Responsible NLP Checklist

Paper title: *DeepGuard: Secure Code Generation via Multi-Layer Semantic Aggregation*
Authors: *Li Huang, Zhongxin Liu, Yifan Wu, Tao Yin, dong li, Jichao Bi, Nankun Mu, Hongyu Zhang, Meng Yan*

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?
We explicitly state in our Ethics Statement section that we "have not identified any significant ethical considerations associated with our work".

☑ **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 dataset consists of vulnerable and secure code snippets in Python and C/C++ addressing software weaknesses like SQL injection and cross-site scripting. It does not contain personally identifying information or offensive content, making anonymization protocols unnecessary.

☑ 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?
Relevant statistics and distributions are thoroughly reported in Appendix A and Figure 8. The training set comprises exactly 1,606 programs forming 803 vulnerable/secure pairs. Tables 6 and 7 detail the specific scenario splits for testing and out-of-distribution generalization.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup and hyperparameters are documented in Appendix C.1 and summarized in Table 8. This includes explicit values such as 5 epochs, a learning rate of 2x10^-5, a LoRA rank of 16, and specific loss weights.

☑ 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?
When evaluating inference latency in Table 11, the paper reports the average time across 5 runs and includes the standard deviation (e.g., 0.0331 0.0010) to clearly indicate variance and error margins.

☒ **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.?
Not applicable, as the evaluation framework relies entirely on automated, executable unit tests and static analysis tools like CodeQL. No human subjects or annotators were utilized.

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)?
Not applicable, as the evaluation framework relies entirely on automated, executable unit tests and static analysis tools like CodeQL. No human subjects or annotators were utilized.

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)?
Not applicable, as the evaluation framework relies entirely on automated, executable unit tests and static analysis tools like CodeQL. No human subjects or annotators were utilized.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable, as the evaluation framework relies entirely on automated, executable unit tests and static analysis tools like CodeQL. No human subjects or annotators were utilized.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
While the paper researches code generation models, there is no declaration that AI assistants were used to conduct the research or write the manuscript itself.