## Responsible NLP Checklist

Paper title: *Context as a Tool: Context Management for Long-Horizon SWE-Agents*
Authors: *Shukai Liu, Bo Jiang, Jian Yang, Yizhi LI, Jinyang Guo, Xianglong Liu, Bryan Dai*

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?
*See the Ethical Statement section, which discusses risks including over-reliance on agent outputs, misuse for unwanted code changes, and privacy risks from repository artifacts.*

☑ **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?
*See the checklist section on personally identifying or offensive content, which explains that the public GitHub-derived data may contain identifiers or offensive language and notes planned automated scrubbing for released artifacts.*

☑ 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?
*See Section 3 and Table 1, which report dataset sizes, the 500-instance evaluation set, the 20k CAT-Instruct and 20k BASE-INSTRUCT training data, and statistics such as trajectory length and token counts.*

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*See Section 3, which specifies the base model, optimizer, weight decay, learning-rate schedule, warm-up ratio, peak learning rate, temperature, context length, tool set, and maximum interaction rounds.*

☑ 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?
*See Table 1, Table 2, Table 3, and Figures 46, which report averages, medians, maxima, Pass@1 on N=500, token usage, and task-difficulty breakdowns, making the reported statistics explicit.*

---

*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?**

☒ 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.?
*N/A, because the work did not involve recruited participants or annotators.*

☒ 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)?
*N/A, because the work did not recruit or pay participants or annotators.*

☑ 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)?
*See the checklist section on data consent, which explains that no new data were collected directly from individuals and that the work relies on publicly available software engineering artifacts under the original platform and dataset terms.*

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*N/A, because the work did not involve human-subject experiments or new data collection from human participants and is described as typically exempt under standard IRB criteria.*

☑ **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?
*See the checklist section on AI assistants*