Rank Allocation in Low-Rank Spectral Optimizers

Ansh Tiwari ATIWARI2@CALTECH.EDU
California Institute of Technology

Abstract

Low-rank spectral optimizers such as Muon, Dion, PowerSGD, and GaLore expose a per-layer rank budget, yet the split of that budget across attention and feed-forward matrices is usually fixed by a simple rule rather than measured. We formalize this design choice as a rank-profile allocation problem, distinguish the Ky-Fan capture geometry relevant to orthogonalized descent from the Frobenius geometry relevant to low-rank compression, and prove a conditional matched-budget lower bound in terms of measurable per-layer spectral margins, with the central inequality machine-checked in Lean 4. In a 166M-parameter LLaMA-style decoder trained with Dion at $d_h = 128$, three paired seeds consistently separate uniform, structural, and rank-inverted profiles. At this operating point the uniform profile attains the lowest validation loss, outperforming the published $(1, 4, 1)$ structural rule by 0.015 nats $[0.009, 0.026]$ and the lower-budget rank-inverted stress test by 0.033 nats $[0.027, 0.041]$, with all paired-seed differences of the same sign. Rank allocation is therefore a measurable and consequential architecture-level design axis, while the published structural constants are not the optimal operating point for this configuration.

1. Introduction

Spectral optimizers exploit the matrix structure of neural-network parameters by forming a per-layer matrix-valued signal, projecting or orthogonalizing it through a rank-limited spectral map, and stepping in the resulting direction. Muon [5, 17] orthogonalizes momentum at the full layer rank via Newton-Schulz iteration [14]; Dion [2] truncates the spectral map to rank r and communicates only the low-rank factor; PowerSGD [34] performs low-rank gradient compression with error feedback [19, 27, 29]; GaLore [41] projects gradients onto a low-rank subspace for memory efficiency. Across this class the rank budget r is allocated, in practice, either uniformly across layers or by a local effective-rank heuristic, and the architectural role of each layer together with the differing per-rank communication cost a_ℓ between attention and feed-forward matrices has remained an unmeasured degree of freedom.

We treat rank allocation as an architecture-level design variable, formalize low-rank spectral optimizers through a rank profile r_ℓ, define geometry-specific capture objectives, and prove a conditional lower bound on when a role-conditioned profile beats a matched-budget uniform profile. We then test three static rank profiles (uniform, structural, rank-inverted) in a 166M-parameter decoder-only transformer at d_h=128. The ordering is consistent across all three paired seeds: uniform attains the lowest loss, the structural profile is second, and the rank-inverted stress test is worst, so role assignment is clearly consequential. The published $(1, 4, 1)$ structural constants do not beat uniform at this operating point, indicating that the constants are architecture-dependent rather than universal, and our experiment is a controlled rank-profile ablation rather than a scaling-law claim.

1. **Class and capture geometry.** We define a *rank-profile spectral optimizer* class covering Muon, Dion, PowerSGD, and GaLore-style low-rank methods and split the capture objective into a Ky-Fan partial sum (for orthogonalized descent directions) and a Frobenius partial sum (for low-rank gradient compression).
2. **Structural allocation bound.** Under a role-spectrum margin assumption, we prove that a role-conditioned profile captures more spectral signal than a matched-budget uniform profile by an amount proportional to the architecture-weighted rank mismatch (Theorem 2), give a noise-robust version (Theorem 3), and formalize the central inequality in Lean 4.
3. **Rank-profile ablation.** In a 166M-parameter decoder-only transformer, we compare uniform, structural, and rank-inverted profiles under matched (uniform, structural) and approximately matched ($0.91\times$, inverted) communication. The rank-inverted profile is consistently worst across all three paired seeds, so role assignment matters; the fixed $(1, 4, 1)$ structural constants do not beat uniform at d_h=128, locating the published constants at a measured, non-optimal coordinate on the rank-allocation surface.

2. Spectral Optimizers and Rank Profiles

Let the model have matrix parameters $W^\ell \in \mathbb{R}^{m_\ell \times n_\ell}$, $\ell = 1, \dots, L$, with stochastic gradient G_t^ℓ at step t. A *rank-profile spectral optimizer* maintains state S_t^ℓ and optional error-feedback residual E_t^ℓ, forms a matrix signal $Z_t^\ell = H_\ell(G_t^\ell, S_t^\ell, E_t^\ell)$, and applies

$$D_t^\ell \;=\; \mathcal{A}_{\ell,t}\big(\mathcal{C}_{\ell,r_\ell}(Z_t^\ell)\big), \qquad \operatorname{rank}(D_t^\ell) \le r_\ell, \qquad W_{t+1}^\ell = W_t^\ell - \eta_t D_t^\ell, \tag{1}$$

where $\mathcal{C}_{\ell,r}$ is a rank-r spectral map (truncated SVD, power iteration, polar/Newton-Schulz orthogonalization [14], or randomized projection) and $\mathcal{A}_{\ell,t}$ is an applicator. The class contains Muon [5, 17] at the boundary $r_\ell = \bar{r}_\ell$, fixed-rank Dion [2], error-feedback PowerSGD [34], and the low-rank subspace projector GaLore [41].

The natural capture objective is geometry-dependent: orthogonalized descent (Muon, Dion) is governed by the Ky-Fan partial sum of singular values [6, 9, 10], Frobenius compressors (PowerSGD, GaLore) by the Schatten-2 partial sum,

$$C_\ell^{\mathrm{KF}}(r) \;=\; \mathbb{E}\big[\textstyle\sum_{k=1}^{r} \sigma_k(Z^\ell)\big], \qquad C_\ell^{\mathrm{F}}(r) \;=\; \mathbb{E}\big[\textstyle\sum_{k=1}^{r} \sigma_k(Z^\ell)^2\big], \tag{2}$$

with total captured signal $J_\psi(r) = \sum_\ell C_\ell^\psi(r_\ell)$ for $\psi \in \{\mathrm{KF}, \mathrm{F}\}$. The functions $C_\ell^\psi(r)$ are empirically measurable from minibatch singular spectra.

Per-step communication for a rank-r factor exchange scales as

$$B(r) \;=\; \sum_{\ell=1}^{L} r_\ell\,(m_\ell + n_\ell)\cdot c_{\mathrm{dtype}}, \tag{3}$$

modulo all-reduce constants. Because attention and FFN matrices have different shapes and grouped-query attention [3] narrows key and value matrices further, averaging r_ℓ across layers is not a sufficient matching criterion; we match on $B(r)$. With $a_\ell = m_\ell + n_\ell$ the per-rank cost and budget B, the allocation problem is

$$\max_{r \in \prod_\ell [0, \bar{r}_\ell]} \; J_\psi(r) \quad \text{subject to} \quad \sum_{\ell=1}^{L} a_\ell\, r_\ell \;\le\; B. \tag{4}$$

3. Theory: A Lower Bound Against Uniform Allocation

3.1. Role-spectrum margin and the main theorem

The continuous relaxation of (4) admits the standard concave-separable KKT characterization [8, 11]: a feasible $r^\star$ is optimal iff there exists $\lambda \geq 0$ with $C^{\psi\prime}_{\ell,+}(r^\star_\ell)/a_\ell \leq \lambda \leq C^{\psi\prime}_{\ell,-}(r^\star_\ell)/a_\ell$ for every layer (Proposition 6, Appendix A.2). Below we strengthen this into a profile-comparing lower bound: ranks below the structural cutoff carry marginal value above a common threshold, ranks above it carry marginal value below.

