# **GRIT**: Guided Relational Integration for Efficient Multi-Table Understanding

Yujin Kang [1], Seong Woo Park [2], Yoon-Sik Cho [1]

[1] Chung-Ang university , [2] Intellivix, Republic of Korea

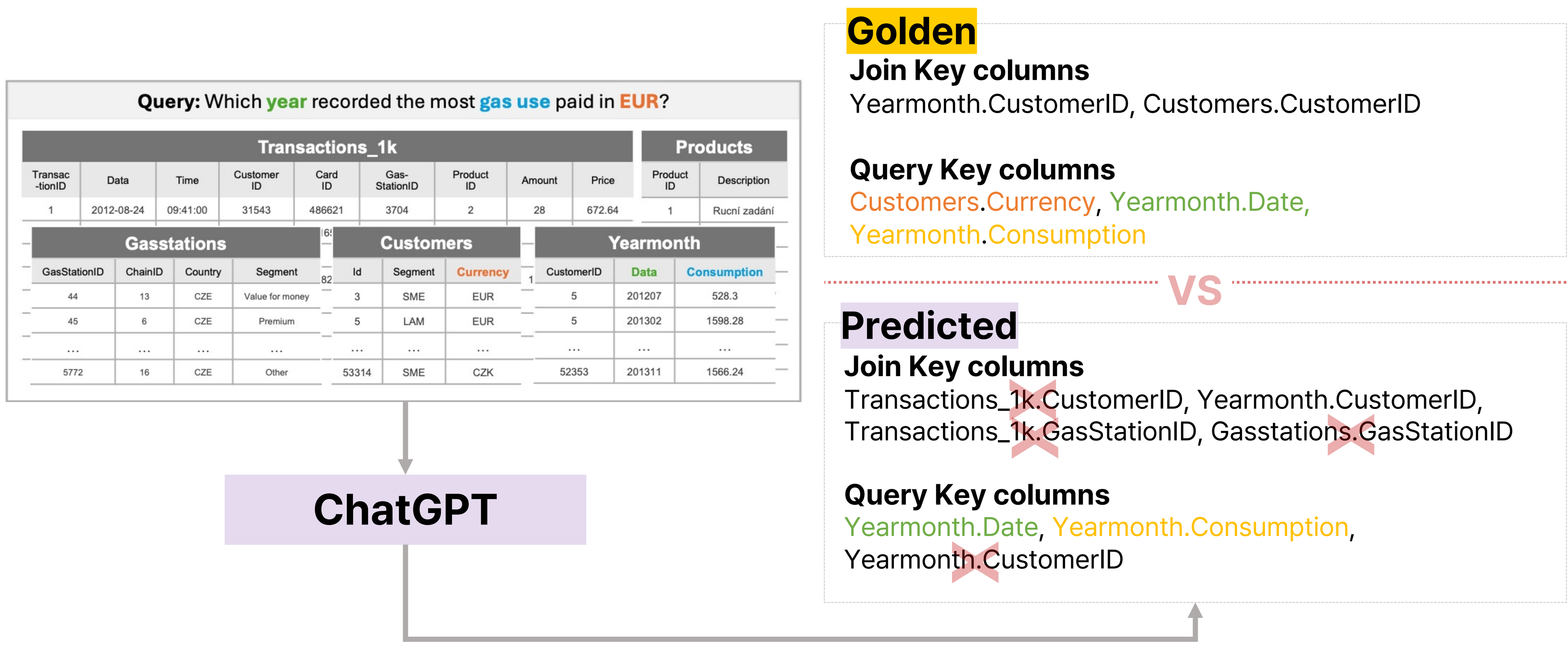
## Motivation

### Limited LLM Input size for Multi-table processing

- Multi-table processing requires integrating information from multiple tables
- Only a portion of the data can be processed when inputting actual databases

### Lack of table structure understanding in LLMs

- Existing LLM training is primarily based on text data.
- Text is sequential and order-dependent, while tables are bi-directional and order-independent.



