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

Zui Chen
Tsinghua

Zihui Gu
Renmin

Lei Cao
University of Arizona
MIT

Ju Fan
Renmin

Sam Madden
MIT

Nan Tang
QCRI

I ❤

Jan 10, 2023 @ CIDR, AMSTERDAM

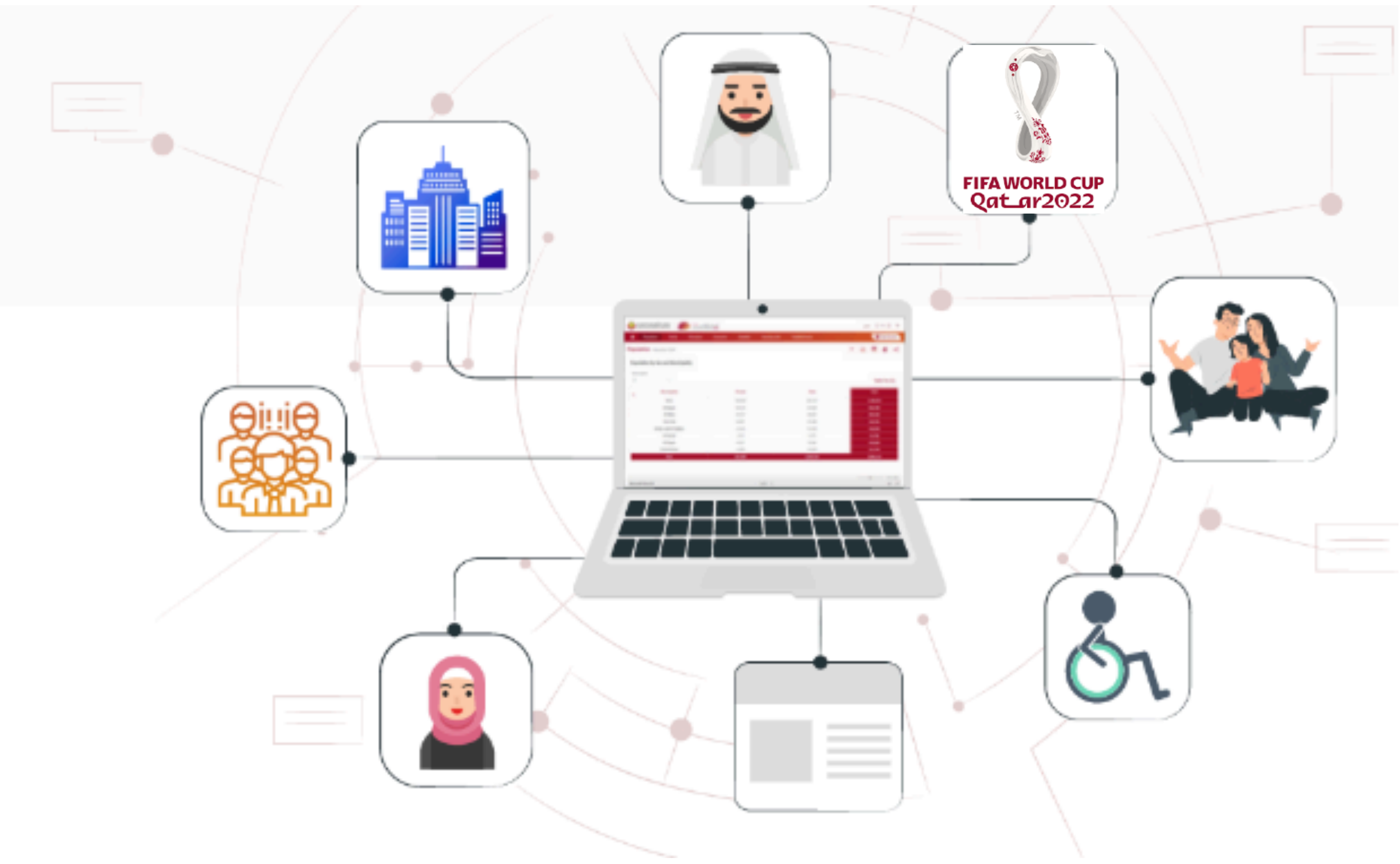
Multi-modal Data Lakes

Text files, CSV files, databases, log files, knowledge graphs …

One Driving Application (PSA)

Qatar

جهـــاز التخطيـــط والإحصـــاء
Planning and Statistics Authority
State of Qatar • دولـــة قطـــر



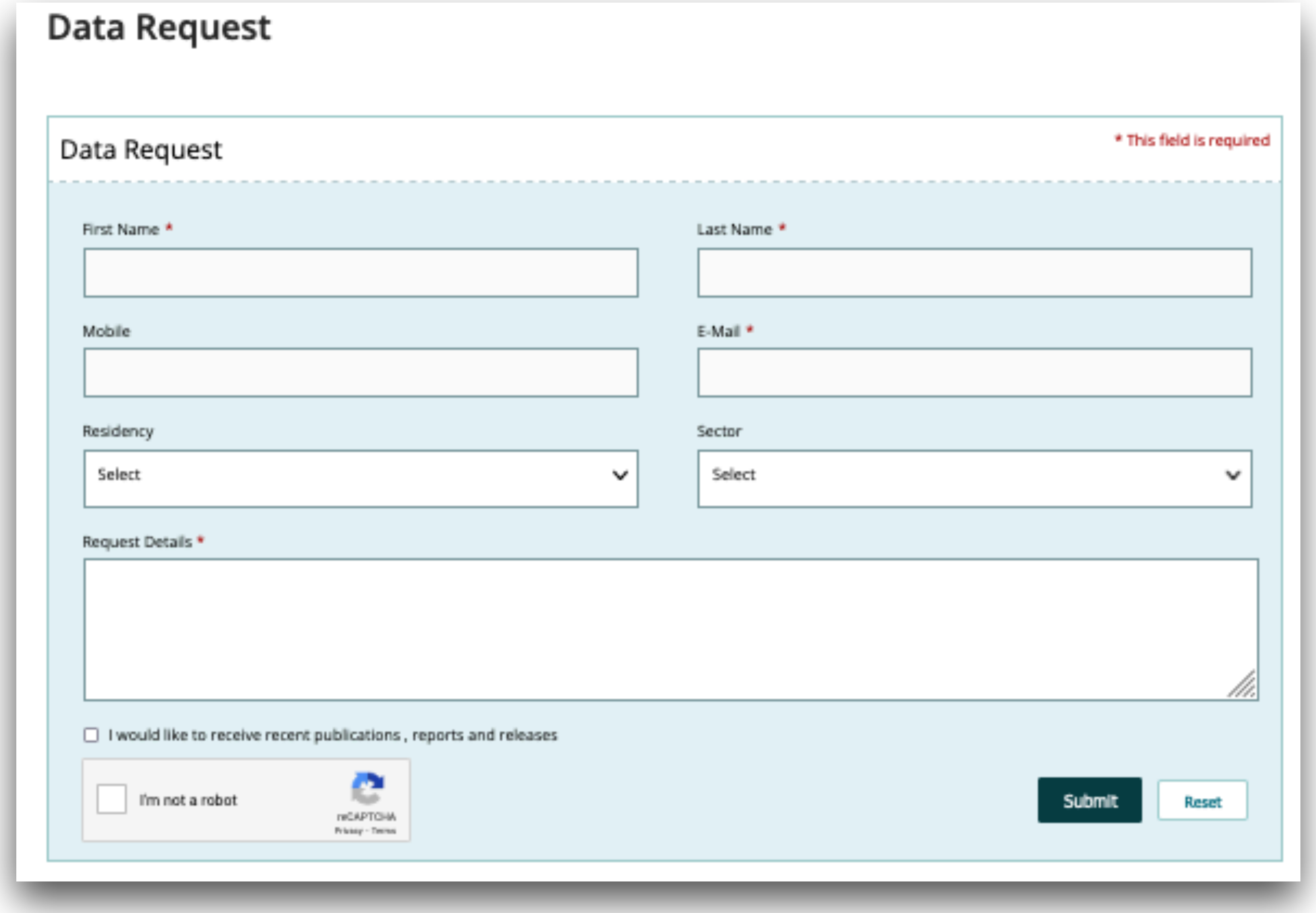
FIFA 2022
population
education
economic activities
disabilities
buildings
…
…

One Driving Application (PSA)

Qatar

جهـــاز التخطيـــط والإحصـــاء
Planning and Statistics Authority
State of Qatar • دولـــة قطـــر

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



Data Request

Data Request

* This field is required

First Name *

Last Name *

Mobile

E-Mail *

Residency: Select

Sector: Select

Request Details *

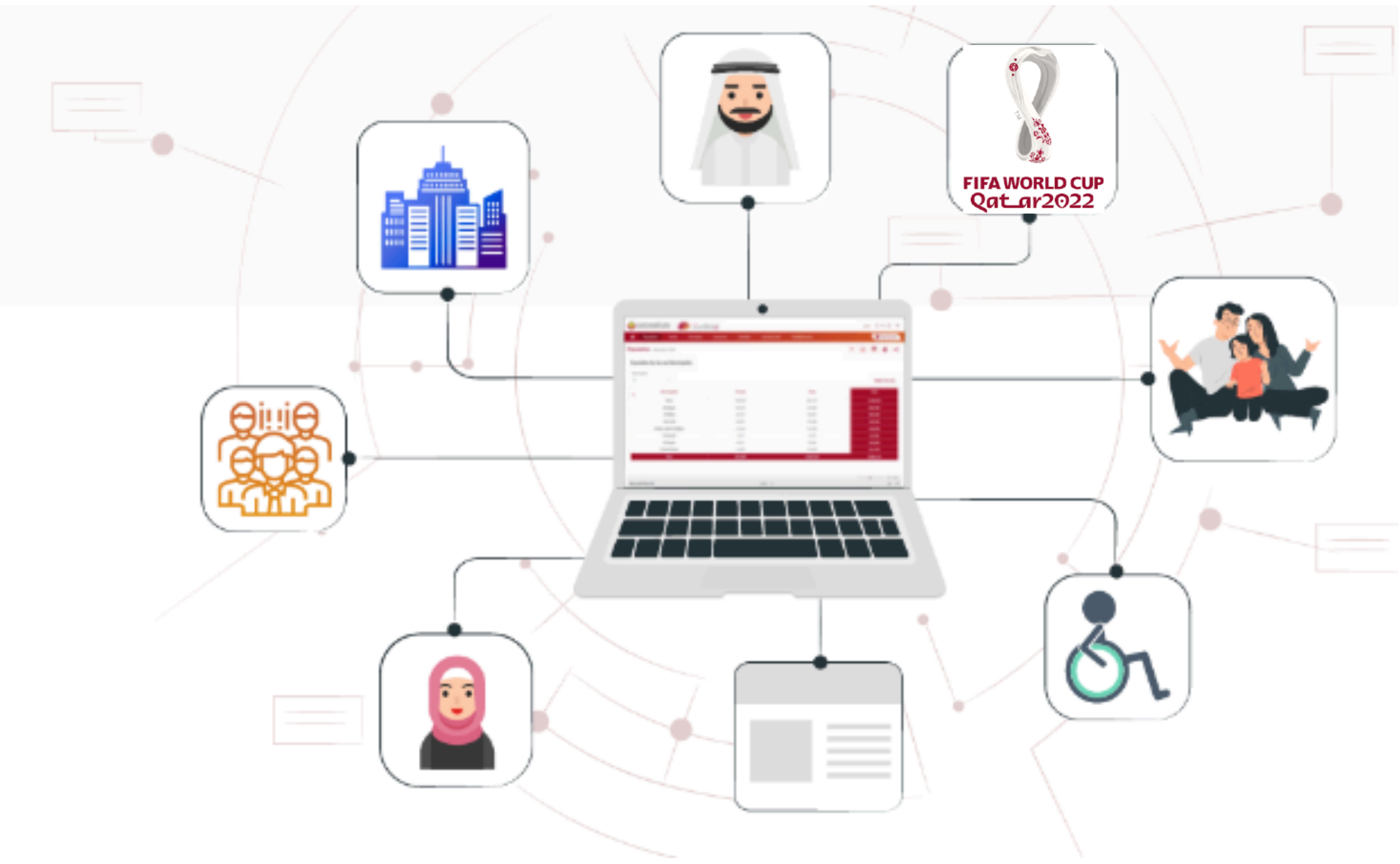
☐ I would like to receive recent publications , reports and releases

☐ I'm not a robot

One Driving Application (PSA)

Qatar

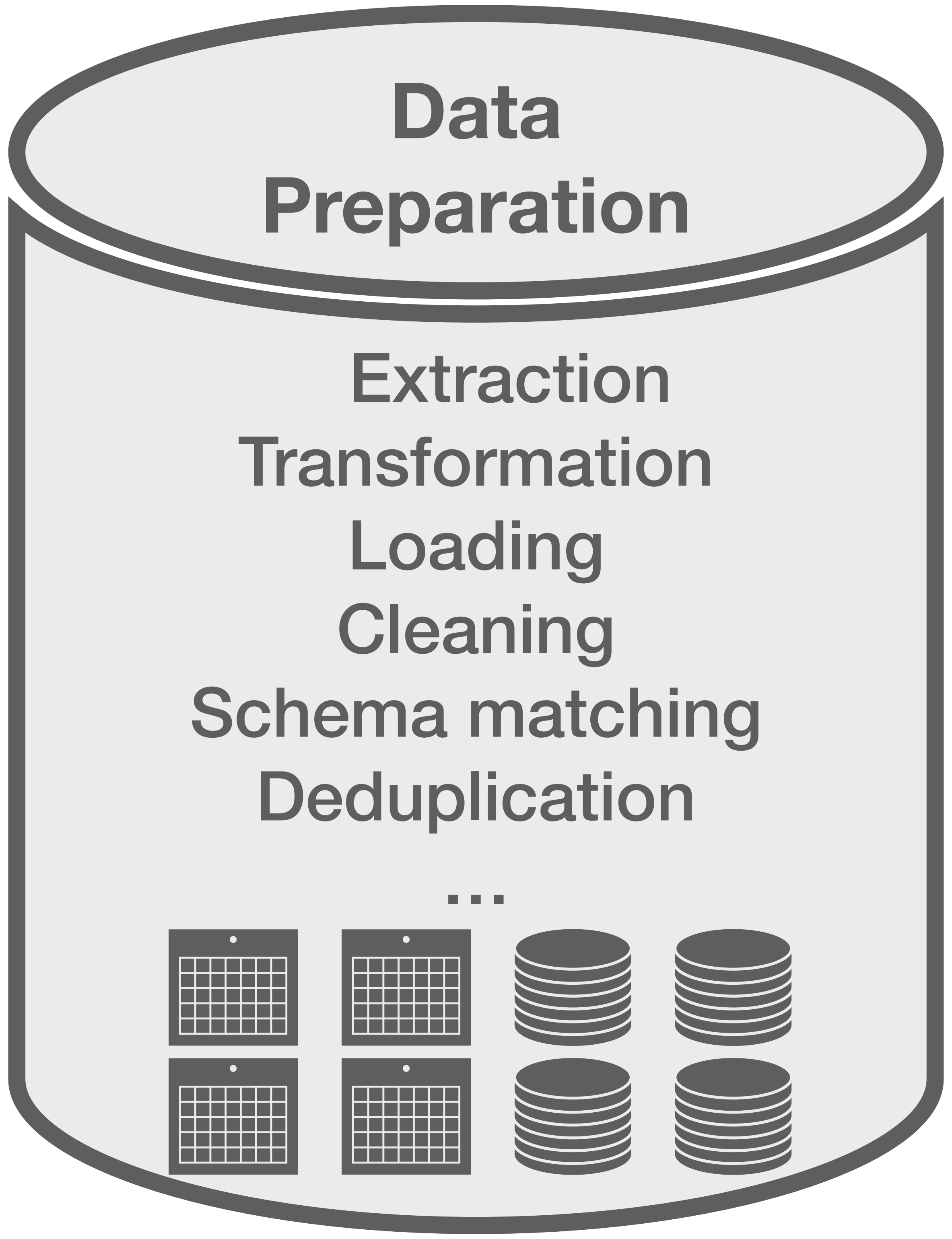
جهاز التخطيط والإحصاء
Planning and Statistics Authority
State of Qatar • دولة قطر



