Responsible NLP Checklist

Paper title: *Multilingual Refusal Alignment for Safer Large Language Models*
Authors: *Aleksandra Krasnodbska, Wojciech Kusa, Aldo Lipani*

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?
Yes we acknowledge potential risks in the Limitations section, noting that multilingual models may behave inconsistently across languages and that the automatically generated RefusEU dataset could propagate biases or be misused in unintended ways.

☑ **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?
No the dataset we created is generated and does not contain any personally identifying information or offensive content. No additional anonymization steps were necessary.

☑ 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 3 and 4, Appendix C

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes details about the computing infrastructure and the number of parameters of the models used are provided in Appendix D

☑ 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?
Yes we report results for each language individually and provide additional mean scores across low-resource, high-resource, and all languages where relevant. All reported results are clearly labeled to indicate whether they represent per-language scores or aggregated means. Error bars or measures of variance are not included, as each score is based on a single evaluation run.

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.?
Appendix C.2

☒ 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)?
No external participants were recruited for this study. All manual analysis and annotation were carried out by the authors, so issues of recruitment and payment do not apply.

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)

☑ **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?
Privately revealed to ACL ARR 2026 January Program Chairs, ACL ARR 2026 January Submission3740 Area Chairs, ACL ARR 2026 January Submission3740 Authors, ACL ARR 2026 January Submission3740 Reviewers, ACL ARR 2026 January Submission3740 Senior Area Chairs Yes we used AI assistants to check grammar and improve writing style, as described in the Limitations section. No content, results, or scientific claims were generated or modified by the AI assistants.