

## Responsible NLP Checklist

Paper title: *TableRAG: A Retrieval Augmented Generation Framework for Heterogeneous Document Reasoning*

Authors: *Xiaohan Yu, Pu Jian, Chong Chen*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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 F Check List in Appendix*

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

- B1. Did you cite the creators of artifacts you used?

*Section 3, 4, 5 and Section F Check List in Appendix and Reference papers*

- B2. Did you discuss the license or terms for use and/or distribution of any artifacts?

*Section F Check List in Appendix*

- 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)?

*Section C and F in Appendix*

- 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 involved in our benchmark is from Wikipedia, which does not contain information that names or uniquely identifies individual people or offensive content.*

- B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

*Section 4 Benchmark Construction, Section C Implementation Details*

- 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 Benchmark Construction*

---

*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 perform LLM inference in our experiments and analyze the statistics of the LLM calls and latency in Section 6.*

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

*Experimental Setup in Section 5.1 and Hyperparameters in Appendix Section E*

- 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 5.1 Experimental Settings*

- 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?

*Irrelevant*

**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 Section B Guidelines for Answer Annotation*

- 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 Section B Guidelines for Answer Annotation (volunteers)*

- 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 Section B Guidelines for Answer Annotation*

- 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?

*Appendix Section B.1 Annotator Profiles*

**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?

*We employ LLMs during the benchmark construction and evaluation phases, as detailed in Section 4 and Section 5.1.2. Specifically, we leverage Claude-3.7-Sonnet for query synthesis as compositions over primitives during benchmark construction and use Qwen 2.5-72B-Instruct to assess answer correctness in our experiments.*