TabSQLify: Enhancing Reasoning Capabilities of LLMs Through Table Decomposition [id-651]

Md Mahadi Hasan Nahid, Davood Rafiei

{mnahid, drafiei}@ualberta.ca Department of Computing Science, University of Alberta, Canada

ALBERTA INNOVATES NSERC CRSNG

1. Overview

TabSQLify, is a method that uses text-to-SQL generation to break down tables into smaller, relevant sub-tables for answering questions or verifying statements. Our approach demonstrates remarkable performance compared to prevailing methods reliant on full tables as input and it significantly reduces input context length, enhancing scalability and efficiency for large-scale table reasoning.

2024 NAACL MEXICO CITY, MEXICO

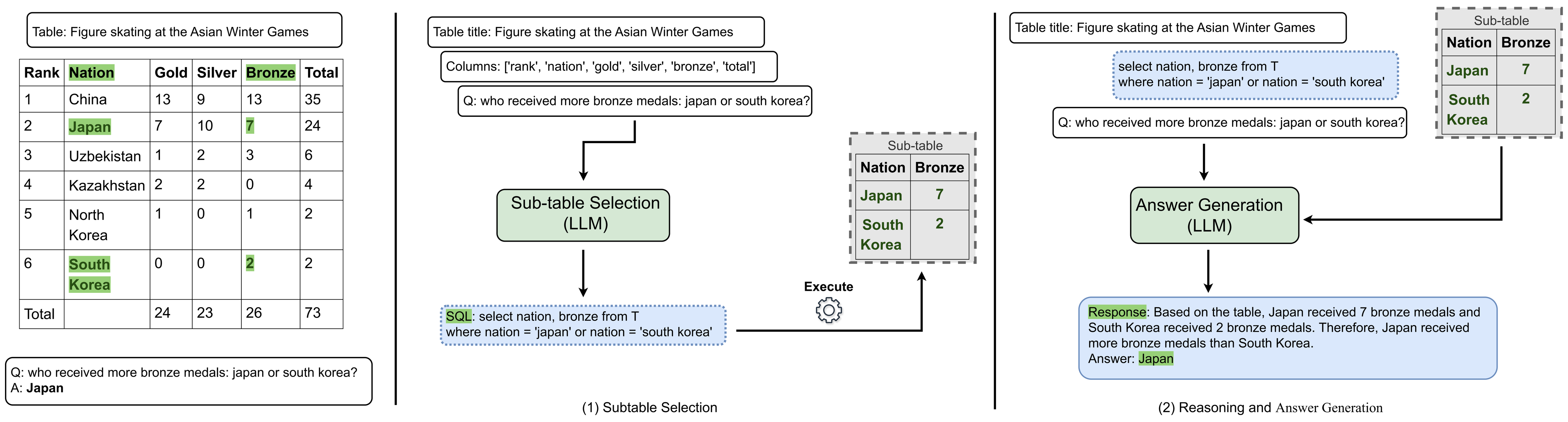
2. Challenges and Key Features

Challenges: (1) Unusual format - table structures, rows, columns, and headers. (2) Prompt Size Limitations - We can only fit a restricted number of tokens. (3) More tokens lead to hallucination and incorrect reasoning. (4) Processing large tables requires additional computational resources and costs.

Key Features: (1) Reducing input length for better scalability and efficiency in reasoning tasks (2) Filtering out irrelevant and redundant information to make the reasoning process more focused (3) Specially useful for large tables (4) Providing an intermediate representation (SQL queries and sub-tables) for improved interpretability and explainability

3. Methodology

TabSQLify consisting of two steps: (1) **Subtable Selection:** generating SQL queries from natural language questions or statements and executing the SQL queries on the original tables to obtain sub-tables containing only essential information, and (2) **Answer Generation:** using LLMs with the sub-table and the question or claim to generate the answer.



4A. Subtable Selection

```
Generate SQL for selecting the required rows and columns, given the question and table to answer the question correctly.
.... .... .... ....
SQLite table properties:

Table: 2012–13 Exeter City F.C. season(row_number,name,league,fa_cup,league_cup,jp_trophy,total)

3 example rows:
select * from T limit 3;
row_number | name | league | fa_cup | league_cup | jp_trophy | total
0 | scot bennett | 5 | 0 | 0 | 0 | 5
1 | danny coles | 3 | 0 | 0 | 0 | 3
2 | liam sercombe | 1 | 0 | 0 | 0 | 1

Q: does pat or john have the highest total?
SQL: select name, total from T where name like '%pat%' or name like '%john%'
.... .... .... ....
SQLite table properties:

Table: Figure skating at the Asian Winter Games (row_number, rank, nation, gold, silver, bronze, total)

3 example rows:
select * from T limit 3;
row_number | rank | nation | gold | silver | bronze | total
0 | 1 | china | 13 | 9 | 13 | 35
1 | 2 | japan | 7 | 10 | 7 | 24
2 | 3 | uzbekistan | 1 | 2 | 3 | 6

Q: who received more bronze medals: japan or south korea?
SQL:
----- ------ ------ ------
response:  select nation, bronze from T where nation = 'japan' or nation = 'south korea'
```

