

CSE 199: UB Seminar (Fall 2026) - Data management

Lecture 5: Natural Language to SQLa

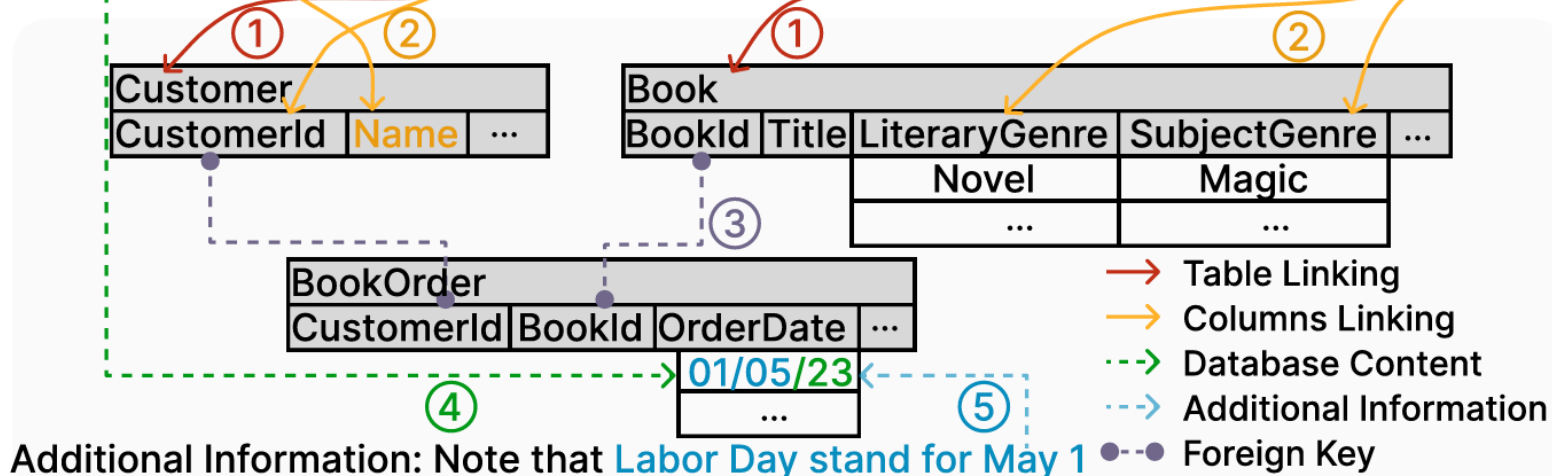
Zhuoyue Zhao
zzhao35 [at] buffalo [dot] edu

How does human translate NL question into SQL?

NL Query:

Find the **names** of all **customers** who checked out **books** on exactly 3 different **genres** on **Labor Day in 2023**.

Database:



Additional Information: Note that **Labor Day** stand for **May 1**

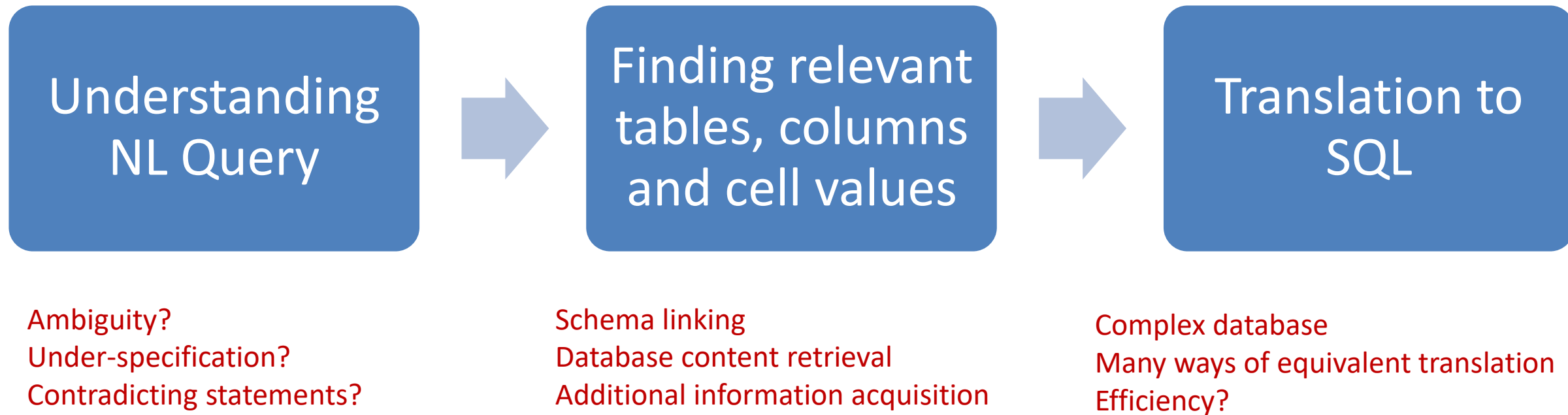
SQL:

```
SELECT Name
FROM Customer
NATURAL JOIN BookOrder
NATURAL JOIN Book
WHERE OrderDate='01/05/23'
GROUP BY CustomerId, Name
HAVING COUNT(DISTINCT SubjectGenre)=3
```

```
SELECT Name
FROM Customer
WHERE CustomerId = (SELECT CustomerId
FROM BookOrder NATURAL JOIN Book
WHERE OrderDate='01/05/23'
GROUP BY CustomerId
HAVING COUNT(DISTINCT SubjectGenre)=3)
```

