Responsible NLP Checklist

Paper title: *PV-SQL: Synergizing Database Probing and Rule-based Verification for Text-to-SQL Agents*
Authors: *Yuan Tian, Tianyi Zhang*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☒ A2. Did you discuss any potential risks of your work?
No. Our work focuses on improving the accuracy of text-to-SQL generation using publicly available benchmarks (e.g., BIRD and Spider) and does not involve human subjects, personal data, or deployment in real-world decision-making systems. Therefore, we do not identify direct risks associated with the use of this work beyond standard limitations of model-based systems, which are discussed in the Limitations section.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

N/A B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
(left blank)

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Yes. We report dataset statistics and evaluation settings in Section 5.1, including the number of examples for each benchmark (e.g., Spider, BIRD, and Mini-Dev) and the evaluation splits used in our experiments.

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. We describe the experimental setup in Section 5, including benchmarks, evaluation metrics, baselines, and base models. Key design choices and hyperparameters, such as the maximum number of probing and repair iterations, are described in Section 4.

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No. Our experiments are conducted with deterministic settings (temperature = 0), and results are

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

reported from a single run. Therefore, we do not report additional descriptive statistics such as variance or error bars.

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
(left blank)

N/A D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
(left blank)

N/A D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
We used AI assistants (e.g., ChatGPT) for limited language editing and phrasing suggestions during the writing process. All technical content, experimental design, and conclusions were conducted by the authors.