Responsible NLP Checklist

Paper title: *LongWeave: A Long-Form Generation Benchmark Bridging Real-World Relevance and Verifiability*
Authors: *Zikai Xiao, Fei Huang, Jianhong Tu, Jianhui Wei, Wen MA, Yuxuan Zhou, Jian Wu, Bowen Yu, Zuozhu Liu, Junyang Lin*

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?
The work focuses on creating a benchmark for evaluating existing language models. It does not introduce new models or applications with direct societal risks, and its primary purpose is to identify and measure technical limitations in long-form generation.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 2.3 cites the source (QreCC) for the original texts used in the Paragraph Reordering task.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The paper provides a public GitHub link in the abstract where the code and data are available, which includes license information.

☑ 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 2.3 explains that texts from QreCC documents were used for the Paragraph Reordering task. This is consistent with the original artifact's purpose as a source of coherent documents.

☒ 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 benchmark data is primarily synthetically generated using rule-based scripts. For the Paragraph Reordering task, public documents from the QreCC dataset were used. No new data was collected from human subjects, and the generation process does not involve personal or sensitive information.

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2 describes the overall pipeline and individual tasks in detail. The Appendix provides further examples and details on the data generation process for each task.

☑ 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 2.4 and Figure 3 provide statistics on input length distribution. Section 3.2 specifies the number of test samples used for the evaluation (200 samples per variant, totaling 5,600 per model).

☑ **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?
Section 3.1 lists the models evaluated. Appendix F ("Evaluation Efficiency") and Appendix Table 10 provide details on the computing infrastructure (NVIDIA A100 GPUs) and model backends.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.1 and Appendix Table 10 describe the experimental setup, including the models, deployment backends (vLLM, API), and specific decoding parameters used for inference.

☑ 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 4.1 and Table 4 report mean performance and standard deviations from a stability analysis to demonstrate the reliability of the benchmark's metrics. The main results in Table 3 are reported as average scores across multiple runs.

☑ 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?
Section 2.3 mentions the use of the Flake8 toolkit for the Code Fixing task, which is a standard, rule-based linter.

☒ **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)

N/A 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)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(left blank)

☑ **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?
The use of AI assistants is mentioned in the methodology. Specifically, Section 2.3 states that an LLM (GPT-4o) was used to generate news topics and statements for the AP Style News Writing task. Section 3.4 notes that an LLM was used for initial labeling of failure patterns, which were then manually verified.