南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Large Language Models are Versatile Decomposers: Decompose Evidence and Questions for Table-based Reasoning

Yunhu Ye[1,2,‡], **Binyuan Hui**[3,‡], **Min Yang**[2†], **Binhua Li**[3], **Fei Huang**[3], **Yongbin Li**[3†]
[1] University of Science and Technology of China
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] DAMO Academy, Alibaba Group
yeyunhu@mail.ustc.edu.cn, min.yang@siat.ac.cn
{binyuan.hby, shuide.lyb}@alibaba-inc.com
https://github.com/AlibabaResearch/DAMO-ConvAI

SIGIR 2023

研究背景：表格推理任务与应用

核心难点

不仅要理解表格结构，还要在表格数据与自然语言问题之间建立复杂的逻辑映射。

表格推理定义

在半结构化表格数据与非结构化自然语言问题之间，进行多步的数值计算、逻辑判断与文本语义理解的过程。

典型应用场景

表格事实核验 (Table-based Fact Verification, 如 TabFact)

表格问答 (Table QA, 如 WikiTableQuestion, FetaQA)

基于表格的信息检索与知识抽取

关键挑战

跨模态理解： 同时处理自由文本问题与半结构化表格数据

复杂推理能力： 涉及数值比较、聚合运算、逻辑否定等组合操作

此前方法的主要痛点

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

推理瓶颈

Reasoning Bottlenecks

- 规模扩展受限
- 逻辑容易出错
- 泛化能力不足

> 面对“巨大”表格和复杂查询时，现有方法的性能往往会急剧下降，且难以保证中间过程的可靠性。

大表引发性能退化

Token 限制：LLM无法一次性处理包含30+行的大型表格。
噪声干扰：大量无关行列信息会干扰模型的注意力机制，导致判断失误。

复杂问题难分解

信息分散：答案所需的关键线索往往散落在表格的不同角落。
多步推理困难：长难句问题需要多层逻辑拆解，直接推理极易断链。

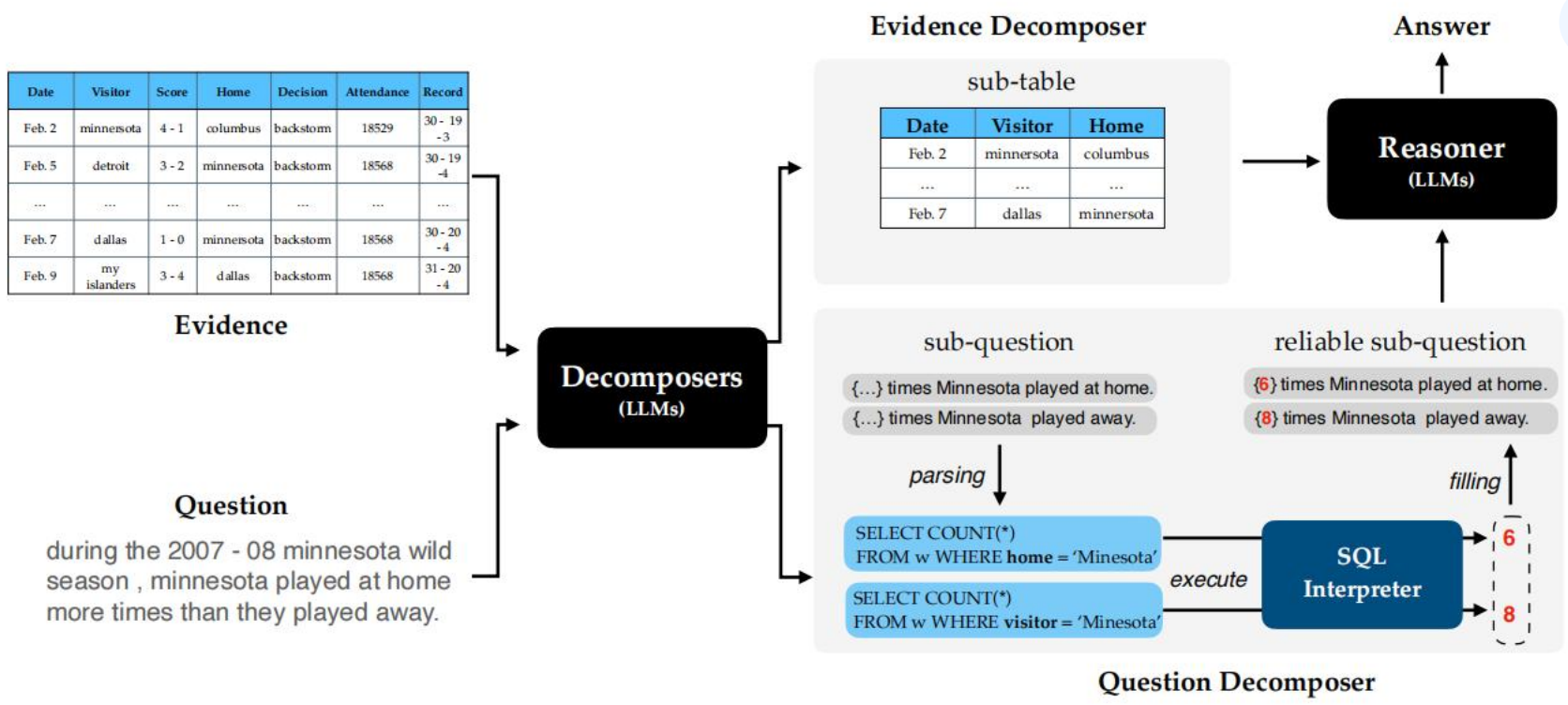
CoT 直接分解易幻觉

符号运算错误：LLM在进行纯文本的数值计算（如加减、比较）时常犯低级错误。
逻辑不一致：生成的推理步骤可能与表格事实相矛盾。

微调依赖与泛化局限

数据依赖：需要大量领域标注数据，成本高昂且难以迁移到新场景。
破坏通用性：针对特定任务微调可能损害大模型原有的In-context Learning能力。

Dater框架



Date	Visitor	Score	Home	Decision	Attendance	Record
Feb. 2	minnesota	4 - 1	columbus	backstorm	18529	30 - 19 - 3
Feb. 5	detroit	3 - 2	minnesota	backstorm	18568	30 - 19 - 4
...	...	...	...	...	...	...
Feb. 7	dallas	1 - 0	minnesota	backstorm	18568	30 - 20 - 4
Feb. 9	my islanders	3 - 4	dallas	backstorm	18568	31 - 20 - 4

Date	Visitor	Home
Feb. 2	minnesota	columbus
...	...	...
Feb. 7	dallas	minnesota

复杂情景分解

将一个复杂任务分解为若干简单任务。

原始复杂问题

全表上的低复杂度证据定位 + 子表上的高复杂度推理

自一致性解码

为了提升推理的稳健性，采用自一致性解码策略。

对同一输入进行多次推理采样

通过投票机制选择出现频率最高的答案，减少随机错误

输入整合与提示

将分解后的**子证据**与经过执行验证的**可靠子问题**作为上下文。

使用少样例提示，引导模型综合信息得出结论

覆盖常见操作：计数、数值比较、逻辑聚合等

全链路可解释性

数据层面：子表剔除无关干扰，行/列索引明确，证据范围可查

逻辑层面：子问题展示推理步骤，中间执行结果清晰可见

