Spectral Scaling Laws in Language Models: *How Effectively Do Feed-Forward Networks Use Their Latent Space?*

Nandan Kumar Jha
New York University
nj2049@nyu.edu

Brandon Reagen
New York University
bjr5@nyu.edu

Abstract

As Large Language Models (LLMs) scale, the question is not just how large they become, but *how much of their capacity is effectively utilized*. Existing scaling laws relate model size to loss, yet overlook how components exploit their latent space. In this work, we focus on Feed-Forward Networks (FFNs) and recast width selection as a spectral utilization optimization problem. Using a lightweight diagnostic suite: Hard Rank (participation ratio), Soft Rank (Shannon Rank), Spectral Concentration, and the composite Spectral Utilization Index (SUI), we quantify how many latent directions are meaningfully activated across LLaMA, GPT-2, and nGPT families. Our *key finding* is an **Asymmetric Spectral Scaling Law**: soft rank follows an almost perfect power law with FFN width, while hard rank grows only sublinearly, with high variance. This asymmetry suggests that widening FFNs mostly adds low-energy tail directions, while dominant-mode subspaces saturate early. Moreover, at larger widths, variance further collapses into a narrow subspace, leaving much of the latent space under-utilized. These results recast FFN width selection as a principled trade-off between tail capacity and dominant-mode capacity, offering concrete guidance for inference-efficient LLM design.

1 Introduction

As Large Language Models (LLMs) continue to grow in scale and complexity, a central blind spot remains: *How effectively is their internal capacity utilized?* Existing empirical scaling laws (Kumar et al., 2025; Tao et al., 2024; Sardana et al., 2024; Kaplan et al., 2020) relate model performance to factors such as width, depth, and data size, but they offer little insight into how different architectural components exploit, or potentially squander, the high-dimensional latent space. These laws treat models as black boxes, abstracting away the internal dynamics of transformer blocks and leaving open questions about representational usage.

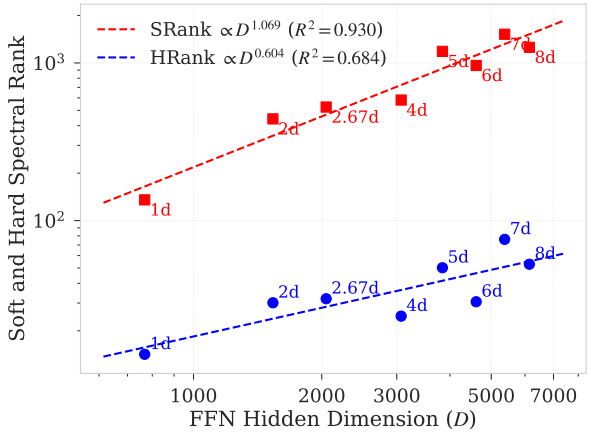


Figure 1: Spectral rank vs. FFN hidden dimension in LLaMA-130M base model, with width sweep $D = \alpha d$ (total parameters therefore differ across α). Log-Log fits: Soft rank follows a linear power-law fit (β=1.06, R^2=0.93), while hard rank grows sublinearly (β=0.60, R^2=0.68), indicating width mainly adds low-energy tail directions rather than enlarging the high-energy dominant-mode subspace.

Among transformer components, FFNs dominate the parameter budges as they can account for as much as 67% of the total parameters in decoder-only models (Pires et al., 2023; Geva et al., 2021). Yet, FFN width is typically set by rules of thumb rather than design principles, e.g., 4× expansion in GPT-2 (Radford et al., 2019) and 2.67× in LLaMA (Touvron et al., 2023). Even in recent LLMs such as Qwen (Hui et al., 2024), the FFN width varies substantially across model sizes (≈2.4-5.8×) underscoring the lack of theoretical grounding.

Despite their prevalence, we still lack a clear understanding of how FFN width affects effective capacity usage. This raises three questions: *Is increasing FFN width always beneficial for expressivity? How many latent directions are actually used in practice? Can we quantify representational efficiency beyond FLOPs and loss?*

We address these questions by reframing FFN width selection as a *spectral utilization* problem. The **intuition** is straightforward: if wider FFNs

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35059–35070
November 4-9, 2025

truly expand usable capacity, then their spectrum should reflect growth in the effective dimensionality of the subspace the model exploits. To test this, we conduct a layer-wise spectral audit across GPT-2, LLaMA, and nGPT (Loshchilov et al., 2025) backbones, analyzing the eigenspectrum of post-activation covariance over training steps and layers.

We quantify utilization using four lightweight, differentiable metrics: Hard Rank (participation ratio) to capture the dimensionality of the high-energy, or the dominant, mode (Gao et al., 2017); Soft Rank (Shannon Rank) to quantify uniformity across all directions (De Domenico and Biamonte, 2016); Spectral Concentration (eigenvalue early enrichment) to quantify how much variance is captured by leading eigenvalues (Marbut et al., 2023); and finally Spectral Utilization Index (SUI), a composite metric that harmonically combines hard and soft rank to balance dominant-mode and tail usage

Through systematic analysis across the FFN width sweep $D{=}\alpha d$, where $\alpha \in \{1, 2, 2.67, 4, 5, 6, 7, 8\}$, and model sizes ranging from 70M to 250M parameters, we uncover an Asymmetric Spectral Scaling Law that fundamentally changes our understanding of capacity allocation. The power law (Log-Log) fits reveal a striking asymmetry (Figure 1): While soft spectral rank scales near-perfectly with FFN width ($\beta \to 1, R^2 \to 1$), hard spectral rank, measuring the dominant subspace, plateaus early with weak, noisy scaling ($\beta \approx 0.5, R^2 \approx 0.5$).

This asymmetry highlights that widening FFNs operates through *tail-first growth*: predominantly adding low-energy directions while the high-energy mode saturates early. In other words, capacity increases, but it is increasingly allocated to directions that carry little variance. This effect resembles the well-known spectral bias in function space, where low input frequencies are learned before high ones (Rahaman et al., 2019). Both perspectives point to the same underlying principle: capacity is allocated unevenly across modes, though expressed in different bases (Fourier vs. activation eigenspectrum).

