Responsible NLP Checklist

Paper title: *RPC-Bench: A Fine-grained Benchmark for Research Paper Comprehension*
Authors: *Yelin Chen, Fanjin Zhang, Suping Sun, Yunhe Pang, Yuanchun Wang, Jian Song, XiaoYan Li, Lei Hou, Shu Zhao, Jie Tang, Juanzi Li*

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?
Ethical Considerations

☑ **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?
Ethical Considerations

☑ 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 3.3

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.1

☒ 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 computational cost constraints, we evaluated the test-set outputs of all 28 models using two independent LLM-based judges with generation temperature fixed at 0. The final reported scores are obtained by averaging the evaluations from the two judges, which mitigates potential bias from any single judge and improves evaluation consistency. This protocol ensures stable and reproducible results under deterministic inference settings.

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.?
Appendix B.9

☑ 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)?
Appendix B.9

☑ 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)?
Appendix B.9

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix B.9

☑ **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 this work, large language models (LLMs) were used solely as generalpurpose assistive tools for language refinement. Specifically, we employed LLMs to polish the phrasing and improve grammatical correctness in the manuscript. No part of the research design, idea generation, data analysis, experimental execution, or substantive technical writing was performed by LLMs. All conceptual contributions, scientific content, and experiments were conceived, implemented, and verified entirely by the authors.