Symphony: Towards Natural Language Query Answering over Multi-modal Data Lakes (Vision)

Zui Chen  
Tsinghua

Zihui Gu  
Renmin

Lei Cao  
University of Arizona

Ju Fan  
Renmin

Sam Madden  
MIT

Nan Tang  
QCRI

Jan 10, 2023 @ CIDR,  
AMSTERDAM
Multi-modal Data Lakes

Text files, CSV files, databases, log files, knowledge graphs …
One Driving Application (PSA)

FIFA 2022
population
education
economic activities
disabilities
buildings
...
...
One Driving Application (PSA)

FIFA 2022
population
education
economic activities
disabilities
buildings

...
One Driving Application (PSA)

FIFA 2022 population education economic activities disabilities buildings ...
...

Qatar Planning and Statistics Authority

State of Qatar
Data Management Perspective

Structured Queries (SQL)

Data Preparation

Extraction
Transformation
Loading
Cleaning
Schema matching
Deduplication
...

High human-cost
Data Management Perspective

Structured Queries (SQL)

Data lake (raw format)

Data Preparation

Extraction
Transformation
Loading
Cleaning
Schema matching
Deduplication
...

High human-cost ➞ No Data Preparation
Data Management Perspective

Structured Queries (SQL) $\rightarrow$ Natural Language Queries

Data Preparation
- Extraction
- Transformation
- Loading
- Cleaning
- Schema matching
- Deduplication

High human-cost $\rightarrow$ No Data Preparation

Data lake (raw format)
Data Management Perspective

Structured Queries (SQL) → Natural Language Queries

Data Preparation
- Extraction
- Transformation
- Loading
- Cleaning
- Schema matching
- Deduplication
...

High human-cost → No Data Preparation

Symphony
- Reasoning
- Retrieval

Data lake (raw format)
Foundation Models?

Foundation Models (GPT3, ChatGPT)

Implicit knowledge

Symphony
- Reasoning
- Retrieval

Explicit knowledge
Foundation Models?

Foundation Models (GPT3, ChatGPT)

Implicit knowledge

(1) Poor (DB) reasoning

Symphony

Reasoning

Retrieval

Explicit knowledge
Foundation Models?

Foundation Models (GPT3, ChatGPT)

Implicit knowledge

(1) Poor (DB) reasoning
(2) Not up-to-date

Symphony

Reasoning

Retrieval

Explicit knowledge
Foundation Models?

**Foundation Models** (GPT3, ChatGPT)

*Implicit knowledge*

---

**Question:** Who gave a talk about “Dutch tomatoes in CIDR 2023?”

**File**

*The Netherlands has basically reinvented the tomato*

Gustavo Alonso keynote at CIDR’23

---

**Symphony**

- **Reasoning**
  - **Retrieval**
  - **Explicit knowledge**

---

(1) Poor (DB) reasoning
(2) Not up-to-date
A Running Example

(1) Index
A Running Example

Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

Faraar (transl. Absconding) is a 1975 Bollywood crime film drama. The film is produced by Alankar Chitra and directed by Shanker Mukherjee. The film stars Amitabh Bachchan, Sharmila Tagore, Sanjeev Kumar, Sulochna Sajjan, Agha and Bhagwan Dada...

Source: https://en.wikipedia.org/wiki/Faraar

<table>
<thead>
<tr>
<th>Year</th>
<th>Song</th>
<th>Film</th>
</tr>
</thead>
</table>
| 1971 | Zindagi Ek Safer | Andaz | ...
| 1971 | Yeh Jo Mohabbat | Kati Patang | ...
| 1975 | Main Phiraa tum | Faraar | ...

Source: https://en.wikipedia.org/wiki/Kishore_Kumar
A Running Example

(1) Index

Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

(2) Data Retrieval

Faraar (transl. Absconding) is a 1975 Bollywood crime film drama. The film is produced by Alankar Chitra and directed by Shanker Mukherjee. The film stars Amitabh Bachchan, Sharmila Tagore, Sanjeev Kumar, Sulochana, Sajjan, Agha and Bhagwan Dada...

