Responsible NLP Checklist

Paper title: *Parameter-Efficient Routed Fine-Tuning: Mixture-of-Experts Demands Mixture of Adaptation Modules*

Authors: *Yilun Liu, Yunpu Ma, Yuetian Lu, Shuo Chen, Zifeng Ding, Volker Tresp*

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?
In the Limitations section we discussed the potential risks including bias, societal risks, and environmental impact.

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

☑ B1. Did you cite the creators of artifacts you used?
In section 3 we cited creators of all major artifacts used including the 2 opensource LLMs and all 14 benchmark datasets used in evaluation.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
All models and datasets are licensed under Apache 2.0 license, which permits research and educational use consistent with this study's methodology.

☑ 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)?
In section 3 we elaborated our usage of the mentioned opensource models and datasets, which aligns with their licensed usage for research and educational purposes.

☑ 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?
In the Limitations section we acknowledge that the training data may contain demographic and geographic skews inherited from web corpora. We used well-established, publicly available models (OLMoE-1B-7B, Mixtral-87B) and benchmark datasets that are widely adopted in NLP research. Following standard practice in parameter-efficient fine-tuning research, we utilized these artifacts as-is for educational and research purposes only. The models and datasets have undergone the data curation processes implemented by their original creators, and our work focuses on methodological improvements to fine-tuning techniques rather than data collection or curation.

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.?
All opensource models and datasets have comprehensive documentation provided in their corresponding papers which we have cited in section 3.

☑ 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 3.1 and Appendix A provide relevant statistics including the use of Commonsense170K and Math50K training sets, evaluation across 14 tasks, and details about model sizes and parameters.

☑ **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?
In Section 3.1 and Appendix A.1 we report model parameters (active / trainable / total parameters) and computing infrastructure (single NVIDIA A100 GPU for OLMoE, 4NVIDIA H100 GPUs for Mixtral).

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.1 describes the experimental setup including datasets, baselines, and evaluation methodology. Appendix A.1 and Table 2 provide detailed hyperparameter configurations including learning rates, batch sizes, epochs, and optimizer settings for both models.

☑ 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 3 and all major results report average performance calculated across individual benchmark test sets (8 commonsense reasoning and 6 arithmetic reasoning tasks), following established evaluation practices in the field. Appendix B provides statistical significance analyses.

☑ 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?
Section 3.1 mentions using benchmark suites and evaluation frameworks from Hu et al. (2023) , and Appendix A.1 mentions specific training frameworks.

☒ **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)