Responsible NLP Checklist

Paper title: *Vulnerability of LLMs' Stated Belief? LLMs Belief Resistance Check Through Strategic Persuasive Conversation Interventions*
Authors: *Fan Huang, Haewoon Kwak, Jisun An*

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?
Discussed in the Ethics Statement.

☑ **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?
(left blank)

☑ 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?
Reported in Section 3.3 (Dataset) with Table 2 showing per-dataset original and post-filtering instance counts.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Reported in Section 6.2 (Fine-Tuning Experiments)

☑ 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 3.4 (Evaluation Metrics) states results are means across instances and appeal types, and then section 4,5,6 report the experimental results.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A 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 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.

No external or crowdsourced participants; the quality-verification annotation was performed in-house by a single grad-level researcher using a simple 1–5 rating scale, with no risks, disclaimers, or informed-consent procedures applicable.

☒ 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)?
No external or crowdsourced participants were recruited or paid; the quality-verification annotation was performed in-house by a single grad-level researcher as part of their regular research duties.

☒ 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)?
No new data was collected from people; all experiments use publicly released benchmark datasets (BoolQ, PubMedQA, LatentHatred) whose authors managed consent at the point of original release.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No new human-subjects data collection took place; the study relies on publicly released benchmark datasets and in-house research-team quality checks, so no ethics-board review was applicable.

☑ **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?
We listed the detail in the Acknowledgements section.