Responsible NLP Checklist

Paper title: *To Mask or to Mirror: Human-AI Alignment in Collective Reasoning*
Authors: *Crystal Qian, Aaron T Parisi, Clmentine Bouleau, Vivian Tsai, Mal Lebreton, Lucas Dixon*

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. Ethical considerations and potential risks are provided in the Limitations section on page 10.

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

☑ B1. Did you cite the creators of artifacts you used?
Deliberate Lab, the experimentation platform, is cited with supplementary Appendix F. All models used are publicly available; usage is summarized in Appendix E.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
All models used are publicly available; usage is summarized in Appendix E.

☒ 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)?
No data derivatives or artifacts were used for this study.

☑ 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 data collection details are provided in Appendix B. No personally identifying info was gathered. Informed consent was obtained and an IRB was consulted.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Demographics of human participants are provided in Table 3. Specifications for models and parameters are provided in Appendix E, with prompts and implementation detailed in Appendix G for reproducibility.

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?
All statistics are provided in the main text (primarily Section 6) with supplementary statistics in Appendix D. All statistical tests are described and cited where relevant.

☑ **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?
Model size and budget are specified in Appendix E.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
There was no hyperparameter search. Our experimental setup (including human experiment and LLM implementation) is detailed in Section 5.

☑ 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?
All statistics include relevant standard errors, deviations, p-values, and labels. They can largely be found in Section 6 and Appendix D.

☒ 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?
No existing packages were used, only off-the-shelve LLM APIs (which are documented in Appendix E).

☑ **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.?
Screenshots of the instruction interface are provided in Figures 1, 2, 10, and 11. Instructions are described as part of Appendix G, and disclaimers are described in 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)?
Recruitment information and payment are provided in Appendix B, specifically providing details on country of residence and adequate compensation.

☑ 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)?
This is described in the first paragraph of Section B.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This is provided in the main text (section 5.2).

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Table 3 (Appendix B) provides descriptive statistics.

☒ **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)