Responsible NLP Checklist

Paper title: *CLAIMCHECK: How Grounded are LLM Critiques of Scientific Papers?*
Authors: *Jiefu Ou, William Gantt Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, Benjamin Van Durme*

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 does not belief our work raises any risk concerns, in collecting the dataset, we have complied with OpenReview licensing and terms of use. Further, since both the papers and the reviews in CLAIMCHECK are anonymized, there is little concern about leakage of personally identifi- able information (PII).

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

☑ B1. Did you cite the creators of artifacts you used?
We properly cite the artifacts in Section 3

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We discuss the license for artifacts in Appendix A.1

☑ 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)?
We discuss our use of existing artifacts was consistent with their intended use in Ethical Considerations section

☒ 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?
Since both the papers and the reviews in CLAIMCHECK are anonymized, there is little concern about leakage of personally identifiable information (PII).

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We discuss the documentation of artifacts in section 3

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.

☑ 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?
We report the relevant statistics of artifacts in section 3

☑ **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?
All the model we use are proprietary models that does not have publicly available information of model size, and since we use these models through API calls, there is no GPU hours and infrastructure-related considerations

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We provide experimental setup and hyperparameters information in Appendix C

☒ 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?
Since the annotation and evaluation of performance on CLAIMCHECK requires heavy labor input with extensive domain knowledge, we are constrained by the time and monetary budgets to collect human evaluation results for multiple runs of experiments to collect descriptive statistics.

☒ 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?
The only existing package we use is PaperMage, we use its default implementation, model and parameter settings.

☑ **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.?
We provide the instructions and annotation interfaces in Section 3 and Appendix B and E.

☒ 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)?
We mentioned in Sec 3.3 that all the annotators are authors of this work

☒ 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)?
We discussed with our annotator during the annotation that how the collected data would be use and obtained their consent

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We report the characteristics of annotators in Sec 3 and Appendix B

☒ **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?
(left blank)