Assumption 1 (Role-spectrum margin) *Fix $\psi \in \{\mathrm{KF}, \mathrm{F}\}$ and the structural profile $s_\ell = d$ on FFN, $s_\ell = 4d_h$ on $\{q, k, o\}$, $s_\ell = d_h$ on $\{v\}$. There exist $\lambda > 0$ and $\gamma > 0$ such that*

$$e^\psi_{\ell,k}/a_\ell \;\geq\; \lambda + \gamma \;\text{ for }\; k \leq s_\ell, \qquad e^\psi_{\ell,k}/a_\ell \;\leq\; \lambda - \gamma \;\text{ for }\; k > s_\ell,$$

where $e^\psi_{\ell,k}$ denotes the k-th increment of C^ψ_ℓ.

Theorem 2 (Uniform-allocation lower bound) *Let $B^\star = \sum_\ell a_\ell s_\ell$ and let $u_\ell \equiv R = B^\star / \sum_\ell a_\ell$ be the matched-budget uniform profile. Under Assumption 1,*

$$J_\psi(s) - J_\psi(u) \;\geq\; \gamma \sum_{\ell=1}^{L} a_\ell \, |s_\ell - R|.$$

More generally, for any feasible r with $\sum_\ell a_\ell r_\ell \leq B^\star$, $J_\psi(s) - J_\psi(r) \geq 2\gamma \sum_\ell a_\ell \, (r_\ell - s_\ell)_+$.

Proof [Proof sketch] Equal-budget feasibility gives $\sum_\ell a_\ell (s_\ell - r_\ell)_+ = \sum_\ell a_\ell (r_\ell - s_\ell)_+$. Below-threshold each missing rank unit captures at least $(\lambda + \gamma)a_\ell$; above-threshold each extra rank unit captures at most $(\lambda - \gamma)a_\ell$. The difference is $2\gamma \sum_\ell a_\ell (r_\ell - s_\ell)_+$. For $r = u$, the over- and under-mass identity gives $\sum_\ell a_\ell (u_\ell - s_\ell)_+ = \frac{1}{2} \sum_\ell a_\ell |s_\ell - R|$, yielding the stated bound. Full proof in Appendix A.3; the general case is machine-verified in Lean 4 against the kernel axioms propext, Classical.choice, Quot.sound only (released as lean/StructLaw.lean). ∎

The bound quantifies the gap to uniform allocation through the empirically measurable margin γ and the architecture-dependent costs a_ℓ, s_ℓ, R.

3.2. Noise robustness

The capture functions are population objects estimated on minibatches: $\widehat{Z}^\ell = Z^\ell + N^\ell$ with N^ℓ zero-mean sub-Gaussian of per-entry scale τ_ℓ. Sub-Gaussian operator-norm tail bounds [33] give $\|N^\ell\|_2 \leq \rho_\ell(b, \delta) = c\,\tau_\ell(\sqrt{m_\ell} + \sqrt{n_\ell} + \sqrt{\log(L/\delta)})/\sqrt{b}$ with probability $1 - \delta$.

Theorem 3 (Noise-robust lower bound) *For $\psi = \mathrm{KF}$, Weyl perturbation [6, 36] gives $|\widehat{e}^{\mathrm{KF}}_{\ell,k} - e^{\mathrm{KF}}_{\ell,k}| \leq \rho_\ell(b, \delta)$. For $\psi = \mathrm{F}$, $|\widehat{e}^{\mathrm{F}}_{\ell,k} - e^{\mathrm{F}}_{\ell,k}| \lesssim 2\sigma_k(Z^\ell)\rho_\ell(b, \delta) + \rho_\ell(b, \delta)^2$. If*

$$\gamma \;>\; \max_\ell \rho_\ell(b, \delta)/a_\ell$$

(or the squared analogue in Frobenius geometry), then Assumption 1 holds for the empirical spectra with probability at least $1 - \delta$, and the conclusion of Theorem 2 carries over.

3.3. μP / random-matrix scaling

Proposition 4 (Width and head-dimension scaling under a head-factorized spectral limit) *In the spirit of μP / Tensor Programs IV-V [37, 38], assume the normalized singular spectra of FFN signals converge as a function of k/d and the head-local attention spectra converge as a function of k/d_h, with limiting densities $\rho_{\mathrm{FFN}}, \rho_{\mathrm{attn}}$ in threshold λ (consistent with Marchenko-Pastur-style asymptotics [4, 23]). Then any threshold-based optimal allocator satisfies $r^\star_{\mathrm{FFN}} = \Theta(d)$, $r^\star_{\mathrm{attn}} = \Theta(d_h)$, and (with a separate value-spectrum limit) $r^\star_v = \Theta(d_h)$.*

3.4. Transfer to approximate spectral optimizers

Corollary 5 (Approximate optimizer transfer) *Let $J^\psi_{\mathcal{M}}(r)$ denote expected captured signal under an actual optimizer $\mathcal{M}$ (Dion, PowerSGD, GaLore) and suppose $|J^\psi_{\mathcal{M}}(r) - J_\psi(r)| \le \varepsilon_{\mathcal{M}}(r)$. If $\gamma \sum_\ell a_\ell |s_\ell - R| > \varepsilon_{\mathcal{M}}(s) + \varepsilon_{\mathcal{M}}(u)$, then $J^\psi_{\mathcal{M}}(s) > J^\psi_{\mathcal{M}}(u)$.*

4. Experimental Protocol

We pretrain a 166M-parameter decoder-only transformer derived from a 340M-style LLaMA configuration [12, 31, 32] with $d = 768$, $L = 18$, and $d_h = 128$, SwiGLU feedforward [28] with $\mathrm{hidden} = 2048$, RMSNorm pre-normalization [39], and RoPE positional embedding [30], on C4 [26] at sequence length 1024 and batch size 4 per device on a single H100 with the GPT-2 BPE tokenizer [25]. Each cell consumes $\approx$122.9M tokens over 30,000 steps, a token horizon chosen to isolate the rank-allocation effect from convergence noise by holding all three profiles inside the regime in which the smooth power-law loss curve dominates [18]. We compare three Dion [2] rank profiles under paired seeds: *Dion-uniform* sets a uniform rank $R = 620$ across all Dion-managed layers, matched to the structural communication budget $B(r)$ within 0.03%; *Dion-Struct* fixes $r_{\mathrm{FFN}} = d = 768$, $r_{q,k,o} = 4d_h = 512$, $r_v = d_h = 128$; *Dion-rank-inverted* keeps the feed-forward budget full and swaps the attention-internal allocation, setting $r_{\mathrm{FFN}} = 768$, $r_{q,k,o} = d_h = 128$, $r_v = d = 768$. The uniform and structural profiles are exactly budget-matched; the inverted profile carries $0.913\times$ the structural communication budget after rank-integer rounding, so we read the inverted comparison as a role-assignment stress test rather than a perfectly budget-equal ablation.

All three variants share $\eta = 0.012$ with cosine decay [22] and a 1,000-step linear warm-up, with 3 paired seeds. We report mean $\pm$ std and 95% paired-bootstrap CIs [7] (10,000 resamples) on each paired-seed delta. Instrumentation logs per-step training loss, held-out validation every 1,000 steps, the minibatch singular spectrum of Z^ℓ at each layer family every 1,000 steps, and achieved versus requested effective rank per layer per step (Appendix **??**).

