Responsible NLP Checklist

Paper title: *Who is the richest club in the championship? Detecting and Rewriting Underspecified Questions Improve QA Performance*
Authors: *Yunchong Huang, Gianni Barlacchi, Sandro Pezzelle*

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?
This work focuses on improving QA benchmark evaluation quality by identifying and rewriting underspecified questions. We do not foresee significant misuse risks, as the method is aimed at improving scientific rigor. No sensitive data or vulnerable populations are involved.

☑ 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 data used in this study consists of samples from established, publicly available QA benchmark datasets which are primarily factual and knowledge-based in nature. Given the domain and the high-quality curation of these source datasets, the risk of encountering personally identifying information or offensive content is minimal. Our research focuses on the linguistic properties of question specification (underspecified vs. fully specified) rather than personal or sensitive topics, hence no additional filtering or anonymization steps were performed beyond using these standard research benchmarks.

☑ 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?
In Section 3.1 and Section 3.2, we report details of datasets we curated, utilized, or sampled from.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
In Section 3 and Appendix B. Our approach is prompt-based. Model selection was performed by evaluating all candidate models on UNDER and UNDER-gold datasets (Section 3.1). Experiment designs are discussed in Experimental Setup subsections throughout Section 3. Prompts we utilized are provided in full in Appendix B.

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.

☑ 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 3.2, Section 3.4 and Appendix D. We report mean F1 and Nvidia AA scores across all conditions (Table 2, Table 6), use independent t-tests to assess statistical significance, and provide violin plots showing result distributions (Figures 310).

☑ **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 was performed by the paper's authors with expertise in formal linguistics. The working taxonomy guiding annotations is provided in Table 1 and detailed in Appendix A.

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)?
Annotation was conducted by the paper's authors themselves, not crowdworkers or paid participants, so recruitment and payment information are not applicable.

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)?
Data annotation was performed by the authors. No external participants were involved, so consent procedures are not applicable.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The annotation involved the paper's authors working with publicly available academic QA datasets. No human subjects research requiring ethics review board approval was conducted.

☑ **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?
In the Acknowledgment, we pointed out that AI assistants (Claude and ChatGPT) were used for code debugging and optimization in the experiments. All scientific content, ideas, and conclusions are entirely the authors' own.