Spectral Scaling Laws in Language Models

How Effectively Do Feed-Forward Networks Use Their Latent Space?

Nandan Kumar Jha & Brandon Reagen (New York University)

ICML International Conference On Machine Learning
AIW@ICML'25

Motivation: Why FFN Width Matters for LLMs?

1. Capacity and Computational Efficiency:
   - Feedforward Networks (FFNs) contain **60-70%** of total parameters
   - FFN FLOPs dominate in **<4K** context length regime
   - FFNs store factual information and directly affect model's capacity
2. Architectural Diversity:
   - Different models use different FFN width (GPT-2 & Pythia: **4×**, LLaMA: **2.67×**)
3. Interpretability Gap:
   - Existing (Loss - parameter) scaling laws ignore how FFN width is utilized

We need principled tools to analyze FFN capacity allocation and scaling efficiency

Our Approach: Spectral Rank Measures

Hard Spectral Rank
(Participation Ratio)

$$\frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}$$

How many dimensions are active in the eigenspectrum, relative to total variance

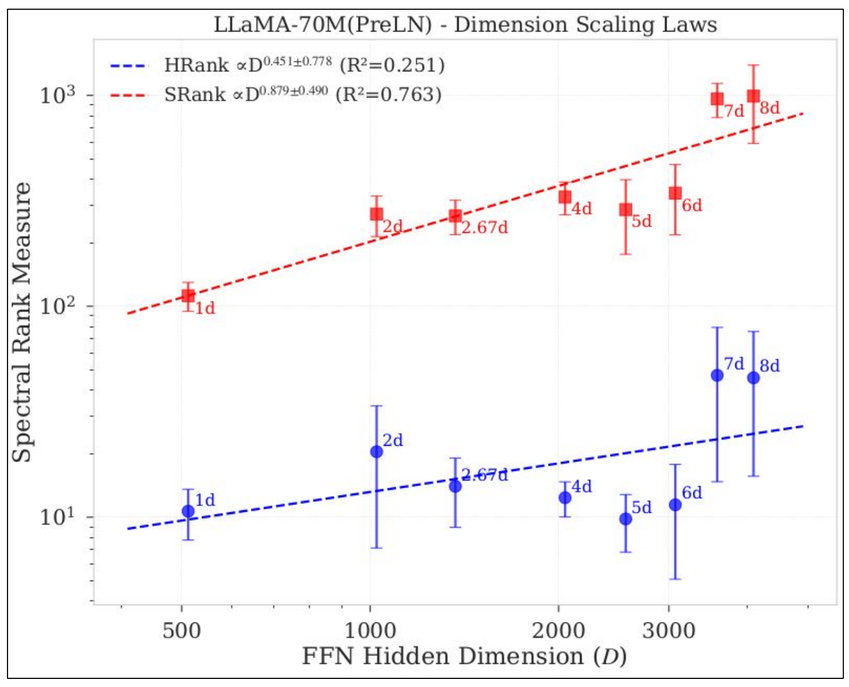
Soft Spectral Rank
(Shannon entropy rank)

$$\exp\left(-\sum_i \tilde{\lambda}_i \log \tilde{\lambda}_i\right)$$

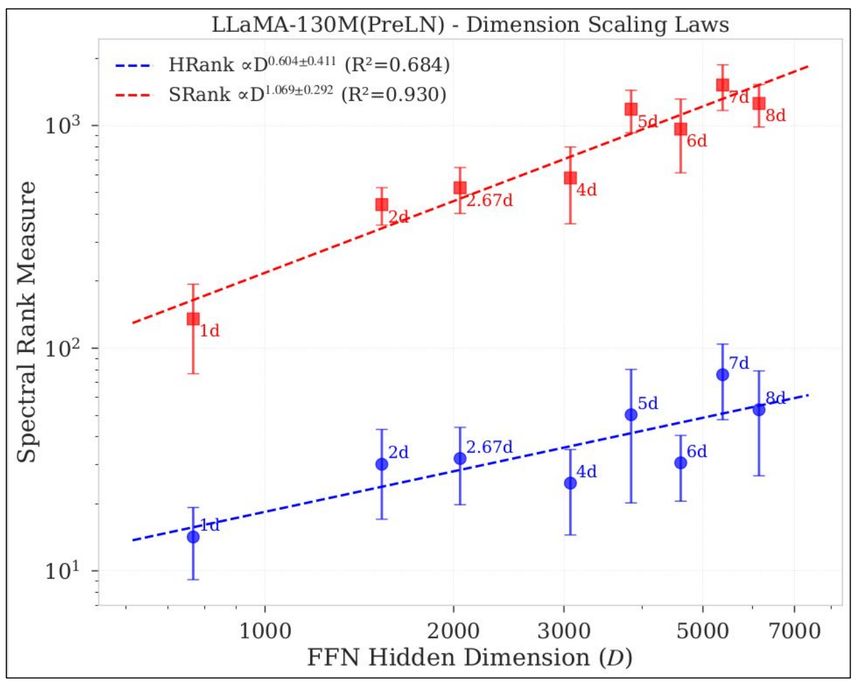
How uniformly variance is distributed in eigenspectrum (insensitive to spikes)

