Responsible NLP Checklist

Paper title: *Prior Beliefs Prejudice LLM-as-Judge: Evidence from Persuasion Evaluation*
Authors: *Pardis Sadat Zahraei, Xiaoning Wang, Nimet Beyza Bozdag, Gokhan Tur, Dilek Hakkani-Tr*

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 risks of the ConvinceQA dataset containing harmful content and potential misuse of persuasion-as-probe.

☑ **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?
The dataset contains synthetically generated arguments and does not include personally identifying information.

☑ 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 3.3 reports dataset statistics (27,756 arguments, 1,263 claims, breakdown by category).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C.3 details training configurations including LoRA rank, learning rate, batch size, epochs, etc.

☑ 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?
The paper reports CWA scores, delta values, and Table 4 breakdowns across models.

☑ **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.?
Annotators used the same prompt templates as LLM evaluation (Appendix A); no additional instructions were given.

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.

☒ 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)?
Annotators were NLP researchers at UIUC involved in the project; no crowdsourcing platform or monetary payment was used.

☑ 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 states all human raters provided informed consent.

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?**

N/A E1. If you used AI assistants, did you include information about their use?
Used for minor editing and grammatical corrections; all research, ideation, and analysis were conducted by the authors.