Contributions. This work makes four main contributions: **Conceptual.** We reframe FFN width selection, traditionally treated as an implementation detail, as a problem of spectral utilization, and introduce the first principled framework for understanding how FFN capacity is allocated with their width scaling. **Theoretical.** We uncover Asymmetric Spectral Scaling Laws that capture divergent growth between soft and hard spectral ranks. These laws reveal that FFN widening follows a *tail-first growth* pattern, explaining why naive width scaling can yields diminishing returns. **Methodological.** We develop a lightweight, differentiable diagnostic suite for tracking layerwise representational usage during training. This includes a closed-form estimator, $K_{\text{eff}} = 1+(D-1)\cdot\text{SUI}$, which links utilization to effective dimension. **Empirical.** Across diverse architectures and scales, we show that (i) soft/hard rank asymmetry persist across model families, (ii) optimal widths are consistently narrower than those used in practice, (iii) LayerNorm placement critically shapes utilization: Post-LN suppresses tail capacity scaling, whereas Mix-LN (Li et al., 2025) improves dominant-mode scaling while preserving near-linear tail growth.

2 Related Work

Cost-aware neural scaling. The foundational work (Kaplan et al., 2020) established the power-law relations between loss and compute, later refined by the Chinchilla laws (Hoffmann et al., 2022), which showed that many models are compute-suboptimal, too wide and under-trained for their budgets. Follow-up studies (Sardana et al., 2024) extended this perspective to deployment: under heavy traffic, the compute-optimal point shifts toward smaller models trained on more tokens, lowering inference cost. Paquette et al. (2024) map the regimes where capacity, optimizer noise, or embedding quality dominate under fixed budgets.

Other orthogonal cost factors have also been identified: vocabulary should scale with width (Tao et al., 2024); reduced numerical precision effectively shrinks parameter count (Kumar et al., 2025); and robust estimation methods enable reliable scaling-law fits from small pilot runs (Choshen et al., 2024). These studies map efficiency trade-offs along multiple axes—compute, traffic, vocabulary, and precision. Our spectral-utilization laws introduce a *complementary axis:* they target latent-space usage, capturing how width is actually employed rather than measured by FLOPs alone.

Universality and representational capacity. After normalizing for efficiency offsets, checkpoints spanning models from GPT-2 to PaLM have been shown to collapse onto a single sigmoidal curve, suggesting a shared scaling trajectory across architectures (Ruan et al., 2024). The *Physics of LMs* series reports a related regularity for factual knowledge: a ≤ 2 bits/parameter ceiling that ap-

pears largely architecture-agnostic (Allen-Zhu and Li, 2025). Earlier work traced such apparent universality to heavy-tailed eigenspectra and implicit self-regularization (Martin and Mahoney, 2021). More recent analyses refine this view: small singular values have been shown to encode critical information in pretrained Transformers (Staats et al., 2024), while spectral collapse has been linked to over-smoothing dynamics in attention stacks (Dovonon et al., 2024).

Architectural and domain-specific scaling Scaling exponents are not architecture-agnostic. Tay et al. (2022) show that the most effective inductive bias shifts with scale: Switch-Transformers (Fedus et al., 2022) dominate in smaller parameter regimes, Performers (Choromanski et al., 2020) at mid-scale, and vanilla attention at large scale. Cabannes et al. (2024) derive exact scaling laws for associative-memory matrices, while Shi et al. (2024) explain why larger models can underperform on time-series tasks by introducing a look-back-aware law. Fort (2025) frames adversarial robustness as a scaling phenomenon, showing that resistance to attack remains nearly constant across two orders of magnitude in model size. Finally, Lyu et al. (2025) present an analytically solvable attention mechanism that yields closed-form power laws, providing a theoretical baseline.

These threads underscore that scaling is multi-faceted, bending with inductive bias, data modality, precision, and security constraints, precisely the facets our spectral scaling laws aim to highlight across GPT-2, LLaMA, and nGPT.

3 Method

In this section, we explain our methodology for extracting layer-wise covariance spectra from FFN internal representation, and describe the four spectral metrics that quantify spectral utilization, and capture various aspect of spectrum (e.g., uniformity vs spikes). We finish with the end-to-end algorithm and a short complexity analysis.

3.1 Preliminaries and Eigendecomposition

Notation Let an L-layer transformer be given. Each transformer consist of an FFN layer whose hidden width is D; the width multiplier (relative to the model's embedding size d) is denoted $\alpha = D/d$. Formally, FFN with gating activation (e.g., SwiGLU in LLaMA (Touvron et al., 2023)) represented as $\text{FFN}(x) = W_{\text{down}}(\sigma(W_{\text{gate}}x) \odot (W_{\text{up}}x))$, where $\odot$ represents element-wise multiplication and σ is activation function such as SiLU (Elfwing et al., 2018). The pre-activation (output of the first linear projection) and pos-activation (before the down-projection) is represented as $\text{PreAct}(X) = W_{\text{gate}}x$ and $\text{PostAct}(X) = \sigma((W_{\text{gate}}x) \odot (W_{\text{up}}x))$.

Activation sampling and co-variance matrix formation During training step t we sample a mini-batch of N tokens from each FFN layer's (ℓ) post-activation $X_{\text{post}}^{(\ell,t)} \in \mathbb{R}^{N\times D}$. We compute the covariance using all N tokens without any sub-sampling or statistical approximations to capture the true behavior of the model. Further, we compute an unbiased covariance matrix for all tokens in the batch as follows:

$$\Sigma = \frac{(X-\mu)^T(X-\mu)}{N-1} \in \mathbb{R}^{D\times D}. \quad (1)$$

For each covariance matrix, we perform eigen-decomposition to obtain the eigenvalues $\Sigma v = \lambda v$. The eigenvalues are sorted in descending order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_D \geq 0$. All subsequent metrics depend only on this spectrum.

3.2 Spectral Rank Metrics

When a feed-forward block is widened, the key question shifts from how many parameters did we add? to how many of those additional directions does the model actually use? To quantify this notion of *use*, we analyze the eigenspectrum of the post-activation covariance matrix and distill it into four metrics, each lies in the range $[0, 1]$ and can be computed in $\mathcal{O}(D)$ time (Table 1).

Hard spectral rank. Participation Ratio (PR) acts as a hard counter of dominant directions. Since PR squares the first spectral moment and divides by the second, it is particularly sensitive to prominent eigenvalues: even a single large spike can significantly cap its value, whereas numerous smaller eigenvalues have minimal impact (Gao et al., 2017; Hu and Sompolinsky, 2022). Hence, PR effectively rounds off all but the strongest axes, a *hard* spike-sensitive estimate.

Soft Spectral Rank. It complements PR by measuring the Shannon entropy of the full eigenvalue distribution (Skean et al., 2025; Wei et al., 2024; Garrido et al., 2023; De Domenico and Biamonte, 2016; Anand et al., 2011; Passerini and Severini, 2008), by converting eigenspectrum into a probability distributions as $p_i = \lambda_i / \sum_j \lambda_j$. Normalizing

Table 1: Spectral utilization metrics for characterizing the FFN latent space utilization. Hard and Soft Rank capture absolute participation and entropy-based ranks in the native $[1, D]$ scale, while their normalized forms yield bounded $[0, 1]$ utilization scores. Spectral concentration measures front-loading of variance, SUI balances hard and soft ranks, and eDim translates spectral patterns into an interpretable effective dimension.