FIFA 2022
population
education
economic activities
disabilities
buildings
…
…

Data Management Perspective

Structured Queries (SQL)



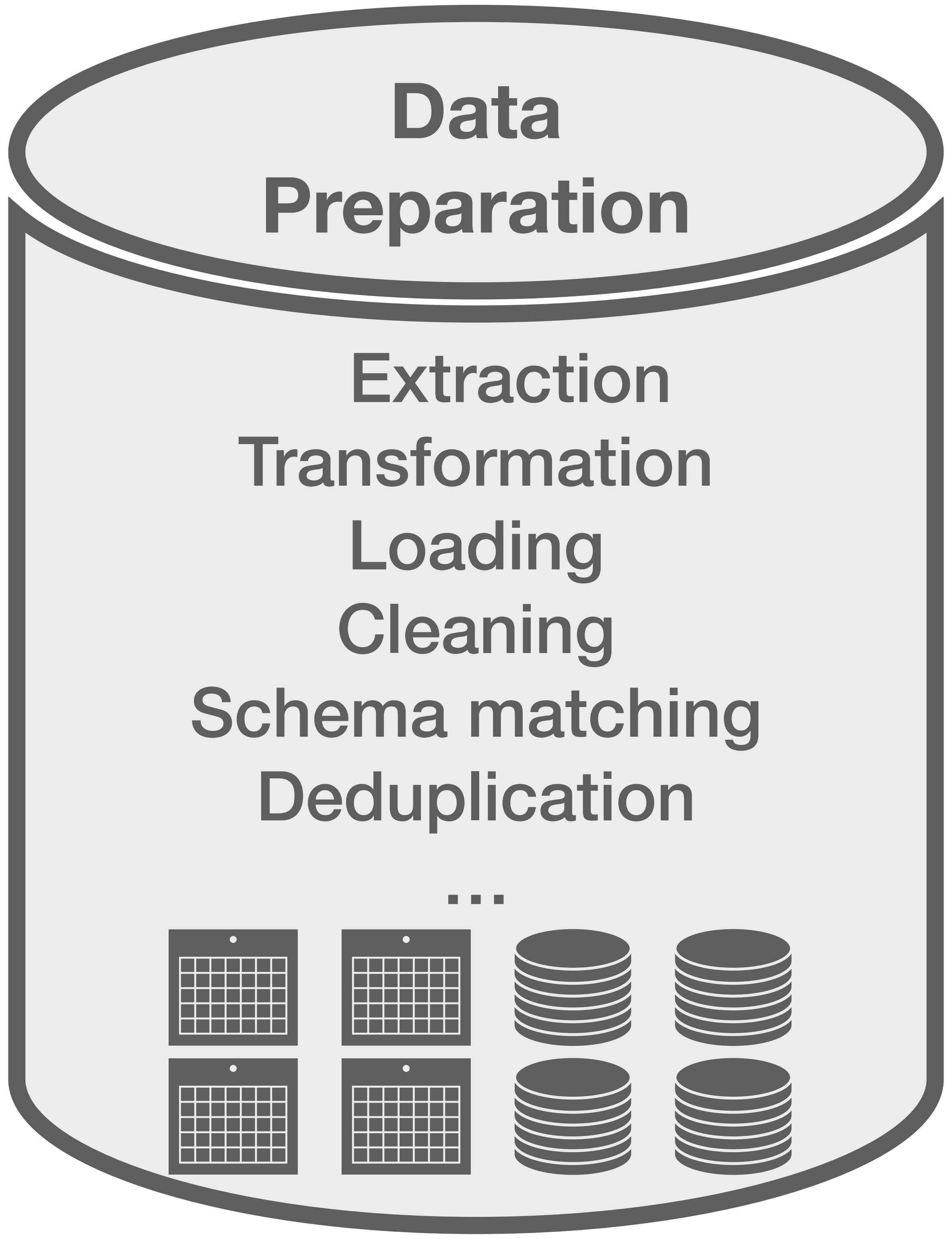
Data Preparation

Extraction
Transformation
Loading
Cleaning
Schema matching
Deduplication
...

High human-cost

Data Management Perspective

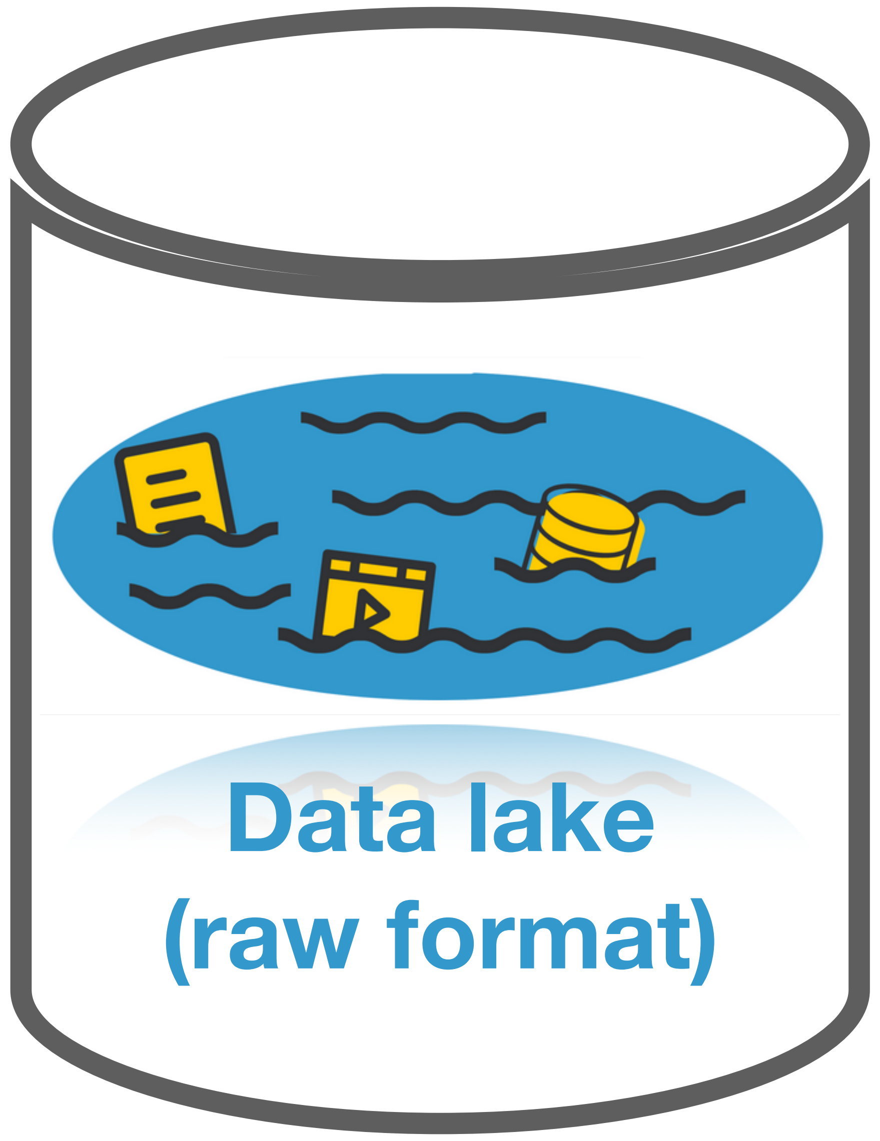
Structured Queries (SQL)



Data Preparation

Extraction
Transformation
Loading
Cleaning
Schema matching
Deduplication
…

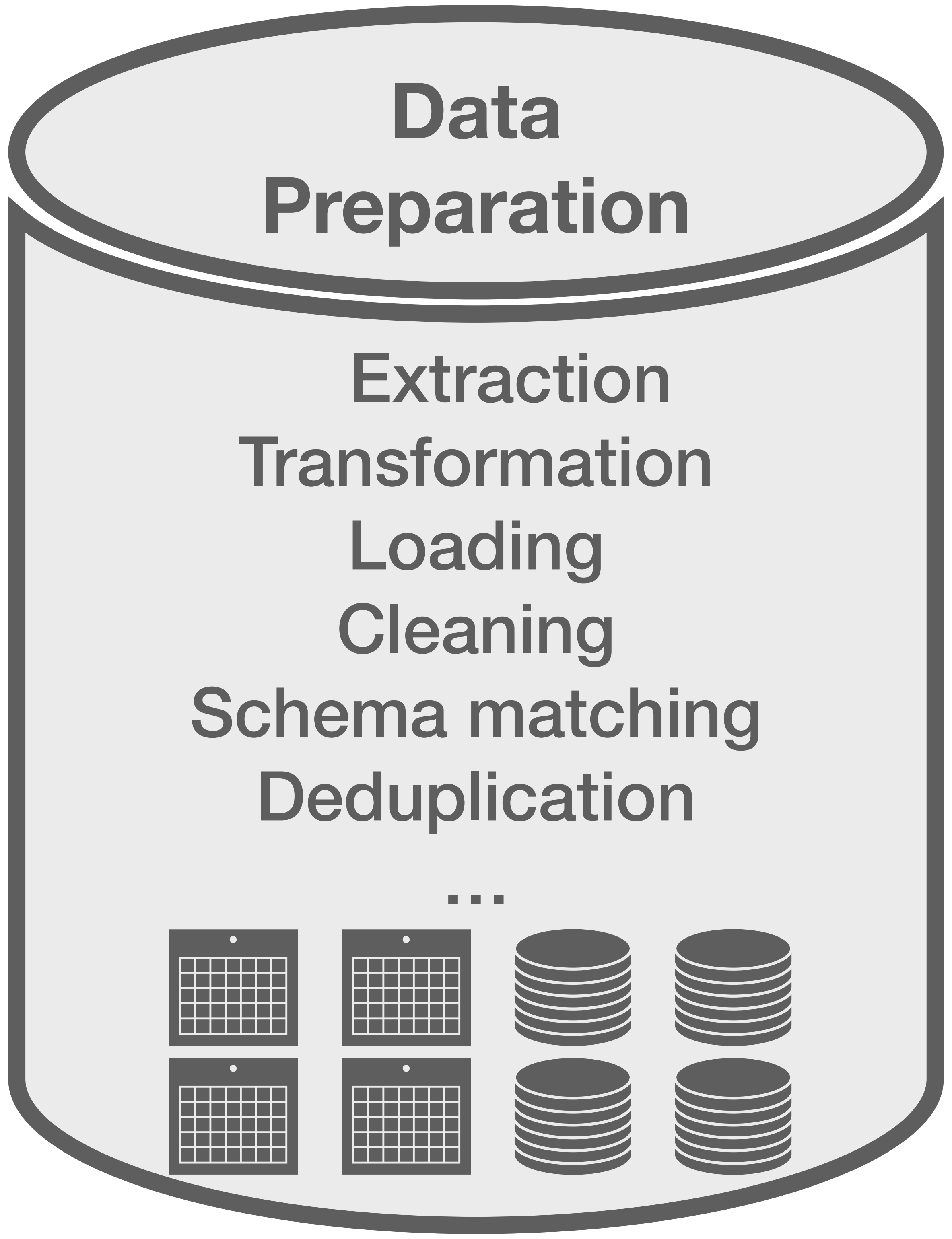
High human-cost → No Data Preparation



Data lake
(raw format)

