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

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) continue to scale, understanding how effectively their internal capacity is utilized becomes increasingly important, especially for inference-time efficiency. While existing scaling laws relate model size to loss and compute, they offer little insight into the representational dynamics of individual components. In this work, we focus on the Feed-Forward Network (FFN), a dominant sub-block in decoder-only transformers, and recast FFN width selection as a *spectral utilization* problem. We introduce a lightweight, differentiable diagnostic suite comprising: **Hard Rank** (Participation Ratio), Soft Rank (spectral entropy), Spectral Concentration, and the composite Spectral Utilization Index (SUI), designed to quantify how many latent directions are meaningfully activated. Our spectral audit across GPT-2, LLaMA, and nGPT models reveals that spectral utilization grows with model size but not monotonically with width, often peaking at intermediate dimensions (e.g., $D = 2048$). We identify clear instances of *spectral collapse*, where wider FFNs concentrate variance into a narrow subspace, leaving much of the latent space unused.

1 Introduction

As Large Language Models (LLMs) continue to grow in scale and complexity, a fundamental question arises: *How effectively is their internal capacity utilized?* Despite the availability of empirical scaling laws that relate model performance to width, depth, and data size, these laws provide little visibility into how efficiently different components operate. They abstract away the internal dynamics of transformer blocks, leaving open critical questions about representational usage.

Among these blocks, the Feed-Forward Network (FFN) subcomponents have received limited analytical attention. Though they constitute a substantial fraction of parameters in decoder-only architectures, their contribution to internal efficiency remains poorly understood. FFNs play a key role in maintaining feature diversity, preserving token isotropy, and enabling signal propagation. Yet width configurations—such as the widely used 4× expansion in GPT or 2.67× in LLaMA—are often adopted as static heuristics, rather than grounded design choices.

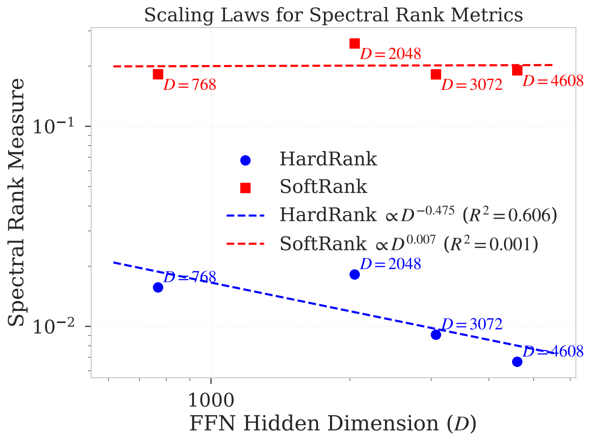


Figure 1: Scaling Laws for LLaMA-130M models

This raise the critical 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. Through a careful analysis of FFN-layer activations, we identify two distinct failure modes: *spectral collapse*, where increased width compresses representational variance into a narrow subspace, and *spectral dilution*, where variance is dispersed thinly across many low-signal dimensions. These behaviors indicate internal inefficiency, regardless of model size or perplexity performance.

To study these dynamics, we conduct a comprehensive spectral audit of FFNs in transformer models spanning GPT-2, LLaMA, and nGPT. Our evaluation covers width multipliers $r \in \{1, 2.67, 4, 6, 8\}$

and model sizes from 70M to 250M parameters. For each FFN layer, we compute the full eigenspectrum of its covariance matrix (post-activation) across training steps, layers.

We quantify spectral utilization using four interpretable and computationally efficient metrics: *Hard Spectral Rank:* based on the Participation Ratio (PR), measuring the number of dominant directions activated; *Soft Spectral Rank:* information-theoretic based rank measure, capturing how variance is distributed across the spectrum; *Spectral Concentration:* via Eigenvalue Early Enrichment, assessing how much variance is captured by leading eigenvalues; and finally *Spectral Utilization Index (SUI):* a composite metric that harmonically combines hard and soft rank to balance localized and distributed representations.

