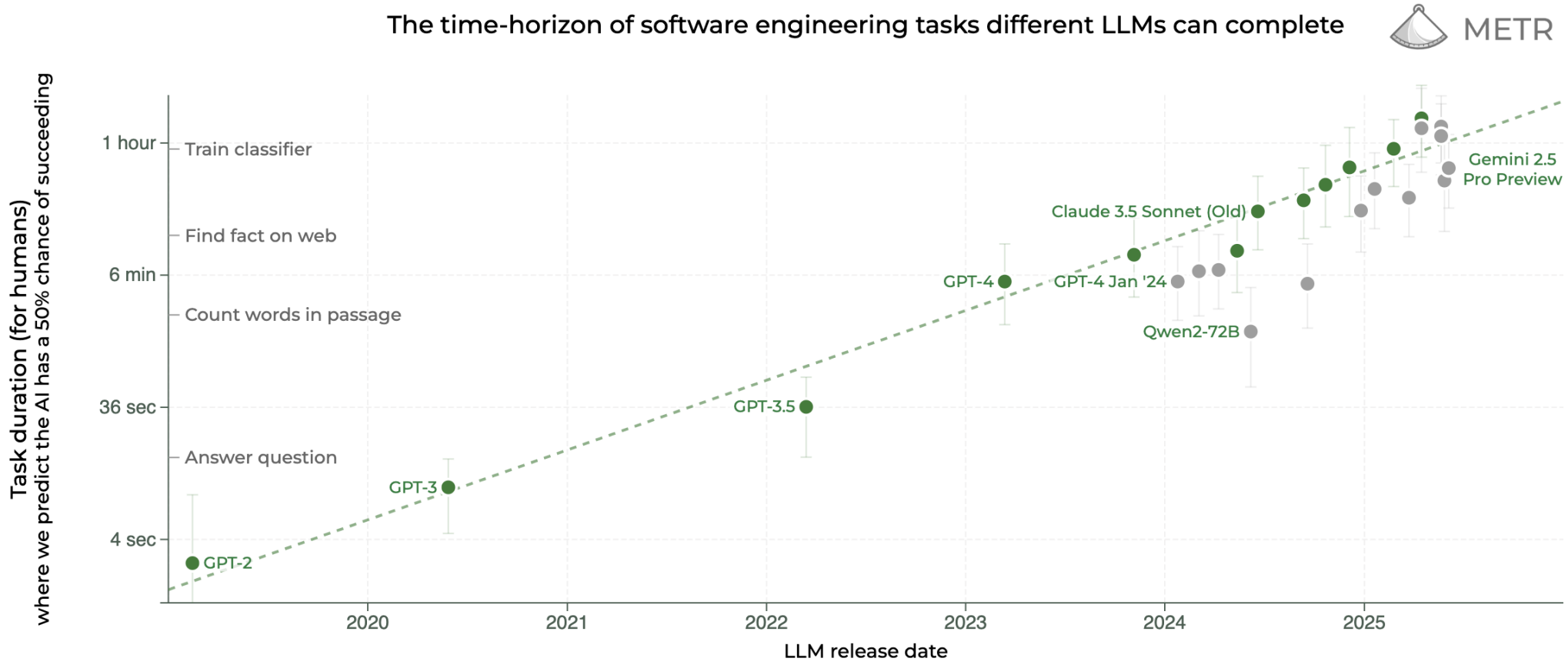
# Text-to-SQL Benchmarks Are Broken

Daniel Kang and others

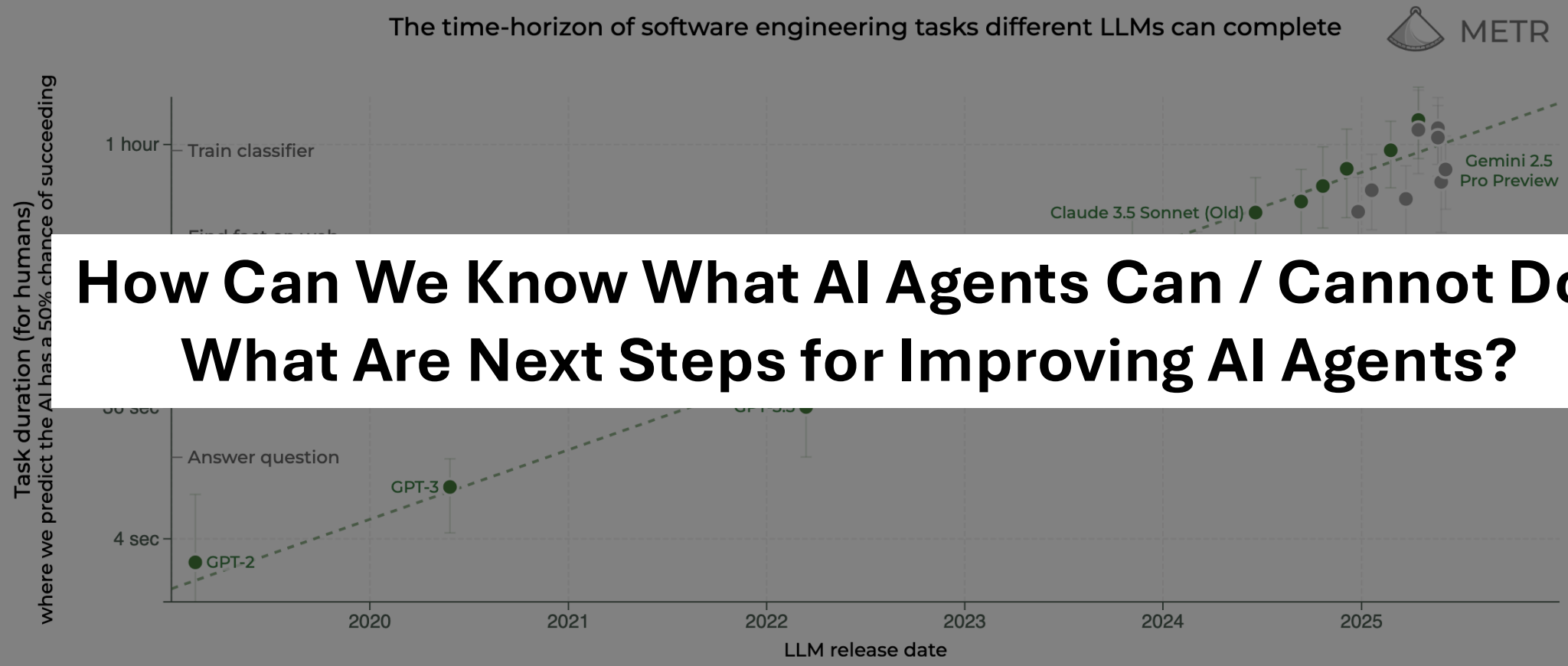
# AI Agents are Becoming Surprisingly Capable






METR, https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/, 19 March 2025

# AI Agents are Becoming Surprisingly Capable




**How Can We Know What AI Agents Can / Cannot Do?**
**What Are Next Steps for Improving AI Agents?**



METR, https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/, 19 March 2025

# David Patterson

*Turing Award Laureate*

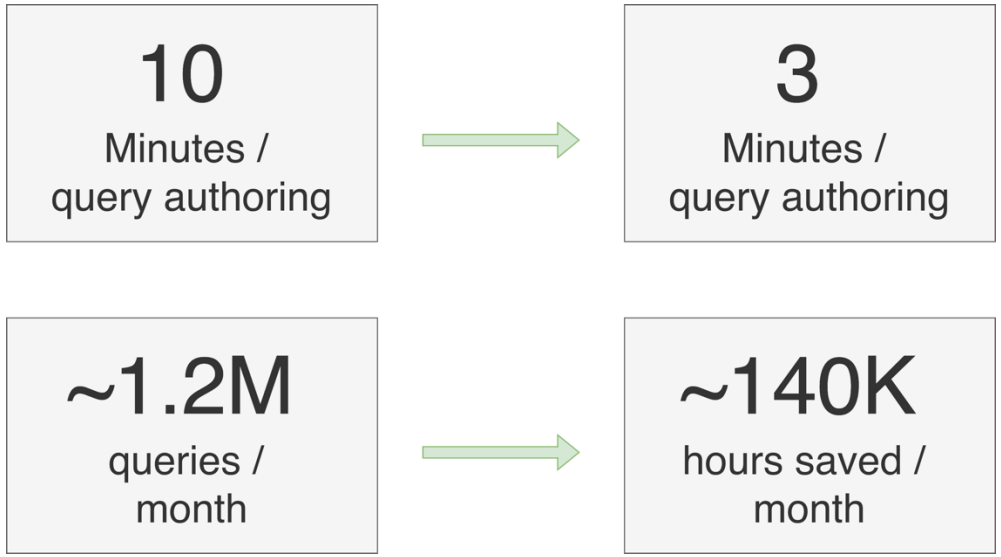
"For better or worse, benchmarks shape a field. … When a field has good benchmarks, we settle debates and the field makes rapid progress."