4B. Answer Generation

```
Based on the table title and execution result of the sql query bellow, find the answer to the given question correctly.
.... .... .... ....
Table_title: Playa de Oro International Airport
SQL: select City, Passengers from T;

City | Passengers
United States, Los Angeles | 14,749
United States, Houston | 5,465
Canada, Calgary | 3,761
Canada, Saskatoon | 2,282
Canada, Vancouver | 2,103
United States, Phoenix | 1,829
Canada, Toronto | 1,202
Canada, Edmonton | 110
United States, Oakland | 107

Question: how many more passengers flew to los angeles than to saskatoon from manzanillo airport in 2013?
A: To find the answer to this question, let's think step by step. Based on the table, in 2013, the number of passengers who flew to Los Angeles from Manzanillo Airport was 14,749, while the number of passengers who flew to Saskatoon was 2,282. So, the difference in the number of passengers between Los Angeles and Saskatoon is 14,749 - 2,282 = 12,467. Therefore, the answer is 12,467.
Answer: 12,467
.... .... .... ....
Table_title: Figure skating at the Asian Winter Games
SQL: select nation, bronze from T where nation = 'japan' or nation = 'south korea'

nation | bronze
japan | 7
south korea | 2

Question: who received more bronze medals: japan or south korea?
A: To find the answer to this question, let's think step by step.
----- ------ ------- -------
response:  Based on the table, Japan received 7 bronze medals and South Korea received 2 bronze medals. Therefore, Japan received more bronze medals than South Korea.
Answer: Japan
```

5. Experimental Setup

LLM: gpt-3.5-turbo (4k context size)
Datasets: (1) WikiTableQuestions (2) TabFact (3) FeTaQA (4) WikiSQL

6A. Results - WikiTQ

Models	Accuracy
Agarwal et al., 2019	44.1
Wang et al., 2019	44.5
TaPas	48.8
GraPPa	52.7
LEVER	62.9
ITR	63.4
GPT-3 CoT	45.7
TableCoT-Codex	48.8
DATER-Codex	65.9
BINDER-Codex	61.9
ReAcTable-Codex	65.8
SQL-Codex	61.1
BINDER-chatgpt	55.4
DATER-chatgpt	52.8
ReAcTable-chatgpt	52.5
SQL-chatgpt	54.1
TableCoT-chatgpt	52.4
StructGPT	52.2
Chain-of-Table	59.9
TabSQLify$_{col}$	62.0
TabSQLify$_{row}$	63.7
TabSQLify$_{col+row}$	**64.7**

Table 1: Accuracy compared to the baselines on WikiTQ with the official evaluator.

6B. Results - TabFact

Model	Accuracy
Table-BERT	68.1
LogicFactChecker	74.3
SAT	75.5
TaPas	83.9
TAPEX	85.9
SaMoE	86.7
PASTA	90.8
Human	92.1
TableCoT-Codex	72.6
DATER-Codex	85.6
BINDER-Codex	85.1
ReAcTable-Codex	83.1
ReAcTable-chatgpt	73.1
TableCoT-chatgpt	73.1
BINDER-chatgpt	79.1
DATER-chatgpt	78.0
Chain-of-Table	80.2
TabSQLify$_{col}$	77.0
TabSQLify$_{row}$	78.5
TabSQLify$_{col+row}$	**79.5**

Table 2: Experimental results on TabFact. Here, "Human" indicates the human performance (Ye et al., 2023)

6C. Results - FeTaQA

Model	R-1	R-2	R-L
T5-small	0.55	0.33	0.47
T5-base	0.61	0.39	0.51
T5-large	0.63	0.41	0.53
TableCoT-Codex	0.62	0.40	0.52
DATER-Codex	0.66	0.45	0.56
ReAcTable	0.71	0.46	0.61
TableCoT-chatgpt	0.62	0.39	0.51
TabSQLify$_{col}$	0.57	0.34	0.47
TabSQLify$_{row}$	0.60	0.37	0.49
TabSQLify$_{col+row}$	0.58	0.35	0.48

Table 3: Experimental results on FeTaQA.

