# SAME ARCHITECTURE, DIFFERENT CAPACITY: OPTIMIZER-INDUCED SPECTRAL SCALING LAWS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Scaling laws have made language-model performance predictable from model size, data, and compute, but they typically treat the optimizer as a fixed training detail. We show that this assumption misses a fundamental axis of representation scaling: how effectively the optimizer converts added FFN width into utilized spectral capacity. Using eigenspectra of feed-forward network representations, measured through soft and hard spectral-ranks, we find that *the same Transformer architecture realizes markedly different spectral scaling laws when trained with different optimizers*. Holding architecture and width schedule fixed, AdamW exhibits weak hard-rank scaling ($\beta$=0.44) on rare-token (TAIL) representations where learning is known to be hardest, whereas Muon achieves linear scaling ($\beta$=**1.02**) in the same regimes, a $2.3\times$ increase in the scaling exponent. This difference is not reducible to validation loss: AdamW configurations can match low-rank Dion variants in perplexity, under extended training, while exhibiting sharply different spectral geometry, demonstrating that *matched loss does not imply matched representation structure*. Hard–soft rank asymmetry further reveals that optimizers differ not only in how much capacity is realized, but also in how that capacity is structured across eigenmodes. To disentangle optimizer effects from architectural ones, we compare against architectural interventions (e.g., attention rank and positional encoding), and find that optimizer-induced spectral shifts often exceed the architectural effects. These results suggest optimization as a first-class axis of representation scaling, motivating optimizer–architecture co-design.

## 1 INTRODUCTION

Classical scaling-law studies showed that language-model loss follows predictable power-law trends with model size, training data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). This resource-centric view has made scaling actionable: given a compute budget, one can estimate how to allocate resources across parameters and data. Yet a central component of training remains largely outside this framework—the optimizer. Growing evidence suggests that optimizers do more than affect convergence speed; they also shape the representations learned by a model through implicit inductive biases (Pascanu et al., 2025; Bernstein & Newhouse, 2024).

Recent work has begun to incorporate optimizer effects into loss-level scaling. (Volkova et al., 2026) show that loss-scaling exponents can remain shared across optimizers while multiplicative efficiency factors differ. This provides a useful abstraction for predicting validation loss, but it leaves open a representation-level question: *if two optimizers exhibit similar loss scaling, do they also learn similar internal representations?* If matched loss can arise from different representation geometries, then optimizer choice is not merely an efficiency knob. It is a design axis that determines how model capacity is structured across eigenmodes, allocated across token regimes, and realized during training.

We study this question through the eigenspectra of feed-forward network (FFN) representations. FFNs provide a natural setting for this analysis: in standard Transformer architectures, they account for roughly two-thirds of model parameters (Geva et al., 2021), and their expansion–nonlinearity–compression structure exposes a spectrally measurable latent space. This structure allows us to ask how efficiently added FFN width is converted into utilized spectral capacity. Building on prior work (Jha & Reagen, 2025), which established spectral scaling laws for FFN utilization under a fixed optimizer, we ask whether these laws are invariant to optimizer choice or instead depend on the



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

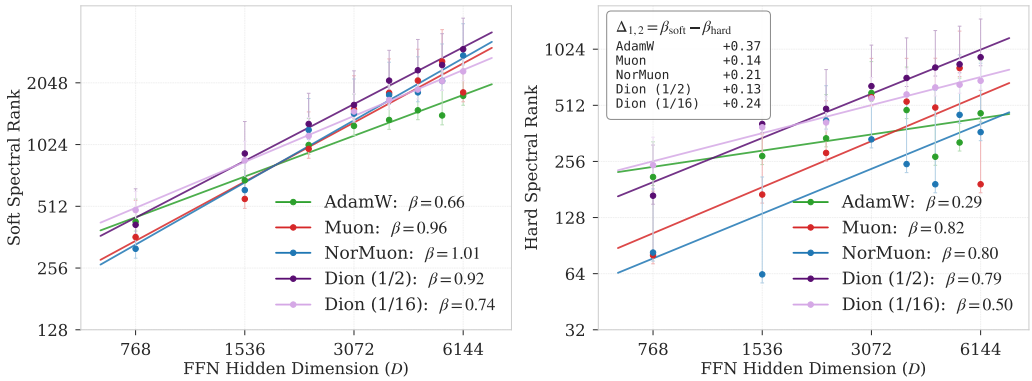



Figure 1: Spectral scaling exponents depend on optimizer choice: Soft (left) and hard spectral rank (right) as a function of FFN width in GPT-2 160M. AdamW exhibits the largest hard-soft asymmetry ($\Delta_{1,2} = 0.37$), indicating concentrated eigenspectra. Muon and Dion (1/2) reduce this asymmetry to $\Delta_{1,2} \approx 0.14$. Moreover, hard-rank scaling exhibits stronger dependence on optimizer choices.

architecture–optimizer pair. We compare AdamW (Loshchilov & Hutter, 2019), Muon (Jordan et al., 2024b; Liu et al., 2025), NorMuon (Li et al., 2026), and rank-constrained Dion variants (Ahn et al., 2025), measuring how spectral capacity is realized during training and how it scales with FFN width.

Holding the architecture and width schedule fixed, changing the optimizer alone yields markedly different spectral scaling laws (see Figure 1). We measure FFN capacity using soft spectral rank, which captures entropy-weighted spectral spread, and hard spectral rank, which is more sensitive to concentration in dominant eigenmodes (Jha & Reagen, 2025). The two metrics respond differently to optimizer choice: soft rank grows substantially with width across all optimizers, with scaling exponents clustered in a narrow range $\beta_{\text{soft}} \in [0.66, 1.01]$, whereas hard rank is strongly optimizer-dependent, spanning $\beta_{\text{hard}} \in [0.29, 0.82]$. AdamW exhibits weak hard-rank scaling ($\beta$=0.29), while Muon achieves near-linear scaling ($\beta$=0.82), under identical architecture and training data.

This result changes how spectral scaling should be interpreted. In prior work, the gap between soft-rank and hard-rank scaling appeared to be a stable property of the Pre-LN Transformer architecture (Jha & Reagen, 2025). We show that this hard–soft asymmetry is itself optimizer-dependent. AdamW exhibits the largest asymmetry ($\Delta_{1,2} = 0.37$), while Muon-style optimizers substantially reduce this gap; in particular, Muon and Dion$(1/2)$ both reach $\Delta_{1,2} \approx 0.14$. Hence, added FFN width is not automatically converted into usable capacity. The optimizer helps determine whether extra dimensions become dominant representational directions or remain diffuse spectral mass.

This spectral divergence is not apparent from validation loss alone. AdamW configurations can match matrix-aware optimizer variants in perplexity under extended training, while their representation spectra remain structurally distinct. Thus, *matched loss does not imply matched representation scaling*, and neither learning-rate tuning nor extended training closes this gap. Therefore, optimizer shapes the geometry of the learned representation, not just the convergence speed.

The optimizer effect is also structured across the token distribution. Language data follows a Zipfian distribution, and LLMs are known to struggle disproportionately with rare and long-tail knowledge (Kandpal et al., 2023). We therefore stratify token representations by frequency and measure spectral scaling separately across frequency regimes. AdamW exhibits especially weak hard-capacity scaling on rare-token representations, while the largest AdamW-to-Muon scaling gain appears in the mid-frequency regime. This shows that optimizer geometry changes not only aggregate representation capacity, but also how capacity is allocated across the token-frequency distribution.

Finally, we compare optimizer-induced effects against architectural interventions in attention rank (Amsel et al., 2025) and positional encoding (Su et al., 2024; Chi et al., 2023). Optimizer-induced spectral shifts often dominate or reshape these architectural effects, increasing per-head attention rank produces smaller spectral changes than switching optimizers, RoPE removal yields optimizer-dependent redistribution, and orthonormal optimizers enable partial PostLN configurations to reach meaningful perplexity where AdamW fails. These empirical observations support the view that the architectural capacity is not realized independently of optimization, architectural effects are modulated by optimizer geometry and vice-versa.

**Contributions.** Our contributions are as follows:

1. **Optimizer-induced spectral scaling laws.** We show that the same Transformer architecture realizes substantially different FFN spectral-capacity scaling laws depending on the optimizer. Hard–soft rank asymmetry is optimizer-dependent, exposing spectral scaling as a property of the architecture–optimizer pair rather than architecture alone. Rényi-entropy spectral analysis further confirms that the optimizer-induced differences persist across concentration regimes.
2. **Matched loss $\neq$ matched geometry.** We demonstrate that optimizer-induced spectral differences are not explained by learning-rate tuning, convergence speed, or final validation loss: configurations with matched perplexity can exhibit distinct spectral geometries.
3. **Frequency- and update-rank-dependent capacity allocation.** We show that optimizer-induced spectral scaling varies across token-frequency regimes, with mid- and low-frequency tokens showing the strongest effects. Dion rank further acts as a control knob for hard-capacity growth.
4. **Optimizer–architecture co-design.** We show that optimizer-induced spectral shifts can exceed or reshape the effects of architectural interventions, motivating joint optimizer–architecture design.

## 2 Related Work

**Scaling laws and optimizer-aware scaling.** Classical scaling laws established predictable power-law relationships between validation loss, model size, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022), while treating optimizer as a fixed training choice (Loshchilov & Hutter, 2019; Kingma, 2015; Shazeer & Stern, 2018). Recent work shows that scaling behavior can also depend on inductive biases beyond raw resources: (Ngo & Ravanbakhsh, 2026) find *architecture-dependent exponents* in neural force fields, where equivariant architectures achieve more favorable power-law slopes. Further, (Volkova et al., 2026) propose optimizer-aware loss scaling for LLM pretraining, modeling optimizer differences as multiplicative efficiency factors on shared exponents across AdamW, Muon, Shampoo (Gupta et al., 2018), SOAP (Vyas et al., 2025), and Scion (Pethick et al., 2025). This provides a useful abstraction in which optimizers act primarily as efficiency rescalings for loss-level prediction.

**Spectral capacity and effective rank.** Spectral measures have been used to characterize the effective dimensionality and utilization of learned representations (Roy & Vetterli, 2007; Garrido et al., 2023; Wei et al., 2024; Skean et al., 2025). Moving from loss to representation scaling, (Jha & Reagen, 2025) introduced spectral scaling laws for FFN latent-space utilization in LLMs under a fixed optimizer. In a complementary direction, (Nakis et al., 2026) showed that utilized capacity is different from the nominal capacity in graph models.

**Optimizer geometry.** Recent optimizer work suggests that training algorithms impose nontrivial geometry on matrix-valued parameters. Muon (Jordan et al., 2024b) orthogonalizes matrix updates via Newton–Schulz iterations (Kim & hwan Oh, 2026; Amsel et al., 2026). Large-scale studies further show that Muon-style and other matrix-based optimizers can be competitive for LLM pretraining, although their gains depend on tuning, scale, and evaluation protocol (Liu et al., 2025; Wen et al., 2026). Dion (Ahn et al., 2025) provides a particularly useful intervention for our analysis since its rank-constrained orthonormalized updates allow us to separate update geometry from update rank.

**Architecture–optimizer interaction.** (Ding et al., 2023) showed that architectural priors can be transferred into optimizers through gradient reparameterization, folding inductive biases that would normally live in the architecture into the optimizer update rule. PoLAR (Lion et al., 2025) provides another example of architecture–optimizer co-design, pairing structured low-rank parameterization with Riemannian optimization on the manifold induced by the parameterization.

## 3 Methodology

We measure how optimizer choice changes the effective latent capacity of FFNs under a fixed Transformer architecture. Three choices are central to the measurement. First, capacity is not one-dimensional: effective ranks with different concentration sensitivities can scale differently with FFN width. We therefore work within the Rényi effective-rank family rather than relying on a single rank estimate (Table 13). Second, probe location matters. Pre-activation states capture optimizer-induced

geometry before the FFN nonlinearity, while post-activation states capture the capacity realized after nonlinear transformation. Third, aggregate spectra can hide token-frequency-dependent effects; hence, we stratify FFN representations by token frequency. Throughout, *spectral capacity* denotes the effective dimensionality of variance-bearing directions in the FFN latent space.

### 3.1 FFN PROBE POINTS AND COVARIANCE SPECTRA

For layer $\ell$, we probe the FFN at two complementary states: the pre-activation state $z_{\ell,t}$ and the post-activation state $a_{\ell,t}$:

$$z_{\ell,t} = W_{\text{in}}^{(\ell)} x_{\ell,t}, \qquad a_{\ell,t} = \phi(z_{\ell,t}). \tag{1}$$

The two probes answer complementary questions. Pre-activation spectra expose the optimizer-shaped linear geometry before the FFN nonlinearity, whereas post-activation spectra measure the realized latent capacity passed to the output projection. Their comparison gives a three-stage view of FFN capacity: the optimizer shapes the linear expansion, the nonlinearity redistributes spectral mass, and the post-activation state determines the capacity available to subsequent layers.

Given FFN representations $h_n \in \mathbb{R}^D$ from either probe point, we compute the empirical covariance and trace-normalized eigenspectrum:

$$C = \frac{1}{N-1}\sum_{n=1}^{N}(h_n - \mu)(h_n - \mu)^\top, \quad \mu = \frac{1}{N}\sum_{n=1}^{N} h_n, \quad p_i = \frac{\lambda_i(C)}{\sum_{j=1}^{D}\lambda_j(C)}. \tag{2}$$

The distribution $p$ describes how variance is allocated across FFN latent directions, with $p_i \geq 0$ and $\sum_i p_i = 1$. Trace normalization makes spectra comparable across probe points, layers, and widths.

### 3.2 RÉNYI-FAMILY EFFECTIVE RANK

**Rényi entropy and the effective-rank.** Using the normalized eigenspectrum $p$, we quantify spectral spread through the Rényi entropy family (Rényi, 1961; Van Erven & Harremos, 2014). The order $\alpha$ controls concentration sensitivity: lower orders give more weight to weak eigendirections and diffuse spectral support, while higher orders increasingly emphasize dominant eigendirections. We define

$$H_\alpha(p) = \begin{cases} \frac{1}{1-\alpha}\log\sum_{i=1}^{D} p_i^\alpha, & \alpha > 0,\ \alpha \neq 1, \\ -\sum_{i=1}^{D} p_i \log p_i, & \alpha \to 1, \end{cases} \qquad R_\alpha(p) = \exp(H_\alpha(p)). \tag{3}$$

For $\alpha \neq 1$, this gives $R_\alpha(p) = (\sum_i p_i^\alpha)^{1/(1-\alpha)}$. Thus, $R_\alpha$ defines a continuum of effective-rank measures with different concentration sensitivities, unifying entropy-based effective rank (Roy & Vetterli, 2007; Garrido et al., 2023; Wei et al., 2024) and participation-ratio effective dimension (Gao et al., 2017; Hu & Sompolinsky, 2022; Xie et al., 2025; Ruscio et al., 2025) within a single information-theoretic framework. Table 13 summarizes how different $\alpha$ probe different aspects of the spectrum.

**Soft and hard rank as primary anchors.** For the main scaling-law analyses, we anchor at $\alpha = 1$ and $\alpha = 2$, which correspond to two standard notions of effective rank:

$$R_1(p) = \exp\Big(-\sum_i p_i \log p_i\Big) \quad \text{(soft rank)}, \qquad R_2(p) = \frac{1}{\sum_i p_i^2} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2} \quad \text{(hard rank)}. \tag{4}$$

The soft rank $R_1$ measures Shannon-like entropy-weighted spectral spread, while the hard rank $R_2$ is the participation ratio and provides a stricter, concentration-sensitive measure of effective dimensionality. These two anchors capture complementary aspects of spectral capacity: $R_1$ is sensitive to diffuse spread across many directions, whereas $R_2$ is more strongly affected by dominant eigendirections. The full Rényi sweep across $\alpha \in \{0.5, 1, 1.5, 2, 3, 5\}$ is reported in Appendix C

**Hard–soft asymmetry.** Since Rényi entropy is non-increasing in $\alpha$, $R_1(p) \geq R_2(p)$ for any spectrum (Van Erven & Harremos, 2014). We define the rank-level hard–soft asymmetry as

$$A_{1,2}(p) = \log R_1(p) - \log R_2(p) \geq 0, \tag{5}$$

with larger values indicating more concentrated eigenspectra. For scaling-law fits, we report the corresponding exponent-level asymmetry

$$\Delta_{1,2} = \beta_{\text{soft}} - \beta_{\text{hard}}, \tag{6}$$

where $\beta_{\text{soft}}$ and $\beta_{\text{hard}}$ are obtained by fitting $R_1(D)$ and $R_2(D)$ as functions of FFN width $D$. Higher asymmetry indicates that added width expands low-variance directions more than the dominant ones.

