Emergent Specialization: Rare Token Neurons in Language Models

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

Large language models struggle with representing and generating rare tokens despite their importance in specialized domains. In this study, we identify neuron structures with exceptionally strong influence on language model's prediction of rare tokens, termed as *rare token neurons*, and investigate the mechanism for their emergence and behavior. These neurons exhibit a characteristic three-regime organization (plateau, power-law, and rapid decay) that emerges dynamically during training, evolving from a homogeneous initial state to a functionally differentiated architecture. In the activation space, rare token neurons form a coordinated subnetwork that selectively co-activates while avoiding co-activation with other neurons. This functional specialization potentially correlates with the development of heavy-tailed weight distributions, suggesting a statistical mechanical basis for emergent specialization.

1. Introduction

While large language models (LLMs) have demonstrated remarkable capabilities in learning statistical patterns of human language, they consistently struggle with representing and generating rare tokens—words or phrases that appear infrequently in training data [13, 17, 31]. This challenge stems from the power-law distributions inherent in natural language [29, 32], where a significant portion of linguistic phenomena appears with extremely low frequency[4, 12]. Recent work has shown this limitation can lead to collapse when training on synthetic data that either truncates or narrows the tail of the distribution [2, 7, 11].

While several extrinsic and operational methods have been proposed to address this limitation—such as retrieval-augmented generation [15], in-context learning [8], and non-parametric memory mechanisms [3]—the intrinsic, mechanistic question remains: do LLMs develop internal mechanisms specialized for processing rare tokens during pre-training? This question parallels human language acquisition, where children demonstrate remarkable "fast mapping" abilities—learning new words after minimal exposure—from as young as 12 months of age [5, 19]. Cognitive neuroscience explains this through the Complementary Learning Systems (CLS) theory [22, 23], which posits that the brain employs two distinct neural systems: a neocortical system for gradual learning of distributed representations, and a hippocampal system specialized for rapid encoding of specific experiences, including rare events [14, 25].

Mechanistic interpretability research has revealed neurons encoding interpretable features ranging from syntactic relationships [18] to semantic concepts [10], but has primarily focused on common patterns. Stolfo et al. [26] discovered neurons that modulate token logits proportionally to frequency, but specialized mechanisms for rare tokens remain underexplored. In this study, we focus on decoder-only Transformer-based models and extend their work to focus on rare tokens and investigate

how individual neurons in the final MLP layer of transformer-based language models specialize in processing rare tokens during training.

Our analysis reveal three key findings: (i) LLMs develop dedicated "rare token neurons" that disproportionately impact the prediction of infrequent tokens; (ii) These specialized neurons emerge through distinct regimes during training; (iii) The emergence of specialized neuron groups correlates with the development of heavy-tailed weight distributions, suggesting a statistical mechanical basis for functional specialization.

2. Methodology

Inspired by prior work on confidence-regulating neurons [26], we hypothesize that certain neurons in language models specialize in modulating token-level probabilities—particularly for ***rare*** tokens that occur infrequently in the training data. To test this hypothesis, we conduct targeted ablation experiments across several language models, including the Pythia family [1], with intermediate checkpoints and training set available (The Pile [9]). Following the intervention approach of Stolfo et al. [26], we assess each neuron's influence by performing mean ablation experiments, that is, fixing a specific neuron's activation to its mean value over a reference dataset. We measure influence as the expected absolute change in token-level loss after ablation, computed from a filtered dataset of 25,088 context-token pairs sampled from the C4 Corpus [24].

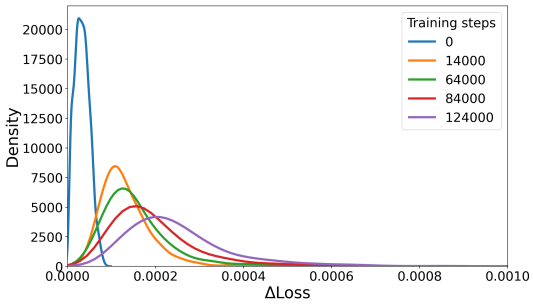


Figure 1: Absolute Δloss across training steps.