Model	Fluency	Correct	Adequate	Faithful
T5-large	94.6	54.8	50.4	50.4
Human (Chen, 2023)	95	92.4	95.6	95.6
TableCoT-chatgpt	96	82	75	87
TabSQLify$_{col}$	98	83	79	85
TabSQLify$_{row}$	96	80	77	89
TabSQLify$_{col+row}$	97	88	84	93

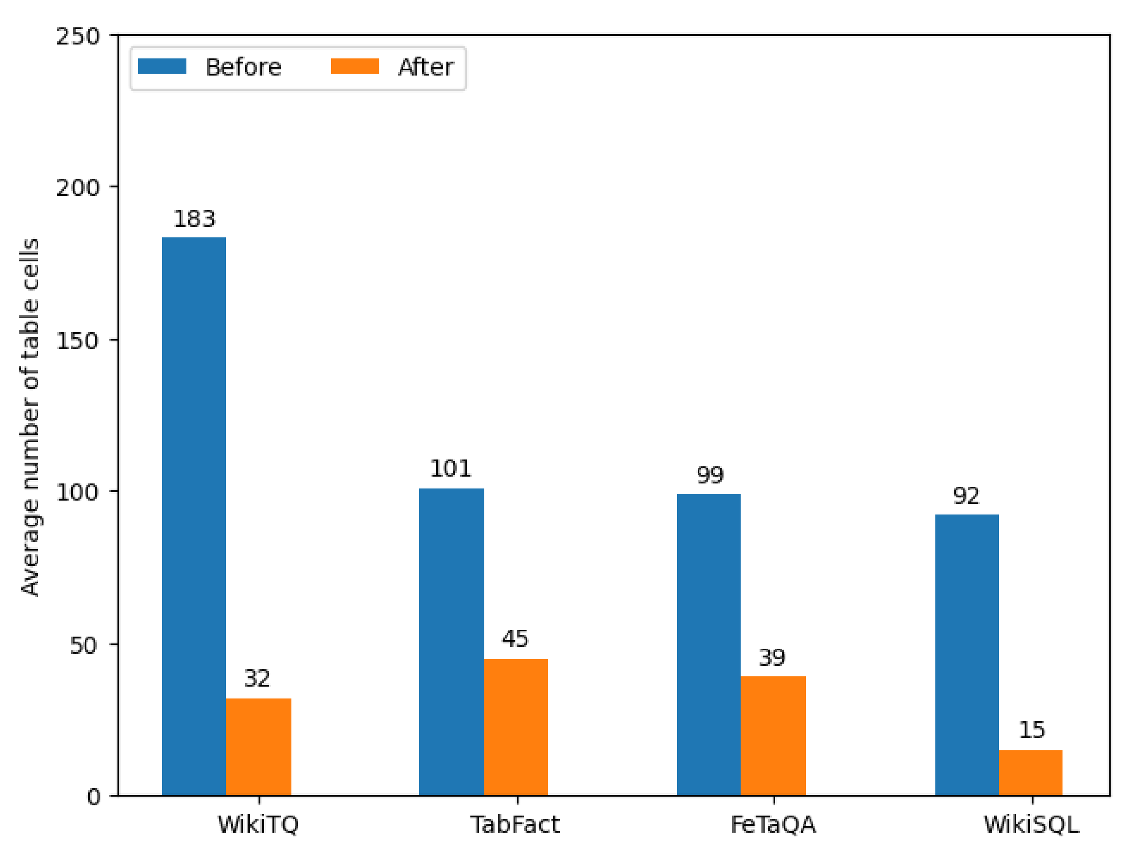
Table 4: Human evaluation results on FeTaQA.

6D. Results - WikiSQL

Model	Accuracy
SEQ2SQL	59.4%
StructGPT	65.6%
RCI (Glass et al., 2021)	89.8%
TabSQLify$_{col+row}$	**76.7%**

Table 5: Experimental results on WikiSQL. RCI is a fine tuning based model, and its results may not be directly comparable due to the model's high reliance on the training set.

7. Reduction



8. Conclusions

Here are some **key takeaways**:

1. Novel approach utilizing Text-to-SQL generation. Decomposes tables into smaller, contextually relevant sub-tables.
2. Substantial reduction in table size. Particularly advantageous for large tables exceeding LLMs' context window.
3. Enhances performance on challenging table reasoning datasets, demonstrating potential for further enhancement.