Responsible NLP Checklist

Paper title: *GuessingGame: Measuring the Informativeness of Open-Ended Questions in Large Language Models*

Authors: *Dylan Hutson, Daniel Vennemeyer, Aneesh Deshmukh, Justin Zhan, Tianyu Jiang*

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?
We did not identify ethical, social, or harmful risks associated with our research.

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

☑ B1. Did you cite the creators of artifacts you used?
Sec 5

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In Sec 5, we cite the dataset correctly. The dataset is under Apache license, open-source for academic use.

☑ 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)?
Sec 5. We cite the dataset correctly. The dataset is under Apache license, open-source for academic use.

☒ 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 does not contain any sensitive or private information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Sec 5

☑ 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?
Sec 5

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.

☑ **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?
We report the parameters used for Llama (70B) in section 5. We did not record our computational budget or the details of our computing infrastructure. Our method does not require significant computational resources, so with the methodology provided in the paper, anyone could reproduce the results.

N/A C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sec 5, Appendix E,F

☑ 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?
Sec 5,6

☑ 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?
Sec 4,5, Appendix D

☑ **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.?
The annotation guidelines is short, and we describe human annotation procedure in Section 6.

☑ 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)?
Sec 6

☑ 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)?
Sec 6

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We do not have an ethics review board as the annotation task does not have any ethical concern.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We used a very small number of annotators, described in Sec 6.

☒ **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?
N/A