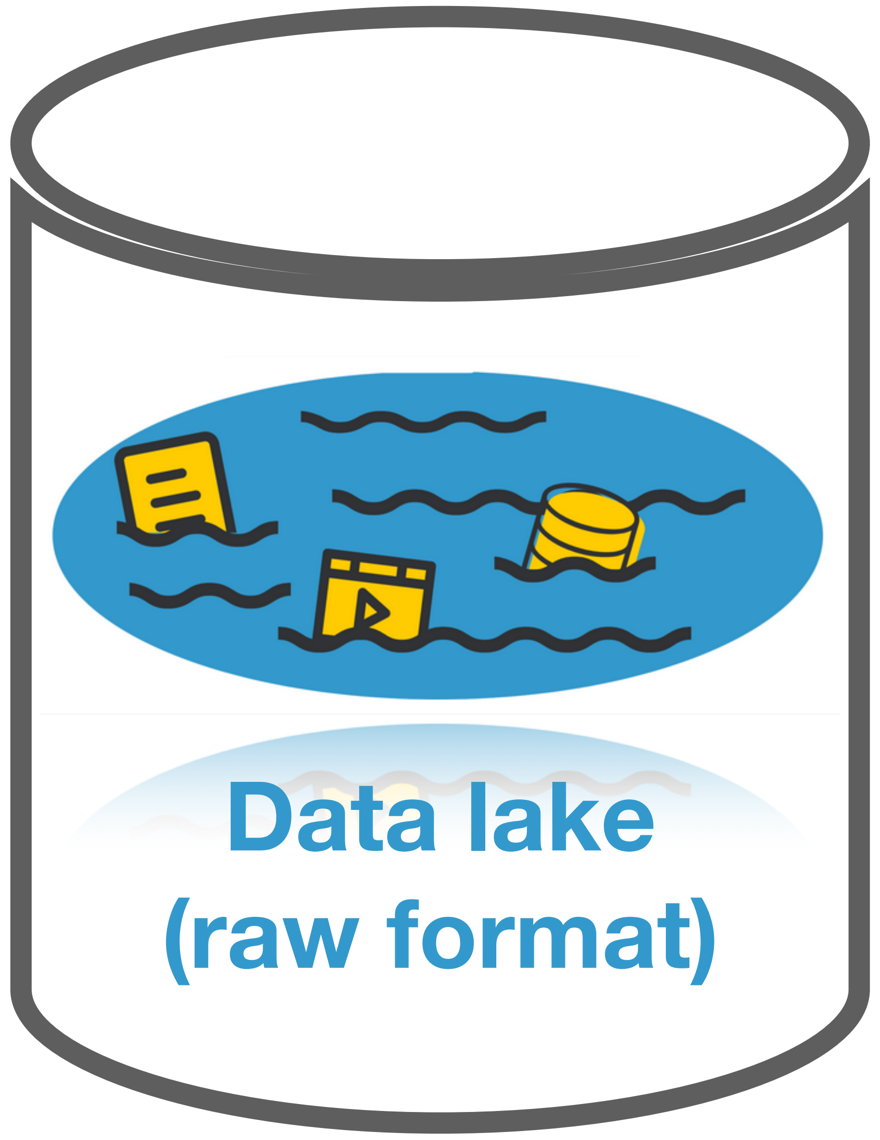
Data Management Perspective

Structured Queries (SQL) ⟶ Natural Language Queries



Data Preparation

Extraction
Transformation
Loading
Cleaning
Schema matching
Deduplication
…

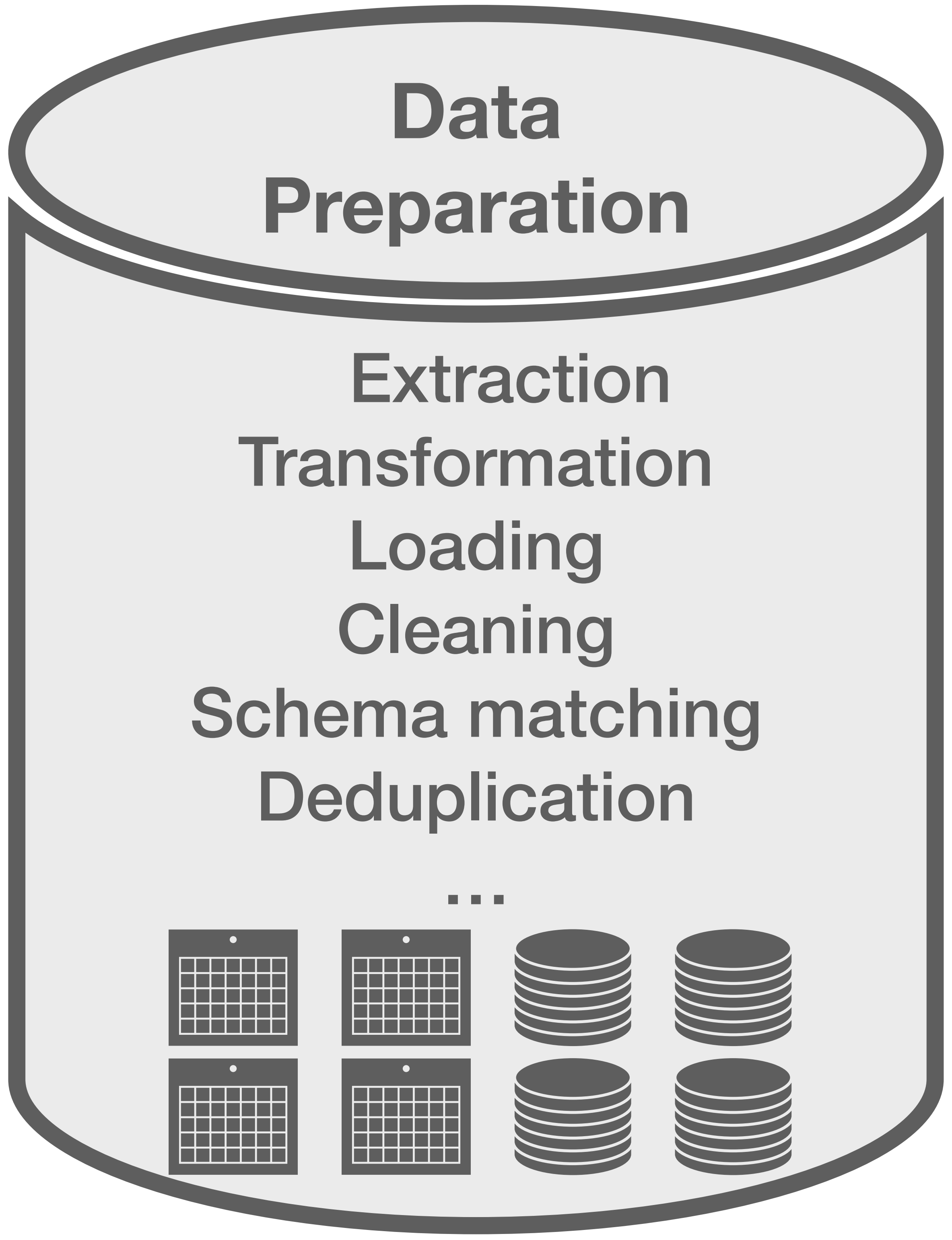


Data lake
(raw format)

High human-cost ⟶ No Data Preparation

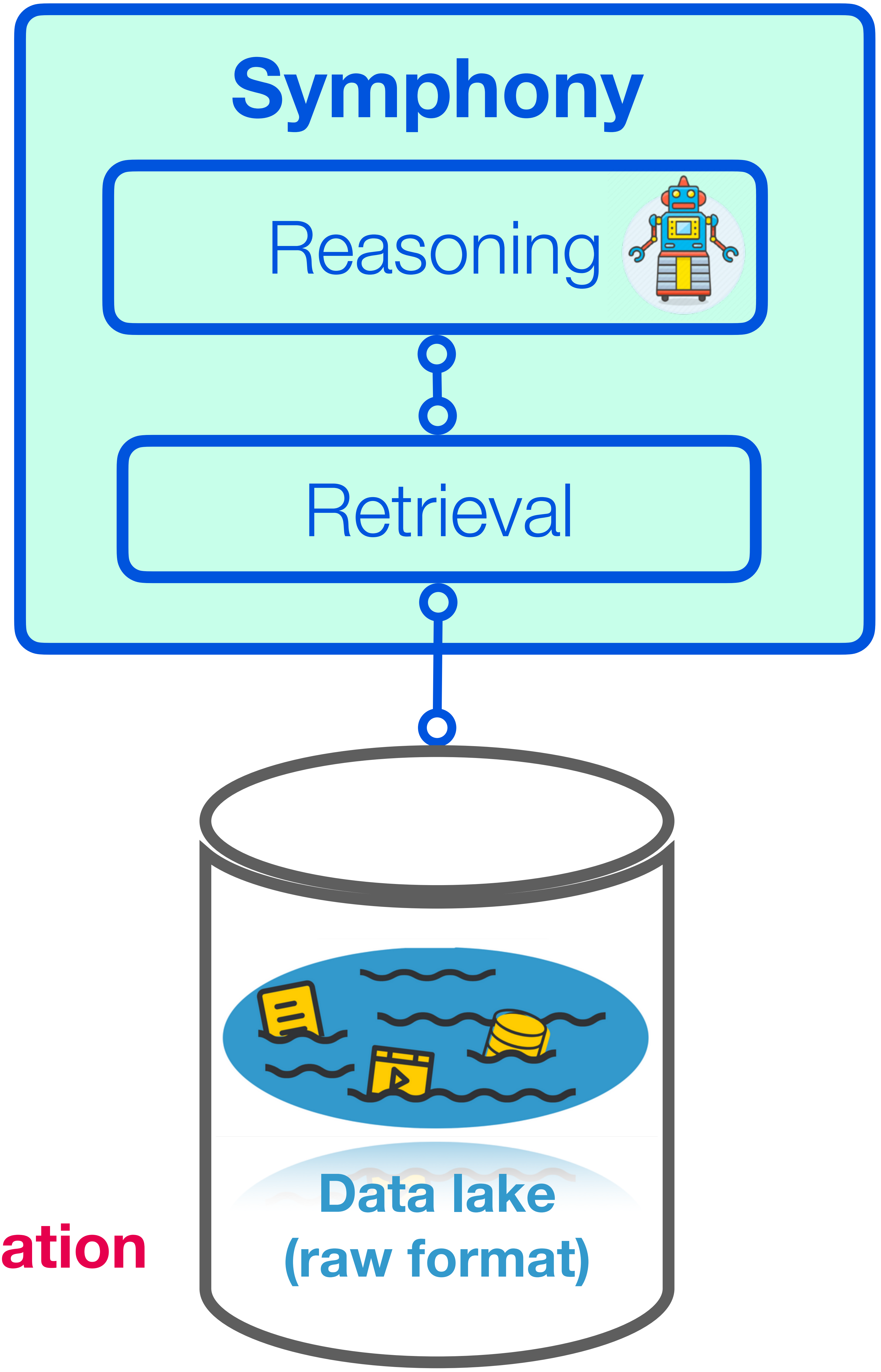
Data Management Perspective

Structured Queries (SQL) ⟶ Natural Language Queries



Data Preparation

Extraction
Transformation
Loading
Cleaning
Schema matching
Deduplication
...



Symphony

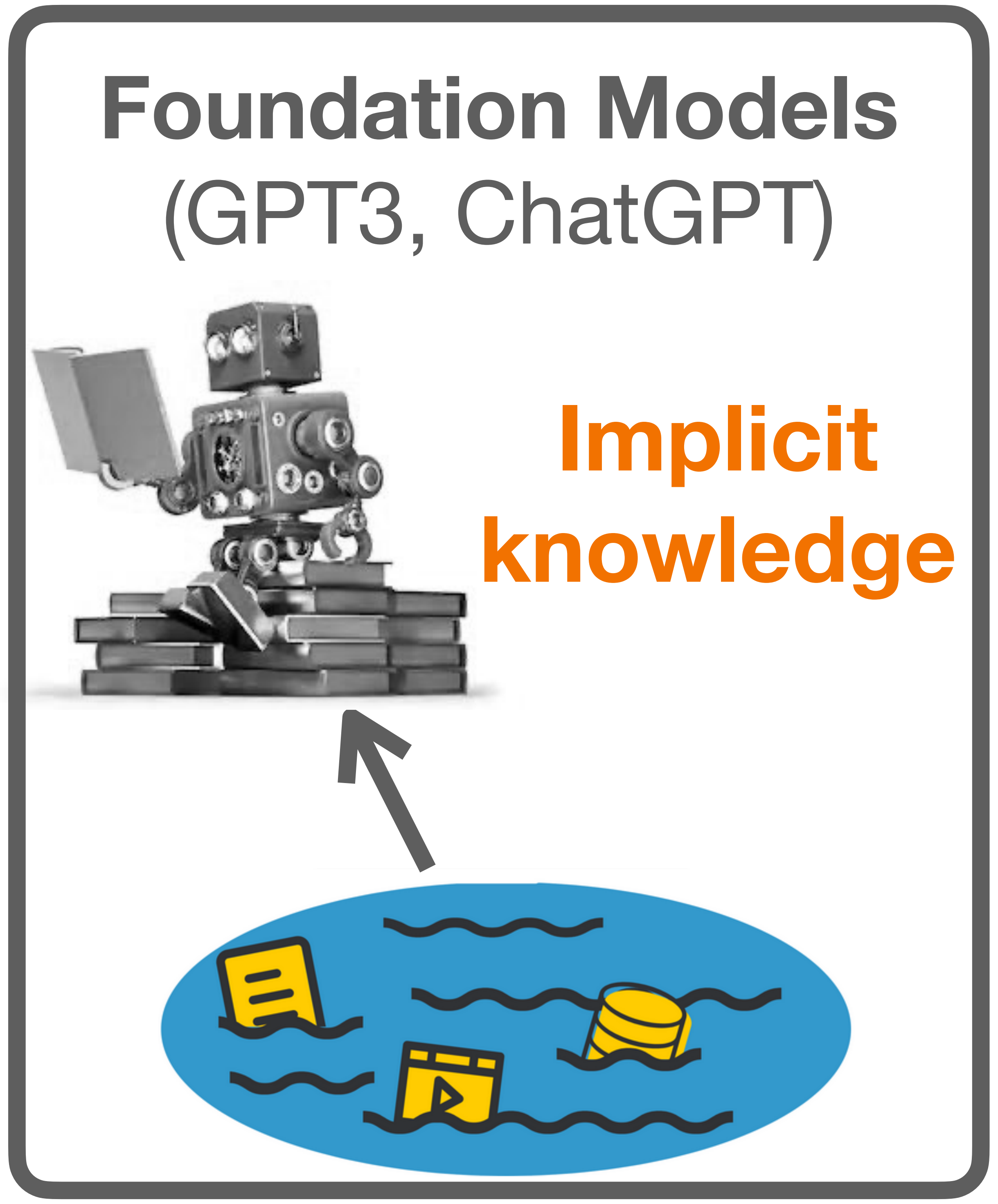
Reasoning

Retrieval

Data lake
(raw format)

High human-cost ⟶ No Data Preparation

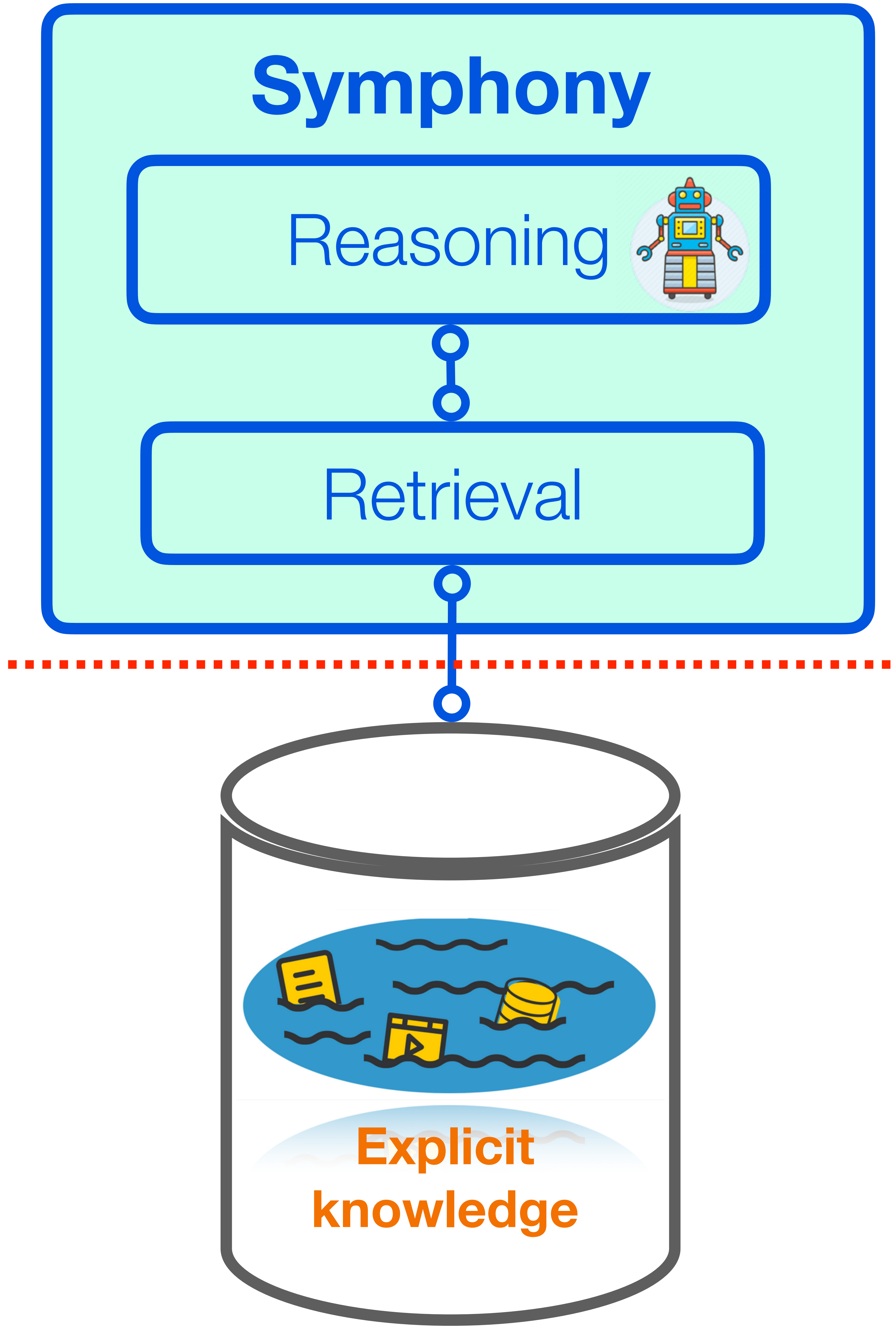
Foundation Models?