(a) An Example of NL2SQL (Ambiguous NL)

How does human translate NL question into SQL?



NL2SQL in LLM era

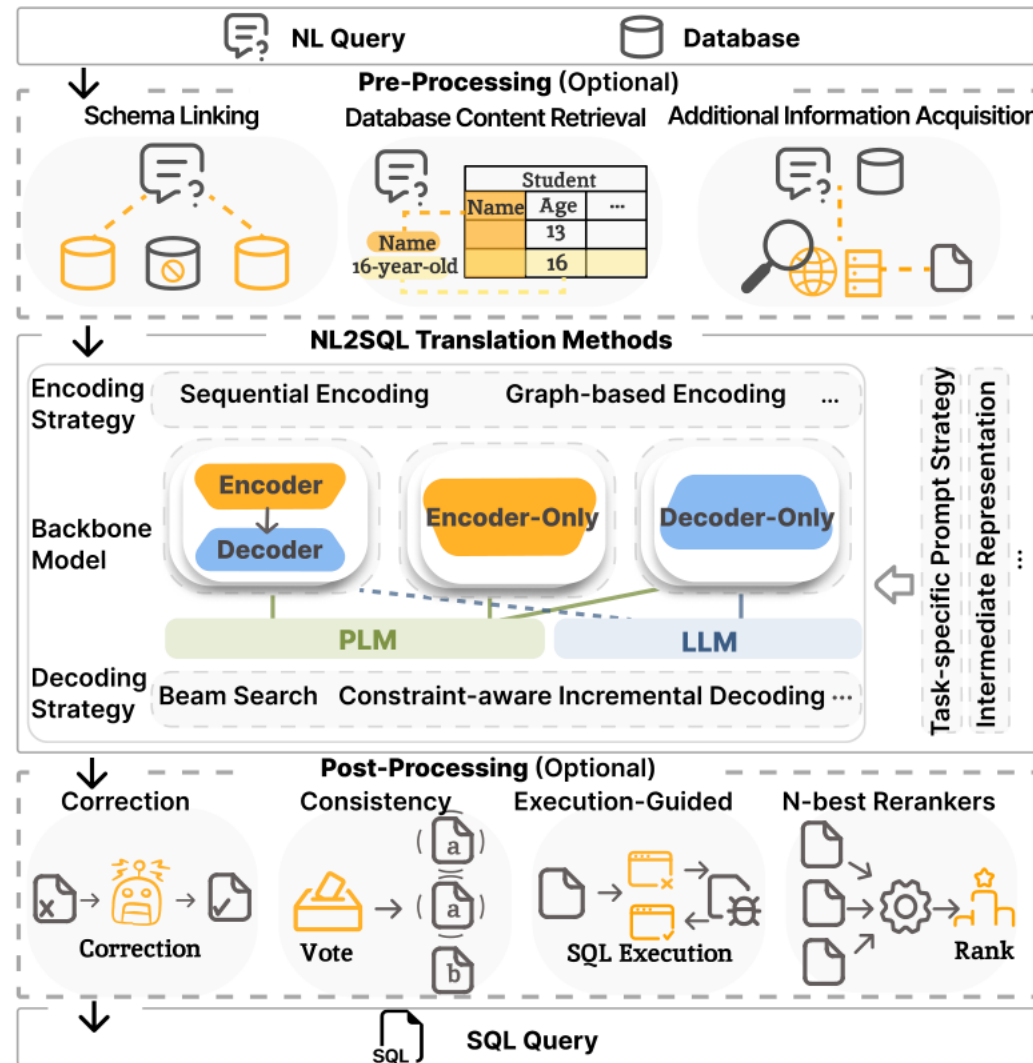


Fig. 5: An Overview of NL2SQL Modules in the LLM Era.

Few-Shot NL2SQL

- DAIL-SQL

```
1 Below is an instruction that describes a task, paired
  ↳ with an input that provides further context. Write a
  ↳ response that appropriately completes the request.
2
3 ### Instruction:
4 Write a sql to answer the question "How many continents
  ↳ are there?"
5
6 ### Input:
7 continents(ContId, Continent)
8 countries(CountryId, CountryName, Continent)
9
10 ### Response:
11 SELECT
```

Question representation

```
1 /* Some example questions and corresponding SQL queries
  ↳ are provided based on similar problems: */
2 /* Answer the following: How many authors are there? */
3 SELECT count(*) FROM authors
4
5 /* Answer the following: How many farms are there?. */
6 SELECT count(*) FROM farm
7
8 ${TARGET_QUESTION}
```

*In-context learning
(providing similar examples)*

Question Representation in DAIL-SQL

```
1 Table continents, columns = [ContId, Continent]
2 Table countries, columns = [CountryId, CountryName,
  ↳ Continent]
3 Q: How many continents are there?
4 A: SELECT
```

Listing 1: Example of Basic Prompt

```
1 Given the following database schema:
2 continents: ContId, Continent