David Patterson, https://dl.acm.org/doi/pdf/10.1145/2209249.2209271, Communications of ACL 2012.

# Text-to-SQL Is Widely Studied Across Industries


...

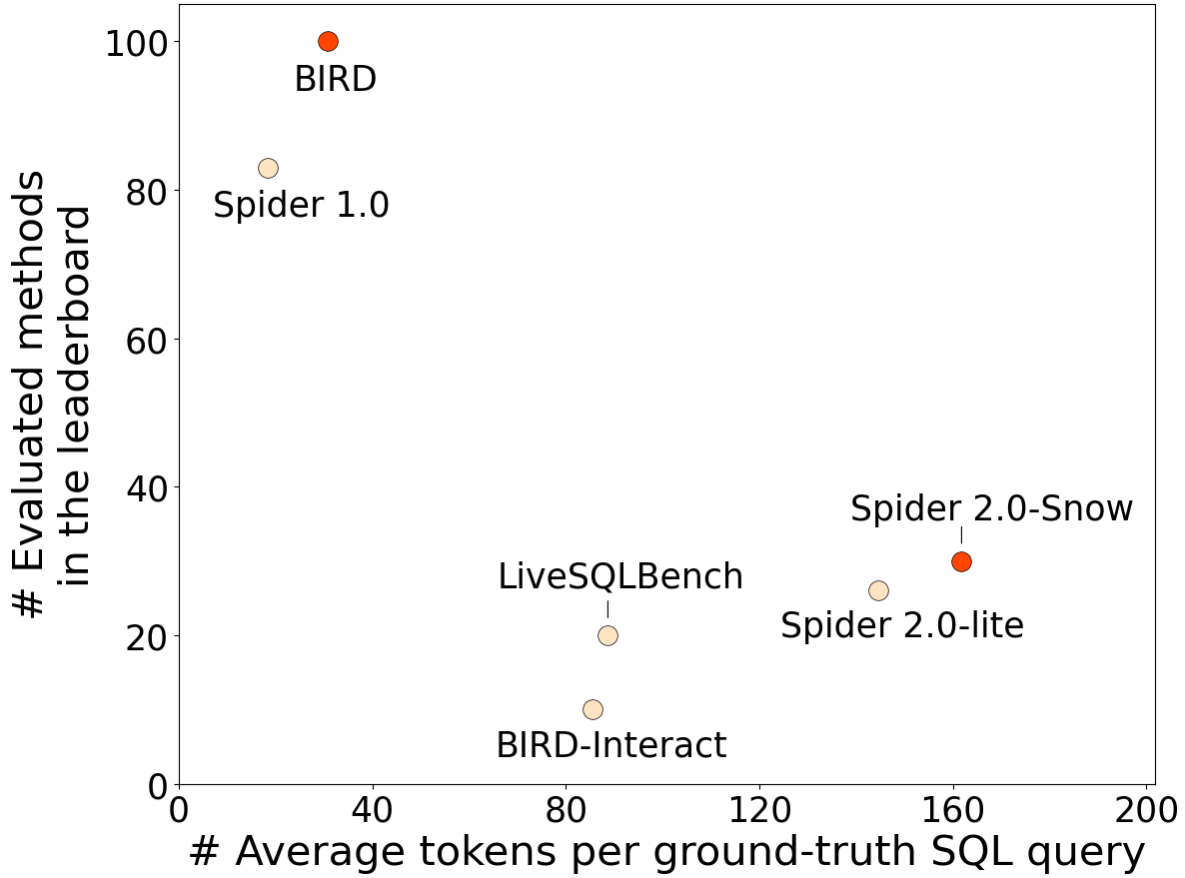
Companies developing text-to-SQL models/agents




The impact of Uber text-to-SQL agent: QueryGPT*

* https://medium.com/wrenai/how-uber-is-saving-140-000-hours-each-month-using-text-to-sql-and-how-you-can-harness-the-same-fb4818ae4ea3

# BIRD and Spider 2.0-Snow Are Widely Used to Measure Text-to-SQL Progress

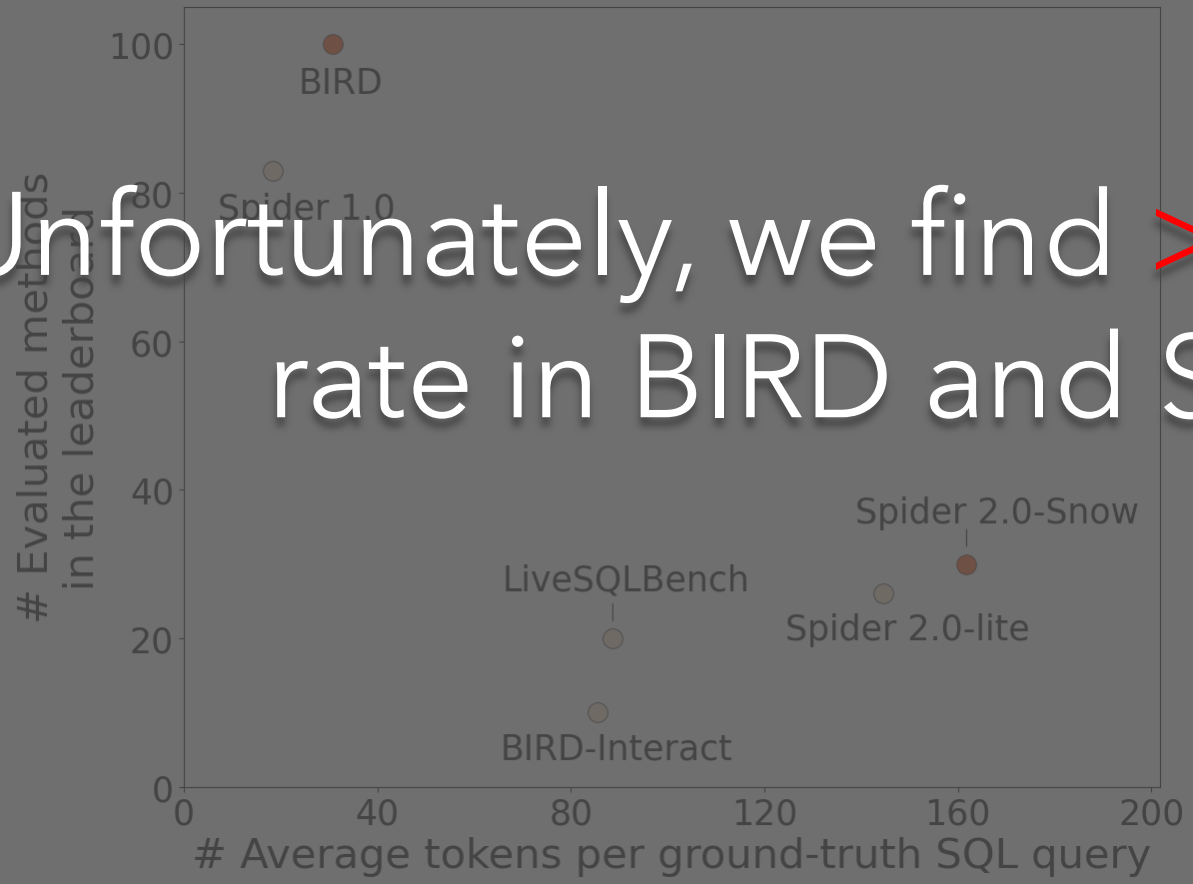



Number of evaluated models and SQL tokens across text-to-SQL benchmarks (data cutoff: Oct. 20, 2025)

