Emergent Specialization: Rare Token Neurons in Language Models

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Large language models (LLMs) struggle with representing and generating rare tokens despite their importance in specialized domains. In this study, we identify a reproducible *three-regime organization* of neuron influence in the final MLP layer for rare-token prediction, composed of: (i) a highly influential plateau of specialist neurons, (ii) a power-law regime of moderately influential neurons, and (iii) a rapid decay regime of minimally contributing neurons. We show that neurons in the plateau and power-law regimes form coordinated subnetworks with distinct geometric and co-activation patterns, while exhibiting heavy-tailed weight distributions, consistent with predictions from Heavy-Tailed Self-Regularization (HT-SR) theory. These specialized subnetworks emerge dynamically during training, transitioning from a homogeneous initial state to a functionally differentiated architecture. Our findings reveal that LLMs spontaneously combine distributed power-law sensitivity with specialized rare-token processing. This mechanistic insight bridges theoretical predictions from sparse coding and superposition with empirical observations, offering pathways for interpretable model editing, efficiency optimization, and deeper understanding of emergent specialization in deep networks.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in modeling complex linguistic patterns, yet they consistently struggle with representing and generating *rare tokens*, that is, words or phrases that appear infrequently in the training data [12, 33, 16]. This limitation arises from the heavy-tailed frequency distributions of natural language [34, 31], where a significant portion of linguistic phenomena appears with extremely low frequency[4, 11]. 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 [7, 10, 2].

While extrinsic approaches have been proposed such as retrieval-augmented generation [14], in-context learning [8], and non-parametric memory mechanisms [3], the intrinsic mechanisms by which LLMs internally organize computation for rare token prediction remain largely unexplored. Understanding these mechanisms is crucial both for theoretical advancement in mechanistic interpretability and for practical applications in model improvement and efficiency optimization.

This challenge parallels human language acquisition, where children rapidly learn new words from minimal exposure [5, 18]. Cognitive neuroscience explains this through Complementary Learning Systems (CLS) theory [21, 13], 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 [13, 26]. Analogously, LLMs may spontaneously develop distributed and specialized circuits for handling low-frequency tokens.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Recent mechanistic interpretability studies have revealed neurons encoding syntactic [17] and semantic [9] features, as well as frequency-sensitive neurons modulating token logits [27]. However, the organizational principles governing rare token processing—particularly how neurons self-organize into functional subnetworks—remain underexplored.

In this study, we investigate decoder-only transformers and focus on the final MLP layer, where feature integration is critical for token prediction [30]. Across multiple model families, we identify neuron populations that disproportionately influence rare token prediction, forming a **three-regime structure** of influence:

1. **Influential plateau:** A small subset of neurons with exceptionally high impact on rare token prediction.
2. **Power-law regime:** A majority of moderately influential neurons following a scale-free distribution, consistent with sparse feature learning.
3. **Rapid decay:** A long tail of minimally contributing neurons.

We further demonstrate that plateau and power-law neurons form coordinated activation subspaces, exhibit structured geometric organization, and develop heavy-tailed weight distributions consistent with Heavy-Tailed Self-Regularization (HT-SR) theory [19, 20]. These findings suggest that LLMs spontaneously self-organize into specialized subnetworks for rare token processing, bridging theoretical predictions from sparse coding and emergent superposition with empirical observations. These results provide insights for understanding how language models spontaneously develop computational specialization for low-frequency linguistic phenomena.

2 Background

2.1 Transformer architecture

We focus on the Multi-Layer Perceptron (MLP) sublayers of decoder-only transformers, where feature integration is critical for token prediction [30]. Given a normalized hidden state $x \in \mathbb{R}^{d_{\text{model}}}$ from the residual stream, the MLP transformation is:

$$\text{MLP}(x) = W_{\text{out}}\, \phi(W_{\text{in}} x + b_{\text{in}}) + b_{\text{out}}, \tag{1}$$

where $W_{\text{in}} \in \mathbb{R}^{d_{\text{mlp}} \times d_{\text{model}}}$, $W_{\text{out}} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{mlp}}}$, and $b_{\text{in}}, b_{\text{out}}$ are learned biases. The nonlinearity ϕ is typically GeLU, though gated activations such as SiLU are common. We refer to the post-activation entries $\phi(W_{\text{in}} x + b_{\text{in}})$ as *neurons*, indexed by their layer and position (e.g., `<layer>.<index>`).