3 countries: CountryId, CountryName, Continent
4
5 Answer the following: How many continents are there?
6 SELECT
```

Listing 2: Example of Text Representation Prompt

```
1 ### Complete sqlite SQL query only and with no
  ↳ explanation
2 ### SQLite SQL tables, with their properties:
3 #
4 # continents(ContId, Continent)
5 # countries(CountryId, CountryName, Continent)
6 #
7 ### How many continents are there?
8 SELECT
```

Listing 3: Example of OpenAI Demostration Prompt

```
1 /* Given the following database schema: */
2 CREATE TABLE continents(
3     ContId int primary key,
4     Continent text,
5     foreign key(ContId) references countries(Continent)
6 );
7
8 CREATE TABLE countries(
9     CountryId int primary key,
10    CountryName text,
11    Continent int,
12    foreign key(Continent) references continents(ContId)
13 );
14
15 /* Answer the following: How many continents are there?
  ↳ */
16 SELECT
```

Listing 4: Example of Code Representation Prompt

```
1 Below is an instruction that describes a task, paired
  ↳ with an input that provides further context. Write a
  ↳ response that appropriately completes the request.
2
3 ### Instruction:
4 Write a sql to answer the question "How many continents
  ↳ are there?"
5
6 ### Input:
7 continents(ContId, Continent)
8 countries(CountryId, CountryName, Continent)
9
10 ### Response:
11 SELECT
```