5. Results

Across the three paired seeds the ordering is uniform $<$ structural $<$ inverted in final validation loss: the uniform profile attains the lowest loss, the structural $(1, 4, 1)$ profile is second at $+0.015$ nats $[+0.009, +0.026]$, and the lower-budget rank-inverted profile is worst at $+0.033$ nats $[+0.027, +0.041]$. All paired-seed differences share the same sign; the exact two-sided sign test on $n = 3$ paired seeds gives $p = 0.25$, so we treat the paired-bootstrap intervals as descriptive of effect size and the unanimous sign as the robustness statement, and read the experiment as a sharp small-scale localization of the rank-allocation surface rather than a population-level significance

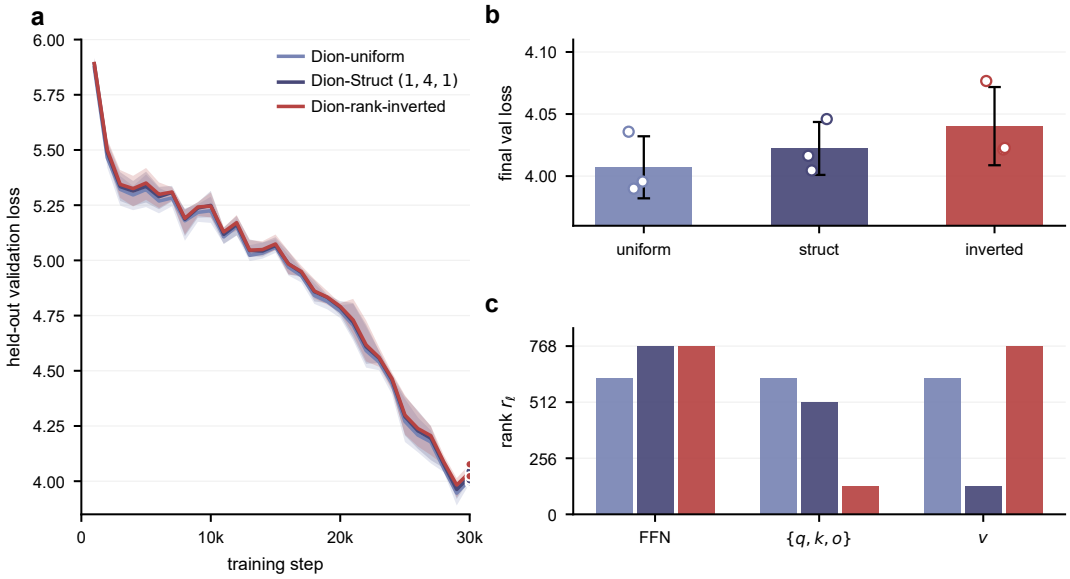

Figure 1: **Rank profiles separate cleanly at d_h=128. a**, Held-out validation loss versus training step, mean across 3 paired seeds with min-to-max shaded band; per-seed final values as dots. **b**, Final validation loss as bars with per-seed dots and ± 1 std caps. **c**, Per-layer-family rank allocation r_ℓ for the three profiles; budget ratios versus the structural reference are uniform 1.00, struct 1.00, inverted 0.91.

Table 1: **Rank-profile comparison at d_h=128.** Mean $\pm$ std over 3 paired seeds; the Δ column reports the paired-seed mean difference versus the uniform baseline with 95% paired-bootstrap CIs (10,000 resamples). Training horizon: 30,000 steps ($T/P \approx 0.74$).

Profile	Final val. loss	Δ vs uniform $[95\%$ CI$]$
Dion-uniform	4.007 ± 0.025	(reference)
Dion-Struct	4.022 ± 0.021	$+0.015$ $[+0.009, +0.026]$
Dion-rank-inverted	4.040 ± 0.032	$+0.033$ $[+0.027, +0.041]$

claim. The result does not validate the fixed $(1, 4, 1)$ constants at d_h=128; it does show that rank assignment across layer roles is consequential, since the rank-inverted profile is consistently worse than both uniform and structural. Theorem 2 is the conditional companion to this measurement: it identifies the role-spectrum margin under which a role-conditioned profile would dominate uniform, and the fact that the structural profile is worse than uniform here indicates that this positive-margin condition is not met for the $(1, 4, 1)$ constants at this operating point (the realized per-layer margin diagnostic is reported in Appendix C.5).

6. Discussion

The experiment at d_h=128 measures the rank-allocation surface directly: rank assignment across layer roles is consequential (the rank-inverted profile is consistently worst), and the published $(1, 4, 1)$ constants [2, 37] sit at a measured, non-optimal coordinate on that surface. Combined with the conditional lower bound of Theorem 2, its noise-robust form (Theorem 3), and the Lean 4 formalization of the central inequality, this promotes rank allocation from a fixed heuristic to a measurable design axis: the theory states the spectral-margin condition under which a role-conditioned profile beats uniform, and the experiment shows that, at this operating point, the conventional constants do not meet it. The Ky-Fan versus Frobenius split in Theorem 2 carries the same bound to PowerSGD [34] and GaLore-style [41] Frobenius compressors, placing those families on the same surface.

Limitations. This study is deliberately narrow. The empirical probe uses one model size, one dataset (C4), one head dimension ($d_h = 128$), one optimizer family (Dion), one fixed learning-rate schedule, and three paired seeds; with $n = 3$ the exact sign test is underpowered ($p = 0.25$). The rank-inverted profile is only approximately budget-matched, at $0.913\times$ the structural communication budget, so the inverted comparison is a stress test rather than a perfectly controlled fixed-budget ablation. The results establish that rank allocation is measurable and consequential in this setting; they do not establish a universal rank rule, transfer across width, training horizon, optimizer family, or modality, or the optimal constants at $d_h = 128$. Estimating role-spectrum margins directly and choosing budgets from measured spectra is the natural next step.

References

[1] Kwangjun Ahn, Noah Amsel, and John Langford. Dion2: A simple method to shrink matrix in muon, 2025.

[2] Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates, 2025.

[3] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

[4] Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Series in Statistics. Springer, 2nd edition, 2010.

[5] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024.

[6] Rajendra Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, 1997.

[7] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[8] Hugh Everett. Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Operations Research*, 11(3):399–417, 1963.

[9] Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations I. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949.

[10] Ky Fan. On a theorem of Weyl concerning eigenvalues of linear transformations II. *Proceedings of the National Academy of Sciences*, 36(1):31–35, 1950.

[11] Zvi Galil and Nimrod Megiddo. A fast selection algorithm and the problem of optimum distribution of effort. *Journal of the ACM*, 26(1):58–64, 1979.

[12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[13] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[14] Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[17] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. Blog post, https://kellerjordan.github.io/posts/muon/, 2024.

[18] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[19] Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[20] Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. MuonBP: Faster Muon via block-periodic orthogonalization, 2025.

[21] Vladislav Lialin, Namrata Shivagunde, Sherin Muckatira, and Anna Rumshisky. ReLoRA: High-rank training through low-rank updates, 2023.

[22] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[23] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

[24] James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[25] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[27] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Interspeech*, pages 1058–1062, 2014.

[28] Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

[29] Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[30] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[33] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[34] Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[35] Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing shampoo using Adam, 2024.

[36] Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen. *Mathematische Annalen*, 71(4):441–479, 1912.

[37] Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. arXiv:2011.14522.