### 3.3 TOKEN-FREQUENCY STRATIFICATION

Aggregate spectra can be dominated by frequent tokens and may obscure the capacity scaling for rarer tokens. Motivated by the long-tailed structure of language data, we stratify FFN representations by token frequency, where frequency regimes are defined over token *types*, rather than knowledge concepts or factual entities. This connects most directly to token-level analyses of frequency-dependent scaling (Kunstner & Bach, 2025), while the broader difficulty of rare and long-tail knowledge in language models is supported by (Kandpal et al., 2023). Let $f(v)$ be the corpus frequency of token type $v \in \mathcal{V}$ and $M = \sum_{v\in\mathcal{V}} f(v)$ be the total occurrence. We sort token types by decreasing frequency and choose thresholds $\tau_{\text{head}}$ and $\tau_{\text{mid}}$ so that the top regime covers approximately one third of occurrences and the top two regimes cover approximately two thirds:

$$\sum_{v:f(v)\geq\tau_{\text{head}}} f(v) \approx \frac{M}{3}, \quad \sum_{v:f(v)\geq\tau_{\text{mid}}} f(v) \approx \frac{2M}{3}, \; b(v) = \begin{cases} \text{HEAD}, & f(v) \geq \tau_{\text{head}}, \\ \text{MID}, & \tau_{\text{mid}} \leq f(v) < \tau_{\text{head}}, \\ \text{TAIL}, & f(v) < \tau_{\text{mid}}. \end{cases} \quad (7)$$

HEAD contains the most frequent token types covering the top third of occurrence mass, MID covers the next third, and TAIL contains the remaining lower-frequency types. For each regime $b$, we compute covariance spectra and rank metrics on $\mathcal{H}_b = \{h_n : b(v_n) = b\}$. This stratification allows us to understand how optimizer-induced spectral-capacity varies across the token distribution.

### 3.4 SCALING LAWS FOR EFFECTIVE FFN CAPACITY

We vary the FFN hidden dimension as $D = md_{\text{model}}$ with $m \in \mathbb{Z}$ and compute post-activation soft and hard ranks for each layer, across token-frequency regime (HEAD, MID, TAIL). We fit

$$R(D) \propto D^{\beta} \qquad \Longleftrightarrow \qquad \log R(D) = \beta \log D + c, \quad (8)$$

where $R \in \{R_1, R_2\}$ is computed on aggregate or frequency-stratified spectra. The same machinery applies to any Rényi order $\alpha$, allowing us to fit a scaling exponent $\beta_\alpha$ at different concentration sensitivities. The main analyses focus on $\alpha = 1$ and $\alpha = 2$, with the broader $\alpha$ sweep reported in Appendix C. The exponent $\beta$ measures how efficiently added FFN width is converted into effective latent capacity in the probed concentration regime.

## 4 OPTIMIZER-INDUCED SPECTRAL SCALING LAWS

**Experimental setup** We train GPT-style decoder-only Transformers on FineWeb-Edu (Penedo et al., 2024) following the modded-nanoGPT (Jordan et al., 2024a) configuration: Pre-RMSNorm, RoPE (Su et al., 2024), squared-ReLU (So et al., 2021) FFNs, no biases, and QK-normalization (Dehghani et al., 2023). Our primary experiments use 160M base models, and we replicate the main scaling trends on 350M base models. The 160M and 350M labels denote the base ($4\,d_{\text{model}}$) configuration; we vary the FFN hidden dimension as $D = m\,d_{\text{model}}$ with $m \in \{1, \ldots, 8\}$ at 160M and $m \in \{1, 2, 3, 4\}$ at 350M, hence total parameter count grows with FFN width across the sweep. We compare AdamW (Loshchilov & Hutter, 2019), Muon (Jordan et al., 2024b), NorMuon (Li et al., 2026), and Dion (Ahn et al., 2025) at rank fractions $r \in \{1/2, 1/4, 1/8, 1/16\}$. The 160M variants are trained on 3.15B tokens, the 350M variants on 4.19B tokens, both with sequence length of 512 and global batch size 1024. Refer to Appendix A for more training details and setup.

### 4.1 SPECTRAL SCALING LAWS ARE OPTIMIZER-DEPENDENT

We perform spectral scaling analysis separately across token-frequency regimes to see how capacity is allocated for frequent vs. rare tokens, and whether the optimizer effects are concentrated in particular frequency regimes. Figure 2 depicts the scaling trends, and Table 1 reports their numerical values.

**Hard-rank scaling is strongly optimizer-dependent.** Optimizer choice strongly affects hard spectral rank scaling, which measures the growth of dominant-mode capacity. In the TAIL regime, AdamW reaches only $\beta_{\text{hard}} = 0.44$, whereas Muon and NorMuon achieve linear scaling with $\beta_{\text{hard}} = 1.02$ and $1.04$, respectively. The separation is also large in the MID regime: AdamW obtains $\beta_{\text{hard}} = 0.24$, compared with $0.93$ for Muon and $0.95$ for NorMuon. HEAD tokens show weaker and less reliable hard-rank separation, with $\beta_{\text{hard}} \in [0.26, 0.59]$ and lower fit quality. Thus, MID and TAIL tokens are the most diagnostic regimes for optimizer-induced scaling effects.

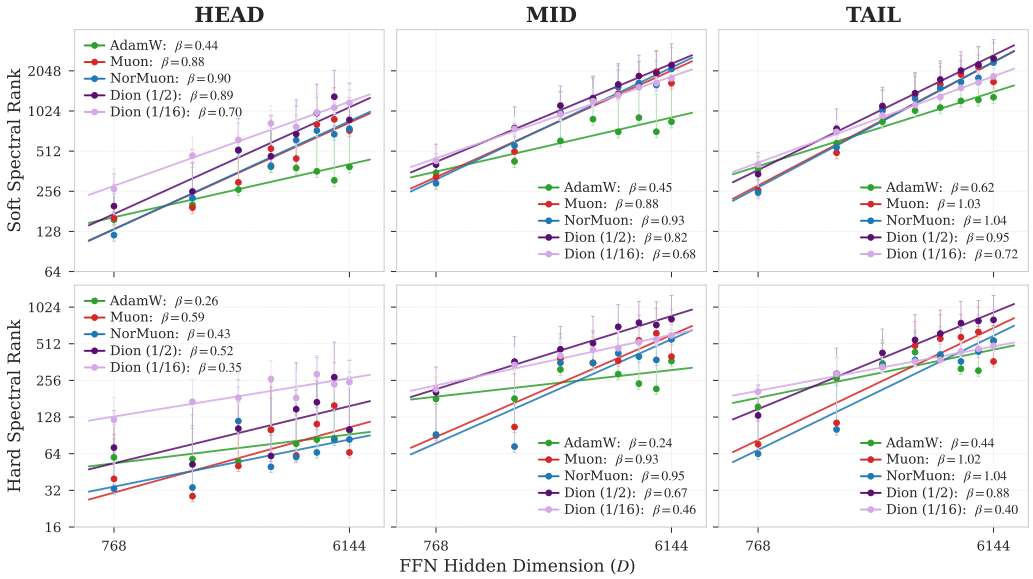



Figure 2: Optimizer-dependent spectral scaling across token-frequency regimes. Soft spectral rank (top) and hard spectral rank (bottom) are shown as functions of FFN hidden width for HEAD, MID, and TAIL tokens in GPT-2 160M. AdamW exhibits persistent positive hard–soft asymmetry, indicating that added width contributes mostly to diffuse spectral capacity rather than dominant-mode capacity. Muon and NorMuon reduce this asymmetry to near zero for MID and TAIL tokens, whereas low-rank Dion ($r = 1/16$) retains AdamW-like asymmetry despite orthonormalized updates.

**MID tokens show the largest AdamW-to-Muon gain.** A finer frequency-dependent pattern appears when comparing AdamW and Muon. Under AdamW, HEAD and MID tokens scale similarly ($\beta_{\text{hard}} = 0.26$ and $0.24$), while TAIL tokens scale more strongly ($\beta_{\text{hard}} = 0.44$). Muon changes this structure: MID rises to $\beta_{\text{hard}} = 0.93$, nearly matching TAIL ($\beta_{\text{hard}} = 1.02$), while HEAD remains lower ($\beta_{\text{hard}} = 0.59$). The AdamW-to-Muon gains are

$$\Delta\beta_{\text{MID}} = +0.69, \qquad \Delta\beta_{\text{TAIL}} = +0.58, \qquad \Delta\beta_{\text{HEAD}} = +0.33.$$

Thus, the MID gain is the largest, roughly $2.1\times$ the HEAD gain.

Table 1: $\beta$ values for soft and hard ranks with $R^2$ in parentheses for GPT-2 160M (Figure 2). Positive $\Delta_{1,2}$ indicates concentrated eigenspectra while lower values indicates better utilization of FFN width. Muon and NorMuon achieve near-zero $\Delta_{1,2}$ for MID/TAIL tokens, while AdamW maintains $\Delta_{1,2} \approx +0.2$ across all the token-frequency regimes.

| | HEAD | | | MID | | | TAIL | | |
|---|---|---|---|---|---|---|---|---|---|
| Optimizer | $\beta_{\text{hard}}$ | $\beta_{\text{soft}}$ | $\Delta_{1,2}$ | $\beta_{\text{hard}}$ | $\beta_{\text{soft}}$ | $\Delta_{1,2}$ | $\beta_{\text{hard}}$ | $\beta_{\text{soft}}$ | $\Delta_{1,2}$ |
| AdamW | 0.26 (0.59) | 0.44 (0.82) | +0.18 | 0.24 (0.36) | 0.45 (0.82) | +0.21 | 0.44 (0.66) | 0.62 (0.97) | +0.18 |
| Muon | 0.59 (0.54) | 0.88 (0.90) | +0.29 | 0.93 (0.82) | 0.88 (0.96) | −0.04 | 1.02 (0.81) | 1.03 (0.94) | +0.01 |
| NorMuon | 0.43 (0.45) | 0.90 (0.92) | +0.47 | 0.95 (0.77) | 0.93 (0.98) | −0.02 | 1.04 (0.89) | 1.04 (0.98) | +0.00 |
| Dion (1/2) | 0.52 (0.43) | 0.89 (0.90) | +0.37 | 0.67 (0.98) | 0.82 (0.99) | +0.15 | 0.88 (0.98) | 0.95 (0.99) | +0.07 |
| Dion (1/16) | 0.35 (0.75) | 0.70 (0.98) | +0.35 | 0.46 (0.98) | 0.68 (1.00) | +0.22 | 0.40 (0.99) | 0.72 (1.00) | +0.31 |

**Hard–soft asymmetry reveals how FFN width is utilized.** AdamW exhibits persistent positive asymmetry across all frequency regimes ($\Delta_{1,2} \approx 0.18$–$0.21$), indicating that added width contributes mostly to diffuse spectral capacity rather than dominant-mode capacity. In contrast, Muon and NorMuon nearly eliminate this asymmetry for MID and TAIL tokens; hence dominant-mode capacity scales at nearly the same rate as entropy-weighted spectral spread in the critical frequency regimes.

**Low-rank optimizer structure constrains hard-capacity scaling.** Dion separates orthonormalization from update rank. With rank fraction $r = 1/2$, Dion approaches Muon/NorMuon in the TAIL regime, reaching $\beta_{\text{hard}} = 0.88$ with small asymmetry ($\Delta_{1,2} = 0.07$). At $r = 1/16$, however, TAIL hard-rank scaling drops to $\beta_{\text{hard}} = 0.40$, comparable to AdamW, while the asymmetry rises to $\Delta_{1,2} = 0.31$. Thus, orthonormalization alone is insufficient: the rank of the optimizer update constrains how efficiently added FFN width becomes usable hard spectral capacity.

**Robustness of the scaling trends.** Soft-rank fits are consistently strong, indicating that added width reliably increases entropy-weighted spectral spread. Some HEAD and MID hard-rank fits have lower $R^2$, so we interpret those exponents as directional evidence of scaling behavior rather than precise constants. To verify that the aggregate trends are not artifacts of layer averaging, we also fit layer-wise exponents $\beta_\ell$ independently for each layer. Appendix Fig. 7 reports the resulting $\beta_\ell$ distributions, and Appendix Fig. 8 shows their depth profiles.

### 4.2 Matched Loss Does Not Imply Matched Spectral Geometry

A plausible explanation for AdamW's lower spectral-scaling exponents is slower convergence: perhaps longer training is required to match the scaling behavior of matrix-aware optimizers. We test this convergence-only explanation by training AdamW for 12K steps and comparing it with Dion($1/16$) at 6K steps, which achieves similar validation perplexity. This comparison tests whether matching loss is sufficient to recover the same spectral-capacity scaling.

**Extended AdamW training does not recover hard-rank scaling.** Table 2 rejects the convergence-only explanation. Although AdamW 12K matches Dion $(1/16)$ in perplexity (Table 9 in Appendix B.1), its aggregate hard-rank scaling nearly vanishes: $\beta_{\mathrm{hard}}$ drops from $0.29$ at 6K steps to $0.03$ at 12K steps. This degradation is also visible in HEAD, MID, and TAIL regimes, as their $\beta_{\mathrm{hard}}$ decrease to $0.13$, $0.17$, and $0.18$, respectively. In contrast, Dion $(1/16)$ maintains reliable power-law scaling, with $\beta_{\mathrm{hard}} = 0.50$ in aggregate and strong fit quality across frequency regimes ($R^2 = 0.75$ in HEAD and $R^2 \geq 0.98$ in MID/TAIL). Note that the aggregate soft-rank scaling decreases only mildly, from $\beta_{\mathrm{soft}} = 0.66$ to $0.58$, while hard–soft asymmetry increases from $\Delta_{1,2} = +0.37$ to $+0.55$. Thus, longer AdamW training fails to convert added FFN width into dominant-mode capacity.

Table 2: Extended AdamW training breaks hard-rank scaling despite matched loss in GPT-2 160M. AdamW 12K and Dion $(1/16)$ 6K achieve similar perplexity across width points, but exhibit different spectral geometry: AdamW's aggregate hard-rank scaling nearly vanishes ($\beta_{\mathrm{hard}}$=0.03, $R^2$=0.01), whereas Dion maintains reliable power-law scaling. $R^2$ values are shown in parentheses.

| | Aggregate | | | HEAD | | MID | | TAIL | |
|---|---|---|---|---|---|---|---|---|---|
| Configuration | $\beta_{\mathrm{hard}}$ | $\beta_{\mathrm{soft}}$ | $\Delta_{1,2}$ | $\beta_{\mathrm{hard}}$ | $\Delta_{1,2}$ | $\beta_{\mathrm{hard}}$ | $\Delta_{1,2}$ | $\beta_{\mathrm{hard}}$ | $\Delta_{1,2}$ |
| AdamW 6K | 0.29 (0.34) | 0.66 (0.97) | +0.37 | 0.26 (0.59) | +0.18 | 0.24 (0.36) | +0.21 | 0.44 (0.66) | +0.18 |
| AdamW 12K | **0.03 (0.01)** | 0.58 (0.91) | **+0.55** | 0.13 (**0.12**) | **+0.28** | 0.17 (0.42) | **+0.30** | 0.18 (**0.29**) | **+0.35** |
| Dion (1/16) 6K | 0.50 (0.97) | 0.74 (1.00) | +0.24 | 0.35 (0.75) | +0.35 | 0.46 (0.98) | +0.22 | 0.40 (0.99) | +0.31 |

**Hard-rank scaling breaks dynamically during training.** Figure 3 shows the diminishing return for hard-rank scaling in TAIL regime. TAIL hard-rank scaling initially improves, peaks ∼4K steps, and then declines. HEAD degrades more steadily after the early transient, while MID shows intermediate degradation. Soft-rank scaling remains comparatively stable throughout, leads to growing hard–soft asymmetry in every frequency regime, including TAIL ($+0.18 \rightarrow +0.35$). The underlying participation-ratio trajectories explain why the hard-rank power law fails: larger FFN widths lose PR capacity faster, breaking the monotonic width–capacity ordering required for a clean scaling law (Appendix E). This show that optimizer-induced spectral scaling is not a transient convergence artifact, and matched perplexity does not imply matched spectral capacity.