Listing 5: Example of Alpaca SFT Prompt

Question Representation in DAIL-SQL

```

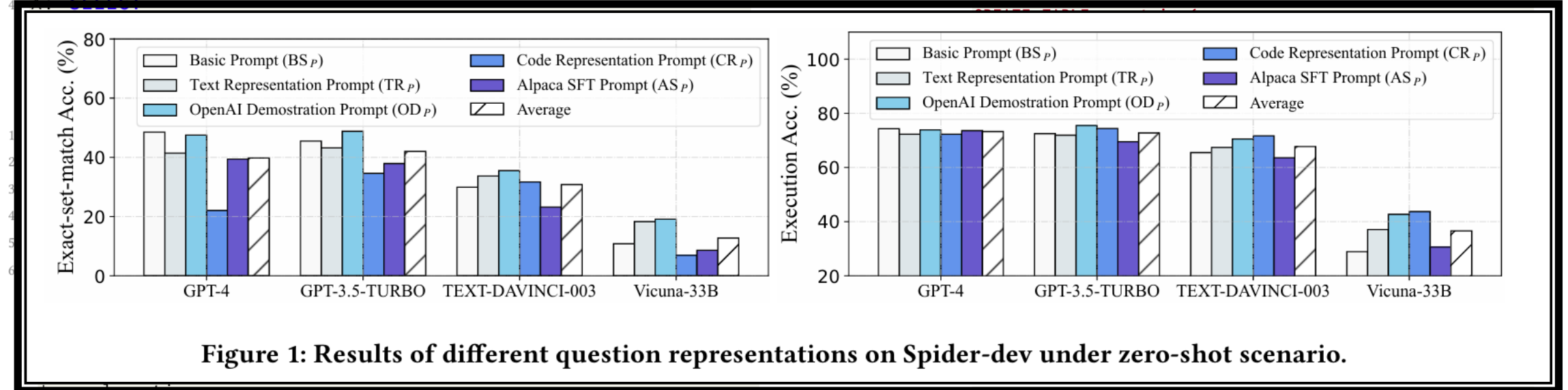
1 Table continents, columns = [ContId, Continent]
2 Table countries, columns = [CountryId, CountryName,
  ↳ Continent]
3 Q: How many continents are there?
4 A: SELECT

```

```

1 /* Given the following database schema: */
2 CREATE TABLE continents(
3     ContId int primary key,
4     Continent text,
5     foreign key(ContId) references countries(Continent)
6 );
7

```



```

1 Explanation:
2 ### SQLite SQL tables, with their properties:
3 #
4 # continents(ContId, Continent)
5 # countries(CountryId, CountryName, Continent)
6 #
7 ### How many continents are there?
8 SELECT

```

Listing 3: Example of OpenAI Demonstration Prompt

```

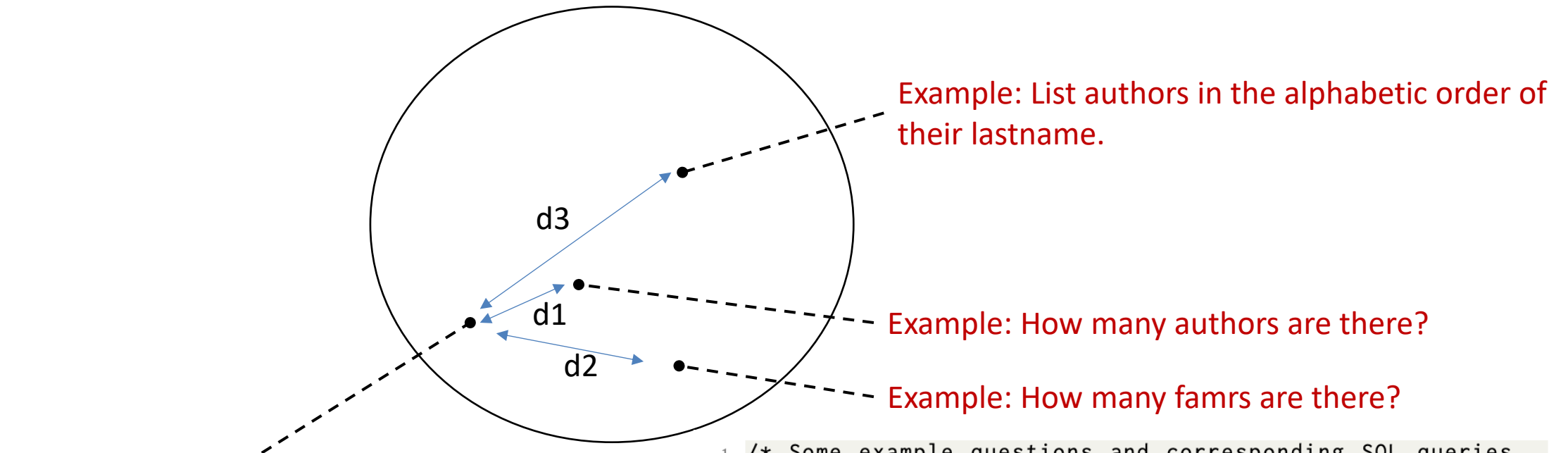
1 ### Instruction:
2 Write a sql to answer the question "How many continents
  ↳ are there?"
3
4 ### Input:
5 continents(ContId, Continent)
6 countries(CountryId, CountryName, Continent)
7
8 ### Response:
9 SELECT

