The Ohio State University

Dr.Spider: A Diagnostic Evaluation Benchmark towards Text-to-SQL Robustness

Shuaichen Chang[1], Jun Wang[2], Mingwen Dong[2], Lin Pan[2], Henghui Zhu[2], Alexander Hanbo Li[2], Wuwei Lan[2], Sheng Zhang[2], Jiarong Jiang[2], Joseph Lilien[2], Steve Ash[2], William Wang[2], Zhiguo Wang[2], Vittorio Castelli[2], Patrick Ng[2], Bing Xiang[2]

[1]Ohio State University, [2] AWS AI Labs

ICRL 2023

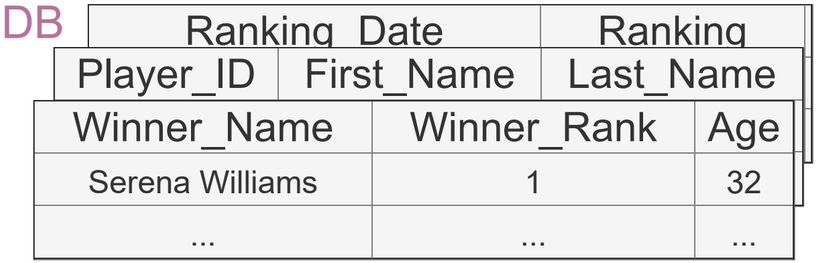
Selected as a notable-top-5% paper

Text-to-SQL

NLQ

Find the name and rank of the 3 youngest winners across all matches.

