Responsible NLP Checklist

Paper title: *HopWeaver: Cross-Document Synthesis of High-Quality and Authentic Multi-Hop Questions*
Authors: *Zhiyu Shen, Jiyuan Liu, Yunhe Pang, Yanghui Rao, Fu Lee Wang, Jianxing Yu*

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?
Our work synthesizes multi-hop questions from publicly available Wikipedia corpora. The framework itself does not introduce potential risks beyond those inherent in the source data. We discuss the possibility of inheriting biases from source corpora in the Limitations section.

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

N/A 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?
(left blank)

☑ 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?
Appendix B (Table 8) reports detailed dataset statistics including dataset size, question type distribution, average question length, and average answer length, with comparisons to baseline datasets.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix H provides comprehensive experimental settings including corpus details, generator LLMs, embedding and reranker models, coarse retrieval parameters (1, 2, 3), default generation parameters, and RAG system settings.

☑ 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?
We report evaluation scores averaged across 5 LLM judges (Section 5.1, Table 1), multi-dimensional quality scores (Figure 4), and LLM judge reliability metrics including standard deviation, Krippendorff's Alpha, and Fleiss' Kappa (Appendix F.3, Table 12).

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 C.1 describes the pairwise human validation study, including the evaluation criteria provided to the three human evaluators and the comparison setup (50 pairwise comparisons with LLM score difference > 0.3).

☒ 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)?
The three human evaluators were Master's students from our research group who participated as part of their academic research activities. They were not recruited through a formal process or crowdsourcing platform, and no monetary payment was provided for this specific validation task.

☒ 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)?
The participants were student co-authors and collaborators on this research project. They provided verbal consent to participate in the validation study. The purpose and use of their evaluation data within this paper were fully understood and agreed upon as part of the collaborative research process.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Formal ethics review board approval was not sought for this part of the study. The protocol involved a small number of expert evaluators (student collaborators) performing a low-risk, non-sensitive task of evaluating text quality. This type of internal validation activity generally does not require formal ethics board approval at our institution.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)