Responsible NLP Checklist

Paper title: *Investigating Counterfactual Unfairness in LLMs towards Identities through Humor*
Authors: *Shubin Kim, Yejin Son, Junyeong Park, Keummin Ka, Seungbeen Lee, Jaeyoung Lee, Hyeju Jang, Alice Oh, Youngjae Yu*

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. Potential risks are discussed in the Ethics Statement. Our work involves identity-based and potentially offensive humor, which may expose annotators or readers to harmful language. We mitigate these risks through controlled experimental design, restricted research use, and careful framing that emphasizes bias diagnosis rather than endorsement of harmful content.

☑ 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?
Yes. We carefully reviewed all data sources to ensure that no personally identifying information is included. The datasets are derived from publicly available corpora (e.g., Reddit and Kaggle), and we filter content to focus on generalized identity categories rather than individuals. While the data may contain potentially offensive content (e.g., disparagement humor), it is used strictly for research purposes, and appropriate safeguards are discussed in Section 3.2 and the Ethics Statement.

☑ 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. We report detailed dataset statistics in Section 3 (Dataset Construction) and Appendix A.3, including the number of humor samples (e.g., 400 identity-agnostic jokes and 737 identity-specific jokes), as well as the number of generated prompts and responses (e.g., 12,320 prompts per model and 61,600 total instances).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. We describe the experimental setup in Section 2.1 and Appendix A.3, including model versions, prompting strategy, and inference settings (e.g., temperature = 0.7). As our study does not involve

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.

training or hyperparameter tuning, we focus on standardized inference settings to reflect realistic deployment conditions.

☑ 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 descriptive statistics throughout Sections 25, including mean values, percentage differences (e.g., ARR), and statistical significance (p-values). Results are aggregated across multiple identity pairs and prompt instances, ensuring robustness beyond single-run outputs.

☑ **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.?
Yes. We provide detailed annotation prompts and evaluation criteria in Appendix A.4 and D.2, including full instructions used for automated and human evaluation.

☒ 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. While we employed trained human annotators for dataset verification, we did not include detailed reporting on recruitment procedures or compensation. This is a limitation of the current work and will be addressed in future revisions.

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)?
N/A. The datasets used in this study are derived from publicly available sources (e.g., Reddit and Kaggle corpora), and do not involve direct interaction with identifiable individuals.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. This study does not involve direct human subject experimentation or intervention, and relies on publicly available data and annotation procedures.

☑ **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?
Yes. We disclose the use of AI assistants in the Ethics Statement (Section Ethics Statement), including their roles in language editing, dataset preprocessing, and evaluation.