```

Listing 5: Example of Alpaca SFT Prompt

Finding similar queries in DAIL-SQL

Embedding (a high dimensional space)



```
1 /* Some example questions and corresponding SQL queries
   ↳ are provided based on similar problems: */
2 /* Answer the following: How many authors are there? */
3 SELECT count(*) FROM authors
4
5 /* Answer the following: How many farms are there?. */
6 SELECT count(*) FROM farm
7
8 ${TARGET_QUESTION}
```

Example representation in DAIL-SQL

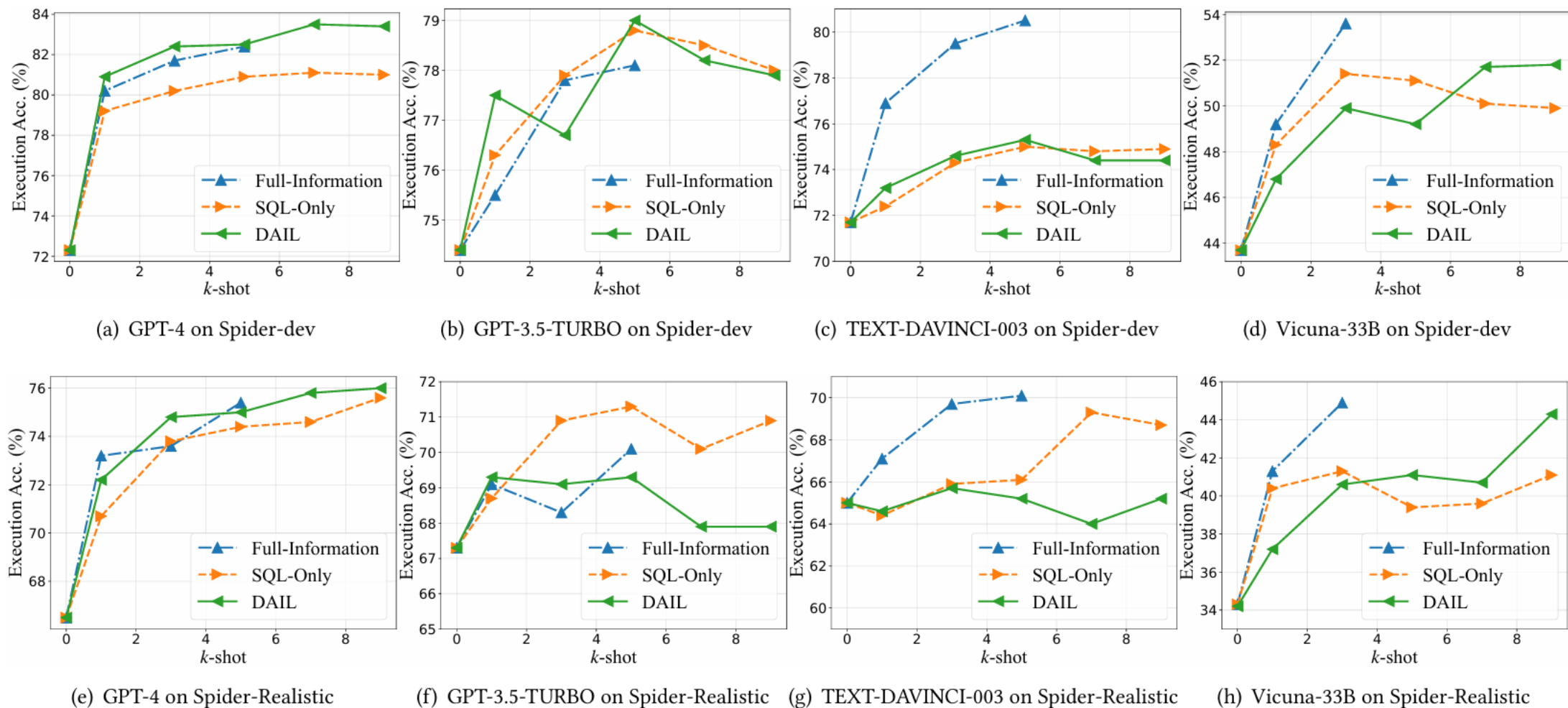
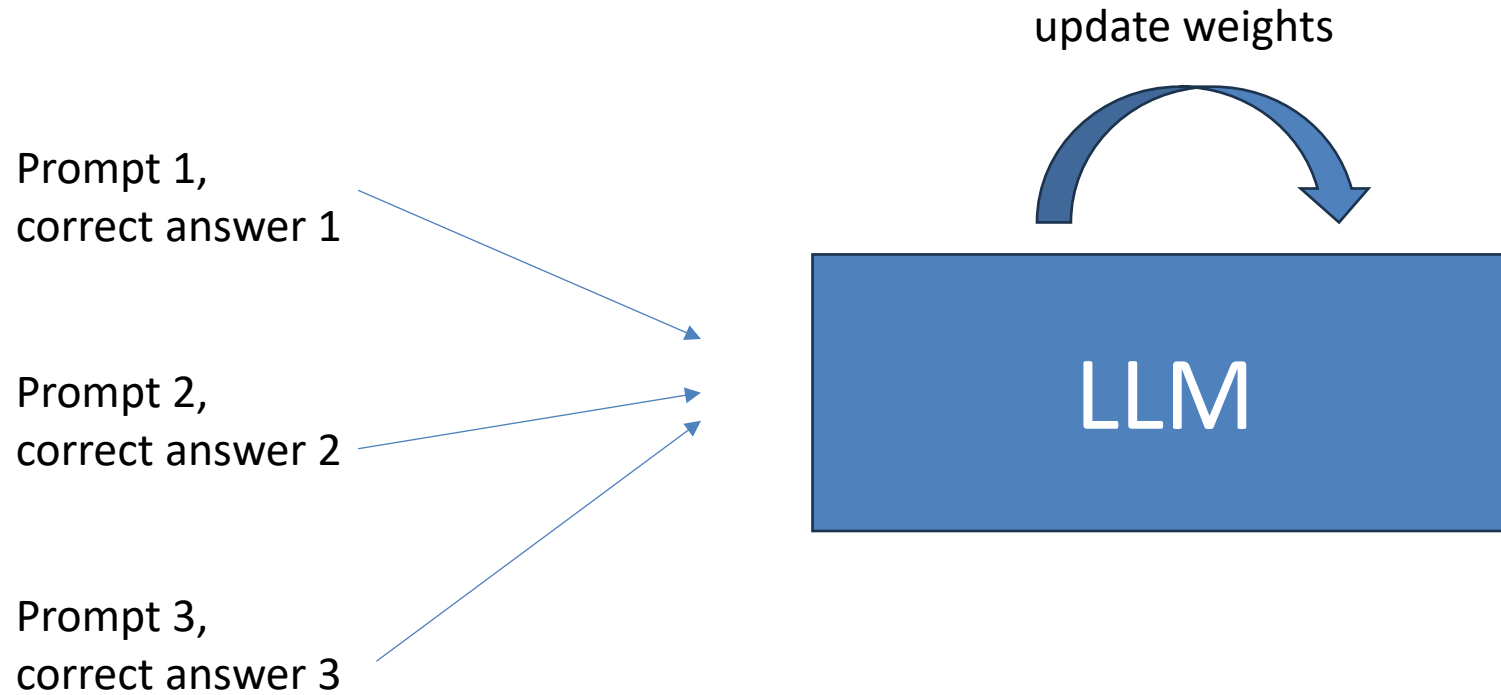


Figure 4: Evaluation on Spider-dev with different example organizations. The selection is fixed to DAIL Selection.

