A Critical Review of GraLoRA (Jung et al., 2025):

Granular Low-Rank Adaptation for Parameter-Efficient Fine-Tuning

Mikaela Connell

CSCI E-222 Foundations of Large Language Models

Harvard Extension School

May 2026

1. Introduction

Low-Rank Adaptation (LoRA) has become a common approach to parameter-efficient fine-tuning (PEFT) of large language models. In "GraLoRA: Granular Low-Rank Adaptation for Parameter-Efficient Fine-Tuning," the authors introduce Granular Low-Rank Adaptation (GraLoRA) as an extension of LoRA. LoRA is often used due to the simplicity of the method, as it freezes the original model weights and trains only small low-rank adapter matrices. Despite the efficiency of LoRA, as argued by Jung et al., LoRA has a structural limitation: higher rank does not always improve performance, and a persistent gap to full fine-tuning (FFT) remains. Jung et al. aim to address this gap through proposing GraLoRA, which divides weight matrices into sub-blocks and assigning each block its own low-rank adapter, demonstrating consistent gains over LoRA across five evaluation domains.

This review argues that GraLoRA is a strong method, as it identifies a specific failure mode in LoRA and suggests an architectural fix. However, this method depends on hyperparameter choices (the rank r and partition factor k), an outlier analysis demonstrated on a single layer of one model, and an empirical evaluation reported without error bars. The remainder of this critical review summarizes the problem, method, and results (Section 2), critically analyzes strengths, limitations, and open questions (Section 3), and connects the paper to core concepts from CSCI E-222 (Section 4).

2. Summary of the Paper

2.1 Problem

The motivation behind PEFT methods, and LoRA specifically, is that full fine-tuning of large generative models is computationally expensive. LoRA freezes the pre-trained weights and

learns a low-rank update R = sBA^T with rank r much smaller than the matrix dimensions. This results in a reduction of trainable parameter count, but the rank-r bottleneck is the cause of a performance gap compared to FFT. The initial response to increase r, but simply increasing r leads to a plateau or decrease in accuracy beyond r ≈ 32–64.

The authors' diagnosis of the root cause is channel dominance in the gradient: a small number of outlier input channels with abnormally high activations steer the gradient of the adapter matrix B. The authors demonstrate this concretely in the down-projection matrix of Layer 1 of LLaMA3.1–8B, where channels 198 and 2427 dominate, and show that the gradient deviation between LoRA and FFT widens as rank increases. This reframes LoRA's rank ceiling as a structural problem in gradient propagation.

2.2 Method

GraLoRA splits each weight matrix into smaller block-wise low-rank adapters, aiming to localize gradient updates and reduce distortion from outlier channels. This works by partitioning the weight matrix into a grid of k × k independent blocks, each cell of the grid with its own local low-rank adapter. For example, k = 2 yields 4 small LoRA in place of one, and k = 4 yields 16. When k = 1, there is one block covering the whole matrix, which is the standard LoRA. Therefore, GraLoRA generalizes LoRA. The full update matrix R is the concatenation of k^2 block-wise products $B_{i,j} A_{i,j}^T$, where each block-level adapter has rank r/k.

The total parameter count stays the same as LoRA. This is because standard LoRA has (M + N) × r parameters, and GraLoRA's k^2 blocks each with (M/k + N/k) × (r/k) parameters expand to k^2 × (M/k + N/k) × (r/k) = (M + N) × r, which is the same count. The core argument for why partitioning works is that each block-level LoRA only handles a slice of the input and a slice of

the output, so outlier influence is confined to the k blocks that intersect the outlier column rather than propagating through the entire adapter.

2.3 Main Results

Jung et al. report improvements over LoRA, MoRA, and RaSA across five domains: code generation (LLaMA3.1–8B on HumanEval+), commonsense reasoning (across LLaMA and Qwen models at multiple scales), mathematical reasoning, general language understanding (RoBERTa-base on GLUE), and personalized image generation (SDXL). The headline result is a +8.5% absolute gain in Pass@1 on HumanEval+ at rank 128 over LoRA. On commonsense reasoning, GraLoRA wins 26 of 32 task-model cells. On GLUE, the best GraLoRA variant averages 86.0% versus LoRA's 84.2%, matching or exceeding full fine-tuning (85.2%) on most sub-tasks.

