Rethinking the Role of Tensor Decompositions in Post-Training LLM Compression

Artur Zagitov [*1 2 3] **Alexander Miasnikov** [*1 2 3] **Maxim Krutikov** [1 2] **Vladimir Aletov** [1 2 4] **Nail Bashirov** [1 2]
Gleb Molodtsov [1 2] **Artem Tsedenov** [1 2] **Aleksandr Beznosikov** [1 2 3]

Abstract

Post-training compression is essential for deploying large language models (LLMs) under tight resource constraints. Tensor decompositions have emerged as a promising direction, offering compact parameterizations well suited to Transformer weight structures. However, existing studies evaluate these methods in narrow settings, leaving unclear whether tensorization is effective at large-scale deployment. We systematically evaluate tensor compression across dense and MoE architectures, establishing performance trade-offs grounded in both empirical analysis and theoretical derivation. We identify a fundamental mismatch between the shared subspaces assumed by tensor decompositions and the heterogeneous representations learned by modern LLMs, thereby delineating their practical limits and clarifying their viable role in large-scale deployment. The code is available at `https://github.com/brain-lab-research/TT-LLM`.

1. Introduction

In recent years, large language models (LLMs) have grown considerably in scale, increasing the storage and deployment costs and limiting their applicability in resource-constrained settings. Consequently, compression techniques are widely used to improve efficiency while preserving quality.

The primary goal of model compression is to reduce redundancy while preserving the model's functional behavior. Standard approaches include pruning, which removes redundant components to decrease model size (Frantar & Alistarh, 2022; Ma et al., 2023; Gromov et al., 2024); quantization, which represents weights and activations with lower-precision data types (Frantar et al., 2022; Chee et al., 2023; Egiazarian et al., 2024; Lin et al., 2024; Zagitov et al., 2026); and knowledge distillation (KD), which trains a compact student model to approximate the behavior of a larger teacher model (Hinton et al., 2015; Tan et al., 2023). However, achieving state-of-the-art performance with these techniques typically requires extensive fine-tuning or data-driven calibration.

A natural alternative to the structural methods is matrix and tensor decompositions, which are appealing due to their established theoretical background. Methods in this category decompose a dense layer into a product of smaller factors, achieving parameter reduction while approximately maintaining the functionality of the original layer. For matrices, truncated singular value decomposition (SVD) provides an optimal low-rank approximation in the Frobenius and spectral norms (Eckart & Young, 1936; Mirsky, 1960), while tensor decompositions extend this idea to multi-dimensional weight tensors – multi-head attention (MHA) (Vaswani et al., 2017) and mixture of experts (MoE) (Fedus et al., 2022a).

However, existing literature on tensor decompositions (Gu et al., 2025; Li et al., 2026) reports positive results under evaluation protocols that do not reflect full-scale deployment constraints, leaving the practical utility of tensor decompositions unclear. We close this gap with a systematic study of tensor-based compression across realistic LLM settings, covering both dense and MoE architectures. We complement our empirical findings with a theoretical analysis that explains why standard decompositions fail to scale.

2. Compression Strategies

2.1. Pruning

Pruning removes parameters or components that contribute least to the model's behavior. It is done by exploiting the highly non-uniform parameter redundancy inherent in LLMs. Notably, both prior research (Gromov et al., 2024) and our own experiments (Figure 1) demonstrate that intermediate layers have relatively little impact on final quality and can tolerate aggressive pruning. This phenomenon

[*]Equal contribution [1]MIRAI, Moscow, Russia [2]BRAIn Lab, Moscow, Russia [3]Innopolis University, Innopolis, Russia [4]HSE, Moscow, Russia. Correspondence to: Nail Bashirov <bashirov.n@miriai.org>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. 2026.

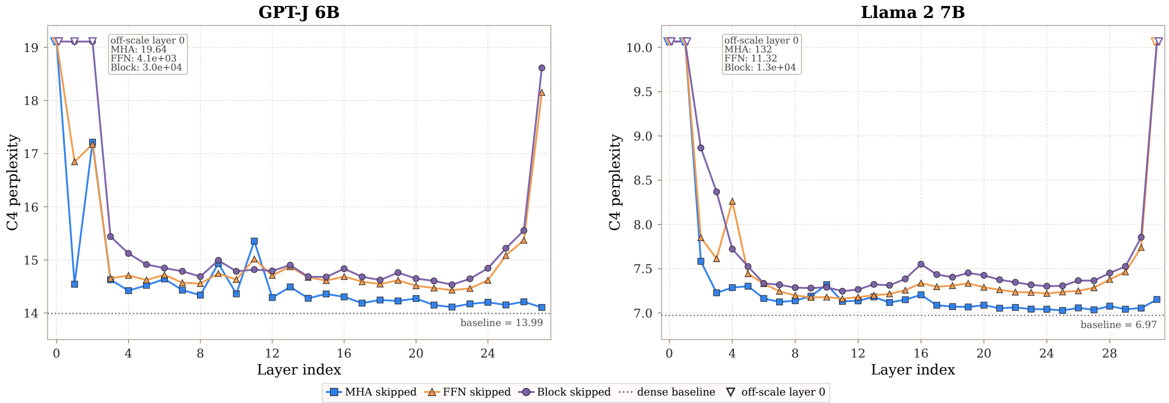

Figure 1. Perplexity (PPL) over pruning. Displays the final PPL over each layer pruned.

arises because the earliest and latest layers disproportionately handle critical tasks, such as forming token representations and generating final predictions, while the intermediate layers play a comparatively smaller role. However, while pruning effectively discards these middle-layer tails, it limits the maximum achievable compression ratio due to the dominant mass of the parameters intact. It does not compactly reparameterize the remaining structure, making it a naïve baseline for more structurally advanced decompositions.

2.2. Matrix decompositions

A natural next step is matrix factorization, which compresses weights via low-rank reparameterization rather than deletion. For Feed-Forward Networks (FFNs) – which contain the majority of model parameters (Geva et al., 2021) – we apply LASER-style truncated SVD (Figure 3) (Sharma et al., 2024).

The fundamental problem of the low-rank matrix approximation is to find a matrix $\widehat{X}$ of restricted rank r that closely approximates a target matrix X. Formally, this is expressed as the optimization problem:

$$\min_{\operatorname{rank}(\widehat{X})\leq r} \left\| X - \widehat{X} \right\|_{\alpha}. \tag{1}$$

According to the Eckart-Young-Mirsky theorem (Mirsky, 1960), for any unitarily invariant norm $\|\cdot\|_{\alpha}$ (e.g., Frobenius or spectral) the global optimum to (1) is given by the truncated SVD.

The SVD factorizes the original matrix into a product of three components:

$$X = U\Sigma V^{\top},$$

where U and V contain the left and right singular vectors, respectively, and Σ is a diagonal matrix of singular values. Truncating this decomposition to the top r singular values yields the theoretically optimal compressed representation:

$$\widehat{X} = U_r \Sigma_r V_r^{\top}.$$

While generally effective, this approach can deteriorate in layers containing functionally critical superweights (Yu et al., 2024). These weights are not merely large-magnitude outliers: they participate in narrow computational pathways that can produce massive activations on specific low-semantic-content tokens (Sun et al., 2026). Since truncated SVD optimizes a global, task-agnostic reconstruction objective, it prioritizes high-energy singular directions rather than task-sensitive directions or entries. Consequently, such layers may require higher ranks or targeted corrections that explicitly preserve functionally critical coordinates.

Applying the same matrix-factorization framework to MHA projections results in even sharper quality degradation. This is because MHA has an inherently tensor-like organization (Elhage et al., 2021): different heads encode distinct functional subspaces. Flattening the corresponding projection tensors into a single 2D matrix forces all heads to share the same low-rank factors, thereby coupling otherwise heterogeneous subspaces and reducing head specialization. This structural mismatch motivates tensor decompositions, which retain the head-wise multi-dimensional organization of MHA instead of imposing an artificial matrix structure.

2.3. Tensor decompositions

An N-th order tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ generalizes matrices to higher dimensions. MHA and MoE blocks admit natural tensor representations, with heads and experts as explicit modes, motivating tensor-based compression.

Tucker decomposition approximates a tensor via a compressed core and per-mode factors:

$$\mathcal{X} \approx \mathcal{G} \times_1 U^{(1)} \times_2 U^{(2)} \cdots \times_N U^{(N)},$$

where $\mathcal{G} \in \mathbb{R}^{R_1 \times \cdots \times R_N}$ is the core tensor, $U^{(n)} \in \mathbb{R}^{I_n \times R_n}$ the n-th mode factor, and $(R_1, \ldots, R_N)$ the Tucker ranks.