**BIRD**: Evaluates the most methods (100)

**Spider 2.0-Snow**: Features complex SQL queries (161.8 tokens/query, on average)

# BIRD and Spider 2.0-Snow Are Widely Used to Measure Text-to-SQL Progress




Number of evaluated models and SQL tokens across text-to-SQL benchmarks (data cutoff: Oct. 20, 2025)

**BIRD**: Evaluates the most methods (100+)

**Spider 2.0-Snow**: Features complex SQL queries (161.8 tokens/query, on average)

Unfortunately, we find >50% annotation error rate in BIRD and Spider 2.0-Snow

# Pervasive Annotation Errors Break Text-to-SQL Benchmarks and Leaderboards

**Annotation error correction**

Corrected 48% of sampled BIRD Dev examples.

→

**Misreported benchmark execution accuracy**

Execution accuracy changes: up to 19%

→

**Misleading leaderboard rankings**

Ranking changes: up to 9 positions (16 agents)

# Recurring Error Patterns in Text-to-SQL Benchmarks

**User query:** Can you provide a daily weather summary for July 2019 within a 5 km radius of latitude 26.75 and longitude 51.5? ...

**Annotated SQL**

```
...  ST_DWITHIN(ST_GEOGFROMWKB("TRI"."geography"),
ST_POINT(26.75, 51.5), 5000) ...
```

**Issue:** Snowflake's ST_POINT expects (longitude, latitude). The query passes ST_POINT(26.75, 51.5), which inverts the order.

sf_bq291 of Spider 2.0-Snow. Annotators mistakenly swapped the longitude–latitude order in the ST_POINT arguments

**E1** Mismatches between the semantics of the SQL query and the logic of the natural language input

# Recurring Error Patterns in Text-to-SQL Benchmarks

**User query:** What is the highest amount of budget spend for an event?

**Annotated external knowledge:** Highest amount of budget spend refers to MAX(spent)

**Annotated SQL**

```
SELECT MAX(spent) FROM budget
```

**Issue:** Each event can have multiple rows in the 'budget' table. The query does not aggregate 'spent' by 'link_to_event' and then take the maximum across those event totals.

Problem 1378 of BIRD Mini-Dev. Annotators ignore the fact that one event can have multiple rows in the 'budget' table

**E2** Mismatches between the semantics of the SQL query and the corresponding database

# Recurring Error Patterns in Text-to-SQL Benchmarks

**User query:** Which state special schools have the highest number of enrollees from grades 1 through 12?

**Annotated external knowledge:** State Special Schools refers to DOC = 31; Grades 1 through 12 means K-12

**Annotated SQL**

```
SELECT T2.School
FROM frpm AS T1 INNER JOIN schools AS T2
ON T1.CDSCode = T2.CDSCode WHERE T2.DOC = 31
ORDER BY T1."Enrollment␣(K-12)" DESC LIMIT 1
```

**Issue:** 'Enrollment (K-12)' includes kindergarten and is therefore not equivalent to enrollment for grades 1 through 12.

Problem 46 of BIRD Mini-Dev. Annotators have limited understanding of "K-12"

**E3** Mismatches between the semantics of the ground-truth SQL query and the related domain knowledge

# Recurring Error Patterns in Text-to-SQL Benchmarks

**User query:** What are the five longest types of highways within the multipolygon boundary of Denmark (as defined by Wikidata ID 'Q35') by total length, analyzed through planet features?

**Annotated SQL**

```
...
SELECT REPLACE(highway_type, '"', '') AS highway_type
FROM highway_info ...
```

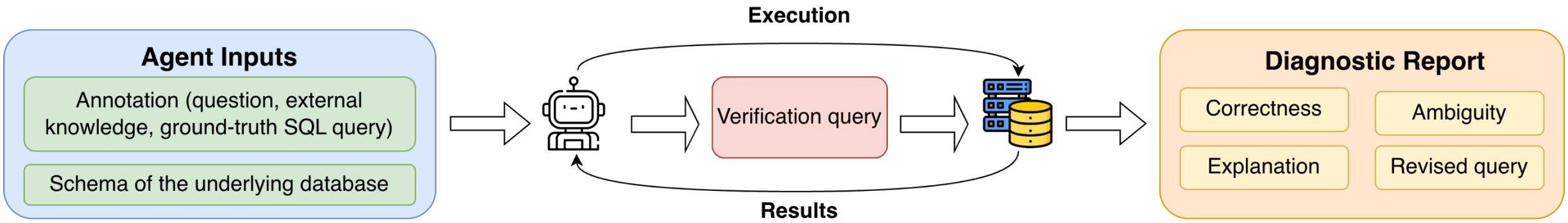
**Issue:** The annotated query includes an additional step: replacing double quotation marks, which is not specified in the question.

sf_bq017 of Spider 2.0-Snow. The annotated query requires a specific output format, but the question doesn't specify it.

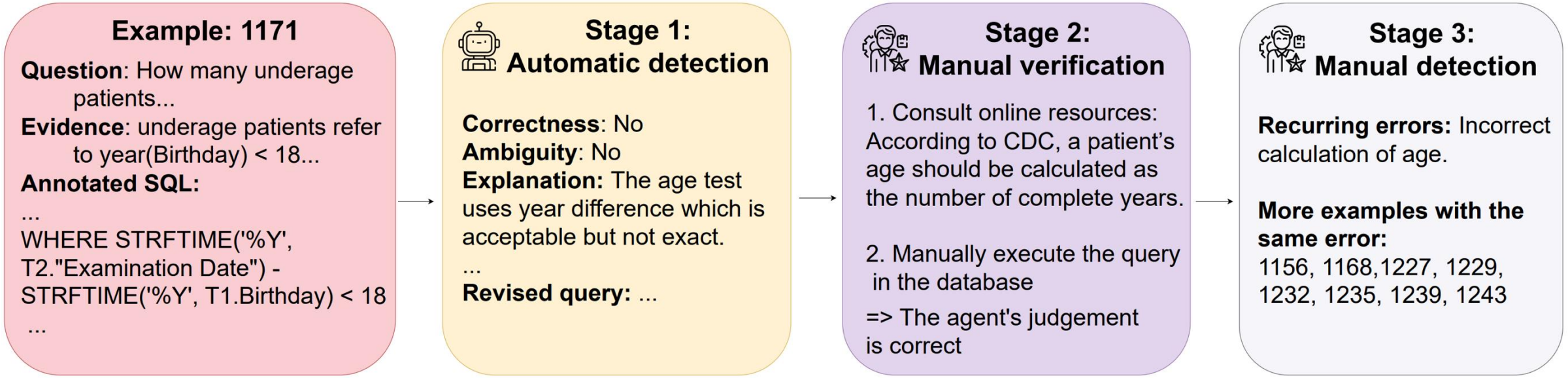
**E4** Ambiguity in the natural language input

Can we improve the current annotation pipeline to reduce annotation errors?

