Responsible NLP Checklist

Paper title: *CODESTRUCT: Code Agents over Structured Action Spaces*
Authors: *Myeongsoo Kim, Chao-Chun Hsu, Dingmin Wang, Shweta Garg, Varun Kumar, Murali Krishna Ramanathan*

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?
The paper includes a dedicated Limitations section, but it does not include a dedicated discussion of broader potential risks or misuse. The current discussion focuses on technical limitations of the approach.

☑ **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?
The work uses existing public software-engineering benchmarks and does not newly collect or release human-subject data, but the paper does not include a dedicated discussion of screening for personally identifying 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?
Section 4.1 reports relevant dataset statistics and characteristics, including 500 SWE-Bench Verified issues and 149 CodeAssistBench tasks across 7 programming languages.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.3 describes the experimental setup, including the models used, default decoding parameters, no model-specific fine-tuning or task-specific adaptation, unchanged prompts, and interaction budgets.

☒ 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?
The paper reports main benchmark results in Tables 1 and 2, along with additional analyses in

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.

the appendix, but it does not report variance measures such as error bars, confidence intervals, or statistics over multiple runs, nor does it explicitly state that results are averaged over repeated runs.

☒ **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.?
No human participants or annotators were used.

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)?
No human participants or annotators were used.

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)?
No human participants or annotators were used.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human participants or annotators were used.

☑ **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?
AI assistants were used only for limited writing and coding support. All technical content, experiments, analysis, and final verification were performed and checked by the authors.