Metric	Definition	Range	Qualitative signal	Interpretation	Cost
Hard Spectral Rank	$\text{PR} = \frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}$, $\tilde{\text{PR}} = \frac{\text{PR}-1}{D-1}$	$[0, 1]$	Spikes $\rightarrow$ collapse	Dominant spikes	$\mathcal{O}(D)^*$
Soft Spectral Rank	$\text{eR} = \exp\left(-\sum_i p_i \log p_i\right)$, $\tilde{\text{eR}} = \frac{\text{eR}-1}{D-1}$	$[0, 1]$	Long tails $\rightarrow$ dilution	Uniformity of spread	$\mathcal{O}(D)$
Spectral Concentration	$\text{SC} = \frac{2}{D} \times \sum_{k=1}^{D} \left(\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{D} \lambda_i} - \frac{k}{D}\right)$	$[0, 1]$	Strength of spikes	Front-loadedness	$\mathcal{O}(D)$
Spectral Utilization Index	$\text{SUI} = \frac{2\tilde{\text{PR}} \cdot \tilde{\text{eR}}}{\tilde{\text{PR}} + \tilde{\text{eR}}}$	$[0, 1]$	Penalizes both extremes	Balanced utilization	$\mathcal{O}(1)^\dagger$
Effective dimension	$\text{eDim} = 1 + (D-1)\text{SUI}$	$[1, D]$	# active PCs	# active dimensions	$\mathcal{O}(1)$

*Once eigenvalues are sorted; †Once ranks known

to $[0, 1]$ yields a smooth measure of dimensionality that captures long-tail variance patterns. Thus, while hard rank is sensitive to dominant peaks, soft rank responds to tail behavior. Describing the pair as hard and soft therefore captures their complementary sensitivities: former reacts sharply to collapse (variance concentrated in a few axes), whereas the latter flags spectral dilution, variance diffused so widely that no direction carries significant weight.

Spectral Utilization Index SUI combines hard and soft spectral ranks into a unified measure of spectral utilization. Hard and soft ranks independently capture opposing failure modes–spectral collapse versus dilution. To effectively combine these metrics, we adopt their harmonic mean, as it strongly penalizes imbalance: the harmonic mean sharply drops if either input is low, ensuring SUI attains high scores only when both metrics indicate balanced utilization. By rewarding spectra that avoid extremes and peak when a moderate number of principal directions carry most variance, SUI thus provides a robust, intuitive, and parameter-free indicator of overall spectral behavior.

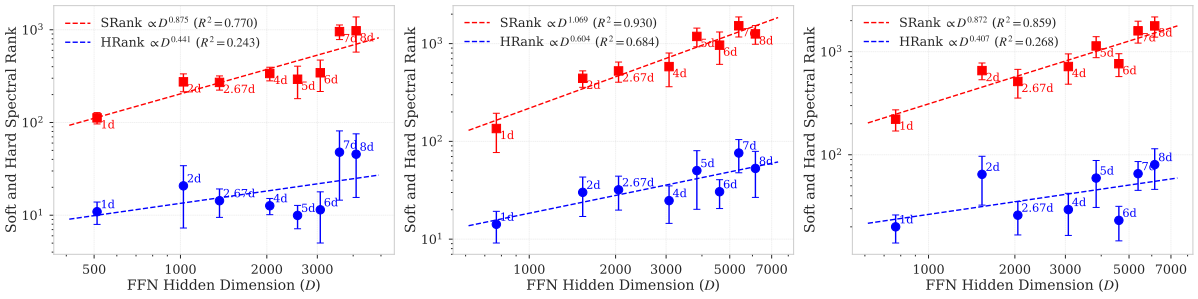
Spectral concentration. Practitioners not just about how many directions are active, but also about where the variance is concentrated. Spectral concentration measures the area between the cumulative eigen-spectrum and a uniform baseline (Marbut et al., 2023), where a higher value indicates that variance predominantly concentrates within the leading principal components, whereas lower value implies a more uniform distribution of variance across the spectrum. Thus, unlike previous metrics, it distinguishes spectra that utilize different fractions of the available latent space.

Finally, we convert SUI into an integer-valued measure called Effective Dimension (eDim), which directly represents the approximate number of active principal components. This makes interpretation more intuitive, particularly it simplifies abstract ratio into an absolute counts over abstract ratios and simplifies comparisons across layers of varying widths.

Why these specific metrics? The hard and soft ranks offer complementary perspectives on spectral utilization: one highlights spectra dominated by a few large eigenvalues, while the other captures cases with many small eigenvalues spread over a long tail. Spectral concentration metric complements these ranks by pinpointing precisely where variance accumulates. SUI unifies the two ranks into a single robust metric, penalizing both spectral extremes, and eDim further translates this into an intuitive count of active principal components. Collectively, these metrics map each layer onto an interpretable three-dimensional spectrum: collapse versus dilution, front-loaded versus dispersed variance, and overall spectral efficiency.

4 Experimental Results

In this section, we present our empirical findings on the spectral scaling laws in by varying the hidden dimension sizes of FFNs. We primarily use Hard and Soft utilization to investigate how each scales with the hidden dimension D for three sizes of LLaMA models (70M, 130M, 250M). To study how effectively FFNs leverage increasing hidden dimensions, we trained LLaMA models from scratch on C4 datasets. For each scale, we varied the hid-



(a) LLaMA-70M (PreLN) (b) LLaMA-130M (PreLN) (c) LLaMA-250M (PreLN)

Figure 2: **Asymmetric spectral scaling** with FFN width in LLaMA-style Pre-LN models. Soft rank (SRank, **red**) and hard rank (HRank, **blue**) vs. FFN hidden dimension D on log-log axes for (a) 70M, (b) 130M, and (c) 250M backbones (fixed d, width sweep $D \in \{1, 2, 2.67, 4, 5, 6, 7, 8\}$). Dashed lines are power-law fits; annotations mark αd. Soft-rank exponents cluster near unity ($\beta = \{0.873, 1.069, 0.872\}$; $R^2 = \{0.770, 0.980, 0.850\}$), while hard-rank exponents are smaller and noisier ($\beta = \{0.441, 0.604, 0.407\}$; $R^2 = \{0.248, 0.684, 0.268\}$). All networks are trained from scratch; markers show layer median values, and error bars indicate across-layer variability.