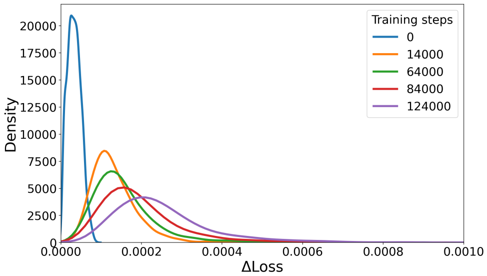


Figure 1: Absolute change in token-level cross-entropy loss (Δloss) per neuron across training steps. Rare-token-specialized neurons gradually emerge, exhibiting disproportionately high influence compared to the majority of neurons with negligible effect.

We select the *last MLP layer* for analysis, as it directly projects into the unembedding matrix that produces token probabilities. While this choice limits generalizability to earlier layers, it provides a principled starting point for studying emergent specialization in rare token processing.

2.2 Heavy-Tailed Self-Regularization (HT-SR) Theory

HT-SR theory provides a spectral lens for understanding functional specialization in neural networks [19, 20, 15, 6]. It connects functional specialization to statistical properties of weight matrices and their eigenvalue distributions.

For 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:

$$X_i := W_i^\top W_i \in \mathbb{R}^{n\times n}, \tag{2}$$

The empirical spectral distribution (ESD) of X_i is defined as:

$$\mu_{X_i} := \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(X_i)}, \tag{3}$$

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 posits that successful training induces heavy-tailed spectral behavior, reflecting self-organization toward a critical regime between order and chaos. The tail-heaviness is quantified via the Hill estimator α_{Hill} (Section 4.3), where low α_{Hill} (typically $\alpha < 2$) indicates strong functional specialization and distributed feature learning. In our context, neurons forming heavy-tailed subnetworks are hypothesized to correspond to rare token specialists, consistent with principles from sparse coding and information bottleneck theory [22, 28].

3 Rare Token neuron analysis framework

3.1 Rare Token Neuron Identification

We hypothesize that a subset of neurons in the final MLP layer functionally specialize to modulate the prediction of **rare tokens**, that is, tokens with low frequency in the training corpus. This hypothesis is grounded in sparse coding [22] and the information bottleneck principle [28], suggesting that neural resources are allocated selectively to efficiently represent low-frequency, high-information content signals.

Ablation methodology To test this hypothesis, we perform targeted ablation experiments following Stolfo et al. [27]. For neuron i with activation n_i in the last MLP layer, we define the *mean-ablated activation*:

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

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. This intervention isolates the contribution of each neuron to token-level predictions. We quantify functional specialization using the *Neuron Effect*:

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

where $\mathcal{L}$ is the token-level cross-entropy loss and $\text{LM}(x)$ is the model output. High $\Delta\text{loss}(i)$ indicates neurons with disproportionate influence on rare token prediction.

Experimental setup We sample 25,088 tokens from the C4 corpus [25], filtering for rare tokens using a two-stage procedure: (i) retaining tokens below the 50th percentile in unigram frequency, and (ii) restricting to valid, correctly spelled English words[1].

Emergence of rare token neurons Figure 1 illustrates the per-neuron influence distribution over training. We observe a heavy-tailed structure: most neurons have negligible impact, while a small group exhibits disproportionately high Δloss. We term these *rare token neurons*, further categorized as *boosting* or *suppressing* based on whether their activity increases or decreases rare token likelihood.

