Responsible NLP Checklist

Paper title: *MUG-Eval: A Proxy Evaluation Framework for Multilingual Generation Capabilities in Any Language*

Authors: *Seyoung Song, Seogyeong Jeong, Eunsu Kim, Jiho Jin, Dongkwan Kim, Jay Shin, Alice Oh*

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?
The work presents an evaluation framework rather than a generative model or application. As a benchmarking tool for assessing existing LLMs, it doesn't introduce new risks beyond those already present in the models being evaluated.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 3 and Appendix A.2

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The paper uses existing publicly available benchmarks and datasets accessed through standard research channels (Hugging Face, APIs). Since these are established research artifacts commonly used in the community and the paper focuses on evaluation methodology rather than redistribution of the datasets, explicit license discussion was not included. The artifacts are properly cited and used for their intended research purposes.

☒ 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 artifacts used are research benchmarks made publicly available for scientific use. The paper adapts these benchmarks into conversational evaluation tasks, which falls within their intended research purpose for evaluating language 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 paper uses existing research benchmarks and standard evaluation datasets that are already established in the research community. These benchmarks consist of general knowledge questions,

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.

word lists, and code samples without personally identifying information or offensive content. As publicly available research artifacts designed for evaluation purposes, they have already undergone appropriate content review by their original creators.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 and Appendix A

☑ 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 and Appendix A

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

N/A C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The paper reports model sizes in model names but does not provide explicit parameter counts for all models or computational budget details. Since the evaluation uses API calls rather than local inference, total GPU hours and detailed infrastructure specifications were not reported.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix B.2

☑ 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 4.1 and 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?
Section 3.2 and Appendix B.2

☑ **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.?
Appendix C.6.1 details the human evaluation methodology, stating that the scoring rubric was adapted from Hada et al. (2024a).

☑ 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 and Appendix C.6.1

☑ 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

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

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix C.6.1

☑ **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?
Acknowledgments