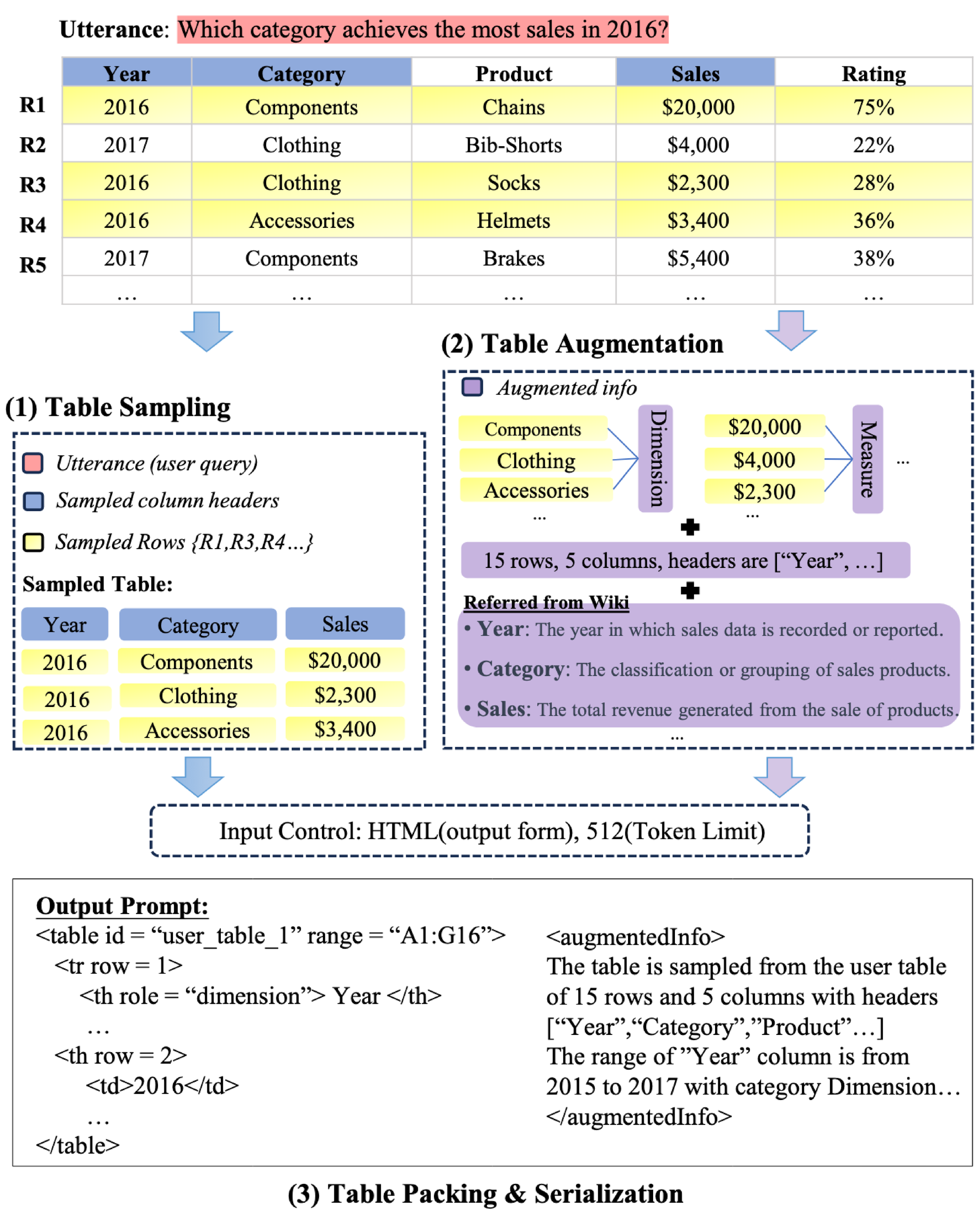
TAP4LLM: Table Provider on Sampling, Augmenting, and Packing Semi-structured Data for Large Language Model Reasoning