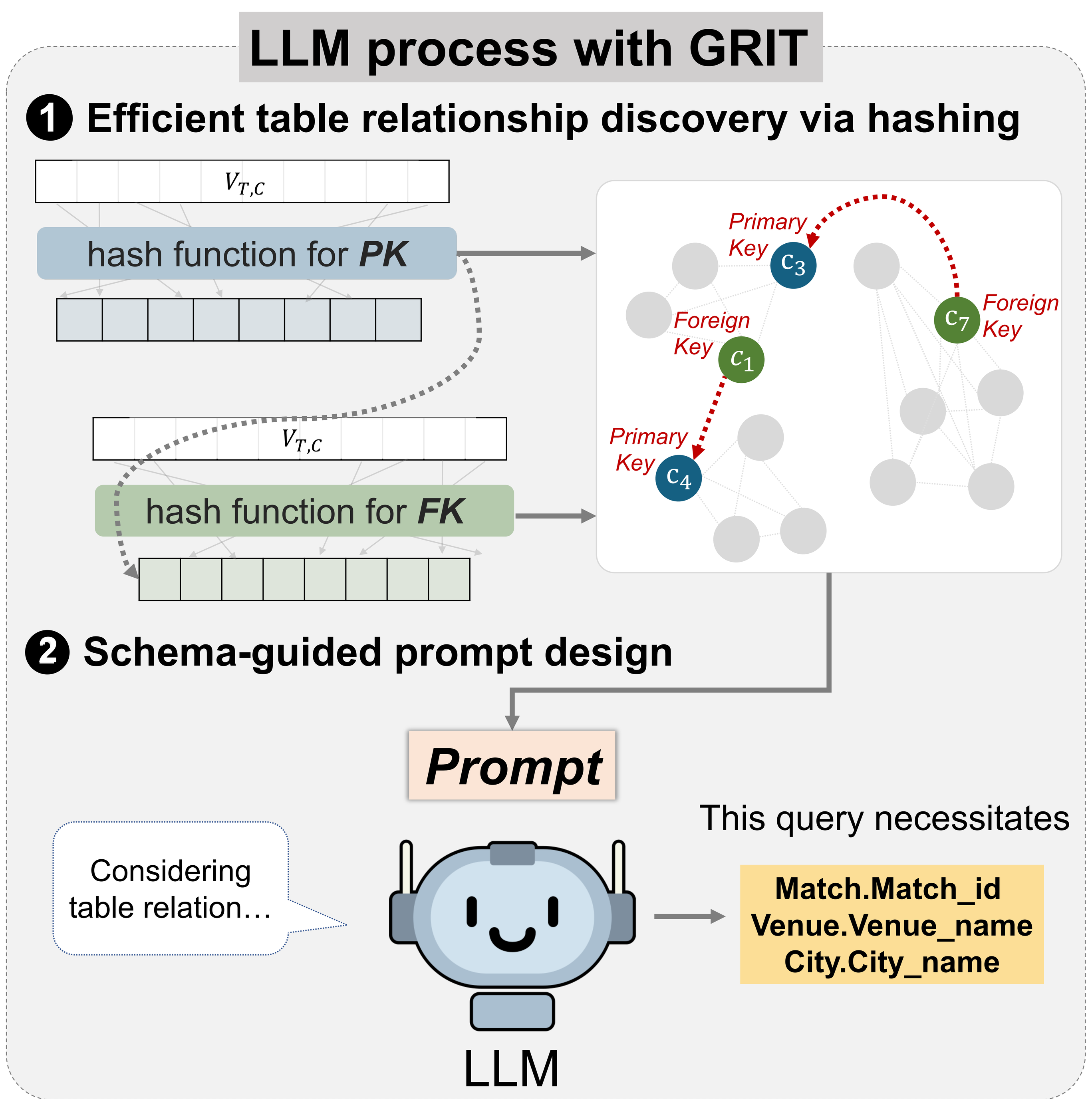

✓ LLMs cannot process the full scale of large multi-table data

✓ LLMs struggle to understand table structures as effectively as text.