Foundation Models

(GPT3, ChatGPT)

Implicit knowledge



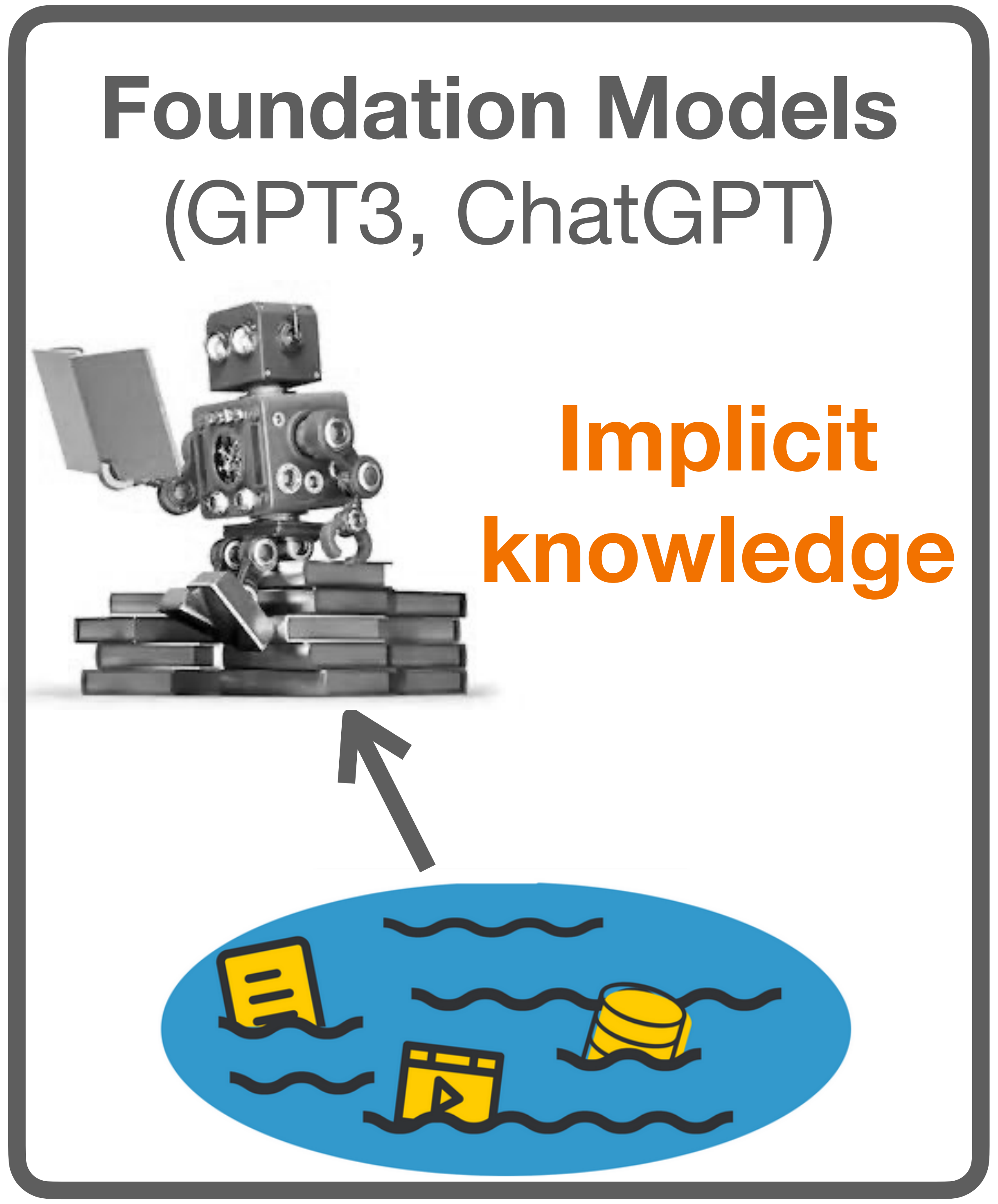
Symphony

Reasoning

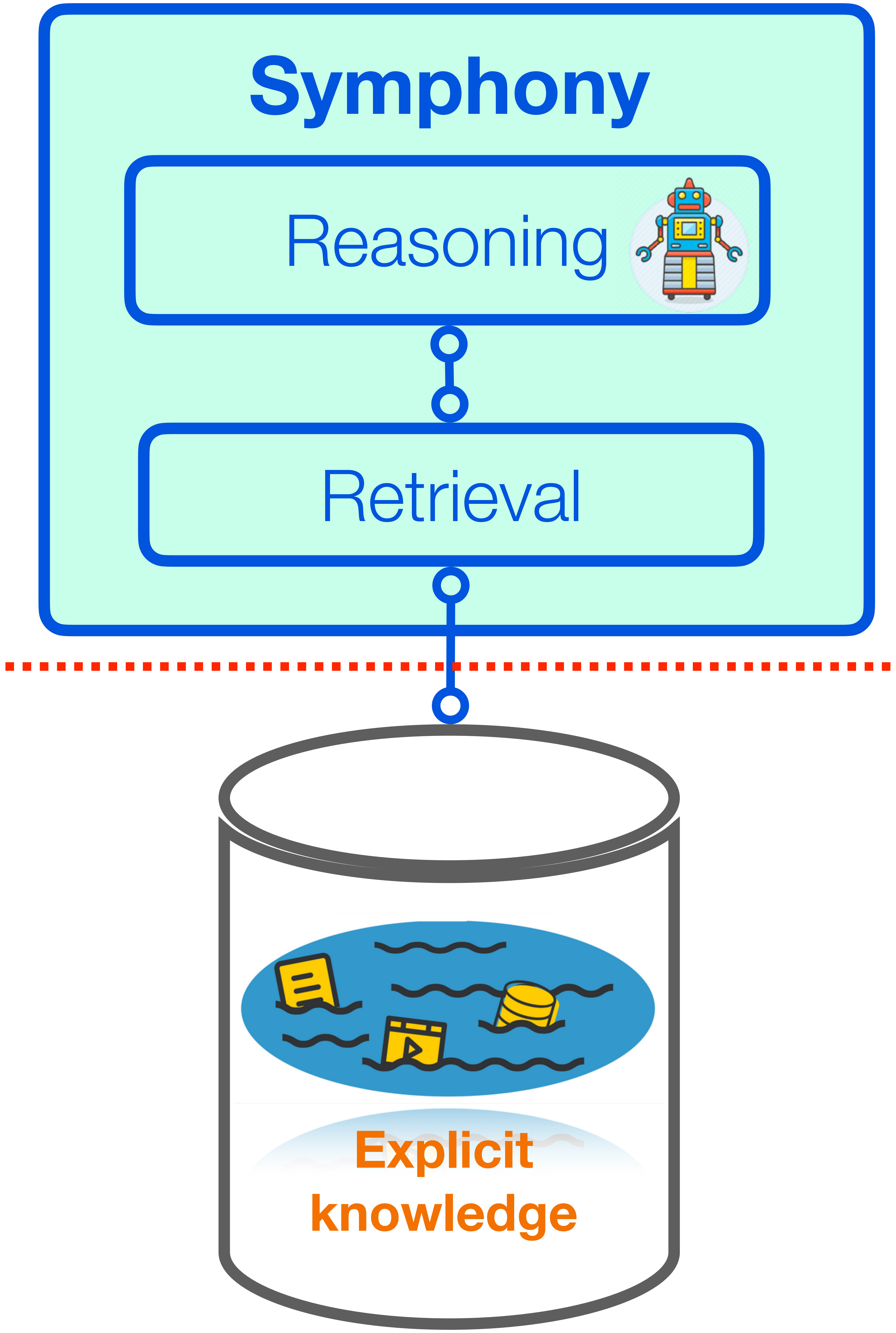
Retrieval

Explicit knowledge

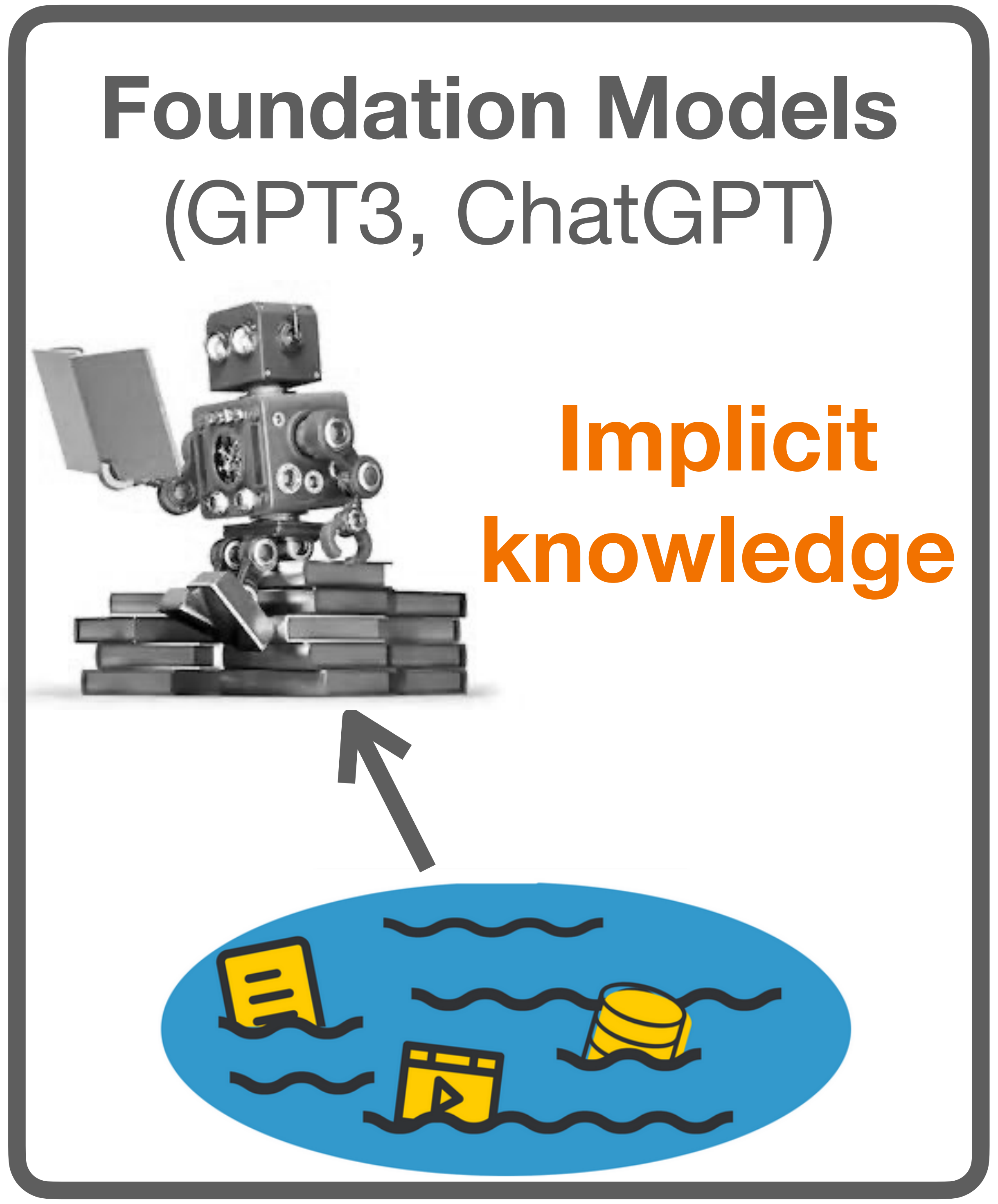
Foundation Models?



(1) Poor (DB) reasoning



Foundation Models?

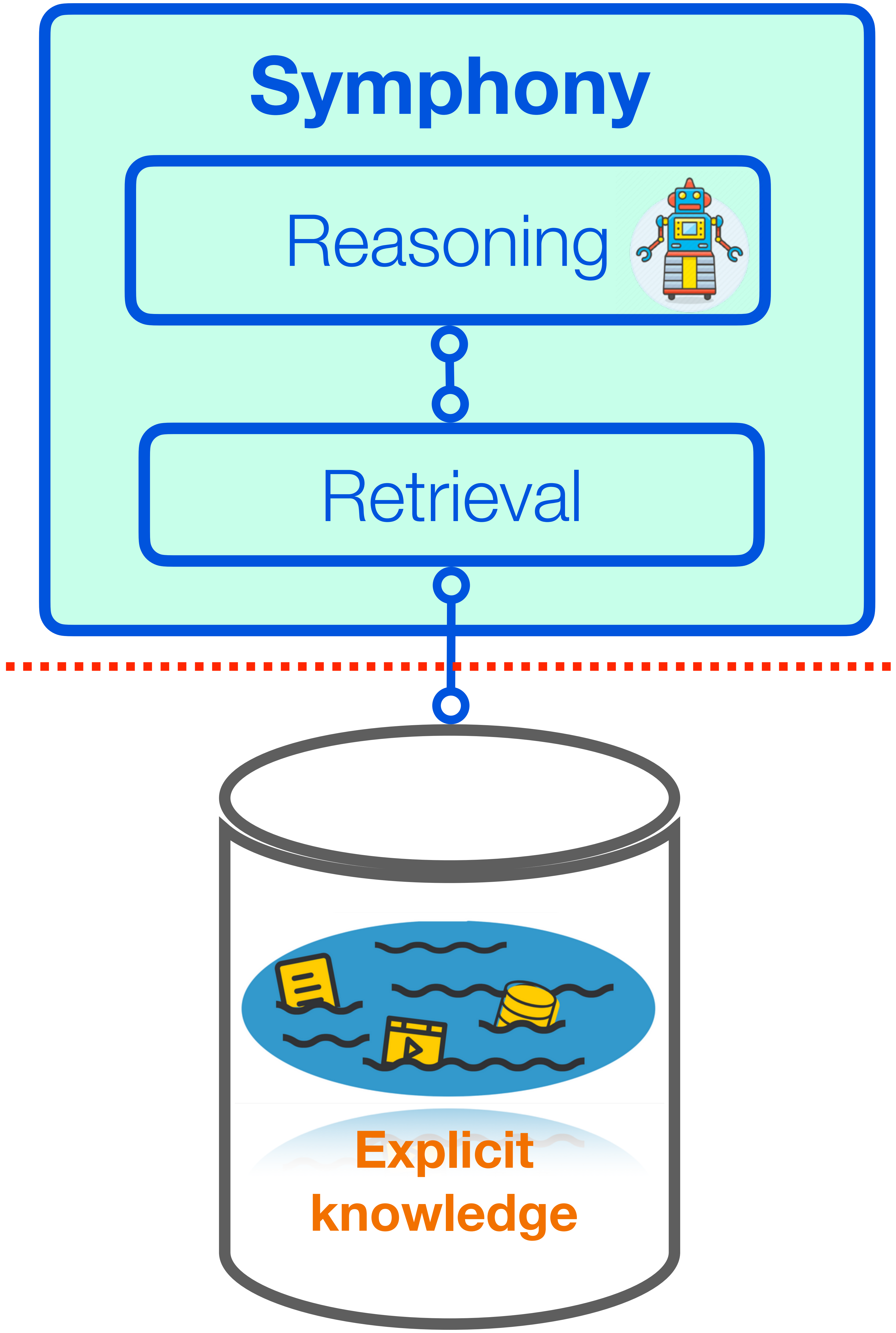


Foundation Models
(GPT3, ChatGPT)