Specifically, for each neuron i, we compute the influence as:

$$\Delta\text{loss}(i) = \mathbb{E}_{x\sim\mathcal{D}}\left|\mathcal{L}(\text{LM}(x), x) - \mathcal{L}(\text{LM}(\tilde{x}^{(i)}), x)\right|, \tag{1}$$

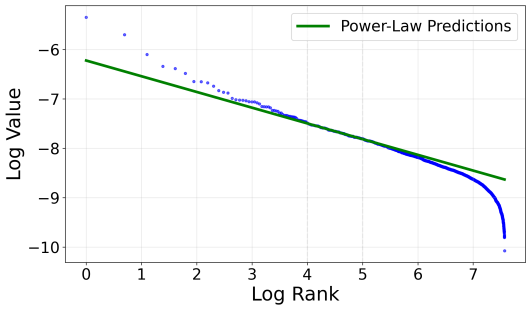
where $\text{LM}(x)$ denotes the model's output after applying LayerNorm and decoding, and $\mathcal{L}$ represents the token-level cross-entropy loss. The slope calculations for identifying structural transitions are performed using finite difference methods with sliding windows, as detailed in Appendix 5.3.1.

Figure 1 shows the distribution of per-neuron influence across training, measured as the absolute change in token-level loss after ablation. The concentration of neurons near zero Δloss, and a tail with large Δloss suggests that a small subset becomes particularly influential for rare tokens during training. We refer to these as *rare token neurons*. Within this subset, we define *boosting neurons* as those that increase the likelihood of rare tokens, and *suppressing neurons* as those that decrease it.

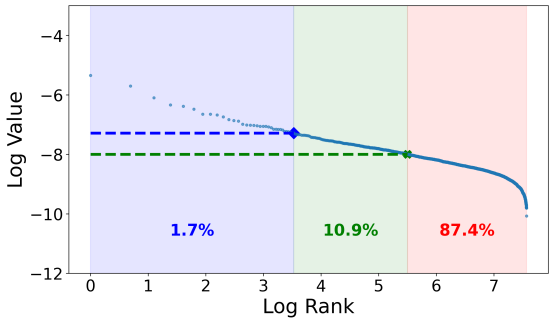
3. Results

3.1. Three-regime Structure in Neuron Influence

Ranking neurons by their ΔLoss reveals a consistent three-regime structure presented in log-log scale across model scales and architectures (Figure 2b; more results in Figure 5.3.4). This structure suggests a functional specialization composed of: i.) **Influential plateau regime** where a small



(a) Absolute Δloss distribution across training steps.



(b) Three-regime structure of neuron influence.

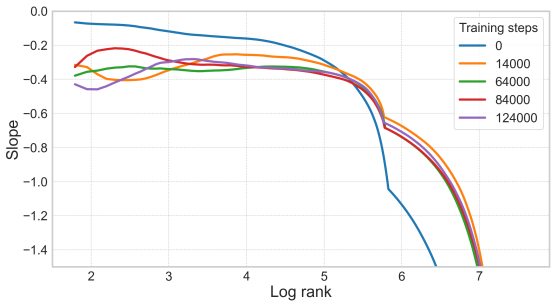
Figure 2: (a) The green line shows the power-law prediction; influence declines faster on the right and deviates on the left due to an emerging bias, though the slope remains within the power-law regime. (b) Illustration of the three-regime structure.

fraction (1.7%) of neurons exhibit consistently large influence, forming a plateau in the leftmost region; ii.) **Power-law regime** where the majority of influential neurons follow a power-law relationship, which appears as a linear relation in log-log coordinates

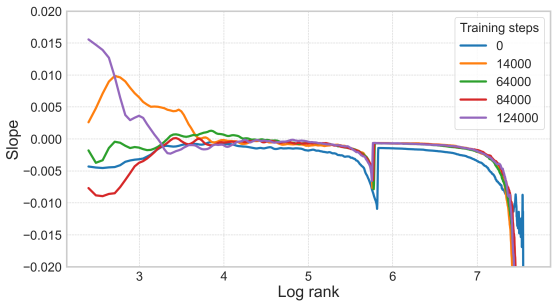
$$\log|\Delta\text{Loss}| \approx -\kappa \log(\text{rank}) + \beta, \tag{2}$$

where the power-law exponent κ appears as the slope of a linear function; and iii.) **Rapid decay tail regime** where the remaining neurons decay more rapidly than power-law predictions, indicating negligible contribution to rare token prediction.