[1] Filtered using the `pyspellchecker` library: `https://pypi.org/project/pyspellchecker/`

3.2 Distribution and Structure of Neuron Influence

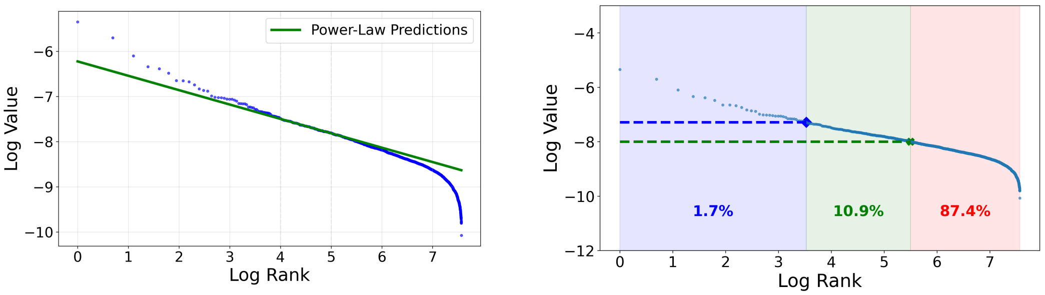
Ranking neurons by Δloss reveals a **three-regime structure** (Figure 2b):

1. **Influential plateau regime:** A small fraction of neurons consistently exhibit high influence, forming a plateau at the top of the distribution.
2. **Power-law decay regime:** Most neurons follow a scale-free power-law:

$$\log |\Delta \text{loss}| \approx -\kappa \log(\text{rank}) + \beta, \tag{6}$$

   where κ reflects the slope. This intermediate regime aligns with theoretical predictions of sparse feature extraction in overparameterized networks [20].
3. **Rapid decay regime:** The remaining neurons contribute negligibly, with influence decaying faster than power-law predictions.

This structure indicates functional specialization: a small subnetwork disproportionately handles rare token prediction. The power-law regime suggests scale-free organization characteristic of self-organized criticality in complex systems [1, 29].



(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.

3.3 Co-activation Patterns Through the Lens of Activation Space Geometry

Having identified rare-token neurons through targeted ablation experiments, we turn to a mechanistic analysis of their behavior, aiming to uncover structural principles that govern the (dis)appearance of rare tokens in model predictions. To this end, we conduct a series of geometric analysis on the activation space. Our approach is motivated by the hypothesis that the internal representations learned by language models encode information in geometrically meaningful ways—such that certain geometric structures (e.g. vectors,subspaces or manifolds) are responsible for particular semantic representations [23, 24].

Construction of the activation space To understand how rare token neurons function collectively, we construct high-dimensional vectors comprising of activations of a certain neuron in response to the selected context-token pairs from the C4 corpus [25].

Two geometric statistics for co-activation detection We hypothesize that rare token neurons do not act in isolation, but instead participate in coordinated subspaces to modulate token-level probabilities. To this end we introduce two statistics in the activation space to measure the potential coordination patterns.

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:

$$d_{\text{eff}} = \min \left\{ d : \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{j=1}^{N} \lambda_j} \geq \tau \right\}, \tag{7}$$

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

The second statistics is the *pairwise cosine similarity* between 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:

$$\cos(\theta_{ij}) = \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|}. \tag{8}$$

Clustering and activation correlations To investigate whether rare token neurons exhibit clustered activation patterns, we compute pairwise correlations of their activations across the selected context-token pairs. 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:

$$D_{ij} = 1 - |\rho_{ij}|, \tag{9}$$

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.

3.4 Weight Eigenspectrum and the Emergence of Functional Substructures

To better understand how rare-token-influential neurons acquire their specialization, we analyze the evolution of weight eigenspectra across training steps. This perspective allows us to track how the model progressively develops functional differentiation, linking representational geometry with weight-space organization.