Key Findings: Asymmetric Spectral Scaling Laws

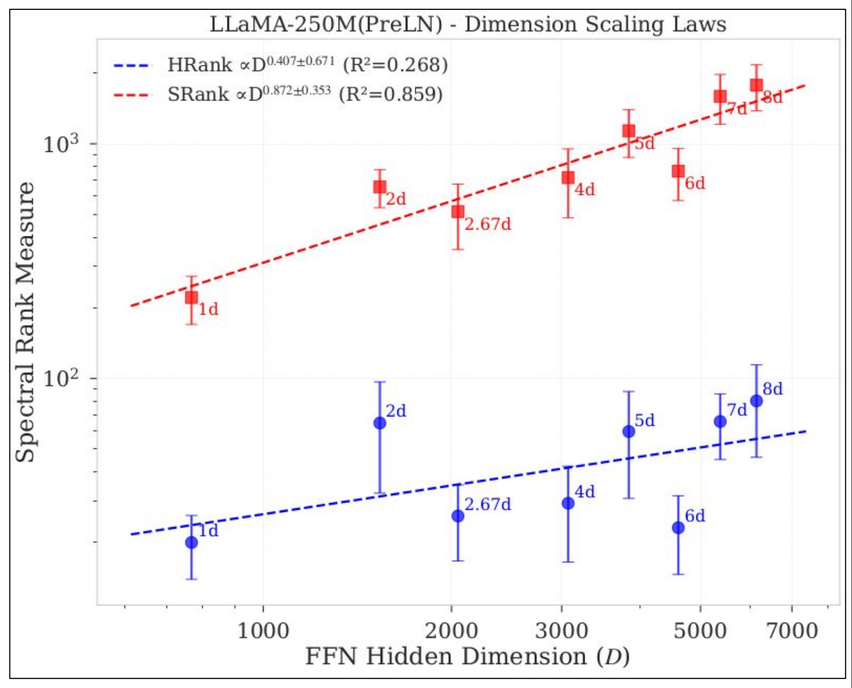
LLaMA-70M (PreLN)



LLaMA-130M (PreLN)



LLaMA-250M (PreLN)

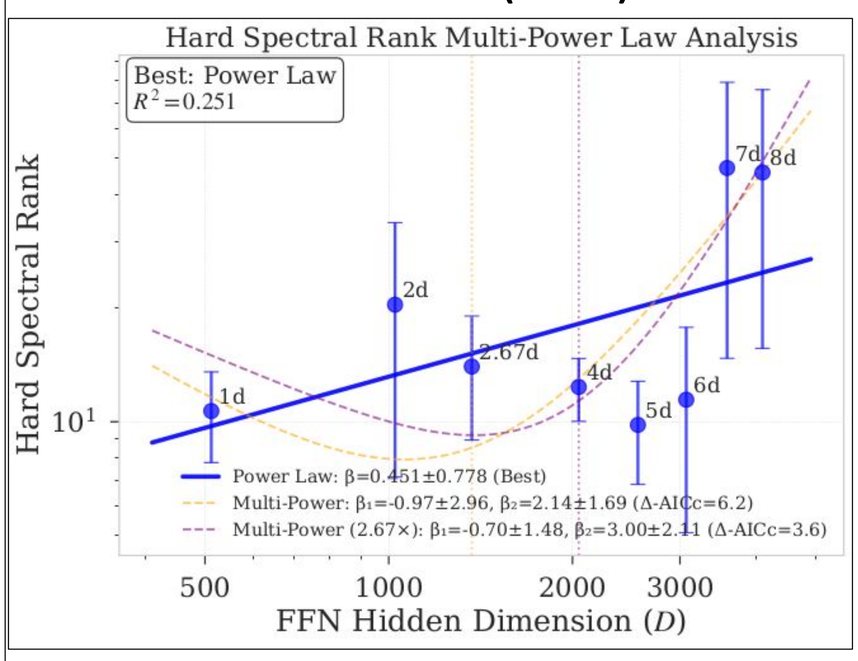


Asymmetric scaling pattern: Soft spectral rank exhibits better power law trend **($\beta \to 1$, $R^2 \to 1$)** than hard spectral rank **($\beta \to 0.5$, $R^2 \to 0.5$)**

