

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

Shuaichen Chang<sup>1</sup>, Jun Wang<sup>2</sup>, Mingwen Dong<sup>2</sup>, Lin Pan<sup>2</sup>, Henghui Zhu<sup>2</sup>, Alexander Hanbo Li<sup>2</sup>, Wuwei Lan<sup>2</sup>, Sheng Zhang<sup>2</sup>, Jiarong Jiang<sup>2</sup>, Joseph Lilien<sup>2</sup>, Steve Ash<sup>2</sup>, William Wang<sup>2</sup>, Zhiguo Wang<sup>2</sup>, Vittorio Castelli<sup>2</sup>, Patrick Ng<sup>2</sup>, Bing Xiang<sup>2</sup> <sup>1</sup>Ohio State University, <sup>2</sup> AWS AI Labs

ICRL 2023 Selected as a notable-top-5% paper



### Text-to-SQL

### NLQ

SQL

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



Text-to-SQL Model

SELECT winner name, winner rank FROM matches ORDER BY winner age ASC LIMIT 3

	Date	Ranking			
st_Name		Last_Name			
	Winner	Age			
	1	32			

### Does the high accuracy imply a robust model?

We perturb the data in Spider<sup>1</sup> to uncover the weaknesses in models that might not be evident when evaluated on the original data.



### **Our Evaluation Benchmark**

Includes 17 perturbation types across all three text-to-SQL components Diagnoses models with task-specific robustness phenomena

Provide insights for developing robust text-to-SQL models

### Text-to-SQL Example



ate	Ranking
Name	Last_Name
er_Rank	Loser_Name
1	Simona Halep

## Database (DB) Perturbations

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

,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,					·····
Pre-perturbation	Rankii	ng Date	Ranking		DB Perturba
Find the name and rank of the 3 youngest winners	Player_ID Winner_Name Serena Williams	First_Name Winner_Rank 1	Last_Name Loser_Name Simona Halep		Find the name the 3 younges
across all matches.					across all m
Pica	Picard				
SELECT winner_name, winner_rank FROM matches ORDER BY winner_age LIMIT 3					SELECT T2.c from match T1.v
Correct				Incorrect: miss	
`````			^^	· ``.	· · · · · · · · · · · · · · · · · · ·



Picard

champ\_name, T2.winner\_rank\_points hes as T1 join players as T3 on winner\_id = T3.player id

the constraints about 3 youngest winners

# Natural Language Question (NLQ) Perturbations

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

						, <b></b>
	Pre-perturbation	Ranki	ng Date	Ranking		NLQ Pertur
	Find the name and rank of	Player_ID	First_Name	Last_Name		Vho are the thr
	the 3 youngest winners	VINNEr_INAME Serena Williams	vvinner_Rank	LOSET_NAME Simona Halep	v	inners across
	across all matches.					And their
	Picard					
	SELECT winner_name, winner_rank FROM matches ORDER BY winner_age LIMIT 3					SELECT winn order
· · · · · · · · · · · · · · · · · · ·	Correct			,		Incorrect: selec
	````			·	`.	``



### **SQL** Perturbations

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

	· · · · · · · · · · · · · · · · · · ·			·····		····
	<b>Pre-perturbation</b>	Ranki	ng Date	Ranking		SQL Pertur
	Find the name and rank of the 3 youngest winners across all matches.	Player_ID Winner_Name Serena Williams 	First_Name Winner_Rank 1 	Last_Name Loser_Name Simona Halep 		Find the name the 8 younge across all r
	Picard					
	SELECT winner_name, winner_rank FROM matches ORDER BY winner_age LIMIT 3					SELECT mate
	Correct					Incorrect: pred
`.				·	``,	``



### 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

Keyword-syn SQL keywor synonyms)

Keyword-can SQL keywor carrier phrase

Column-sync column india onyms)

Column-carr column indic phrases)

Column-att: column indic gated attribut

Column-valu indicators by

Value-synon value indica onyms)

Multitype ( phenomena a

Others

Category	Example				
nonym (Replace rd indicators with	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.				
rrier (Imply ord indicators by es)	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.				
onym (Replace Question: List the name of teachers whose hometown is not cators with syn- Lever Urban District. Paraphrases: Find the name of teachers who were not b Little Lever Urban District.					
rier (Imply cators by carrier	mply Question: Show the name of teachers aged either 32 or 33? arrier Paraphrases: Which teachers are aged either 32 or 33.				
ribute (Imply cators by aggre- tes)	Question: What is the name of the conductor who has worked the greatest number of year? Paraphrases: Who has worked the longest as conductor?				
ue (Imply column 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.				
nym (Replace ators with syn-	Question: Find all airlines that have at least 10 flights. Paraphrases: Show the number of airlines that have at least ten flights.				
(Contain multiple bove)	Question: Find number of pets owned by students who are older than 20? Paraphrases: How many pets are owned by students over 20?				
	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.				

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

- 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

bation type DBcontent-equivalence arning is nd SQL perturbations perturbations

- 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

- 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 <u>https://openreview.net/pdf?id=Wc5bmZZU9cy</u> lacksquare