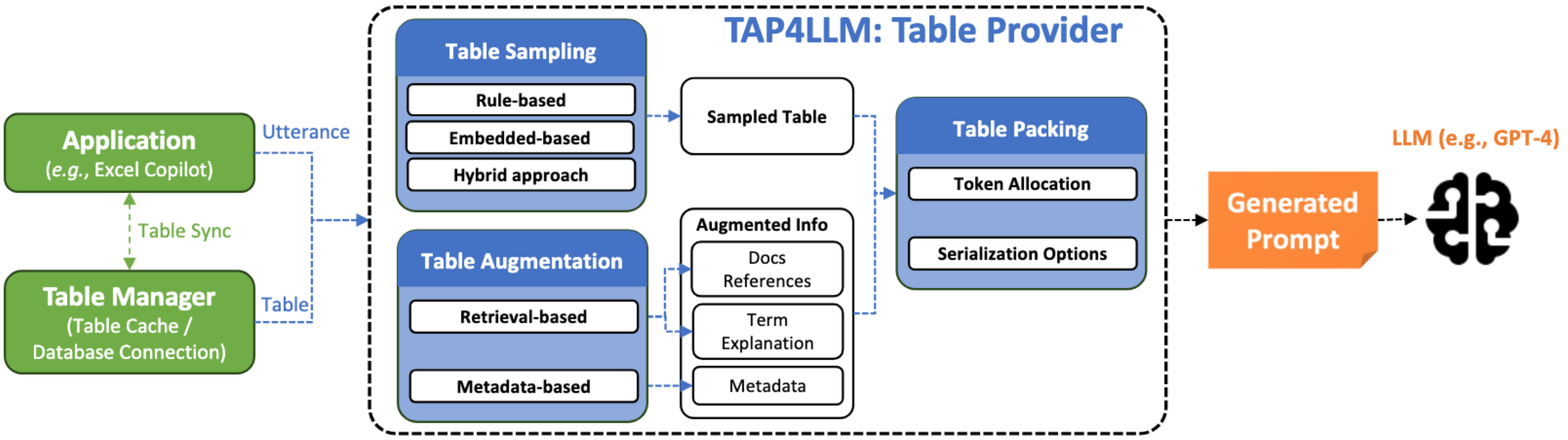
Components of TAP4LLM	SQA		FEVEROUS		TabFact		HybridQA		ToTTo	
	Acc	Delta	Acc	Delta	Acc	Delta	Acc	Delta	BLEU-4	Delta
All	34.12%	0.00%	68.32%	0.00%	64.78%	0.00%	27.87%	0.00%	54.93%	0.00%
w/o table sampling	**26.54%**	-7.58%	**61.54%**	-6.78%	**58.12%**	-6.66%	**24.12%**	-3.75%	**48.47%**	-6.46%
w/o table augmentation - all	29.12%	-5.00%	63.74%	-4.58%	60.23%	-4.55%	25.14%	-2.73%	53.42%	-1.51%
w/o table augmentation - metadata-based	33.87%	-0.25%	64.38%	-3.94%	62.78%	-2.00%	26.98%	-0.89%	53.42%	-1.51%
w/o table augmentation - retrieval-based	31.42%	-2.7%	66.23%	-2.09%	62.97%	-1.81%	26.33%	-1.54%	52.67%	-2.26%
w/o table packing	31.87%	-2.25%	67.42%	-0.90%	63.28%	-1.50%	26.32%	-1.55%	52.87%	-2.06%

Larger Table Analysis

- For smaller table(s), table augmentation typically yields better results, while for larger tables, sampling performs better. This aligns well with human intuition and our understanding of information entropy.



Broader Application & Plugin Module



- **Table manager** acts as in intermediary, managing the data that is either stored locally in a cache or accessed through a database connection.
- **Table sync** is crucial for "interactive table reasoning" and for maintaining data integrity.

Conclusion

Conclusion

- TAP4LLM (Table Provider for LLM) is a powerful toolkit designed to enhance the interaction between LLMs and structured table data.

Conclusion

- TAP4LLM (Table Provider for LLM) is a powerful toolkit designed to enhance the interaction between LLMs and structured table data.
- It provides optimized prompt designs and robust functionalities to ensure high-quality outputs when LLMs process table-related inputs.

Conclusion

- TAP4LLM (Table Provider for LLM) is a powerful toolkit designed to enhance the interaction between LLMs and structured table data.
- It provides optimized prompt designs and robust functionalities to ensure high-quality outputs when LLMs process table-related inputs.
- It enables high flexibility and can serve as a plugin module for various table reasoning pipelines.