Responsible NLP Checklist

Paper title: *Idea First, Code Later: Disentangling Problem Solving from Code Generation in Evaluating LLMs for Competitive Programming*
Authors: *Sama Hadhoud, Alaa Elsetohy, Frederikus Hudi, Jan Christian Blaise Cruz, Steven Halim, Alham Fikri Aji*

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 7

☑ **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?
he dataset consists exclusively of competitive programming problem statements, expert-written editorials, and official test cases released as part of programming contests or course assessments. These materials are technical in nature and do not contain personally identifying information, references to identifiable individuals, or user-generated content. As a result, no explicit PII screening, anonymization, or offensive-content filtering was required

☑ 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 2.3; Appendix B

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 2.2; Appendices C, E, F. No hyperparameter search was performed; the paper reports the actual inference setup/configuration used.

☑ 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?
Section 3; Appendix H

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.?
Section 2.4; 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)?
The study involved a single expert annotator who is also one of the papers authors. The annotation was conducted as part of the authors research contribution, was not crowdsourced, involved no vulnerable population, and did not involve any monetary compensation. Therefore, recruitment and payment details were not applicable or necessary to report.

☒ 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)?
The data consists exclusively of competitive programming problems, editorials, and test cases released as part of contests or course assessments. No personal data, user-generated content, or human subject data is included. Therefore, individual consent was not applicable.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This work does not involve human-subject experimentation, personal data collection, or intervention studies. The dataset is composed of technical artifacts (programming problems and editorials). For any materials that were not publicly available, we obtained explicit permission from the respective course instructors or contest organizers prior to use. The single expert annotation task poses minimal risk. As such, ethics review board approval was not required.

☑ **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 8 Use of Generative AI