Source: https://en.wikipedia.org/wiki/Faraar

Prompt 1: The passage P1 has the following content: ...

The table T1 has the following columns: Year, Song, Film, Music Director, Lyricist.

Based on P1 and T1, the question is “Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?”.

What sub-questions can it be broken down into?

- P1: What is the name of the film produced by Alankar Chitra and directed by Shanker Mukherjee. It can be answered by P1.

- T1: The first sub-question is “What is the name of the film produced by Alankar Chitra and directed by Shanker Mukherjee?”, it can be answered by P1.

- T1: The second sub-question is “What is the name of the song in the film?”, it can be answered by T1.

(3) Query Decomposition

Source: https://en.wikipedia.org/wiki/Kishore_Kumar
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

Faraar (transl. Absconding) is a 1975 Bollywood crime film drama. The film is produced by Alankar Chitra and directed by Shanker Mukherjee. The film stars Amitabh Bachchan, Sharmila Tagore, Sanjeev Kumar, Sulochna, Sajjan, Agha and Bhagwan Dada...

Source: https://en.wikipedia.org/wiki/Faraar

Prompt 1: The passage P1 has the following content: ...

The table T1 has the following columns: Year, Song, Film, Music Director, Lyricist.

Based on P1 and T1, the question is “Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?”. What sub-questions can it be broken down into?

GPT-3: q1 What is the name of the film produced by Alankar Chitra and directed by Shanker Mukherjee. It can be answered by P1.

Prompt 2: The first sub-question is “What is the name of the film produced by Alankar Chitra and directed by Shanker Mukherjee?”, it can be answered by P1.

GPT-3: q2 the second sub-question is “What is the name of the song in the film?”, it can be answered by T1.

with NLP Question Answering with NL2SQL
(1) Index (X-to-Vec)

Data Retrieval

Query Decomposition

Sub-query execution

Encoder

Index: FAISS

Meta

Vectors

X
(text, tables, databases, …)
(1) Index (X-to-Vec)

(2) Data Retrieval

(3) Query Decomposition

(4) Sub-query execution

$X$ (text, tables, databases, …)

Encoder → 

Vectors

Index: FAISS

Meta

AutoEncoders

Fixed-size vector $\text{vec}(X)$

Encoder

Decoder

serialize($X$) → 

Compress

Reconstruct

$X$
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

Prompt 1: The passage P1 has the following content: …
The table T1 has the following columns: Year, Song, Film, Music Director, Lyricist. Based on P1 and T1, the question is “Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?”. What sub-questions can it be broken down into?

GPT-3: q1 What is the name of the film produced by Alankar Chitra and directed by Shanker Mukherjee. It can be answered by P1.

Prompt 2: The first sub-question is “What is the name of the film produced by Alankar Chitra and directed by Shanker Mukherjee?”, it can be answered by P1.

GPT-3: q2 the second sub-question is “What is the name of the song in the film?”, it can be answered by T1.
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

```
Initial Prompt
prompt_1 = Serialize(d'_1); Serialize(d'_2); \ldots \text{Serialize}(d'_m)
Based on d'_1, d'_2, \ldots and d'_m, the question is Q, what sub-questions can it be broken down into?
```

```
Prompt Generation
(qi, di)
prompt_1
```

```
Next Prompt
prompt_i = The \text{[N]} sub-query is \text{[Q]}, it can be answered by \text{[D]}
```

GPT-3
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

**Q** Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

**T1**

**P1**

**Initial Prompt**

```
prompt_1 = Serialize(d'_1); Serialize(d'_2); ... Serialize(d'_m)
```

Based on \(d'_1, d'_2, \ldots\) and \(d'_m\), the question is \(Q\), what sub-questions can it be broken down into?