DB



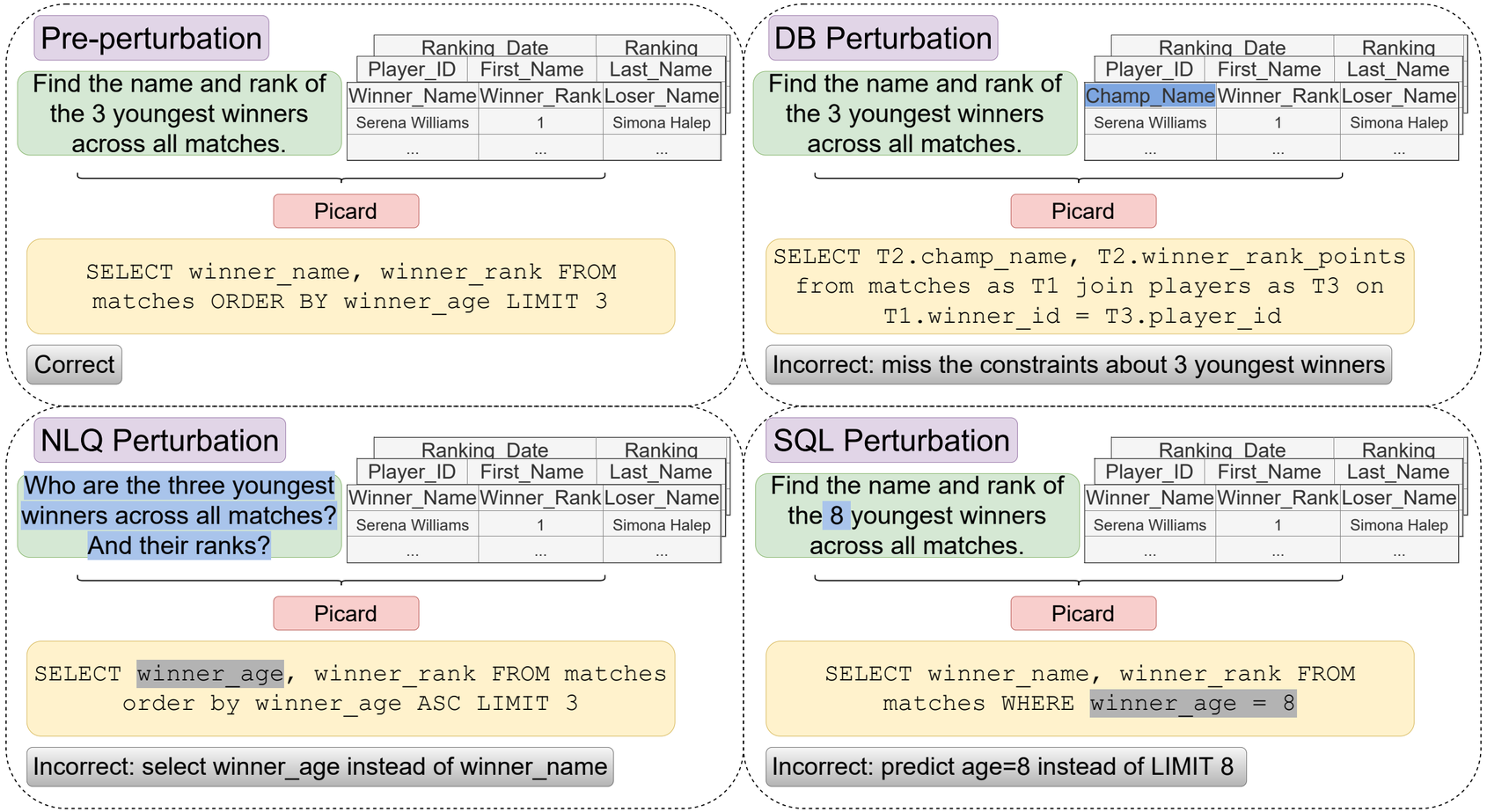
Text-to-SQL Model

SQL

```
SELECT winner_name, winner_rank FROM
matches ORDER BY winner_age ASC LIMIT 3
```

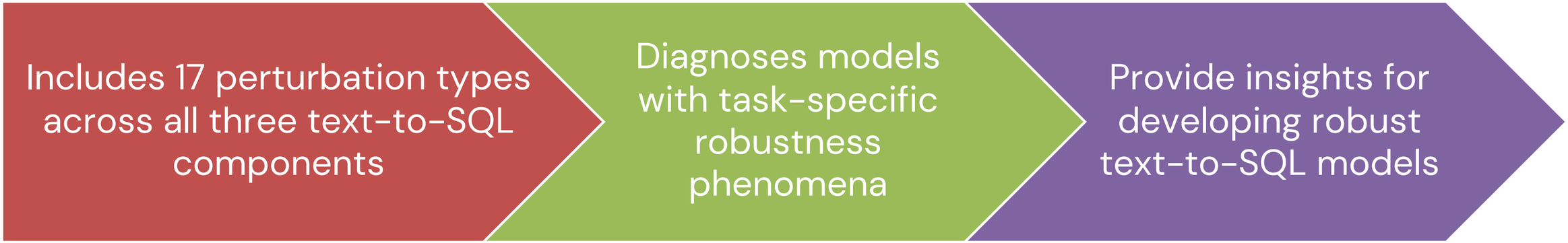
Does the high accuracy imply a robust model?

We perturb the data in Spider[1] to uncover the weaknesses in models that might not be evident when evaluated on the original data.