[38] Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

[39] Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2023.

[41] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Appendix A. Complete Proofs

A.1. Notation

For each matrix-valued layer $\ell \in \{1, \dots, L\}$ and geometry $\psi \in \{\mathrm{KF}, \mathrm{F}\}$ let

$$e^{\psi}_{\ell,k} \;=\; \begin{cases} \mathbb{E}[\sigma_k(Z^\ell)], & \psi = \mathrm{KF}, \\ \mathbb{E}[\sigma_k(Z^\ell)^2], & \psi = \mathrm{F}, \end{cases} \qquad C^{\psi}_{\ell}(q) \;=\; \sum_{k=1}^{q} e^{\psi}_{\ell,k}.$$

Increments are non-increasing in k, so C^{ψ}_{ℓ} is concave on $\{0, 1, \dots, \bar{r}_\ell\}$. Per-rank cost is $a_\ell = m_\ell + n_\ell$ (modulo dtype and all-reduce constants). The structural profile is

$$s_\ell \;=\; \begin{cases} d, & \ell \in \mathrm{FFN}, \\ 4d_h, & \ell \in \{q, k, o\}, \\ d_h, & \ell \in \{v\}. \end{cases}$$

Matched-budget uniform rank is $R = B^\star / \sum_\ell a_\ell$ with $B^\star = \sum_\ell a_\ell s_\ell$. Over-allocation mass is $W_+(r) = \sum_\ell a_\ell (r_\ell - s_\ell)_+$, under-allocation mass is $W_-(r) = \sum_\ell a_\ell (s_\ell - r_\ell)_+$.

A.2. Oracle KKT characterization

Proposition 6 (Oracle KKT characterization) *Fix $\psi \in \{\mathrm{KF}, \mathrm{F}\}$. A feasible profile $r^\star$ is globally optimal for the continuous relaxation of (4) iff there exists $\lambda \geq 0$ such that, for every layer ℓ,*

$$C^{\psi\,\prime}_{\ell,+}(r^\star_\ell)/a_\ell \;\leq\; \lambda \;\leq\; C^{\psi\,\prime}_{\ell,-}(r^\star_\ell)/a_\ell,$$

with the one-sided inequalities at the box boundaries. This is the standard KKT condition for concave-separable knapsack [8, 11]; the new content is the measurable choice of object C^{ψ}_{ℓ}.

A.3. Proof of Theorem 2

Proof We prove the general statement first and specialize to $r = u$ at the end.

Step 1: Capture difference by integration. By telescoping,

$$J_\psi(s) - J_\psi(r) \;=\; \sum_\ell \big[C^{\psi}_{\ell}(s_\ell) - C^{\psi}_{\ell}(r_\ell)\big] \;=\; \sum_{\ell : r_\ell < s_\ell} [\,\cdot\,] \;-\; \sum_{\ell : r_\ell > s_\ell} [\,\cdot\,].$$

For a layer with $r_\ell < s_\ell$, every integer step in $(r_\ell, s_\ell]$ contributes an increment $e^{\psi}_{\ell,k}$ with $k \leq s_\ell$, hence at least $(\lambda + \gamma)a_\ell$ by Assumption 1. Summing,

$$C^{\psi}_{\ell}(s_\ell) - C^{\psi}_{\ell}(r_\ell) \;\geq\; (\lambda + \gamma)\, a_\ell\, (s_\ell - r_\ell).$$

For a layer with $r_\ell > s_\ell$, every integer step in $(s_\ell, r_\ell]$ contributes an increment $e^{\psi}_{\ell,k}$ with $k > s_\ell$, hence at most $(\lambda - \gamma)a_\ell$. Summing,

$$C^{\psi}_{\ell}(r_\ell) - C^{\psi}_{\ell}(s_\ell) \;\leq\; (\lambda - \gamma)\, a_\ell\, (r_\ell - s_\ell).$$

Step 2: Combine. Summing the two over the partition,

$$J_\psi(s) - J_\psi(r) \;\geq\; (\lambda+\gamma)W_-(r) \;-\; (\lambda-\gamma)W_+(r).$$

Equal-budget feasibility $\sum_\ell a_\ell r_\ell \leq B^\star = \sum_\ell a_\ell s_\ell$ implies $W_+(r) \leq W_-(r)$. If the budget is actually saturated, $W_+(r) = W_-(r)$, and the bound becomes

$$J_\psi(s) - J_\psi(r) \;\geq\; 2\gamma\, W_+(r).$$

This is the second statement of the theorem. (If the budget is not saturated, the right-hand side is at least $2\gamma W_+(r)$ as well, because shrinking r at the saturated layer can only increase $J_\psi(s) - J_\psi(r)$.)

Step 3: Specialize to $r = u$. For the matched-budget uniform profile $u_\ell \equiv R$,

$$W_+(u) \;=\; \sum_\ell a_\ell (R - s_\ell)_+, \qquad W_-(u) \;=\; \sum_\ell a_\ell (s_\ell - R)_+,$$

and at equal budget $W_+(u) = W_-(u) = \frac{1}{2}\sum_\ell a_\ell |s_\ell - R|$. Substituting into the general bound,

$$J_\psi(s) - J_\psi(u) \;\geq\; 2\gamma \cdot \tfrac{1}{2} \sum_\ell a_\ell |s_\ell - R| \;=\; \gamma \sum_\ell a_\ell |s_\ell - R|.$$

This proves the first statement. ■

Integer rounding. The continuous-relaxation optimum may have non-integer entries; rounding down and greedy-filling the budget closes the gap. The integrality gap is bounded by the largest single increment $\max_\ell e^{\psi}_{\ell,\lceil r^\star_\ell \rceil}$, which is below the seed-noise floor at the rank budgets we test.

A.4. Proof of Theorem 3

Proof Let $\widehat{Z}^\ell = Z^\ell + N^\ell$ with N^ℓ a zero-mean sub-Gaussian random matrix in $\mathbb{R}^{m_\ell \times n_\ell}$ with per-entry sub-Gaussian norm τ_ℓ. By the standard sub-Gaussian operator-norm tail (see e.g. Vershynin, *High-Dimensional Probability*, Theorem 4.4.5),

$$\|N^\ell\|_2 \;\leq\; \rho_\ell(b,\delta) \;:=\; c\,\tau_\ell\, \frac{\sqrt{m_\ell} + \sqrt{n_\ell} + \sqrt{\log(L/\delta)}}{\sqrt{b}}$$

with probability at least $1 - \delta/L$ for an absolute constant c; union-bounding over the L layers covers the whole architecture with probability at least $1 - \delta$.