Implicit knowledge

(1) Poor (DB) reasoning

(2) Not up-to-date



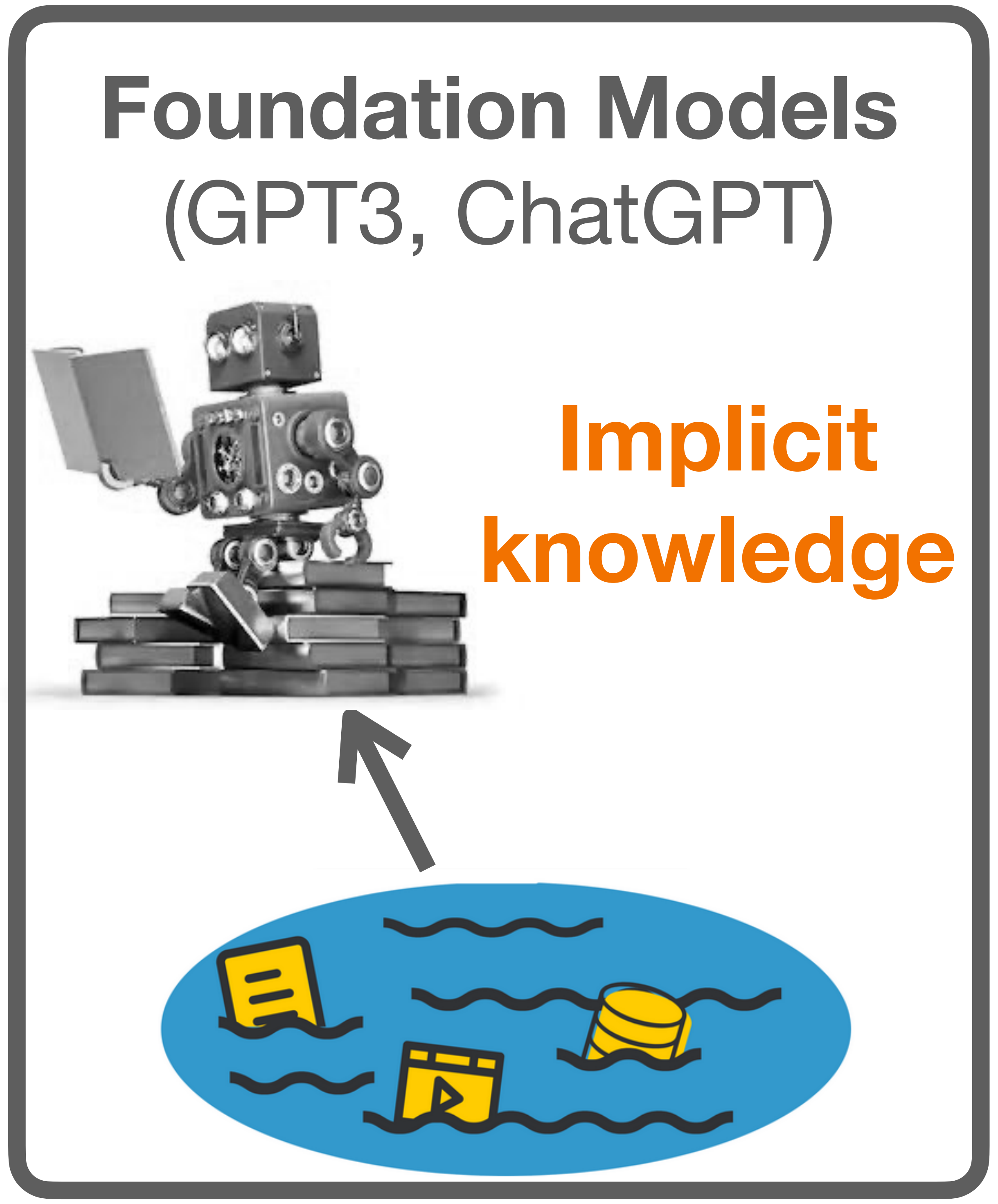
Symphony

Reasoning

Retrieval

Explicit knowledge

Foundation Models?



Foundation Models (GPT3, ChatGPT)

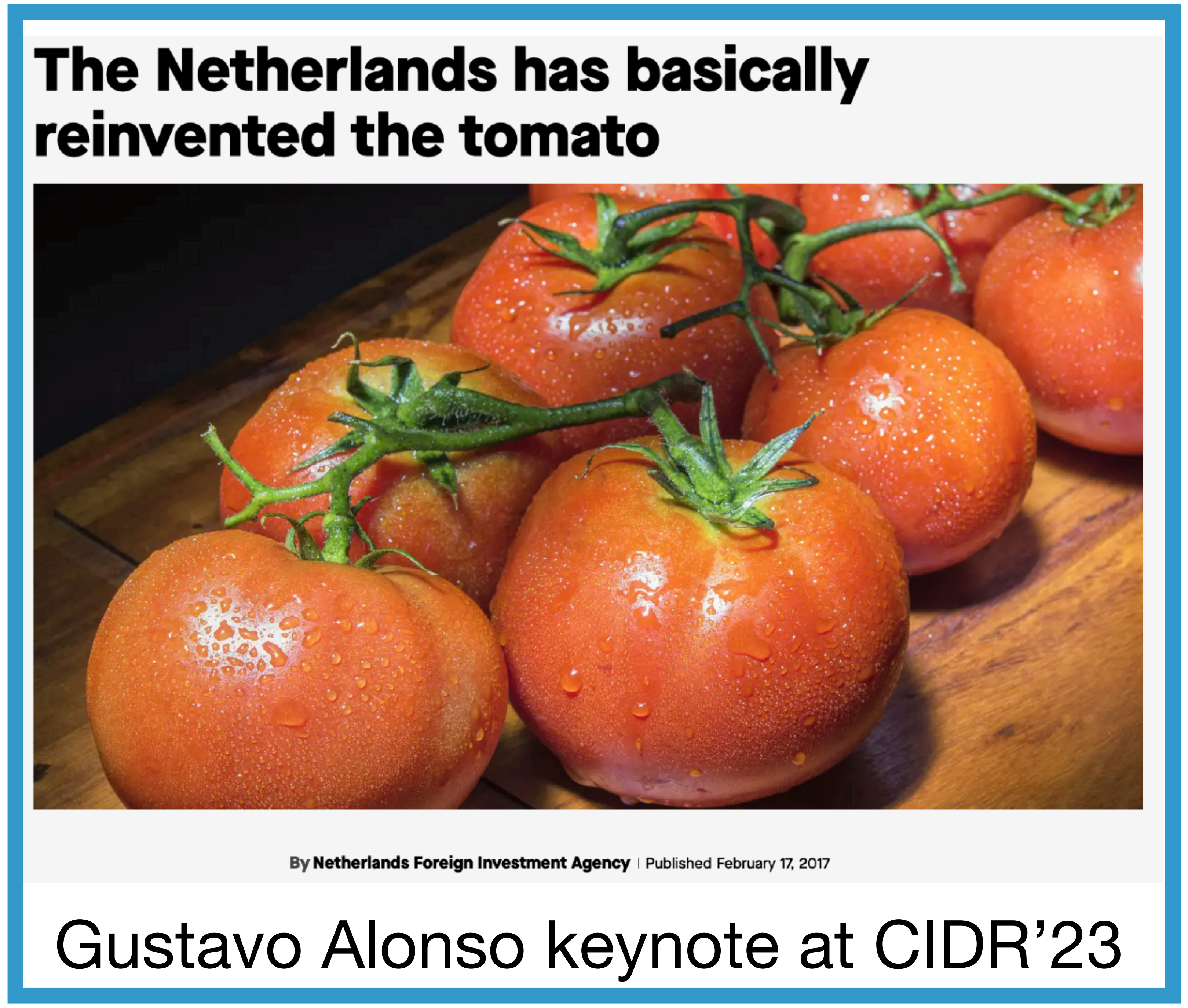
Implicit knowledge

(1) Poor (DB) reasoning

(2) Not up-to-date

Question: Who gave a talk about "Dutch tomatoes in CIDR 2023?"

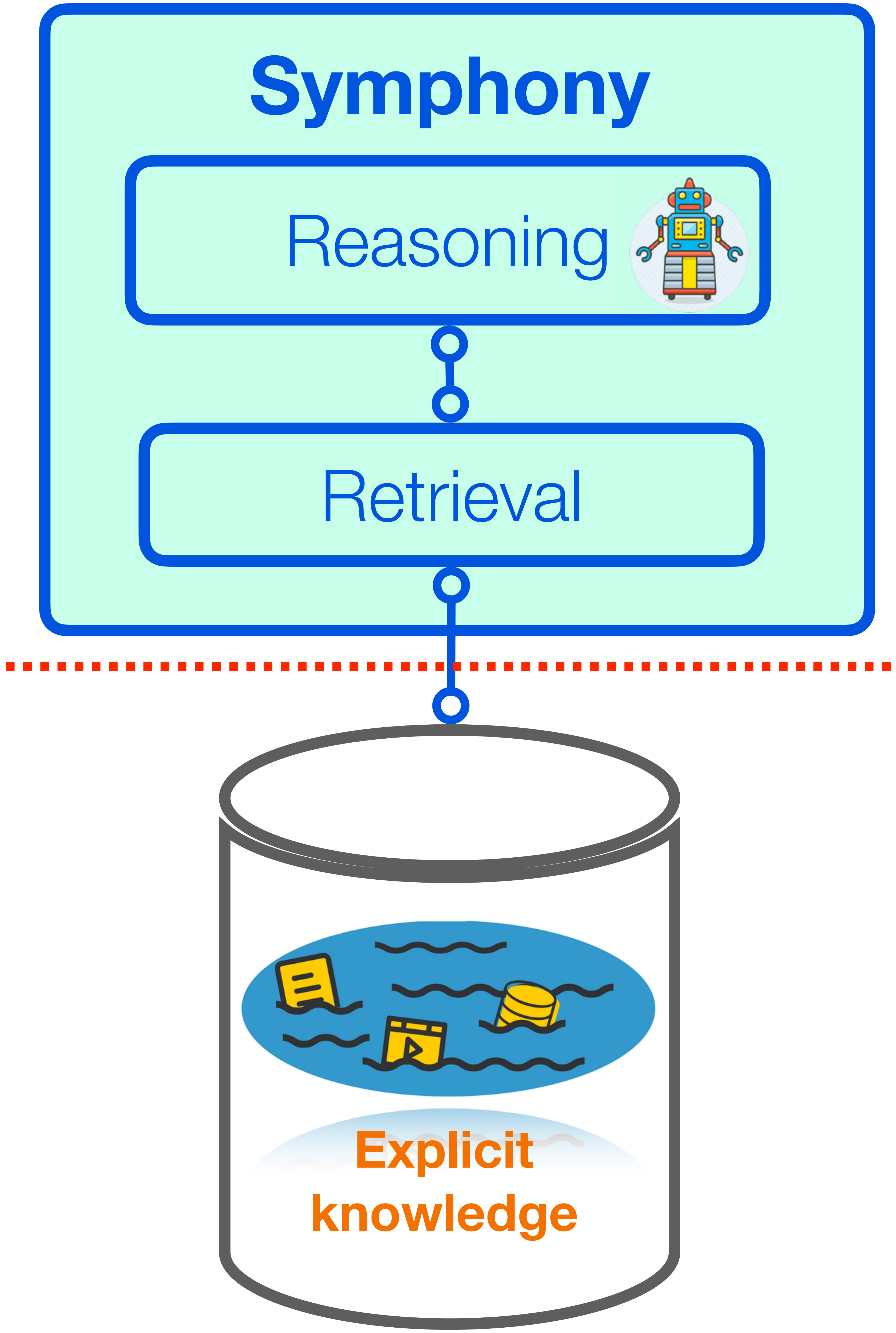
File



The Netherlands has basically reinvented the tomato

By Netherlands Foreign Investment Agency | Published February 17, 2017

Gustavo Alonso keynote at CIDR'23



Symphony

Reasoning

Retrieval

Explicit knowledge

A Running Example

(1) Index

A Running Example

(1) Index

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

A Running Example

(1) Index

Q 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, Sulochna, Sajjan, Agha and Bhagwan Dada…

P1

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

T1

Year	Song	Film	…
1971	Zindagi Ek Safer	Andaz	…
1971	Yeh Jo Mohabbat	Kati Patang	…
1975	Main Pyaasa tum	Faraar	…
…	…	…	…

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

A Running Example

(1) Index

Q 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, Sulochna, Sajjan, Agha and Bhagwan Dada… P1

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

T1

Year	Song	Film	…
1971	Zindagi Ek Safer	Andaz	…
1971	Yeh Jo Mohabbat	Kati Patang	…
1975	Main Pyaasa tum	Faraar	…
…	…	…	…

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

(3) Query Decomposition

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.**

A Running Example

(1) Index

Q 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, Sulochna, Sajjan, Agha and Bhagwan Dada… P1

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

T1

Year	Song	Film	…
1971	Zindagi Ek Safer	Andaz	…
1971	Yeh Jo Mohabbat	Kati Patang	…
1975	Main Pyaasa tum	Faraar	…
…	…	…	…

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

(3) Query Decomposition

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.**

(4) Sub-query Evaluation

"Faraar" *"Main Pyaasa tum"*

res(q1) over P1 → res(q2) over T1 → res(Q)

with NLP Question Answering **with NL2SQL**

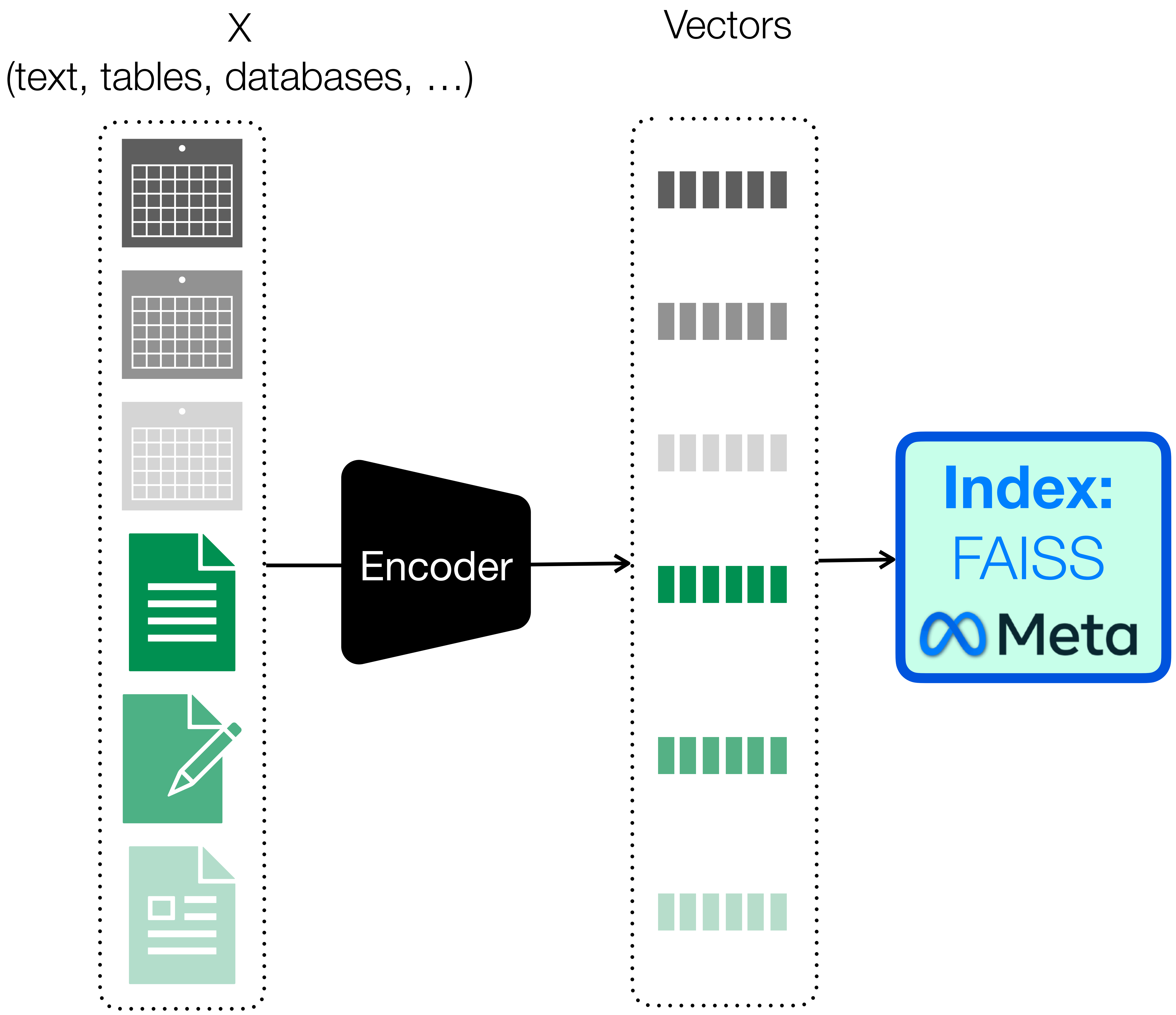
(1) Index (X-to-Vec)

