

Responsible NLP Checklist

Paper title: *GATEAU: Selecting Influential Samples for Long Context Alignment*

Authors: *Shuzheng Si, Haozhe Zhao, Gang Chen, Yunshui Li, Kangyang Luo, Chuancheng Lv, Kaikai An, Fanchao Qi, Baobao Chang, Maosong Sun*

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.

N/A A2. Did you discuss any potential risks of your work?

Our work is designed to select influential samples for long context alignment, thus there is no potential risks for research communities.

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

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

We cite the related papers in the Section References and the Appendix.

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

We provide the related information in the Appendix. Meanwhile, we follow the licenses to use the open-source LLMs.

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 provide the related information in the Appendix. Our use of existing artifact(s) are consistent with their intended use.

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?

Our collected data is from open-source communities that do not contain any information that names or uniquely identifies individual people.

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

We provide the detailed introduction of our used datasets in the Appendix.

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?
We provide the detailed introduction of our used datasets in the Appendix.
- 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 provide the detailed introduction of model size and budget in the Appendix and Section Experiments.
- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We provide the detailed introduction of experimental setup and hyperparameters in the Appendix and Section Experiments.
- 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?
We provide the detailed introduction of descriptive statistics in the Appendix and Section Experiments.
- 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?
We provide the detailed introduction of packages in the Appendix and Section Experiments.
- 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.?
We provide the full text of instructions in the Appendix.
- 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)?
We provide the such information in the Appendix.
- 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)?
We provide the such information in the Appendix. Meanwhile, we will open-source our selected training data.
- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our selected and used data is from open-source communities, thus there is no ethics issue.
- D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We provide the such information in the Appendix.
- 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 use AI assistants in our experiments for comparisons, e.g., GPT-4. We provide the related information in the Appendix and Section Experiment.