den dimension D across 8 values, D=αd, where $\alpha \in \{1, 2, 2.67, 4, 5, 6, 7, 8\}$

4.1 Asymmetric Spectral Scaling Laws

Asymmetric scaling across widths. Across all three backbones LLaMA networks (Figure 2), the soft spectral rank follows a near-linear power law with width, whereas the hard spectral rank grows sublinearly and with greater variability. Quantitatively, SRank slopes are $\beta \approx 0.88$ (70M), $\beta \approx 1.07$ (130M), and $\beta \approx 0.87$ (250M), all with strong fits ($R^2 \approx 0.77, 0.93, 0.86$). In contrast, HRank slopes are much smaller ($\beta \approx 0.44, 0.60, 0.41$) and substantially noisier ($R^2 \approx 0.24, 0.68, 0.27$). The persistent vertical separation between SRank and HRank trends spans orders of magnitude, indicating that widening FFNs consistently inflates entropy-sensitive spectral rank more than the core participation-ratio-defined subspace.

Tail-first growth. The disparity in slopes and lower R^2 values for HRank point to a *tail-first allocation of capacity*: as width D increases, models primarily populate low-energy directions (raising SRank), while the high-energy subspace expands slowly and irregularly (limited HRank gains). The 130M case comes closest to linear SRank scaling ($\beta \approx 1.07$, $R^2 \approx 0.93$), yet even here the hard-rank response remains sublinear ($\beta \approx 0.60$). This asymmetry supports the interpretation that width first buys coverage of many fine-grained, low-variance modes before it substantially grows the dominant, high-variance core.

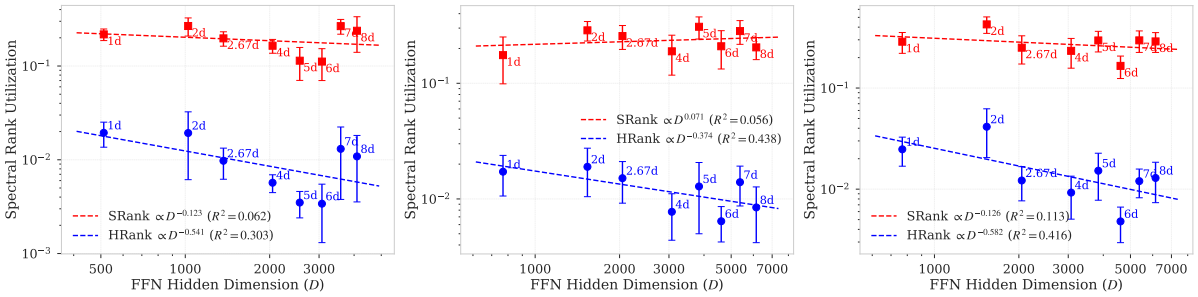
Design implications. Because widening predominantly enlarges the low-energy tail, returns on the dominant-mode subspace diminish with D. Practically, this suggests width schedules should avoid excessive tail growth, favoring tail-aware pruning (to preserve core modes and trim diffuse directions) and MoE designs that allocate experts to tail capacity rather than uniformly inflating a single dense FFN. In short, width is best understood not as a monotone "bigger is better" knob, but as a trade-off between *tail coverage* and *core strength*.

4.2 Spectral Rank Utilization

From capacity to efficiency. Normalizing ranks by D turns them into utilization fractions, $\tilde{HR}$ and $\tilde{SR}$. Across scales, $\tilde{HR}$ declines reliably with width, confirming that the high-energy mode occupies a shrinking share of dimensions as D grows (e.g., slopes around -0.5 across 70M/130M/250M). By contrast, $\tilde{SR}$ is nearly *scale-invariant* (slopes ≈ 0), showing that the low-energy tail keeps pace with widening.

Consistency with the asymmetric law. Algebraically, if SRank $\propto D^{\beta_{\text{soft}}\sim 1}$ and HRank $\propto D^{\beta_{\text{hard}}<1}$, then $\frac{\text{SRank}}{D} \propto D^{\beta_{\text{soft}}-1} \approx D^0$ and $\frac{\text{HRank}}{D} \propto D^{\beta_{\text{hard}}-1} \downarrow$, exactly matching the observed near-flat soft utilization and negative hard utilization slopes. Put simply, widening allocates capacity *tail-first*: coverage expands, but the fraction devoted to the core contracts.

Failure modes in utilization space. This view cleanly separates two regimes. *Spectral dilution* arises when $\tilde{SR}$ remains flat (or slightly increasing) while $\tilde{HR}$ falls, most visible at 130M. *Spectral collapse* appears when both utilizations decrease, pronounced at large D for 250M. These patterns are consistent across backbones and independent of absolute width, making them a compact efficiency



(a) LLaMA-70M (PreLN) (b) LLaMA-130M (PreLN) (c) LLaMA-250M (PreLN)

Figure 3: **Spectral-rank utilization vs. FFN width** in LLaMA-style Pre-LN models. We plot soft-rank utilization (SRank$/(D-1)$, **red**) and hard-rank utilization (HRank$/(D-1)$, **blue**) vs. FFN hidden dimension D on log-log axes for 70M, 130M, and 250M backbones (fixed depth; width sweep $D = \alpha d, \alpha \in \{1, 2, 2.67, 4, 5, 6, 7, 8\}$). Dashed lines show power-law fits, highlighting that SRank scales nearly linearly with width while HRank grows more slowly and with higher variability. All networks are trained from scratch; markers indicate layer median, and error bars denote across-layer variability.

Table 2: Summary of spectral-utilization metrics and fitted scaling exponents. The right-most columns list the power-law slope for HardRank and SoftRank, quantifying how sharply each metric saturates as width increases.

	D=768				D=2048				D=3072				D=4608				**Scaling Laws Parameters**			
	HRank	SRank	SUI	**eDim**	HRank	SRank	SUI	**eDim**	HRank	SRank	SUI	**eDim**	HRank	SRank	SUI	**eDim**	HRank(β,R^2)		SRank(β,R^2)	
70M	0.011	0.118	0.021	17	0.007	0.113	0.013	27	0.006	0.121	0.011	36	0.003	0.078	0.005	25	-0.72	0.864	-0.172	0.410
130M	0.016	0.182	0.029	23	0.018	0.259	0.034	71	0.009	0.182	0.017	54	0.007	0.190	0.013	60	-0.475	0.606	0.007	0.001
250M	0.030	0.272	0.054	42	0.012	0.226	0.024	49	0.012	0.232	0.022	69	0.005	0.156	0.009	44	-0.928	0.923	-0.261	0.730

diagnostic.