We also show that AdamW–Muon spectral scaling gap is not closed by learning-rate tuning. Across the learning-rate sweep, Muon's lowest TAIL hard-rank exponent remains above AdamW's highest valid exponent, indicating that the gap is not a simple hyperparameter artifact (Appendix F).

### 4.3 Update Rank Constrains Hard-Capacity Scaling

We next isolate which component of orthonormalized-update optimizers drives spectral capacity growth. Dion provides a controlled intervention: varying the rank fraction $r$ changes the rank of the projected update while preserving the orthonormalized-update structure. This lets us test whether update rank controls how added FFN width is converted into soft and hard spectral capacity.

**Low update rank primarily limits rare-token hard capacity.** Figure 4 shows the scaling effect in the TAIL tokens. As the Dion rank fraction decreases from $r = 1/2$ to $r = 1/16$, TAIL hard-rank

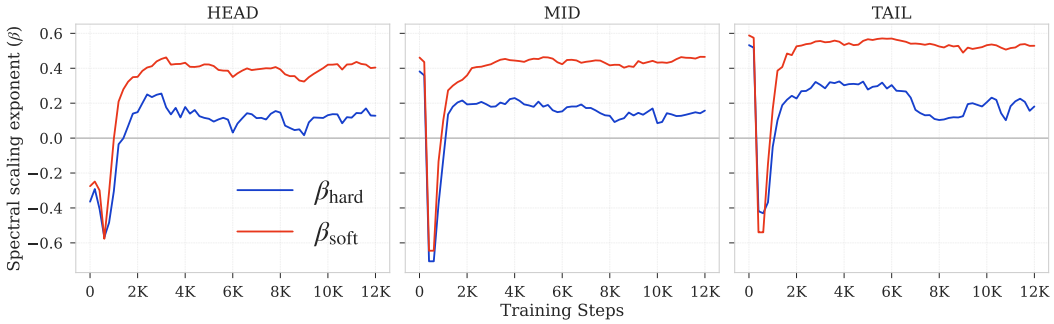



Figure 3: Extended AdamW training weakens hard-rank scaling in GPT-2 160M. Hard-rank scaling ($\beta_{\text{hard}}$) exhibits diminishing return for TAIL token scaling. Soft-rank scaling remains comparatively stable, leading to increasing hard–soft asymmetry at 12K (see Table 2).

scaling drops from $\beta_{\text{hard}} = 0.88$ to $0.40$, bringing it closer to the AdamW. By contrast, TAIL soft-rank scaling degrades more gradually, from $\beta_{\text{soft}} = 0.95$ to $0.72$, and remains above AdamW throughout the sweep. Thus, aggressive rank reduction does not eliminate diffuse spectral growth, it primarily limits the conversion of added FFN width into dominant-mode hard capacity.

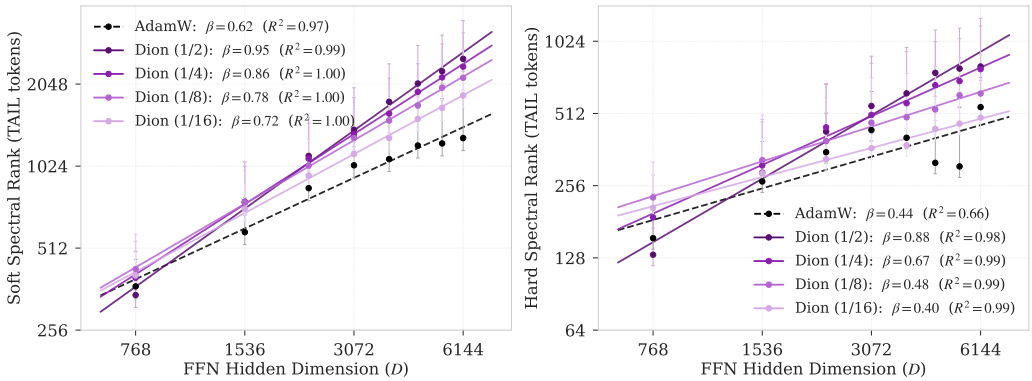



Figure 4: Soft spectral rank (left) and hard spectral rank (right) scaling is shown for TAIL tokens in GPT-2 160M, with AdamW as a reference. As the Dion rank fraction $r$ decreases, hard-rank scaling drops from $\beta = 0.88$ at $r = 1/2$ to $\beta = 0.40$ at $r = 1/16$, falling into the AdamW regime. Soft-rank scaling degrades more gradually ($0.95 \rightarrow 0.72$) and remains above AdamW, indicating that low update rank primarily limits dominant-mode hard capacity rather than all spectral growth.

**Hard–soft asymmetry exposes the rank bottleneck.** The selective degradation of hard rank appears directly in the hard–soft asymmetry. In the TAIL regime, $\Delta_{1,2}$ increases from $+0.07$ at $r = 1/2$ to $+0.31$ at $r = 1/16$, indicating that added FFN widths are less effectively converted into dominant-mode capacity. This effect is frequency-dependent: TAIL shows the clearest asymmetry increase, MID changes more mildly, and HEAD is non-monotonic across rank fractions. This shows that update rank acts as a control knob between high-rank, Muon-like capacity scaling and low-rank, AdamW-like scaling. With sufficient update rank, orthonormalized updates support growth in both soft and hard spectral capacity. Under aggressive rank reduction, soft-rank growth remains relatively robust, but rare-token hard-rank scaling falls toward the AdamW regime. Thus, *optimizer-update rank is part of the representation-scaling design space, not merely an efficiency parameter.* Full scaling exponents across frequency regimes are reported in Appendix Table 19.

Table 3: Low update rank increases hard–soft scaling asymmetry (Figure 4). Aggregate $\Delta_{1,2}$ rises from $+0.13$ at $r{=}1/2$ to $+0.24$ at $r{=}1/16$; TAIL shows the largest rise ($+0.07 \rightarrow +0.31$).

| | | Dion rank $r$ | | | |
|---|---|---|---|---|---|
| $\Delta_{1,2}$ | AdamW | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{16}$ |
| Aggregate | $+0.37$ | $+0.13$ | $+0.12$ | $+0.20$ | $+0.24$ |
| HEAD | $+0.18$ | $+0.37$ | $+0.17$ | $+0.19$ | $+0.35$ |
| MID | $+0.21$ | $+0.15$ | $+0.19$ | $+0.20$ | $+0.22$ |
| TAIL | $+0.18$ | $+0.07$ | $+0.18$ | $+0.29$ | $+0.31$ |

### 4.4 Optimizer-Dependent Spectral Scaling Persists at Larger Scale

**The optimizer-dependent structure persists at 350M.** Figure 5 shows TAIL-token spectral scaling at 350M, and the qualitative ordering matches the 160M results. Muon achieves near-linear hard-rank scaling ($\beta_{\text{hard}} = 1.13$, $R^2 = 0.94$), NorMuon remains strong ($\beta_{\text{hard}} = 0.88$, $R^2 = 0.98$), and AdamW remains clearly sublinear ($\beta_{\text{hard}} = 0.39$, $R^2 = 0.82$). Dion $(1/16)$ is also sublinear ($\beta_{\text{hard}} = 0.48$), far below Muon and NorMuon, showing that the low-rank update bottleneck persists at larger scale. Thus, the key optimizer-dependent scaling trends observed at 160M, strong $\beta_{\text{hard}}$ for Muon/NorMuon and weaker for AdamW/low-rank Dion, also appear at 350M.

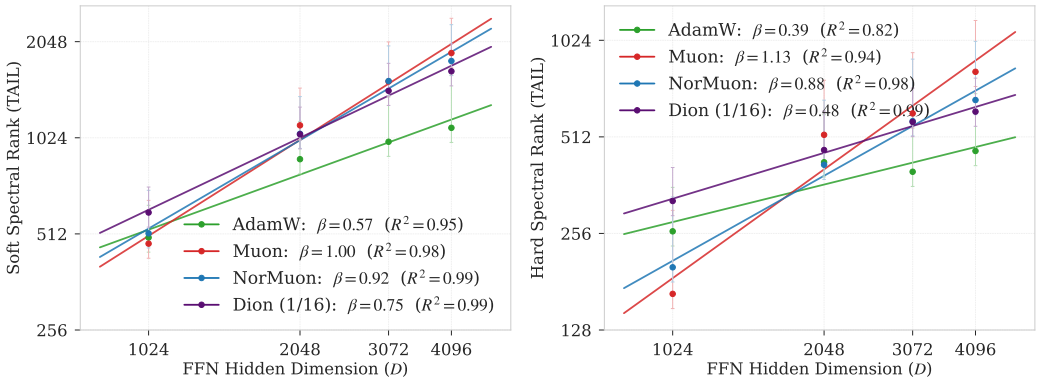



Figure 5: Optimizer-dependent TAIL spectral scaling persists at 350M scale. Soft spectral rank (left) and hard spectral rank (right) are shown across a four-point FFN-width sweep for TAIL tokens in GPT-2 350M. Muon and NorMuon maintain stronger hard-rank scaling than AdamW, while low-rank Dion$(1/16)$ remains in a lower hard-capacity ($\beta_{\text{hard}}$), while their soft-rank scaling incurred less degradation, indicating that low update rank primarily limits dominant-mode hard capacity.

**Hard–soft asymmetry also generalizes.** Table 4 compares $\beta_{\text{hard}}$ and $\Delta_{1,2}$ between 160M and 350M models. AdamW maintains positive TAIL asymmetry across scale ($+0.18/+0.19$), indicating that added FFN width converts primarily into the diffuse capacity scaling rather than dominant-mode capacity scaling. Dion $(1/16)$ also retains elevated positive asymmetry ($+0.31/+0.27$), consistent with the rank-bottleneck behavior in Section 4.3. By contrast, Muon reaches slightly negative asymmetry at 350M in both aggregate and TAIL spectra ($\Delta_{1,2} = -0.16$ and $-0.13$), while NorMuon remains close to zero asymmetry. Overall, the 350M experiment supports the conclusion that optimizer geometry shapes representation scaling beyond a single model size.

Table 4: Scale consistency from 160M to 350M: each cell reports the 160M/350M value. Muon and NorMuon maintain stronger hard-rank scaling than AdamW and low-rank Dion across scale, while AdamW and Dion$(1/16)$ retain positive asymmetry ($\Delta_{1,2}$).

| | Aggregate | | TAIL | |
|---|---|---|---|---|
| Optimizer | $\beta_{\text{hard}}$ | $\Delta_{1,2}$ | $\beta_{\text{hard}}$ | $\Delta_{1,2}$ |
| AdamW | 0.29 / 0.46 | $+0.37$ / $+0.17$ | 0.44 / 0.39 | $+0.18$ / $+0.19$ |
| Muon | 0.82 / 1.21 | $+0.14$ / $-0.16$ | 1.02 / 1.13 | $+0.01$ / $-0.13$ |
| NorMuon | 0.80 / 0.95 | $+0.21$ / $+0.02$ | 1.04 / 0.88 | $+0.00$ / $+0.04$ |
| Dion (1/16) | 0.50 / 0.53 | $+0.24$ / $+0.30$ | 0.40 / 0.48 | $+0.31$ / $+0.27$ |

## 5 Discussion and Conclusion

Spectral scaling laws are not properties of architecture alone, optimizer geometry changes how added FFN width is expressed in representation space. The main distinction between optimizers is not whether spectral rank grows, but which kind of spectral capacity grows. Under AdamW, added FFN width contributes primarily to diffuse spectral capacity rather than dominant-mode capacity. Muon-style optimizers make this conversion substantially more efficient, with the largest gains concentrated in MID- and TAIL-frequency token regimes.

**Limitations.** Our study establishes spectral scaling laws within a controlled empirical regime, using GPT-style decoder-only models with 160M and 350M parameters. Although the 350M runs replicate the main 160M patterns, testing at 1B+ scale would further strengthen the evidence for frontier-scale generality.

### AI USE STATEMENT

In this work, we used generative AI tools for editing and polishing the written text. We have not used generative AI tools for research ideation, experimental design, or the analysis and interpretation of results, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for refining the presentation of plots, figures, and tables. We have reviewed all AI-assisted work. Specifically, LLMs were used only to improve grammar, clarity, and style of text written by the authors, and all suggested edits were manually reviewed to ensure they did not alter the technical content or claims. For figures and tables, LLMs assisted only with formatting and visual presentation; all underlying data and results were produced by the authors' own experiments and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We provide the basic training details in the very-beginning of Section 4. The other implementation details are provided in Appendix A, including hyperparameter configurations.

## REFERENCES

Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

Noah Amsel, Gilad Yehudai, and Joan Bruna. Quality over quantity in attention layers: When adding more heads hurts. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Ta-Chung Chi, Ting-Han Fan, Li-Wei Chen, Alexander Rudnicky, and Peter Ramadge. Latent positional information is in the self-attention variance of transformer language models without positional embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning (ICML)*, 2023.

Xiaohan Ding, Honghao Chen, Xiangyu Zhang, Kaiqi Huang, Jungong Han, and Guiguang Ding. Re-parameterizing your optimizers rather than architectures. In *International Conference on Learning Representations (ICLR)*, 2023.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. *BioRxiv*, 2017.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in neural information processing systems (NeurIPS)*, 2022.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning (ICML)*, 2023.

Yoav Gelberg, Koshi Eguchi, Takuya Akiba, and Edoardo Cetin. Extending the context of pretrained LLMs by dropping their positional embedding. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning (ICML)*, 2018.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Yu Hu and Haim Sompolinsky. The spectrum of covariance matrices of randomly connected recurrent neuronal networks with linear dynamics. *PLoS computational biology*, 2022.

Nandan Kumar Jha and Brandon Reagen. Spectral scaling laws in language models: How effectively do feed-forward networks use their latent space? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL `https://github.com/KellerJordan/modded-nanogpt`.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cecista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. *URL https://kellerjordan. github. io/posts/muon*, 2024b.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning (ICML)*, 2023.

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.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Gyu Yeol Kim and Min hwan Oh. Convergence of muon with newton-schulz. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Diederik P Kingma. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under zipf's law. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Itay Lavie, Guy Gur-Ari, and Zohar Ringel. Towards understanding inductive bias in transformers: A view from infinity. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. In *International Conference on Machine Learning (ICML)*, 2026.

Kai Lion, Liang Zhang, Bingcong Li, and Niao He. PoLAR: Polar-decomposed low-rank adapter representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Nikolaos Nakis, Panagiotis Promponas, Konstantinos Tsirkas, Katerina Mamali, Eftychia Makri, Leandros Tassiulas, and Nicholas A Christakis. Rank is not capacity: Spectral occupancy for latent graph models. *arXiv preprint arXiv:2605.11142*, 2026.

Khang Ngo and Siamak Ravanbakhsh. Scaling laws and symmetry, evidence from neural force fields. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Razvan Pascanu, Clare Lyle, Ionut-Vlad Modoranu, Naima Elosegui Borras, Dan Alistarh, Petar Velickovic, Sarath Chandar, Soham De, and James Martens. Optimizers qualitatively alter solutions and we should leverage this. *arXiv preprint arXiv:2507.12224*, 2025.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. In *International Conference on Machine Learning (ICML)*, 2025.

Jose C Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, 1961.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European signal processing conference*, 2007.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning (ICML)*, 2018.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning (ICML)*, 2025.

David So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Searching for efficient transformers for language modeling. In *Advances in neural information processing systems (NeurIPS)*, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. In *Neurocomputing*, 2024.

Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. In *IEEE Transactions on Information Theory*, 2014.

Alexandra Volkova, Mher Safaryan, Christoph H Lampert, and Dan Alistarh. Towards robust scaling laws for optimizers. *arXiv preprint arXiv:2602.07712*, 2026.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing shampoo using adam for language modeling. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Kaiyue Wen, David Leo Wright Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Yi Xie, Stefan Mihalas, and Łukasz Kuśmierz. Slow transition to low-dimensional chaos in heavy-tailed recurrent neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning (ICML)*, 2020.

Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

# APPENDIX

**A Experimental Setup** **15**
A.1 Model architecture . . . 15
A.2 Training protocol . . . 15
A.3 Optimizer hyperparameters . . . 15
A.4 Token-frequency stratification . . . 16
A.5 Spectral measurements and reporting . . . 16
A.6 Compute . . . 16

**B Validation Perplexity Across Spectral-Scaling Runs** **16**
B.1 GPT-2 160M FFN-Width Sweep . . . 16
B.2 Architectural Intervention: Increasing Attention Ranks in GPT-2 160M . . . 17
B.3 Spectral Scaling at Scale: GPT-2 350M FFN-Width Sweep . . . 17
B.4 Architectural Intervention: Removing RoPE in GPT-2 350M . . . 17
B.5 Summary . . . 18

**C Rényi Effective Rank Analysis: Where Optimizer-Induced Capacity Forms** **18**

**D Layer-Wise Robustness of Optimizer-Induced Spectral Scaling** **20**

**E Hard-Rank Dynamics Under Extended AdamW Training** **22**

**F The AdamW–Muon Spectral-Scaling Gap Persists Across Learning Rates** **23**

**G Scaling Exponents for Dion Rank Sweep** **24**

**H Attention-Rank Effects and Optimizer–Architecture Interactions** **24**
H.1 Optimizer-Induced Scaling Effects Dominate the Attention-Rank Interventions . . . 25

**I Removing RoPE Reshapes Optimizer-Dependent Spectral Scaling** **27**

**J Optimizer Geometry Expands the Trainable Normalization Space** **28**

**K Position-Dependent FFN Spectral Transformation Under RoPE and NoPE** **29**
K.1 Positional Amplification Tracks TAIL Hard-Rank Scaling . . . 29
K.2 NoPE Changes the Amount of TAIL Positional Processing . . . 29
K.3 NoPE Changes the Depth Localization of Positional Processing . . . 30
K.4 NoPE Reverses the Frequency Bias of Positional Processing . . . 30

# A Experimental Setup

## A.1 Model architecture

All models follow the modded-nanoGPT (Jordan et al., 2024a) configuration, summarized in Table 5. All linear layers (attention projections, FFN, and the LM head) are bias-free. Attention uses rotary position embeddings (Su et al., 2024) with base $10{,}000$, applied per-head, and a non-parametric RMSNorm is additionally applied to the query and key tensors after RoPE (QK-norm). The FFN block uses two linear projections with squared-ReLU (So et al., 2021) between them. Pre-Norm and Post-Norm placements are controlled by a single `postln_frac` parameter $k \in [0, 1]$: the first $\lfloor k \cdot L \rfloor$ layers use Post-RMSNorm and the remainder use Pre-RMSNorm, with $k = 0$ recovering pure Pre-RMSNorm; this parameterization is used in the partial PostLN configurations. Weights are initialized with the spectral-condition scheme of Yang et al. (2023), $\sigma = (1/\sqrt{d_{\text{in}}}) \min(1, \sqrt{d_{\text{out}}/d_{\text{in}}})$.

Table 5: Architectural configurations. All models use the GPT-2 byte-pair tokenizer with padded vocabulary size 50,304. The 160M and 350M labels denote the base $4\times$ FFN configuration; FFN-width sweeps vary only the FFN intermediate dimension.

| Model | Layers | Heads | $d_{\text{model}}$ | Base FFN width ($D$) |
|---|---|---|---|---|
| 160M | 12 | 12 | 768 | 3072 |
| 350M | 24 | 32 | 1024 | 4096 |

## A.2 Training protocol

Table 6 summarizes the training details. All models are trained with a constant learning rate followed by a 20% linear cooldown to zero, no warmup, BF16 mixed precision, and no gradient clipping. Each (optimizer, FFN width) combination is trained as an independent run. For the extended-training 4.2, AdamW models are trained for 12,000 iterations with all other settings held fixed. Validation perplexity is computed on a held-out FineWeb-Edu subset of $10.5$M tokens.

Table 6: Training protocol. Tokens per step $= 1024 \times 512 = 524{,}288$. The 160M and 350M labels denote the base $4\times$ FFN configuration; FFN width is varied as $D = m\, d_{\text{model}}$ with $m$ as listed below, so total parameter count and per-iteration cost scale with $D$ while all other settings are held fixed.

| | 160M | 350M |
|---|---|---|
| Dataset | FineWeb-Edu (Penedo et al., 2024) | FineWeb-Edu (Penedo et al., 2024) |
| Sequence length | 512 | 512 |
| Global batch size | 1024 | 1024 |
| Iterations | 6,000 | 8,000 |
| Total tokens | 3.15B | 4.19B |

## A.3 Optimizer hyperparameters

All matrix-aware optimizers (Muon (Jordan et al., 2024b), NorMuon (Li et al., 2026), Dion (Ahn et al., 2025)) follow the parameter-update recipe described in (Ahn et al., 2025): Lion (Chen et al., 2023) is used as the scalar optimizer for non-matrix parameters (embeddings, normalization scales, the LM head), and the per-parameter learning-rate scaling follows the spectral-condition prescription of (Yang et al., 2023). Optimizer hyperparameters are listed in Table 7. Learning-rate ablations (Appendix F) additionally sweep AdamW learning rate over $\{1, 3, 6\} \times 10^{-3}$ and Muon learning rate over $\{1, 2, 4\} \times 10^{-2}$, holding all other hyperparameters fixed.

Table 7: Optimizer hyperparameters. Weight decay is $0.01$ throughout; momentum $\mu = 0.95$ for Muon, NorMuon, and Dion. AdamW (Loshchilov & Hutter, 2019) uses $(\beta_1, \beta_2) = (0.9, 0.95)$.

| | AdamW | Muon | NorMuon | Dion |
|---|---|---|---|---|
| Learning rate | $3\times10^{-3}$ | $2\times10^{-2}$ | $2\times10^{-2}$ | $2\times10^{-2}$ |
| LR scaling | — | spectral-norm | spectral-norm | spectral-norm |
| Scalar optimizer | — | Lion (Chen et al., 2023) | Lion (Chen et al., 2023) | Lion (Chen et al., 2023) |

Table 8: Frequency-regime statistics for FineWeb-Edu (GPT-2 BPE, vocab 50,304; corpus size $\approx 10.26$B tokens). Tertiles are set by cumulative occurrence mass.

| regime | Frequency threshold | Token types | % of vocabulary | % of occurrences |
|---|---|---|---|---|
| HEAD | $f(v) \geq 32{,}815{,}898$ | 30 | 0.06% | 32.79% |
| MID | $900{,}178 \leq f(v) < 32{,}815{,}898$ | 1,215 | 2.42% | 34.21% |
| TAIL | $f(v) < 900{,}178$ | 49,059 | 97.53% | 33.00% |

### A.4 Token-frequency stratification

Token frequencies $f(v)$ are computed once from the FineWeb-Edu tokenized training shards (10.26B tokens total) by histogram aggregation over GPT-2 byte-pair token IDs and held fixed across all experiments, ensuring that HEAD/MID/TAIL regime assignments are identical across optimizers, FFN widths, and model scales. Tertile thresholds $\tau_{\text{head}}$ and $\tau_{\text{mid}}$ are set by cumulative occurrence mass: sorting token types by decreasing $f(v)$ and taking cumulative sums, $\tau_{\text{head}}$ is the smallest frequency such that cumulative mass remains below $M/3$, and $\tau_{\text{mid}}$ is defined analogously at $2M/3$ (Eq. 7). The resulting regimes each cover approximately one third of total occurrences but span dramatically different numbers of token *types*, reflecting the heavy-tailed Zipfian structure of natural language (Kandpal et al., 2023; Kunstner & Bach, 2025): 30 token types (0.06% of vocabulary) carry the HEAD third of occurrences, 1,215 types (2.42%) carry the MID third, and 49,059 types (97.53%) carry the TAIL third. regime statistics are summarized in Table 8.

### A.5 Spectral measurements and reporting

Pre- and post-activation FFN representations (Section 3.1) are collected on the held-out validation batch ($\approx$10.5M tokens), aggregated across batch and sequence positions before computing the empirical covariance and its normalized eigenspectrum (Eq. 2). Spectral-rank quantities are computed per layer and then averaged across layers; frequency-stratified results are computed separately within each HEAD/MID/TAIL regime before layer averaging. Eigen-metric statistics are logged every 200 training steps at the 160M scale and every 400 steps at the 350M scale. For each scaling fit, we report the fitted exponent $\beta$ and the corresponding coefficient of determination $R^2$. All scaling-law fits use single seeds per (optimizer, FFN width) configuration; error bars in figures correspond to the inter-layer standard deviation of per-layer measurements.

### A.6 Compute

We use 4$\times$NVIDIA RTX 3090 GPUs (24 GB each) for the 160M-scale experiments and 8$\times$NVIDIA RTX 3090 GPUs for the 350M-scale experiments. The main FFN-width sweep comprises 40 training runs at 160M (five optimizers $\times$ eight widths) and 20 runs at 350M (five optimizers $\times$ four widths), complemented by the learning-rate, normalization-placement, positional-encoding, and extended-training ablations.

## B Validation Perplexity Across Spectral-Scaling Runs

In this section, we report the validation perplexities (PPL) for the training runs used in the spectral-scaling analyses. All values are computed on the held-out validation split of FineWeb-Edu (Penedo et al., 2024) with context length 512, following the evaluation protocol in Appendix A. These values provide loss-level context for the representation-geometry results. Note that the main comparisons are not between failed and successful training runs, rather between runs that can achieve comparable validation perplexity while exhibiting different spectral-scaling behavior.

### B.1 GPT-2 160M FFN-Width Sweep

Table 9 reports validation perplexity across the full eight-point FFN-width sweep for the 160M base configuration. The AdamW 12K row is included as the extended-training experiments described in Section 4.2. AdamW improves substantially with longer training and reaches perplexity comparable to Dion ($r = 1/16$), but remains well behind Muon and NorMuon optimizers.

Table 9: Validation perplexity for the GPT-2 160M FFN-width sweep. Perplexity is reported across FFN widths on the held-out validation set. AdamW 12K is included as an extended-training control.

| | FFN width ($d = 768$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Optimizer | $d$ | $2d$ | $3d$ | $4d$ | $5d$ | $6d$ | $7d$ | $8d$ |
| AdamW (6K) | 38.15 | 36.79 | 34.12 | 34.34 | 33.85 | 31.68 | 32.17 | 32.43 |
| AdamW (12K) | 34.25 | 32.15 | 30.40 | 30.47 | 29.43 | 29.04 | 28.45 | 28.29 |
| Muon | 31.79 | 29.83 | 28.65 | 27.90 | 27.27 | 26.85 | 26.43 | 26.15 |
| NorMuon | 31.77 | 29.82 | 28.63 | 27.82 | 27.23 | 26.74 | 26.29 | 25.94 |
| Dion ($r$=1/2) | 32.06 | 30.19 | 28.96 | 28.28 | 27.67 | 27.16 | 26.83 | 26.51 |
| Dion ($r$=1/4) | 32.61 | 30.69 | 29.58 | 28.79 | 28.23 | 27.68 | 27.41 | 27.13 |
| Dion ($r$=1/8) | 33.21 | 31.34 | 30.19 | 29.51 | 28.94 | 28.53 | 28.22 | 27.89 |
| Dion ($r$=1/16) | 34.18 | 32.41 | 31.33 | 30.66 | 30.08 | 29.71 | 29.40 | 29.23 |

### B.2 Architectural Intervention: Increasing Attention Ranks in GPT-2 160M

Table 10 reports the validation perplexities for the 6-head architecture used in Section H.1. This intervention changes the attention-head configuration while preserving the same FFN-width sweep, at fixed parameter count. Muon and NorMuon remain the top-performing optimizers, while the Dion variants and AdamW preserve the same broad ordering as in the original 12-head setting.

Table 10: Validation perplexity for the GPT-2 160M 6-head sweep. Perplexity is reported across FFN widths for the reduced-head-count architecture.

| | FFN width ($d = 768$) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Optimizer | $d$ | $2d$ | $3d$ | $4d$ | $5d$ | $6d$ | $7d$ | $8d$ |
| AdamW | 37.95 | 35.40 | 34.36 | 33.34 | 32.95 | 32.65 | 32.49 | 32.19 |
| Muon | 31.56 | 29.69 | 28.59 | 27.84 | 27.22 | 26.71 | 26.36 | 26.09 |
| NorMuon | 31.59 | 29.62 | 28.55 | 27.67 | 27.08 | 26.69 | 26.21 | 25.89 |
| Dion ($r$=1/2) | 31.93 | 29.96 | 28.99 | 28.26 | 27.59 | 27.20 | 26.77 | 26.47 |
| Dion ($r$=1/16) | 33.91 | 32.09 | 31.16 | 30.50 | 29.94 | 29.50 | 29.27 | 28.94 |

### B.3 Spectral Scaling at Scale: GPT-2 350M FFN-Width Sweep

Table 11 reports validation perplexity for the 350M scale-replication runs in Section 4.4. The sweep is coarser than the 160M sweep, but the loss-level ordering remains consistent: Muon and NorMuon achieve best perplexity, followed by Dion ($r = 1/16$), with AdamW remaining higher in perplexity.

Table 11: Validation perplexity for the GPT-2 350M FFN-width sweep. Perplexity is reported across the four FFN widths used in the 350M scale-replication experiment.

| | FFN width ($d = 1024$) | | | |
|---|---|---|---|---|
| Optimizer | $d$ | $2d$ | $3d$ | $4d$ |
| AdamW | 31.27 | 29.00 | 28.38 | 28.35 |
| Muon | 26.07 | 24.41 | 23.51 | 22.97 |
| NorMuon | 26.06 | 24.37 | 23.44 | 22.80 |
| Dion ($r$=1/16) | 27.95 | 26.33 | 25.45 | 24.90 |

### B.4 Architectural Intervention: Removing RoPE in GPT-2 350M

Table 12 reports validation perplexity for the NoPE runs used in Section I. Removing RoPE increases perplexity for all optimizers, but Muon and NorMuon remain at the top. These values provide the loss-level context for the RoPE vs NoPE architectural intervention.

Table 12: Validation perplexity for GPT-2 350M without RoPE (Section I).

| | FFN width ($d = 1024$) | | | |
|---|---|---|---|---|
| Optimizer | $d$ | $2d$ | $3d$ | $4d$ |
| AdamW | 33.24 | 30.90 | 29.27 | 29.66 |
| Muon | 27.05 | 25.33 | 24.37 | 23.75 |
| NorMuon | 27.08 | 25.26 | 24.29 | 23.60 |

### B.5 Summary

Across the validation-perplexity tables, Muon and NorMuon consistently form the strongest loss-level optimizer family, while AdamW remains substantially higher in perplexity. The extended AdamW control improves validation perplexity and matches Dion ($r = 1/16$) closely, however their spectral geometry remains distinct. Thus, the spectral-scaling differences reported in the paper are not reducible to a simple failed-optimization explanation. Architectural interventions such as increasing attention ranks or removing RoPE shift absolute perplexity, but do not remove the need to analyze optimizer-dependent representation geometry.

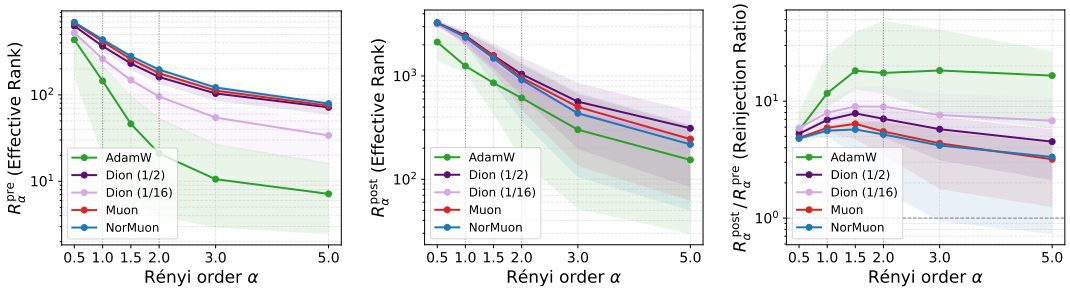
## C Rényi Effective Rank Analysis: Where Optimizer-Induced Capacity Forms

The methodology section 3.2 defines the Rényi effective-rank family $R_\alpha$ as a continuum of spectral-capacity probes. From information-theoretic viewpoint, the order $\alpha$ changes the resolution at which the normalized eigenspectrum is probed, lower orders give greater relative weight to weak eigendirections (see Table 13), while higher orders increasingly emphasize dominant modes (Rényi, 1961; Roy & Vetterli, 2007; Van Erven & Harremos, 2014; Principe, 2010). We uses two anchors of this family, $R_1$, the soft-rank measure of diffuse spectral capacity, and $R_2$, the hard-rank measure of dominant-mode capacity. Now, we perform full Rényi-order sweep to study where optimizer-induced spectral capacity emerges, and how it varies across rank measures.