Power-Law to Rapid Decay Transition The transition from power-law to rapid decay can be identified through the slope behavior of $\log|\Delta\text{Loss}|$ vs. $\log(\text{rank})$. Using the finite difference method detailed in Appendix 5.3.4), we estimate the local slope $\kappa(r)$ with a sliding window approach. As shown in Figure 3.1a, the first derivative starts to decrease around $\log(\text{rank}) \approx 5$, marking the breakdown of the power-law and the onset of rapid decay. This characterizes a functional boundary between moderately and minimally influential neurons.



(a) First derivative of the loss contribution in $\log - \log$ coordinates.



(b) Second derivative of the loss contribution in $\log - \log$ coordinates.

Emergence of the Plateau Regime Unlike the rapid decay transition, the plateau regime is not readily distinguishable by the first derivative alone. Neurons in the range $\log(\text{rank}) \in (2, 5)$ exhibit a relatively consistent power-law behavior but with increased baseline influence. Such increment is quantified by

$$\delta := \log|\Delta\text{Loss}| - (-\kappa \log \text{rank} + \beta),$$

which measures to what extend the power-law underestimates the influence of top-ranked neurons. Figure 3 shows that *the highly-influential plateau emerges progressively during training.*—implying functional specialization through training.

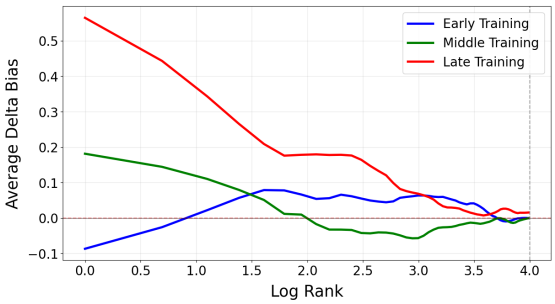


Figure 3: Absolute Δloss across training steps.

Second-Order Derivative Analysis Our slope analysis reveals a notable feature in the derivative structure. While the first derivative of $\log|\Delta\text{Loss}|$ versus $\log(\text{rank})$ remains continuous, the second derivative exhibits a discontinuity around the power-law to rapid decay transition (see Figure 3.1)b. This pattern provides additional evidence for the structural transition between regimes.

The emergence of this three-regime organization during training suggests that language models spontaneously develop specialized computational strategies for rare token processing, with different neuron populations serving distinct functional roles.

3.2. Co-activation Patterns Through the Lens of Activation Space Geometry

We analyze the behavior of rare-token-influential neurons through the geometry of their activation patterns. Despite being selected via individual ablation experiments, these neurons exhibit systematic organizational structure that differs from random neuron groups: they co-activate strongly with each other while systematically avoiding co-activation with neurons less involved in rare token prediction as shown in within-group and cross-group correlations in Table 1. While the absolute correlation values are modest, statistical testing reveals meaningful differences: for Pythia-410M, rare token boosting neurons show within-group correlation of 0.036 ± 0.008 compared to 0.007 ± 0.003 for random neurons ($p < 0.001$, Wilcoxon rank-sum test with Bonferroni correction). The cross-group correlation between boosting and suppressing neurons (0.040 ± 0.009) exceeds both individual group correlations with random neurons, suggesting these functionally opposing groups operate within a shared computational framework rather than independently.

To further investigate this structure, we construct high-dimensional activation vectors for each neuron using context-token pairs from the C4 corpus [24]. We then examine the geometric patterns with effective dimension and cosine similarity.

Effective dimension analysis reveals that rare-token neurons lie on a significantly lower-dimensional manifold than random neurons. Across model families, rare token neurons show 8-13% reduction in effective dimensionality needed to explain 95% of activation variance. This compression suggests that they activate in a more coordinated, structured manner rather than independently(see results in Table 2).

Pairwise cosine similarity provides additional evidence for functional organization(see results in Table 3). Random neuron pairs show near-zero similarity (mean $\approx$ 0.05), consistent with

uncorrelated activation patterns. Rare-token boosting and suppressing neurons exhibit higher within-group similarity ($0.09 - 0.17$), indicating some degree of coordinated activation.

Notably, these two groups also show substantial cross-group similarity, despite their opposing effects—suggesting they operate in coordinated, antagonistic roles.

3.3. Weight Eigenspectrum

To investigate how the network progressively develops functional differentiation, we apply Heavy-Tailed Self-Regularization (HT-SR) theory [20, 21] to analyze the eigenspectral properties of neuron groups. This analysis examines whether rare token neurons develop distinct weight matrix characteristics compared to random neuron populations.

