

Responsible NLP Checklist

Paper title: *LLM-FK: Multi-Agent LLM Reasoning for Foreign Key Detection in Large-Scale Complex Databases*

Authors: *Zijian Tang, Ying Zhang, Sibao Cai, Ruoxuan Wang*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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?

The research focuses on automated metadata discovery and schema analysis within databases. It uses statistical methods to identify structural relationships (foreign keys) and does not involve human subjects, personal sensitive data, or any real-world deployment that could cause social harm.

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

- 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?

We utilized widely-studied, publicly available benchmark datasets that are standard in the database community and do not contain sensitive personal information or offensive content.

- 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?

Detailed statistics are documented in Appendix C.

C. Did you run computational experiments?

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

We describe the experimental setup in Section 4.1. Note that our proposed method is hyperparameter-free and does not require extensive hyperparameter tuning or search. All fixed model configurations and environment settings are detailed in Section 4.1 and Appendix D (Reproducibility Statement).

- 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?

Yes. The results reported are highly consistent across multiple runs due to the inherent stability and deterministic nature of our proposed method. Since the performance variance is negligible, the reported scores represent a reliable and reproducible measure of the model's capability. Detailed experimental configurations are provided in Section 4.1 and Appendix D to ensure full reproducibility.

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.?
(left blank)

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)?
We did not recruit or pay any participants as our study relies exclusively on established, publicly available benchmark datasets.

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)?
No, because we only used standard, publicly available benchmark datasets. Consent was handled by the original creators of these datasets, and our use complies with their existing licenses.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The datasets used are standard benchmarks and are publicly available.

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?
We only use AI to assist with coding and writing, including grammar checking.