Specifically, we sweep $\alpha \in \{0.5, 1, 1.5, 2, 3, 5\}$ and measure

$$R_\alpha^{\mathrm{pre}}, \qquad R_\alpha^{\mathrm{post}}, \qquad \rho_\alpha = \frac{R_\alpha^{\mathrm{post}}}{R_\alpha^{\mathrm{pre}}},$$

where $R_\alpha^{\mathrm{pre}}$ is computed before the FFN nonlinearity, $R_\alpha^{\mathrm{post}}$ is computed after the nonlinearity, and $\rho_\alpha$ measures nonlinear reinjection ratio. This separates the optimizer-induced geometry entering the nonlinearity from the representation geometry after activation. We use GPT-2 350M models since it is the largest model used in this work, however, we observe the same qualitative trends at 160M scale.




(a) Pre-activation spectrum (b) Post-activation spectrum (c) Nonlinear reinjection

Figure 6: Rényi-family view of optimizer-shaped spectral capacity in GPT-2 350M, with FFN width $4\times$. We report pre-activation rank $R_\alpha^{\mathrm{pre}}$, post-activation rank $R_\alpha^{\mathrm{post}}$, and nonlinear reinjection ratio $\rho_\alpha = R_\alpha^{\mathrm{post}}/R_\alpha^{\mathrm{pre}}$. Curves show the median across 24 layers, and shaded bands denote the interquartile range. Vertical guides mark the soft-rank anchor $\alpha = 1$ and the hard-rank anchor $\alpha = 2$.

**Pre-activation spectrum shows optimizer-induced geometry.** Figure 6 shows the Rényi-family profiles for a GPT-2 350M, and we observe that the strongest optimizer separation appears before the FFN nonlinearity. Across all tested Rényi orders, the pre-activation rank follows a stable hierarchy:

$$\text{NorMuon} > \text{Muon} > \text{Dion}(r = 1/2) > \text{Dion}(r = 1/16) > \text{AdamW}.$$

Table 13: Interpretation of Rényi order $\alpha$ for normalized FFN eigenspectra.

| Order | Sensitivity | Spectral interpretation |
|---|---|---|
| $0 < \alpha < 1$ | Weak-direction-sensitive | Gives more credit to small nonzero eigenvalues |
| $\alpha = 1$ | Shannon-balanced | Measures entropy-weighted spectral spread |
| $1 < \alpha < 2$ | Mildly concentration-sensitive | Interpolates Shannon spread and order-2 concentration |
| $\alpha = 2$ | Quadratic / participation-ratio regime | Captures concentration-sensitive effective rank |
| $\alpha > 2$ | Strongly concentration-sensitive | Increasingly dominated by leading eigendirections |

This hierarchy aligns with validation perplexity at width $4\times$ (Table 14).

At the soft-rank anchor $\alpha = 1$, pre-activation rank increases from $126.0$ for AdamW to $411.3$ for NorMuon. At the hard-rank anchor $\alpha = 2$, the same monotonic structure holds:

$$R_2^{\text{pre}} = 177.4,\ 164.8,\ 148.0,\ 84.5,\ 28.0$$

for NorMuon, Muon, Dion $(r = 1/2)$, Dion $(r = 1/16)$, and AdamW, respectively. Thus, Muon and NorMuon enter the FFN nonlinearity with substantially richer pre-activation spectra than AdamW, indicating that optimizer-induced spectral geometry is already visible before activation.

**The FFN nonlinearity re-organize the optimizer hierarchy.** The post-activation spectra do not simply preserve the pre-activation ordering. Although NorMuon has the largest pre-activation rank across the Rényi family, Dion $(r = 1/2)$ achieves the largest post-activation rank for $\alpha \geq 1$. At the hard-rank anchor $\alpha = 2$, the post-activation ranks are

$$R_2^{\text{post}} = 989.1,\ 864.4,\ 811.3,\ 798.5,\ 546.7$$

for Dion $(r = 1/2)$, Muon, NorMuon, Dion $(r = 1/16)$, and AdamW, respectively. Thus, post-activation effective rank is not determined by pre-activation rank alone. The FFN nonlinearity redistributes variance in an optimizer-dependent way, causing rank-order changes between the pre- and post-activation spectra. In other words, pre-activation ranks capture the optimizer-induced precursor geometry, while post-activation ranks capture the effective latent capacity available to subsequent layers in Transformer.

**Large nonlinear reinjection is compensatory.** The reinjection ratio $\rho_\alpha$ explains the post-activation reordering. AdamW exhibits the largest reinjection across the Rényi family, peaking at the hard-rank anchor with $\rho_2 = 51.9$. In contrast, Muon and NorMuon require much smaller reinjection: $\rho_2 = 4.8$ for Muon and $\rho_2 = 4.3$ for NorMuon. Thus, at $\alpha = 2$, AdamW's reinjection is roughly $10.8\times$ larger than Muon's and $12.1\times$ larger than NorMuon's. However, AdamW still has the lowest post-activation $R_2$ and the worst validation perplexity. Thus, large reinjection does not implies better spectral capacity; rather, it indicates that the FFN nonlinearity is compensating for a more collapsed optimizer-shaped precursor.

Table 14: Pre-activation rank and validation perplexity for GPT-2 350M at FFN width $4\times$. Lower PPL is better, while higher pre-activation rank indicates larger effective spectral capacity before the FFN nonlinearity.

| Optimizer | PPL ↓ | $R_1^{\text{pre}}$ | $R_2^{\text{pre}}$ | $\rho_1$ | $\rho_2$ |
|---|---|---|---|---|---|
| AdamW | 28.35 | 126.0 | 28.0 | 22.6 | 51.9 |
| Dion $(r = 1/16)$ | 24.92 | 238.0 | 84.5 | 8.9 | 12.4 |
| Dion $(r = 1/2)$ | 23.27 | 344.3 | 148.0 | 6.5 | 6.4 |
| Muon | 22.97 | 389.1 | 164.8 | 5.2 | 4.8 |
| NorMuon | 22.80 | 411.3 | 177.4 | 5.2 | 4.3 |

Dion further illustrates this effect as a controllable update-rank intervention. At $\alpha = 2$, decreasing the Dion rank fraction from $r = 1/2$ to $r = 1/16$ reduces $R_2^{\text{pre}}$ from $148.0$ to $84.5$, while increasing the reinjection ratio from $6.4$ to $12.4$, and their post-activation rank reduce from $989.1$ to $798.5$. Thus, aggressive low-rank updates reduce the richness of the pre-activation geometry, but the FFN nonlinearity partially recovers post-activation capacity.

This effect becomes especially visible in the high-$\alpha$ regime. At $\alpha = 5$, Dion $(r = 1/16)$ has much lower pre-activation rank than Muon and NorMuon ($31.8$ versus $68.3$ and $71.4$), however its post-activation rank slightly exceeds both ($221.7$ versus $217.3$ and $198.7$). Thus, *Rényi sweep reveals rank-order reversals that would be hidden by a single effective-rank metric*.

Overall, the Rényi-family diagnostics separate three roles inside the FFN block. Pre-activation spectra reveal the optimizer-shaped geometry entering the nonlinearity; post-activation spectra show the representation passed to the FFN output projection; and $\rho_\alpha$ measures how strongly the nonlinearity

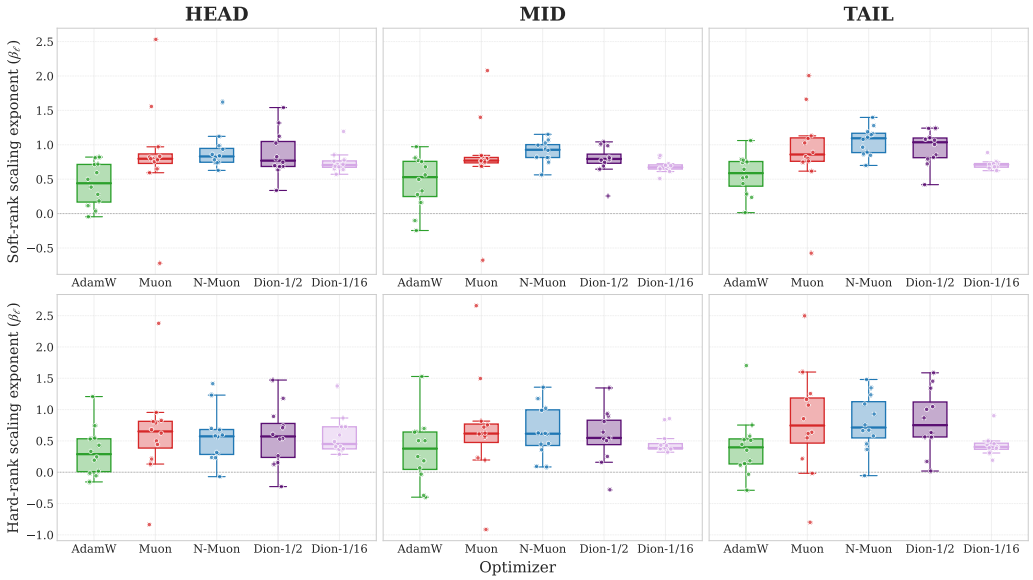



Figure 7: Distribution of layer-wise scaling exponents for GPT-2 160M. For each layer $\ell$, optimizer, token-frequency regimes, and spectral rank metric, we fit $R_\ell(D) \propto D^{\beta_\ell}$ across FFN widths. Box plots summarize the distribution of $\beta_\ell$ across layers, and the horizontal dotted line marks $\beta_\ell = 0$.

reshapes the spectrum. In the 350M comparison, the best-performing optimizers enter the nonlinearity with high pre-activation effective rank and require only moderate reinjection. This links optimizer geometry to the spectral scaling laws and suggests that Rényi-family profiles can serve as useful diagnostics for optimizer–architecture interaction.

## D Layer-Wise Robustness of Optimizer-Induced Spectral Scaling

The main results in Section 4.1 fit scaling exponents from layer-aggregated spectral ranks. Here, we investigate whether those aggregate trends reflect broad behavior across depth, or they are driven by a small number of layers. For each layer $\ell$, optimizer, frequency regimes, and spectral rank metric, we independently fit $R_\ell(D) \propto D^{\beta_\ell}$ across the same FFN widths used in the aggregate scaling-laws. This produces one scaling exponent $\beta_\ell$ per layer. We summarize the layer-wise distributions using the median, interquartile range (IQR), and the fraction of layers with positive scaling exponent. A high positive-layer fraction indicates that the aggregate scaling trend is broadly reflected across depth.

Figure 7 shows the distribution of layer-wise exponents, and Figure 8 shows their depth profiles. Tables 15 and 16 report the corresponding robust summary statistics. These diagnostics demonstrate the spread of layer-wise exponents and the extent to which aggregate trends are reflected across depth.

Table 15: Layer-wise soft-rank scaling exponents. For each optimizer and token-frequency regimes, we report the median, interquartile range (IQR), and fraction of layers with positive $\beta_\ell$.

| | HEAD | | | MID | | | TAIL | | |
|---|---|---|---|---|---|---|---|---|---|
| Optimizer | Med. | IQR | Frac.$> 0$ | Med. | IQR | Frac.$> 0$ | Med. | IQR | Frac.$> 0$ |
| AdamW | 0.441 | 0.547 | 0.917 | 0.530 | 0.511 | 0.833 | 0.588 | 0.358 | 1.000 |
| Muon | 0.797 | 0.138 | 0.917 | 0.770 | 0.079 | 0.917 | 0.859 | 0.340 | 0.917 |
| NorMuon | 0.831 | 0.201 | 1.000 | 0.927 | 0.188 | 1.000 | 1.095 | 0.281 | 1.000 |
| Dion (1/2) | 0.770 | 0.363 | 1.000 | 0.794 | 0.133 | 1.000 | 1.036 | 0.286 | 1.000 |
| Dion (1/16) | 0.705 | 0.093 | 1.000 | 0.673 | 0.061 | 1.000 | 0.711 | 0.053 | 1.000 |

**Added width broadly expands soft-rank capacity across depth.** A positive layer-wise exponent $\beta_\ell > 0$ implies that the effective capacity of layer $\ell$ increases with added FFN width. More importantly, it show that the near-linear aggregate $\beta_{\text{soft}}$ is not produced by a small subset of layers. Soft-rank exponents are positive for nearly all layers, and their medians remain large for Muon-style optimizers. NorMuon attains the largest median soft-rank exponents across all regimes, reaching

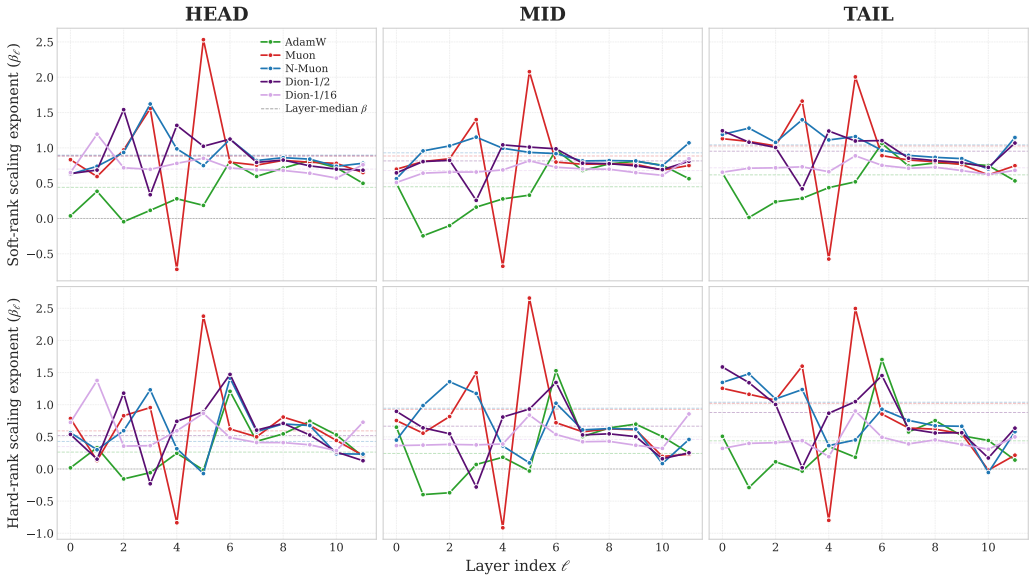



Figure 8: Depth profiles of layer-wise scaling exponents for GPT-2 160M. Each curve shows $\beta_\ell$ as a function of layer index for a fixed optimizer, frequency regimes, and rank metric. Dashed horizontal lines denote the corresponding aggregate exponent from the main analysis.

$0.831$ in HEAD, $0.927$ in MID, and $1.095$ in TAIL. AdamW also shows positive scaling in most layers, *but with weaker medians and wider layer-wise spread*. Dion $(1/16)$ is more constrained but highly stable, with small IQRs across MID and TAIL regimes (Table 15). Thus, aggregate $\beta_{\text{soft}}$ trends in the scaling laws are not only directionally positive across depth, their magnitude and optimizer ordering are also reflected in the layer-wise fits.

**Hard-rank scaling is more heterogeneous but remains optimizer-dependent.** Hard-rank exponents show stronger variation across depth than soft-rank exponents, consistent with dominant-mode capacity being a more selective spectral notion. The optimizer-dependent structure observed in the aggregate fits remains visible at the layer level. For AdamW, $75\%$ of layers have positive hard-rank scaling in HEAD and MID, and its median hard-rank exponents remain below $0.40$ in all regimes. Muon, NorMuon, and Dion $(1/2)$ produce substantially larger median hard-rank exponents, although with greater layer-wise spread than in the soft-rank case. The separation is clearest in TAIL, where their median exponents are $0.746$, $0.714$, and $0.751$, respectively (Table 16). This reinforce the aggregate scaling laws result that orthonormal-update optimizers convert added FFN width into dominant-mode capacity more effectively than AdamW.