Our analysis is grounded in the framework of *Heavy-Tailed Self-Regularization* (HT-SR) theory, introduced in Section 2.2. HT-SR predicts that during training, the emergence of heavy-tailed structures in weight matrices reflects effective feature learning: directions in weight space with larger variance correspond to dimensions carrying strong learned signals, while lighter-tailed directions capture noise or redundant patterns. From this perspective, neuron groups exhibiting heavier-tailed eigenvalue spectra are expected to encode richer, more structured computations.

Formally, for a given neuron group $\mathcal{G}$, we extract the corresponding slice of the weight matrix $\mathbf{W}_{\mathcal{G}} \in \mathbb{R}^{|\mathcal{G}| \times d}$ and compute its correlation matrix:

$$\mathbf{\Xi}_{\mathcal{G}} = \frac{1}{d} \mathbf{W}_{\mathcal{G}} \mathbf{W}_{\mathcal{G}}^{\top}. \tag{10}$$

We then analyze the eigenvalue distribution $\{\lambda_i\}$ of $\mathbf{\Xi}_{\mathcal{G}}$ to characterize the internal dimensionality and structure of the group's learned representations.

To quantify the shape of the eigenspectrum, we estimate the power-law exponent α_{Hill} using the Hill estimator:

$$\alpha_{\text{Hill}} = \left[\frac{1}{k} \sum_{i=1}^{k} \log \left(\frac{\lambda_i}{\lambda_k} \right) \right]^{-1}, \tag{11}$$

where k is a tunable parameter that determines the truncation point for tail estimation. Following prior work on pruning via spectral statistics [15, 32], we adopt the Fix-finger method to automatically set k, aligning the truncation point λ_{min} with the peak of the empirical spectral density.

Tracking α_{Hill} across steps provides insight into the emergence of functional specialization. Lower α_{Hill} values (heavier tails) indicate that a neuron group's weight directions concentrate more variance into a few dominant modes, consistent with coordinated feature learning. Conversely, higher α_{Hill} values (lighter tails) suggest weaker specialization and less structured computations.

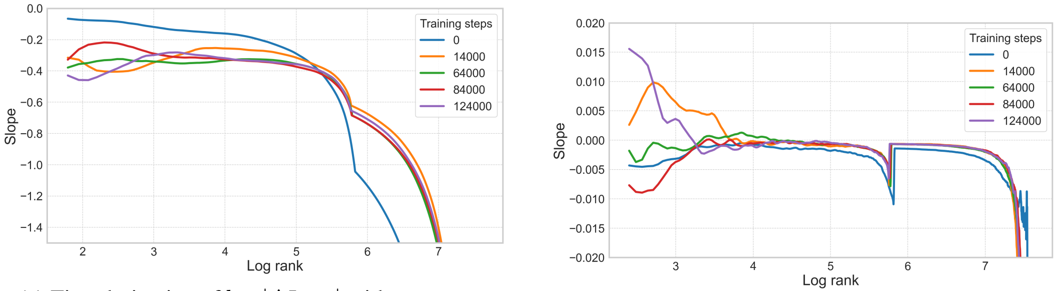
4 Results

4.1 Three-regime Structure in Neuron Influence

Ranking neurons by their ΔLoss reveals a consistent three-regime structure presented in log-log scale. Such regime distinction is observed across model scales and architectures (Figure 2b; more results in Figure 6). This structure suggests a functional specialization composed of: i.) **Influential plateau regime** where a small fraction (1.7%) of neurons exhibit consistently exponentially larger 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

$$\log |\Delta\text{Loss}| \approx -\kappa \log(\text{rank}) + \beta, \tag{12}$$

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.



(a) First derivative of $\log|\Delta\text{Loss}|$ with respect to $\log(\text{rank})$ for Pythia-410M. The plot highlights the transition between the power-law regime and the rapid decay regime, indicating the functional boundary between moderately and minimally influential neurons.

(b) Second derivative of $\log|\Delta\text{Loss}|$ versus $\log(\text{rank})$. Discontinuities mark structural transitions between regimes, providing additional evidence for the emergence of the three-regime organization of rare-token neuron influence.

