## Rebuttal PDF for Paper #4226

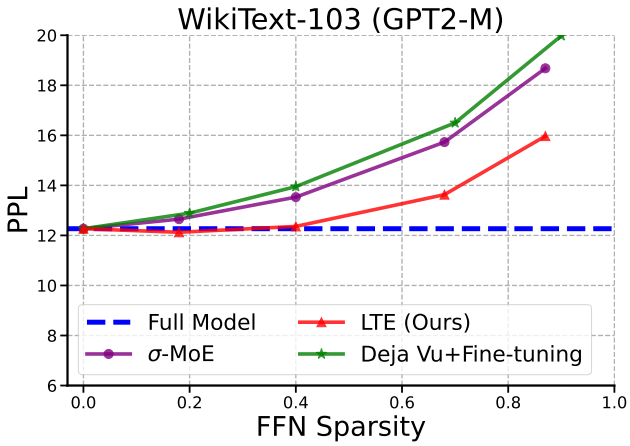


Figure 1: Performance comparison with new baselines suggested in reviews: 1) $\sigma$-MoE [1] and 2) Deja VU [2] with further fine-tuning. We evaluate two baselines by fine-tuning GPT2 medium with WikiText103. For $\sigma$-MoE, we implement it for MoEfication and train for the same training time as we train LTE models. For Deja Vu + Fine-tuning, we first fine-tune the GPT2-M model and then apply Deja Vu. Then, we further fine-tune the MoEfied models. As illustrated in the figure, LTE consistently outperforms the other two baselines across various levels of sparsity.

Table 1: Performance comparison with model pruning. We apply Wanda [3] on WikiText103 fine-tuned LLaMA-7B with the author-implemented code. Besides using C4 as the calibration data suggested by the Wanda paper, we also try to use WikiText for calibration to improve this baseline. The sparsity reported here is the overall sparsity of the entire model, which is different from the FFN sparsity reported in the paper. For LTE, we recalculate the overall sparsity based on the FFN sparsity. The evaluation results show that LTE achieves a lower perplexity than Wanda.

| Method | Wanda (2:4) | Wanda (4:8) | Wanda (Unstructured) | Wanda (2:4) | Wanda (4:8) | Wanda (Unstructured) | LTE ($\eta = 2$) |
|---|---|---|---|---|---|---|---|
| Overall Sparsity | 50% | 50% | 50% | 50% | 50% | 50% | 52% |
| Calibration Data | C4 | C4 | C4 | Wiki | Wiki | Wiki | - |
| PPL | 11.45 | 8.39 | 7.04 | 10.82 | 8.17 | 6.87 | **5.95** |

## References


[1] Róbert Csordás, Kazuki Irie, and Jürgen Schmidhuber. Approximating two-layer feedforward networks for efficient transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 674–692, 2023.

[2] Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR, 2023.

[3] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.