Responsible NLP Checklist

Paper title: *EVM-QuestBench: An Execution-Grounded Benchmark for Natural-Language Transaction Code Generation*
Authors: *Pei Yang, wanyi Chen, Ke Wang, Lynn Ai, Eric Yang, TIANYU SHI*

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?
Section 1 discusses financial and safety risks of LLM-driven on-chain transaction automation (e.g., incorrect parameters leading to reverts or irreversible loss). Section 7 further discusses evaluation risks such as RPC/fork instability and unpinned fork block height affecting cross-run comparability.

☑ **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?
Section 8 (Section 8 (Ethics Statement) states that all experiments run on a locally forked blockchain without real-fund transfers or live state modification, and that the benchmark does not collect or process personal data.Ethics Statement)

☑ 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?
Section 3 reports benchmark composition and data statistics, including the number of tasks (107), the atomic/composite split (62/45), template counts and sampling ranges, workflow step statistics for composite tasks, and difficulty distributions.

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

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We did not perform hyperparameter search or report best-found hyperparameters from a sweep. We instead document a fixed evaluation protocol (e.g., decoding temperature, multi-round evaluation, fork/snapshot settings) in Section 4 and additional reproducibility settings in Appendix A.

☑ 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,

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.

etc. or just a single run?
Section 5 reports descriptive statistics over five independent evaluation rounds per model, including the mean Atomic, Composite, and Total scores, standard deviations, coefficient of variation (CV%), and observed minmax ranges (Table 1).

☒ **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?**

☑ E1. If you used AI assistants, did you include information about their use?
Section 3 (Instruction Templates and Parameterization); Introduction (contributions bullet on LLM-assisted benchmark development).