Figure 3: Derivatives of loss contribution with respect to loss rank, in $\log - \log$ coordinates

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, we estimate the local slope $\kappa(r)$ with a sliding window approach. As shown in Figure 3a, 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.

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 4 shows that *the highly-influential plateau emerges progressively during training*—implying functional specialization through training.

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)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.

Emergence of the Plateau Regime In Figure 4, we illustrate the dynamics of $\delta(r)$ through the training process. It shows that *the plateau phase emerges progressively during training*. The deviation is most pronounced for highest-ranked neurons and develops gradually as training proceeds, becoming increasingly significant in the later training stages. This evolution demonstrates a process of progressive functional differentiation, where a small subset of neurons gains disproportionate influence beyond what would be predicted by the power-law relationship.

Notably, these plateau-phase neurons maintain slope characteristics similar to those in the power-law phase but operate at a higher baseline level of influence. As training proceeds, they acquire an additional positive bias term that causes systematic deviation from power-law scaling. The temporal development of these phases indicates that language models progressively form a specialized neuron subnetwork for rare token processing.

4.2 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.

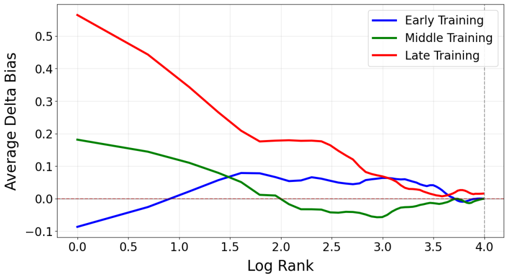


Figure 4: Dynamics of deviation $\delta(r)$ between observed neuron influence and power-law predictions across training. The plot illustrates the progressive emergence of the influential plateau phase: top-ranked neurons gradually gain disproportionate influence beyond the power-law expectation. These plateau-phase neurons maintain similar slope characteristics to the power-law regime but operate at a higher baseline, indicating the formation of a specialized subnetwork for rare-token processing over the course of training.

To further investigate this structure, we construct high-dimensional activation vectors for each neuron using context-token pairs from the C4 corpus [25]. 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 ≈ 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.

4.3 Weight Eigenspectrum

To investigate how the network progressively develops functional differentiation, we apply Heavy-Tailed Self-Regularization (HT-SR) theory [19, 20] 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.

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.

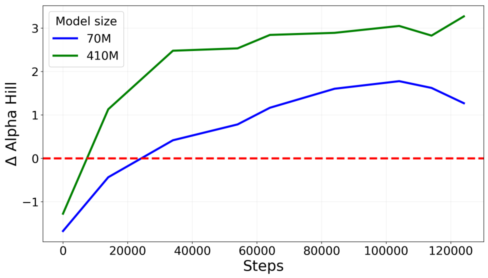


Figure 5: Hill-estimator α_{Hill} values across training for rare-token-influential neurons versus random neurons. Lower α_{Hill} (heavier tails) indicates stronger functional specialization. Rare-token neurons consistently develop heavier-tailed weight distributions, consistent with HT-SR theory.

Figure 5 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.

5 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 5.1 (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 5.2 (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 [21, 13], 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.

6 Limitation

Identification of regimes Our current framework for detecting the three-regime structure relies on ablation-based proxies, such as neuron-wise changes in token-level loss. While these measures capture statistical influence, they do not fully establish functional specialization. Future work should investigate whether influential neurons across different model families co-activate in rare-token-related contexts, which would provide stronger evidence that these neurons are genuinely specialized rather than statistically salient.

Neuron coordination within regimes The observed coordination effects, such as activation correlations among rare-token neurons, are modest in magnitude. Although correlation values indicate weak-to-moderate coupling, their consistency across multiple models, metrics, and statistical tests suggests systematic organization beyond random expectation. This likely reflects the inherently distributed computation in transformer architectures, where even modest coordination across a large neuron population can generate significant functional effects. More sophisticated measures—e.g., mutual information, causal interventions, or network-level connectivity analyses—could provide a more rigorous assessment of functional specialization.