Tensor Train (TT) decomposition represents a tensor as a

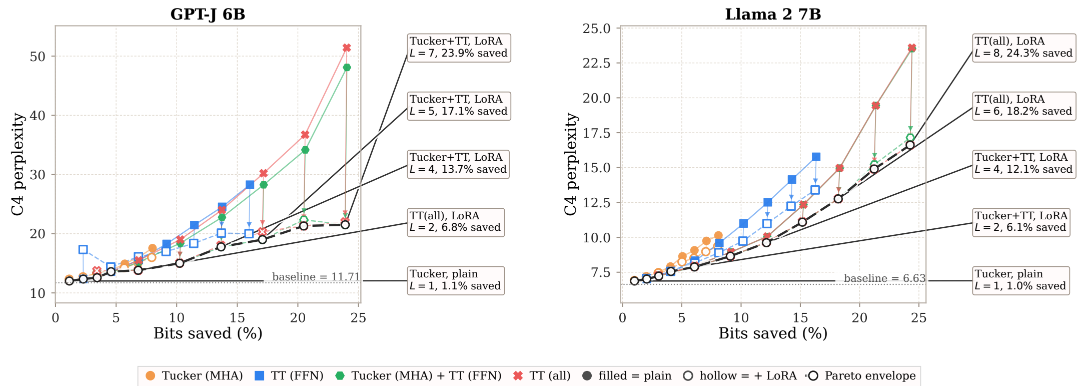

Figure 2. C4 perplexity versus bits saved, excluding embeddings, for GPT-J 6B and LLaMA 2 7B. Each point is one compression run, and L denotes the number of consecutive compressed transformer blocks. Arrows connect each decomposition to its LoRA-repaired variant, and the Pareto frontier marks the best observed trade-offs.

chain of three-dimensional cores:

$$\mathcal{X}_{i_1,\ldots,i_N} \approx \sum_{r_1,\ldots,r_{N-1}} \mathcal{G}^{(1)}_{1,i_1,r_1}\, \mathcal{G}^{(2)}_{r_1,i_2,r_2} \cdots \mathcal{G}^{(N)}_{r_{N-1},i_N,1},$$

where $\mathcal{G}^{(n)} \in \mathbb{R}^{R_{n-1}\times I_n \times R_n}$, $R_0 = R_N = 1$, and $(R_1,\ldots,R_{N-1})$ are the TT ranks.

Since MHA is equivalent to a structured convolution-like operator (Cordonnier et al., 2020), Tucker decomposition naturally fits attention projections by preserving head-wise structure, an approach explored by TensorLLM (Gu et al., 2025) and LeSTD (Li et al., 2026). For FFN layers, which lack an explicit head mode, TT decomposition is the natural counterpart.

We evaluate GPT-J 6B and LLaMA 2 7B under four compression schemes: Tucker on attention only (reproducing TensorLLM-style factorization), TT on FFN only, Tucker+TT jointly, and TT on all projections, with a maximum Tucker rank of 64. Following our layer-sensitivity analysis (Figure 1), we compress contiguous middle blocks: starting at block 14 for GPT-J 6B (range 14–20) and block 16 for LLaMA 2 7B (range 16–23), and progressively extend the range to measure error accumulation. Each compressed model is evaluated before and after a lightweight LoRA repair ($r = 16$, trained on WikiText-2). Compression is measured in bits saved over non-embedding parameters. We extend this evaluation to MoE architectures by applying TD-MoE (Xu et al., 2026) on Qwen3-30B-A3B and GPT-OSS-20B, tracking both perplexity and downstream accuracy (Appendices B.1, B.2).

Figure 2 summarizes the results. Attention-only compression preserves perplexity but yields negligible size reduction; compressing FFN layers achieves higher compression at the cost of sharp quality degradation. LoRA repair partially recovers quality but does not resolve the fundamental trade-off. Despite structural alignment between tensor formats and LLM components, all methods exhibit poor compression-quality trade-offs at practical compression ratios, pointing to a deeper mismatch between standard tensor assumptions and learned LLM representations. We also provide trade-offs for WikiText-2 perplexity and macro LM-Eval accuracy drop in Appendix C.2.

For FFN layers, where Tucker degenerates to standard matrix factorization, we additionally provide a direct comparison against LASER-style truncated SVD on LLaMA 2 7B. Figure 3 shows that TT and matrix rank reduction follow nearly identical trend, with TT reaching marginally higher bits saved at the cost of the same monotonic quality drop. Appendix C.4 further shows that decomposition sensitivity is layer-dependent, with early layers being the most fragile.

Activation analysis confirms reduced angular diversity and distorted norm distributions, indicating that TT compresses FFN representations into a restricted latent structure no more faithfully than its matrix counterpart. Even this flexible tensorization fails to escape the core trade-off between compression ratio and representation diversity (Figure 4).

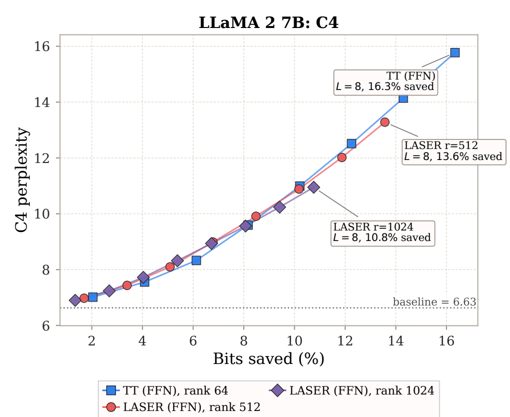

Figure 3. TT versus LASER for FFN compression on LLaMA 2 7B. Both methods compress the same middle-to-late block ranges.

Next, we consider the same question in sparse expert architectures. MoE layers are a natural tensor target: expert index provides an explicit tensor mode, thus, we evaluate TD-MoE (Xu et al., 2026) as a complementary setting on Qwen3-30B-A3B and GPT-OSS-20B. We track both perplexity and downstream accuracy where reliable; full results are in Appendices B.1 and B.2.

Despite structural alignment, tensor decompositions yield poor compression-quality trade-offs under stronger compression (Figure 2). This suggests a mismatch between standard tensor assumptions and learned LLM representations.

2.4. Structural Mismatch of Tensor Decompositions in LLMs

To explain this behavior, we compare residual-stream activations produced by the original dense model and by the model in which a consecutive range of Transformer blocks has been decomposed. For each compression run, the activation is measured after the last decomposed block, so the diagnostic captures the accumulated effect of all decomposed layers up to the target compression ratio rather than the local error of a single layer. We quantify this deviation by the mean angle between the two residual-stream activation vectors and by the ratio of their norms, averaged over evaluation tokens.

Figure 4 shows that runs with low perplexity remain close to the dense-model trajectory in both direction and scale, whereas high-perplexity runs exhibit larger angular drift, norm shrinkage, or both. The same pattern appears for LLaMA 2 7B in Figure 10, suggesting that tensorization degrades quality by progressively distorting the geometry of the residual stream.

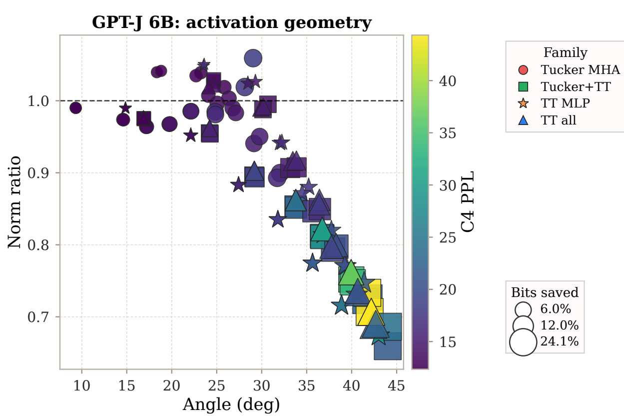

Figure 4. Activation geometry of GPT-J 6B compression runs. Each point is one run across all tested layer ranges. The x-axis shows the mean angle between dense and compressed after-block activations at the last compressed block, and the y-axis shows the compressed-to-dense norm ratio. Color indicates C4 perplexity, and marker size indicates bits saved excluding embeddings.

Thus, the main limitation is not the local expressivity of Tucker or TT decompositions, but their inability to preserve the heterogeneous representation geometry induced by LLMs.

3. Partial Explanation Through Operator Norms

The representation collapse across tensor formats can be partially explained through operator norms. Reshaping a weight matrix W into a tensor $\mathcal{T}$ preserves the ambient Euclidean geometry ($\|W\|_F = \|\mathcal{T}\|_F$), but fundamentally alters the operator isometry.

Setup. Let $W \in \mathbb{R}^{m\times n}$ be a layer weight and $\varphi : \mathbb{R}^{m\times n} \to \mathbb{R}^{I_1\times\cdots\times I_d}$ a reshape with $\prod_{\ell=1}^{d} I_\ell = mn$ and $d \geq 2$. Set $\mathcal{T} = \varphi(W)$, let $\widehat{\mathcal{T}} \in \mathcal{M}_{\mathbf{r}}$, where $\mathcal{M}_{\mathbf{r}}$ is a fixed rank manifold, be a Frobenius-optimal low-rank approximation (Tucker rank $\mathbf{r} = (r_1, \ldots, r_d)$ via HOSVD or TT rank $(r_1, \ldots, r_{d-1})$ via TT-SVD), and $\widehat{W} = \varphi^{-1}(\widehat{\mathcal{T}})$. Singular values of the mode-ℓ unfolding $\mathcal{T}_{(\ell)}$ are $\sigma_{\ell,1} \geq \sigma_{\ell,2} \geq \ldots$; we drop the mode subscript when clear.