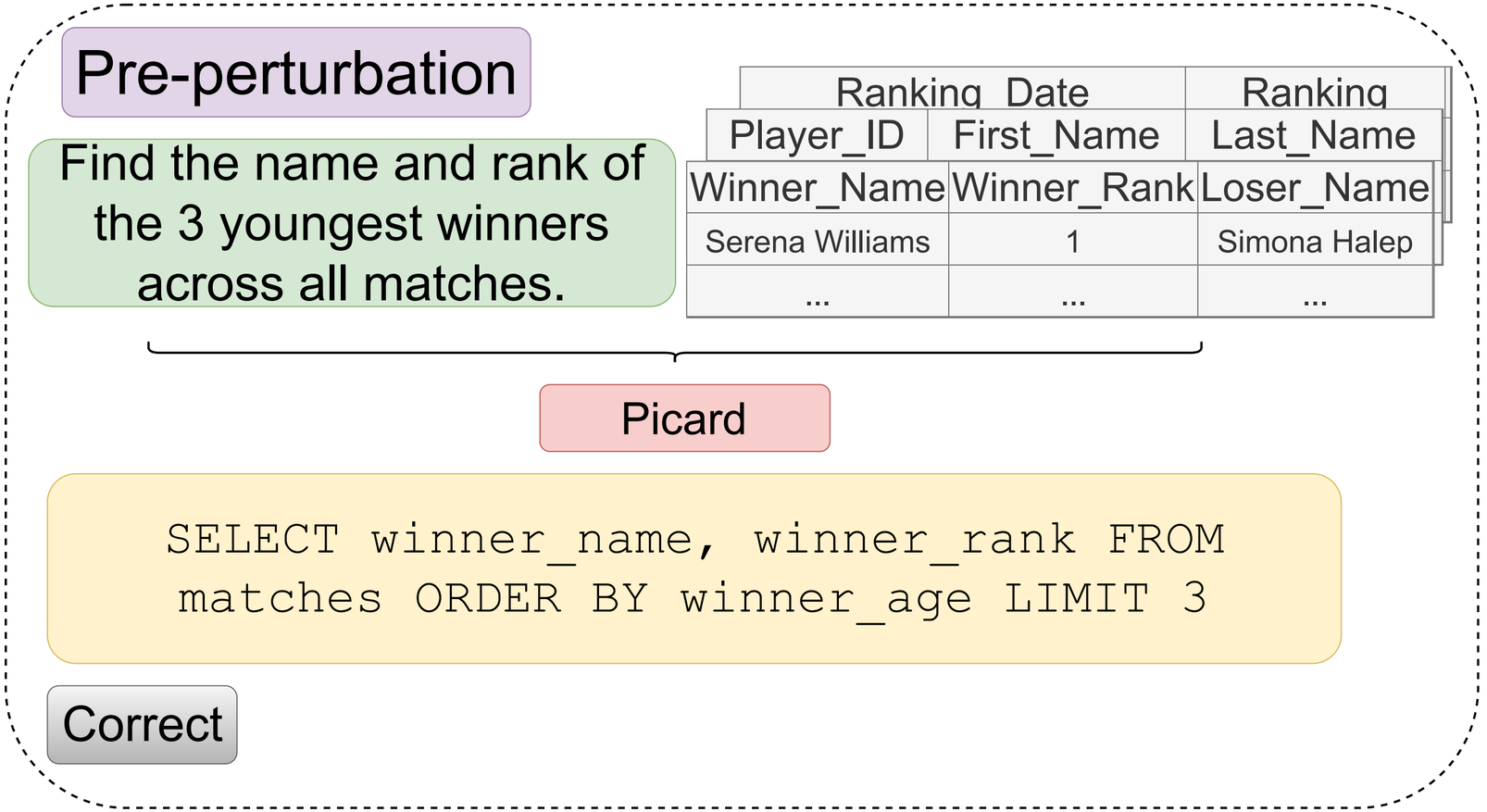


[1] Yu, Tao, et al. "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task." *EMNLP*. 2018.