(2) Data Retrieval

(3) Query Decomposition

(4) Sub-query execution

(1) Index (X-to-Vec) | (2) Data Retrieval | (3) Query Decomposition | (4) Sub-query execution



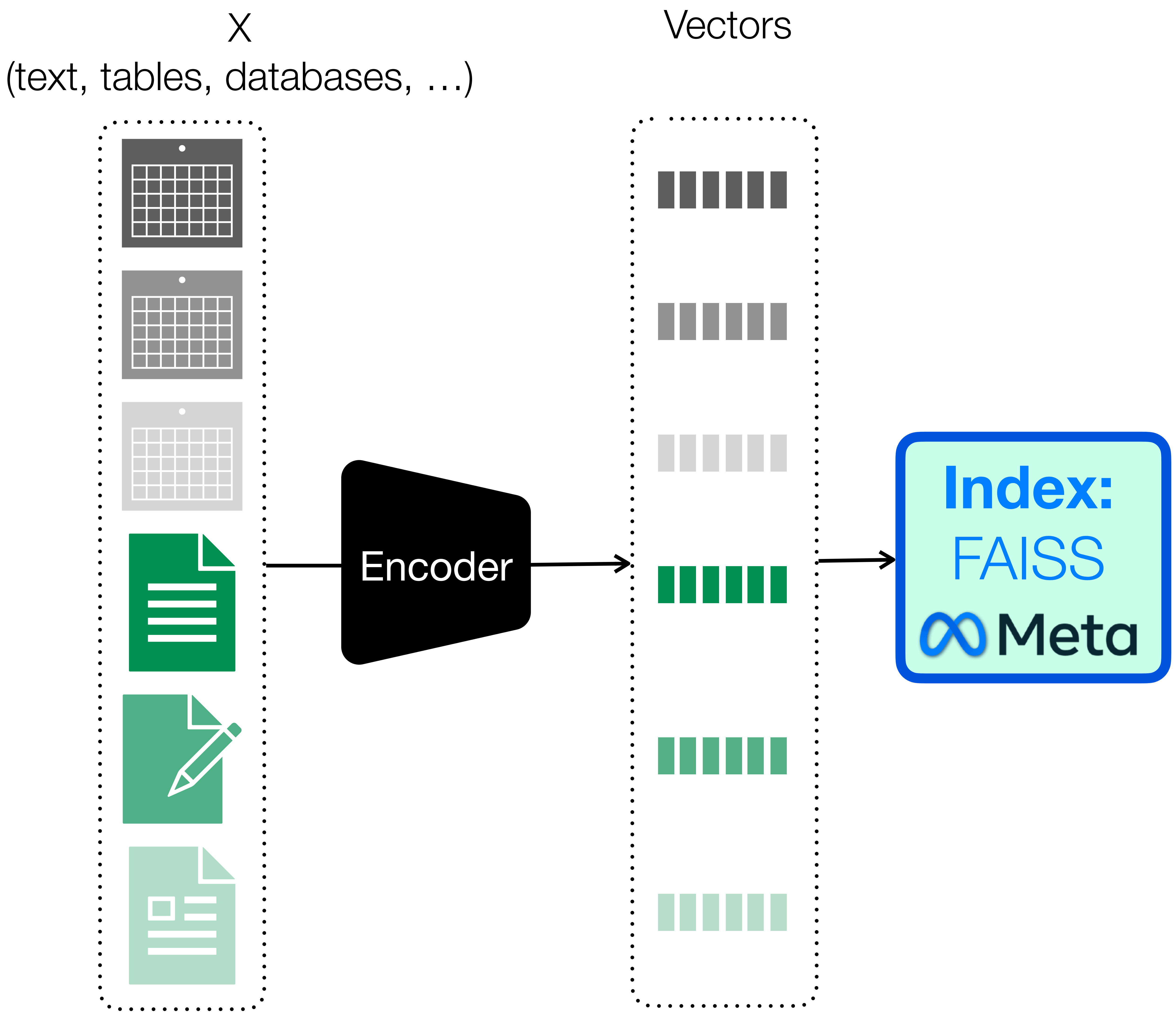
X
(text, tables, databases, …)

Vectors

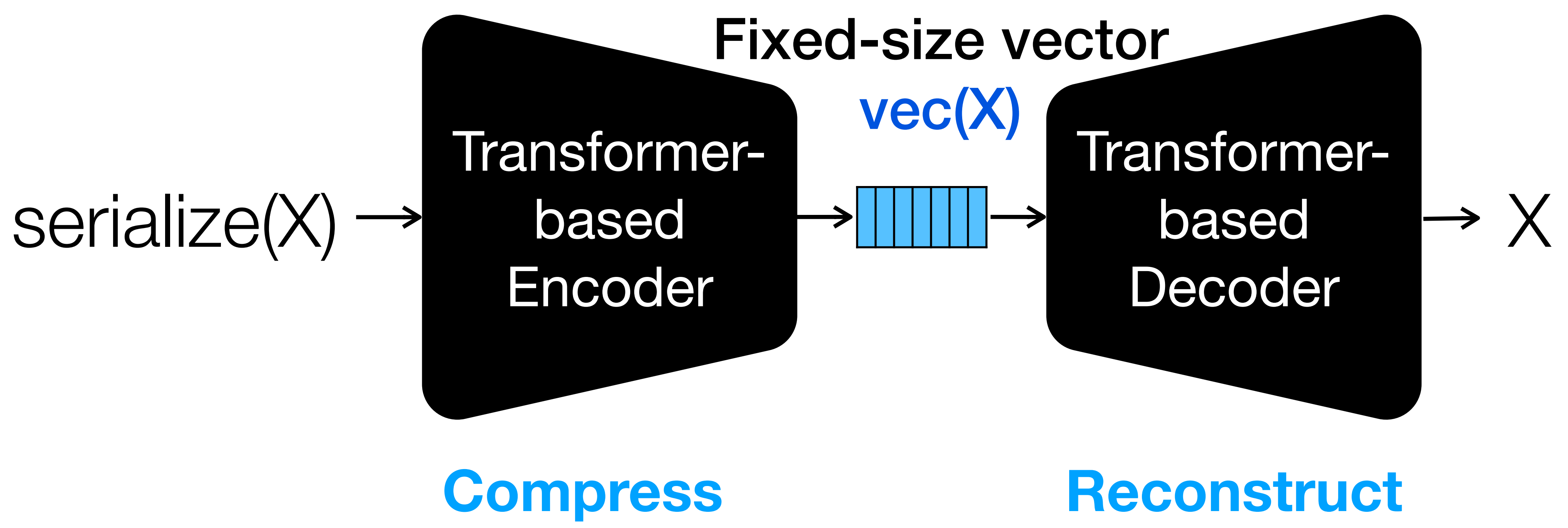
Encoder

Index:
FAISS
Meta

(1) Index (X-to-Vec) | (2) Data Retrieval | (3) Query Decomposition | (4) Sub-query execution



AutoEncoders

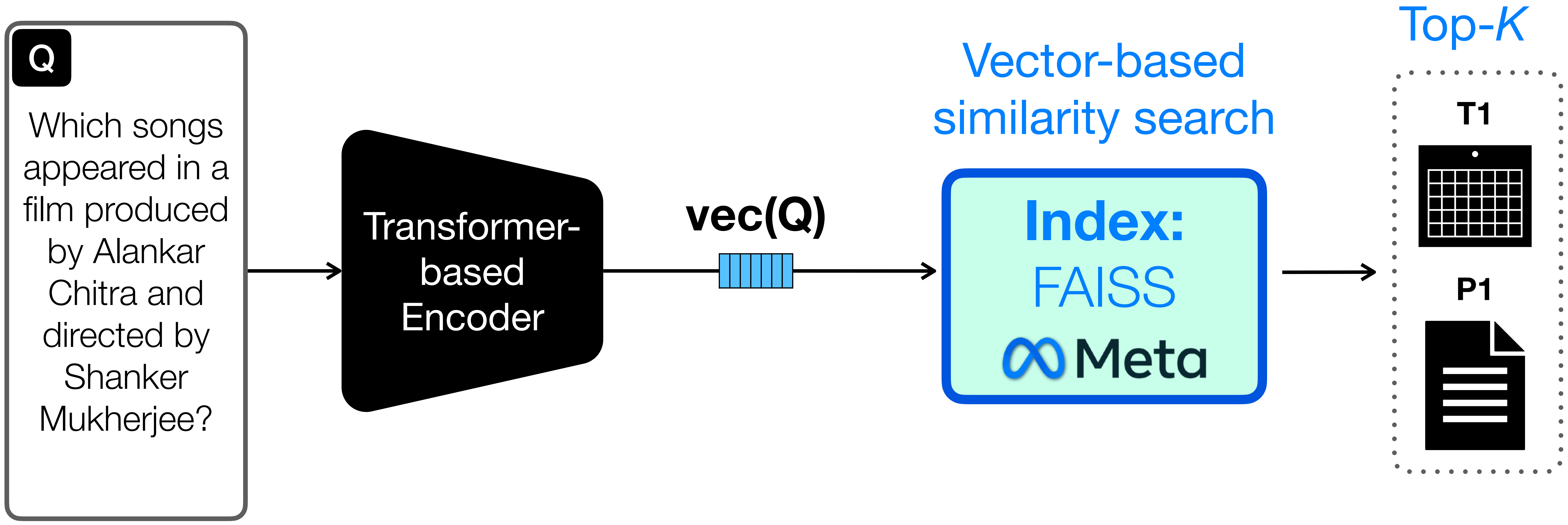


(1) Index (X-to-Vec)

(2) Data Retrieval

(3) Query Decomposition

(4) Sub-query execution



Q

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

Transformer-based Encoder

vec(Q)

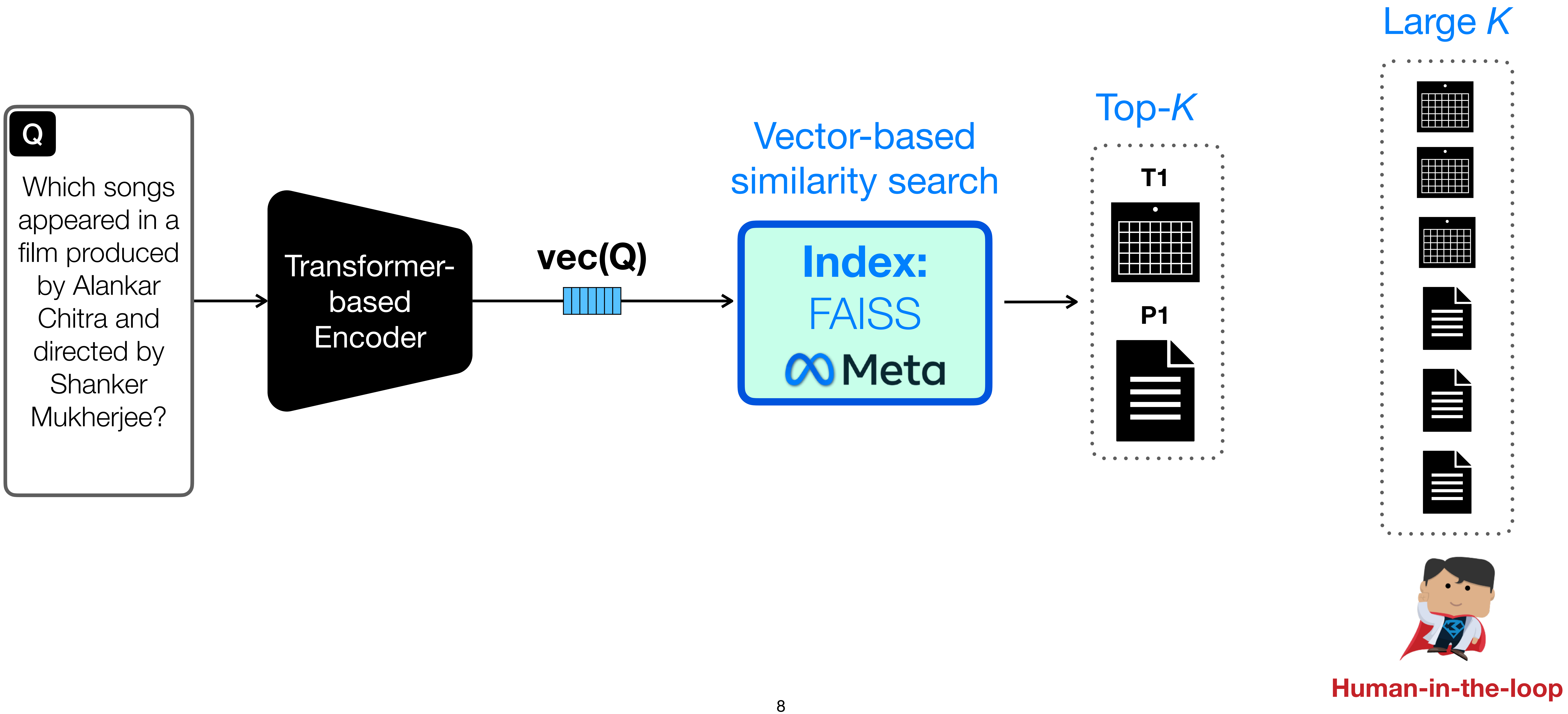
Vector-based similarity search

Index: FAISS Meta

Top-*K*

T1

P1



Q

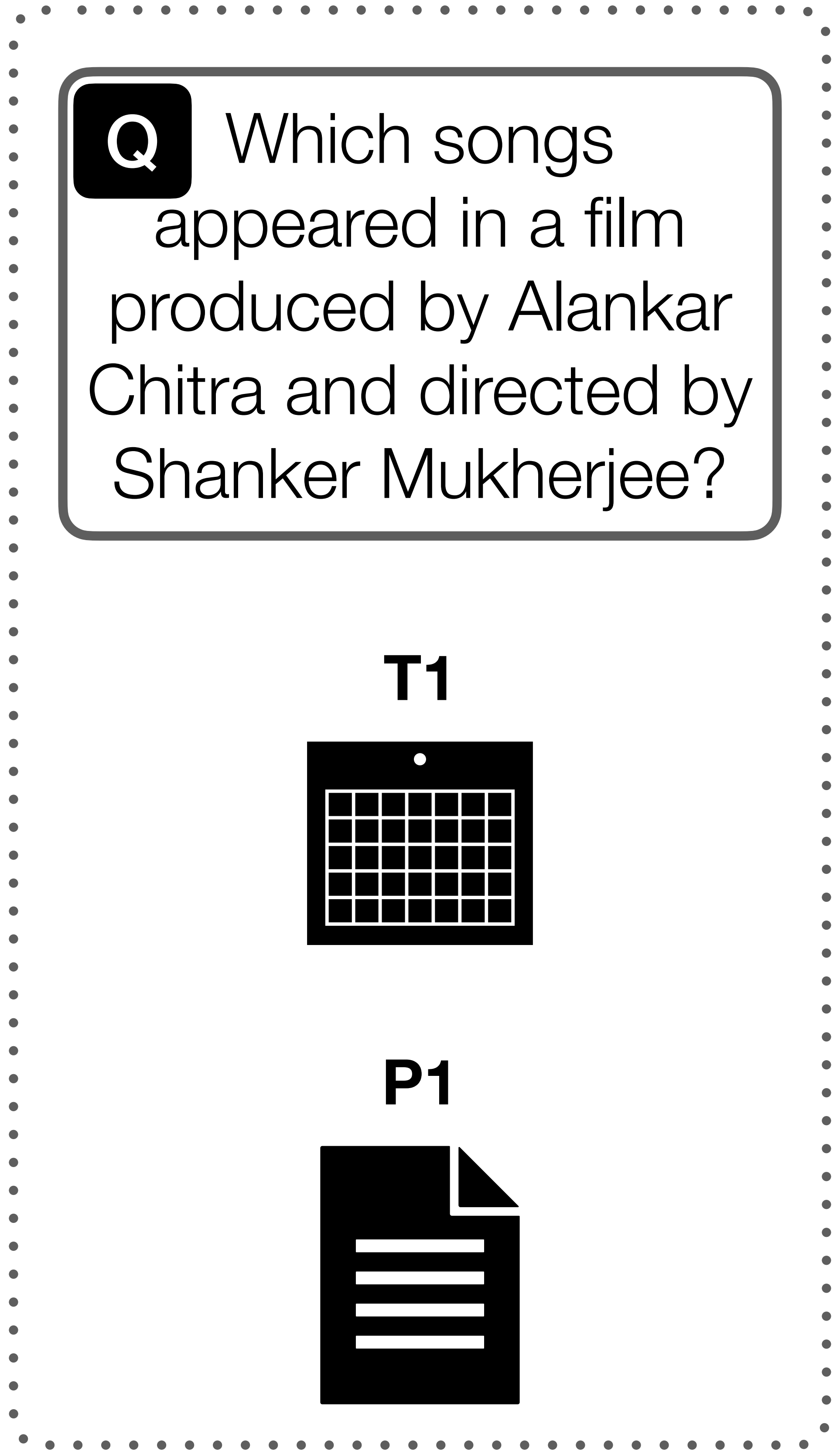
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?