Tensorization preserves Frobenius but not operator geometry. Since φ permutes entries, $\|\varphi(A)\|_F = \|A\|_F$ for all A. The tensor spectral (injective) norm,

$$\|\mathcal{T}\|_\sigma := \sup_{\|u^{(\ell)}\|_2=1} \langle \mathcal{T}, u^{(1)} \otimes \cdots \otimes u^{(d)} \rangle, \tag{2}$$

takes its supremum over rank-one unit tensors: a strict subset of the unit operator-norm ball when $d \geq 3$. Hence $\|\varphi(W)\|_\sigma \leq \|W\|_2$, with equality only when $\varphi(W)$ is rank-one. For $d \geq 3$ computing $\|\mathcal{T}\|_\sigma$ is computationally hard, and the two norms can be polynomially separated.

4. Conclusion

This study shows that tensor decompositions fail as a standalone post-training LLM compression method because they optimize the wrong geometry. Tucker and TT preserve Frobenius mass while distorting the spectral (head-, neuron-, and expert-specific) subspaces that carry model function. These perturbations accumulate across depth, and manifest as activation drift and downstream degradation.

Impact Statement

This work aims to clarify the practical scope of tensor decomposition methods for post-training LLM compression and to contextualize prior results obtained under questionable evaluation settings. By highlighting key limitations and constraints, we hope to support more rigorous research on tensor decomposition methods for LLM compression.

Acknowledgments

The work was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-15-2025-013, dated June 20, 2025, IGK 000000C313925P4B0002).

References

Chee, J., Cai, Y., Kuleshov, V., and De Sa, C. M. Quip: 2-bit quantization of large language models with guarantees. *Advances in neural information processing systems*, 36: 4396–4429, 2023.

Chen, X., Ha, M., Lan, Z., Zhang, J., and Li, J. Mobe: Mixture-of-basis-experts for compressing moe-based llms. *arXiv preprint arXiv:2508.05257*, 2025.

Cordonnier, J. B., Loukas, A., and Jaggi, M. On the relationship between self-attention and convolutional layers. In *8th International Conference on Learning Representations, ICLR 2020*, 2020.

Eckart, C. and Young, G. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Egiazarian, V., Panferov, A., Kuznedelev, D., Frantar, E., Babenko, A., and Alistarh, D. Extreme compression of large language models via additive quantization. *arXiv preprint arXiv:2401.06118*, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022a.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022b.

Frantar, E. and Alistarh, D. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Gromov, A., Tirumala, K., Shapourian, H., Glorioso, P., and Roberts, D. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations*, 2024.

Gu, Y., Zhou, W., Iacovides, G., and Mandic, D. Tensorllm: Tensorising multi-head attention for enhanced reasoning and compression in llms. In *International Joint Conference on Neural Networks (IJCNN)*, 2025. URL https://arxiv.org/abs/2501.15674.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Li, Y., Guo, Z., and Miao Yin, B. L. LeSTD: Learning sparse Tucker decomposition for efficient large language models. arXiv preprint arXiv:2601.01123, 2026.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Ma, X., Fang, G., and Wang, X. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Mirsky, L. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1): 50–59, 1960.

Molodtsov, G., Miasnikov, A., and Beznosikov, A. Hierarchical mixture-of-experts with two-stage optimization. In *ICML 2026 Workshop on Weight-Space Symmetries: from Foundations to Practical Applications*, 2026.

OpenAI. Introducing gpt-oss. https://openai.com/index/introducing-gpt-oss/, 2025. Accessed: 2026-05-08.

Sharma, P., Ash, J., and Misra, D. K. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. In *International Conference on Learning Representations*, volume 2024, pp. 17632–17651, 2024.

Su, Z., Li, Q., Zhang, H., Qian, Y., Xie, Y., and Yuan, K. Unveiling super experts in Mixture-of-Experts large language models. *arXiv preprint arXiv:2507.23279*, 2025. URL https://arxiv.org/abs/2507.23279.

Sun, S., Canziani, A., LeCun, Y., and Zhu, J. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

Tan, S., Tam, W. L., Wang, Y., Gong, W., Zhao, S., Zhang, P., and Tang, J. Gkd: A general knowledge distillation framework for large-scale pre-trained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 134–148, 2023.

Tang, Y., Li, X., Liu, F., Guo, W., Zhou, H., Wang, Y., Han, K., Yu, X., Li, J., Zang, H., et al. Pangu pro moe: Mixture of grouped experts for efficient sparsity. *arXiv preprint arXiv:2505.21411*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xu, Y., Wang, Y., Peng, X., Zang, H., Chen, M., Xia, P., and Wen, Z. Td-moe: Tensor decomposition for moe models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=D9cnZNZfxX. ICLR 2026.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yu, M., Wang, D., Shan, Q., Reed, C. J., and Wan, A. The super weight in large language models. *arXiv preprint arXiv:2411.07191*, 2024.

Zagitov, A., Molodtsov, G., and Beznosikov, A. Harp: Hadamard-preconditioned adaptive rotation processor for extreme llm quantization. 2026. URL https://arxiv.org/abs/2605.29843.

A. Notation

General objects and conventions. Scalars are denoted by lowercase letters such as r, p, q, ρ, and L. Matrices are denoted by uppercase letters such as X, $\widehat{X}$, A, W, $\widehat{W}$, U, V, and Σ. Tensors are denoted by calligraphic letters such as $\mathcal{X}$, $\mathcal{G}$, and $\mathcal{T}$. The notation $\widehat{\cdot}$ denotes an approximation of the corresponding dense object. The symbol $\approx$ denotes an approximate factorization or reconstruction.

Compression and evaluation variables. L denotes the number of consecutive compressed Transformer blocks in the quality–compression trade-off experiments. The LoRA repair rank is denoted $r = 16$. The TD-MoE per-layer compression ratio is denoted by ρ, with experiments using $\rho \in \{0.2, 0.4\}$. In the GPT-OSS-20B expert-mode comparison, PRESERVE fixes $r_1 = K$, while COMPRESS uses $r_1 < K$. Example selected TD-MoE ranks are $(32, 1720, 2664)$ for PRESERVE and $(20, 2496, 2880)$ for COMPRESS.

Method abbreviations. Table 1 lists the method-specific abbreviations used throughout the paper.

Table 1. Method abbreviations.

Abbreviation	Meaning / role in the paper
HOSVD	Higher-order singular value decomposition
TT-SVD	SVD-based algorithm for constructing Tensor Train decompositions
TD-MoE	Tensor-decomposition compression of Mixture-of-Experts layers
LASER	Truncated-SVD-based post-training compression baseline for FFN compression
LeSTD	Prior Tucker-style tensor-compression method for Transformer attention projections
TensorLLM	Prior Tucker-style tensor-compression method for Transformer attention projections

B. TD-MoE Experiments

Mixture-of-Experts architectures are a particularly natural target for compression because expert parameters are expected to contain substantial redundancy. Since the introduction of Switch Transformers (Fedus et al., 2022b), MoE models have been trained under a tension between expert specialization and load balancing: experts must learn distinct functions while remaining sufficiently interchangeable to ensure stable routing. This trade-off often leads to overlapping representations across experts. The effect is even more pronounced in grouped expert architectures (Tang et al., 2025; Molodtsov et al., 2026), where experts are explicitly organized into shared structures. As a result, many recent MoE compression methods are based on the premise that expert weights admit a compact shared representation and that cross-expert redundancy can be removed with limited quality degradation. TD-MoE follows this intuition by treating the expert dimension as a tensor mode and factorizing it jointly with the feature dimensions.

B.1. Progressive Layer Compression on Qwen3-30B-A3B

TD-MoE is a post-training compression method designed for Mixture-of-Experts layers. Instead of decomposing each expert weight matrix independently, it stacks all experts in a layer into a three-dimensional tensor over expert, input, and output modes, then applies a joint Tucker factorization:

$$\mathcal{X} \approx \mathcal{G} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}.$$

Here $U^{(1)} \in \mathbb{R}^{K \times r_1}$ acts on the expert mode and represents r_1 latent meta-experts, $U^{(2)} \in \mathbb{R}^{d_{\text{in}} \times r_2}$ spans the compressed input-feature subspace, and $U^{(3)} \in \mathbb{R}^{d_{\text{out}} \times r_3}$ spans the compressed output-feature subspace. The core tensor $\mathcal{G}$ couples these three latent modes.