➔ Therefore, to enable **effective** multi-table reasoning, the information needs to be transformed into a text form that LLMs can **efficiently** interpret.

## Methodology

**Let's efficiently extract the relationships among multi-tables and deliver them to the LLM in a form it can easily understand.**




### 1 Efficient table relationship discovery via hashing

**1. Primary key detection**

- Use **HyperLogLog (HLL)** to approximate distinct counts efficiently in large-scale tables

**2. Foreign key detection**

To avoid false matches, combine multiple signals:

1) **Containment score**
   - Use **Bloom Filter** for efficient membership testing

$$\phi_{\text{con}}(C, C_{\text{pk}}) = \frac{1}{|V_{T,C}|} \sum_{v \in V_{T,C}} \mathbf{1}\{BF(v) = 1\}$$

2) **Cardinality ratio**: unique(FK) / unique(PK)
   - Use **HyperLogLog** to count uniqueness efficiently.

$$\phi_{\text{card}}(C, C_{\text{pk}}) = \frac{\text{HLL}(V_{T,C})}{\text{HLL}(V_{T_{\text{pk}}, C_{\text{pk}}})}$$

3) **Name similarity**: token overlap between column names

$$\phi_{\text{name}}(C, C_{\text{pk}}) = \frac{2 \cdot f(\mathcal{T}_C, \mathcal{T}_{C_{\text{pk}}})}{|\mathcal{T}_C| + |\mathcal{T}_{C_{\text{pk}}}|}.$$

### 2 Schema-guided prompt design

- Provide **only table headers with schema**
- Explicitly encode **PK–FK relationships** discovered by GRIT
- Transform schema into **LLM-friendly textual representation**

## Experiment

| Model | Input | Bird | | Spider | |
|---|---|---|---|---|---|
| | | Join-key | Query-key | Join-key | Query-key |
| *Closed-source* LLM | | | | | |
| GPT-3.5 Turbo | Header | 64.05 | 59.39 | 69.81 | 71.28 |
| | + GRIT | 67.98 (+ 6.14%) | 60.74 (+ 2.27%) | 73.32 (+ 5.03%) | 71.86 (+ 0.82%) |
| Claude Haiku | Header | 64.68 | 73.12 | 81.43 | 79.41 |
| | + GRIT | 73.08 (+ 12.98%) | 74.13 (+ 1.38%) | 86.22 (+ 5.88%) | 82.32 (+ 3.67%) |
| Claude Sonnet | Header | 80.63 | 74.86 | 87.06 | 80.48 |
| | + GRIT | 82.16 (+ 1.90%) | 75.33 (+ 0.62%) | 89.73 (+ 3.06%) | 82.26 (+ 2.21%) |
| Grok3 | Header | 78.76 | 74.38 | 84.33 | 80.74 |
| | + GRIT | 79.88 (+ 1.43%) | 75.33 (+ 1.28%) | 86.55 (+ 2.64%) | 81.38 (+ 0.80%) |
| *Open-source* LLM | | | | | |
| LLaMA3 | Header | 24.00 | 33.77 | 26.86 | 52.75 |
| | + GRIT | 34.02 (+ 41.76%) | 36.48 (+ 8.03%) | 44.95 (+ 67.32%) | 55.51 (+ 5.23%) |
| Mistral | Header | 4.28 | 44.45 | 19.33 | 55.62 |
| | + GRIT | 8.63 (+ 101.74%) | 48.88 (+ 9.98%) | 37.30 (+ 92.95%) | 56.86 (+ 2.22%) |