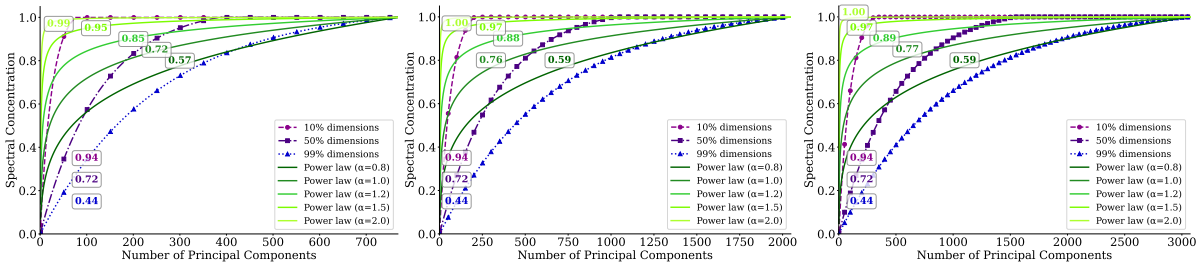
Composite diagnostics. Table 2 shows that SUI decreases monotonically for every checkpoint (e.g., 70 M: $0.021 \rightarrow 0.005$), while eDim saturates around 40–50 regardless of D. Because SUI penalizes a drop in either rank, its steady decline confirms that *no part of the spectrum scales proportionally with width*.

Implications for model design. Our findings suggest three key principles for efficient model design: (1) *Stop widening early*–for Pre-LN LLaMA, increasing D beyond $\sim$3,000 yields diminishing spectral returns; (2) *Monitor SUI during training*–it offers a one-line diagnostic that flags wasted parameters before full convergence; and (3) *Layer-wise adaptation beats uniform scaling*–the heterogeneous behavior across checkpoints suggests allocating width dynamically, pruning collapsing layers and selectively widening those still far from dilution. By grounding width decisions in spectral utilization rather than parameter counts, practitioners can trim model size without sacrificing representational power, a crucial step towards efficient-inference at scale.

4.3 Scaling Laws for Spectral Concentration

We investigate the spectral concentration of FFNs activation covariance matrices by modeling their eigenvalue distribution via a truncated power-law: $\lambda_k \propto k^{-\alpha}, \quad k = 1, \ldots, D$, where the exponent α controls how variance is distributed across eigen-directions. While traditional rank-based metrics (e.g., Hard and Soft Spectral Ranks) integrate information from *all* eigenvalues, they often overlook crucial details in the distribution's shape, such as distinguishing between sharply peaked spectra with extensive flat tails and those smoothly decaying. The proposed power-law scaling framework directly addresses this limitation, isolating the shape characteristics of spectral distributions. Higher values of α yield spectra sharply concentrated (front-loaded) among leading directions, indicating incipient collapse, whereas lower values produce more uniform (diluted) distributions, indicative of suboptimal variance allocation (Fig. 4).

Empirically, several robust trends emerge from our analysis. Spectral concentration, monotonically increases with α: as α rises from 0.8 to 2.0, it grows consistently from around 0.57 to nearly 0.99 (Table 3). Once eigenvalues decay faster than k^{-2}, variance is predominantly concentrated in the initial directions, becoming effectively dimension-invariant and independent of model width. This invariance enables meaningful comparisons of FFN efficiency across models of different sizes by aligning them on a common spectral utilization axis.



(a) $D = 768$ (b) $D = 2048$ (c) $D = 3072$

Figure 4: Power-law templates for spectral concentration. Cumulative-variance curves generated from synthetic power-law spectra $\lambda_k \propto k^{-\alpha}$ for three latent sizes ($D = 768, 2048, 3072$). Larger exponents (α) front-load variance and push the curve upward. Coloured call-outs report the concentration value reached by benchmark cut-offs.

Table 3: Quantitative summary of the curves in Fig 4. For each α and hidden size D we list the variance carried by the top-1 eigenvalue, and cumulative variance captured by the first 10%, 25% and 50% principal components, along with the concentration score. The results show sharp transition around $\alpha \approx 1.2$: below it at least half the spectrum is needed to explain 80% of the variance (dilution), above it fewer than 10% directions suffice (collapse).

α	Top-1 eigenvalue			Variance @ 10% dimensions			Variance @ 25% dimensions			Variance @ 50% dimensions			Spectral Concentration		
	768	2048	3072	768	2048	3072	768	2048	3072	768	2048	3072	768	2048	3072
0.8	6.9%	5.4%	4.9%	51.9%	54.3%	55.2%	68.4%	70.0%	70.5%	83.1%	84.0%	84.3%	0.57	0.59	0.59
1.0	13.8%	12.2%	11.6%	68.2%	72.0%	73.3%	80.8%	83.1%	83.9%	90.4%	91.6%	91.9%	0.72	0.76	0.77
1.2	23.4%	22.2%	21.8%	81.9%	85.9%	87.2%	90.1%	92.3%	93.0%	95.4%	96.4%	96.7%	0.85	0.88	0.89
1.5	39.4%	38.9%	38.8%	93.9%	96.3%	97.0%	97.2%	98.3%	98.6%	98.8%	99.3%	99.4%	0.95	0.97	0.97
2.0	60.8%	60.8%	60.8%	99.3%	99.7%	99.8%	99.8%	99.9%	99.9%	99.9%	100.0%	100.0%	0.99	1.00	1.00

For larger $\alpha \geq 1.5$, over 90% of variance resides within merely the top 10% of principal components (Table 3). Conversely, at smaller values ($\alpha \approx 0.8$), capturing the same variance requires more than 50% of components, leading to a state we term "spectral dilution." Notably, activations in prevalent models such as LLaMA typically exhibit intermediate spectral concentration ($\alpha \approx 1.1$–1.3), thereby balancing effective dimensionality and representational compactness, avoiding the extremes of either spectral dilution or collapse.

4.4 Spectral Scaling Dynamics

As shown in Figure 5, during the first 2K to 3K training steps the spectral landscape is still fluid: both Hard- and Soft-Rank curves rise steeply and the fitted β coefficients fluctuate, accompanied by low R^2. This early volatility warns against drawing scaling-law conclusions from partially trained checkpoints. Around step 5K the exponents settle and R^2 surpasses 0.6, suggesting that a stable power-law relation has emerged. Averaged over the final 1K steps we obtain $\beta_{\text{hard}} \approx -0.38$ and $\beta_{\text{soft}} \approx +0.07$.