We evaluate TD-MoE on Qwen3-30B-A3B by compressing MoE layers one at a time, starting from the middle, the 24th layer. We first extend the compressed set toward later layers for 12 MoE layers, reaching roughly the 75% depth point of the model, and then extend toward earlier layers for another 12 MoE layers, reaching roughly the 25% depth point. The final setting therefore covers 24 MoE layers in total. Figure 5 shows perplexity on WikiText-2 and C4 for two per-layer compression ratios $\rho \in \{0.2, 0.4\}$, where ρ controls how aggressively each expert is compressed via Tucker decomposition.

Figure 6 reports the corresponding downstream accuracy curves. A plausible explanation for the model-dependent behavior

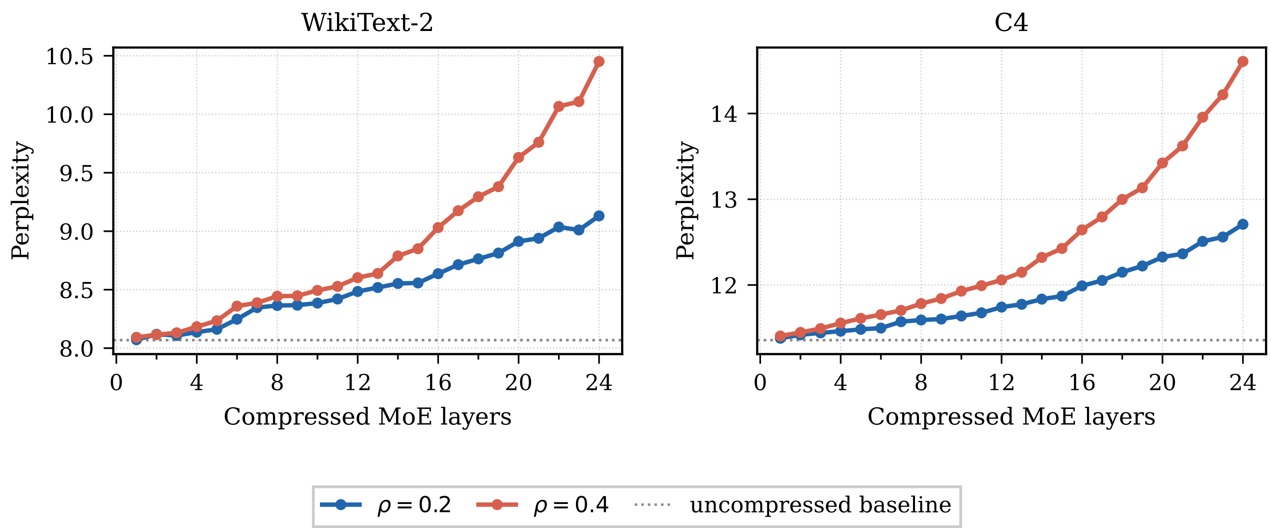

Figure 5. Perplexity on WikiText-2 (left) and C4 (right) as the number of TD-MoE-compressed MoE layers increases on Qwen3-30B-A3B.

is that Qwen3-30B-A3B is a fine-grained MoE with 128 experts and 8 active experts per token (Yang et al., 2025). Recent work identifies three shallow *super experts* in layers 1–3 whose pruning raises WikiText-2 perplexity from 8.70 to 59.86 and collapses reasoning, while randomly pruning non-super experts has negligible effect (Su et al., 2025). Our schedule starts at mid-depth and does not touch those shallow super experts, so moderate TD-MoE may act mainly as denoising in less critical expert subspaces, consistent with prior observations that selective rank reduction can sometimes improve LLM accuracy by removing harmful higher-order components (Sharma et al., 2024). GPT-OSS-20B has a smaller MoE structure, with 24 layers, 32 experts, and 4 active experts per token (OpenAI, 2025), so the same intervention has less expert-mode redundancy to exploit before it removes functional diversity (Figure 7).

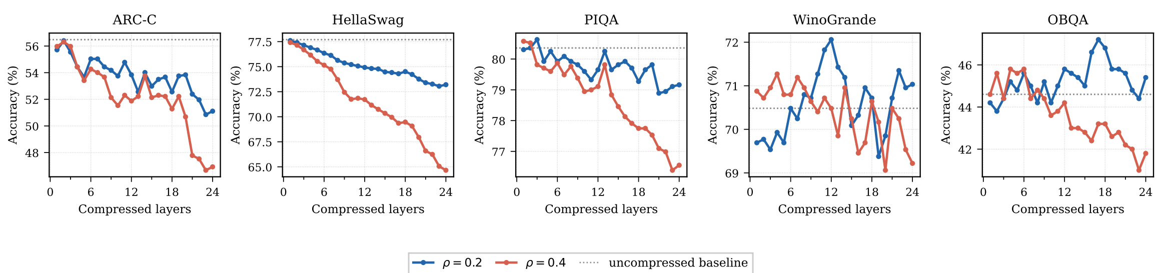

Figure 6. Downstream task accuracy (%) as the number of TD-MoE-compressed MoE layers increases on Qwen3-30B-A3B.

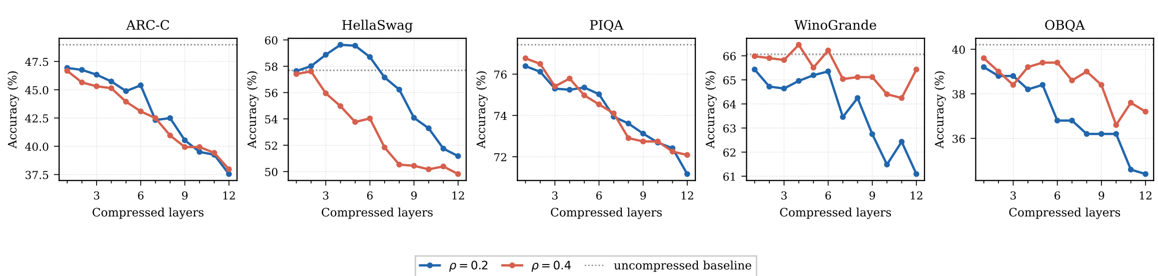

Figure 7. Downstream task accuracy (%) as the number of TD-MoE-compressed MoE layers increases on GPT-OSS-20B for $\rho \in \{0.2, 0.4\}$.

B.2. Expert-Mode Comparison

We evaluate TD-MoE on OpenAI GPT-OSS-20B at per-layer compression ratio $\rho{=}0.4$, varying the *expert mode*: PRESERVE fixes the expert-dimension Tucker rank to $r_1{=}K$ (all experts are kept as distinct latent directions), whereas COMPRESS also compresses the expert dimension ($r_1 < K$). At the same target compression ratio, this reallocates the saved expert-mode budget to the feature modes: in our GPT-OSS-20B runs, PRESERVE selects ranks $(32, 1720, 2664)$, while COMPRESS selects $(20, 2496, 2880)$. Thus COMPRESS trades expert-mode capacity for higher-rank input and output factors, allowing the decomposition to exploit cross-expert redundancy at the cost of reduced expert diversity.

Figure 8 reports the corresponding downstream accuracy curves for PRESERVE and COMPRESS as the number of TD-MoE-compressed GPT-OSS-20B layers increases.

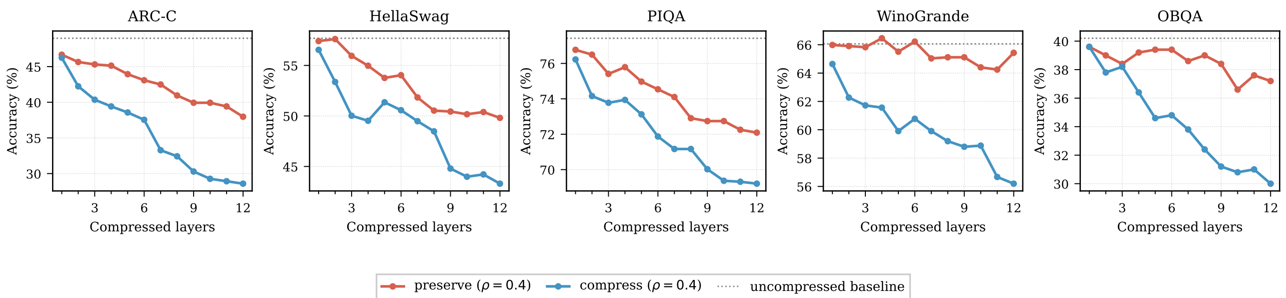

Figure 8. Downstream accuracy (%) versus number of TD-MoE-compressed layers on GPT-OSS-20B at $\rho{=}0.4$. PRESERVE keeps the expert Tucker rank equal to K; COMPRESS additionally reduces the expert dimension.

B.3. Comparison against Matrix Decomposition (MoBE)

We compare TD-MoE against Mixture-of-Basis-Experts (Chen et al., 2025) at two compression ratios, $\rho \in \{0.2, 0.4\}$, and perform an activation-geometry diagnostic. Figure 9 shows that, under both compression settings, TD-MoE has larger activation angles and greater residual-stream relative L2 error, while also exhibiting worse perplexity.

