Responsible NLP Checklist

Paper title: *LOGICAL-COMMONSENSEQA: A Benchmark for Logical Commonsense Reasoning*
Authors: *Obed Junias, Maria Leonor Pacheco*

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. See Limitations and Ethical Considerations sections.

☑ 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?
No. The dataset consists of generic commonsense questions and answer options derived from CommonsenseQA dataset and model-generated alternatives. It does not include personally identifying information, references to individual people or offensive content. The content is abstract and non-sensitive, and does not require anonymization.

☑ 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. See Section 3 (The Logical Commonsense Dataset) and Appendix A for dataset size, relation-wise instance counts, and train/dev/test splits.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. See Section 4 and 5 (Task Formulation and Experiments) and Appendix A for model configurations, prompting settings, decoding parameters, and fine-tuning details.

☒ 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?
No. We report single-run evaluation results (accuracy and macro-F1) for each model and setting on fixed train/dev/test splits, without descriptive statistics across multiple runs.

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.

☑ **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.?
No. Annotators were given task-specific guidelines based on the awarenessconsensus framework, which are described in the paper, and the annotation guidelines provided to annotators are included in Appendix A. As the validation was conducted internally by lab annotators, full instruction texts are omitted for brevity.

☒ 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. The validation study was conducted internally by lab annotators for analysis purposes. Therefore, recruitment and payment details are not applicable and are not reported.

☒ 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 dataset is derived from the publicly available CommonsenseQA benchmark and model-generated content. No personal data from individuals was collected, and no additional consent was required.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. The work involves low-risk annotation of synthetic and benchmark-derived content and does not require ethics review board approval.

☑ **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. AI assistants were used to assist with coding tasks and for language polishing. All scientific content, design, experiments, analysis, and conclusions were produced and verified by the authors.