Responsible NLP Checklist

Paper title: *Beyond Explicit Refusals: Soft-Failure Attacks on Retrieval-Augmented Generation*
Authors: *Wentao Zhang, Yan Zhuang, ZhuHang Zheng, Mingfei Zhang, Jiawen Deng, Fuji Ren*

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?
Section: Ethics Statement

☑ **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?
Ethics Statement

☑ 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 5.1

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 Appendix A.6.2 (Page 6, 14)

☑ 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 5.4

☑ **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 A.2 (Table 6, Page 13). We provide the full Answer Utility Score (AUS) rubric used by our human annotators.

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.

☑ 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)?
Ethics Statement and Appendix C.6 (Page 19). We recruited four graduate students with NLP backgrounds for the study.

☑ 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 C.6 (Page 19). The annotators were recruited graduate students who were informed of the research purpose and the double-blind evaluation task.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study involved a small-scale expert evaluation (4 students) for the purpose of metric validation and did not involve sensitive personal data or vulnerable populations

☑ **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?
AI assistants were used for language editing and clarity improvement. All technical content, experimental design, and conclusions were authored and verified by the authors.