Our empirical findings reveal three key insights. First, *spectral utilization increases with model size*, confirming that larger models activate more dimensions and better exploit their representational capacity. Second, *increasing FFN width does not always improve utilization.* In fact, we observe a clear peak at intermediate widths (e.g., $D = 2048$), beyond which metrics such as SUI and effective dimension (eDim) stagnate or decline—indicating spectral dilution. Third, we uncover *sublinear spectral scaling laws* linking FFN width to effective rank, with best-fit exponents $\beta < 1$ for both hard and soft rank metrics. These patterns suggest that widening FFNs beyond a certain point results in diminishing returns, and that internal dimensionality can often be trimmed without performance loss.

Contributions. This work makes four key contributions: **Conceptual.** We reframe FFN width selection as a spectral utilization problem and formalize two failure modes—*spectral collapse* and *spectral dilution*—to diagnose under- and over-capacity conditions in transformer blocks. **Architectural.** We conduct the first layer-wise spectral audit across GPT-2, LLaMA, and nGPT variants, analyzing FFNs under varying width multipliers and normalization configurations. Our results reveal strong layer-wise heterogeneity and challenge the use of fixed-width heuristics. **Algorithmic.** We introduce a lightweight, differentiable diagnostic suite—*Participation Ratio*, *Spectral Entropy*, *Eigenvalue Early Enrichment*, and the *Spectral Utilization Index*—for quantifying representational usage. We also provide a closed-form estimator of effective dimension: $K_{\text{eff}} = 1 + (D - 1) \cdot \text{SUI}$

Empirical. Across 70M, 130M, and 250M parameter models, we establish sub-linear power laws relating FFN width to spectral ranks, quantify how LayerNorm placement modulates utilization dynamics.

2 Related Work

Cost-aware neural scaling. Early work (Kaplan et al., 2020) formalised the now-canonical loss-vs-compute power law for language models, later refined by Chinchilla scaling laws (Hoffmann et al., 2022), who showed that existing models are compute-sub-optimal because they are too large and under-trained. Sardana et al. (2024) extend this frontier to the deployment setting: for traffic ∼10B requests the compute-optimal point shifts to smaller models trained on more tokens, lowering inference cost. (Paquette et al., 2024) derive a four-plus-three-phase diagram that predicts which factor—capacity, optimizer noise, or feature embedding—dominates under a fixed budget Orthogonal cost axes matter too: (Tao et al., 2024) reveal that vocabulary size must grow with width, while Kumar et al. (Kumar et al., 2025) introduce precision-aware laws that treat lower-precision training as shrinking the effective parameter count. (Choshen et al., 2024) offer statistically robust procedures for fitting such laws from small pilot runs. Collectively, these studies motivate our spectral-utilization laws as a *complementary efficiency axis*, tracking latent-space usage rather than surface-level FLOPs.

Universality and representational capacity. (Ruan et al., 2024) show that once efficiency offsets are normalized, ∼100 heterogeneous checkpoints, from GPT-2 to PaLM, collapse onto a single sigmoidal curve. The Physics of LMs series reaches a similar conclusion for factual knowledge, finding a near-constant ~ 2 bits/parameter ceiling across architectures (Allen-Zhu and Li, 2025). Martin et al. (Martin and Mahoney, 2021) first linked such universality to heavy-tailed eigenspectrum and implicit self-regularization; (Staats et al., 2024) refine this by showing that small singular values encode critical information in pretrained transformers, and (Dovonon et al., 2024) connect spectrum collapse to transformer over-smoothing.

Architectural and domain-specific scaling Scaling exponents are not architecture-agnostic: () find that the best inductive bias flips with scale—Switch-Transformers rule at small N, Performers at mid-scale, vanilla attention later. (Cabannes et al., 2024) derive precise laws for associative-memory matrices, while (Shi et al., 2024) explain why larger models sometimes underperform on time-series by introducing a look-back-aware law. Fort (Fort, 2025) frames robustness itself as a scaling phenomenon, showing adversarial attack resistance stays nearly constant across two orders of magnitude in model size. Finally, Li et al. (Lyu et al., 2025) present an analytically solvable attention that yields closed-form power laws, offering a theoretical baseline.