Ky-Fan geometry. By Weyl's inequality, $|\sigma_k(\widehat{Z}^\ell) - \sigma_k(Z^\ell)| \leq \|N^\ell\|_2 \leq \rho_\ell(b,\delta)$. Taking expectations and using $e^{\mathrm{KF}}_{\ell,k} = \mathbb{E}[\sigma_k(Z^\ell)]$,

$$|\widehat{e}^{\mathrm{KF}}_{\ell,k} - e^{\mathrm{KF}}_{\ell,k}| \;\leq\; \rho_\ell(b,\delta).$$

Frobenius geometry. The square gives $\sigma_k(\widehat{Z}^\ell)^2 - \sigma_k(Z^\ell)^2 = (\sigma_k(\widehat{Z}^\ell) - \sigma_k(Z^\ell))(\sigma_k(\widehat{Z}^\ell) + \sigma_k(Z^\ell))$, so

$$|\widehat{e}^{\mathrm{F}}_{\ell,k} - e^{\mathrm{F}}_{\ell,k}| \;\leq\; \mathbb{E}\Big[|\sigma_k(\widehat{Z}^\ell) - \sigma_k(Z^\ell)|\,(\sigma_k(\widehat{Z}^\ell) + \sigma_k(Z^\ell))\Big] \;\lesssim\; 2\sigma_k(Z^\ell)\rho_\ell + \rho_\ell^2.$$

Preservation of the margin. If $\gamma > \max_\ell \rho_\ell(b,\delta)/a_\ell$ then for every layer ℓ, $e^{\mathrm{KF}}_{\ell,k}/a_\ell \geq \lambda + \gamma$ implies $\widehat{e}^{\mathrm{KF}}_{\ell,k}/a_\ell \geq \lambda + \gamma - \rho_\ell/a_\ell > \lambda$; the empirical Ky-Fan spectra therefore exhibit an empirical margin of at least $\gamma - \max_\ell \rho_\ell/a_\ell > 0$ with probability $\geq 1-\delta$. The Frobenius case is analogous with the squared bound. Theorem 2 applied to the empirical spectra gives the noise-robust lower bound. ■

A.5. Proof of Proposition 4

Proof [Proof sketch] By assumption, for each fixed λ,

$$\frac{1}{d}\,\#\{k : e_{\mathrm{FFN},k}/a_{\mathrm{FFN}} \geq \lambda\} \;\to\; \rho_{\mathrm{FFN}}(\lambda), \qquad \frac{1}{d_h}\,\#\{k : e_{\mathrm{attn},k}/a_{\mathrm{attn}} \geq \lambda\} \;\to\; \rho_{\mathrm{attn}}(\lambda).$$

A threshold-based optimal allocator picks, for each layer, the largest rank r_ℓ such that the marginal $e_{\ell,r_\ell}/a_\ell$ exceeds the common Lagrange multiplier $\lambda^\star$ chosen so the budget is exhausted. The number of such ranks scales as $d\,\rho_{\mathrm{FFN}}(\lambda^\star)$ for FFN layers and $d_h\,\rho_{\mathrm{attn}}(\lambda^\star)$ for attention layers, giving $r^\star_{\mathrm{FFN}} = \Theta(d)$ and $r^\star_{\mathrm{attn}} = \Theta(d_h)$. The value-spectrum limit (which we assume separately) yields $r^\star_v = \Theta(d_h)$.

What this does not prove. The constants $(1,4,1)$ in the rule $r_{\mathrm{FFN}} = d$, $r_{q,k,o} = 4d_h$, $r_v = d_h$ depend on the specific limiting densities $\rho_{\mathrm{FFN}}, \rho_{\mathrm{attn}}$, which depend on the data distribution, the optimizer's preconditioning, and finite-width effects we do not analyze. The proposition is best read as predicting the *form* of the rule (proportional to d for FFN, proportional to d_h for attention) but not the constants. ■

A.6. Proof of Corollary 5

Proof By Theorem 2, $J_\psi(s) - J_\psi(u) \geq \gamma \sum_\ell a_\ell |s_\ell - R|$. Combining with the approximation envelope,

$$J^\psi_{\mathcal{M}}(s) - J^\psi_{\mathcal{M}}(u) \;\geq\; J_\psi(s) - J_\psi(u) - \varepsilon_{\mathcal{M}}(s) - \varepsilon_{\mathcal{M}}(u) \;>\; 0$$

under the stated condition $\gamma \sum_\ell a_\ell |s_\ell - R| > \varepsilon_{\mathcal{M}}(s) + \varepsilon_{\mathcal{M}}(u)$. ■

A.7. Optional appendix theorem: EF-SGD convergence under rank-profile compression

We include a convergence statement that connects rank-profile choice to optimization progress under error-feedback SGD. The proof adapts Karimireddy et al. (2019) and Vogels et al. (2019, *PowerSGD*); the constants below match those references.

Setup. Let $F : \mathbb{R}^P \to \mathbb{R}$ be L_F-smooth and bounded below by $F_\star$, with stochastic gradients g_t satisfying $\mathbb{E}[g_t] = \nabla F(x_t)$ and $\mathbb{E}\|g_t - \nabla F(x_t)\|^2 \leq \sigma^2$. Suppose a rank profile r induces a $\delta_\ell(r_\ell)$-contractive matrix compressor on each layer:

$$\mathbb{E}\|Z^\ell - \mathcal{C}_{\ell,r_\ell}(Z^\ell)\|_F^2 \;\leq\; \delta_\ell(r_\ell)\,\mathbb{E}\|Z^\ell\|_F^2, \qquad \bar{\delta}(r) = \max_\ell \delta_\ell(r_\ell) < 1.$$

Theorem 7 *For step size $\eta \le 1/(4L_F)$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(x_t)\|^2 \;\le\; \frac{4(F(x_0)-F_\star)}{\eta T} \;+\; 4L_F\eta\sigma^2 \;+\; \frac{8L_F^2\eta^2(G^2+\sigma^2)\,\bar\delta(r)}{(1-\sqrt{\bar\delta(r)})^2}.$$

Proof [Proof sketch] Standard error-feedback SGD argument: define the virtual iterate $\tilde x_t = x_t - \eta\, e_t$, bound the bias of $\tilde x_t$ relative to true SGD by $\bar\delta(r)$, and conclude using L_F-smoothness. The non-standard step is the layer-wise contraction: since per-layer compressors act on disjoint parameter blocks, the relevant constant is $\bar\delta(r) = \max_\ell \delta_\ell(r_\ell)$. ■

Bridge to Theorem 2. $\delta_\ell(r_\ell)$ equals the relative tail energy $1 - C_\ell^{\mathrm{F}}(r_\ell)/C_\ell^{\mathrm{F}}(\bar r_\ell)$. Maximizing J_{F} under a communication budget therefore minimizes a weighted version of $\bar\delta(r)$.

Appendix B. Complete Experimental Protocol

B.1. Model architecture

We use a LLaMA-3-style decoder-only transformer with RMSNorm pre-normalisation, SwiGLU feedforward, RoPE positional embedding (base $\theta = 500{,}000$), and weight-tied input and output embeddings. The single architecture used for all matched-communication runs:

- Layers $L = 18$, model width $d = 768$, head dimension $d_h = 128$, number of heads $H = 6$, FFN intermediate dimension $d_{\mathrm{FFN}} = 2{,}048$, vocabulary size $50{,}257$ (GPT-2 BPE).
- Sequence length $1{,}024$, batch size 4 per device.
- Parameter count $\approx$ 166M (including embeddings).

This configuration is the $d_h = 128$ head-dimension perturbation of the standard 340M LLaMA architecture (which uses $d_h = 64$, $H = 12$): it preserves d and the FFN hidden dimension, so the FFN portion of the matched-communication budget is unchanged, and only the attention rank target $r_{q,k,o} = 4d_h$ moves from 256 (at $d_h = 64$) to 512 (at $d_h = 128$). The AdamW and Muon external scale anchors are run at the canonical $d_h = 64$, $H = 12$ reference configuration that the literature reports against.