Scope limited to the final MLP layer Our analysis focuses exclusively on neurons in the last MLP layer, where feature integration directly impacts token prediction. Rare-token processing, however, may involve interactions across multiple layers and attention mechanisms. Extending the analysis to earlier MLP layers, attention heads, and residual streams would offer a more comprehensive understanding of how LLMs coordinate distributed and specialized processing for low-frequency tokens.

Applicability to downstream tasks We evaluate specialization in the context of next-token prediction on language modeling corpora. The practical impact of these rare-token neurons on downstream applications—such as question-answering, reasoning, or domain-specific generation—remains untested. Future work should examine whether the identified subnetworks meaningfully contribute to task performance and whether interventions on these neurons can improve model efficiency or controllability in applied settings.

References

[1] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.*, 59:381–384, 1987.

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

[3] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. B. Van Den Driessche, J.-B. Lespiau, B. Damoc, A. Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240. PMLR, 2022.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

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

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

[7] E. Dohmatob, Y. Feng, P. Yang, F. Charton, and J. Kempe. A tale of tails: Model collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024.

[8] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, Z. Sui, W. Liu, Y. Yang, et al. A survey of in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[9] W. Gurnee, A. Raghunathan, and N. Nanda. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

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

[11] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[12] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR, 2023.

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

[14] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[15] H. Lu, Y. Zhou, S. Liu, Z. Wang, M. W. Mahoney, and Y. 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.

[16] S. Mallen, J. Hou, E. Wallace, M. Dredze, and N. Hegde. Not all knowledge is created equal: Tracking the impact of memorization across pre-training and fine-tuning. *arXiv preprint arXiv:2310.02173*, 2023.

[17] C. D. Manning, K. Clark, J. Hewitt, U. Khandelwal, and O. Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054, 2020.

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

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

[20] C. H. Martin and M. 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.

[21] J. L. McClelland, B. L. McNaughton, and R. 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.

[22] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[23] K. Park, Y. J. Choe, and V. Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv:2311.03658v2*, 2024.

[24] K. Park, Y. J. Choe, Y. Jiang, and V. Veitch. The geometry of categorical and hierarchical concepts in large language models. *arXiv:2406.01506v3*, 2025.

[25] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. 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.

[26] A. C. Schapiro, N. B. Turk-Browne, M. M. Botvinick, and K. 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.

[27] A. Stolfo, B. Wu, W. Gurnee, Y. Belinkov, X. Song, M. Sachan, and N. Nanda. Confidence regulation neurons in language models. *Advances in Neural Information Processing Systems*, 37:125019–125049, 2024.

[28] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[29] N. W. Watkins, G. Pruessner, S. C. Chapman, N. B. Crosby, and H. J. Jensen. 25 years of self-organized criticality: Concepts and controversies. *Space Science Reviews*, 198:3–44, 2016.

[30] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

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

[32] Y. Yang, R. Theisen, L. Hodgkinson, J. E. Gonzalez, K. Ramchandran, C. H. Martin, and M. 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.

[33] C. Zhang, G. Almpanidis, G. Fan, B. Deng, Y. Zhang, J. Liu, A. Kamel, P. Soda, and J. Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[34] G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, 1949.

A Appendix

A.1 Results

Activation Correlation Analysis We examined activation patterns across neuron groups in five pre-trained language models, comparing rare token neurons (boost and suppress groups) against random baseline controls. Table 1 presents pairwise activation correlations within and between neuron groups (group size = 50).

The results demonstrate consistently higher intra-group correlations for both boost neurons (range: 0.004–0.036) and suppress neurons (range: 0.011–0.052) compared to random controls (range: -0.007–0.017). Notably, cross-group correlations between boost and suppress neurons (B vs. S) show positive values (0.010–0.040), suggesting coordinated but potentially antagonistic functionality. Random baseline comparisons (R1 vs. R2) exhibit near-zero correlations, confirming the specificity of our identified neuron groups.