These threads underscore that scaling is multifaceted, 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 $r = D/d$. Formally, FFN with gating activation (e.g., SwiGLU in LLaMA (Touvron et al., 2023)) represented as $\mathrm{FFN}(x) = W_{\text{down}}(\sigma(W_{\text{gate}}x) \odot (W_{\text{up}}x))$, 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 $\mathrm{PreAct}(X) = W_{\text{gate}}x$ and $\mathrm{PostAct}(X) = \sigma((W_{\text{gate}}x) \odot (W_{\text{up}}x))$.

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 $X^{(\ell,t)}_{\text{post}} \in \mathbb{R}^{N\times D}$. 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 eigendecomposition 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 Utilization 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 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

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

*Once eigenvalues are sorted; †Once ranks known

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.

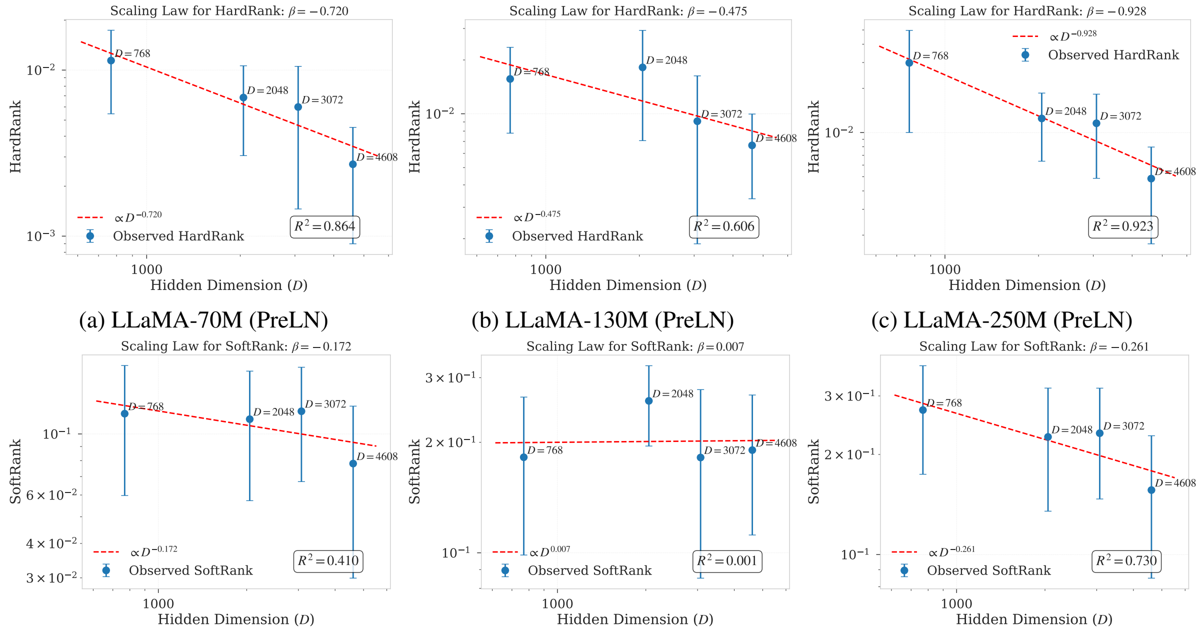
4 Experimental Results

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 hidden dimension D across four values: 768, 2048, 3072, and 4608.

4.1 Spectral Rank Scaling Laws

HardRank obeys a steep negative power law. Figure 2 (a–c) shows a clear power-law decay of HardRank with width, with exponents $\beta_{\mathrm{HR}} \in [-0.47, -0.93]$ and $R^2 = 0.61$–0.92. Dominant directions grow sub-linearly; beyond $D \approx 3\mathrm{k}$ the model gains < 0.01 additional high-variance axes per $1\,000$ new parameters. This early saturation signals that *over-provisioning* starts well before the $4\,608$-dimensional setting commonly used in practice.

SoftRank reveals heterogeneous tail behavior. The bottom row of Figure 2 shows a more nu-



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

(d) LLaMA-70M (PreLN) (e) LLaMA-130M (PreLN) (f) LLaMA-250M (PreLN)

Figure 2: Empirical scaling laws for the HardRank (top row) and SoftRank (bottom row) across different FFN hidden dimensions D in LLaMA variants. Each subplot shows the observed rank utilization values (blue markers with error bars measured across layers) alongside a fitted power-law trend (red dashed line). The negative exponents (β) and high R^2 values support a "spectral collapse" pattern, whereby effective rank declines as D increases.