Table 16: Layer-wise hard-rank scaling exponents. For each optimizer and frequency regime, we report the median, interquartile range (IQR), and fraction of layers with positive $\beta_\ell$.

| | HEAD | | | MID | | | TAIL | | |
|---|---|---|---|---|---|---|---|---|---|
| Optimizer | Med. | IQR | Frac.$>0$ | Med. | IQR | Frac.$>0$ | Med. | IQR | Frac.$>0$ |
| AdamW | 0.288 | 0.525 | 0.750 | 0.377 | 0.597 | 0.750 | 0.397 | 0.400 | 0.833 |
| Muon | 0.651 | 0.427 | 0.917 | 0.617 | 0.294 | 0.917 | 0.746 | 0.720 | 0.833 |
| NorMuon | 0.573 | 0.398 | 0.917 | 0.615 | 0.569 | 1.000 | 0.714 | 0.579 | 0.917 |
| Dion (1/2) | 0.571 | 0.543 | 0.917 | 0.548 | 0.390 | 0.917 | 0.751 | 0.559 | 1.000 |
| Dion (1/16) | 0.450 | 0.353 | 1.000 | 0.386 | 0.086 | 1.000 | 0.404 | 0.098 | 1.000 |

**Low-rank Dion is stable but capacity-limited.** The layer-wise analysis also clarifies the behavior of Dion $(1/16)$. This optimizer has positive hard-rank scaling in every layer across all regimes, with especially narrow IQRs in MID and TAIL (Table 16). However, its median hard-rank exponents are much lower than those of Muon, NorMuon, and Dion $(1/2)$. For instance, in TAIL, Dion $(1/16)$ has median $\beta_\ell = 0.404$, compared with $0.746$ for Muon, $0.714$ for NorMuon, and $0.751$ for Dion $(1/2)$. Thus, aggressive low-rank update structure yields stable but constrained hard-rank scaling.

**Depth profiles reveal structured heterogeneity.** The depth profiles in Figure 8 show that layer-wise exponents vary across depth, especially for hard rank. This variation helps explain why some aggregate hard-rank fits have lower $R^2$ in the main table. Even with this heterogeneity, the optimizer ordering remains visible across many layers: AdamW generally exhibits weaker hard-rank scaling, whereas Muon, NorMuon, and Dion $(1/2)$ frequently produce larger positive exponents. These layer-wise fits support the aggregate scaling laws trends. Thus, optimizer-induced scaling differences are not artifacts of layer aggregation; rather, *hard-rank capacity is more depth-sensitive and optimizer-sensitive than soft-rank spread*.

## E Hard-Rank Dynamics Under Extended AdamW Training

Section 4.2 shows that extended AdamW training breaks the power-law relationship between FFN width and hard rank, *even when validation perplexity continues to improve*. In this section, we investigate this through the post-activation hard-rank trajectories. We focus on hard rank because the beta-dynamics analysis in Figure 3 shows the diminishing return in $\beta_{\text{hard}}$(TAIL) for AdamW extended training. For a clean scaling law, larger FFN widths should preserve a systematic width-capacity ordering. Extended AdamW training disrupts this ordering. For example, the $8d$ model initially gains hard-rank capacity, but later drops below narrower models in some frequency regimes. As a result, larger FFN width no longer corresponds to larger hard-rank capacity, breaking the width–capacity relationship required for reliable power-law fits.

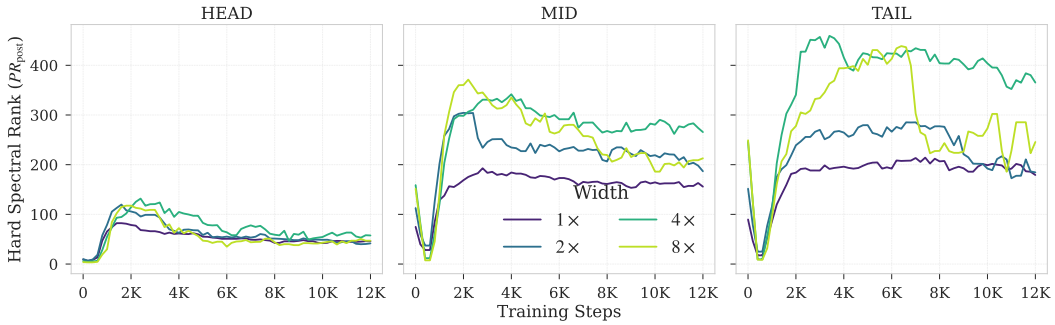



Figure 9: Hard-rank dynamics for AdamW extended training with GPT-2 160M. Post-activation hard ranks are shown across 12K training steps for representative FFN widths ($1\times$, $2\times$, $4\times$, $8\times$). In HEAD regime, all widths rise early and then decline toward similar values. In MID regimes, width ordering is partially violated at the end of training—$8\times$ trajectory falling below the $4\times$ trajectory. In TAIL regimes, the breakdown is **strongest:** the $8\times$ trajectory peaks early, drops sharply after roughly 7K steps, falls below the $2\times$ trajectory around 8K, and remains far below the $4\times$ trajectory at 12K.

**Differential hard-rank erosion explains the fit breakdown.** Figure 9 and Table 17 show that the hard-rank scaling failure, induced by a breakdown of width–capacity ordering. In the TAIL regimes, hard rank initially increases with width, consistent with a width-scaling relation. During extended training, however, wider FFN configurations can lose dominant-mode capacity more rapidly than narrower ones. The $8d$ model rises from 278 at 2K to 394 at 4K, but then drops to 228 by 8K, falling below the $2d$ model at 277. By 12K, the $1d$ and $2d$ models have nearly converged (179 vs. 185), whereas the $4d$ model remains substantially higher (365). Thus, the hard-rank fit does not merely become noisy, the relationship between width and dominant-mode capacity is itself disrupted.

Table 17: TAIL hard-rank dynamics under extended AdamW training for GPT-2 160M. Post-activation hard rank is reported at selected checkpoints for representative FFN widths. The $8d$ trajectory rises early but drops below the $2d$ model by 8K steps and remains far below the $4d$ model at 12K, disrupting the width–capacity relationship needed for a reliable power-law fit.

| | $1d$ | | | | $2d$ | | | | $4d$ | | | | $8d$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 2K | 4K | 8K | 12K | 2K | 4K | 8K | 12K | 2K | 4K | 8K | 12K | 2K | 4K | 8K | 12K |
| Hard-rank | 184 | 196 | 207 | 179 | 239 | 265 | 277 | 185 | 341 | 416 | 404 | 365 | 278 | 394 | 228 | 246 |

The same pattern appears, more mildly, outside the TAIL regime. In HEAD regime, the $8\times$ trajectory declines from roughly $117$ near its early peak to about $47$ by 12K, while the smaller-width trajectories converge toward similar values. In MID regime, the $8\times$ trajectory falls below the $4\times$ trajectory by the end of training. These dynamics support the interpretation in Section 4.2, extended AdamW training does not shift all widths uniformly, but preferentially erodes hard-rank capacity in wider FFN configurations. This breaks the width–capacity ordering needed for hard-rank scaling, **despite the continuous improvement in validation perplexity.**

# F The AdamW–Muon Spectral-Scaling Gap Persists Across Learning Rates

A natural concern is that the spectral-scaling gap between AdamW and Muon could be the artifact of learning-rate tuning rather than optimizer geometry. We investigate this by sweeping learning rates for both optimizers and analyzing their spectral scaling behavior. AdamW is evaluated at $\{10^{-3}, 3\times10^{-3}, 6\times10^{-3}\}$, with $3\times10^{-3}$ used as the default in the main experiments. Muon is evaluated at $\{10^{-2}, 2\times10^{-2}, 4\times10^{-2}\}$, with $2\times10^{-2}$ used as the default. This sweep is intended as a scalar learning-rate control, rather than an exhaustive hyperparameter search.

**Learning-rate tuning does not recover Muon-like hard-rank scaling.** Table 18 shows that the AdamW–Muon gap is not closed by the tested learning-rate sweep. The clearest comparison is TAIL hard rank, AdamW's largest reliable exponent is $\beta = 0.44$ at $3\times10^{-3}$, whereas Muon's lowest exponent across the sweep is $\beta = 0.80$ at $10^{-2}$. Thus, even Muon's weakest learning-rate setting exceeds AdamW's strongest reliable setting by $0.36$. Moreover, this gap is larger than the within-AdamW reliable range ($0.44-0.32 = 0.12$) and the within-Muon range ($1.02-0.80 = 0.22$). That is, the TAIL hard-rank exponent ranges do not overlap across the tested learning rates. AdamW learning-rate sweep redistribute weak scaling across frequency regimes, at $10^{-3}$, $\beta_{\mathrm{hard}}$(HEAD) improves to $0.43$ but MID collapses to $\beta$=0.00. At $3\times10^{-3}$, $\beta_{\mathrm{hard}}$(TAIL) improves to $0.44$ while HEAD and MID remain weak, and at $6\times10^{-3}$, the TAIL hard-rank fit becomes unreliable ($R^2 = 0.04$).

Table 18: Spectral scaling exponents across learning rates for GPT-2 160M. Hard-rank and soft-rank scaling exponents $\beta$ are reported across token-frequency regimes, with $R^2$ in parentheses.

| | | Hard Rank | | | Soft Rank | | |
|---|---|---|---|---|---|---|---|
| Optimizer | LR | HEAD | MID | TAIL | HEAD | MID | TAIL |
| AdamW | $10^{-3}$ | 0.43 (0.57) | 0.00 (0.00) | 0.32 (0.64) | 0.63 (0.89) | 0.52 (0.76) | 0.70 (0.99) |
| AdamW | $3\times10^{-3}$ | 0.26 (0.59) | 0.24 (0.36) | 0.44 (0.66) | 0.44 (0.82) | 0.45 (0.82) | 0.62 (0.97) |
| AdamW | $6\times10^{-3}$ | 0.25 (0.19) | 0.28 (0.29) | 0.12 (0.04) | 0.43 (0.47) | 0.36 (0.43) | 0.34 (0.33) |
| Muon | $10^{-2}$ | 0.75 (0.90) | 0.77 (0.87) | 0.80 (0.97) | 0.99 (0.99) | 0.85 (1.00) | 0.91 (0.99) |
| Muon | $2\times10^{-2}$ | 0.59 (0.54) | 0.93 (0.82) | 1.02 (0.81) | 0.88 (0.90) | 0.89 (0.96) | 1.03 (0.94) |
| Muon | $4\times10^{-2}$ | 0.69 (0.43) | 0.71 (0.82) | 0.96 (0.93) | 0.82 (0.94) | 0.83 (0.97) | 0.97 (0.99) |

**Soft-rank scaling shows the same ordering with a smaller gap.** The same qualitative separation appears for $\beta_{\mathrm{soft}}$, although the gap is smaller than $\beta_{\mathrm{hard}}$. AdamW's largest TAIL soft-rank exponent is $\beta = 0.70$ at $10^{-3}$, whereas Muon's lowest TAIL soft-rank exponent is $\beta = 0.91$ at $10^{-2}$. The ordering remains unchanged across the tested learning rates. Soft-rank fits are also more stable for Muon, with TAIL $R^2 \geq 0.94$ across the sweep.

Thus, the above learning-rate sweep ablations rule out a simple scalar learning-rate explanation for the AdamW–Muon gap. Combined with the extended-training control in Appendix E, this shows that AdamW's weaker hard-rank scaling is not explained by insufficient training time or learning-rate mistuning. This supports the interpretation that optimizer geometry changes how added FFN width is converted into usable capacity.

## G Scaling Exponents for Dion Rank Sweep

Section 4.3 shows that reducing Dion's update rank lowers hard-capacity scaling, especially for TAIL tokens. Table 19 reports the full fitted exponents underlying that comparison. The clearest trend appears in MID and TAIL, where decreasing the rank fraction from $r = 1/2$ to $r = 1/16$ reduces hard-rank scaling more sharply than soft-rank scaling. For instance, $\beta_{\mathrm{hard}}$(TAIL) drops from $0.88$ to $0.40$, while $\beta_{\mathrm{soft}}$ decreases gradually from $0.95$ to $0.72$. HEAD trends are less monotonic and have larger uncertainty, so the rank-bottleneck interpretation is most reliable in the MID and TAIL token regimes.

Table 19: Scaling exponents across Dion rank fractions for GPT-2 160M. For each optimizer, token-frequency regimes, and rank metric, we report the fitted exponent $\beta$ with fit uncertainty and $R^2$ in parentheses. Decreasing Dion's rank fraction primarily reduces hard-rank scaling, especially in MID and TAIL, while soft-rank scaling degrades more gradually.

| | HEAD | | MID | | TAIL | |
|---|---|---|---|---|---|---|
| Optimizers | Hard Rank | Soft Rank | Hard Rank | Soft Rank | Hard Rank | Soft Rank |
| AdamW | $0.26 \pm 0.22$ ($R^2 = 0.59$) | $0.44 \pm 0.21$ ($R^2 = 0.82$) | $0.24 \pm 0.32$ ($R^2 = 0.36$) | $0.45 \pm 0.21$ ($R^2 = 0.82$) | $0.44 \pm 0.31$ ($R^2 = 0.66$) | $0.62 \pm 0.10$ ($R^2 = 0.97$) |
| Dion (1/2) | $0.52 \pm 0.60$ ($R^2 = 0.43$) | $0.89 \pm 0.30$ ($R^2 = 0.90$) | $0.67 \pm 0.10$ ($R^2 = 0.98$) | $0.82 \pm 0.06$ ($R^2 = 0.99$) | $0.88 \pm 0.13$ ($R^2 = 0.98$) | $0.95 \pm 0.07$ ($R^2 = 1.00$) |
| Dion (1/4) | $0.72 \pm 0.34$ ($R^2 = 0.82$) | $0.90 \pm 0.16$ ($R^2 = 0.97$) | $0.58 \pm 0.06$ ($R^2 = 0.99$) | $0.77 \pm 0.03$ ($R^2 = 1.00$) | $0.67 \pm 0.06$ ($R^2 = 0.99$) | $0.86 \pm 0.03$ ($R^2 = 1.00$) |
| Dion (1/8) | $0.56 \pm 0.27$ ($R^2 = 0.81$) | $0.75 \pm 0.16$ ($R^2 = 0.96$) | $0.53 \pm 0.08$ ($R^2 = 0.98$) | $0.74 \pm 0.03$ ($R^2 = 1.00$) | $0.48 \pm 0.04$ ($R^2 = 1.00$) | $0.78 \pm 0.02$ ($R^2 = 1.00$) |
| Dion (1/16) | $0.35 \pm 0.20$ ($R^2 = 0.75$) | $0.70 \pm 0.09$ ($R^2 = 0.98$) | $0.46 \pm 0.07$ ($R^2 = 0.98$) | $0.68 \pm 0.05$ ($R^2 = 1.00$) | $0.40 \pm 0.04$ ($R^2 = 0.99$) | $0.72 \pm 0.03$ ($R^2 = 1.00$) |

## H Attention-Rank Effects and Optimizer–Architecture Interactions

Section H.1 compares optimizer-induced spectral-scaling gains with the effect of architectural intervention, attention-rank, reducing the number of attention heads from 12 to 6 at fixed parameter count. Here, we report the exact effect sizes, and interaction diagnostics for both, the optimizer-induced and architecture -induced spectral scaling shift. In particular, the goal is to separate three quantities: (i) the optimizer-induced gain over AdamW, (ii) the absolute attention-rank architectural shift, and (iii) the change in optimizer gain induced by the architectural intervention.

Table 20 reports these quantities. The optimizer-induced gain exceeds the absolute attention-rank shift in 28 of 30 regime–optimizer comparisons (Figure 10). The only exceptions occur for HEAD hard-rank scaling under AdamW and Muon, where increasing per-head attention rank produces a larger shift than the best optimizer-induced gain over AdamW. Thus, across nearly all regimes, changing optimizer produces a larger spectral-scaling shift than this controlled attention-rank intervention.