方法：证据分解器 (Evidence Decomposer)

南京航空航天大學 Nanjing University of Aeronautics and Astronautics

目标与方法

核心目标

从包含大量冗余信息的"巨型表格"中，精准提取
问题相关的"子表格"，降低干扰。

实现方法

利用LLM的In-context Learning能力，
示（Few-shot Prompting）预测相关的

```
/*
{
  "table_caption": "gambrinus liga",
  "columns": ["season", "champions", "runner - up", "third place", "top goalscorer", "club"],
  "table_column_priority": [
    ["season", "1993 - 94", "1994 - 95", "1995 - 96"],
    ["champions", "sparta prague (1)", "sparta prague (2)", "slavia prague (1)"],
    ["runner - up", "slavia prague", "slavia prague", "sigma olomouc"],
    ["third place", "ban\u00edk ostrava", "fc brno", "baumit jablonec"],
```

```
/*
table caption : 1972 vfl season.
col : home team | home team score | away team | away team score | venue | crowd | date
row 1 : st kilda | 13.12 (90) | melbourne | 13.11 (89) | moorabbin oval | 18836 | 19 august 1972
row 2 : south melbourne | 9.12 (66) | footscray | 11.13 (79) | lake oval | 9154 | 19 august 1972
row 3 : richmond | 20.17 (137) | fitzroy | 13.22 (100) | mcg | 27651 | 19 august 1972
row 4 : geelong | 17.10 (112) | collingwood | 17.9 (111) | kardinia park | 23108 | 19 august 1972
row 5 : north melbourne | 8.12 (60) | carlton | 23.11 (149) | arden street oval | 11271 | 19 august 1972
row 6 : hawthorn | 15.16 (106) | essendon | 12.15 (87) | vfl park | 36749 | 19 august 1972
*/
statement : what is the away team with the highest score?
explain : the statement want to ask the  away team of highest away team score.
the highest away team score is 23.11 (149). it is on the row 5.so we need row 5.
The answer is : f_row([row 5])
```

流程与优势

输入 完整表格 + 用户问题

输出 行索引集合 {Row_i} + 列索引

> Example: col(name, cost), row(1, 3, 13)
> → 生成仅包含姓名、成本列及第1/3/13行数据的子表

- 关键词匹配
- 粗粒度筛选
- 适配超长表格
- 具备可解释性

方法：问题分解器 (Question Decomposer)

南京航空航天大學 Nanjing University of Aeronautics and Astronautics

分解流程示例

原始复杂问题
"Minnesota played at home more times than..."

↓

Cloze 抽象
"Minnesota played {...} times at home."

↓

SQL 生成
SELECT COUNT(*) FROM w WHERE home='min'

↓

回填结果
"Minnesota played { 6 } times at home."

