Responsible NLP Checklist

Paper title: *Code Reffix: A Benchmark for Reflection-Guided Code Repair with Large Language Models*
Authors: *Zaiyuan Di, Jianting Chen, Yunxiao Yang, Xiaoying Gao, Li Yang, Zhihao Wang, Yang Xiang*

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?
This paper introduces a benchmark for code repair and is not directly tied to a specific application or deployment setting. The benchmark is constructed from existing public and widely used datasets. Beyond the common limitations of benchmark construction, such as possible data distribution or domain coverage imbalance, we did not identify significant privacy, safety, or broader societal risks.

☑ 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 utilize open-source code datasets that exclusively consist of LLM-generated code and do not contain any personally identifiable information.

☑ 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?
3.2

☑ C. Did you run computational experiments?

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

☑ 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?
4.2, 4.3, 4.4

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.?
The manual audit was conducted by the authors themselves, and no formal written participant instructions were prepared or reported.

☒ 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)?
The manual audit was conducted by the authors themselves, rather than by external participants. It did not involve recruitment or payment issues.

☒ 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)?
This work relies on open-source datasets and LLM synthetic data. We did not collect data directly from individuals.

☐ 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?
We used ChatGPT to translate some sentences in this paper.