Table 20: Optimizer-induced gains over AdamW and absolute attention-rank architectural shifts for GPT-2 160M. $\Delta\beta^{\star}_{\mathrm{opt}}$ measures the best optimizer-induced gain over AdamW under the 12-head architecture. Columns to the right report $A_{\mathrm{arch}}(o) = |\beta_{o,6h} - \beta_{o,12h}|$ for each optimizer.

| | | | $A_{\mathrm{arch}}(o)$ | | | | |
|---|---|---|---|---|---|---|---|
| Regime | Best opt. | $\Delta\beta^{\star}_{\mathrm{opt}}$ | AdamW | Muon | NorMuon | Dion (1/2) | Dion (1/16) |
| $\beta_{\mathrm{hard}}$(HEAD) | Muon | 0.330 | 0.345 | 0.651 | 0.060 | 0.130 | 0.039 |
| $\beta_{\mathrm{soft}}$(HEAD) | NorMuon | 0.457 | 0.250 | 0.419 | 0.133 | 0.206 | 0.048 |
| $\beta_{\mathrm{hard}}$(MID) | NorMuon | 0.703 | 0.429 | 0.294 | 0.131 | 0.086 | 0.093 |
| $\beta_{\mathrm{soft}}$(MID) | NorMuon | 0.481 | 0.147 | 0.230 | 0.104 | 0.015 | 0.011 |
| $\beta_{\mathrm{hard}}$(TAIL) | NorMuon | 0.600 | 0.313 | 0.367 | 0.023 | 0.014 | 0.011 |
| $\beta_{\mathrm{soft}}$(TAIL) | NorMuon | 0.424 | 0.051 | 0.198 | 0.019 | 0.034 | 0.002 |

The architectural intervention also changes the optimizer gap itself. We quantify this through

$$I_\star^{(b,m)} = \Delta\beta_{\text{opt}}^{\star,(b,m)}(6h) - \Delta\beta_{\text{opt}}^{\star,(b,m)}(12h). \tag{9}$$

Positive values indicate that reducing the number of heads increases the best achievable optimizer-induced gain over AdamW. Table 21 shows that $I_\star$ is positive in five of six regimes, with the largest increase in HEAD hard-rank scaling ($+0.306$). The only negative interaction occurs in MID hard rank, where the best gain decreases by $0.153$ but remains large. Thus, the architectural intervention changes not only the scaling exponents themselves, but also the size of the optimizer gap.

Table 21: Interaction of attention-rank architecture with the AdamW-to-best optimizer gain for GPT-2 160M. We report the best optimizer-induced gain over AdamW under the 12-head and 6-head architectures, together with $I_\star = \Delta\beta_{\text{opt}}^\star(6h) - \Delta\beta_{\text{opt}}^\star(12h)$.

| Regime | $\Delta\beta_{\text{opt}}^\star(12h)$ | $\Delta\beta_{\text{opt}}^\star(6h)$ | Interaction $I_\star$ |
|---|---|---|---|
| $\beta_{\text{hard}}(\text{HEAD})$ | 0.330 | 0.636 | +0.306 |
| $\beta_{\text{soft}}(\text{HEAD})$ | 0.457 | 0.612 | +0.155 |
| $\beta_{\text{hard}}(\text{MID})$ | 0.703 | 0.550 | -0.153 |
| $\beta_{\text{soft}}(\text{MID})$ | 0.481 | 0.519 | +0.038 |
| $\beta_{\text{hard}}(\text{TAIL})$ | 0.600 | 0.637 | +0.037 |
| $\beta_{\text{soft}}(\text{TAIL})$ | 0.424 | 0.555 | +0.131 |

Table 22 reports the fitted exponents used to compute both the signed architectural effects and the optimizer-induced gains. These values show a **stronger form of optimizer–architecture interaction**, the attention-rank intervention changes which optimizer is best matched to the architecture. Under the 12-head baseline, NorMuon attains the largest scaling exponent in five of six regimes, with Muon leading only in HEAD hard rank. Under the 6-head architecture, Muon attains the largest exponent in all six regimes. Thus, reducing the number of heads does not act as a uniform offset applied to all optimizers. It changes the optimizer ordering itself.

Table 22: Fitted scaling exponents under 12-head and 6-head architectures for GPT-2 160M. Each entry reports $(\beta_{12h}, \beta_{6h})$ for the corresponding optimizer, frequency regime, and rank metric.

| Regime | AdamW | Muon | NorMuon | Dion (1/2) | Dion (1/16) |
|---|---|---|---|---|---|
| $\beta_{\text{hard}}(\text{HEAD})$ | (0.263, 0.608) | (0.593, 1.244) | (0.429, 0.489) | (0.520, 0.390) | (0.349, 0.388) |
| $\beta_{\text{soft}}(\text{HEAD})$ | (0.441, 0.691) | (0.884, 1.303) | (0.898, 0.765) | (0.886, 0.680) | (0.699, 0.651) |
| $\beta_{\text{hard}}(\text{MID})$ | (0.242, 0.671) | (0.927, 1.221) | (0.945, 1.076) | (0.666, 0.580) | (0.460, 0.367) |
| $\beta_{\text{soft}}(\text{MID})$ | (0.449, 0.596) | (0.885, 1.115) | (0.930, 1.034) | (0.815, 0.830) | (0.679, 0.668) |
| $\beta_{\text{hard}}(\text{TAIL})$ | (0.438, 0.751) | (1.021, 1.388) | (1.038, 1.015) | (0.879, 0.865) | (0.404, 0.393) |
| $\beta_{\text{soft}}(\text{TAIL})$ | (0.618, 0.669) | (1.026, 1.224) | (1.042, 1.061) | (0.952, 0.918) | (0.718, 0.716) |

**These diagnostics support the arguments for architecture-optimizer co-design.** Optimizer-induced gains are usually larger than the direct attention-rank-induced shift, and the same architectural intervention has optimizer-dependent effects, increasing per-head attention rank changes which optimizer achieves the largest spectral-scaling exponent. In other words, attention-rank changes and optimizer geometry act as coupled, rather than separable, axes of representation scaling.

### H.1 Optimizer-Induced Scaling Effects Dominate the Attention-Rank Interventions

Having established that optimizer choice changes how FFN width is converted into realized capacity for a fixed architecture, we next ask whether these optimizer-induced effects are comparable to, or larger than, a controlled architectural intervention. Motivated by recent work showing that attention rank can limit expressivity (Amsel et al., 2025; Garg et al., 2022), we increase per-head attention rank by reducing the number of attention heads, which increases the attention ranks while preserving the total parameter count.

For each frequency regime $b$ and spectral rank metric $m$, we compare two effects on the fitted scaling exponent $\beta$. The optimizer-induced gain over AdamW in the original 12-head architecture as:

$$\Delta\beta_{\text{opt}}^{\star,(b,m)} = \max_o \beta_{o,12h}^{(b,m)} - \beta_{\text{AdamW},12h}^{(b,m)}. \tag{10}$$

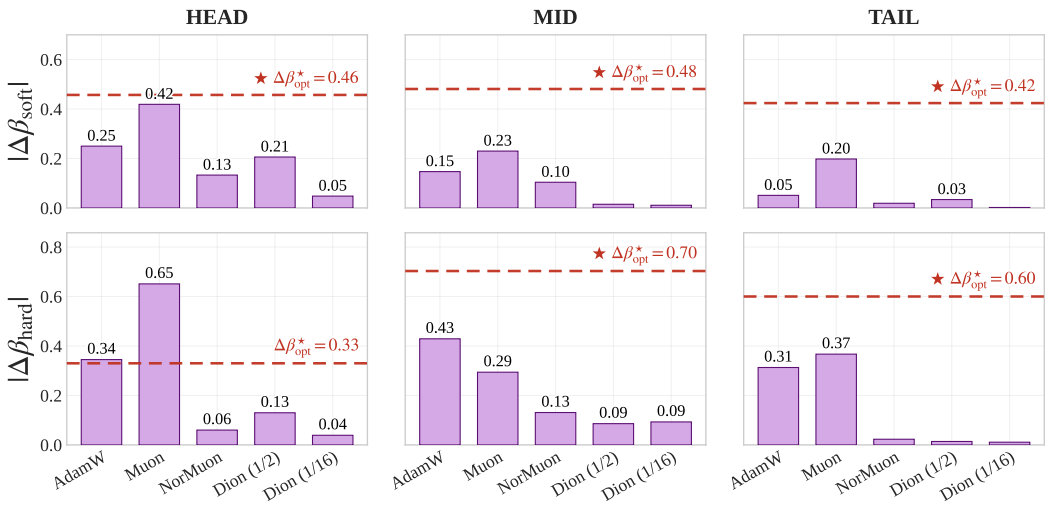



Figure 10: Optimizer-induced shifts in spectral-scaling exceed attention-rank shifts in GPT-2 160M. We compare the optimizer-induced spectral scaling shift for AdamW in the original 12-head architecture, $\Delta\beta^{\star}_{\text{opt}}$ (shown as red dashed lines), and the spectral scaling shift induced by architectural intervention, increasing the attention ranks at fixed total parameter count, across each optimizer (shown as bars). Optimizer-induced gains exceed attention-rank shifts in 5 of 6 frequency regimes (marked with ★), across 28 of 30 scaling shifts; the only exceptions occur in HEAD hard-rank scaling.

where the maximum is taken over optimizers. The attention-rank induced spectral scaling shift for optimizer $o$ is

$$A^{(b,m)}_{\text{rank}}(o) = \left|\beta^{(b,m)}_{o,6h} - \beta^{(b,m)}_{o,12h}\right|. \tag{11}$$

Further, to test how uniformly architectural intervention is expressed across optimizers, we report the signed architectural effect for each optimizer $o$, frequency regime $b$, and rank metric $m$ as:

$$\Delta\beta^{(b,m)}_{\text{arch}}(o) = \beta^{(b,m)}_{o,6h} - \beta^{(b,m)}_{o,12h}. \tag{12}$$

These comparisons investigate whether optimizer choice changes the spectral-scaling exponents more than the tested architectural intervention, such as increasing attention-rank under fixed parameter count, and whether the intervention is expressed uniformly across optimizers.

**Optimizer effects exceed attention-rank effects in nearly all regimes.** Figure 10 shows that the optimizer-induced gain is larger than the attention-rank shift in 28 of 30 comparisons. The only exceptions occur in $\beta_{\text{hard}}(\text{HEAD})$, where the attention-rank shifts under AdamW (0.345) and Muon (0.651) exceed the optimizer-induced gain (0.330). Thus, *HEAD hard-rank scaling is the most architecture-sensitive regime under this intervention.* However, in MID and TAIL regimes the optimizer-induced shift exceeds the attention-rank-induced shifts for every optimizer. For instance, the optimizer-induced $\beta_{\text{hard}}(\text{MID})$ gain is 0.703, exceeding the largest attention-rank shift (0.43 under AdamW); and $\beta_{\text{hard}}(\text{TAIL})$ gain is 0.6, exceeding the largest shift of 0.367 under Muon.

**Attention-rank-induced scaling shifts depend on optimizer geometry.** The signed effect (Eq. 12) shows that this architectural intervention is not expressed uniformly across optimizers. For $\beta_{\text{hard}}(\text{TAIL})$, increasing per-head attention rank raises the exponent under AdamW and Muon by $+0.313$ and $+0.367$, respectively, but has near-zero effect under NorMuon and Dion variants ($|\Delta\beta_{\text{arch}}| \leq 0.023$). Moreover, this effect is also selective across spectral ranks. For TAIL tokens, attention-rank shift is larger for $\beta_{\text{hard}}$ than for $\beta_{\text{soft}}$ under AdamW ($+0.313$ vs. $+0.051$) and Muon

Table 23: Signed attention-rank architectural effects (Eq. 12) in GPT-2 160M. Both, the sign and magnitude of the architectural effect are strongly optimizer-dependent.

| Regime | AdamW | Muon | NorMuon | Dion(1/2) | Dion(1/16) |
|---|---|---|---|---|---|
| $\beta_{\text{hard}}(\text{HEAD})$ | +0.345 | +0.651 | +0.060 | -0.130 | +0.039 |
| $\beta_{\text{soft}}(\text{HEAD})$ | +0.250 | +0.419 | -0.133 | -0.206 | -0.048 |
| $\beta_{\text{hard}}(\text{MID})$ | +0.429 | +0.294 | +0.131 | -0.086 | -0.093 |
| $\beta_{\text{soft}}(\text{MID})$ | +0.147 | +0.230 | +0.104 | +0.015 | -0.011 |
| $\beta_{\text{hard}}(\text{TAIL})$ | +0.313 | +0.367 | -0.023 | -0.014 | -0.011 |
| $\beta_{\text{soft}}(\text{TAIL})$ | +0.051 | +0.198 | +0.019 | -0.034 | -0.002 |

(+0.367 vs. +0.198). This suggest that increased per-head rank *primarily affects dominant-mode capacity* rather than diffuse spectral spread.

This intervention also changes the optimizer–architecture match. In the original 12-head architecture, NorMuon attains the largest scaling exponent in five of six token-regimes; however, in 6-head architecture, Muon attains the largest exponent in all six regimes. Thus, increasing per-head attention rank does not impose a fixed spectral shift across optimizers. Rather, the architectural intervention changes which optimizer most effectively converts added FFN width into spectral capacity. Refer to Appendix H for more in-depth analysis.

# I Removing RoPE Reshapes Optimizer-Dependent Spectral Scaling

We next study a second architectural intervention, removing explicit positional signal. RoPE injects relative positional structure into self-attention (Su et al., 2024), while recent work shows that Transformers can operate without explicit positional embeddings and may benefit from NoPE designs (Kazemnejad et al., 2023; Chi et al., 2023; Gelberg et al., 2026). We evaluate this intervention at 350M scale for Muon and NorMuon, whose power-law fits, for both RoPE and NoPE, are reliable across frequency regimes. More importantly, this comparison separates two related optimizer geometries: both optimizers use orthonormalized updates, while NorMuon additionally imposes per-neuron normalization.

**NoPE produces optimizer-dependent capacity redistribution.** Table 24 shows that removing RoPE does not induce a uniform change in spectral scaling. The clearest sign reversal occurs in HEAD hard-rank scaling, NoPE increases Muon's exponent from 1.083 to 1.286 ($\Delta\beta = +0.203$), but decreases NorMuon's from 1.023 to 0.656 ($\Delta\beta = -0.367$). Under Muon, NoPE shifts hard-capacity scaling toward HEAD tokens while reducing MID and TAIL scaling ($1.008 \to 0.744$ and $1.127 \to 0.836$). Under NorMuon, NoPE decreases hard-rank scaling in all three frequency regimes, with the largest drop in HEAD. Thus, hard–soft asymmetry also changes: under Muon, MID and TAIL asymmetry shift from negative to positive ($-0.085 \to +0.096$ and $-0.129 \to +0.091$), while under NorMuon asymmetry increases in all three regimes, with the largest increase in HEAD ($+0.045 \to +0.315$).

Table 24: Removing RoPE induces optimizer-dependent spectral redistribution. Hard-rank scaling exponents and hard–soft asymmetry are reported for GPT-2 350M RoPE/NoPE models. Removing RoPE under Muon increases $\beta_{\text{hard}}$ for HEAD, but decreases for MID and TAIL tokens. Under NorMuon it decreases across all token regimes. All reported fits have $R^2 \geq 0.93$.

| Optimizer | | $\beta_{\text{hard}}$ | $\Delta\beta_{\text{hard}}$ | $\Delta_{1,2}$ |
|---|---|---|---|---|
| Muon | HEAD | 1.083 / 1.286 | +0.203 | −0.201 / −0.004 |
| | MID | 1.008 / 0.744 | −0.264 | −0.085 / +0.096 |
| | TAIL | 1.127 / 0.836 | −0.291 | −0.129 / +0.091 |
| NorMuon | HEAD | 1.023 / 0.656 | −0.367 | −0.045 / +0.315 |
| | MID | 0.723 / 0.569 | −0.154 | +0.145 / +0.276 |
| | TAIL | 0.881 / 0.739 | −0.142 | +0.038 / +0.136 |

In other words, positional-signal removal changes not only the magnitude of hard-rank scaling, but also how added FFN width is converted into dominant-mode vs diffuse spectral capacity. Appendix K provides a position-dependent FFN analysis supporting this interpretation.

**NoPE redistributes optimizer gaps across token-frequency regimes.** Removing RoPE also changes where optimizer differences appear, across the token-frequency regimes. The Muon–NorMuon hard-rank gap widens sharply in HEAD tokens, from $|1.083 - 1.023| = 0.06$ under RoPE to $|1.286 - 0.656| = 0.63$ under NoPE, because NoPE improves Muon while degrading NorMuon. In contrast, the gap narrows in MID ($|1.008 - 0.723| = 0.285$ to $|0.744 - 0.569| = 0.175$) and TAIL ($|1.127 - 0.881| = 0.246$ to $|0.836 - 0.739| = 0.097$). Thus, removing positional signal redistributes optimizer-dependent spectral capacity across the token-frequency spectrum.