**Prompt Generation**

**Next Prompt**

```
prompt_i = The \([N]\) sub-query is \([Q]\), it can be answered by \([D]\)
```

**Stop**

**GPT-3**
Natural Language Query over Text

NLP
Question Answering

Natural Language Query over Table(s)

NL2SQL
(SIGMOD’23)
Natural Language Query over Text

- NLP
- Question Answering

Natural Language Query over Table(s)

- NL2SQL (SIGMOD’23)
- TableQA (EMNLP’22)

NLDBs
"Database reasoning over multiple text files"
Preliminary Result

• Data lake
  • 400K web tables
  • 6M passages (text files)
• Queries: 18, manually designed
  • Goal: ensuring each query can be answered by the data lake
• Test:
  • Is the decomposition correct?
  • Is the sub-query evaluation correct?
Sample Successful Case (12/18)

- **Question**: Which is taller, the tallest building in the UK or the tallest building in South Korea?
- **Retrieved datasets**:

<table>
<thead>
<tr>
<th>Rank</th>
<th>Official Name</th>
<th>Height (m)</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The Shard</td>
<td>310</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>22 Bishopsgate</td>
<td>278</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>One Canada Square</td>
<td>235</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Height (m)</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lotte World Tower</td>
<td>550</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Landmark Tower</td>
<td>412</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Tower A</td>
<td>339</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

**Decomposed queries**:
1. “What is the height of the tallest building in the UK?”, it can be answered by T1;
2. “What is the height of the tallest building in the South Korea?”, it can be answered by T2.

**GPT-3 for aggregation**: Question, result of subquery-1, result of subquery-2
Sample Failure Case (6/18)

- **Question**: What year was the first German film that won the Academy Award for Best Foreign language Film released?
- **Retrieved datasets**:

  **Passage P1**

  *The Tin Drum* (German: *Die Blechtrommel*) is a 1979 film adaptation of Günter Grass' novel of the same title, directed by Volker Schlöndorff from a screenplay co-written with Jean-Claude Carrière and Franz Seitz. …

  Source: https://en.wikipedia.org/wiki/The_Tin_Drum_(film)

  **Table T1**

<table>
<thead>
<tr>
<th>ceremony</th>
<th>film title used in nomination</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>51st</td>
<td>the glass cell</td>
<td>nominee</td>
</tr>
<tr>
<td>52nd</td>
<td><em>the tin drum</em></td>
<td>won academy award</td>
</tr>
<tr>
<td>53rd</td>
<td>fabian</td>
<td>not nominated</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

  Source: https://en.wikipedia.org/wiki/List_of_German_submissions_for_the_Academy_Award_for_Best_Foreign_Language_Film

- **Decomposed queries**:
  1. “What year was the film The Tin Drum released?”, it can be answered by T1;
  2. “What was the first German film that won the Academy Award for Best Foreign language Film?”, it can be answered by T1.
Sample Failure Case (6/18)

- **Question**: What year was the first German film that won the Academy Award for Best Foreign language Film released?
- **Retrieved datasets**: 

  - **Passage P1**
    
    *The Tin Drum* (German: *Die Blechtrommel*) is a 1979 film adaptation of Günter Grass' novel of the same title, directed by Volker Schlöndorff from a screenplay co-written with Jean-Claude Carrière and Franz Seitz. ...

    [Source: https://en.wikipedia.org/wiki/The_Tin_Drum_(film)]

  - **Table T1**
    
    | ceremony | film title used in nomination | result             |
    |----------|-------------------------------|--------------------|
    | 51st     | the glass cell                | nominee            |
    | 52nd     | **the tin drum**              | won academy award  |
    | 53rd     | fabian                        | not nominated      |
    | ...      | ...                           | ...                |

    [Source: https://en.wikipedia.org/wiki/List_of_German_submissions_for_the_Academy_Award_for_Best_Foreign_Language_Film]

- **Decomposed queries**:
  1. “What year was the film *The Tin Drum* released?”, it can be answered by T1; P1
  2. “What was the first German film that won the Academy Award for Best Foreign language Film?”, it can be answered by T1.
Conclusion

• **New way** of exploring multi-modal data lakes

• **AI-Assistant**: Political/business leaders surrounded by advisors

Future Work

• Improving the index (X-to-Vec)

• Contrastive learning

• Combining with traditional string similarity search

• Sub-query execution module

• NL2SQL