Responsible NLP Checklist

Paper title: *LCES: Zero-shot Automated Essay Scoring via Pairwise Comparisons Using Large Language Models*
Authors: *Takumi Shibata, Yuichi Miyamura*

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?
Limitations and Ethics Statement

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

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We used publicly available datasets and LLM APIs under their respective terms of use, but did not explicitly discuss licensing terms in the paper.

☒ 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 used public datasets and commercial LLM APIs under standard research conditions, but did not explicitly discuss intended use alignment in the paper.

☑ 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?
Ethics Statement

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We described the datasets and models used in our experiments but did not provide detailed documentation such as demographic or linguistic coverage.

☑ 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?
Section 4

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 reported the model sizes and computing infrastructure (e.g., API-based access to commercial LLMs), but did not quantify GPU hours or total computational budget.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, 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?
Due to the high computational cost of running large language models multiple times, we only report results from a single run.

☒ 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?
We used standard packages for evaluation and preprocessing (e.g., scikit-learn, pytorch), but did not report parameter settings in detail.

☒ **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.?
(left blank)

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)?
(left blank)

☒ 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 used publicly available benchmark datasets (e.g., ASAP) that are assumed to be collected with appropriate consent for research use, but we did not explicitly discuss consent in the paper.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We used publicly available benchmark datasets collected by third parties. We did not collect any new data ourselves, and therefore did not seek ethics review board approval.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We used existing benchmark datasets with human-provided scores, but the demographic or geographic characteristics of the annotators are not available in the dataset and were not reported.

☑ **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?
Acknowledgments