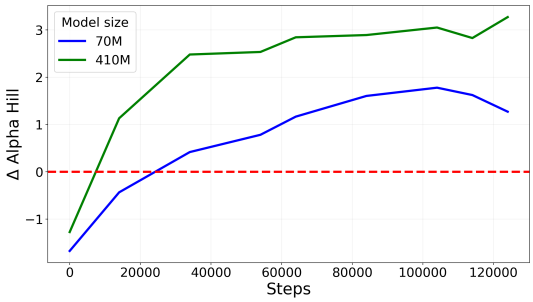


Figure 4: Absolute Δloss across training steps.

For each neuron group G, we compute its correlation matrix and then analyze the eigenvalue spectrum $\{\lambda_i\}$ of Ξ_G to assess the internal structure of the group's learned representations. To quantify spectral shape, we use the Hill estimator to measure the power-law exponent in the tail of the eigenvalue distribution. Details are provided in Appendix 5.3.3

Figure 4 shows that specialized neurons consistently exhibit lower α_{Hill} values—i.e., heavier-tailed distributions—compared to random neurons after the initial training phase. This pattern holds across model families and sizes (see results in Table 4). This persistent separation provides strong evidence for functional differentiation through implicit regularization. Despite fluctuations during training, the fundamental pattern remains: neurons that significantly impact rare token prediction consistently develop more pronounced heavy-tailed characteristics than neurons with random or general functionality.

4. Discussion and Conclusion

Based on our empirical observations, we propose two mechanistic conjectures to explain the emergence of rare token processing capabilities in language models:

Conjecture 4.1 (Dual-Regime Organization) *The emergence of power-law regime and its distinction from the rapid decay regime suggest a spontaneous specialization of influential neurons. Among the rare token neurons, the power-law structure, the α_{Hill} behavior, and the co-activation patterns indicate self-organization phenomena that exceed random expectations.*

Conjecture 4.2 (Parallel Mechanism Conjecture) *The plateau regime emerges through a mechanism that parallels the mechanism for the emergence of power-law regime. Through training, it further differentiates a small subset of neurons within the power-law group, by increasing their influence to form the influential plateau.*

The Dual-Regime Organization conjecture is motivated by the observation that the $\log \Delta$loss–$\log$ rank slope remains consistent across both the power-law and plateau regimes. This consistency suggests

that an underlying power-law structure governs both regimes, with the plateau reflecting an additional, distinct mechanism operating on top of this foundation. The Parallel Mechanism conjecture proposes that rare token processing relies on two complementary computational strategies: a distributed regime (power-law) for general rare token sensitivity and a specialized subnetwork (plateau) for exceptional cases. This resembles the Complementary Learning Systems (CLS) theory in cognitive neuroscience [14, 22], where general statistical learning coexists with specialized mechanisms for encoding exceptions and novel experiences.

However, we emphasize that these conjectures are preliminary hypotheses based on our empirical observations. The modest effect sizes in our coordination analyses and the indirect nature of our weight eigenspectrum measurements suggest that stronger evidence would be needed to definitively establish these mechanisms.

This paper presents a systematic investigation into the emergent neuronal mechanisms that language models develop for processing rare tokens—a fundamental challenge requiring a balance between learning and low-frequency generalization. Through ablation experiments and geometric analysis, we identified neuron groups with disproportionate influence on rare token prediction, organized through co-activation and heavy-tailed statistics. Our analysis revealed a three-regime structure of influence: a specialized influential plateau regime, a power-law regime following efficient coding principles, and a rapid decay regime with minimal contribution to rare token processing. These regimes emerge progressively during training, suggesting spontaneous functional differentiation rather than predetermined architectural specialization. While our evidence supports the existence of rare token neurons and their organizational structure, we acknowledge that the underlying mechanisms remain partially understood. The modest coordination effects and indirect spectral measures indicate that stronger theoretical frameworks and measurement techniques are needed to fully characterize these phenomena.

These results highlight the emergence of computational specialization in large language models, with implications for interpretability, efficiency optimization, and targeted model improvement. As language models continue to scale, understanding how they spontaneously develop specialized capabilities will become increasingly important for both theoretical advancement and practical applications.

Our findings also suggest new directions for interpretability research. Instead of examining individual neurons in isolation, future work could investigate how specialized subnetworks coordinate to handle low-frequency linguistic phenomena. The emergence of these structures during training raises questions about whether similar specialization occurs for other phenomena beyond rare tokens.