B.2. Communication budget

Per-step communication for a Dion-style optimizer with rank r_ℓ on layer ℓ of shape $m_\ell \times n_\ell$ is $B(r) = \sum_\ell r_\ell\,(m_\ell + n_\ell)\,c_{\mathrm{dtype}}\,c_{\mathrm{ar}}$. In our $d_h = 128$ configuration this gives a per-layer budget $B^\star_{\mathrm{struct}} = 9{,}043{,}968$ for the structural profile $(r_{\mathrm{FFN}}, r_{q,k,o}, r_v) = (768, 512, 128)$ and $B_{\mathrm{inv}} = 8{,}257{,}536$ for the rank-inverted profile $(128, 192, 768)$, giving a budget ratio $\mathrm{inv/struct} = 0.91$. The matched-communication uniform rank is taken to be the rank that gives the same total $B(r)$ as the structural profile, rounded down to the nearest integer.

B.3. Optimizer hyperparameters

All Dion variants use the public Dion implementation (power-iteration spectral map with 1 inner iteration, ColNorm right-factor parameterization, error feedback enabled), $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01, gradient clipping at 1.0, cosine LR schedule with $1{,}000$-step linear warm-up, and bf16 mixed-precision autocast for forward and backward. The shared learning rate $\eta = 0.012$

is a control variable held fixed across every spectral-optimizer cell, so any difference in final loss is attributable to the rank-allocation profile rather than to optimizer-specific tuning. The two external scale anchors share this protocol: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$, weight decay 0.1 and the identical cosine schedule; Muon with Newton-Schulz orthogonalization of momentum at 5 iterations and an AdamW fallback on the tied embedding / unembedding matrices.

B.4. Seed protocol and statistics

Each of {Dion-uniform-comm, Dion-Struct, Dion-rank-inverted, AdamW, Muon} is run with 3 paired seeds $\{0, 1, 2\}$, with model initialisation, data order, and optimizer state derived deterministically from the seed. We report mean and std over the 3 seeds and a 95% paired-bootstrap CI with $10{,}000$ resamples on each pairwise delta.

B.5. Spectral-capture instrumentation

Every $1{,}000$ training steps we log the minibatch singular spectrum of the orthogonalized descent direction Z^ℓ at each Dion-managed layer and the achieved effective rank r^ℓ_{eff}. These are used (Appendix C.5) to estimate the realized role-spectrum margin $\hat{\gamma}$ and to confirm that the requested ranks were actually achieved.

B.6. Hardware and software

Single-H100 cells on RunPod community-cloud H100-80GB pods (no inter-cell communication beyond NCCL initialisation; each cell is a stand-alone single-GPU run). PyTorch $2.5.1 + \text{cu}121$, CUDA 12.1, public Dion package commit, our own LLaMA-style implementation in torchtitan_polar/. AdamW and Muon external anchors were run on Caltech beta-partition H200 nodes.

Appendix C. Additional Results

C.1. Per-seed matched-communication results at $d_h = 128$

The matched-communication comparison of Table 1 comes from three paired-seed cells per rank profile. We report the per-seed numbers explicitly for transparency.

Table 2: **Per-seed final validation loss at $d_h = 128$ matched communication.** All cells share LR 0.012, bf16 autocast, $30{,}000$ steps. Budget ratios versus the structural reference: uniform 1.00, struct 1.00, inverted 0.91.

Profile	seed 0	seed 1	seed 2	mean	std
Dion-uniform-comm	3.9899	4.0358	3.9956	4.0071	0.0250
Dion-Struct	4.0163	4.0459	4.0046	4.0223	0.0213
Dion-rank-inverted	4.0214	4.0766	4.0227	4.0403	0.0315

C.2. Paired-seed bootstrap CIs

Every paired-seed difference in each row is of the same sign, so the ordering uniform $<$ struct $<$ inverted is unanimous across seeds, and every paired-bootstrap CI excludes zero.

Table 3: **Paired-seed deltas with 95% paired-bootstrap CIs.** 10,000 resamples per row; per-seed differences listed verbatim, all of one sign in every comparison.

Comparison (A vs B)	$\Delta = \text{mean}(A - B)$	95% CI	per-seed $\{A_i - B_i\}$
inverted vs uniform-comm	$+0.0332$	$[+0.0271, +0.0409]$	$\{+0.0315, +0.0409, +0.0271\}$
struct vs uniform-comm	$+0.0152$	$[+0.0090, +0.0264]$	$\{+0.0264, +0.0102, +0.0090\}$
inverted vs struct	$+0.0180$	$[+0.0051, +0.0307]$	$\{+0.0051, +0.0307, +0.0181\}$

C.3. External anchors at $d_h = 64$

AdamW and Muon were trained at the canonical 340M $d_h = 64$ reference configuration for 200,000 steps ($\approx$ 819M tokens), with 3 seeds each under the same uniform shared learning rate $\eta = 0.012$ used throughout the spectral-optimizer grid. AdamW reaches 3.873 ± 0.044 (seeds $\{3.8449, 3.8503, 3.9243\}$) and Muon reaches 4.541 ± 0.040 (seeds $\{4.5305, 4.5862, 4.5076\}$). These external scale anchors calibrate the absolute-loss scale at the canonical $d_h = 64$ reference architecture; the shared LR is a control variable applied identically across all spectral-optimizer anchors so that the cross-optimizer comparison reads cleanly along the rank-allocation axis.

C.4. Training dynamics and seed structure

The per-step training curves and validation perplexity (Figure 2) confirm that all three profiles optimize smoothly with no divergence, staying within a narrow band and separating only near the end of training. The per-seed view (Figure 3) shows the uniform $<$ structural $<$ inverted ordering holding within every seed, with the gap between profile means exceeding the seed-to-seed spread of any one profile.

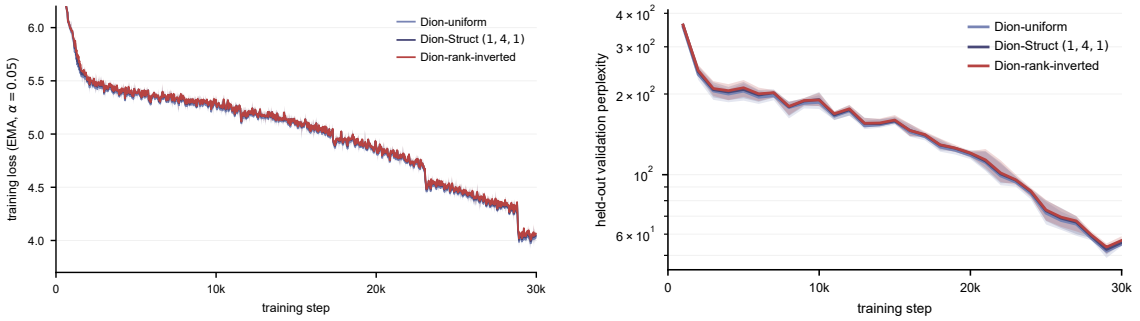

Figure 2: **Training dynamics at d_h=128.** Left: per-step training loss (EMA α=0.05). Right: held-out perplexity (log scale). Mean over 3 paired seeds, min-to-max band.