消除幻觉，确保一致性

Cloze-style 完形填空式抽象

使用 LLM 将自然语言中的具体数值片段用占位符 {...} 掩蔽，仅保留逻辑骨架。

将"值"与"逻辑"分离，避免直接生成错误数值。

逻辑到 SQL 转换

将抽象后的逻辑语句转化为可执行的 SQL 查询语句（类似 Text-to-SQL 解析）。

利用编程语言的严谨性来规范推理过程。

执行与回填 (Execution & Filling)

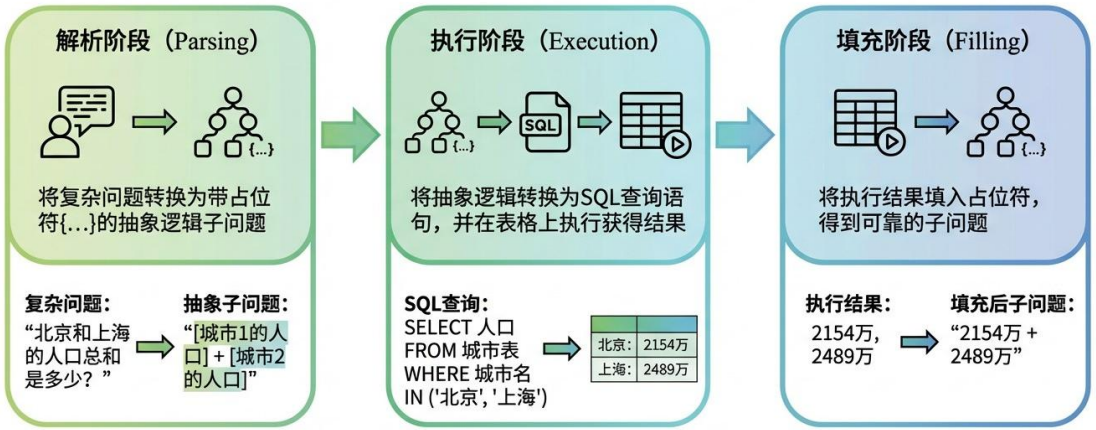
在子证据（Sub-table）上执行 SQL，获取真实结果并回填至占位符中。

生成“可靠子问题 (Reliable Sub-questions)”，作为后续推理的坚实基础。

核心优势

相比传统 CoT，通过引入符号执行（Symbolic Execution），解决了大模型在复杂计算和逻辑推理中常见的幻觉问题。

方法：问题分解器 (Question Decomposer)



Parsing-Execution-Filling 三步策略

1

Parsing (解析)

复杂问题 → 抽象逻辑子问题

将自然语言问题中的具体数值片段用占位符 {...} 掩蔽，保留问题的逻辑骨架，避免直接生成数值导致的幻觉。

2

Execution (执行)

抽象逻辑 → SQL执行结果

将抽象逻辑转化为可执行的SQL查询语句（如COUNT, SUM, FILTER），并在证据表/子表上运行，获取精确的计算结果。

3

Filling (回填)

执行结果 → 可靠子问题

将SQL执行得到的准确数值回填入占位符，形成语义完整且数值可靠的子问题，作为后续推理的坚实基础。

消除数值幻觉 | 逻辑可追溯 | 计算精确性

实验结果

MODEL	TEST
♡ *Fine-tuning based Methods*	
Table-BERT	68.1
LogicFactChecker	74.3
SAT	75.5
TaPas	83.9
TAPEX	85.9
SaMoE	86.7
PASTA	90.8
w/ DATER	**93.0** (↑ 2.2)
Human	92.1
♠ *LLM based Methods*	
Binder	85.1
Codex	72.6
w/ DATER	**85.6** (↑ 13.0)

Table 1: Experimental results on the official small test set of TabFact. Here, "*Human*" indicates the human performance.

MODEL	DEV	TEST
♡ *Fine-tuning based Models*		
MAPO	42.7	43.8
MeRL	43.2	44.1
LatentAlignment	43.7	44.5
IterativeSearch	43.1	44.7
T5-3B	-	49.3
TaPas	49.9	50.4
TableFormer	51.3	52.6
TAPEX	58.0	57.2
ReasTAP	58.3	58.6
TaCube	60.9	60.8
OmniTab	61.3	61.2
w/ DATER	**62.5** (↑ 1.2)	**62.5** (↑ 1.3)
♠ *LLM based Methods*		
Binder	62.6	61.9
Codex	49.3	47.6
w/ DATER	**64.8**(↑ 15.5)	**65.9** (↑ 18.3)

Table 2: Experimental results on WikiTableQuestion with the official evaluator.

- DATER 在 TabFact 与 WTQ 两类任务上均实现稳定性能提升
- 对强基线模型仍有效（TabFact +2.2%，WTQ +1.3%）
- 对 LLM 提升显著（最高 +18.3%），有效弥补表格推理能力不足
- TabFact 上超过人类水平，WTQ 上达到 SOTA
- 证明"表格分解 + 问题分解"具有良好的通用性与扩展性

Thanks