Responsible NLP Checklist

Paper title: *Temporal Scaling Law for Large Language Models*

Authors: *Yizhe Xiong, Xiansheng Chen, Xin Ye, Hui Chen, Zijia Lin, Haoran Lian, Zhenpeng Su, Wei Huang, Jianwei Niu, Jungong Han, Guiguang Ding*

> 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?
Our paper primarily proposes a novel Temporal Scaling Law for pre-training Large Language Models (LLMs). Currently, we are attentive to any potential risks.

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

☑ B1. Did you cite the creators of artifacts you used?
The Pile dataset: Section 2.2. The C4 dataset: Section 2.2. The LAMBADA and WikiText datasets: Section 2.3. The evaluation benchmarks: Section 3, Appendix Section B.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We discuss the license of all used artifacts in Appendix Section B.

☑ 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)?
As discussed in Appendix Section B, all usages of scientific artifacts obey the corresponding licenses.

☒ 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?
To the best of our knowledge, the Pile, the C4, the LAMBADA, and the WikiText are popular datasets for LLM pre-training, and lm-evaluation-harness is a well-known evaluation tool for LLMs. According to their reports, this step is thoroughly applied to the large-scale datasets before release.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The Pile dataset: Section 2.2. The C4 dataset: Section 2.2. The LAMBADA and WikiText datasets: Section 2.3. The evaluation benchmarks: Section 3, Appendix Section B.

☑ 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?
The Pile dataset: Section 2.2. The C4 dataset: Section 2.2. The LAMBADA and WikiText datasets: Section 2.3. The evaluation benchmarks: Section 3, Appendix Section B.

☑ **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 report the information of the models used in Appendix B. However, we did not report GPU hours as we believe it could potentially leak sensitive information about the company's computational resources.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Experiment Setup in Section 2.2 and Appendix B.

☑ 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?
Section 2.3, 2.4, 3, 4, and Appendix Section B have thoroughly described the statistics of our results.

☑ 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?
Experiment Setup in Appendix B.

☒ **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?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)