References

[1] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[2] Maty Bohacek and Hany Farid. Nepotistically trained generative image models collapse. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*, 2023.

[3] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Breviglieri Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240. PMLR, 2022.

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[5] Susan Carey and Elsa Bartlett. Acquiring a single new word. *Papers and Reports on Child Language Development*, 15:17–29, 1978.

[6] Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.

[7] Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

[8] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Zhifang Sui, Wangbo Liu, Yiming Yang, et al. A survey of in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[9] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[10] Wes Gurnee, Aditi Raghunathan, and Neel Nanda. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

[11] Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20555–20565, 2023.

[12] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[13] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR, 2023.

[14] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in Cognitive Sciences*, 20(7):512–534, 2016.

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[16] Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang Wang, Michael W Mahoney, and Yaoqing Yang. Alphapruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. *Advances in Neural Information Processing Systems*, 37: 9117–9152, 2024.

[17] Stella Mallen, Jennifer Hou, Eric Wallace, Mark Dredze, and Nadia Hegde. Not all knowledge is created equal: Tracking the impact of memorization across pre-training and fine-tuning. *arXiv preprint arXiv:2310.02173*, 2023.

[18] Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

[19] Lori Markson and Paul Bloom. Children's fast mapping of word meaning. *Cognitive Psychology*, 33(1):73–110, 1997.

[20] Charles H Martin and Michael W Mahoney. Traditional and heavy-tailed self regularization in neural network models. *arXiv preprint arXiv:1901.08276*, 2019.

[21] Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

[22] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419, 1995.

[23] Randall C O'Reilly, Rajan Bhattacharyya, Michael D Howard, and Nicholas Ketz. Complementary learning systems. *Cognitive Science*, 38(6):1229–1248, 2014.

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

[25] Anna C Schapiro, Nicholas B Turk-Browne, Matthew M Botvinick, and Kenneth A Norman. Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711):20160049, 2017.

[26] Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. *Advances in Neural Information Processing Systems*, 37:125019–125049, 2024.

[27] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[28] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[29] Ronald E Wyllys. Empirical and theoretical bases of zipf's law. *Library Trends*, 30(1):53–64, 1981.

[30] Yaoqing Yang, Ryan Theisen, Liam Hodgkinson, Joseph E Gonzalez, Kannan Ramchandran, Charles H Martin, and Michael W Mahoney. Test accuracy vs. generalization gap: Model selection in nlp without accessing training or testing data. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3011–3021, 2023.

[31] Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binquan Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[32] George Kingsley Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

5. Appendix

5.1. Limitations

Our study has several limitations.

First, the observed coordination effects, are modest in magnitude. The correlation values and similarity measures suggest weak-to-moderate coordination rather than the strong coupling that might be expected from highly specialized circuits. However, the consistency of these patterns across multiple models, measures, and statistical tests provides evidence for systematic organization that exceeds random expectations. This may reflect the distributed nature of neural computation in transformers, where even modest coordination across many neurons can produce significant functional effects.

Future work incorporating more sophisticated measures of functional connectivity, such as mutual information or causal intervention analysis, could provide stronger evidence for the proposed specialization mechanisms.

Additionally, we focus exclusively on neurons in the final MLP layer, while rare token processing likely involves multiple components and mechanisms throughout the model architecture. A comprehensive analysis of these distributed mechanisms could provide deeper insights into how language models handle infrequent tokens.

What's more, we lack a precise theoretical framework for quantifying neuron effects and instead rely on ablation-based proxies such as change in loss. More principled measures of individual neuron contributions would strengthen our mechanistic understanding and enable more robust conclusions about functional specialization.

Finally, our analysis centers on next-token prediction in language modeling contexts. Investigating rare token processing in downstream tasks such as question-answering, reasoning, or domain-specific applications would illuminate the practical implications of these specialized mechanisms and their role in real-world model performance.