Performance comparison of LLMs in table-column retrieval

| Database | Profile | | | Efficiency (MB) | |
|---|---|---|---|---|---|
| | # Rows | # Columns | # Tables | Jaccard-join | GRIT |
| financial | 1,079,680 | 55 | 8 | 3486.36 | 26.63 (-99.24%) |
| card games | 803,451 | 117 | 6 | 1881.34 | 13.63 (-99.28%) |
| codebase community | 740,646 | 71 | 8 | 2107.29 | 9.80 (-99.53%) |
| formula 1 | 514,297 | 96 | 13 | 768.31 | 11.17 (-98.55%) |
| debit card specializing | 423,051 | 23 | 5 | 1260.92 | 10.00 (-99.21%) |
| european football 2 | 222,803 | 201 | 7 | 2613.49 | 5.75 (-99.78%) |
| toxicology | 49,813 | 11 | 4 | 43.21 | 0.78 (-98.19%) |
| california schools | 29,941 | 89 | 3 | 232.67 | 0.76 (-99.67%) |
| thrombosis prediction | 15,952 | 64 | 3 | 101.66 | 0.52 (-99.49%) |
| superhero | 10,614 | 31 | 10 | 9.03 | 0.23 (-97.45%) |

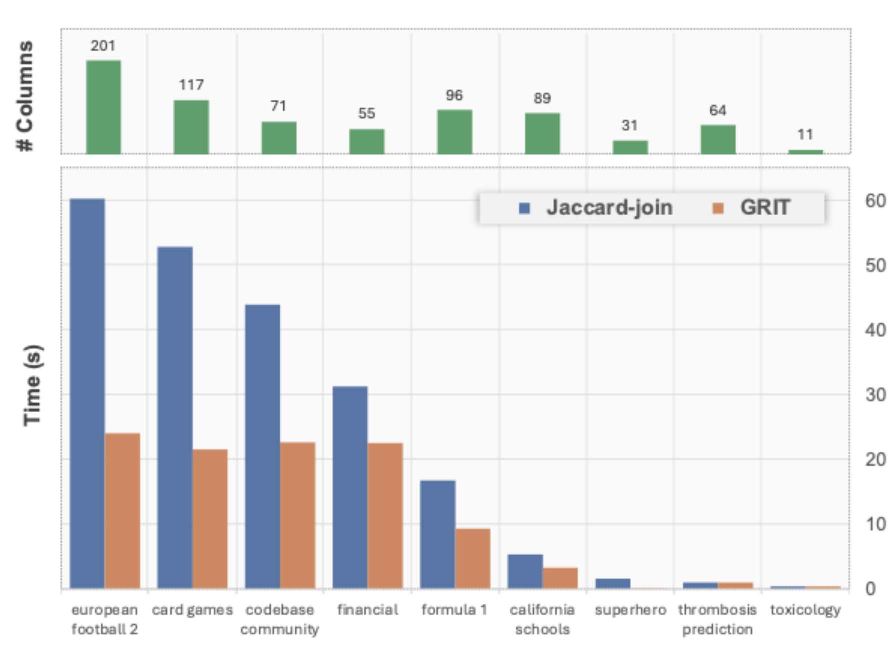
Memory consumption comparison between Jaccard-join and GRIT

| Model | Input | |
|---|---|---|
| | Header | + GRIT |
| GPT-3.5 Turbo | 34.68 | 53.71 (+ 54.89%) |
| Claude Haiku | 39.11 | 55.14 (+ 41.0%) |

Performance in text-to-SQL task

| | Precision | Recall | F1 |
|---|---|---|---|
| Primary Key | 72.23 | **99.17** | 82.19 |
| Foreign Key ($\phi_{\text{cont}} + \phi_{\text{name}} + \phi_{\text{card}}$) | 95.13 | **98.85** | 96.81 |
| $-\phi_{\text{name}}$ | 83.85 | 83.51 | 82.86 |
| $-\phi_{\text{card}}$ | 74.36 | 93.95 | 80.51 |
| $-\phi_{\text{name}} - \phi_{\text{card}}$ | 71.59 | 87.70 | 76.71 |

Detection performance of PK and FK




Comparison of PF-FK relationship construction time across databases

## Conclusion

- Proposed **GRIT**, a lightweight hashing-based method for efficient PK–FK detection and schema extraction
- Achieves **higher accuracy** in multi-table reasoning while drastically reducing **time and memory costs**
- Provides **LLM-interpretable prompts** that improve table-column retrieval and text-to-SQL performance across diverse models

paper

github