The most consequential pattern is not the absolute gain at any single rank but the trajectory: LoRA's Pass@1 on HumanEval+ peaks at rank 32 and falls at rank 128, while GraLoRA continues to improve with higher ranks, reaching 64.3% at rank 128. This is direct empirical evidence for the paper's central claim that GraLoRA removes the structural ceiling that prevents LoRA from benefiting from higher ranks.

3. Critical Analysis

3.1 Strengths

The article's main strength is its contribution to the cross-application of outlier-channel analysis from quantization to fine-tuning. The impact of outliers was first studied in quantization, with methods like SmoothQuant that reduce model precision (Xiao et al., 2024). Related to that,

outliers result in large errors when activations are compressed to lower bit-widths. Jung et al. are among the first to show that the same channels distort the gradient dynamics of LoRA. They acknowledge this transfer: "while the negative impact of outliers has been well recognized in the context of quantization, their influence on LoRA's behavior has not been systematically studied until now." This reframes a known phenomenon as the explanation for a previously unexplained empirical observation (LoRA's rank ceiling).

3.2 Limitations

One limitation of the paper is the outlier-channel analysis. This is demonstrated primarily in one specific location: the down-projection matrix of Layer 1 in LLaMA3.1–8B (Figures 3a, 4, 6). Figure 3a does show that the outlier signature is concentrated in Layer 1 and is much milder in the surrounding layers. This raises the question of whether GraLoRA's mechanism is doing the work the authors claim throughout the network or whether it is primarily fixing one bad layer and the rest of the gains come from increased effective rank. The paper would be strengthened by a layer-wise ablation isolating these two effects, or by demonstrating outlier dynamics on a model family with a different activation profile (for example, Mistral or Gemma).

3.3 Open Questions

Several questions follow from the analysis above. First, would adaptive or learned partitioning, where k is chosen per layer or per matrix based on the local outlier profile, deliver further gains? The paper's own Layer 1 emphasis suggests it should. Second, how does GraLoRA interact with advanced initialization methods (PiSSA, LoRA-GA, EVA)?

4. Connections to Course Concepts

This paper is connected to course content from CSCI E-222. First the connection to LoRA (Hu et al., 2022), the required reading from Lecture 8 (Fine-Tuning Pre-Trained Models). GraLoRA preserves LoRA's core insight, that the original weights can remain frozen, but replaces the single low-rank decomposition with a partitioned one. Working through the GraLoRA expansion makes the LoRA parameter accounting concrete: it is because $k^2 \times (M/k + N/k) \times (r/k)$ collapses back to $(M + N) \times r$ that the method is free in parameter count.

A second connection is to the transformer architecture itself. GraLoRA is applied to all four attention projection matrices (W_q, W_k, W_v, W_o) and all three feed-forward matrices (W_up, W_down, W_gate). The paper's gradient analysis specifically singles out the down-projection matrix as the location of the most severe outliers, which connects to the course discussion of how MLP layers behave differently from attention layers during fine-tuning and what role they play in storing factual knowledge. The $k \times k$ partitioning is, in effect, a way of imposing a coarse spatial structure on what are otherwise dense attention and feed-forward updates.

5. Conclusions

GraLoRA is a strong extension of LoRA because it identifies and addresses a structural weakness in the standard low-rank adaptation. By partitioning weight updates into block-wise adapters, GraLoRA reduces the influence of outlier channels and allows higher-rank adaptation to become more useful. Overall, GraLoRA is a valuable contribution to PEFT research because it shows that efficient fine-tuning is not only about reducing parameter count but also about designing adapter structures that better match the gradient behavior of full fine-tuning.

6. References

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171–4186). https://arxiv.org/abs/1810.04805

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 2790–2799).

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2106.09685

Jung, Y., Ahn, D., Kim, H., Kim, T., & Park, E. (2025). GraLoRA: Granular low-rank adaptation for parameter-efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS 2025)*.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2024). SmoothQuant: Accurate and efficient post-training quantization for large language models. arXiv:2211.10438.