5.2. Background

5.2.1. TRANSFORMER ARCHITECTURE

In this study, we focus on the Multi-Layer Perceptron (MLP) sublayers. Given a normalized hidden state $x \in \mathbb{R}^{d_{\text{model}}}$ from the residual stream, the MLP transformation is defined as:

$$\text{MLP}(x) = W_{\text{out}}\phi(W_{\text{in}}x + b_{\text{in}}) + b_{\text{out}}, \tag{3}$$

where $W_{\text{in}} \in \mathbb{R}^{d_{\text{mlp}} \times d_{\text{model}}}$ and $W_{\text{out}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{mlp}}}$ are learned weight matrices, and $b_{\text{in}}, b_{\text{out}}$ are biases. The nonlinearity ϕ is typically a GeLU activation. We refer to individual entries in the hidden activation vector $\phi(W_{\text{in}}x+b_{\text{in}})$ as *neurons*, indexed by their layer and position (e.g., `<layer>.<index>`). The activations n represent post-activation values of these neurons. We selected the last layer as it directly projects into the unembedding matrix that produces token probabilities, which creates a computational bottleneck where feature integration must occur [28].

5.2.2. HEAVY-TAILED SELF-REGULARIZATION (HT-SR) THEORY

Heavy-Tailed Self-Regularization (HT-SR) theory offers a spectral lens on neural network generalization[6, 16, 20, 21]. Specifically, consider a neural network with L layers, let W_i denote a weight matrix

extracted from the i-th layer, where $W_i \in \mathbb{R}^{m \times n}$ and $m \geq n$. We define the correlation matrix associated with W_i as:

$$X_i := W_i^\top W_i \in \mathbb{R}^{n \times n},$$

which is a symmetric, positive semi-definite matrix. The empirical spectral distribution (ESD) of X_i is defined as:

$$\mu_{X_i} := \frac{1}{n} \sum_{j=1}^{n} \delta_{\lambda_j(X_i)},$$

where $\lambda_1(X_i) \leq \cdots \leq \lambda_n(X_i)$ are the eigenvalues of X_i, and δ is the Dirac delta function. The ESD μ_{X_i} represents a probability distribution over the eigenvalues of the weight correlation matrix, characterizing its spectral geometry.

HT-SR theory proposes that successful neural network training exhibits heavy-tailed spectral behavior in the ESDs of certain weight matrices, due to self-organization toward a critical regime between order and chaos. This phenomenon is quantitatively captured through **Shape metrics**, which quantify the geometry of the ESD through PL α_{Hill} (our primary metric), PL $\hat{\alpha}$, Spectral entropy, and Stable rank. Among these, the power-law (PL) exponent α_{Hill} is particularly informative, as it estimates the tail-heaviness of the eigenvalue distribution using a robust Hill estimator. Low values of α_{Hill} (typically $\alpha < 2$) indicate heavy-tailed behavior, often interpreted as signs of functional specialization and self-organized criticality [30]. A formal definition of α_{Hill} and the associated estimation procedure is provided in Section 3.3.

5.3. Details on rare token neuron analysis framework

5.3.1. ABLATION EXPERIMENT

Formally, let $i \in \{1, 2, \ldots, d_{\text{mlp}}\}$ index a neuron in the MLP layer, and let $n_i \in \mathbb{R}$ denote its activation. For a given input $x \in \mathcal{X}$, let x represent the final hidden state (i.e., the output of the last transformer block). The mean-ablated hidden state $\tilde{x}^{(i)}$ is then given by:

$$\tilde{x}^{(i)} = x + (\bar{n}_i - n_i) w_{\text{out}}^{(i)}, \tag{4}$$

where $\bar{n}_i$ is the mean activation of neuron i across a reference subset of inputs, and $w_{\text{out}}^{(i)}$ is the corresponding output weight vector.

The neuron effects are computed as:

$$\Delta\text{loss}(i) = \mathbb{E}_{x \sim \mathcal{D}} \left| \mathcal{L}(\text{LM}(x), x) - \mathcal{L}(\text{LM}(\tilde{x}^{(i)}), x) \right|, \tag{5}$$

where $\text{LM}(x)$ denotes the model's output after applying LayerNorm and decoding, and $\mathcal{L}$ represents the token-level cross-entropy loss.

To filter the tokens, we implement a two-stage filtering process: at stage one, we retain tokens below the 50th percentile in the unigram frequency distribution of the training set; then at stage two, we restrict our analysis to valid, correctly spelled English words[1], eliminating potential noise from malformed tokens. This is mainly due to our primary focus on rare tokens.

1. Token was filtered with pyspellchecker library: https://pypi.org/project/pyspellchecker/

5.3.2. GEOMETRIC ANALYSIS STATISTICS

Analysis details For each neuron pair (i, j), we first calculate the Pearson correlation coefficient ρ_{ij} between their activation vectors, then transform it into a distance metric:

$$D_{ij} = 1 - |\rho_{ij}|, \tag{6}$$

which captures dissimilarity while remaining agnostic to the direction of correlation.