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

Introduction & Demonstration

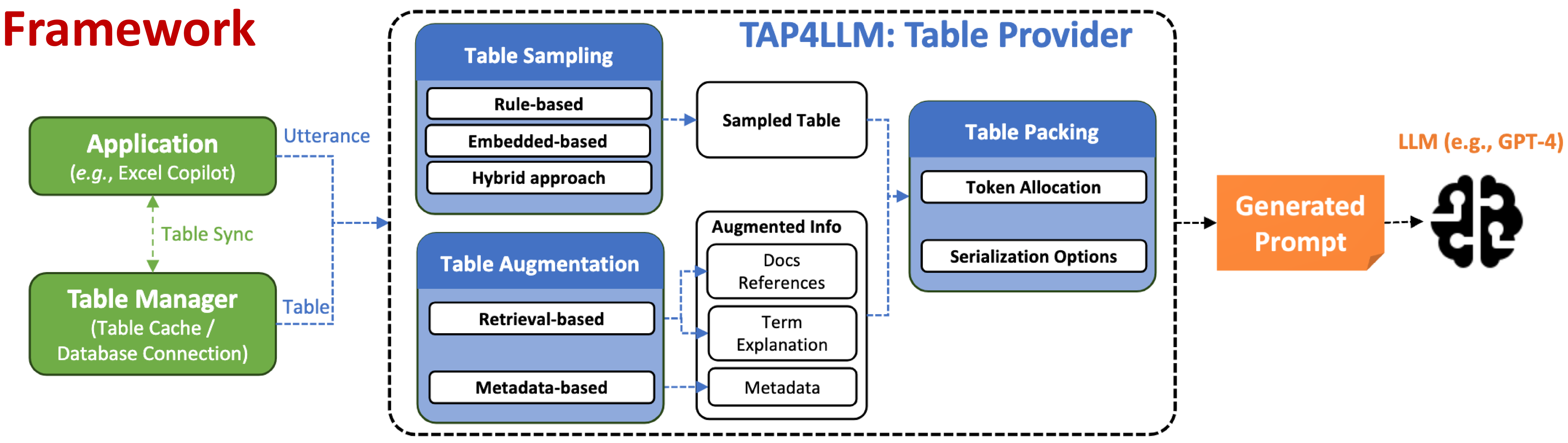


(3) Table Packing & Serialization

Challenges when leveraging LLMs for table reasoning

- Which part of a table should be kept in the prompt?
- What additional/external knowledge could help LLMs better understand a table?
- How to encode the table into a prompt, balancing table augmentation and table sampling?

Framework



- **Table sampling:** decompose a large table into a sub-table with specific rows and columns
- **Table augmentation:** incorporate relevant external knowledge, metadata, and attributes about the original table explicitly
- **Table packing:** convert table(s) into various formats suitable for LLMs' understanding while control the token allocation for table sampling and 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**.