Our Evaluation Benchmark

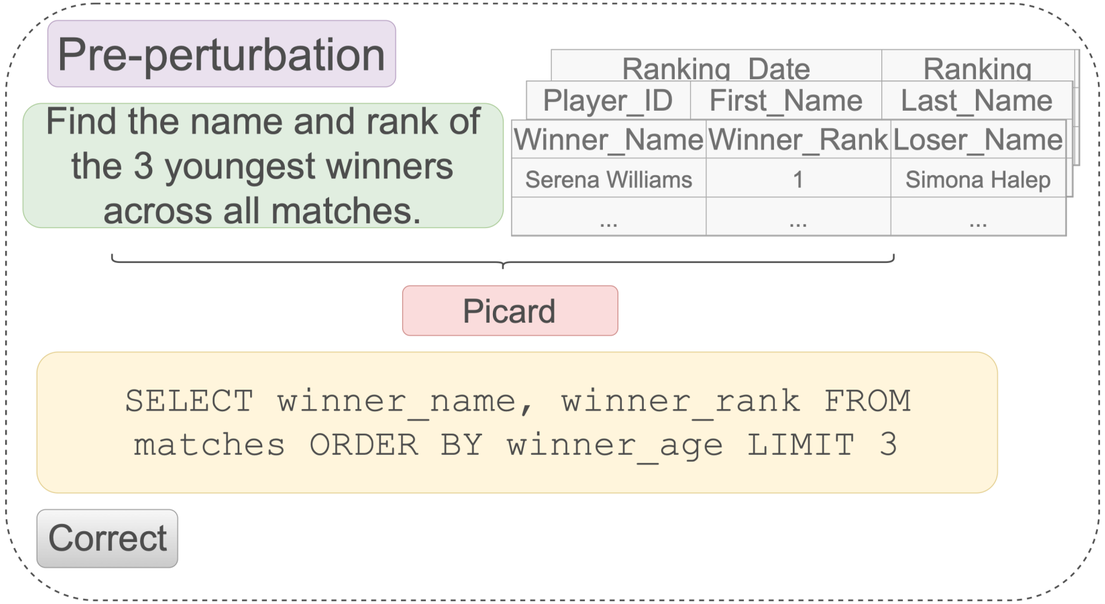


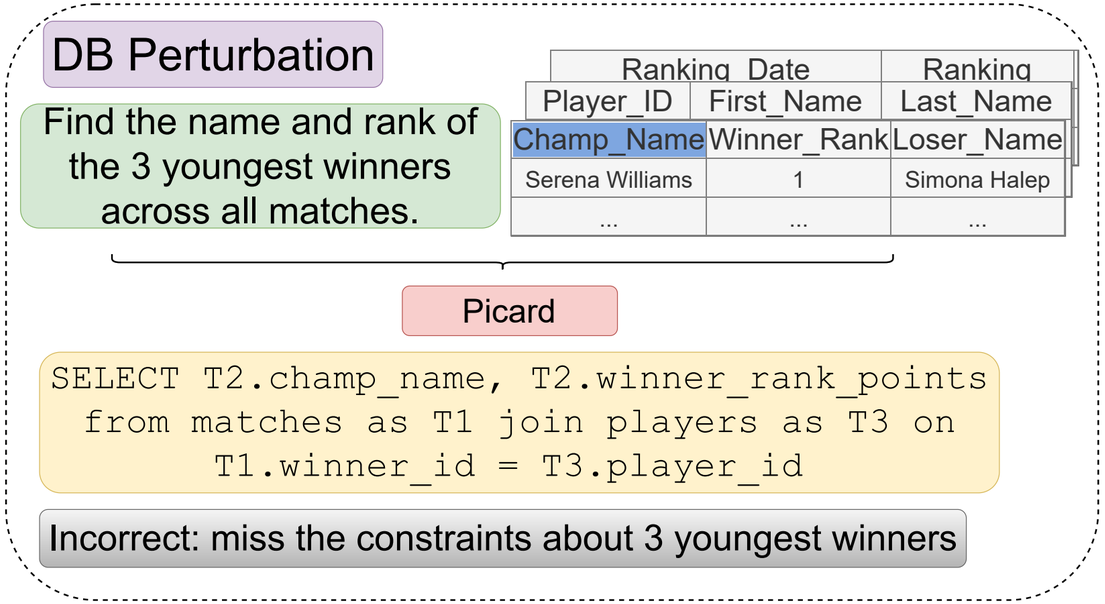
Text-to-SQL Example



Database (DB) Perturbations

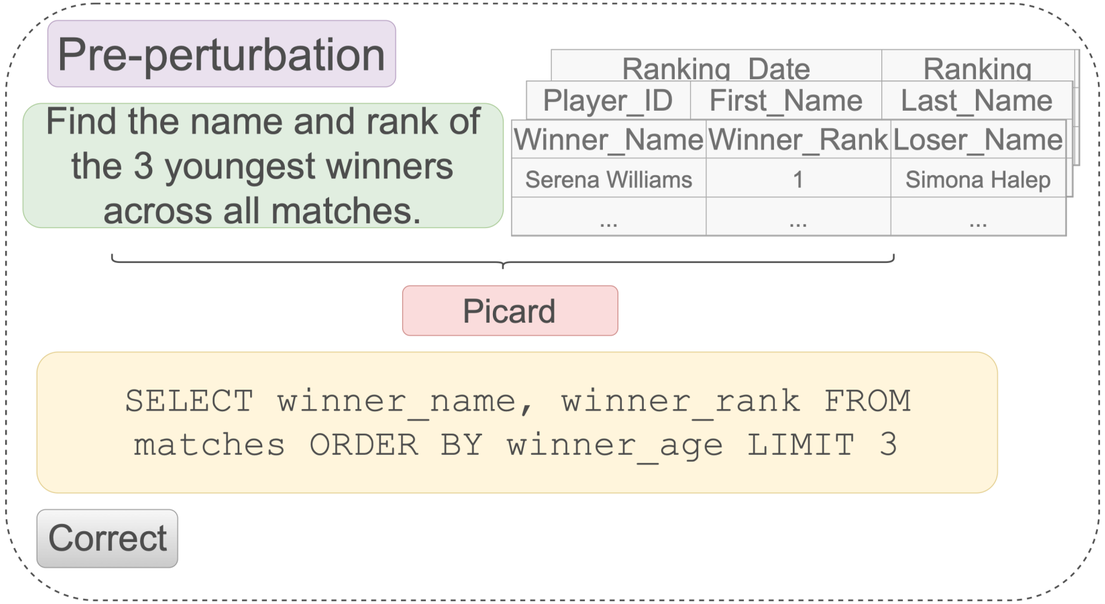
- Data in a database can be represented in various ways.

