Responsible NLP Checklist

Paper title: *CodeWiki: Evaluating AIs Ability to Generate Holistic Documentation for Large-Scale Codebases*
Authors: *Anh Nguyen Hoang, Minh Le-Anh, Bach Le, Nghi D. Q. Bui*

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.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

N/A 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?
(left blank)

☑ 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?
4

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

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

☑ 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

☑ **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 pilot study involved a simple preference evaluation task. Participants received access to official codebases, official documentation, and anonymized outputs from two systems, then provided binary preference judgments. We described the evaluation protocol in Appendix F but did not include

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.

verbatim instructions as the task was straightforward: assess which documentation better reflected the official documentation and repository architecture.

☒ 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)?
Participants were colleagues recruited through professional networks on a voluntary basis. As this was a small-scale pilot study (3 participants, approximately 2 hours total involvement) conducted to provide preliminary validation of our automated methodology, no monetary compensation was provided. We acknowledge this limitation and note that comprehensive human evaluation with proper compensation structures represents important future work.

☑ 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)?
Appendix F. Participants were informed that their preference judgments would be used to validate the automated evaluation methodology and reported in aggregate form without identifying information.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The pilot study involved only preference judgments on publicly available open-source code documentation, presenting minimal risk to participants. No personally identifiable information was collected beyond professional roles. The study falls under minimal risk evaluation activities that do not require formal ethics review at our institution.

☑ **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 during both the coding implementation and manuscript writing phases of this research. For coding, AI assistants helped with software development tasks related to the CodeWiki framework implementation. For writing, AI assistants were used to help draft and refine portions of the manuscript text. The core research contributions, experimental design, analysis, and scientific conclusions represent the original work of the authors, with AI assistants serving as productivity tools rather than primary contributors to intellectual content.