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.

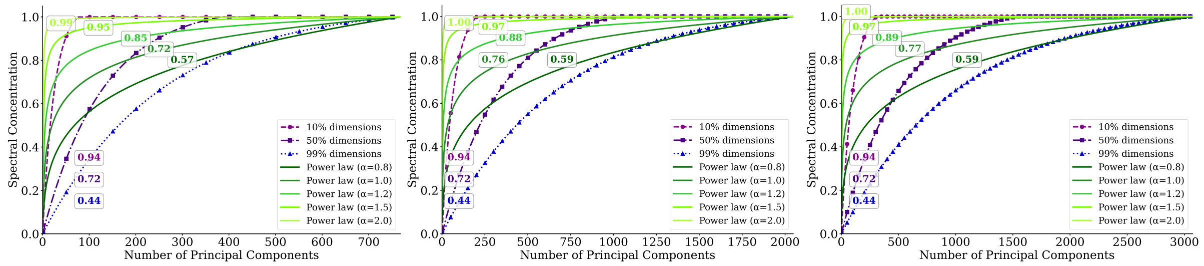
	D=768				D=2048				D=3072				D=4608				**Scaling Laws Parameters**			
	HRank	SRank	SUI	**eDim**	HRank	SRank	SUI	**eDim**	HRank	SRank	SUI	**eDim**	HRank	SRank	SUI	**eDim**	HRank(β,R^2)		SRank(β,R^2)	
70M	0.011	0.118	0.021	17	0.007	0.113	0.013	27	0.006	0.121	0.011	36	0.003	0.078	0.005	25	-0.72	0.864	-0.172	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

anced trend: SoftRank exhibits strikingly different scaling patterns across model sizes. For the 70M model, SoftRank decays mildly ($\beta_{\text{SR}} = -0.17$), indicating that the spectrum's tail is progressively under-utilized. In contrast, the 130M variant shows a statistically zero slope ($\beta_{\text{SR}} \approx 0$), where variance is merely *diluted* across many faint modes while the tail remains populated. The largest 250M model demonstrates a different pattern, with both Hard- and Soft-Rank declining ($\beta_{\text{SR}} = -0.26$), signaling true *spectral collapse* in which neither head nor tail keeps pace with width.

Two failure modes: spectral dilution vs. collapse Comparing the trajectories of Hard- and Soft-Rank allows us to disambiguate under-capacity phenomena: *Spectral dilution* occurs when HardRank falls but SoftRank stays flat (LLaMA-130 M). Widening introduces numerous *weak* directions without increasing dominant variance, spreading information thinly. Conversely, *spectral collapse* manifests when both ranks fall (LLaMA-250 M). Even the low-energy subspace is left empty, indicating genuine over-capacity. These modes are invisible to either metric alone; the joint view is essential.

Composite diagnostics. Table 2 shows that SUI decreases monotonically for every checkpoint (e.g., 70 M: 0.021 → 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 ∼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 hetero-



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

Figure 3: 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 3. 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

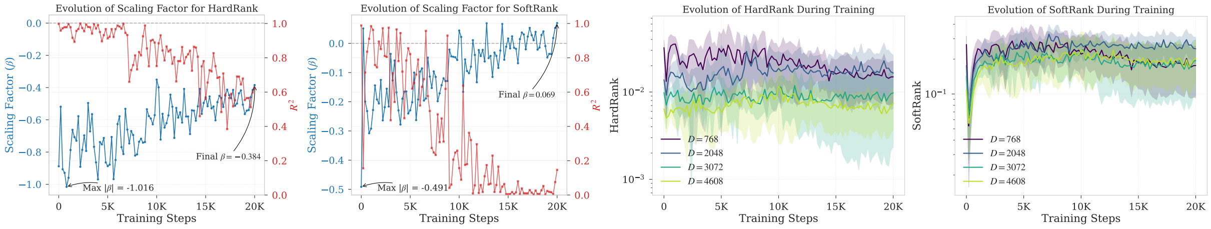
geneous behavior across checkpoints suggests allocating width dynamically, pruning collapsing layers and selectively widening those still far from dilution. By grounding width decisions in spectral utilization rather than parameter counts, practitioners can trim model size without sacrificing representational power, a crucial step towards efficient-inference at scale.