Supervised fine-tuning in DAIL-SQL



Few-Shot NL2SQL

- DAIL-SQL

```
1 Below is an instruction that describes a task, paired
  ↳ with an input that provides further context. Write a
  ↳ response that appropriately completes the request.
2
3 ### Instruction:
4 Write a sql to answer the question "How many continents
  ↳ are there?"
5
6 ### Input:
7 continents(ContId, Continent)
8 countries(CountryId, CountryName, Continent)
9
10 ### Response:
11 SELECT
```

```
1 /* Some example questions and corresponding SQL queries
  ↳ are provided based on similar problems: */
2 /* Answer the following: How many authors are there? */
3 SELECT count(*) FROM authors
4
5 /* Answer the following: How many farms are there?. */
6 SELECT count(*) FROM farm
7
8 ${TARGET_QUESTION}
```

Achieves 86.6% accuracy in Spider, a NL2SQL benchmark.
Does it generalize?

Few-Shot NL2SQL

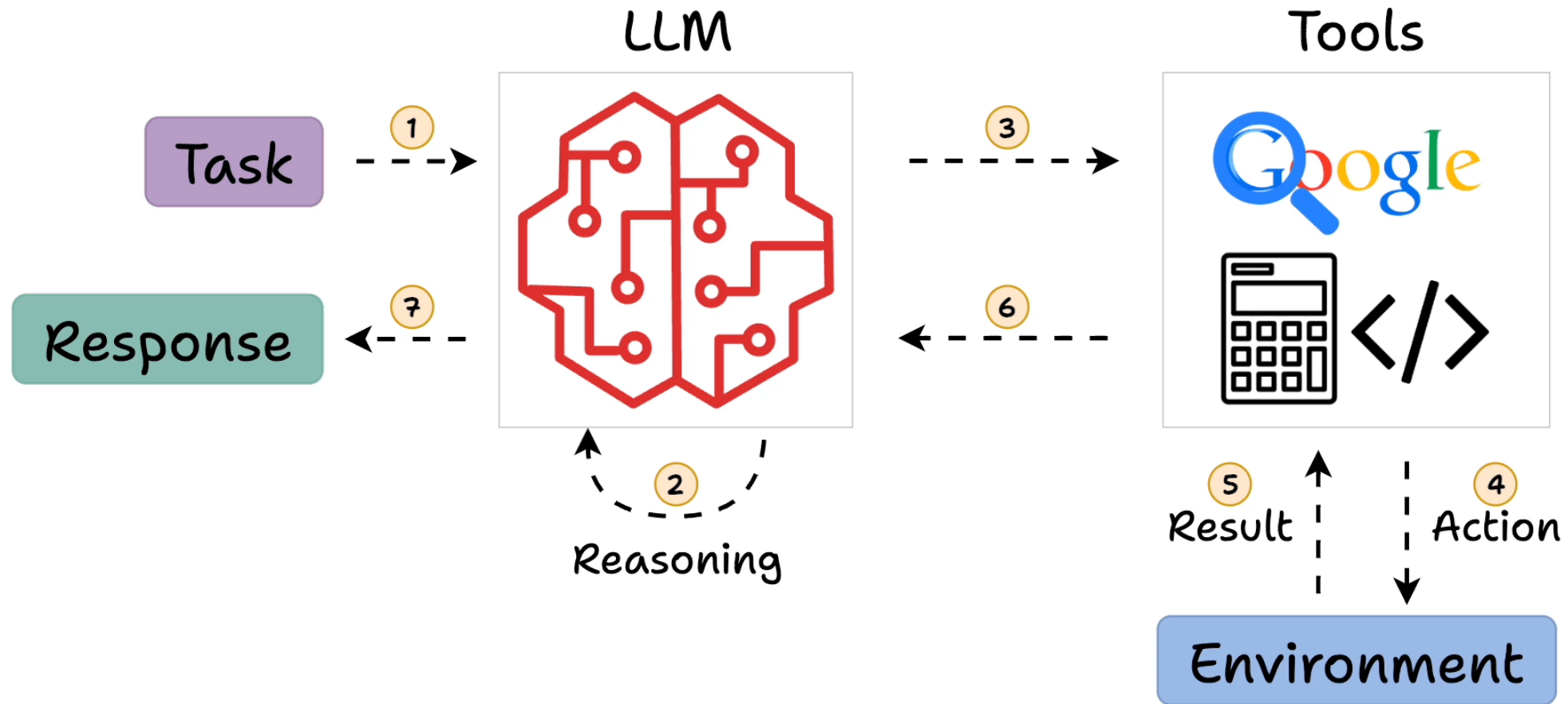
- **DAIL-SQL**

[illegible]

