Responsible NLP Checklist

Paper title: *RoZO: Geometry-Aware Zeroth-Order Fine-Tuning on Low-Rank Adapters for Black-Box Large Language Models*
Authors: *Zichen Song, Weijia Li*

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?
See Section Limitations (p. 8). The paper explicitly discusses risks such as scalability challenges for very large models, limited task coverage, and sensitivity to hyperparameters

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

☑ B1. Did you cite the creators of artifacts you used?
See Section Related Work (p. 23). The paper cites prior methods (LoRA, MeZO, LOZO, etc.) whose artifacts were used for comparison

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The paper does not mention licenses of external code or datasets. The baselines are standard published methods, but license details are not discussed.

☑ 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)?
Section Experiments (p. 57). All baseline artifacts (datasets and models) are used for research purposes in line with their intended use

N/A 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?
(left blank)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section Experiments and Hyperparameters (p. 57). The paper reports which datasets and tasks are covered, and specifies evaluation setups

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?
Section Experiments (p. 57). Reports number of training samples (e.g., k = 16, k = 512, 1000 examples per task) and train/test evaluation splits

☑ **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?
Section Experiments (p. 57) and Hyperparameters (p. 67). Reports model scales (RoBERTa-large, OPT-13B, OPT-30B, OPT-66B, Cydonia-24B) and GPU infrastructure (A100 40GB)

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section Hyperparameters (p. 67). Full details on perturbation radius, LoRA rank, learning rates, batch size, epochs, momentum, adaptive settings

☑ 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 Experiments (p. 57). Results are averaged across three random seeds, and descriptive statistics (accuracy, memory usage, etc.) are provided

☒ 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?
The paper does not explicitly document preprocessing package settings (e.g., tokenizers, libraries). It assumes standard benchmark implementations.

☒ **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?**

☑ E1. If you used AI assistants, did you include information about their use?
help to write