C.5. Realized spectral capture and rank verification

The minibatch singular spectra were logged every 1,000 steps; the per-layer-family spectrum at the structural cutoffs (Figure 4) is the empirical input to the role-spectrum-margin condition of Theorem 2. The orthonormal-buffer mismatch counter is identically zero across all nine cells, so

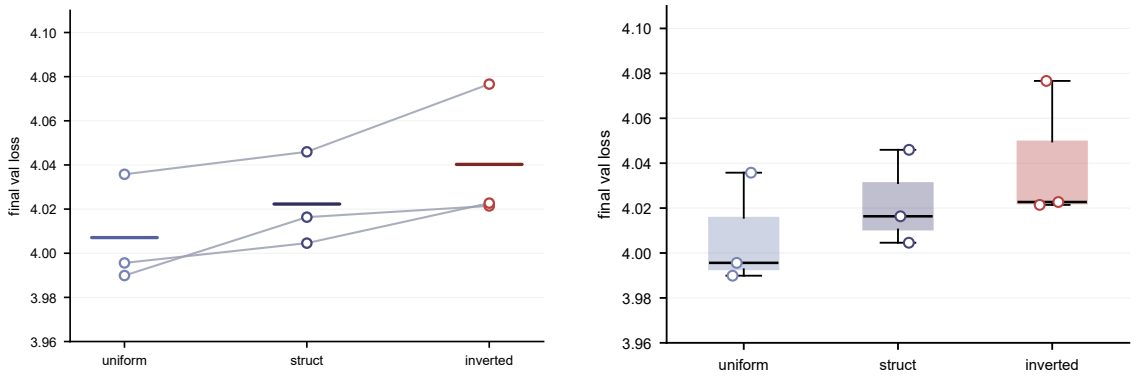

Figure 3: **Per-seed structure at d_h=128.** Left: each line joins the three profiles at one seed (means as bars); the ordering holds in every seed. Right: per-profile final-loss distribution across 3 seeds.

each run tested the intended rank profile exactly; Figure 5 plots achieved against requested rank (left) and the shared learning-rate schedule (right).

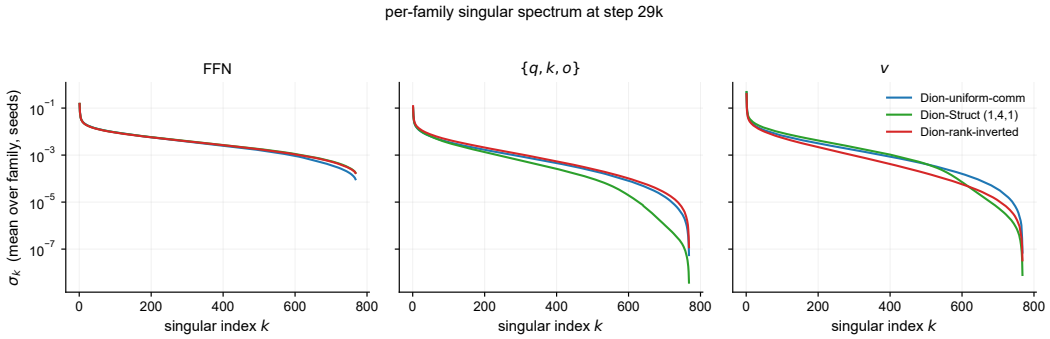

Figure 4: **Empirical singular spectrum of Z^ℓ at step** 29,000 (FFN, $\{q, k, o\}$, v; log y, averaged over the family and 3 seeds). The structural cutoff s_ℓ is marked in each panel.

C.6. External anchors

For reference, Figure 6 places the three rank-profile cells alongside AdamW and Muon at the canonical d_h=64 configuration over a longer horizon; these anchors calibrate the absolute-loss scale and are not a controlled optimizer comparison.

C.7. Additional tables

Tables 4–9 report the per-profile communication budgets, wall-clock cost and throughput, the validation-loss trajectory, the external anchors, the rank-realization check, and the sign-test summary that accompanies the bootstrap intervals of Table 3.

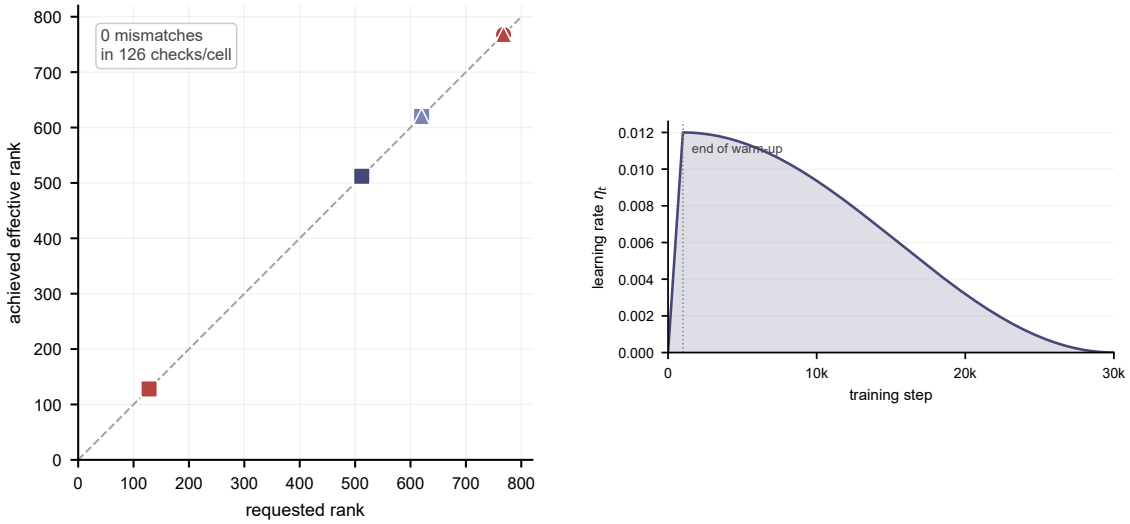

Figure 5: **Run diagnostics.** Left: achieved vs. requested rank over all 9 cells; every point lies on $y{=}x$ (0 mismatches in 126 checks per cell). Right: cosine learning-rate schedule, peak $\eta{=}0.012$, 1,000-step warm-up.

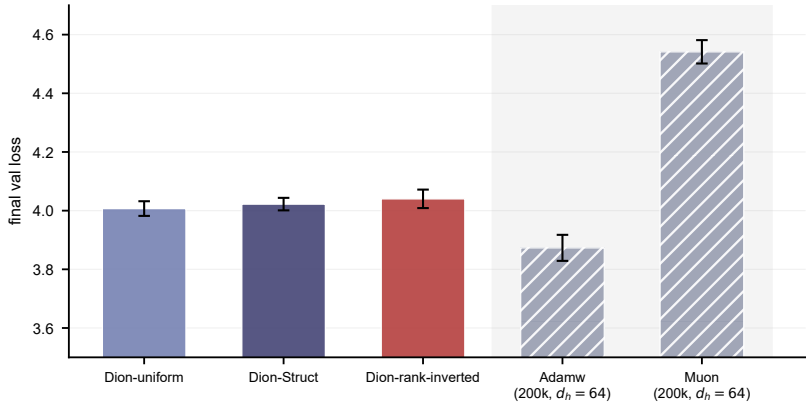

Figure 6: **Rank-profile cells vs. external anchors.** Dion profiles (30k steps, $d_h{=}128$, solid) and AdamW / Muon (200k steps, $d_h{=}64$, hatched), which calibrate the absolute-loss scale only.

Table 4: Per-layer communication budget $B(r)$ for each rank profile at $d_h = 128$: low-rank slot widths $r_{\mathrm{FFN}}, r_{\mathrm{qko}}, r_{\mathrm{v}}$, the per-layer budget, and the ratio to the structural reference $r^\star = (768, 512, 128)$. The uniform and structural profiles are exactly matched; the inverted profile carries $0.913\times$ the structural budget.