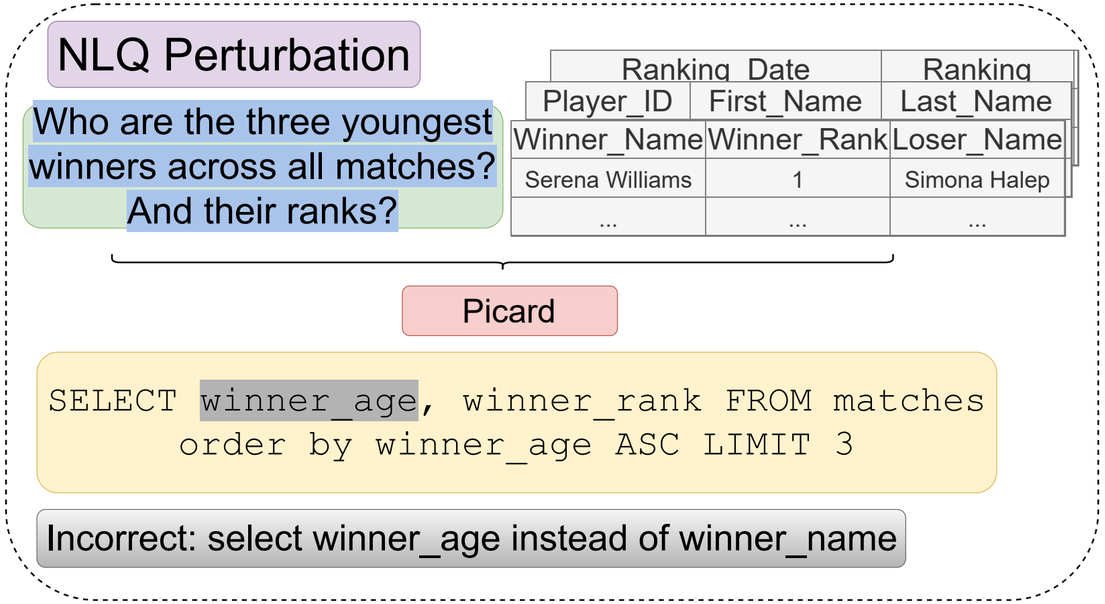




Natural Language Question (NLQ) Perturbations

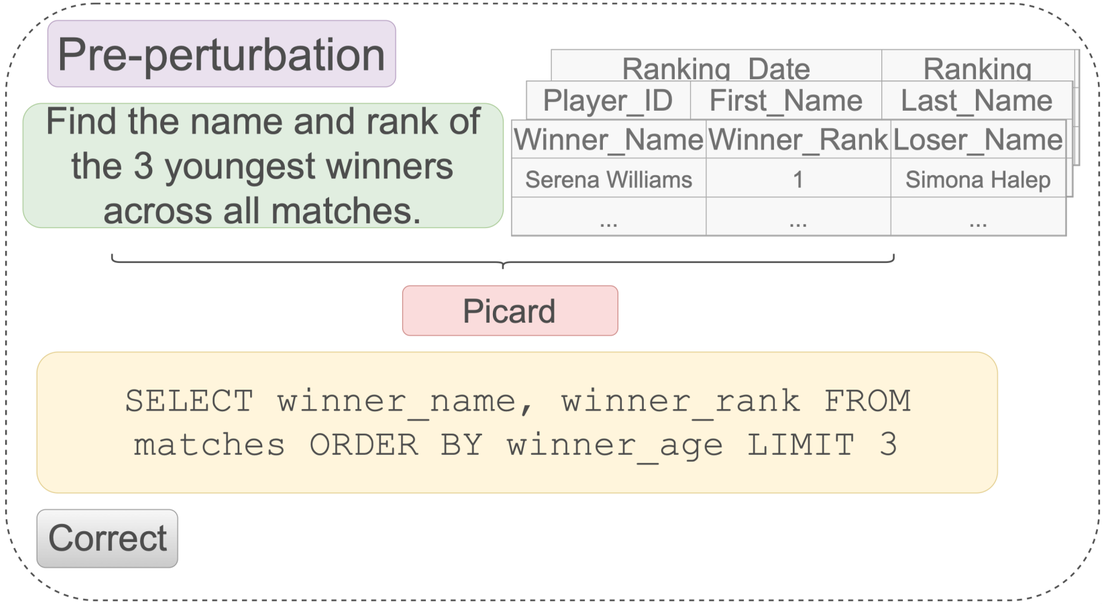
- The same query intent can be expressed with different phrasings.

