## Responsible NLP Checklist

Paper title: *Safety of Large Language Models Beyond English: A Systematic Literature Review of Risks, Biases, and Safeguards*
Authors: *Aleksandra Krasnodbska, Katarzyna Dziewulska, Karolina Seweryn, Maciej Chrabaszcz, Wojciech Kusa*

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?
*Our work is a systematic review of the safety evaluation of large language models beyond English. It does not involve original experimentation, human subjects, or the handling of sensitive or private data. As such, no new risks are introduced, and the review solely synthesizes and analyzes findings from existing, publicly available research. Therefore, a discussion of potential risks was not applicable in this context*

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

☑ B1. Did you cite the creators of artifacts you used?
*All datasets and models reviewed are properly cited in Sections 1, 4, 7 and Appendix E.*

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
*Licensing information for each dataset is documented in Appendix E. For our created artifact (Streamlit app), we host only metadata and references no redistribution of licensed datasets thus respecting all original licenses.*

☑ 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)?
*All datasets and models were used solely for research/survey purposes, consistent with their intended use as stated by their creators. Our dashboard only provides aggregated, cited information (no raw data) as described in Section 2 (Systematic Review Protocol) and Section 4.*

☒ 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?
*As a survey, we do not collect new raw data.*

---

*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.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*The paper documents each datasets scope, languages, size, and categories in Section 4 and tables therein. For the Streamlit app, we provide interactive access and metadata documentation see link in Section 1.*

☑ 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?
*Yes, dataset statistics (sample counts, language coverage,) are presented in Section 4. These same statistics are available interactively through the Streamlit dashboard: https://multilingualrt.streamlit.app/*

☒ **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?
*(left blank)*

N/A C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*(left blank)*

N/A 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?
*(left blank)*

N/A 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?
*(left blank)*

☒ **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.?
*All steps of the systematic review were conducted by the co-authors of the publication, with no involvement from external members.*

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)?
*(left blank)*

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)?
*(left blank)*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*(left blank)*

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*(left blank)*

☒ **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?
*We used AI assistants to help improve the grammar of the paper's writing. Additionally, as described in Section 2.2, we experimented with the PaperFinder tool (https://paperfinder.allen.ai); however, the results of this search were independently validated by two researchers.*