Requires further data annotation and fine-tuning

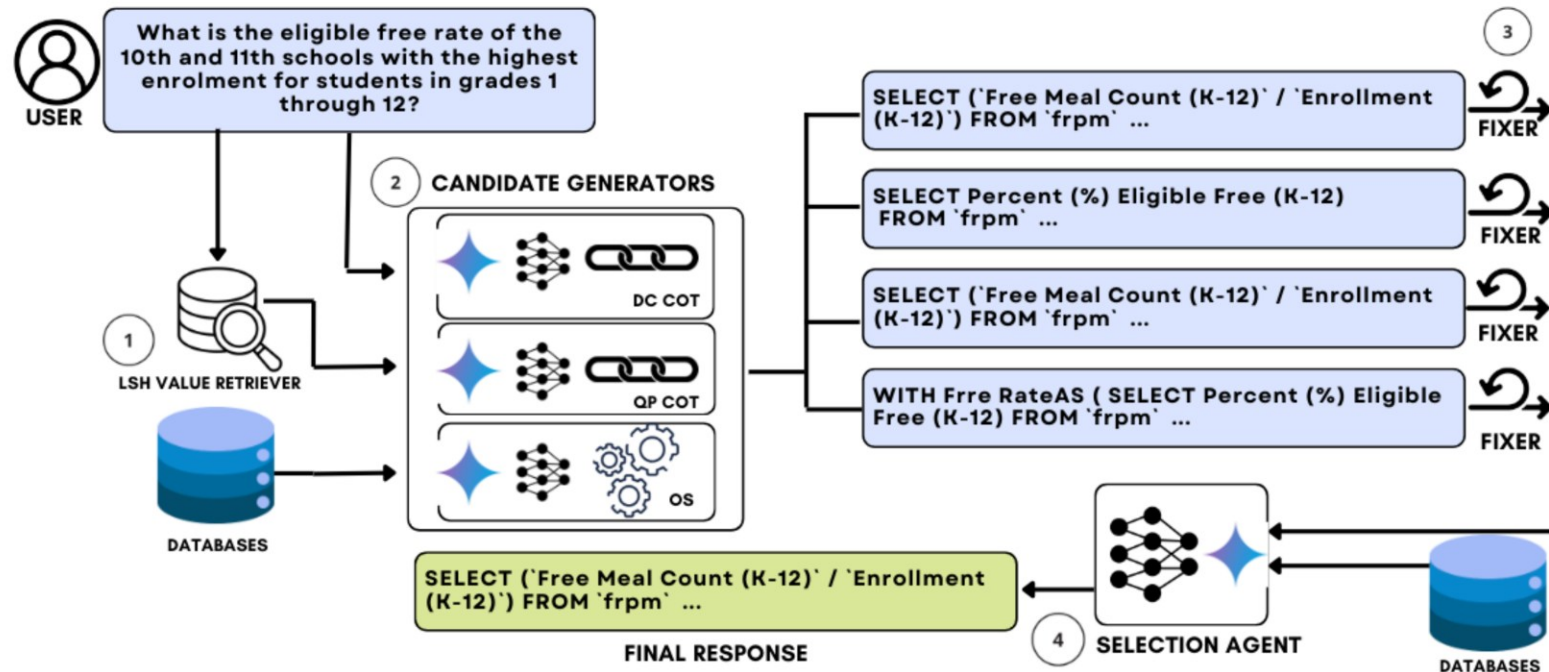
Gao, Dawei, et al. "Text-to-sql empowered by large language models: A benchmark evaluation." PVLDB, 17(5): 1132-1145, 2024.

Agentic NL2SQL



Chase-SQL

- **CHASE-SQL** (*ICLR 2025, Google Cloud and Stanford*)
 - Utilizes the **MinHash LSH** to search for values related to the user query
 - Multiple prompting strategies to **generate various candidate SQL queries** using LLMs, and corrects SQL queries with execution errors through prompting LLMs.
 - Employs an **SQL selection agent** fine-tuned specifically for the database to select the final SQL from multiple candidates.



Quiz time

- Submit your answers in tophat