(1) Index (X-to-Vec)

(2) Data Retrieval

(3) Query Decomposition

(4) Sub-query execution



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

(1) Index (X-to-Vec)

(2) Data Retrieval

(3) Query Decomposition

(4) Sub-query execution

GPT-3

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

T1

P1

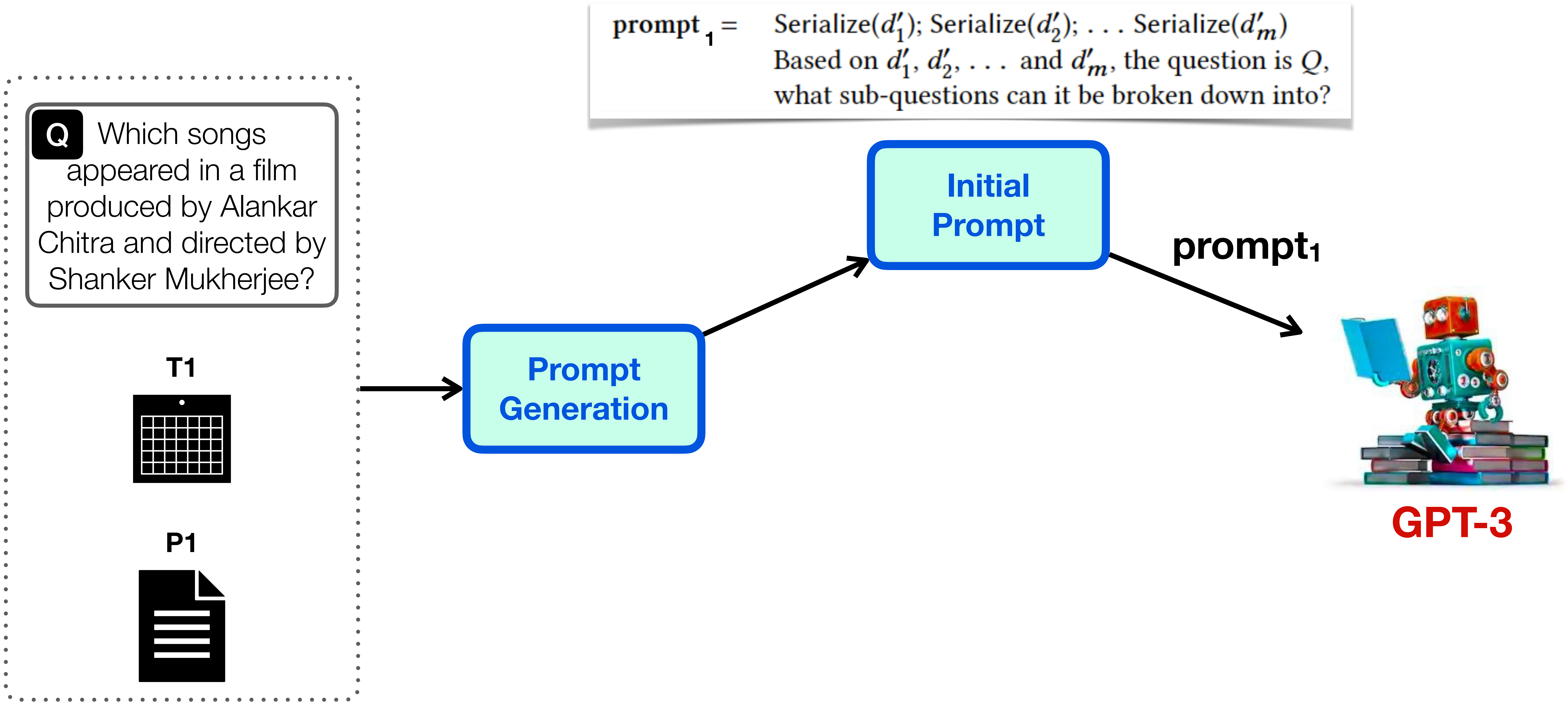
Prompt 1: **The passage P1** has the following content: …
The table T1 has the following columns: Year, Song, Film, Music Director, Lyricist.
Based on P1 and T1, the question is “Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?”.
What sub-questions can it be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

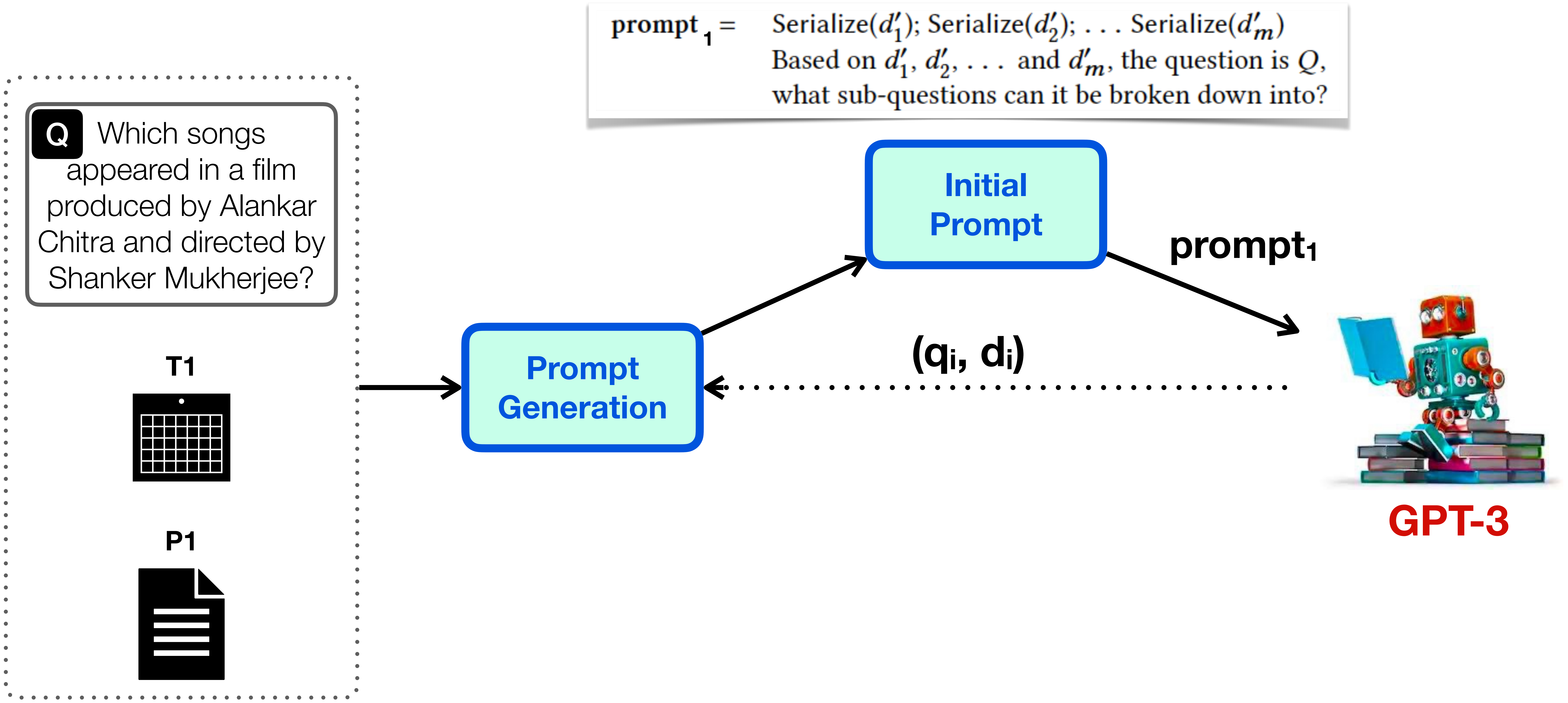
(1) Index (X-to-Vec) | (2) Data Retrieval | (3) Query Decomposition | (4) Sub-query execution



prompt$_1$ = Serialize(d'_1); Serialize(d'_2); . . . Serialize(d'_m)
Based on d'_1, d'_2, . . . and d'_m, the question is Q, what sub-questions can it be broken down into?

Q Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
T1
P1
Prompt Generation
Initial Prompt
prompt₁
GPT-3

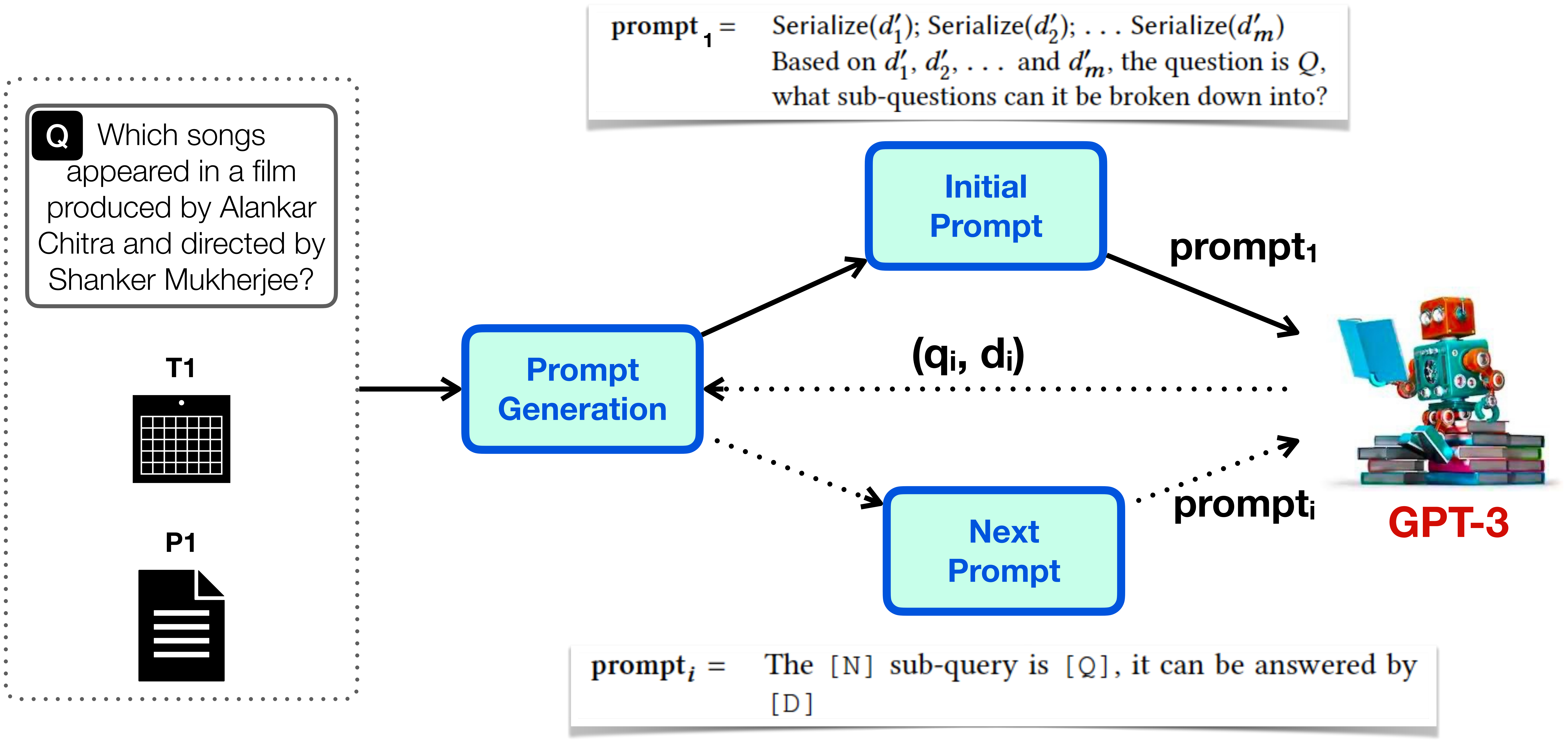
(1) Index (X-to-Vec) | (2) Data Retrieval | (3) Query Decomposition | (4) Sub-query execution



prompt$_1$ = Serialize(d'_1); Serialize(d'_2); . . . Serialize(d'_m)
Based on d'_1, d'_2, . . . and d'_m, the question is Q,
what sub-questions can it be broken down into?

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

(1) Index (X-to-Vec) | (2) Data Retrieval | (3) Query Decomposition | (4) Sub-query execution

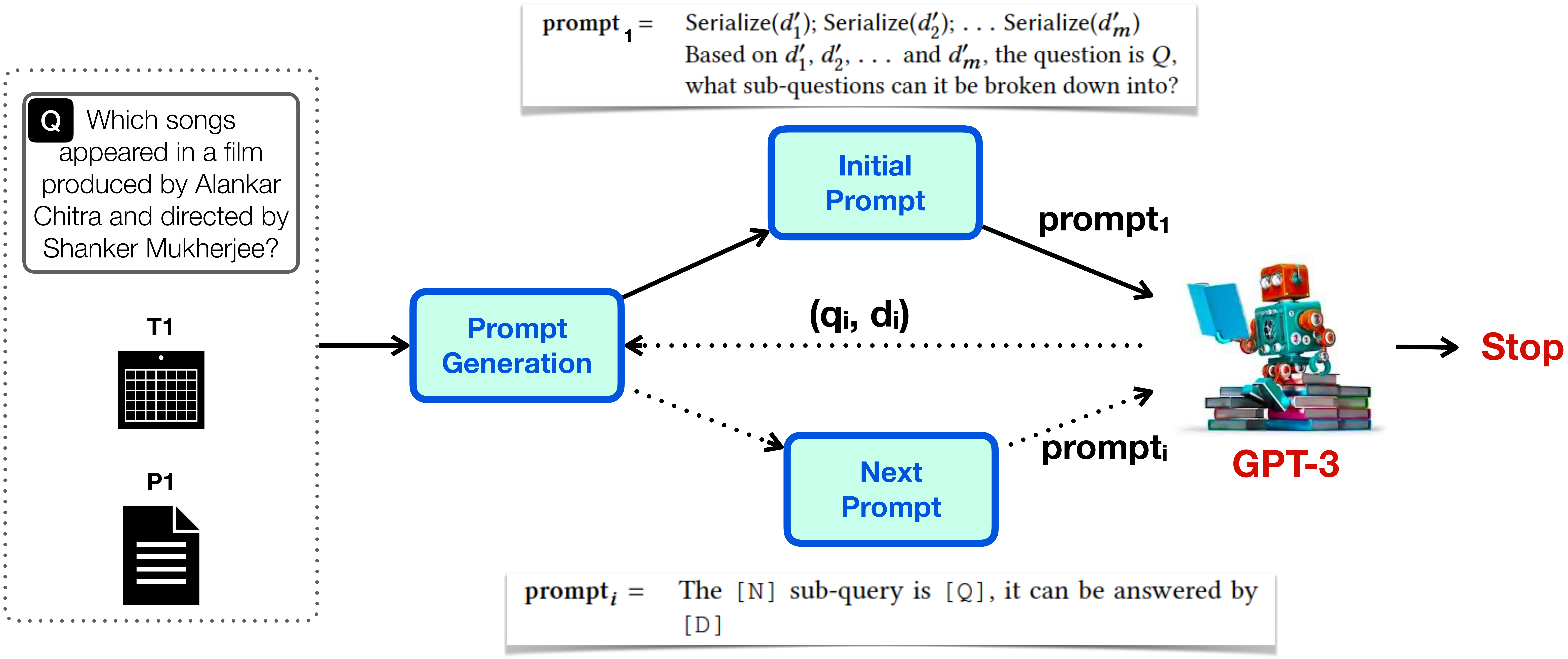


prompt$_1$ = Serialize(d'_1); Serialize(d'_2); . . . Serialize(d'_m)
Based on d'_1, d'_2, . . . and d'_m, the question is Q, what sub-questions can it be broken down into?