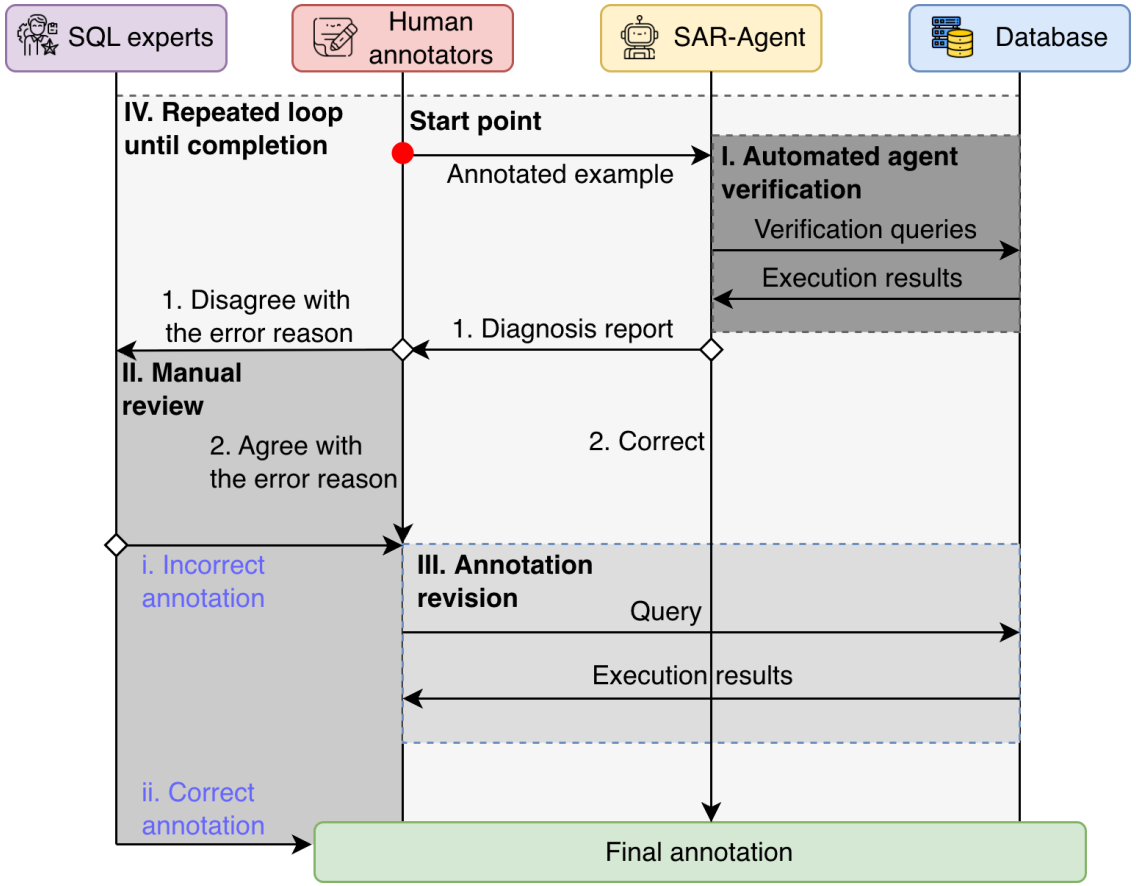
# SAR-Agent - SQL Annotation Reviewer Agent



# Error Detection: Human-In-The-Loop, Three-Stage examination



# Error Correction: SAPAR




The architecture of SAPAR

1. Automated agent verification
2. Manual review
3. Annotation revision
4. Loop until completion

   Judged correct by the AI agent or SQL experts.

# Overview of Our Experiments

- Annotation error rates in text-to-SQL benchmarks
- The impact of annotation errors on the evaluation
- The effectiveness of SAR-Agent in detecting annotation errors

# Experimental settings

**Data:** BIRD Mini-Dev (498 examples) and Spider 2.0-Snow (121 examples)

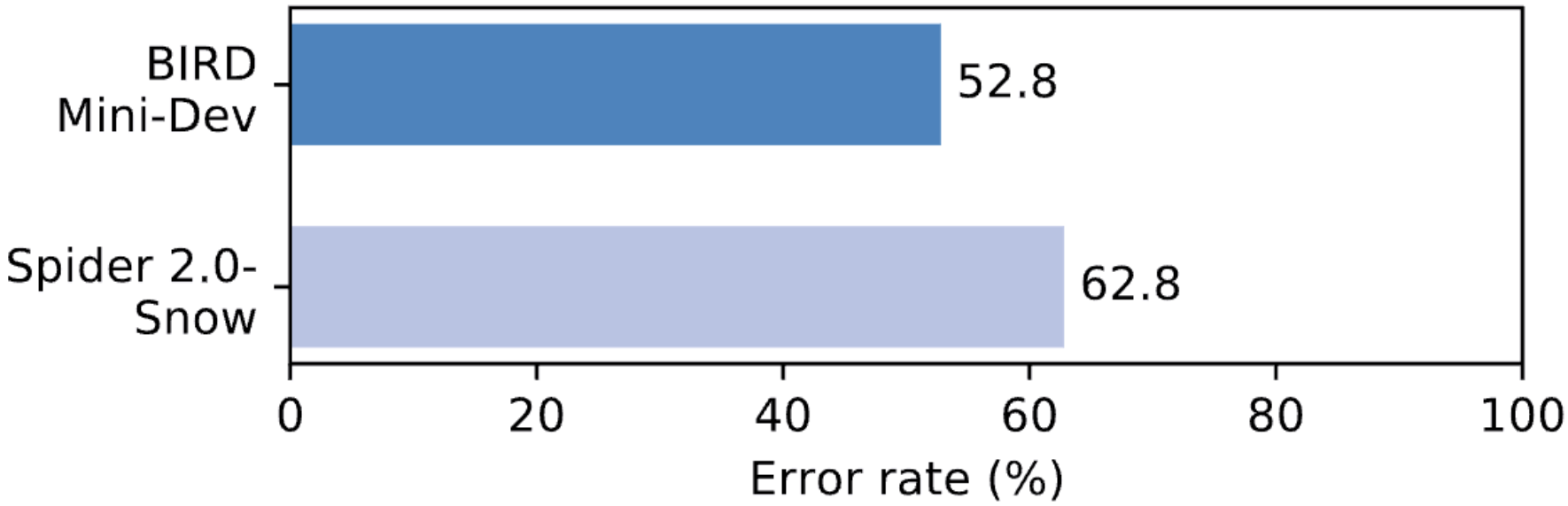
**Metrics:** Error rate

**SAR-Agent settings:**

Model: o3

Maximum iterations: 30

# BIRD and Spider 2.0 Have Pervasive Annotation Errors




BIRD Mini-Dev and Spider 2.0-Snow have error rates of 52.8% and 62.8%, respectively.

# Overview of Our Experiments

- Annotation error rates in text-to-SQL benchmarks
- The impact of annotation errors on the evaluation
- The effectiveness of SAR-Agent in detecting annotation errors

