## Responsible NLP Checklist

Paper title: *Igniting Creative Writing in Small Language Models: LLM-as-a-Judge versus Multi-Agent Refined Rewards*
Authors: *Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, Dawei Yin*

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?
*8 Limitations*

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

☑ B1. Did you cite the creators of artifacts you used?
*References*

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
*We have not discussed specific licensing information within the body of the paper. This information will be made available in our GitHub repository.*

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*The paper's scope is centered on the technical methodology and its experimental validation. Discussions regarding licensing, compliance with the intended use of third-party artifacts, and the specific access conditions of the original data sources were considered outside the primary focus of this research paper.*

☑ 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?
*4.2 Datasets*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*4.1 Task Design,A.2 Scope and Characteristics of Chinese Greetings,A.3 Details on the Human Evaluation Protocol*

*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.*

☑ 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?
*4.2 Datasets*

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*A.1 Hyperparameters*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*A.1 Hyperparameters*

☑ 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?
*Table 1,Table 4,Table 5*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
*4.4 Implementation Details,A.1 Hyperparameters*

☑ **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.?
*4.3 Rubric Design*

☑ 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)?
*A.3 Details on the Human Evaluation Protocol*

☒ 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)?
*The paper does not contain a discussion on the process of obtaining user consent. The data is described as internal company data sourced from online interactions. For such proprietary data, user consent is typically governed by the platform's Terms of Service and Privacy Policy, which are agreed to by users upon registration and use of the service. These agreements generally cover the use of anonymized and aggregated data for internal research and product improvement. The manuscript operates on the assumption that these standard corporate procedures are in place and thus focuses on the technical steps of data anonymization rather than the legal and procedural details of consent acquisition.*

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*The manuscript does not mention approval from an ethics review board. The research utilizes proprietary corporate data from online interactions, which is typically governed by internal company data policies and user agreements.*

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*A.3 Details on the Human Evaluation Protocol*

☑ **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?
*3 Methodology*