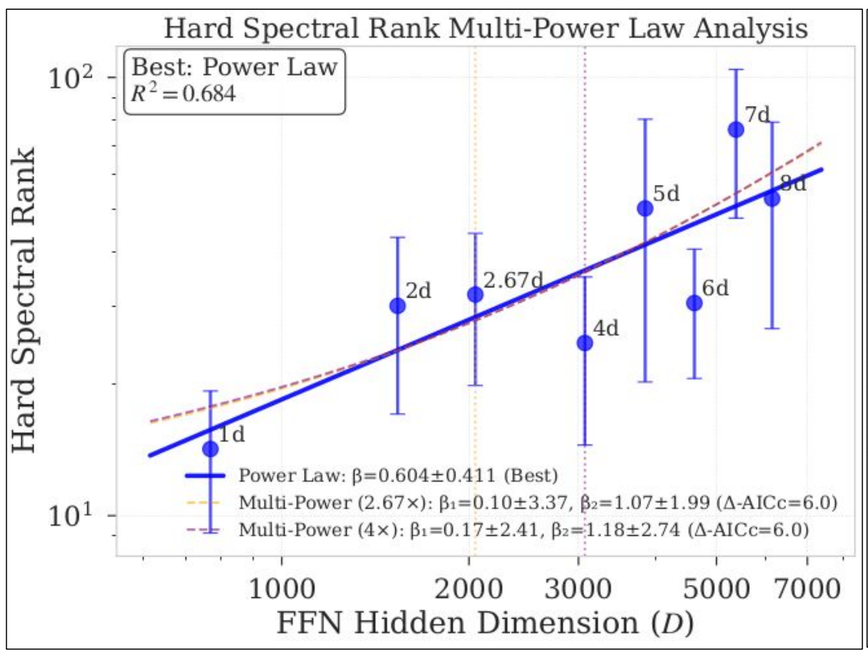
Takeaways: Widening the FFN keeps adding low-energy directions (tail capacity) while the high-energy subspace reaches diminishing returns

Multi-Power Laws for Hard Spectral Rank

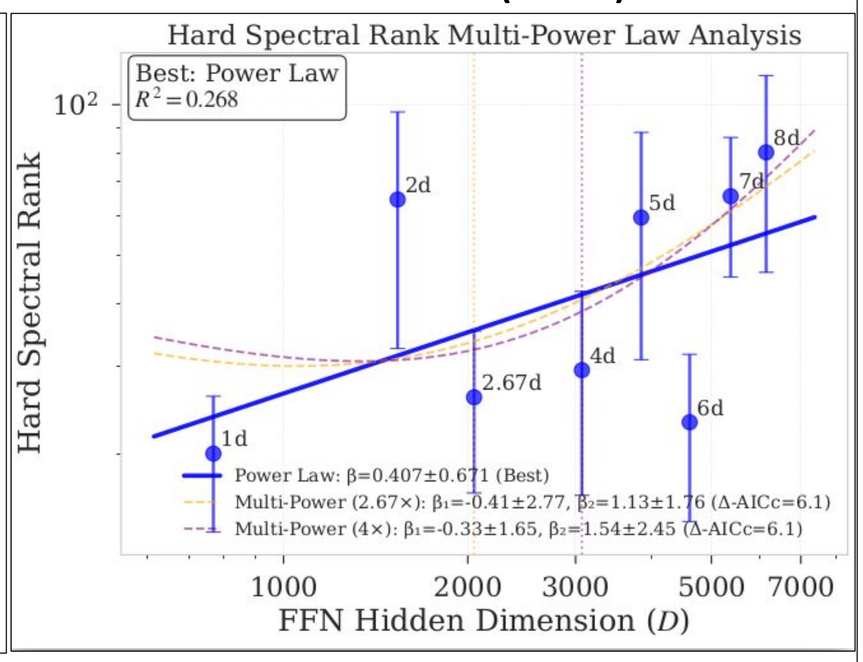
LLaMA-70M (PreLN)



LLaMA-130M (PreLN)



LLaMA-250M (PreLN)



Takeaways: Hard spectral rank does not exhibit a sharp knee; a single sub-linear power law ($\beta \approx 0.5$) explains LLaMA-70M $\longrightarrow$ 250M

Impact of LayerNorm Positioning on Spectral Scaling Laws

Model	PreLN		PostLN		MixLN	
	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 $(R^2 = 0.837)$

Takeaways: PostLN suppresses tail scaling; whereas MixLN achieves optimal capacity allocation by raising dominant-mode capacity and maintaining near-linear tail growth, delaying capacity saturation

Actionable Takeaways

Our spectral analysis demonstrate how widening the FFN primarily expands low-energy directions while dominant modes saturate earlier. We need architectural techniques (e.g., MixLN) to improve dominant mode energy without suppressing tail capacity

Contact:
nj2049@nyu.edu