**Optimizer choice remains the larger TAIL hard-rank effect.** Table 25 compares effect sizes in the TAIL hard-rank regime. The AdamW→Muon optimizer shift ($|\Delta\beta| = 0.74$) is about $2.0\times$ the attention-rank effect and $2.5$–$5.2\times$ the RoPE→NoPE effects. Thus, optimizer

Table 25: Effect-size synthesis for TAIL hard-rank scaling. We compare the magnitude of the AdamW→Muon optimizer gap with TAIL hard-rank shifts induced by attention-rank (160M) and RoPE/NoPE interventions (350M).

| Intervention | $\lvert\Delta\beta_{\text{hard}}\rvert$ | Relative size |
|---|---|---|
| AdamW → Muon | 0.74 | 1.00× |
| $12h \to 6h$ (Muon) | 0.37 | 0.50× |
| RoPE → NoPE (Muon) | 0.29 | 0.39× |
| RoPE → NoPE (NorMuon) | 0.14 | 0.19× |

Table 26: Validation perplexity for partial PostLN configurations in GPT-2 160M at $4\times$ FFN width. In PostLN-$k$, the first $k\%$ of layers use PostLN and the remaining layers use PreLN. Diverged runs (PPL $> 1000$) are marked with $\times$. Lower is better.

| Optimizer | PostLN-25 | PostLN-50 | PostLN-75 | Full PostLN |
|---|---|---|---|---|
| AdamW (lr=$10^{-4}$) | 64.6 | 65.6 | 106.7 | $\times$ |
| AdamW (lr=$3 \times 10^{-4}$) | 41.9 | $\times$ | $\times$ | $\times$ |
| Muon | 28.7 | 30.1 | 40.9 | $\times$ |
| NorMuon | 28.7 | 29.9 | **32.8** | $\times$ |
| Dion ($r = 1/2$) | 29.1 | 31.7 | $\times$ | $\times$ |
| Dion ($r = 1/16$) | 30.7 | 34.7 | $\times$ | $\times$ |

choice remains the dominant source of variation in the TAIL hard-rank comparison among the measured interventions. At the same time, the RoPE/NoPE results show that architectural signal removal reshapes the optimizer gap itself.

These results show that the effect of architectural interventions on representational scaling is mediated by optimizer geometry. Increasing per-head attention rank can alter optimizer ordering, while removing RoPE redistributes optimizer gaps across token-frequency regimes. Thus, both capacity-axis and positional-signal interventions reshape spectral scaling through the architecture–optimizer pair rather than acting as independent perturbations.

## J Optimizer Geometry Expands the Trainable Normalization Space

Sections H.1 and Appendix H show that optimizer geometry changes the spectral scaling behavior of trainable architectures. Here, we study a complementary form of optimizer–architecture coupling: whether optimizer choice changes which normalization-placement architectures can be trained at useful perplexity in the first place. This is not a spectral-scaling experiment as all models use the same $4\times$ FFN width. Instead, it tests whether optimizer geometry changes the feasible region of the architecture search space before spectral scaling is measured.

We use normalization placement as a trainability stress test. In a partial PostLN configuration, denoted PostLN-$k$, the first $k\%$ of layers use PostLN, while the remaining layers use PreLN. This creates a controlled interpolation between the more trainable PreLN regime and the more difficult PostLN regime. Since PostLN is known to be difficult to train at scale due to gradient amplification across layers (Xiong et al., 2020), varying the PostLN fraction lets us test whether optimizers differ in the normalization configurations they can train to useful perplexity.

**Muon-family optimizers remain trainable in higher-PostLN regimes.** Table 26 shows a clear trainability gap. At lr=$3 \times 10^{-4}$, AdamW trains PostLN-25 but diverges for PostLN-50 and PostLN-75. Lowering the AdamW learning rate to $10^{-4}$ avoids divergence up to PostLN-75, but it reaches to PPL $= 106.7$, compared with PPL $= 40.9$ for Muon and PPL $= 32.8$ for NorMuon. Thus, AdamW can be made stable only by moving to a substantially worse optimization regime, whereas Muon-family optimizers train these partial PostLN configurations at useful perplexity.

**Neuron-wise normalization helps in the most aggressive partial-PostLN regime.** Muon and NorMuon perform similarly at lower PostLN fractions, both achieve PPL $= 28.7$ at PostLN-25 and nearly identical perplexity at PostLN-50. The difference appears at PostLN-75, where NorMuon improves over Muon from PPL $= 40.9$ to PPL $= 32.8$. This suggests that NorMuon's neuron-wise normalization of orthogonalized updates provides additional stability when the PostLN fraction is large and gradient amplification is more severe. However, this stabilization is not unlimited, full PostLN remains unstable for all optimizers.

**Trainability and spectral-capacity scaling are related but distinct.** The Dion variants clarify that the trainability frontier is not determined only by update rank. Both Dion(1/2) and Dion(1/16) train PostLN-50, but both fail at PostLN-75. This contrasts with the spectral-scaling results in Section 4.3,

where Dion rank strongly modulates capacity exponents. This distinction suggests that orthonormal update structure is important for stabilizing difficult normalization configurations, while update rank controls how much of the available capacity is converted into effective dominant-mode capacity.

These results broaden the co-design interpretation. Optimizer choice changes how trainable architectures convert added FFN width into useful spectral capacity. Here, optimizer choice changes which normalization-placement architectures are trainable at useful perplexity in the first place. Thus, optimizer geometry affects both *capacity utilization* within a fixed architecture and the *feasible architecture space* itself.

# K Position-Dependent FFN Spectral Transformation Under RoPE and NoPE

Section I shows that removing RoPE induces optimizer-dependent spectral redistribution. In this section, we provide a position-dependent analysis, and investigate whether the redistribution is accompanied by changes in how the FFN nonlinearity transform position-dependent information.

**Symmetry-ratio metric.** Following (Lavie et al., 2024), we quantify how much sequence position explains FFN activation variance. For activations within a given layer and token-frequency regime, we compute an ANOVA-style position-dependence score

$$\eta^2 = \frac{\mathrm{Var}_{\text{between position}}}{\mathrm{Var}_{\text{total}}}. \tag{13}$$

We report its complement, the symmetry ratio

$$\mathrm{SR} = 1 - \eta^2. \tag{14}$$

Higher $\mathrm{SR}$ means the representation is more position-independent, lower $\mathrm{SR}$ means sequence position explains more activation variance. We compute this statistic before and after the FFN nonlinearity: $\mathrm{SR}_{\mathrm{pre}}$ and and $\mathrm{SR}_{\mathrm{post}}$. We define

$$\Delta\mathrm{SR} = \mathrm{SR}_{\mathrm{post}} - \mathrm{SR}_{\mathrm{pre}}. \tag{15}$$

Thus, $\Delta\mathrm{SR} < 0$ indicates that the FFN nonlinearity makes the representation more position-dependent, while $\Delta\mathrm{SR} > 0$ indicates suppression of position-dependence. In the tables below, $\overline{|\Delta\mathrm{SR}|}$ measures the layer-averaged magnitude of FFN-induced positional processing, while the sign of $\Delta\mathrm{SR}$ determines whether this processing amplifies or suppresses position-dependence. **This analysis separates three effects:** whether NoPE changes the amount of FFN-induced positional processing, where in depth this processing is localized, and which token-frequency regime receives stronger amplification.

## K.1 Positional Amplification Tracks TAIL Hard-Rank Scaling

Before analyzing RoPE/NoPE, we first check whether the symmetry-ratio diagnostic tracks the TAIL hard-rank scaling behavior observed in the 160M baseline. Table 27 shows that *optimizers with stronger TAIL hard-rank scaling also exhibit stronger TAIL positional amplification.* NorMuon, Muon, and Dion $(1/2)$ have the largest $\overline{|\Delta\mathrm{SR}|}$ values and the largest TAIL hard-rank exponents, whereas AdamW and Dion $(1/16)$ show near-zero amplification and weaker hard-rank scaling. This does not establish causality, but it suggests that improved TAIL dominant-mode scaling is accompanied by stronger FFN-mediated position-dependent processing.

## K.2 NoPE Changes the Amount of TAIL Positional Processing

We next compare RoPE and NoPE at 350M. Table 28 summarizes the layer-averaged magnitude of TAIL positional processing at the smallest and largest FFN widths in the sweep. Removing RoPE substantially increases TAIL positional amplification for Muon and NorMuon. At $1d$, Muon increases from $0.0380$ to $0.1252$ and NorMuon from $0.0419$ to $0.1313$, corresponding to roughly $3\times$ larger TAIL positional processing. At $4d$, Muon and NorMuon remain $2.31\times$ and $2.93\times$ above their RoPE baselines, respectively. The NoPE-induced effect is strongest at $1d$, suggesting that learned positional compensation is most pronounced when FFN width is more constrained. AdamW does not follow this trend; at $4d$, its TAIL amplification decreases from $0.0183$ under RoPE to $0.0133$ under NoPE.

Table 27: TAIL positional amplification and hard-rank scaling at 160M. We report layer-averaged $\overline{|\Delta\mathrm{SR}|}$ for TAIL tokens, the most negative layer-wise $\Delta\mathrm{SR}$ value, its peak layer, and the corresponding TAIL hard-rank scaling exponent. More negative peak values indicate stronger FFN-induced amplification of position-dependence.

| Optimizer | $\overline{\lvert\Delta\mathrm{SR}\rvert}$ | Peak $\Delta\mathrm{SR}$ | Peak Layer | $\beta_{\mathrm{hard}}$ |
|---|---|---|---|---|
| NorMuon | 0.087 | $-0.711$ | 5 | 1.04 |
| Muon | 0.078 | $-0.529$ | 4 | 1.02 |
| Dion (1/2) | 0.075 | $-0.441$ | 3 | 0.88 |
| Dion (1/16) | 0.005 | $-0.041$ | 6 | 0.40 |
| AdamW | 0.003 | $-0.008$ | 10 | 0.44 |

Table 28: TAIL positional amplification under RoPE and NoPE for GPT-2 350M. Values are layer-averaged $\overline{|\Delta\mathrm{SR}|}$ at the smallest and largest FFN widths. Ratios compare NoPE to RoPE.

| | $1d$ | | | $4d$ | | |
|---|---|---|---|---|---|---|
| Optimizer | RoPE | NoPE | Ratio | RoPE | NoPE | Ratio |
| AdamW | 0.0096 | 0.0212 | 2.21× | 0.0183 | 0.0133 | 0.73× |
| Muon | 0.0380 | 0.1252 | 3.29× | 0.0196 | 0.0453 | 2.31× |
| NorMuon | 0.0419 | 0.1313 | 3.13× | 0.0213 | 0.0624 | 2.93× |

The signed values confirm that this effect corresponds to positional amplification rather than only a change in magnitude. Under NoPE, Muon and NorMuon have negative signed $\Delta\mathrm{SR}(\mathrm{TAIL})$ across all widths, with much larger magnitudes than AdamW. Across settings, $\mathrm{SR}_{\mathrm{pre}}(\mathrm{TAIL})$ remains high, typically between $0.95$ and $0.98$, indicating that the largest differences arise in how the FFN nonlinearity changes position-dependence rather than from different pre-activation symmetry.

### K.3 NoPE Changes the Depth Localization of Positional Processing

Table 29 reports where TAIL positional amplification is strongest across layers. Under RoPE, peak TAIL processing for Muon often occurs in deeper layers (Layer $8, 23, 23, 23$ across the $1d$–$4d$ sweep). Under NoPE, Muon shifts this peak to the earliest layers (Layer $2, 2, 1, 1$). NorMuon also shifts earlier under NoPE, but its peak is more distributed across layers (Layer $3, 11, 10, 2$). AdamW remains concentrated in mid-to-deeper layers under NoPE. Thus, *NoPE changes where positional processing is concentrated in depth, and this localization differs by optimizer.*

Table 29: Peak TAIL positional amplification under RoPE and NoPE for GPT-2 350M. Each entry reports the most negative TAIL $\Delta\mathrm{SR}$ value, with the peak layer in parentheses. Under NoPE, Muon concentrates peak positional processing in the earliest layers, while NorMuon shifts earlier but remains more distributed. Layer indices are zero-based.

| | RoPE: Peak $\Delta\mathrm{SR}$ (Layer) | | | | NoPE: Peak $\Delta\mathrm{SR}$ (Layer) | | | |
|---|---|---|---|---|---|---|---|---|
| Optimizer | $1d$ | $2d$ | $3d$ | $4d$ | $1d$ | $2d$ | $3d$ | $4d$ |
| AdamW | $-0.049$ (23) | $-0.039$ (7) | $-0.046$ (23) | $-0.036$ (23) | $-0.257$ (15) | $-0.150$ (15) | $-0.038$ (18) | $-0.127$ (17) |
| Muon | $-0.172$ (8) | $-0.196$ (23) | $-0.085$ (23) | $-0.160$ (23) | $-0.823$ (2) | $-0.586$ (2) | $-0.456$ (1) | $-0.582$ (1) |
| NorMuon | $-0.373$ (7) | $-0.174$ (23) | $-0.149$ (23) | $-0.109$ (6) | $-0.817$ (3) | $-0.582$ (11) | $-0.374$ (10) | $-0.470$ (2) |

This depth-localization pattern provides a useful lens for the spectral scaling behavior of RoPE and NoPE configurations. Muon concentrates strong position-dependent processing very early, whereas NorMuon produces strong but more distributed processing. This difference is consistent with the observation that Muon and NorMuon exhibit different spectral-scaling behavior under NoPE.

### K.4 NoPE Reverses the Frequency Bias of Positional Processing

We also examine whether the FFN nonlinearity amplifies position-dependence more for frequent or rare tokens. Since negative $\Delta\mathrm{SR}$ indicates positional amplification, we report

$$\Delta\mathrm{SR}_{\mathrm{HEAD}} - \Delta\mathrm{SR}_{\mathrm{TAIL}}. \tag{16}$$

Positive values mean that TAIL tokens are amplified more strongly than HEAD tokens. Under RoPE, all optimizers and widths have negative values, indicating a HEAD-favored positional-processing

bias. Under NoPE, Muon and NorMuon flip to positive values across all widths, indicating stronger TAIL positional amplification. AdamW remains negative under NoPE (Table 30).

This reversal shows that removing explicit positional signal changes which frequency regimes receive stronger FFN-induced positional amplification. Under NoPE, Muon and NorMuon reorganize toward TAIL-favored positional processing, whereas AdamW does not show the same reversal. This provides a direct frequency-level signature of optimizer-dependent positional reorganization under NoPE.

Table 30: HEAD–TAIL positional-processing bias under RoPE and NoPE for GPT-2 350M. Values report $\Delta\mathrm{SR}_{\mathrm{HEAD}} - \Delta\mathrm{SR}_{\mathrm{TAIL}}$. Positive values mean TAIL tokens receive stronger positional amplification. Under RoPE, all optimizers are HEAD-biased; under NoPE, Muon and NorMuon become TAIL-biased.

| | RoPE | | | | NoPE | | | |
|---|---|---|---|---|---|---|---|---|
| Optimizer | $1d$ | $2d$ | $3d$ | $4d$ | $1d$ | $2d$ | $3d$ | $4d$ |
| AdamW | −0.0214 | −0.0300 | −0.0380 | −0.0325 | −0.0212 | −0.0428 | −0.0238 | −0.0060 |
| Muon | −0.0064 | −0.0173 | −0.0207 | −0.0151 | 0.0262 | 0.0256 | 0.0147 | 0.0127 |
| NorMuon | −0.0040 | −0.0088 | −0.0148 | −0.0117 | 0.0438 | 0.0275 | 0.0031 | 0.0032 |

Hence, above position-dependence analysis provides a structured view of the capacity redistribution, as shown in Section I. Precisely, under NoPE, Muon increases HEAD hard-rank scaling while decreasing MID and TAIL. Whereas NorMuon decreases hard-rank scaling across all token regimes. The position-dependence analysis provides the explanation; under NoPE, Muon and NorMuon amplify TAIL position-dependence, shift peak processing toward earlier layers, and reverse the HEAD–TAIL positional-processing bias. However, their localization differs, Muon concentrates peak processing in the earliest layers, while NorMuon is more distributed.

Thus, NoPE changes not only the amount of FFN-induced positional processing, but also its allocation across depth and token frequency. These patterns support the broader conclusion that architectural signal changes are expressed through optimizer-dependent FFN processing rather than as optimizer-independent perturbations.