Responsible NLP Checklist

Paper title: *EvolvR: Self-Evolving Pairwise Reasoning for Story Evaluation to Enhance Generation*
Authors: *Xinda Wang, Zhengxu Hou, Yangshijie Zhang, yanbingren, Jialin Liu, ChenZhuo Zhao, Zhibo Yang, Bin-Bin Yang, Feng Xiao*

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?
Appendix C.1.1. We discussed the impact of repeatedly selecting random pairs for pairwise evaluation on the results.

☑ **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?
N/A. All datasets utilized in this work (HANNA, StoryER, and OpenMEVA) are publicly accessible open-source resources that have been carefully curated to exclude personal identifying information and offensive material.

☑ 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. We adhere to the official data split protocol established by StoryER, with a train:validation:test ratio of 8:1:1.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C.2.3. We report the weights for each dimension in the multi-dimensional evaluation as well as the weights for the three reward components.

☑ 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 4 and Appendix C.1.1. We report results from both single evaluation runs and the averaged scores from multiple pairwise evaluation iterations.

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.?
Ethical Considerations. Expert annotators were provided with the following evaluation guidelines: "You will evaluate AI-generated stories on six dimensions (creativity, coherence, relevance, complexity, surprise, and engagement) using a 1-5 scale. Each story should be read completely before scoring. Focus solely on story quality without considering the AI system that generated it. Your evaluations will be used to validate automatic evaluation metrics. No personal information will be collected."

☑ 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)?
Ethical Considerations. Human evaluators were professional annotators from our research team with specialized training in creative writing assessment and extensive experience in NLP evaluation tasks. These were not crowdsourced workers but dedicated team members with relevant academic backgrounds. Compensation was provided according to institutional standards for professional research assistance, ensuring fair remuneration commensurate with the specialized nature of the task and the evaluators' expertise level.

☑ 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)?
Ethical Considerations. All evaluators provided informed consent before participation. They were explicitly informed that: (1) they would be evaluating AI-generated content only, (2) their judgments would be used solely for research purposes to validate automatic metrics, (3) no personal information beyond their quality assessments would be recorded, and (4) they could withdraw from the evaluation process at any time without penalty. The evaluation data contains only story quality scores without any annotator identification.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. Our human evaluation only involved assessment of AI-generated stories and did not collect personal data or involve human subjects research. The evaluators assessed story quality using standard criteria, which does not require ethics board approval as it does not involve research on human subjects.

☒ **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?
(left blank)