A further observation is that the rank trajectories in panels (c) and (d) preserve their vertical ordering throughout training: wider configurations always sit above narrower ones for Soft-Rank and below for Hard-Rank. Hence the eventual utilization hierarchy is determined surprisingly early, suggesting that practitioners can estimate the utility of a width choice long before full convergence.

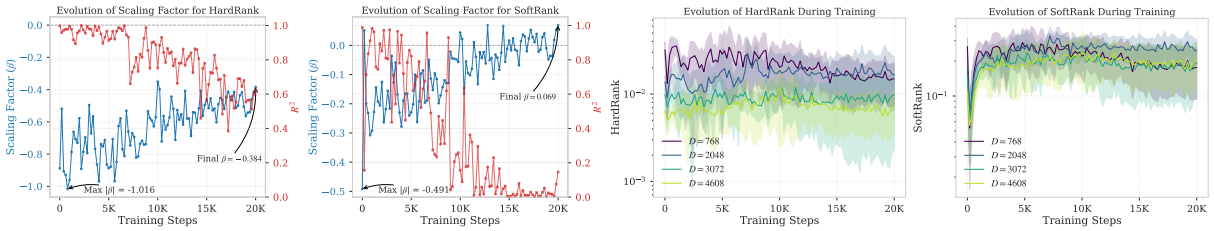
5 Case Study for Spectral Rank Scaling and Utilization

5.1 LayerNorm and Spectral Rank

Pre-LN shows the classic asymmetry. With Pre-LN, soft-rank scales close to linearly with width ($\beta \approx 0.88$ at 70M; $\beta \approx 1.07$ at 130M, high R^2), while hard-rank is clearly sublinear ($\beta \approx 0.45/0.60$, lower R^2). This is the baseline *tail-first growth*: widening expands low-energy directions, while the high-energy core lags behind (Table 4).

Post-LN suppresses tail growth. Shifting LayerNorm after the sub-blocks lowers soft-rank slopes to $\sim 0.71 - 0.82$ with stronger R^2, effectively *dampening tail inflation*. Hard-rank slopes rise modestly to $\sim 0.52 - 0.56$ with better R^2, suggesting more orderly,but still sublinear,growth of the dominant subspace. Intuitively, normalizing after each transformation curbs variance spread, limiting activation of faint directions as width increases.

Mix-LN balances core and tail. Mix-LN restores near-linear soft-rank scaling ($\beta \approx 0.97 -$



(a) Hard Rank β (b) Soft Rank β (c) Hard Rank β (d) Soft Rank β

Figure 5: Training-time evolution of spectral scaling laws (Rank utilization) for LLaMA-130M (Pre-LN). (a) and (b) track, at every logged step, the power-law exponent β (blue, left axis) obtained by regressing $\log(\text{Hard/Soft Rank})$ against $\log D_{\text{FFN}}$ across the width multipliers; the red curve (right axis) is the corresponding coefficient of determination R^2. (c) and (d) show the raw layer-averaged Hard- and Soft-Rank trajectories for each width to illustrate the data being fit. Shaded bands are ± 1 s.d. over layers.

Table 4: Spectral rank scaling across normalization schemes

	PreLN		PostLN		MixLN	
Model	**Hard Rank**	**Soft Rank**	**Hard Rank**	**Soft Rank**	**Hard Rank**	**Soft Rank**
LLaMA-70M	0.451 ± 0.778 ($R^2 = 0.251$)	0.879 ± 0.490 ($R^2 = 0.763$)	0.556 ± 0.358 ($R^2 = 0.706$)	0.712 ± 0.273 ($R^2 = 0.872$)	0.593 ± 0.668 ($R^2 = 0.440$)	0.972 ± 0.477 ($R^2 = 0.805$)
LLaMA-130M	0.604 ± 0.411 ($R^2 = 0.684$)	1.069 ± 0.292 ($R^2 = 0.930$)	0.521 ± 0.294 ($R^2 = 0.758$)	0.818 ± 0.372 ($R^2 = 0.829$)	0.626 ± 0.484 ($R^2 = 0.626$)	1.096 ± 0.484 ($R^2 = 0.837$)

1.10, high R^2) while maintaining hard-rank growth above Pre-LN/Post-LN levels ($\beta \approx 0.59 - 0.63$, moderate R^2). In effect, it *preserves tail coverage* while also improving dominant-mode scaling, avoiding both the over-tailing of Pre-LN and the excessive tail suppression of Post-LN.

5.2 LLaMA-250M PostLN

Spectral collapse in Post-LayerNorm blocks. We observe a strong correlation between spectral health and the performance of LLaMA-250M when the FFN width is increased. In the vanilla Post-LayerNorm setup, spectral dynamics remain stable only for the narrowest FFN width (1d). However, scaling the width to 2.67d or 4d leads to a rapid collapse of spectral diversity: the hard-rank plunges to $\lesssim 10^{-3}$ and the concentration saturates to ≈ 1.0 within the first few thousand steps (Figure 6a). This spectral collapse signifies that most of the variance is funneled into one or two dominant directions, leaving the majority of the ~ 3000 latent dimensions inactive. As a result, model performance deteriorates sharply, with test perplexity exceeding consistent with the figures reported in Table 5.

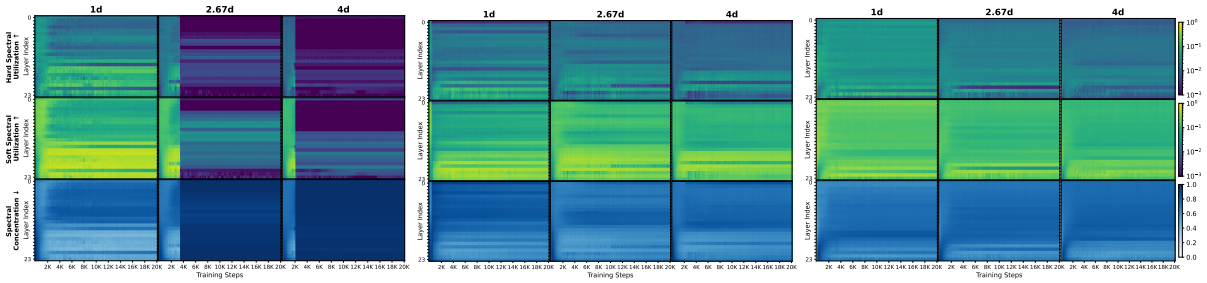
Weight Normalization enables high-rank spectra and best perplexity. Employing weight normalization (WNorm) (Salimans and Kingma, 2016) within each FFN significantly mitigates this collapse. The hard-rank stabilizes in the 10^{-2}–10^{-1} range, while spectral concentration settles around 0.25–0.3, indicating that hundreds of latent directions carry meaningful variance. This richer and more distributed latent basis translates into notably better performance: perplexities of 25.1 (at 2.67d) and 24.3 (at 4d), both outperforming the vanilla 1d baseline (27.1). These results affirm that maintaining a non-degenerate spectrum not only prevents collapse but actively enhances downstream predictive performance.