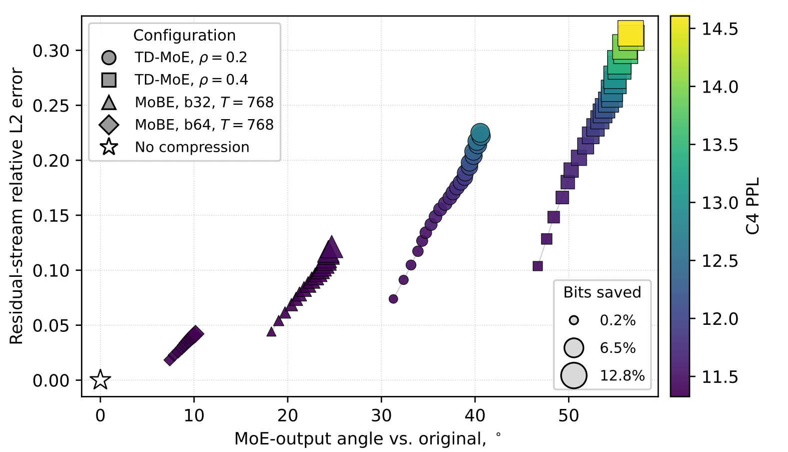

Figure 9. Activation geometry of Qwen3-30B-A3B compression runs. Each point is one run across all tested layer ranges. Angle measures directional drift from the dense model and residual-stream relative L2 error measures reconstruction distortion in the stream.

C. Additional Tensor-Decomposition Results

C.1. Activation-Geometry Diagnostics

Figure 10 repeats the activation-geometry diagnostic for LLaMA 2 7B. As in GPT-J 6B, high-perplexity runs move away from the dense model either by increasing the activation angle or by shrinking the activation norm. This supports the interpretation that the degradation is tied to changes in intermediate representation geometry rather than to compression ratio alone.

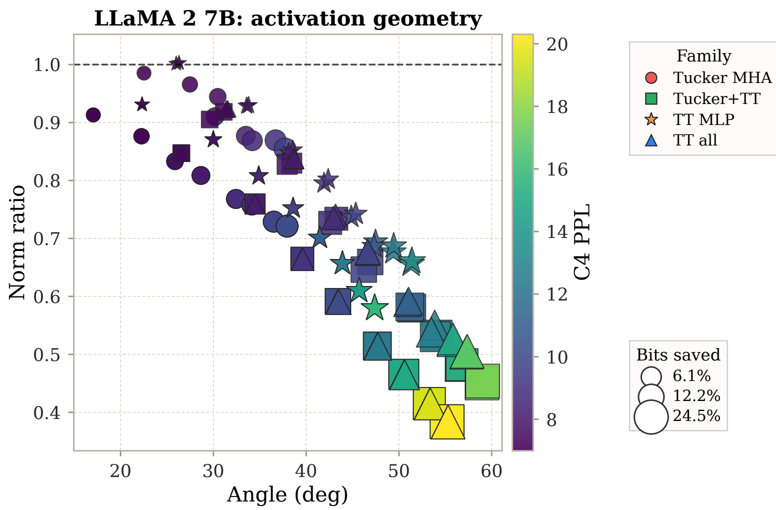

Figure 10. Activation geometry of LLaMA 2 7B compression runs. Each point is one run across all tested layer ranges. Angle measures directional drift from the dense model, norm ratio measures activation-scale distortion, color shows C4 perplexity, and marker size shows bits saved excluding embeddings.

C.2. Quality–Compression Trade-offs

In addition to C4 perplexity, we evaluate WikiText-2 perplexity and zero-shot LM-Eval accuracy. The LM-Eval score is computed on ARC-Challenge, HellaSwag, OpenBookQA, PIQA, and WinoGrande. For each task, we measure the accuracy drop relative to the dense model in percentage points, and report the unweighted average across tasks as the macro LM-Eval accuracy drop. Lower values are better for all metrics shown in this section.

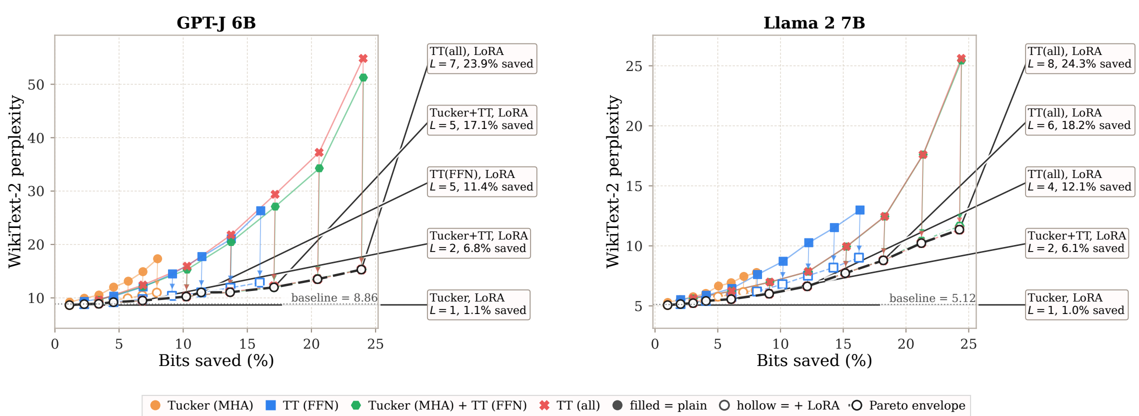

Figure 11. WikiText-2 perplexity versus bits saved for GPT-J 6B and LLaMA 2 7B. Each point is one compression run, and L denotes the number of consecutive compressed transformer blocks. Arrows connect each decomposition to its LoRA-repaired variant; the Pareto frontier marks the best observed trade-offs.

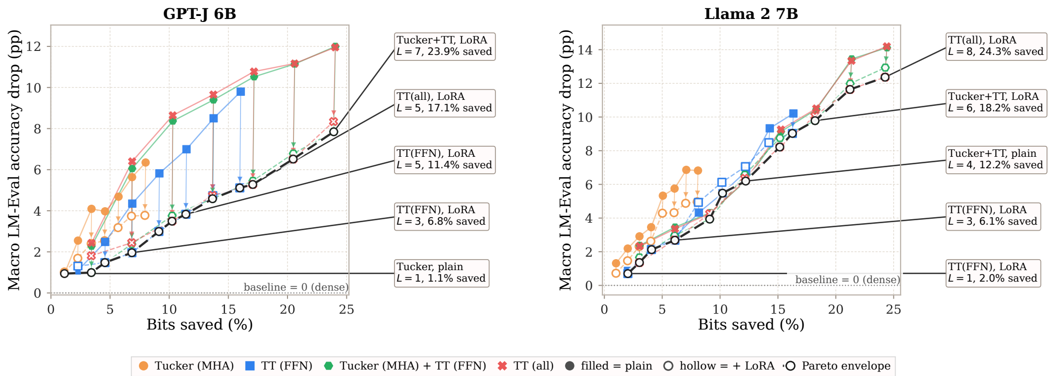

Figure 12. Macro LM-Eval accuracy drop versus bits saved for GPT-J 6B and LLaMA 2 7B. Macro drop is the unweighted average accuracy drop, in percentage points, across ARC-Challenge, HellaSwag, OpenBookQA, PIQA, and WinoGrande. Lower is better.

C.3. GPT-J and LLaMA 2 Tensor-Decomposition Protocol

This section describes the post-training tensor-decomposition experiments on GPT-J 6B and LLaMA 2 7B reported in Figure 2. We use the progressive block schedule: GPT-J 6B is compressed from block 14 through block 20, and LLaMA 2 7B from block 16 through block 23, adding one consecutive block at a time. All runs are evaluated relative to a fixed dense baseline for the same model.

For attention compression, we use a TensorLLM-style Tucker factorization of the query, key, value, and output projections. Since these projections have matching shapes in both models, we stack them into a tensor with input-feature, head, head-dimension, and projection-type modes. We apply partial Tucker factorization to the input-feature, head-dimension, and projection-type modes, leaving the head mode explicit. The maximum input-feature rank is 64, the head-dimension rank is 4, and the projection-type rank is 2.

For TT compression of FFN projections, and for the TT-all setting, each linear weight matrix is tensorized into twelve paired input-output modes. The input and output dimensions are split into twelve approximately balanced factors, interleaved into input-output pairs, and compressed with TT ranks capped at 64. Bias terms, when present, are kept dense.

LoRA repair is applied after decomposition with the decomposed modules frozen. We train rank-16 LoRA adapters with scaling factor 32 on WikiText-2 using AdamW, learning rate 2×10^{-4}, 100 optimizer steps, and gradient accumulation over 8 micro-batches. This stage tests whether limited data-dependent adaptation can recover quality lost by factorization, rather than performing full fine-tuning.