We apply hierarchical agglomerative clustering with Ward linkage to this distance matrix. Specifically, we measure the number of distinct clusters that emerge at a distance threshold of $t = 0.5$. A larger number of clusters would indicate greater functional modularity within the rare-token neuron population, while fewer clusters would suggest more globally coordinated behavior.

Firstly, we introduce the *effective dimensionality* of each neuron's activation distribution using Principal Component Analysis (PCA). Formally, the effective dimension d_{eff} is defined as the smallest d such that the cumulative variance explained exceeds a fixed threshold τ:

$$d_{\text{eff}} = \min \left\{ d : \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{j=1}^{N} \lambda_j} \geq \tau \right\},$$

where λ_i denotes the i-th eigenvalue of the activation covariance matrix.

The second statistic is the *pairwise cosine similarity* between the activation vectors, measuring the activation similarity between neurons, regardless of their activation intensities. Let $\mathbf{h}_i, \mathbf{h}_j \in \mathbb{R}^T$ denote activation traces across T token contexts:

$$\cos(\theta_{ij}) = \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|}.$$

Analysis results We analyze multiple models with differnet parameter sizes and the results exhibit the higher activation correlation within the selected neuron groups while no such effect in the random control group.

Table 1: Activation correlation within and between neuron groups (group_size=50)

Model	Model size	boost	suppress	random	b v.s. r	s v.s. r	b v.s. s	r1 v.s. r2
Pythia-70M	70M	0.028	0.045	0.011	0.002	0.005	0.027	0.009
Pythia-410M	410M	0.036	0.052	0.007	0.005	0.006	0.040	0.007
GPT2-Small	124M	0.017	0.019	0.017	-0.001	-0.001	0.021	0.023
GPT2-Large	774M	0.004	0.011	0.012	-0.004	-0.004	0.010	0.016
GPT2-XL	1.5B	0.036	0.016	-0.007	0.003	-0.0004	0.020	0.008

We also found consistently higher effective dimension proportion in the random baseline group compared with the selected neuron group as shown in table 2.

The cosine similarity also show the mostly positive alignment within the same neuron group while negative alignment between the random baseline and the selected neuron group.

Table 2: Proportion of effective dimensions across neuron groups (group_size=50)

Model	Model size	boost	suppress	random
Pythia-70M	70M	33.5	32.6	36.2
Pythia-410M	410M	33	32.2	37.3
GPT2-Small	124M	37	40	45
GPT2-Large	774M	43	43	46
GPT2-XL	1.5B	40	42	46

Table 3: Cosine similarity between and within neuron groups (group_size=50)

Model	Model size	boost	suppress	random	b v.s. r	s v.s. r	b v.s. s	r1 v.s. r2
Pythia-70M	70M	0.141	0.165	0.021	-0.017	-0.014	0.146	0.021
Pythia-410M	410M	0.107	0.133	0.054	-0.032	-0.041	0.114	0.058
GPT2-Small	124M	0.109	0.122	0.089	-0.100	-0.105	0.120	0.099
GPT2-Large	774M	0.028	0.092	0.041	-0.034	-0.063	0.054	0.052
GPT2-XL	1.5B	0.095	0.095	0.009	-0.010	-0.015	0.090	0.012

5.3.3. WEIGHT EIGENSPECTRUM

To understand the emergence of specialized neuron groups, we analyze model checkpoints across different training steps. This analysis enables us to track how the network progressively develops functional differentiation through the lens of Heavy-Tailed Self-Regularization (HT-SR) theory.

HT-SR theory, introduced in Section 5.2.2 suggests that heavy-tailed structures emerge from feature learning, where useful correlations are extracted during optimization. Neuron groups with more heavy-tailed ESDs which contain more learned signals, are assigned lower sparsity, while neuron groups with light-tailed ESDs are assigned higher sparsity. In practice, for each neuron group $\mathcal{G}$, we compute its correlation matrix as

$$\mathbf{\Xi}_{\mathcal{G}} = \frac{1}{d}\mathbf{W}_{\mathcal{G}}\mathbf{W}_{\mathcal{G}}^{\top},$$

where $\mathbf{W}_{\mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|\times d}$ denotes the slice of the weight matrix corresponding to the group $\mathcal{G}$. We then analyze the eigenvalue spectrum $\{\lambda_i\}$ of $\mathbf{\Xi}_{\mathcal{G}}$ to assess the internal dimensionality and structure of the group's learned representations.

