Responsible NLP Checklist

Paper title: *Neutral Is Not Unbiased: Evaluating Implicit and Intersectional Identity Bias in LLMs Through Structured Narrative Scenarios*
Authors: *Saba Ghanbari Haez, Mauro Dragoni*

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.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
Section 1 Introduction

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We made our dataset and code available for reproducibility but did not explicitly specify or discuss licensing terms in the paper or repository.

☑ 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)?
Our artifacts (scenario dataset and evaluation code) were newly created for bias evaluation in LLMs, intended solely for research purposes. Their use in our study is consistent with this intended purpose.

☑ 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?
All scenarios were fictional and constructed with gender-neutral names; no personally identifying information or offensive content was included. We explicitly verified that the dataset contains no sensitive or harmful material.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We documented the dataset in detail, including scenario domains, identity variations, neutral naming strategy, and linguistic statistics (e.g., token counts, readability, sentiment distribution).

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.

☑ 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?
Dataset statistics are reported in Section 3.1 (Scenario Design) and Appendix A (Tables 913, Figures 47).

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

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
While we specified the models used and their parameter sizes where applicable, we did not report the computational budget (e.g., GPU hours) or infrastructure details.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.2.2 (Prompting Strategy) details the experimental setup and generation parameters (Temperature = 0.5, Top-p = 0.95, Frequency/Persistence penalties = 0.1, Max tokens = 400), applied identically across models. No hyperparameter search was performed; we used fixed settings.

☑ 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?
We report means, deltas, and bootstrapped 95% confidence intervals for CDA reviewer scores (Section 4.1.1, Tables 13), and summary statistics for quantitative analyses (Sections 4.34.4, Tables 47, Figures 3, 912).

☒ 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?
We relied on standard NLP tools (e.g., sentiment analysis and similarity metrics) but did not report detailed implementation choices, package versions, or parameter settings.

☑ **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.?
we provided the annotator instructions via the full CDA-34 scoring rubric in Appendix B.1 and described the reviewer protocol in Section 4.1.1 (two expert reviewers, item scales, scoring procedure). No crowdworker-facing risk disclaimers/screenshots were needed since annotators were experts rating fictional model outputs with no personal data.

☒ 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. We did not recruit or pay participants; CDA annotations were performed by two expert reviewers (one internal collaborator and one external volunteer). Consequently, there was no crowdsourcing platform, recruitment process, or payment to report.

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)

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Our annotators were two expert collaborators (one internal, one external volunteer) who applied CDA ratings. Since this was not a crowdworker population and demographic/geographic characteristics were not relevant to the studys goals, we did not collect or report such information.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)