Responsible NLP Checklist

Paper title: *Not Lost After All: How Cross-Encoder Attribution Challenges Position Bias Assumptions in LLM Summarization*
Authors: *Elahe Rahimi, Hassan Sajjad, Domenic Rosati, Abeer Badawi, Elham Dolatabadi, Frank Rudzicz*

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?
This work presents a methodological improvement for analyzing summarization systems and does not pose identifiable risks to individuals or society that require discussion.

☑ **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 and References. We cite creators of all datasets used (Hermann et al. for CNN/DailyMail, Narayan et al. for XSum, Gliwa et al. for SAMSum, etc.) and the cross-encoder model.

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
(left blank)

☑ 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 3 and methodology sections. We used all datasets for their intended summarization evaluation purposes and employed the cross-encoder model for semantic similarity assessment as designed.

☒ 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?
We used established academic datasets that have been previously vetted and are widely used in the research community without additional PII or offensive content concerns.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix E.1 (Dataset Statistics). We provide comprehensive documentation including domains (news, dialogue, scientific, government), sample sizes, document lengths, and linguistic characteristics.

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.

☑ 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?
Table in Appendix E.1 and throughout results sections. We report sample sizes, document lengths, and detailed statistics for all datasets used.

☑ **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?
ppendix E.2 and E.6. we report model parameters and mention A40 GPUs with 48GB memory.

N/A C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
(left blank)

☑ 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?
throughout results sections and tables. we report means, standard deviations, confidence intervals, and statistical significance.

☒ 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?
While we used standard implementations of evaluation metrics (ROUGE, BERTScore, G-EVAL), we focused on methodological comparisons rather than absolute metric values, making specific package versions less critical to reproducibility of our core findings.

☒ **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?
AI assistance was limited to minor writing improvements and was not considered substantial enough to warrant formal citation in the manuscript.