

Responsible NLP Checklist

Paper title: *An Empirical Study of Position Bias in Modern Information Retrieval*

Authors: Ziyang Zeng, Dun Zhang, Jiacheng Li, zoupanxiang, Yudong Zhou, Yuqing Yang

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?

No, because the work does not involve human subjects, animals, hazardous materials, or sensitive data, and therefore presents no significant risks.

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

☒ B1. Did you cite the creators of artifacts you used?

I

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?

We did not discuss separate license terms. The dataset is already distributed under the CC BY-SA 4.0 license and the Open Data Commons Attribution License (ODC-By v1.0). We also cited the relevant papers.

☒ 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)?

We did not discuss this separately, since our use of the artifacts is solely for scientific research. Our intended use is fully consistent with the conditions under which the original artifacts were made available.

☒ 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 did not conduct additional checks, since our work did not introduce any new data beyond the original dataset. The dataset we used does not contain personally identifiable information or offensive content, so no further anonymization steps were required.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

I

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?

Appendix A

☒ **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?

We did not explicitly report the number of parameters, total computational budget, or specific computing infrastructure. Instead, we provide references to the original model descriptions as well as the complete evaluation code to ensure reproducibility.

- ☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Our experiments did not involve hyperparameter search. We used fixed settings without tuning, so no best-found hyperparameter values are reported.

- ☒ 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?

Our results are easily reproducible through the released code, so a single run is sufficient to replicate the reported outcomes.

- ☐ 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?

(left blank)

☒ **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.?

(left blank)

- ☐ 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)

- ☐ 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)

- ☐ 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?

We used ChatGPT only for minor syntactic polishing of the text. It was not involved in designing experiments, analyzing results, or generating substantive content.