4.2 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 eigendirections. 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. 3).

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.

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.



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

Figure 4: Training-time evolution of spectral scaling laws for LLaMA-130M (Pre-LN). (a) and (b) track, at every logged step, the power-law exponent β (blue, left axis) obtained by regressing $\log(\text{Hard}/\text{Soft Rank})$ against $\log D_{\text{FFN}}$ across the four width multipliers $\{1\times, 2.67\times, 4\times, 6\times\}$; 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.

4.3 Spectral Scaling Dynamics

As shown in Figure 4, 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 Utilization

Training Stability in PostLN LLaMA-250M

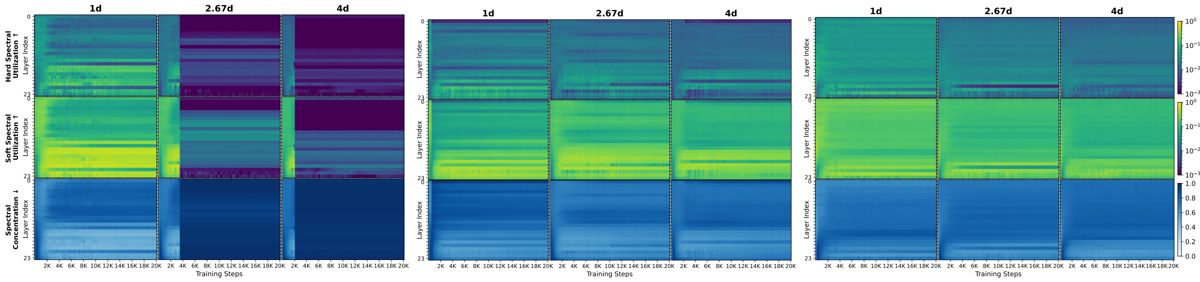
Spectral collapse in Post-LayerNorm blocks. We observe a strong correlation between spectral health and the performance of LLaMA-250M when the FFN width is increased. In the vanilla Post-LayerNorm setup, spectral dynamics remain stable only for the narrowest FFN width (1d). However, scaling the width 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 5a). This spectral collapse signifies that most of the variance is funneled into one or two dominant directions, leaving the majority of the $\sim$ 3000 latent dimensions inactive. As a result, model performance deteriorates sharply, with test perplexity exceeding consistent with the figures reported in Table 4.

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 4: 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).

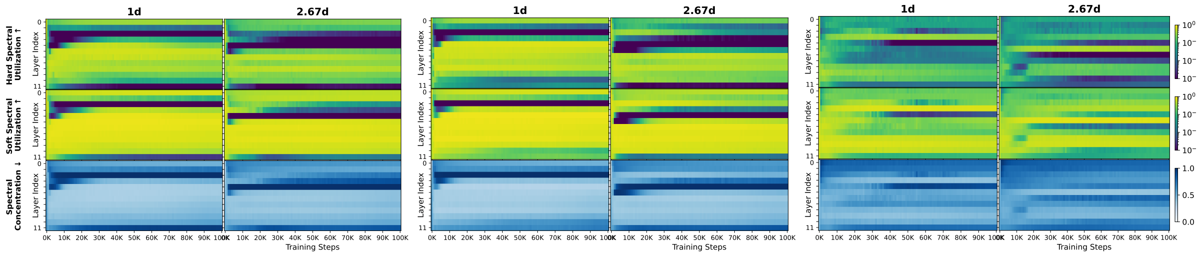
PostLN	$1d$	$2.67d$	$4d$
Vanilla	27.10	1427.91	1431.01
WeightNorm	28.89	25.08	24.27
HypersphericalNorm	31.66	27.92	26.48

Hyperspherical normalization provides stability but with conservative rank. 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). The hard-rank remains roughly an order of magnitude above the collapse threshold, yet $\sim$30% lower than the WNorm trace. Spectral concentration is marginally higher,



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

Figure 5: LLaMA models



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

Figure 6: GPT-2 vs nGPT

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 the full potential of wider FFNs.