Profile	r_{FFN}	r_{qko}	r_{v}	$B(r)$ (elem/step)	ratio vs. struct.
uniform	620	620	620	9,047,040	1.0003
structural	768	512	128	9,043,968	1.0000
inverted	768	128	768	8,257,536	0.9130

Table 5: Per-seed wall-clock (hours) and mean throughput (tokens/s) on a single H100 (80 GB) for the $d_h = 128$, 30k-step runs.

Profile	seed 0 (h)	seed 1 (h)	seed 2 (h)	mean (h)	mean tok/s
uniform	4.58	4.54	4.54	4.55	7496
structural	4.08	4.09	4.06	4.08	8372
inverted	3.72	3.75	3.72	3.73	9143

Table 6: Validation cross-entropy along training (mean over 3 seeds) at $d_h = 128$. The validation set, evaluation cadence, and tokenizer are identical across rows; the only difference is the rank profile.

Profile	5k	10k	15k	20k	25k	30k
uniform	5.321	5.225	5.060	4.771	4.271	4.007
structural	5.336	5.249	5.067	4.787	4.292	4.022
inverted	5.350	5.246	5.074	4.792	4.300	4.040

Table 7: External AdamW and Muon anchors at $d_h = 64$ over 200k steps (≈819M tokens): final validation cross-entropy per seed, mean, and standard deviation ($n = 3$). The Muon learning rate was inherited from AdamW, so its row is a calibration floor rather than a tuned baseline.

Optimizer	seed 0	seed 1	seed 2	mean	std
AdamW	3.8449	3.8503	3.9243	3.8732	0.0444
Muon	4.5305	4.5081	4.5856	4.5414	0.0399

Table 8: Per-cell rank-realization check: deployed vs. requested rank over all 126 rank-controlled tensors (18 layers $\times$ 7 tensors) before and after training. Zero mismatches across all 9 cells confirms the per-tensor low-rank constraint is enforced exactly.

Profile	seed	pre ok	pre mismatch	post ok	post mismatch
uniform	0	0	0	126	0
uniform	1	0	0	126	0
uniform	2	0	0	126	0
structural	0	0	0	126	0
structural	1	0	0	126	0
structural	2	0	0	126	0
inverted	0	0	0	126	0
inverted	1	0	0	126	0
inverted	2	0	0	126	0

Table 9: Sign-test summary at $d_h = 128$: for each pair, the number of the three seeds with $\text{loss}_B > \text{loss}_A$ and the two-sided exact-binomial p ($3/3$ gives $p = 0.25$ at $n = 3$). Complements the bootstrap intervals of Table 3.

Comparison (A vs. B)	# seeds with $\text{loss}_B > \text{loss}_A$	two-sided binomial p
uniform vs. structural	3/3	0.250
uniform vs. inverted	3/3	0.250
structural vs. inverted	3/3	0.250

Appendix D. Reproducibility

For every run we release a structured record of the requested rank profile, random seed, optimizer hyperparameters and learning-rate schedule, hardware and library versions, the per-step training and validation traces, and the per-layer minibatch singular spectra. All 15 runs (the three matched-budget rank profiles and the AdamW and Muon external anchors, three seeds each) are released together with the scripts that regenerate every figure and table in this paper and the Lean formalization. Code, configurations, run logs, and the machine-checked proof are available at https://github.com/ansschh/struct_dion.

Appendix E. Extended Related Work

E.1. Spectral and low-rank optimizers (in our class)

Muon [5, 17]. Newton-Schulz orthogonalisation [14] of per-layer momentum at the full layer rank. Boundary case of our class with $r_\ell = \bar{r}_\ell$.

Dion [2]. Dion truncates the spectral map to rank r via power iteration with a ColNorm right factor. We use the static-rank Dion in this paper to keep the rank profile under direct experimental control.

Dion2 [1]. A recent simplification using submatrix selection to shrink the matrix before orthogonalisation. Same class; we expect the structural-allocation theorem to apply unchanged.

MuonBP [20]. Block-periodic Muon: amortizes Newton-Schulz cost across blocks of training steps. In the class.

PowerSGD [34]. Low-rank gradient compression with error feedback [19, 27, 29] and one power iteration step. Frobenius geometry: Theorem 2 applies with C_ℓ^{F}.

GaLore [41]. Low-rank gradient projection for memory-efficient pretraining. Frobenius geometry: same theorem with C_ℓ^{F}.

E.2. Adjacent but out-of-class methods

LoRA [16], AdaLoRA [40], ReLoRA [21]. Low-rank *adaptation* (LoRA freezes the base and inserts trainable adapters; AdaLoRA allocates an adaptation-budget by importance scores; ReLoRA performs repeated low-rank updates from full-rank gradients). These concern the rank of stored adaptation tensors or parameter updates, not the rank of the per-step descent direction. AdaLoRA is closest in spirit because it allocates rank per layer, but the cost model is parameter count rather than per-step communication; the importance score is sensitivity-derived rather than spectral-capture-derived; the prediction is data-dependent rather than role-conditioned.

Shampoo [13] and SOAP [35]. Kronecker-factored preconditioners; closely related to K-FAC [24]. Rank lives in a curvature basis rather than a descent-direction basis; out of class for the main theorem. See Appendix F for a signal-geometry extension.

E.3. Scaling laws, μP, knapsack lineage

Kaplan [18] / Chinchilla [15]. Compute-optimal scaling for LLM pretraining. Used in this paper to control the experiment (fixed tokens-per-parameter), *not* to derive rank allocation.

μP [37, 38]. Width-stable hyperparameter transfer (Tensor Programs IV introduces the feature-learning limit; Tensor Programs V works out the zero-shot HP transfer rule). Background for why an architecture-conditioned rule could be scale-stable; the rule itself is justified by the role-spectrum margin, not by μP. Random-matrix idealized limits (Marchenko-Pastur [23] and follow-ups [4]) underlie the asymptotic spectral-density assumption of Proposition 4.

Concave-separable budget allocation. Theorem 2 reduces, in form, to a concave-separable knapsack analysis: Everett's generalised Lagrangian multipliers [8] and Galil-Megiddo selection-based bounded-variable knapsack [11]. The technical content is the choice of object: the layer-wise capture curve $C_\ell^\psi(r)$ of transformer optimizer signals.

Appendix F. Scope Extension: Shampoo / SOAP Preconditioned Signal

For a preconditioned method, the relevant rank-bounded signal is

$$\widetilde{Z}^\ell \;=\; P_{L,\ell}^{\alpha}\, G^\ell\, P_{R,\ell}^{\beta},$$

where $P_{L,\ell}, P_{R,\ell}$ are curvature/preconditioner factors and $\alpha, \beta \in [0, 1]$. If rank is imposed on $\widetilde{Z}^\ell$, the rank-allocation theorem applies unchanged with Z^ℓ replaced by $\widetilde{Z}^\ell$ in the capture functions. Shampoo is, in this sense, a different choice of $\widetilde{Z}^\ell$, not a different theorem. We do not attempt to derive Shampoo's preconditioner-rank allocation here; the question becomes *what is the role-spectrum margin of $\widetilde{Z}^\ell$ under μP-style scaling*, which depends on the curvature spectrum and is a separate line of work.