

## Responsible NLP Checklist

Paper title: *EGSS: Entropy-guided Stepwise Scaling for Reliable Software Engineering*

Authors: *Chenhui Mao, Yuanting LEI, Zhixiang Wei, Ming Liang, Zhixiang Wang, Jingxuan Xu, Dajun Chen, wei jiang, Yong Li*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- <sup>N/A</sup> 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?  
*There is no risk inherent in our approach itself.*

### 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?  
*Our work relies solely on the SWE-Bench benchmark, which consists of public GitHub issues and code patches from open-source software projects. These data contain only technical software engineering content (e.g., bug reports, code diffs, and test cases) and do not include personally identifiable information (such as names, emails, or account identifiers) or offensive content. Therefore, no additional anonymization or content filtering was necessary, and we did not include a dedicated discussion in the paper.*

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?  
*The study employs commonly used, open-source datasets (e.g. Swe-Bench Verified) within the domain, eliminating the need for further statistical analysis.*

### C. Did you run computational experiments?

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?  
*Section 5.3 Experiment Setup*

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 5.4 Experimental Results*

*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.?

*In this work, we did not create a dataset, so there were no annotators involved. All data used in this paper come from publicly available open-source datasets.*

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)?

*In this work, we did not create a dataset, so there were no annotators involved. All data used in this paper come from publicly available open-source datasets.*

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)?

*In this work, we did not create a dataset, so there were no annotators involved. All data used in this paper come from publicly available open-source datasets.*

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

*In this work, we did not create a dataset, so there were no annotators involved. All data used in this paper come from publicly available open-source datasets.*

**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 assistant was used during code development to generate initial implementations of data preprocessing functions. All code was manually verified and adapted by the authors.*