Table 1: Pairwise activation correlations within and between neuron groups

Model	Size	Boost	Suppress	Random	B vs. R	S vs. R	B vs. S	R1 vs. 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

Effective Dimensionality We computed the effective dimensionality of activation patterns using the participation ratio metric to assess functional specialization. Table 2 shows that rare token neuron groups consistently exhibit lower effective dimensionality compared to random baselines. This reduction in effective dimensions (boost: 33.0–43.0%, suppress: 32.2–43.0%) relative to random controls (36.2–46.0%) indicates more concentrated, specialized activation patterns within rare token neurons.

Table 2: Effective dimensionality proportions across neuron groups

Model	Size	Boost	Suppress	Random
Pythia-70M	70M	33.5	32.6	36.2
Pythia-410M	410M	33.0	32.2	37.3
GPT2-Small	124M	37.0	40.0	45.0
GPT2-Large	774M	43.0	43.0	46.0
GPT2-XL	1.5B	40.0	42.0	46.0

Cosine Similarity of Weight Vectors Weight vector cosine similarities provide insight into the geometric organization of rare token neurons in parameter space. Table 3 reveals strong positive alignment within boost (0.028–0.141) and suppress (0.092–0.165) neuron groups, while showing consistent negative alignment between these specialized groups and random controls. The negative cross-correlations (boost vs. random: -0.100 to 0.003, suppress vs. random: -0.105 to -0.014) suggest that rare token neurons occupy distinct regions of weight space, supporting our hypothesis of functional specialization.

Weight Distribution Analysis We analyzed the heavy-tailed properties of weight distributions using power-law exponents (α) estimated via the Hill estimator. Table 4 demonstrates that rare token neuron groups exhibit significantly lower α values compared to random controls, indicating heavier-tailed weight distributions. This finding supports our hypothesis that functional specialization correlates with the emergence of heavy-tailed statistical properties in neural networks.

Phase Transition Dynamics Figure 6 illustrates the evolution of neuron influence distributions throughout training across the GPT-2 model family. The suppress neuron populations exhibit

Table 3: Cosine similarity between weight vectors within and across neuron groups

Model	**Size**	**Boost**	**Suppress**	**Random**	**B vs. R**	**S vs. R**	**B vs. S**	**R1 vs. 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

Table 4: Power-law exponents (α) for weight distributions across neuron groups

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

characteristic three-phase development: an initial plateau phase with uniform low influence, followed by a power-law scaling regime, and concluding with rapid decay at high influence values. This pattern emerges consistently across model scales, suggesting a universal mechanism underlying rare token neuron specialization.

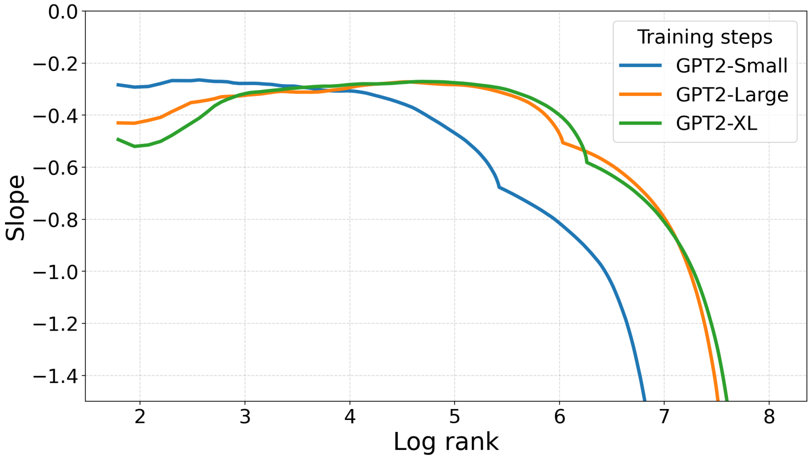


Figure 6: Distribution of neuron influence slopes for suppress neurons across the GPT-2 model family, showing characteristic three-phase organization emerging during training.