# Experimental settings

**Data:** 100 examples sampled from the BIRD Dev set,
manually corrected 48 examples based on SAPAR

**Metrics:** Execution accuracy (EX), Rankings, Spearman's rank correlation coefficient

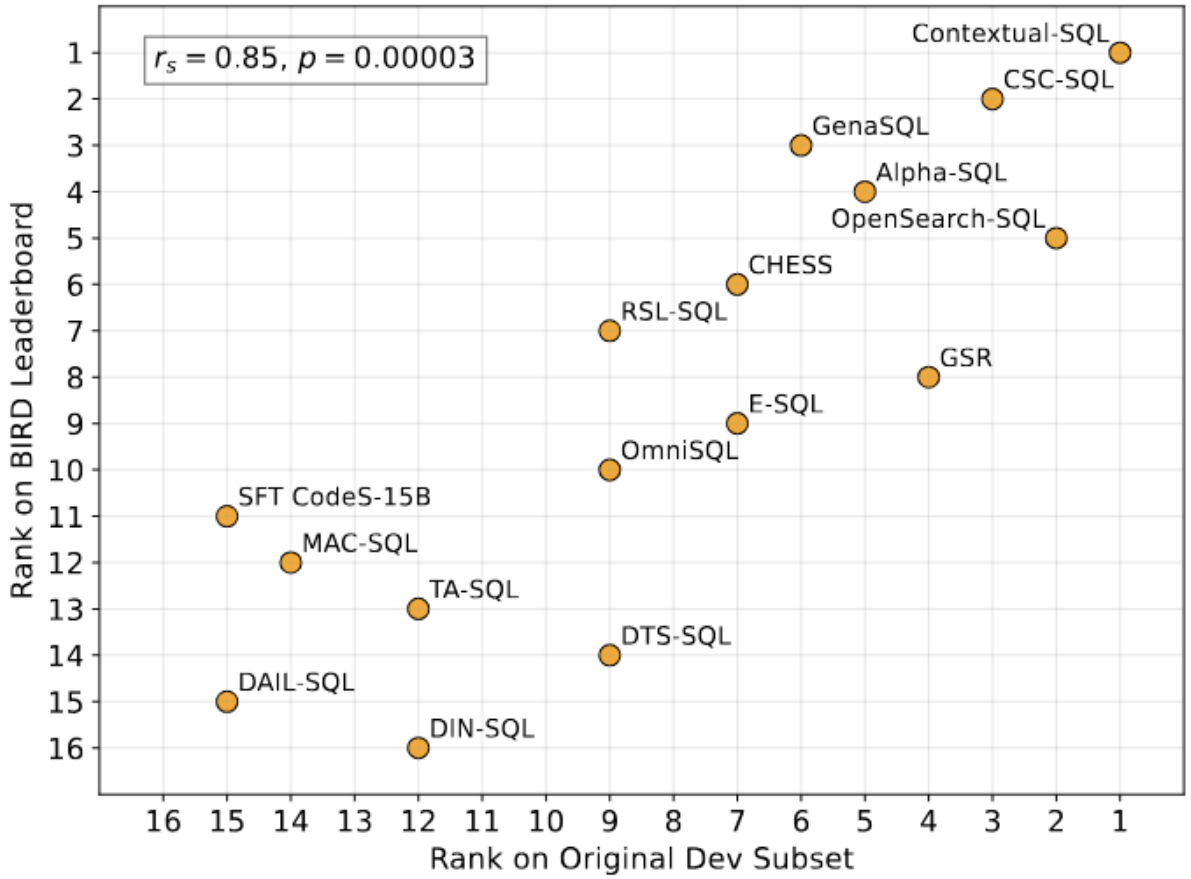
**Text-to-SQL Agents:** All 16 open-source agents from the BIRD leaderboard

# Statistics of Manually Corrected Dev Subset

| Statistics | # Examples (total=100) |
|---|---|
| **Difficulty of selected examples** | |
| Easy | 62 |
| Moderate | 28 |
| Challenging | 10 |
| **Correction type of Dev subset** | |
| Corrected examples | 48 |
| Question correction | 19 |
| External knowledge correction | 17 |
| SQL correction | 41 |
| Schema correction | 1 |
| Database modification | 6 |

We corrected a total of 48 text-to-SQL examples. Each example may involve more than one types of corrections.

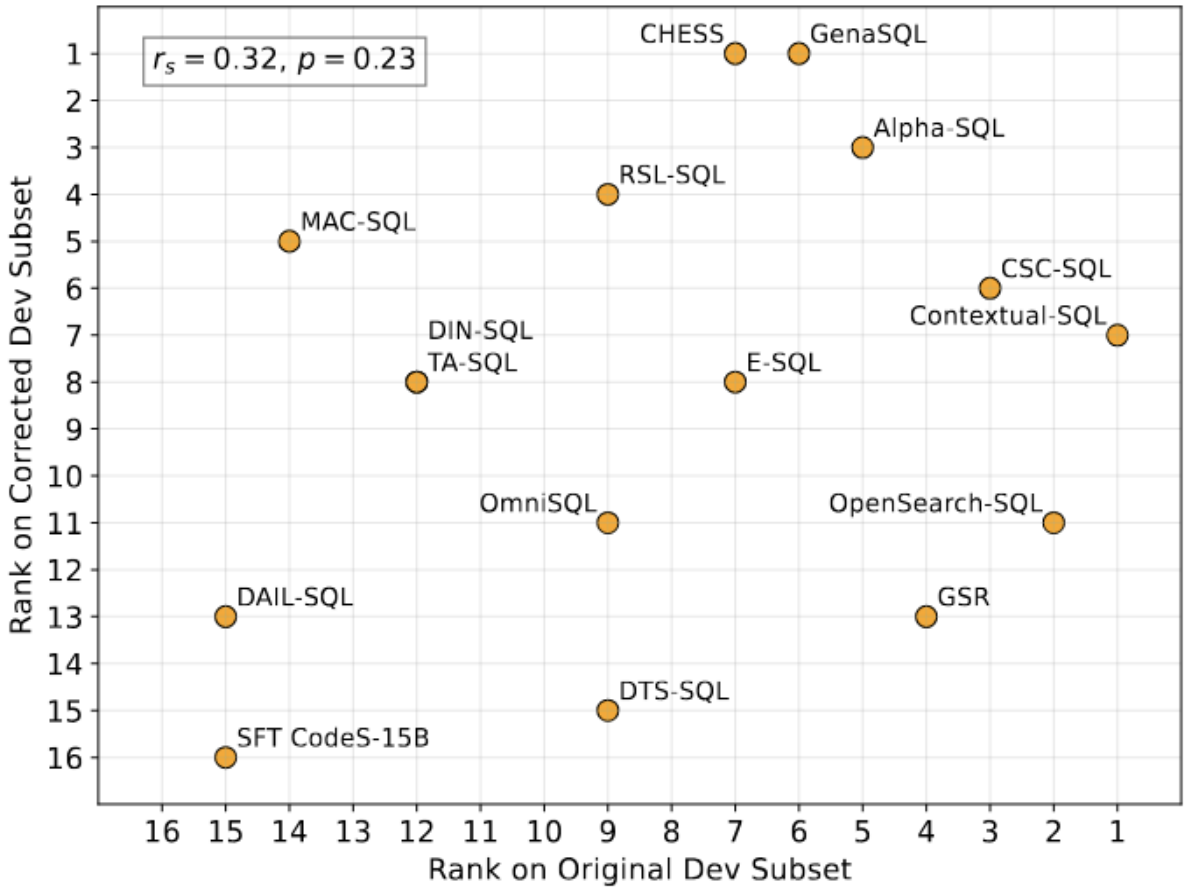
# Ranking Correlation Analysis




Correlation between EX rankings on the original Dev subset and the full BIRD Dev set

The rankings of text-to-SQL agents on the original subset correlate strongly with those on the full Dev set (Spearman's $r_s$=0.85, $p$=3e-5).

