Responsible NLP Checklist

Paper title: *PoSum-Bench: Benchmarking Position Bias in LLM-based Conversational Summarization*
Authors: *XU SUN, Lionel Delphin-Poulat, Christle Tarnec, Anastasia Shimorina*

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?
at Page 9, Ethics Statement , where we delcare the potential risk

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

☑ B1. Did you cite the creators of artifacts you used?
Section 3.1 (Data Collection) cites all original sources of datasets used in our benchmark.

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Our paper focuses on the benchmark methodology rather than licensing details. We use only publicly available datasets that have been previously published in peer-reviewed venues, and our derived benchmark will be made openly accessible to researchers.

☑ 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 3.1 describes our use of existing conversational datasets, which is consistent with their intended research purposes. In our Ethics Statement (at the end of the paper), we specify that our benchmark will be openly accessible for research purposes

☑ 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?
Section 5 (Ethics Consideration) explicitly addresses this. We exclusively used established benchmark datasets that have already undergone anonymization procedures by their original authors before public release. Our work did not collect any new data from human subjects.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.3 (Dataset Preparation) and Table 1 provide comprehensive details about the conversational corpora in our benchmark, including domain types, languages, and conversation characteristics.

☑ 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 1 in Section 3.3 reports comprehensive statistics including number of instances, average word counts, and average turns for each dataset in our benchmark

☑ **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 3.4 (Summary Generation) describes the models used, including Qwen2.5B-instruct in various sizes (1.5B, 3B, 7B, 14B), Google Gemma3-instruct (1B, 4B), and other models with their parameter sizes and Appendix A declared which GPUs we used and how many GPU-hours cost to run the experiments.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 (PoSum-Bench Methodology) details our experimental setup, including similarity calculation methods, thresholds, and sampling approaches. Appendix D further validates our methodology.

☑ 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?
Tables 2-3 in Section 5 report detailed statistics on positional bias across different models, languages, and text lengths, showing percentages for leading, recency, and neutral bias types.

☑ 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?
Appendix A describes our use of sentence transformer models (sentence-transformers/all-MiniLM-L6-v2) for semantic similarity calculations

☒ **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?
As mentioned in our Ethics Statement, Claude 3.7 was used only for text polishing in manuscript preparation.