

Responsible NLP Checklist

Paper title: *Benchmarking Agentic Newswriting via Journalistic Workflows*

Authors: *Yen-Che CHIEN, Kuang-Da Wang, Wei-Yao Wang, Wen-Chih Peng*

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?
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?
The benchmark is constructed from publicly available English-language news articles from BBC and APNews. We do not collect private user data or conduct new data collection involving personally identifying information beyond what is already publicly available in the source articles. Because the benchmark is derived from published news content, we do not perform additional anonymization.

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 3.1.2, Section 3.1.3

C. Did you run computational experiments?

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1, Section 4.1.3, Appendix B.2

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 3.3, Table 2, Section 4.2, Appendix C

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.?

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.

We summarize the annotation task and evaluation criterion in Appendix C, but we do not include the full verbatim annotator instructions or screenshots. The validation study was a small-scale internal preference annotation used to assess evaluator reliability rather than a full human-subject experiment.

- 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 paper does not report formal recruitment or payment information for annotators. The human validation was conducted as a small-scale internal annotation study for evaluation reliability rather than a large-scale paid annotation effort.

- 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)

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