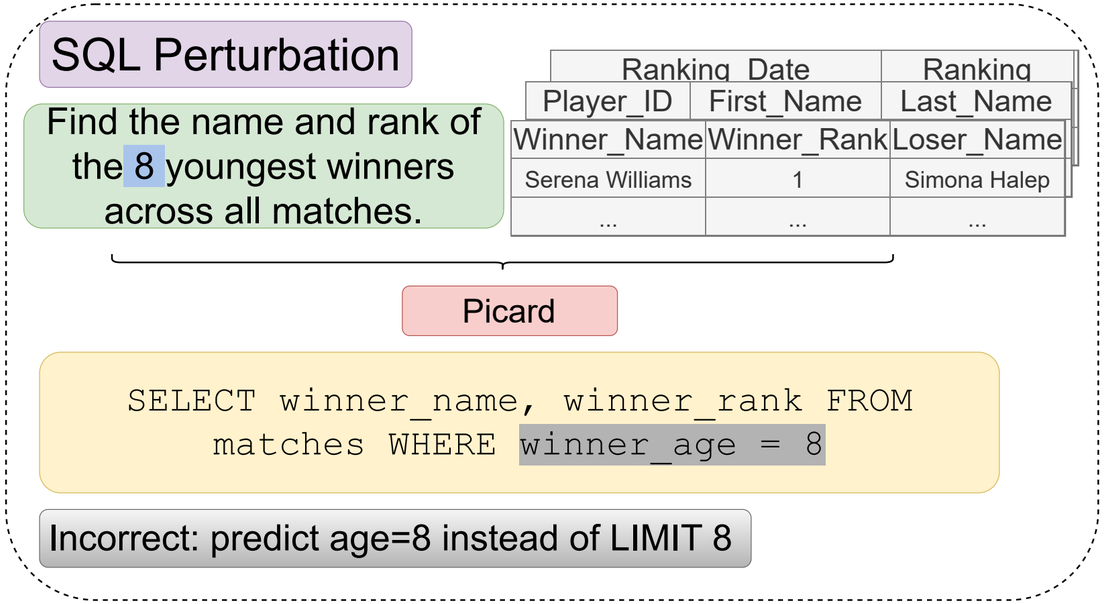




SQL Perturbations

- Modifying logical and symbolic units in a question should not impact understanding of other components.





