Responsible NLP Checklist

Paper title: *EngiBench: A Benchmark for Evaluating Large Language Models on Engineering Problem Solving*
Authors: *Xiyuan Zhou, Xinlei Wang, Yirui He, Ruixi Zou, Yang Wu, Yuheng Cheng, Yulu Xie, Wenxuan Liu, Huan Zhao, Yan Xu, Jinjin Gu, Junhua Zhao*

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?
Yes. The discussion of ethical considerations and potential risks is provided in Appendix B.

☑ **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?
No. The released dataset does not contain any personally identifying information or offensive content. Any personal information collected from human participants (e.g., for acknowledgments) was obtained with explicit consent, stored separately, and is not included in the released data. Details are provided in Appendix B.

☑ 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. Dataset statistics, including the number of problems and task composition across difficulty levels, are reported in Section 3.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. The experimental setup and inference configurations 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. The evaluation reports deterministic benchmark scores without variance across repeated runs.

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.?
Yes. The task setup and instructions for human participants are described across Section 3, Section 4, and Appendix D.

☑ 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)?
Yes. Human participants contributed voluntarily without monetary compensation. See Appendix B (Ethical Considerations).

☑ 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)?
Yes. Human participants contributed voluntarily with explicit consent for research use. See Appendix B (Ethical Considerations).

☒ 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?
Yes. Details are provided in Appendix A (The Use of Large Language Models).