Perplexity is evaluated on WikiText-2 and C4, using sequence length 2048 for GPT-J 6B and 4096 for LLaMA 2 7B. For activation geometry diagnostics, we collect WikiText-2 sequences from the dense and compressed models and compare activations at the last compressed block. We report the mean angular deviation from the dense activations and the compressed-to-dense activation norm ratio. Storage is always computed from the compressed representation, excluding embeddings; for benchmarking, compressed modules are reconstructed to dense linear layers so that the reported quality reflects the compressed weights rather than implementation-specific runtime kernels.

C.4. Single-Layer Decomposition on LLaMA 2 7B

To separate local layer sensitivity from error accumulation across depth, we also evaluate a single-layer protocol on LLaMA 2 7B. In each run, only one transformer layer is modified, and all other layers remain dense. We repeat this for all 32 layers and evaluate WikiText-2 and C4 perplexity. For attention layers, we compare LASER-style matrix factorization, TensorLLM-style Tucker factorization, and a per-head TT variant. For FFN layers, we compare LASER-style matrix factorization with TT factorization. Tucker always refers to the TensorLLM-style decomposition of the attention projections.

Figure 13 shows that sensitivity is highly non-uniform across depth. The earliest layers are especially fragile: decomposing the first attention layer causes very large perplexity spikes for Tucker and per-head TT, while TT on FFN is most unstable in the first two FFN layers. Away from these early layers, attention decompositions usually stay much closer to the dense baseline, although they save little of the total model size. FFN decompositions save more parameters, but their effect grows toward the final layers, especially for TT. This supports the progressive middle-to-late compression schedule used in the main experiments: middle layers are less sensitive locally, but quality still degrades once errors are accumulated over multiple compressed blocks.

Table 2. Storage accounting for the LLaMA 2 7B single-layer decomposition study. Each run compresses one layer at a time. Bits saved are measured over non-embedding model parameters.

Method	Target	Rank	Layer-local CR	Bits saved (%)
LASER	MHA	1024	2.00	0.51
Tucker	MHA	64	240.49	1.01
TT/head	MHA	64	1.98	0.50
LASER	FFN	1024	2.92	1.34
TT	FFN	64	398.63	2.04

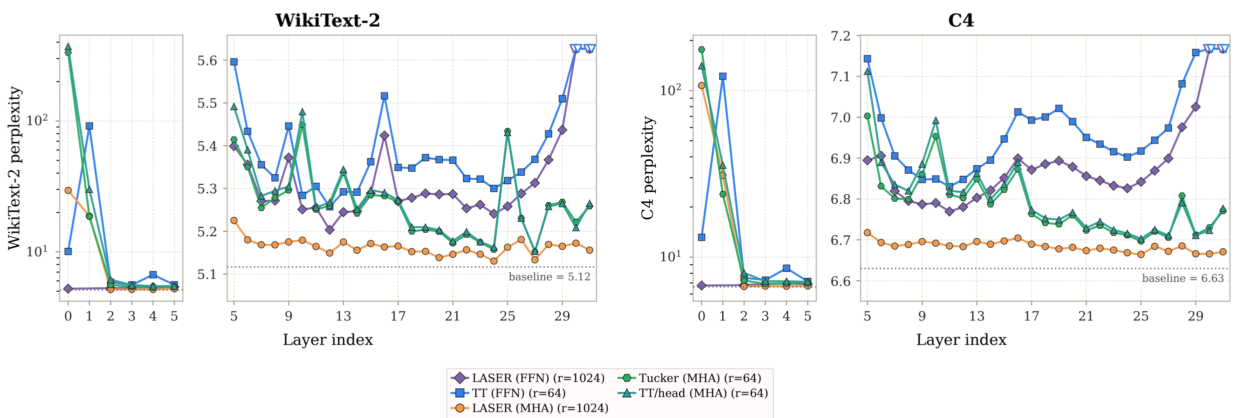

Figure 13. Single-layer decomposition sensitivity on LLaMA 2 7B. Each point compresses only one transformer layer. The left and right panels report WikiText-2 and C4 perplexity, respectively; early layers are shown separately because of large perplexity spikes.

C.5. Detailed GPT-J and LLaMA 2 Results

Tables 3 and 4 report the GPT-J 6B and LLaMA 2 7B tensor-decomposition experiments. Each cell is shown as *direct / +LoRA*, where +LoRA denotes the lightweight rank-16 LoRA trained on WikiText-2. The maximum-L table gives the most compressed setting for each model, while the full table reports all middle-to-late block ranges. For activation geometry, *Angle* is the mean angle between dense and compressed residual-stream activations after the last compressed block, and *Norm* is the compressed-to-dense activation norm ratio. Lower Angle and Norm closer to 1 are better.

Table 3. Maximum-L GPT-J 6B and LLaMA 2 7B tensor-decomposition results. Compressed rows report *direct / +LoRA*. Bits saved are measured over all non-embedding model parameters; bold values are best among compressed runs within each model group.

			Compression ↑	Perplexity ↓		LM-Eval accuracy (%) ↑						Activation geometry	
L	Blocks	Method	Bits saved (%)	WT2	C4	Macro	ARC-C	HSwag	OBQA	PIQA	WinoG	Angle ↓	Norm → 1
GPT-J 6B Dense baseline			0.0	8.86	11.71	50.5	33.9	49.5	29.0	75.5	64.4	–	–
7	14–20	Tucker MHA	8.0 / 7.9	17.33 / **10.98**	17.56 / **15.99**	44.1 / **46.7**	26.5 / **30.3**	41.3 / **45.4**	21.0 / **24.2**	70.0 / **71.7**	**61.9 / 61.9**	**19.4** / 24.2	**0.98** / 0.98
		TT FFN	16.0 / 16.0	26.31 / 12.86	28.28 / 20.00	40.7 / 45.4	24.1 / 29.2	37.3 / 43.4	17.8 / 23.8	66.2 / 71.1	57.9 / 59.3	30.5 / 30.3	0.83 / 0.87
		Tucker+TT	**24.1** / 23.9	51.26 / 15.34	48.09 / 21.51	38.5 / 42.6	21.8 / 24.5	33.3 / 40.6	16.8 / 20.4	62.8 / 68.8	57.5 / 58.8	31.8 / 35.1	0.85 / 0.85
		TT all	24.0 / 23.9	54.84 / 15.30	51.42 / 21.99	38.5 / 42.1	21.9 / 24.8	33.0 / 40.3	17.8 / 19.8	62.8 / 68.4	57.1 / 57.3	31.8 / 35.1	0.86 / 0.85
LLaMA 2 7B Dense baseline			0.0	5.12	6.63	56.2	43.0	57.1	33.4	78.1	69.4	–	–
8	16–23	Tucker MHA	8.1 / 8.0	7.78 / **6.30**	10.14 / **9.36**	49.4 / **51.2**	33.0 / **35.8**	49.5 / **52.0**	24.8 / **26.8**	71.6 / **73.7**	**68.0 / 68.0**	**29.4** / 29.9	0.80 / **0.90**
		TT FFN	16.3 / 16.2	12.98 / 9.02	15.77 / 13.40	46.0 / 47.2	31.6 / 30.7	43.4 / 45.9	22.8 / 23.6	67.4 / 68.9	64.7 / 66.7	38.0 / 41.7	0.74 / 0.78
		Tucker+TT	**24.4** / 24.3	25.44 / 11.64	23.53 / 17.15	42.1 / 43.3	27.8 / 27.7	37.7 / 40.9	18.8 / 18.8	62.6 / 64.6	63.5 / 64.3	43.8 / 48.7	0.58 / 0.65
		TT all	24.4 / 24.3	25.61 / 11.35	23.59 / 16.61	42.0 / 43.8	27.7 / 27.8	37.7 / 41.1	18.4 / 20.6	62.6 / 64.5	63.5 / 65.2	43.9 / 47.1	0.58 / 0.68

Table 4. Detailed GPT-J 6B and LLaMA 2 7B tensor-decomposition results. Compressed rows are grouped by the number of consecutive decomposed blocks L and report *direct / +LoRA*. Bits saved are measured over all non-embedding model parameters. Bold values are best among compressed runs within each model-L group.