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

prompt$_i$ = The [N] sub-query is [Q], it can be answered by [D]

(1) Index (X-to-Vec) | (2) Data Retrieval | (3) Query Decomposition | (4) Sub-query execution



prompt$_1$ = Serialize(d'_1); Serialize(d'_2); . . . Serialize(d'_m)
Based on d'_1, d'_2, . . . and d'_m, the question is Q, what sub-questions can it be broken down into?

prompt$_i$ = The [N] sub-query is [Q], it can be answered by [D]

Q
Which songs appeared in a film produced by Alankar Chitra and directed by Shanker Mukherjee?
T1
P1
Prompt Generation
Initial Prompt
Next Prompt
prompt₁
(qᵢ, dᵢ)
promptᵢ
GPT-3
Stop

(1) Index (X-to-Vec)

(2) Data Retrieval

(3) Query Decomposition

(4) Sub-query execution

Natural Language Query over Text

NLP
Question
Answering

Natural Language Query over Table(s)

NL2SQL
(SIGMOD'23)

(1) Index (X-to-Vec)

(2) Data Retrieval

(3) Query Decomposition

(4) Sub-query execution

Natural Language Query over Text

NLP Question Answering

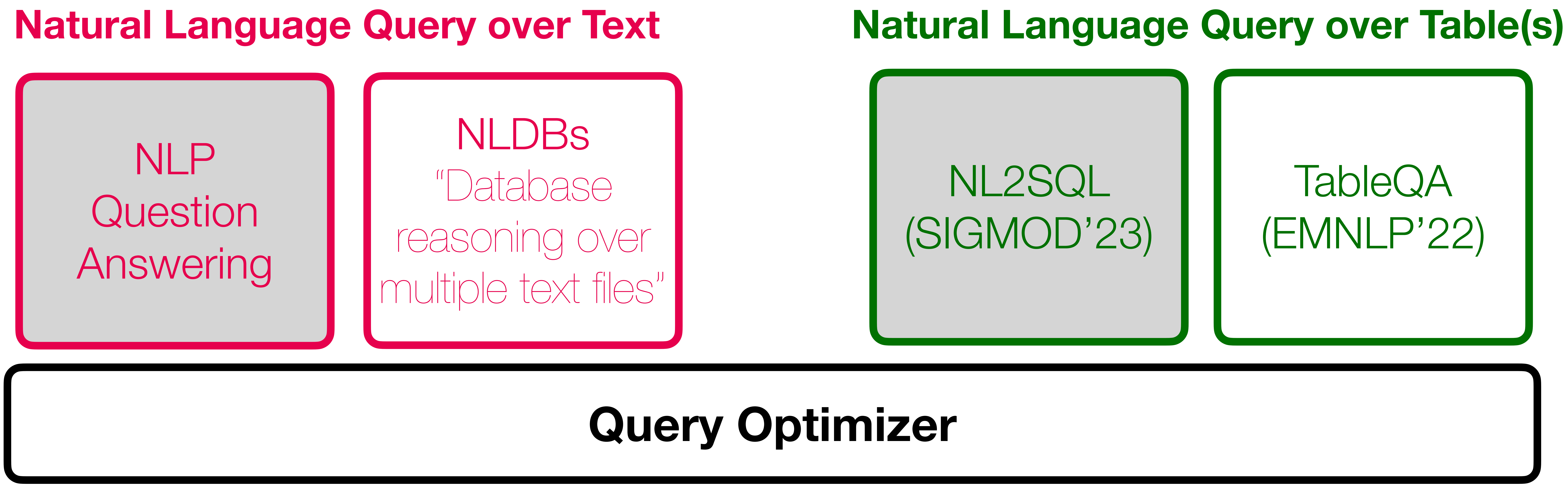
NLDBs "Database reasoning over multiple text files"

Natural Language Query over Table(s)

NL2SQL (SIGMOD'23)

TableQA (EMNLP'22)

(1) Index (X-to-Vec) | (2) Data Retrieval | (3) Query Decomposition | (4) Sub-query execution



Natural Language Query over Text

NLP Question Answering

NLDBs "Database reasoning over multiple text files"

Natural Language Query over Table(s)

NL2SQL (SIGMOD'23)

TableQA (EMNLP'22)

Query Optimizer

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 T1

Rank	Official Name	Height (m)	…
1	The Shard	310	…
2	22 Bishopsgate	278	…
3	One Canada Square	235	…
…	…	…	…

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

Table T2

Rank	Name	Height (m)	…
1	Lotte World Tower	550	…
2	Landmark Tower	412	…
3	Tower A	339	…
…	…	…	…

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

- 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

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

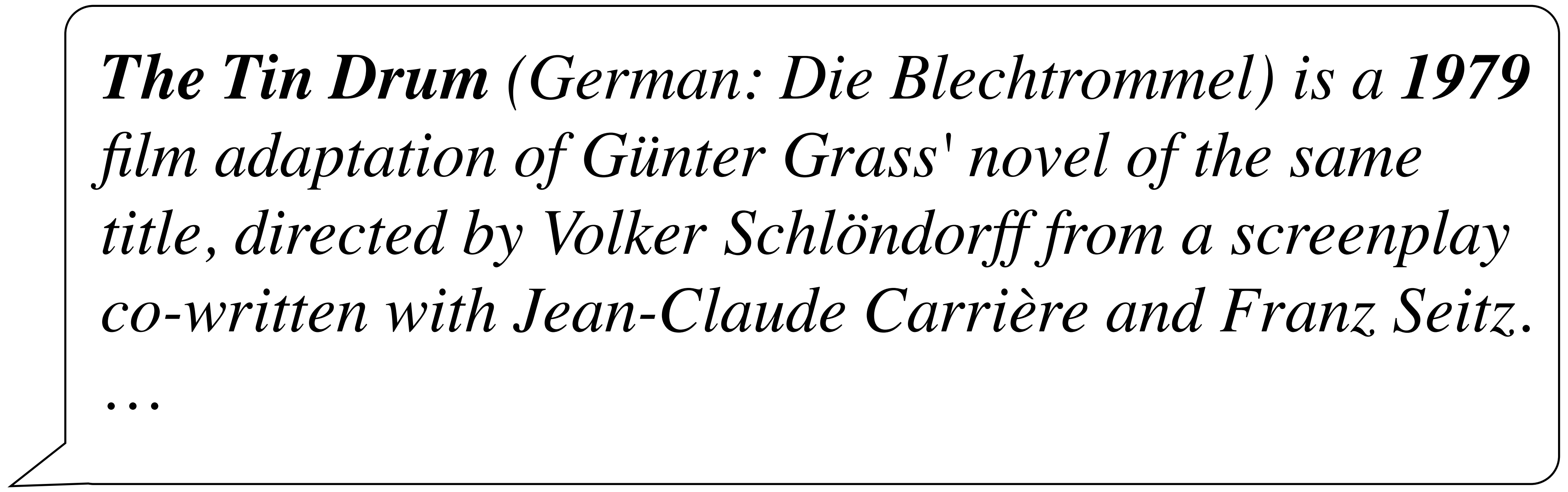
Source: https://en.wikipedia.org/wiki/List_of_German_submissions_for_the_Academy_Award_for_Best_Foreign_Language_Film

- Decomposed queries:
  1. "What year was the film The Tin Drum released?", it can be answered by T1;
  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:

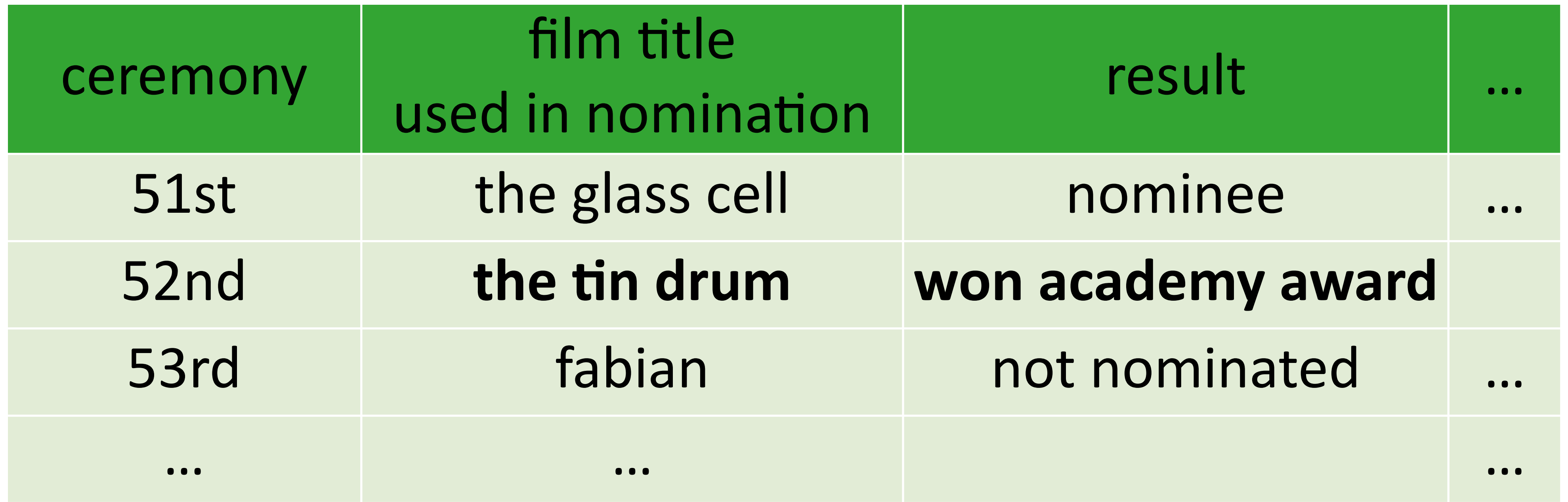
Passage P1



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



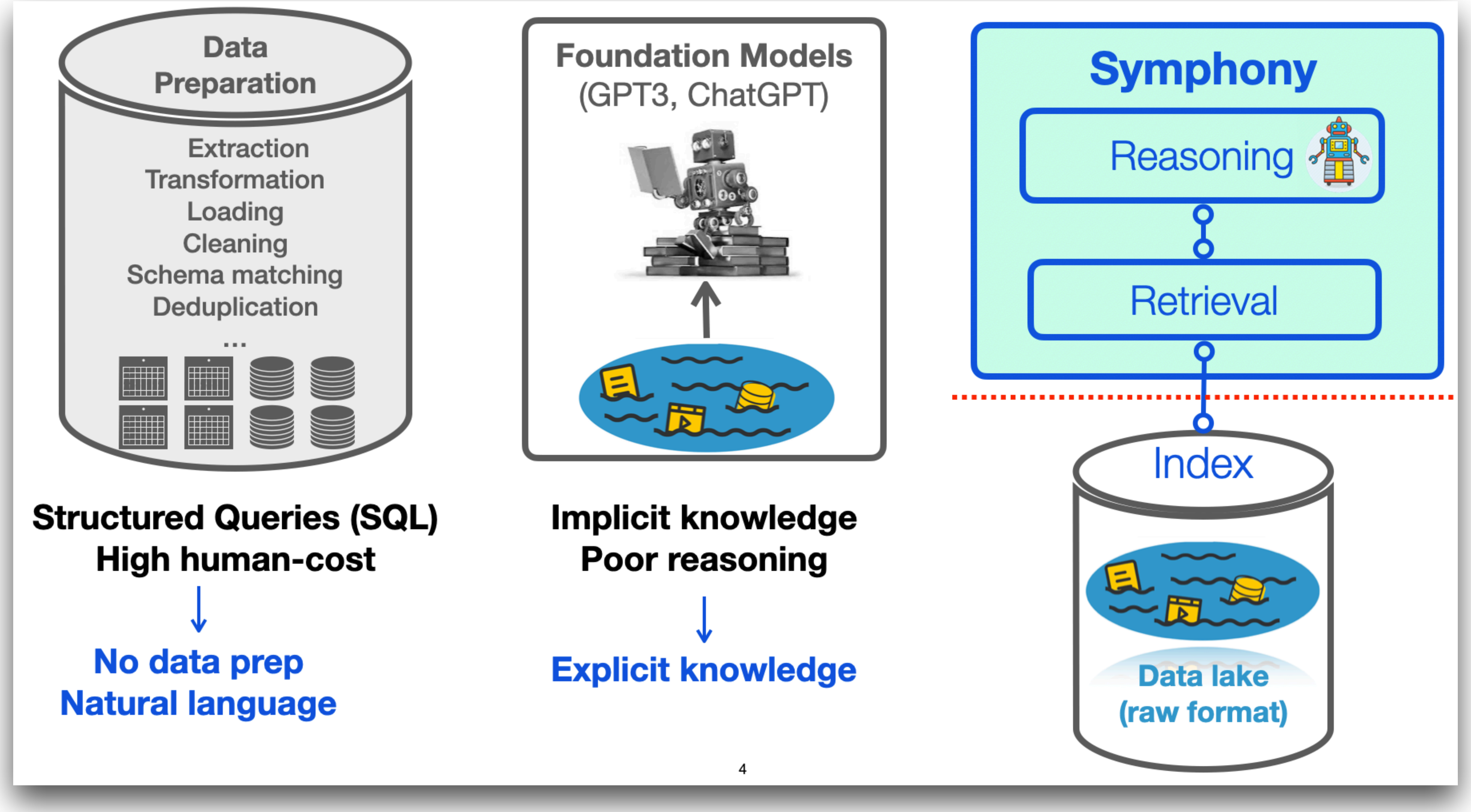
ceremony	film title used in nomination	result	…
51st	the glass cell	nominee	…
52nd	**the tin drum**	**won academy award**	
53rd	fabian	not nominated	…
…	…		…

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



Data Preparation

Extraction
Transformation
Loading
Cleaning
Schema matching
Deduplication
...

Structured Queries (SQL)
High human-cost
↓
No data prep
Natural language

Foundation Models
(GPT3, ChatGPT)

Implicit knowledge
Poor reasoning
↓
Explicit knowledge

Symphony

Reasoning

Retrieval

Index

Data lake
(raw format)

Future Work

- Improving the index (X-to-Vec)
  - Contrastive learning
  - Combining with traditional string similarity search
- Sub-query execution module
  - NL2SQL