Perturbation Data Creation

- For DBs and SQL queries, we perturb them programmatically, using their inherent structure.
- For NLQ perturbations, we propose to use LLMs to simulate human paraphrasing with various task-specific phenomena that we find in crowdsourcing annotations.

Paraphrase Category	Example
`Keyword-synonym` (Replace SQL keyword indicators with synonyms)	Question: What is the code of airport that has the highest number of flights? Paraphrases: Show me the code for the airport that currently has the most flights.
`Keyword-carrier` (Imply SQL keyword indicators by carrier phrases)	Question: Show the name and theme for all concerts and the number of singers in each concert. Paraphrase: List the names and themes for all concerts and how many singers are in each.
`Column-synonym` (Replace column indicators with synonyms)	Question: List the name of teachers whose hometown is not Little Lever Urban District. Paraphrases: Find the name of teachers who were not born in Little Lever Urban District.
`Column-carrier` (Imply column indicators by carrier phrases)	Question: Show the name of teachers aged either 32 or 33? Paraphrases: Which teachers are aged either 32 or 33.
`Column-attribute` (Imply column indicators by aggregated attributes)	Question: What is the name of the conductor who has worked the greatest number of year? Paraphrases: Who has worked the longest as conductor?
`Column-value` (Imply column indicators by values)	Question: What are the ids of the students who do not own cats as pets? Paraphrases: Find the IDs of students who don't own cats.
`Value-synonym` (Replace value indicators with synonyms)	Question: Find all airlines that have at least 10 flights. Paraphrases: Show the number of airlines that have at least ten flights.
`Multitype` (Contain multiple phenomena above)	Question: Find number of pets owned by students who are older than 20? Paraphrases: How many pets are owned by students over 20?
`Others`	Question: what is the name and nation of the singer who have a song having 'Hey' in its name? Paraphrases: Which singers have 'Hey' in their song's name? List their name and nation.

Our Main Findings

- SOTA supervised learning model (Picard) surfers
  - **14%** overall performance drop
  - **50%** performance drop to the most challenging perturbation type `DBcontent-equivalence`

Our Main Findings

- SOTA supervised learning model (Picard) surfers
  - 14% overall performance drop
  - 50% performance drop to the most challenging perturbation type `DBcontent-equivalence`
- Code-pretrained LLM (GPT-3 Codex) with in-context learning is
  - **more robust** than supervised learning models to **DB and SQL perturbations**
  - **less robust** than supervised learning models to **NLQ perturbations**

Our Main Findings

- SOTA supervised learning model (Picard) surfers
  - 14% overall performance drop
  - 50% performance drop to the most challenging perturbation type `DBcontent-equivalence`
- LLM (GPT-3 Codex) with in-context learning is
  - more robust than supervised learning models to DB and SQL perturbations
  - less robust than supervised learning models to NLQ perturbations
- Some model designs are beneficial to model robustness, e.g. **larger model size**
- Some model designs create a **reversed performance** to certain perturbations
  - strengthen the model performance on the original data but weaken the model robustness

Our Main Findings

- SOTA supervised learning model (Picard) surfers
  - 14% overall performance drop
  - 50% performance drop to the most challenging perturbation type `DBcontent-equivalence`
- LLM (GPT-3 Codex) with in-context learning is
  - more robust than supervised learning models to DB and SQL perturbations
  - less robust than supervised learning models to NLQ perturbations
- Some model designs are beneficial to model robustness, e.g. larger model size
- Some model designs create a reversed performance to certain perturbations
  - strengthen the model performance on the original data but weaken the model robustness
- More findings can be found in our paper https://openreview.net/pdf?id=Wc5bmZZU9cy