Responsible NLP Checklist

Paper title: *GBV-SQL: Guided Generation and SQL2Text Back-Translation Validation for Multi-Agent Text2SQL*
Authors: *Daojun Chen, Xi Wang, Shenyuan Ren, Qingzhi Ma, Pengpeng Zhao, An Liu*

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. We do not include a dedicated discussion of potential risks in the current paper. Our work focuses on Text2SQL generation and benchmark quality analysis rather than deployment in high-stakes real-world settings.

☑ 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. Section 4.1.1 describes the datasets used (Spider and BIRD), and Table 3 reports the number of samples in the BIRD dev evaluation subset and full dev set.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. Section 4.1.4 describes the implementation details, including backbone models (Deepseek-v3 and GPT-4o), temperature set to 0, and the maximum number of SQLChecker repair iterations set to 3.

☒ 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. We report single-run benchmark results and ablation results, but do not provide uncertainty estimates such as error bars, confidence intervals, or repeated-run summary statistics.

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.

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

☒ 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.?
No. The paper states that three SQL-proficient graduate students independently inspected and classified data items according to the proposed typology, but the full annotation instructions, screenshots, or participant-facing materials are not included in the paper.

☒ 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)?
No. The paper only states that the annotation/audit process was conducted by three SQL-proficient graduate students, but does not report recruitment procedures or compensation.

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)?
N/A. The study uses public benchmark datasets (Spider and BIRD) rather than newly collected personal data from participants.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. The study does not involve collection of sensitive personal data or human-subject experimentation beyond expert-style dataset auditing on public benchmarks.

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

☒ E1. If you used AI assistants, did you include information about their use?
No. AI assistants were used to support manuscript drafting, language polishing, and translation. However, we did not include an explicit disclosure statement about this usage in the current manuscript.