Activation gating and normalization in GPT2. Figure 6 tracks the spectral evolution, and Table 5 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 5: 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 Limitations

This work establishes spectral utilization as a reliable proxy for FFN width selection, showing that effective rank stabilizes early and peaks around 2.5–3d. Normalization prevents collapse, and spectral metrics consistently predict perplexity, offering insights for efficient LLM design .

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, eigen-computations 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)*.

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*.

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)*.

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*.

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. *arXiv preprint arXiv:2502.15280*.

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*.

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)*.

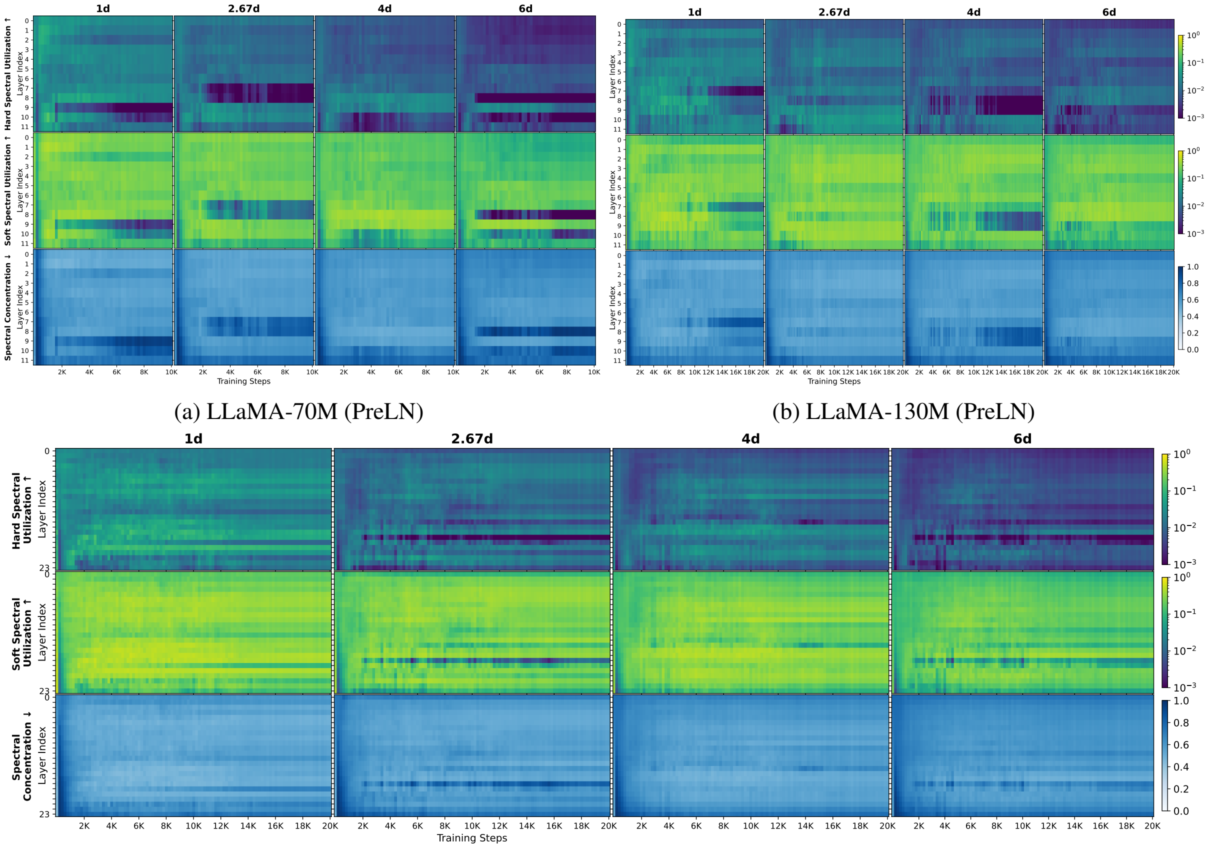
Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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 6: 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 7: LLaMA models