Distributed Specialization: How Transformers Process Rare Tokens Through Parameter Differentiation

Large language models (LLMs) often underperform on **rare token prediction**, even though such tokens carry critical information in specialized domains [1]. Prior work has identified **frequency-sensitive neurons** that modulate token logits [2], but the organizational principles underlying these neurons' functional roles remain poorly understood. In this work, we analyze how transformers self-organize computation for rare versus common tokens, conducting a systematic neuron-level study across GPT-2 and Pythia model families. Our approach combines **neuron ablation**, **activation-space analysis**, and **weight-space spectral characterization** to reveal how functional specialization emerges dynamically during training.

Neuron Influence Analysis and Training Dynamics. We measure each neuron's contribution via the **loss increase after ablation** for the MLP layer before the unembedding layer and uncover a **three-regime structure** in influence distributions for rare token prediction: (1) a **specialist plateau** of highly influential neurons, (2) a **power-law regime** of moderately contributing neurons, and (3) a **rapid-decay regime** of minimally relevant neurons. In contrast, common tokens exhibit only the power-law and decay regimes, demonstrating that rare tokens recruit additional high-influence neurons. Tracking model states during training shows that these plateau neurons emerge progressively, suggesting spontaneous functional differentiation rather than pre-defined specialization.

Distributed Specialization vs. Modularity. We next ask whether specialist neurons form **modular clusters** or operate within a **distributed computation** framework. Using graph-based community detection [3] in activation space, we find that these neurons are **spatially dispersed** across the last MLP layer, with modularity scores close to random baselines (Q = 0.05–0.11 vs. control Q = 0.04–0.09, $p > 0.4$). Attention-head ablation further shows that rare token processing **does not depend on specialized routing**: removing individual heads yields minimal performance impact (~7–8%), whereas **full-layer ablation** causes large drops (~42–45%). Despite their spatial distribution, specialist neurons exhibit **structured co-activation** and **reduced effective dimensionality**, revealing coordinated computation without topological modularity.

Weight Space Analysis. To understand how specialization emerges, we analyze neuron weight spectra using **Heavy-Tailed Self-Regularization theory** [4]. Rare token neurons develop **heavier-tailed weight distributions** than controls (Hill α = 1.57–4.30 vs. 6.37–9.33), indicating **stronger functional specialization**. This spectral separation grows throughout training, suggesting that **implicit regularization mechanisms** drive the observed differentiation.

Our findings reveal that transformers achieve rare token specialization through **distributed parameter differentiation** rather than modular organization. By recruiting additional high-influence neurons that are **spatially dispersed yet functionally coordinated**, models preserve **context-sensitive flexibility** while allocating adaptive capacity for rare token processing. These results challenge modular views of neural computation and provide insights for **interpretable model editing** and **efficiency optimization**.

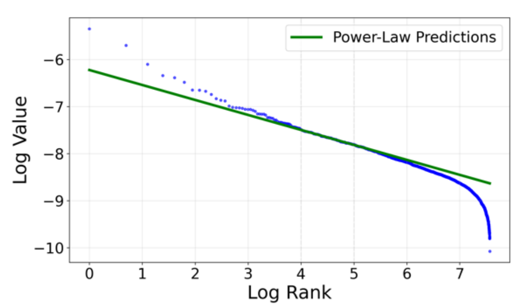

Figure Neuron influence distributions in the Pythia-410M model illustrating a three-regime structure for rare token prediction: a specialist plateau of high-impact neurons, a power-law regime of moderately contributing neurons, and a rapid-decay regime of minimal influence.

References: [1] Kandpal et al. Large language models struggle to learn long-tail knowledge. *ICML* 2023. [2] Stolfo et al. Confidence regulation neurons in language models. *NeurIPS* 2024. [3] Blondel et al. Fast unfolding of communities in large networks. *Journal of statistical mechanics*. 2008. [4] Martin & Mahoney. Implicit self-regularization in deep neural networks. *JMLR* 2021.