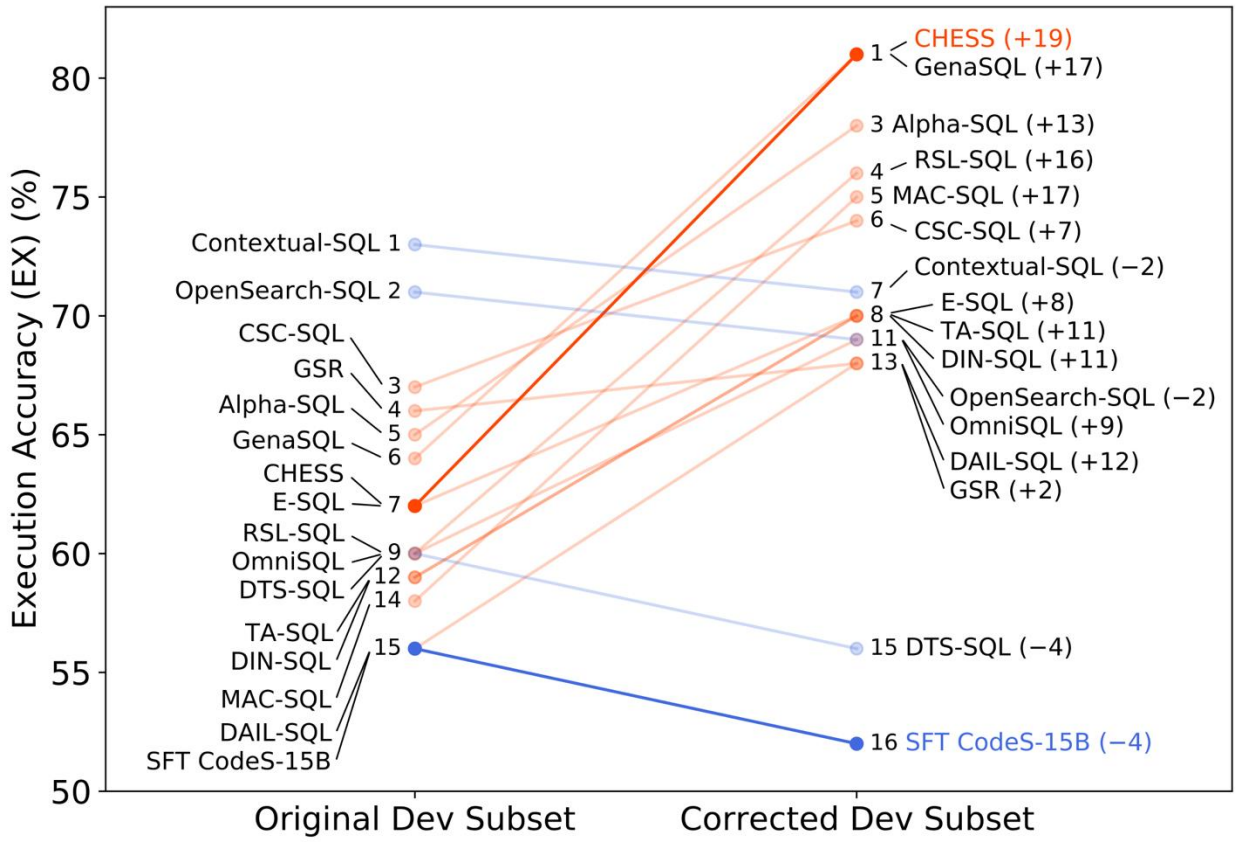
# Ranking Correlation Analysis




Correlation between EX rankings on the original and corrected Dev subset

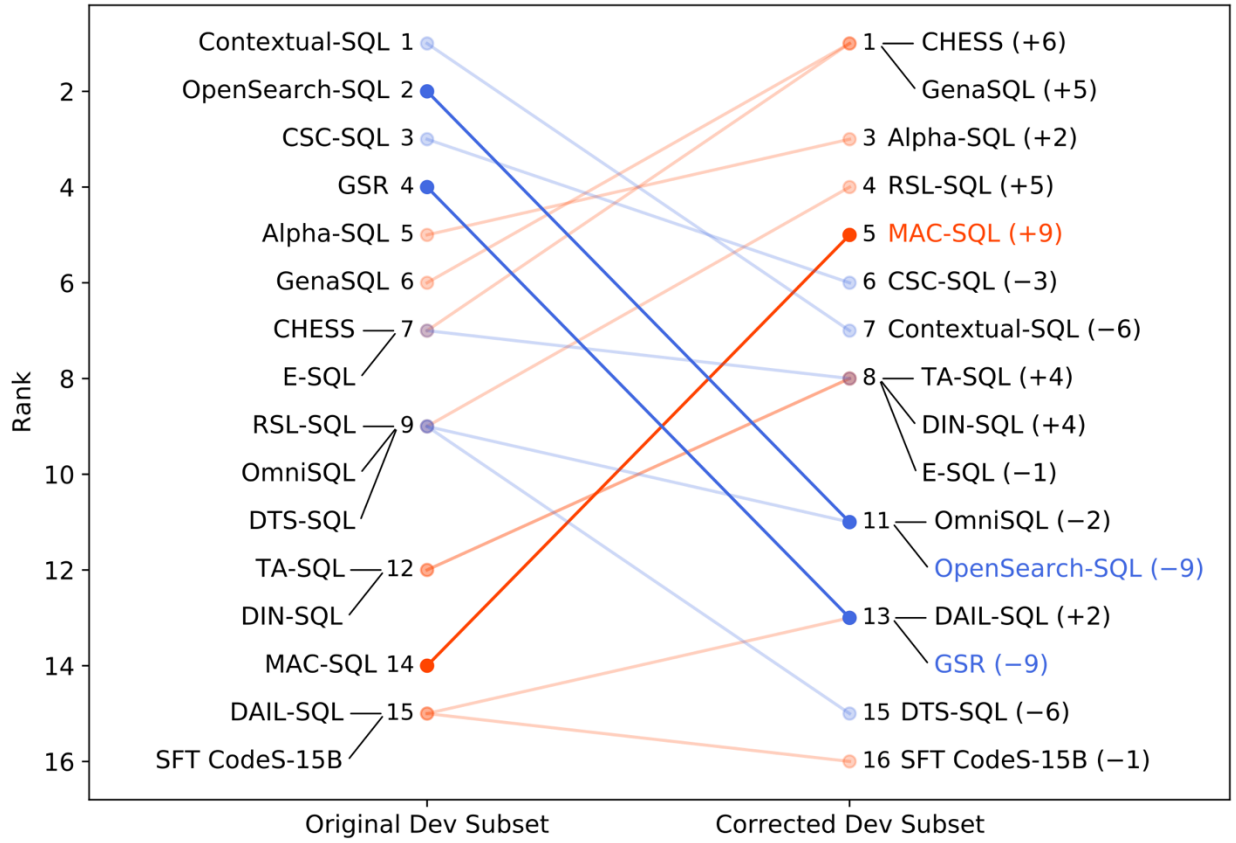
The rankings of text-to-SQL agents on the original subset correlate weakly with those on the corrected subset (Spearman's $r_s$=0.32, $p$=0.23).

# Change of Execution Accuracy




Execution accuracy of agents on original and corrected BIRD Dev subsets. Execution accuracy changes range from -4% to 19% , with an average change of 9.6%.

# Change of Agent Ranking




Agent ranking changes from original to corrected BIRD Dev subsets. Ranking shifts range from -9 to +9 positions, with an average change of 5 positions.

# A Recurring Annotation Error Type

User question: For the patient with an abnormal alkaliphophatase level, how many of them are admitted to the hospital?

```
SELECT COUNT(T1.ID)
FROM Patient AS T1
INNER JOIN Laboratory AS T2
ON T1.ID = T2.ID
WHERE T2.ALP >= 300
AND T1.Admission = '+'
```

An incorrectly annotated query by the BIRD team.

```
SELECT COUNT(DISTINCT T1.ID)
FROM Patient AS T1
INNER JOIN Laboratory AS T2
ON T1.ID = T2.ID
WHERE T2.ALP >= 300
AND T1.Admission = '+'
```

The query after correction.

Original vs. corrected queries for Question 1286 in BIRD Dev; the original omits *DISTINCT* in *COUNT*.
Questions 310, 416, 605, and 1302 also have a similar issue.

Among corrected-query examples, CHESS is misjudged in 41.7% (5/12) due to the missing DISTINCT in COUNT.

