Responsible NLP Checklist

Paper title: *UniGeM: Unifying Data Selection and Mixing via Geometric Exploration and Mining*
Authors: *Changhao Wang, yunfeiyu, Xinhao Yao, Jiaolong Yang, Lu Yu, JunpengFang, Chaobo Li, Riccardo Cantoro, Qing Cui, JUN ZHOU*

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?
No. We do not introduce a new deployment scenario; we focus on offline data curation for pretraining. General risks of LLMs (bias, misuse, privacy leakage) remain and are not uniquely introduced by our method.

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

☒ 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 use existing large-scale pretraining corpora (e.g., The Stack Dedup and Common Crawl) and do not collect new human-subject data. We did not conduct an additional dedicated audit for PII or offensive content beyond the standard preprocessing/filters provided by the source datasets and our benchmark decontamination.

☑ 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?
Dataset composition, token counts, corpus mixing ratios, and training data statistics are described in Section 3.1 and Section 3.2.

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup, model architectures, training configurations, and hyperparameter values are described in Section 3.23.4, with additional details provided in Appendix E.

☑ 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 report aggregated performance metrics across multiple benchmarks, compare different training

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. ACL 2026 used a subset of ARR checklist form.

durations (0.5 vs. 1.0 epochs), and provide ablation studies and scaling analyses in Section 4, making it explicit which values correspond to single runs and comparative settings.

☒ **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)

☑ **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?
AI assistants were used for minor language polishing and clarity improvements during the writing process. All scientific ideas, theoretical formulations, experimental design, implementation, results, and conclusions were developed and verified solely by the authors.