Responsible NLP Checklist

Paper title: *Probing Social Identity Bias in Chinese LLMs with Gendered Pronouns and Social Groups*
Authors: *GENG LIU, Li Feng, Junjie Mu, Mengxiao Zhu, Francesco Pierri*

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 Ethical Considerations. We discuss the risks of reinforcing stereotypes and harmful group descriptions, and clarify that the results reflect model behavior rather than valid descriptions of social groups

☑ **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 Ethical Considerations. The study uses synthetic prompts and no personally identifiable information, but the generated outputs may contain harmful or offensive content toward social groups.

☑ 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 Data and Methods reports model set, prompt design, number of social groups, generated responses, and the WildChat supplementary dataset.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section Data and Methods and the Appendix. We describe prompt construction, model selection, contextual scaffolding, sentiment and toxicity measurement, and the regression specifications.

☑ 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 Results and the Appendix. We report odds ratios / regression coefficients with 95% confidence intervals, along with descriptive statistics for the data and supplementary analyses.

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.?
The manual validation was conducted internally by the authors using simple sentiment judgment guidelines, but the full annotator instructions were not reported verbatim in the paper.

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)

☑ **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?
AI assistants were used only for language polishing and minor code debugging. Research design, experiments, statistical analysis, and interpretation are by the authors.