# Overview of Our Experiments

- Annotation error rates in text-to-SQL benchmarks
- The impact of annotation errors on the evaluation
- The effectiveness of SAR-Agent in detecting annotation errors

# Experimental settings

**Data:** BIRD Mini-Dev (498 examples) and Spider 2.0-Snow (121 examples)

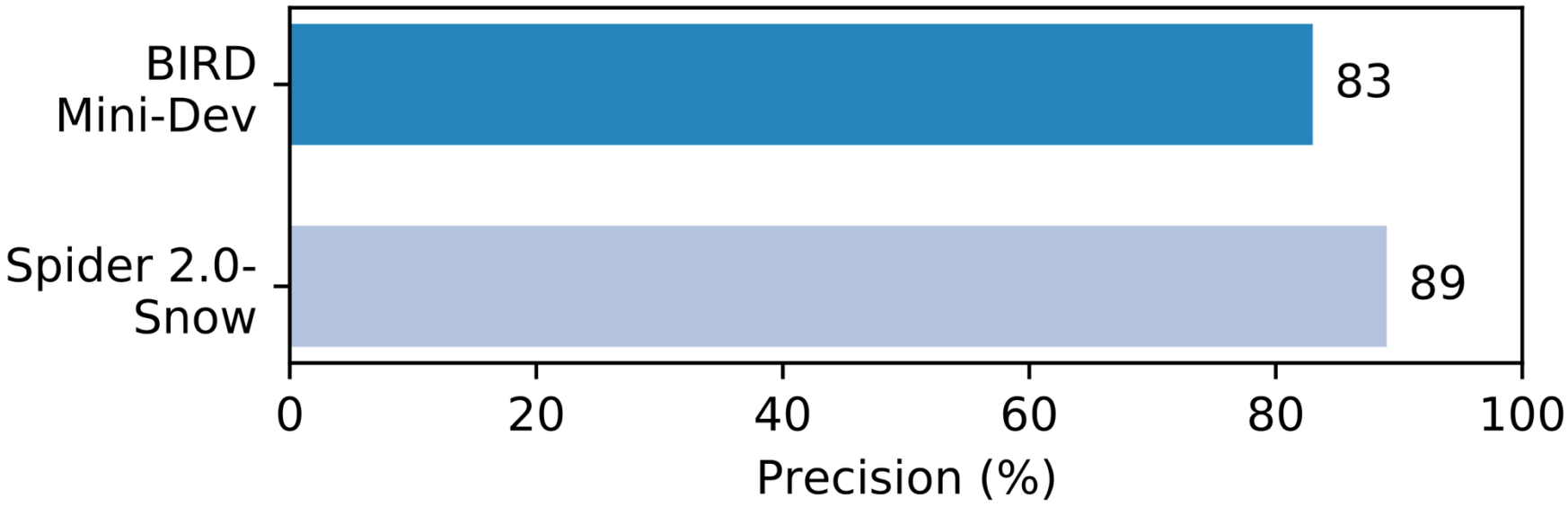
**Metrics:** Precision (%), average cost ($)

**SAR-Agent settings:**

Model: o3

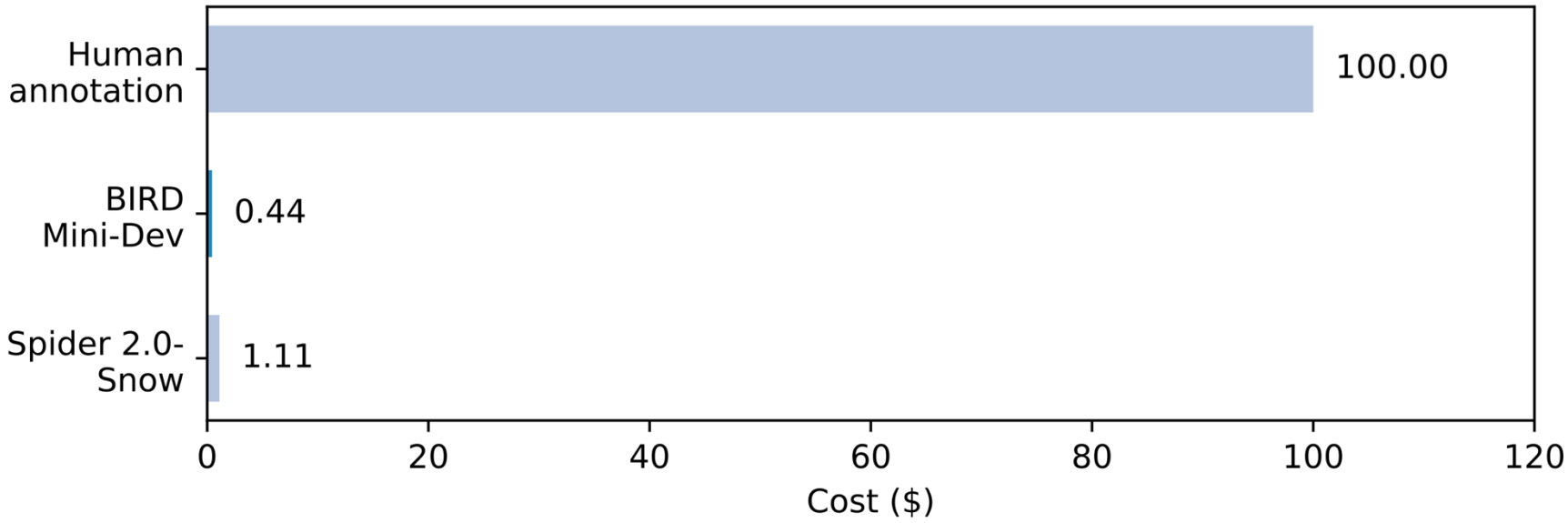
Maximum iterations: 30

# The Precision of SAR-Agent




SAR-Agent achieves 83% precision on BIRD Mini-Dev and 89% on Spider 2.0-Snow.

# The Cost of SAR-Agent




Human annotation costs $100 per example.*
A SAR-Agent report costs less than 1.2% of that per example.

* https://ddkang.substack.com/p/human-data-is-probably-more-expensive

# Experimental settings

**Data:** BIRD Mini-Dev (498 examples)

**Metrics:** Hit rate (%)

$$\frac{\#\ (SAR - Agent\ detections\ \cap\ Arcwise\ detections)}{\#\ Arcwise\ detections} * 100\%$$

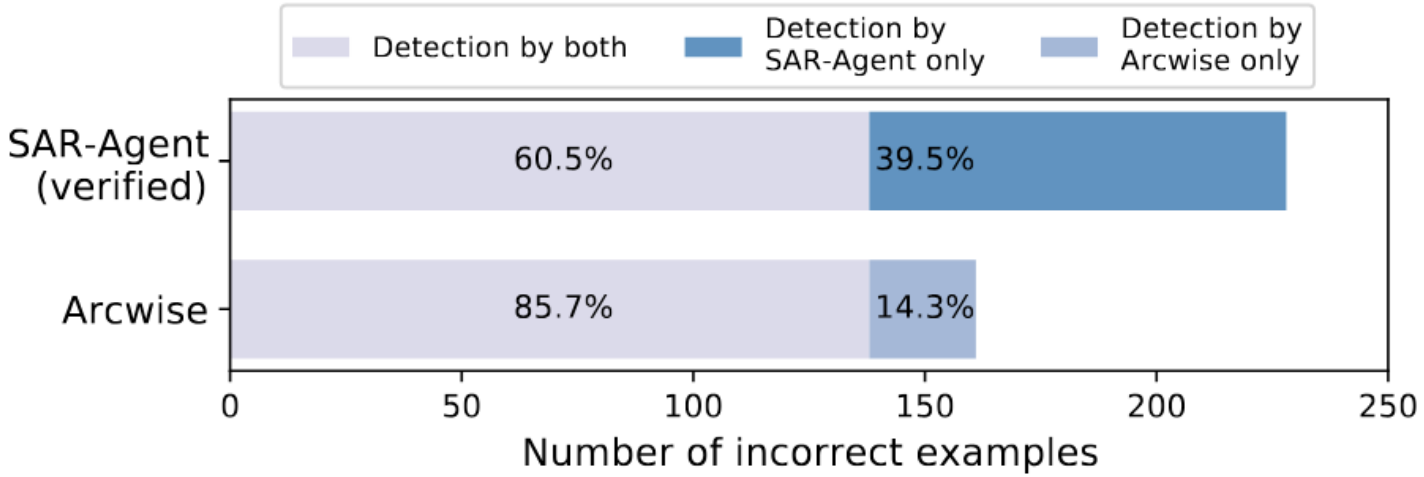
**SAR-Agent settings:**

Model: o3

Maximum iterations: 30

**Baseline:** SQL experts from the Arcwise team

# SAR-Agent Excels at Detecting Annotation Errors




Number of detected incorrect examples in BIRD Mini-Dev by SAR-Agent (after manual verification) and the Arcwise team

