Responsible NLP Checklist

Paper title: *Following Length Constraints in Instructions*
Authors: *Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason E Weston, Jing Xu*

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?
We discuss limitations and potential risks in the Section 8

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

☑ B1. Did you cite the creators of artifacts you used?
Yes we cite all artifacts used in the paper (e.g. Llama2, Llama3, Open Assistant, MT-Bench, AlpacaEval, IFEval) in Section 1-6.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We didn't explicitly discuss the terms or license for use of the existing datasets as they are well-known research artifacts and are used in our work according to their intended terms of use. We cite the existing works to offer readers access to license and terms of use on datasets used in our work.

☑ 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)?
To the best of our knowledge, the existing datasets and models are used consistent with their license and 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?
We didn't explicitly discuss the steps taken to check whether the data that was collected/used contains personally identifying info or offensive content, as they are well-known research artifacts and are used in our work according to their intended terms of use. We cite the existing works to offer readers access to license and terms of use on datasets used in our work.

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.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We didnt collect any new data for finetuning in this work. For existing datasets/models used in our work, we cite the original paper to provide readers access to further details about the construction of those artifacts.

☑ 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?
Yes we report the statistics of our datasets in Section 5.

☑ **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?
Yes we report our model size and budget in Section 5.2.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes we discuss our experiment setup and hyperparameter choices in Section 5.2 and Appendix.

☑ 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?
Yes we report descriptive statistics about our results in Section 6, and we specify whether it's single run or not in Section 3.

☑ 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?
Yes we report the use of existing packages in Appendix.

☒ **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.?
We don't use human annotators in the paper.

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)?
We don't use human annotators in the paper.

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)?
We don't use human annotators in the paper.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We don't use human annotators in the paper.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
We don't use human annotators in the paper.

☒ **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?
Yes we include the use of AI assistants in Section 3.