Table 5: Vanilla PostLN in LLaMa-250M becomes unstable at higher FFN dimensions, causing spikes in PPL values. Adding Weight Normalization or Hyperspherical Normalization to the FFN linear layers stabilizes training (former outperforms the latter across all scales).

PostLN	$1d$	$2.67d$	$4d$
Vanilla	27.10	1427.91	1431.01
WeightNorm	28.89	25.08	24.27
HypersphericalNorm	31.66	27.92	26.48

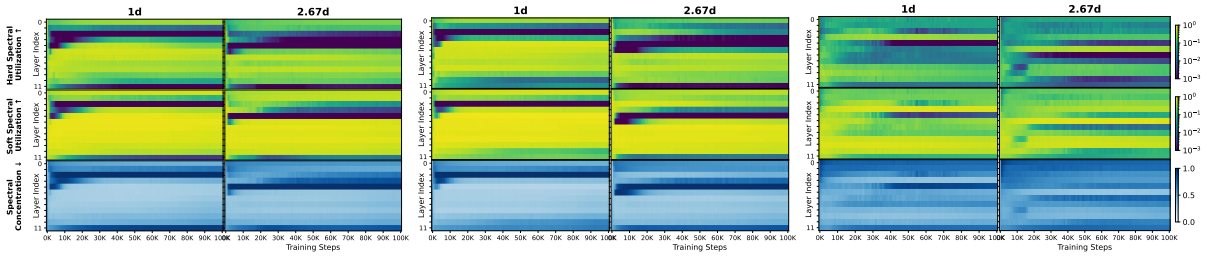
5.3 Hyperspherical Normalization

Hyperspherical normalization (HNorm) also prevents collapse and promotes training stability but results in more conservative spectral utilization (Loshchilov et al., 2025; Lee et al., 2025; Karras et al., 2024; Wang and Isola, 2020; Liu et al., 2017).



(a) LLaMA-250M (PostLN) (b) LLaMA-250M (PostLN) +WNorm (c) LLaMA-250M (PostLN) + HNorm

Figure 6: LLaMA models



(a) GPT-2 (GeLU) (b) GPT-2 (SiLU) (c) nGPT-2 (SiLU)

Figure 7: GPT-2 vs nGPT

The hard-rank remains roughly an order of magnitude above the collapse threshold, yet ∼30% lower than the WNorm trace. Spectral concentration is marginally higher, suggesting a somewhat narrower effective basis. Consequently, while HNorm yields stable performance (27.9 at 2.67d and 26.5 at 4d), it does not match the perplexity gains achieved with WNorm. These findings highlight that collapse prevention is a necessary condition, but further lifting the rank and ensuring richer variance distribution is critical for unlocking full potential of wider FFNs.

Activation gating and normalization in GPT2. Figure 7 tracks the spectral evolution, and Table 6 shows perplexity outcomes of GPT-2 variants using different activation and normalization schemes under two FFN widths (1d and 2.67d). The baseline GPT-2 with GeLU shows early hard-rank growth that quickly saturates around 10^{-2}, while spectral concentration remains high (≈ 0.7). This indicates a narrow set of dominant directions and leads to moderate perplexity (14.07 at 2.67d), with limited gain over the 1d baseline (15.63).

The nGPT configuration augments SwiGLU with hyperspherical weight and activation normalization and a learnable residual eigen-learning rate (eigen-LR) (Loshchilov et al., 2025). This combination substantially enhances spectral health: hard-rank remains two orders of magnitude above collapse, soft-rank saturates earlier with less fluctuation, and concentration reduces to ≈ 0.4–a 20% improvement over GPT-2. These gains are mirrored in performance, with perplexity dropping to 13.60 at 2.67d and stabilising to 15.01 at 1d, outperforming both prior setups.

Table 6: Perplexity (PPL) comparison of GPT-2 and nGPT (Loshchilov et al., 2025) with different activation functions and FFN dimensions.

	GPT-2(GeGLU)		GPT-2(SwiGLU)		nGPT(SwiGLU)	
	1d	2.67d	1d	2.67d	1d	2.67d
PPL	15.63	14.07	15.60	14.05	15.01	13.60

6 Conclusion

We reframed FFN width selection as a spectral utilization problem, showing that widening follows a consistent tail-first pattern: soft-rank utilization remains near-linear while hard-rank utilization declines. This asymmetry, formalized as spectral scaling laws, reveals two efficiency failures, spectral dilution and spectral collapse, that limit naïve width growth. LayerNorm placement modulates these dynamics: Pre-LN amplifies tails, Post-LN suppresses them, and Mix-LN balances both. Together, these results highlight spectral utilization as a new efficiency axis, motivating width-efficient designs via layer-wise scheduling and pruning.

Limitations

The study is limited to English decoder-only models up to 250M parameters and does not validate spectral behavior in multilingual or encoder-decoder settings. While spectral metrics correlate with perplexity, causality remains unproven, and finer-grained subspace analysis may be needed beyond scalar metrics like SUI. Additionally, eigencomputations could pose challenges at extreme scales.

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2025. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Kartik Anand, Ginestra Bianconi, and Simone Severini. 2011. Shannon and von neumann entropy of random networks with heterogeneous expected degree. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. 2024. Scaling laws for associative memories. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, and 1 others. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Leshem Choshen, Yang Zhang, and Jacob Andreas. 2024. A hitchhiker's guide to scaling law estimation. *arXiv preprint arXiv:2410.11840*.

Manlio De Domenico and Jacob Biamonte. 2016. Spectral entropies as information-theoretic tools for complex network comparison. *Physical Review X*.

Gbètondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. 2024. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. In *Journal of Machine Learning Research (JMLR)*.

Stanislav Fort. 2025. Scaling laws for adversarial attacks on language model activations and tokens. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

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

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

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

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

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. 2024. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tanishq Kumar, Zachary Ankner, Benjamin Frederick Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Re, and Aditi Raghunathan. 2025. Scaling laws for precision. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Hojoon Lee, Youngdo Lee, Takuma Seno, Donghu Kim, Peter Stone, and Jaegul Choo. 2025. Hyperspherical normalization for scalable deep reinforcement learning. In *International conference on machine learning (ICML)*.