			Compression ↑	Perplexity ↓		LM-Eval accuracy (%) ↑						Activation geometry	
L	Blocks	Method	Bits saved (%)	WT2	C4	Macro	ARC-C	HSwag	OBQA	PIQA	WinoG	Angle ↓	Norm → 1
GPT-J 6B Dense baseline			0.0	8.86	11.71	50.5	33.9	49.5	29.0	75.5	64.4	–	–
1	14	Tucker MHA	1.1 / 1.1	9.19 / **8.65**	**12.01** / 12.30	**49.5** / 49.4	32.6 / **34.1**	47.8 / **49.2**	28.0 / 25.8	74.6 / **74.9**	**64.6** / 63.1	**9.3** / 18.3	**0.99** / 1.04
		TT FFN	2.3 / 2.3	9.32 / 8.83	12.35 / 17.33	49.4 / 49.2	32.5 / 33.4	47.7 / 49.1	27.8 / 27.2	74.5 / 73.6	64.2 / 62.6	14.9 / 23.5	0.99 / 1.05
		Tucker+TT	**3.4** / 3.4	9.60 / 8.89	12.57 / 13.41	48.2 / 49.5	29.8 / 33.0	46.2 / 48.4	26.6 / **29.4**	74.3 / 73.9	64.1 / 62.6	16.9 / 24.6	0.98 / 1.03
		TT all	3.4 / 3.4	9.62 / 8.93	12.59 / 13.74	48.0 / 48.7	29.7 / 31.7	46.2 / 48.8	26.4 / 26.2	74.4 / 73.9	63.5 / 62.7	16.9 / 24.7	0.98 / 1.02
2	14–15	Tucker MHA	2.3 / 2.3	9.93 / **8.93**	**12.57** / 12.71	47.9 / 48.8	30.2 / 32.8	46.5 / **48.1**	25.6 / 25.0	73.7 / **74.1**	63.5 / **63.9**	**11.9** / 20.3	0.98 / 1.04
		TT FFN	4.6 / 4.6	10.31 / 9.20	13.59 / 14.39	48.0 / **49.0**	31.1 / **33.5**	45.5 / 48.0	**27.8** / 27.6	73.2 / 73.5	62.3 / 62.3	18.5 / 23.3	0.97 / 1.01
		Tucker+TT	**6.9** / 6.8	12.13 / 9.50	15.20 / 13.91	44.4 / 48.1	26.3 / 31.7	41.7 / 47.0	23.0 / 26.0	71.3 / 73.8	59.8 / 62.0	20.5 / 26.2	0.97 / **0.99**
		TT all	6.9 / 6.8	12.41 / 9.50	15.50 / 13.82	44.1 / 48.0	25.6 / 32.6	41.5 / 46.7	23.0 / 24.2	70.8 / 73.7	59.4 / 63.0	20.5 / 25.9	0.97 / 0.99
3	14–16	Tucker MHA	3.4 / 3.4	10.53 / **9.26**	**13.28** / 13.30	46.4 / 48.1	28.5 / 31.8	44.9 / **47.2**	24.4 / 25.4	72.6 / **73.3**	61.4 / **62.6**	**13.7** / 19.9	0.98 / **1.01**
		TT FFN	6.9 / 6.8	12.06 / 9.77	15.54 / 16.12	46.1 / **48.5**	28.4 / **32.7**	43.2 / 46.9	25.4 / **28.4**	72.9 / 73.2	60.7 / 61.4	21.5 / 24.7	0.94 / 0.97
		Tucker+TT	**10.3** / 10.2	15.30 / 10.27	18.41 / 15.01	42.1 / 46.7	23.3 / 30.4	39.3 / 45.5	20.0 / 25.0	69.4 / 72.1	58.6 / 60.5	23.4 / 27.0	0.94 / 0.95
		TT all	10.3 / 10.2	15.93 / 10.25	19.01 / 15.04	41.8 / 47.0	23.0 / 30.8	38.9 / 45.4	19.2 / 25.6	69.3 / 72.9	58.7 / 60.2	23.4 / 27.1	0.95 / 0.95
4	14–17	Tucker MHA	4.6 / 4.5	12.01 / **9.57**	13.79 / **13.61**	46.5 / **47.9**	28.8 / 31.5	44.4 / **47.0**	24.8 / 25.2	72.9 / **73.4**	61.5 / **62.5**	**15.2** / 20.3	0.97 / **1.00**
		TT FFN	9.2 / 9.1	14.49 / 10.38	18.28 / 16.99	44.6 / 47.5	26.7 / **32.3**	41.4 / 46.0	22.6 / **26.4**	71.5 / 72.8	61.0 / 59.9	24.1 / 26.1	0.92 / 0.94
		Tucker+TT	**13.7** / 13.7	20.46 / 11.06	22.77 / 17.78	41.1 / 45.9	22.9 / 29.4	37.4 / 44.3	17.8 / 23.8	68.7 / 71.3	58.6 / 60.6	26.0 / 28.6	0.92 / 0.91
		TT all	13.7 / 13.7	21.80 / 11.12	24.01 / 18.05	40.8 / 45.7	23.0 / 29.4	37.1 / 44.1	17.6 / 23.2	68.3 / 71.9	58.1 / 60.1	26.0 / 29.2	0.93 / 0.92
5	14–18	Tucker MHA	5.7 / 5.7	13.12 / **9.91**	14.92 / **14.20**	45.8 / **47.3**	27.6 / 30.1	43.1 / **46.2**	24.6 / **25.6**	72.1 / **73.1**	**61.5** / **61.5**	**16.6** / 21.6	0.98 / **1.00**
		TT FFN	11.5 / 11.4	17.73 / 11.03	21.46 / 18.36	43.5 / 46.6	27.0 / **32.1**	40.2 / 45.3	20.8 / 24.8	70.2 / 71.7	59.1 / 59.3	26.4 / 27.4	0.89 / 0.91
		Tucker+TT	**17.2** / 17.1	27.10 / 11.97	28.26 / 19.00	39.9 / 45.0	22.3 / 29.7	35.9 / 43.2	17.8 / 22.6	65.8 / 70.9	57.9 / 58.6	28.1 / 30.2	0.90 / 0.89
		TT all	17.2 / 17.1	29.39 / 12.16	30.22 / 20.33	39.7 / 45.2	22.1 / 29.4	35.6 / 43.1	17.4 / 23.4	65.7 / 71.0	57.6 / 59.0	28.2 / 31.2	0.91 / 0.89
6	14–19	Tucker MHA	6.9 / 6.8	14.90 / **10.48**	16.04 / **15.11**	44.8 / **46.7**	26.8 / **30.1**	42.4 / **45.9**	22.4 / 23.2	71.2 / **73.1**	**61.4** / 61.3	**17.9** / 23.8	0.98 / **0.99**
		TT FFN	13.8 / 13.7	21.05 / 11.89	24.54 / 20.10	42.0 / 45.7	24.7 / 28.8	38.6 / 43.9	20.2 / **24.6**	68.3 / 71.4	58.0 / 60.0	28.5 / 28.9	0.86 / 0.88
		Tucker+TT	**20.6** / 20.5	34.27 / 13.62	34.16 / 22.35	39.3 / 43.7	21.6 / 26.8	34.6 / 41.3	18.0 / 21.0	64.4 / 69.9	58.0 / 59.4	30.1 / 33.3	0.87 / 0.86
		TT all	20.6 / 20.5	37.25 / 13.50	36.73 / 21.31	39.3 / 43.9	22.0 / 26.2	34.5 / 41.4	17.6 / 22.6	64.3 / 70.8	58.1 / 58.8	30.1 / 33.0	0.88 / 0.86
7	14–20	Tucker MHA	8.0 / 7.9	17.33 / **10.98**	17.56 / **15.99**	44.1 / **46.7**	26.5 / **30.3**	41.3 / **45.4**	21.0 / **24.2**	70.0 / **71.7**	**61.9** / **61.9**	**19.4** / 24.2	**0.98** / 0.98
		TT FFN	16.0 / 16.0	26.31 / 12.86	28.28 / 20.00	40.7 / 45.4	24.1 / 29.2	37.3 / 43.4	17.8 / 23.8	66.2 / 71.1	57.9 / 59.3	30.5 / 30.3	0.83 / 0.87
		Tucker+TT	**24.1** / 23.9	51.26 / 15.34	48.09 / 21.51	38.5 / 42.6	21.8 / 24.5	33.3 / 40.6	16.8 / 20.4	62.8 / 68.8	57.5 / 58.8	31.8 / 35.1	0.85 / 0.85
		TT all	24.0 / 23.9	54.84 / 15.30	51.42 / 21.99	38.5 / 42.1	21.9 / 24.8	33.0 / 40.3	17.8 / 19.8	62.8 / 68.4	57.1 / 57.3	31.8 / 35.1	0.86 / 0.85
LLaMA 2 7B Dense baseline			0.0	5.12	6.63	56.2	43.0	57.1	33.4	78.1	69.4	–	–
1	16	Tucker MHA	1.0 / 1.0	5.28 / **5.06**	**6.87** / 6.91	54.9 / 55.5	41.0 / **42.7**	55.3 / **56.0**	32.4 / 33.2	76.8 / 76.9	69.0 / 68.5	**17.1** / 22.5	0.91 / 0.99
