Responsible NLP Checklist

Paper title: *MonCulture-Eval: A Hierarchical Benchmark for Evaluating Mongolian Cultural Capabilities of Large Language Models across Scripts and Regions*
Authors: *Quulgan Minggad, XiaoZinan, Yuan Sun*

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?
Section 10 (Ethical Considerations) discusses dual-use risks and the potential for generating culturally targeted disinformation.

☑ **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?
Section 10 (Ethical Considerations) explains that the dataset focuses on localized social logic and historical norms rather than identifying individual people.

☑ 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 4.3 and Table 1 provide comprehensive statistics on domain distribution and task counts.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 details the models evaluated, prompt engineering strategies, and the LLM-as-a-judge protocol.

☑ 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 5.3 and 6 report accuracy percentages, average judge scores, Pearson correlation (r), and Inter-Annotator Agreement (IAA) rates.

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 B

☑ 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)?
Section B.1 details the recruitment of 12 native experts and their compensation rates (30-50 RMB/hour).

☑ 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)?
Section 10 (Ethical Considerations) notes the "Indigenous-First" protocol and the use of paid native experts

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The research recruited experts in language and history for text proofreading and cultural knowledge assessment. It does not fall within the scope of traditional human subject research that requires IRB approval, such as studies involving patients or psychological experiments.

☑ **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?
Section 5.3 describes using Gemini-2.5-Pro and GPT-5.2 as automated evaluators