Pengxiang Li, Lu Yin, and Shiwei Liu. 2025. Mix-LN: Unleashing the power of deeper layers by combining pre-LN and post-LN. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Weiyang Liu, Yan-Ming Zhang, Xingguo Li, Zhiding Yu, Bo Dai, Tuo Zhao, and Le Song. 2017. Deep hyperspherical learning. In *Advances in neural information processing systems*.

Ilya Loshchilov, Cheng-Ping Hsieh, Simeng Sun, and Boris Ginsburg. 2025. nGPT: Normalized transformer with representation learning on the hypersphere. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Bochen Lyu, Di Wang, and Zhanxing Zhu. 2025. A solvable attention for neural scaling laws. In *The Thirteenth International Conference on Learning Representations*.

Anna Marbut, Katy McKinney-Bock, and Travis Wheeler. 2023. Reliable measures of spread in high dimensional latent spaces. In *International Conference on Machine Learning (ICML)*.

Charles H Martin and Michael W Mahoney. 2021. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 2024. 4+3 phases of compute-optimal neural scaling laws. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Filippo Passerini and Simone Severini. 2008. The von neumann entropy of networks. *arXiv preprint arXiv:0812.2597*.

Telmo Pessoa Pires, António V Lopes, Yannick Assogba, and Hendra Setiawan. 2023. One wide feedforward is all you need. In *Proceedings of the Eighth Conference on Machine Translation*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. 2019. On the spectral bias of neural networks. In *International conference on machine learning (ICML)*.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of langauge model performance. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Tim Salimans and Durk P Kingma. 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in neural information processing systems*.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. 2024. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *International Conference on Machine Learning (ICML)*.

Jingzhe Shi, Qinwei Ma, Huan Ma, and Lei Li. 2024. Scaling law for time series forecasting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *International conference on machine learning (ICML)*.

Max Staats, Matthias Thamm, and Bernd Rosenow. 2024. Locating information in large language models via random matrix theory. *arXiv preprint arXiv:2410.17770*.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. 2024. Scaling laws with vocabulary: Larger models deserve larger vocabularies. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q Tran, Dani Yogatama, and Donald Metzler. 2022. Scaling laws vs model architectures: How does inductive bias influence scaling?

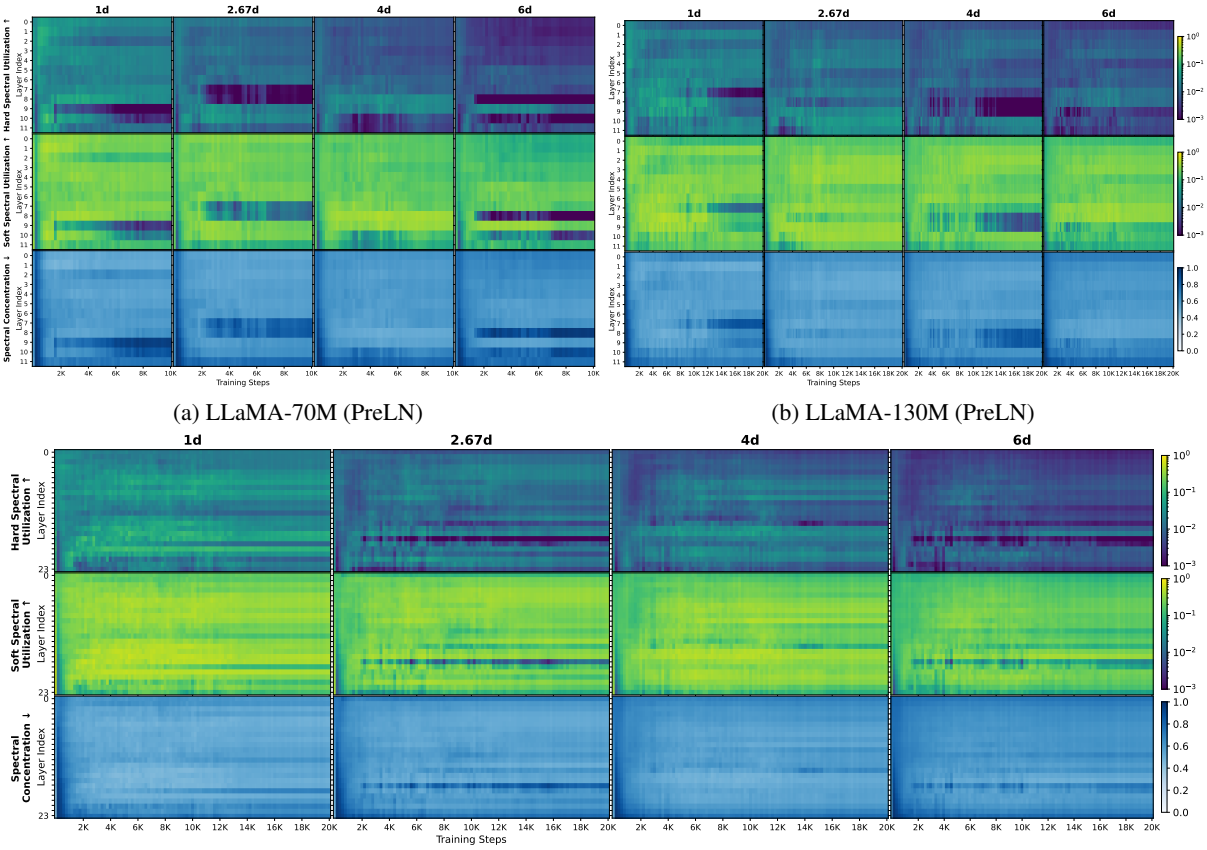
Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning (ICML)*.

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

Table 7: Evaluation perplexity (PPL) for LLaMA models across different normalization positioning and FFN dimensions. The columns $1d$, $2.67d$, $4d$, and $6d$ represent different FFN width, where d is the model dimension. The unusually high PPL in PostLN LLaMA-250M indicate training instability.

Model	**PreLN**				**PostLN**				**MixLN**			
	$1d$	$2.67d$	$4d$	$6d$	$1d$	$2.67d$	$4d$	$6d$	$1d$	$2.67d$	$4d$	$6d$
LLaMA-70M	38.6	34.2	32.4	31.1	38.2	33.6	32.3	31.1	38.7	33.9	32.0	30.7
LLaMA-130M	29.6	26.4	25.8	24.6	29.2	26.7	25.8	25.1	29.2	26.8	25.3	24.3
LLaMA-250M	26.7	24.5	23.3	22.5	27.1	**1427.9**	**1431.0**	**1436.7**	26.8	24.2	23.0	22.5



(a) LLaMA-70M (PreLN)

(b) LLaMA-130M (PreLN)

(c) LLaMA-250M (PreLN)

Figure 8: LLaMA models