1. Of the 161 noisy examples identified by Arcwise, SAR-Agent achieves a hit rate of 85.7%.

2. SAR-Agent identifies 41.6% more incorrect examples than Arcwise.

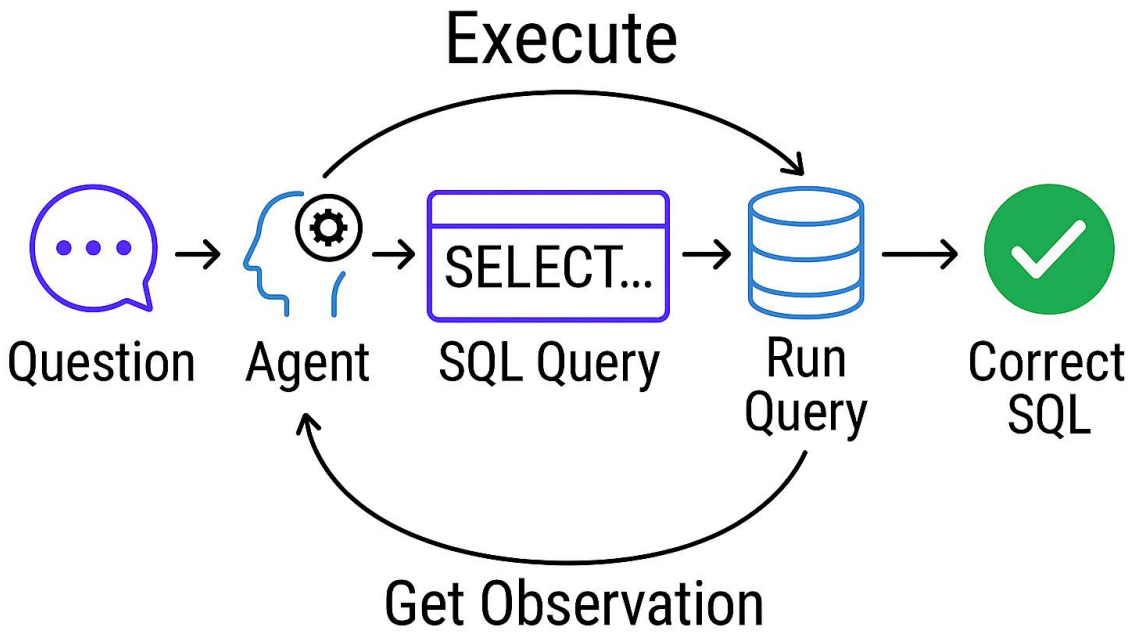
# What happens when you train with noise?

# Training with Noisy Annotations of BIRD

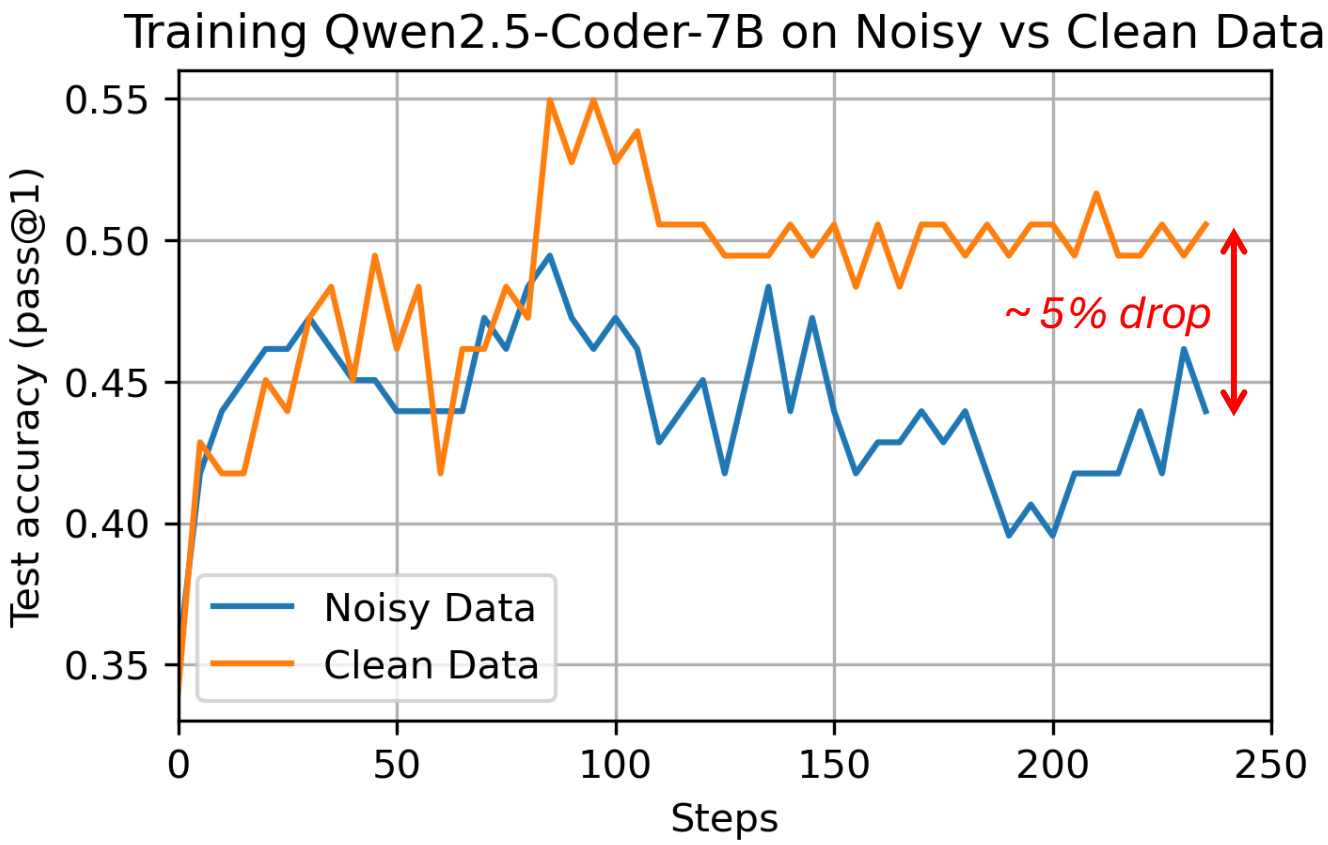
- **Model**: Qwen2.5-Coder-7B-Instruct
- **Train data**: 164 examples from BIRD mini Dev
  - *Noisy* (*original*): 53.3% SQL annotations are incorrect
  - *Clean*: all incorrect annotations are corrected and verified
- **Test data**: 91 examples from BIRD mini Dev

# Training Algorithm (SkyRL-SQL)

- Outcome supervision GRPO
- 5 execute-observe turns per rollout
- 200 steps or convergence



# Learning on Incorrect Annotations = Misleading Models by Human Mistakes




**Train**

**Q**: Among players …, how many of them …?
**Incorrect SQL annotation**:

```
SELECT COUNT(DISTINCT player_api_id)
FROM … WHERE …
```

**Test**

**Q**: Among customers …, how many of them?
**SQL generation by the trained model**:

```
SELECT COUNT(DISTINCT T1.CustomerID)
FROM … INNER JOIN … ON … WHERE …
```

`DISTINCT` is incorrectly omitted in the human annotation / by the trained model

# Conclusion

- Introducing SAPAR and SAR-Agent to assist in annotation review
- Error rates: Bird Mini-Dev (52.8%), Spider 2.0-Snow (62.8%)
- Reevaluation of all 16 open-source text-to-SQL agents on the BIRD Dev subset
  - Relative performance: -7% to 31%
  - Ranking: -9 to +9

code

Daniel Kang
E-mail: ddkang@g.Illinois.edu
Twitter: @daniel_d_kang