		TT FFN	2.0 / 2.0	5.52 / 5.15	7.01 / 7.05	55.3 / **55.5**	41.9 / 42.2	55.9 / 55.9	31.4 / 32.4	77.6 / **78.0**	**69.9** / 69.0	22.3 / 26.3	0.93 / **1.00**
		Tucker+TT	**3.1** / 3.0	5.61 / 5.24	7.23 / 7.25	53.8 / 54.5	38.9 / 40.2	54.3 / 54.8	30.4 / 32.6	76.4 / 77.0	69.1 / 68.1	26.5 / 31.1	0.85 / 0.92
		TT all	3.1 / 3.0	5.61 / 5.23	7.23 / 7.26	53.9 / 54.8	39.1 / 40.7	54.4 / 54.8	30.4 / **33.6**	76.4 / 77.1	69.1 / 68.0	26.5 / 31.5	0.85 / 0.92
2	16–17	Tucker MHA	2.0 / 2.0	5.52 / **5.20**	**7.15** / 7.18	54.0 / **54.7**	38.8 / **40.1**	54.1 / **55.2**	31.6 / **32.8**	76.6 / **77.0**	**69.0** / 68.6	**19.7** / 23.9	0.89 / **0.96**
		TT FFN	4.1 / 4.1	5.88 / 5.41	7.56 / 7.56	54.1 / 54.1	39.9 / 39.1	54.4 / 54.6	31.2 / 31.4	76.6 / 76.6	68.3 / 68.7	26.2 / 28.7	0.90 / 0.94
		Tucker+TT	**6.1** / 6.1	6.24 / 5.56	7.97 / 7.88	52.7 / 53.2	37.8 / 38.0	52.2 / 53.1	29.8 / 30.4	75.2 / 76.3	68.7 / 68.4	30.5 / 34.1	0.80 / 0.86
		TT all	6.1 / 6.1	6.24 / 5.56	7.97 / 7.90	52.8 / 53.5	38.0 / 38.7	52.2 / 53.2	29.8 / 31.4	75.2 / 76.3	**69.0** / 68.0	30.5 / 34.5	0.80 / 0.87
3	16–18	Tucker MHA	3.0 / 3.0	5.76 / **5.34**	7.50 / **7.45**	53.3 / **53.9**	37.7 / 38.8	53.3 / **54.6**	30.8 / **31.4**	76.3 / **76.6**	68.4 / 68.0	**21.7** / 25.8	0.87 / **0.96**
		TT FFN	6.1 / 6.1	6.44 / 5.73	8.33 / 8.17	53.3 / 53.5	**39.4** / 38.9	52.8 / 53.3	29.6 / **31.4**	76.0 / 75.7	**68.7** / 68.3	29.1 / 31.8	0.87 / 0.90
		Tucker+TT	**9.2** / 9.1	6.98 / 6.03	8.95 / 8.66	51.9 / 52.3	35.5 / 36.5	50.5 / 51.4	30.8 / 30.0	74.7 / 76.0	67.9 / 67.5	33.5 / 36.7	0.76 / 0.81
		TT all	9.2 / 9.1	6.98 / 6.01	8.95 / 8.65	51.9 / 51.9	35.5 / 36.1	50.5 / 51.3	30.8 / 30.0	74.8 / 75.0	68.1 / 67.2	33.6 / 36.8	0.76 / 0.82
4	16–19	Tucker MHA	4.0 / 4.0	6.03 / **5.46**	7.91 / **7.68**	52.7 / **53.6**	36.9 / **38.1**	52.6 / **54.0**	30.2 / **32.0**	76.0 / **76.2**	**68.0** / 67.6	**23.5** / 25.8	0.86 / **0.94**
		TT FFN	8.2 / 8.1	7.61 / 6.20	9.59 / 8.90	51.9 / 51.3	37.0 / 35.9	51.1 / 51.7	29.0 / 26.8	74.9 / 74.1	67.3 / 67.8	31.5 / 35.0	0.84 / 0.89
		Tucker+TT	**12.2** / 12.1	7.85 / 6.63	10.06 / 9.62	50.0 / 49.6	35.0 / 34.4	49.0 / 49.4	25.6 / 24.6	73.6 / 73.8	66.9 / 66.0	36.0 / 39.3	0.72 / 0.77
		TT all	12.2 / 12.1	7.86 / 6.63	10.06 / 9.65	50.0 / 49.8	34.9 / 34.1	49.0 / 49.5	25.2 / 25.6	73.6 / 73.6	67.0 / 66.4	36.0 / 38.9	0.72 / 0.78
5	16–20	Tucker MHA	5.1 / 5.0	6.65 / **5.77**	8.64 / **8.25**	50.9 / **51.9**	34.5 / 36.4	51.0 / **53.0**	26.0 / **27.6**	74.1 / **74.6**	**68.8** / 68.0	**25.3** / 27.2	0.84 / **0.92**
		TT FFN	10.2 / 10.2	8.71 / 6.80	10.99 / 9.73	50.7 / 50.1	**36.5** / 33.9	49.2 / 50.4	**27.6** / 26.2	73.6 / 73.0	66.8 / 66.9	33.5 / 36.1	0.81 / 0.84
		Tucker+TT	**15.3** / 15.2	9.94 / 7.74	12.34 / 11.09	47.1 / 47.4	30.6 / 30.8	45.2 / 46.8	23.2 / 23.2	70.0 / 71.0	66.3 / 65.4	38.3 / 42.3	0.68 / 0.74
		TT all	15.3 / 15.2	9.94 / 7.72	12.35 / 11.11	46.9 / 48.0	30.5 / 31.1	45.2 / 46.8	22.8 / 24.0	70.0 / 70.9	66.2 / 67.1	38.4 / 41.5	0.68 / 0.74
6	16–21	Tucker MHA	6.1 / 6.0	6.93 / **5.94**	9.07 / **8.50**	50.4 / **51.9**	34.4 / **36.3**	50.3 / **52.2**	25.8 / **28.2**	73.2 / **74.1**	68.5 / **68.6**	**26.7** / 28.1	0.83 / **0.91**
		TT FFN	12.3 / 12.2	10.26 / 7.52	12.51 / 10.98	49.4 / 49.1	35.6 / 32.3	47.2 / 48.8	26.2 / 25.6	71.3 / 72.0	66.9 / 66.9	35.2 / 38.3	0.79 / 0.81
		Tucker+TT	**18.3** / 18.2	12.45 / 8.81	14.96 / 12.77	45.7 / 46.4	29.8 / 28.7	43.0 / 45.2	23.0 / 24.4	67.6 / 68.6	65.3 / 65.2	40.4 / 44.5	0.64 / 0.71
		TT all	18.3 / 18.2	12.46 / 8.79	14.97 / 12.76	45.7 / 46.1	29.5 / 29.4	43.0 / 45.0	23.0 / 23.0	67.6 / 68.9	65.3 / 64.5	40.4 / 43.8	0.64 / 0.71
7	16–22	Tucker MHA	7.1 / 7.0	7.45 / **6.15**	9.74 / **8.97**	49.3 / **51.3**	32.8 / **35.6**	49.8 / **51.9**	23.4 / **27.2**	72.0 / **74.1**	**68.7** / 67.8	**28.1** / 29.3	0.81 / **0.91**
		TT FFN	14.3 / 14.2	11.52 / 8.19	14.14 / 12.24	46.9 / 47.7	32.0 / 31.1	45.1 / 47.5	23.0 / 23.8	69.3 / 70.3	65.0 / 65.9	36.7 / 40.7	0.76 / 0.80
		Tucker+TT	**21.4** / 21.2	17.60 / 10.39	19.45 / 15.22	42.7 / 44.2	27.2 / 27.5	39.5 / 42.4	19.2 / 21.2	64.3 / 66.2	63.5 / 63.8	42.2 / 46.8	0.61 / 0.68
		TT all	21.4 / 21.2	17.60 / 10.23	19.44 / 14.90	42.9 / 44.6	27.6 / 28.1	39.5 / 42.3	19.2 / 22.0	64.4 / 66.3	63.6 / 64.2	42.2 / 45.4	0.61 / 0.70
8	16–23	Tucker MHA	8.1 / 8.0	7.78 / **6.30**	10.14 / **9.36**	49.4 / **51.2**	33.0 / **35.8**	49.5 / **52.0**	24.8 / **26.8**	71.6 / **73.7**	**68.0** / **68.0**	**29.4** / 29.9	0.80 / **0.90**
		TT FFN	16.3 / 16.2	12.98 / 9.02	15.77 / 13.40	46.0 / 47.2	31.6 / 30.7	43.4 / 45.9	22.8 / 23.6	67.4 / 68.9	64.7 / 66.7	38.0 / 41.7	0.74 / 0.78
		Tucker+TT	**24.4** / 24.3	25.44 / 11.64	23.53 / 17.15	42.1 / 43.3	27.8 / 27.7	37.7 / 40.9	18.8 / 18.8	62.6 / 64.6	63.5 / 64.3	43.8 / 48.7	0.58 / 0.65
		TT all	24.4 / 24.3	25.61 / 11.35	23.59 / 16.61	42.0 / 43.8	27.7 / 27.8	37.7 / 41.1	18.4 / 20.6	62.6 / 64.5	63.5 / 65.2	43.9 / 47.1	0.58 / 0.68