To quantify the spectral shape, we use the Hill estimator to measure the power-law exponent α_{Hill} in the tail of the eigenvalue distribution:

$$\alpha_{\text{Hill}} = \left[\frac{1}{k}\sum_{i=1}^{k}\log\left(\frac{\lambda_i}{\lambda_k}\right)\right]^{-1}, \tag{7}$$

where k is a tunable parameter that adjusts the lower eigenvalue threshold λ_{min} for (truncated) PL estimation. Following prior work on layer-wise pruning [16], we apply the Fix-finger method [30] to select the k, which sets k to align λ_{min} with the peak of the ESD. By tracking the evolution of α_{Hill} across training, we can infer how specialized substructures or subnetworks progressively form and adapt.

Table 4: Alpha hills of neuron groups (group_size=50)

Model	Model size	boost	suppress	random
Pythia-70M	70M	4.30	3.97	6.37
Pythia-410M	410M	3.80	3.43	7.56
GPT2-Small	124M	2.12	1.57	6.74
GPT2-Large	774M	3.30	1.84	8.31
GPT2-XL	1.5B	2.01	1.68	9.33

5.3.4. REGIME TRANSITION DETAILS

Regime identification To precisely identify regime boundaries and track their evolutions during training, it is critical to understand the power-law exponent, appearing as a slope. We employ the finite difference method with a sliding window for estimating this slope:

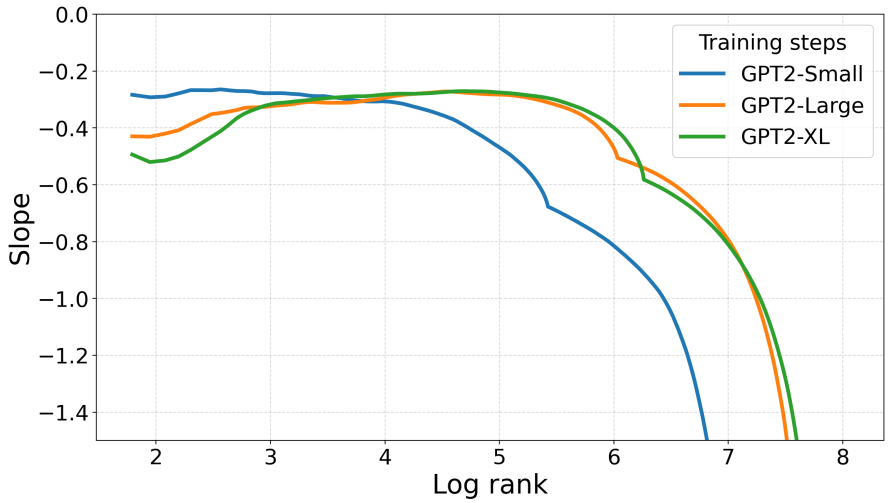
$$-\kappa(r) \approx -\frac{\log|\Delta\text{Loss}(r \cdot e)| - \log|\Delta f\text{Loss}(r)|}{\log(e)} \tag{8}$$

where r is the rank and e is Euler's number. This finite-difference approximation provides a robust estimate of the local slope in log-log space, thus enabling the identification of the behavior of $-\kappa(r)$, in particular transition points where it changes significantly. The three regimes are then identified using an automated change point detection algorithm [27] applied to the $\kappa(r)$ curve, which identifies transition points where the slope changes dramatically. We validate these automatically detected boundaries through manual inspections for distribution differences on either side of the boundaries.

Plateau regime identification We characterize the plateau regime by calculating the difference between observed influence values $\log|\Delta\text{Loss}(r)|$ and the power-law prediction $(-\kappa\log(r) + \beta)$:

$$\delta(r) = \log|\Delta\text{Loss}(r)| - (-\kappa\log(r) + \beta) \tag{9}$$

where κ and β are parameters estimated from the power-law regime region. The quantity $\delta(r)$ illustrates how much the ranked neuron r deviates from the power-law prediction. *A plateau regime is therefore characterized as a* $\log$ *rank range where* $\delta(r)$ *is bounded above a positive value, hence the name "plateau".*



Neuron slope distributions of gpt2 model family