Summarization of change for the paper: “Assessing the Macro and Micro Effects of Random Seeds on Fine-Tuning Large Language Models”-Submission #356

We have updated our papers based on the previous reviewers’s comments which can be summarized as follows:

- **Limited number of random seeds**: We have increased the number of random seeds from 5 to 10 and rerun all experiments. The updated experimental setting is described in Section 4.2 and detailed setting is reported in Table 5 and Table 6 in the Appendix. The new results and analysis are presented in Section 5, with additional results shown in Table 7 and Table 8 in Appendix.
- **Mainly focus on classification tasks**: We have revised the paper to clarify that our contributions primarily target classification tasks. To illustrate the extensibility of our framework, we include a summary of standard evaluation metrics and the possible corresponding consistency metrics for various NLP task types in Table 2 (Appendix). A further discussion on how this summary can guide mitigation of seed-induced variability is provided in Appendix A.3. We hope this offers valuable context and facilitates future extensions of our framework to other task categories.
- **Only experimenting with one LLM, should demonstrate the impact of random seeds on larger language models**: We conducted additional experiments to study the impact of random seeds on a larger model, LLaMA 3.2B. These results are now included in the paper.
- **Lack of discussion about randomness from other sources**: We added a brief discussion in Section 1 covering other sources of randomness, including prompt formatting [1], in-context learning [2], and model initialization [3].
- **Lack of discussion about how to mitigate seed-induced variability**: We added Section 5.3 to discuss possible approaches to mitigate seed-induced variability, including prior work focused on reducing macro-level variation, and potential strategies motivated by our findings for addressing both macro and micro variability. We also mentioned that this will be the future work to explore.

[1] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? arXiv preprint arXiv:2411.10541. 378
[2] Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023. Coverage-based example selection for in context learning. arXiv preprint arXiv:2305.14907.
[3] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. The impact of initialization on lora finetuning dynamics. Advances in Neural Information Processing Systems, 37:117015–117040.