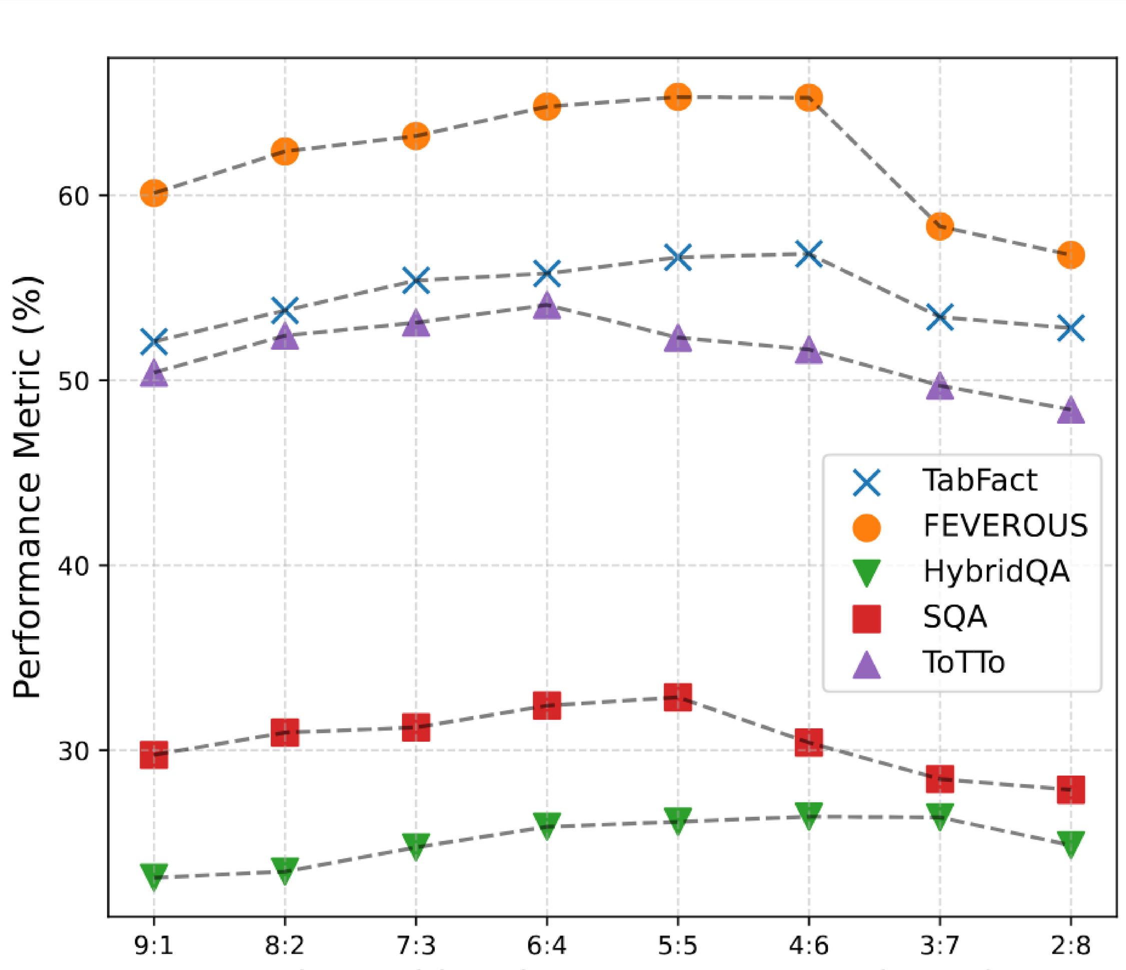
Experiment Results & Findings

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%

- Focusing on **key** rows/columns can improve LLMs' comprehension of tables
- 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)

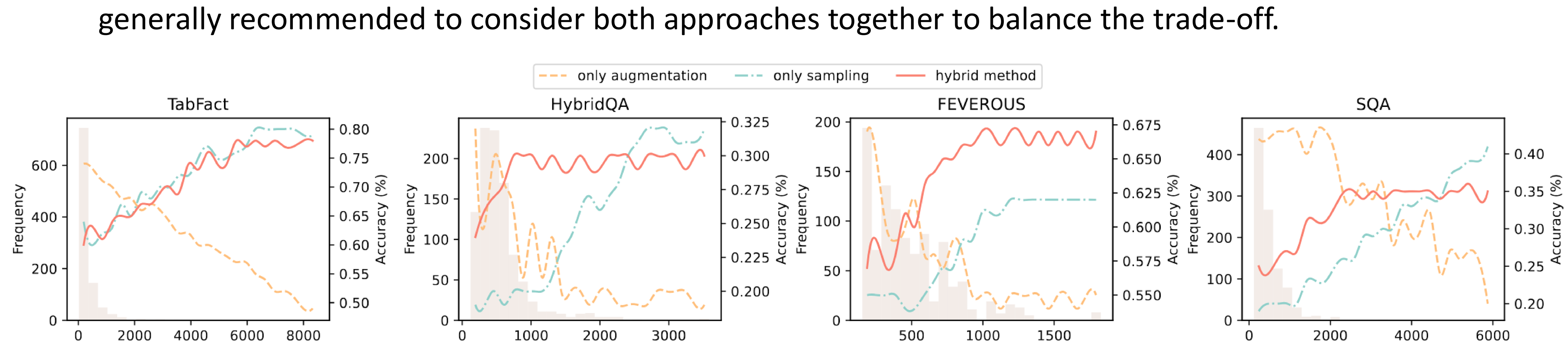
Augmentation Aspect	SQA		FEVEROUS		TabFact		HybridQA		ToTTo	
	Acc	Delta	Acc	Delta	Acc	Delta	Acc	Delta	BLEU-4	Delta
baseline	28.32%	0.00%	63.32%	0.00%	59.80%	0.00%	24.32%	0.00%	49.14%	0.00%
D/M + SF	30.12%	1.80%	65.72%	2.40%	**62.67%**	2.87%	26.12%	1.80%	51.25%	2.11%
Table Size	28.85%	0.53%	63.40%	0.08%	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%

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%



Trade-off between Table Sampling & Augmentation

- **A balanced token distribution** between the table and augmentation (approximately 5:5 or 4:6, referred to as the balanced T:A ratio) generally achieves the best performance across all five datasets.
- 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. It is generally recommended to consider both approaches together to balance the trade-off.



Contact

Yuan Sui (yuansui@comp.nus.edu.sg)
Mengyu Zhou (mezho@microsoft.com)

Acknowledgement

- * indicates equal contribution, # indicates corresponding author.
- The contributions by Yuan Sui, Jiaru Zou and Xinyi He have been conducted and completed during their **interships** at Microsoft.
- The contributions by Lun Du have been conducted and
- completed when he was a **full-time researchers** at Microsoft.