From Tokens to Semantics: The Emergence and Stabilization of Polysemanticity in Language Models

Sharvil Limaye[*]
Rutgers University

Aniruddhan Ramesh[*]
University of Cincinnati

Aiden Zhou[*]
Yale University

Akshay Bhaskar
University of Michigan

Jonas Rohweder
TU Darmstadt
Zuse School ELIZA

Ashwinee Panda
Algoverse AI Research

Vasu Sharma
Algoverse AI Research

Abstract

Polysemanticity—neurons activating for seemingly unrelated features—has long been viewed as a key obstacle for interpretable AI. We show instead that it follows a structured, hierarchical developmental trajectory, offering a principled perspective on how networks allocate scarce representational capacity. We present three interdependent analyses of Pythia models of different sizes across training checkpoints: clustering of top-activating excerpts, Jensen–Shannon divergence over frequency buckets, and a geometric characterization (polytope density and participation ratio). First, we trace representational dynamics over training: early layers encode token- and frequency-specific signals, with high- and low-frequency n-grams occupying distinct regions of activation space that mostly re-converge over training; deeper layers—and larger models—progressively shift toward representations that are invariant to token frequency and organized by semantic content. Second, we identify a coverage principle: neuron coverage (the fraction of positions in which a neuron participates), not raw frequency preference, predicts specialization. High-coverage neurons specialize, while low-coverage neurons remain generalists. Third, we observe that activation manifolds transition from fragmented to consolidated. Together, these results recast polysemanticity not as a static nuisance, but as a structured, evolutionary process that distributes scarce capacity efficiently and abstracts towards meaning. Our code is available at https://github.com/sharvillimaye/from-tokens-to-semantics.

1 Introduction

In a perfectly interpretable neural network, every neuron would represent a unique and understandable feature. Yet, constructing such models beyond toy sizes has proven to be a hard task, even when one enforces a sparse architecture [3]. This is because neurons often encode (or activate strongly to) a set of unrelated features—a phenomenon known as polysemanticity [19, 18]. As stated by the *superposition* explanation, this is probably to avoid the storage concerns of a purely linear architecture, which can only represent as many features as it has neurons [8].

So why is this a concern? It is not only that predicting network behavior is harder when the basis of representations is non-standard. Adversarial techniques can target shared topological structures (arising out of superposition) even when model internals are black boxes, presenting an AI safety risk [10]. Recent work has also diagnosed superposition as a key factor underlying observed neural scaling laws[13]. Thus, frontier models are by nature vulnerable to these consequences.

[*] Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mech Interp.

Past work on the topic has focused on disentangling representations in neural networks. One popular perspective is to use sparse autoencoders (SAEs) or transcoders to extract interpretable features [4, 11, 6]. Alternatively, O'Mahony et al. [20] trained linear classifiers for concept discovery, and Dreyer et al. [5] sought to extract distinct circuits that match a "pure feature."

Yet, no existing work tracks the evolution of feature superposition. This paper strives to fix that gap, providing a longitudinal, multi-scale geometric account of representation consolidation during model pretraining. First, we present a study of how neuron behaviors evolve during pretraining, using cluster geometry as a proxy for polysemanticity. We discover consistent trajectories in neuron specialization that appear to be shared across model sizes. To enhance clarity, we also release an interactive web app that displays the highest-activating text clusters of each neuron over 15 checkpoints for Pythia-70M and Pythia-160M [1].[2] Next, to test theories that input frequency drives polysemanticity, we curate a dataset of country n-grams with varying corpus frequencies, and apply Jensen-Shannon divergence to track the evolution of frequency groups during pretraining. Lastly, we characterize the shape of activation spaces via polytope analysis, revealing how representational geometry transitions from fragmented token-level to consolidated semantic manifolds.

2 Related Works

Neuron-Level Interpretability. Beyond sparse autoencoders, linear probes, and circuit analyses, researchers have developed additional tools for probing individual neurons. Recent work on continual sparse autoencoders, such as the SAE-Track framework [22], has shown that feature representations reorganize substantially across training checkpoints: some features emerge, split, or vanish, while others shift from token-level patterns to more semantic abstractions. Complementary approaches, including neuron embeddings and clustering of top-activating text excerpts, extract interpretable semantic features from high-activation contexts [9]. While these methods excel at characterizing learned representations and have begun to illuminate feature dynamics, they focus primarily on feature-level changes rather than on how neurons systematically transition from encoding multiple unrelated features to more specialized representations across the network.

Frequency and Statistical Analyses. Token frequency and co-occurrence statistics remain central to how language models internalize linguistic structure [21, 17]. Tools like infini-gram [12] enable efficient corpus frequency retrieval, while studies demonstrate that frequency-based regularities dominate early-layer representations [21] and that transformers exhibit sensitivity to token statistics at multiple granularities [17]. However, existing analyses focus primarily on static frequency effects rather than probing how frequency interacts with representational geometry, neuron specialization, or how these relationships evolve across training checkpoints and model scales.

Geometric Interpretability Frameworks. The polytope lens framework [2] advanced geometric analysis of neural networks by proposing that activation spaces be understood through convex polytope structures rather than individual neurons. Polytopes naturally emerge in ReLU networks as regions of identical affine transformations, providing principled geometric foundations for interpretability. Despite their analytical power, geometric lenses have not been systematically applied to study temporal dynamics or developmental patterns in language model training.

Neuron-level probes, corpus statistics, and geometric lenses elucidate complementary aspects of representation, yet each leaves open the question of *dynamics*. Our work unifies these threads into a longitudinal framework that explains how polysemanticity evolves with training and scale.

3 Methods

3.1 Tracking Feature Clusters over Pretraining

Neuron embeddings is a metric that preserves, for a given neuron and input, the feature directions that drive its activation [9]. Let $h^{(l-1)}(x) \in \mathbb{R}^d$ be the representation of input x before layer l, and let $w_k^{(l)} \in \mathbb{R}^d$ be the input weight vector of neuron k in layer l. Then, the neuron embedding is

$$e_k^{(l)}(x) \;=\; h^{(l-1)}(x) \odot w_k^{(l)}, \tag{1}$$

[2] `https://modelevolution.streamlit.app`

where $\odot$ is the Hadamard product. For a dataset, iteratively applying this operation generates an *embedding space* in which a notion of distance can be defined. Given any data points x_i, x_j, let

$$d(x_i, x_j) \; = \; 1 - \cos\big(e_k^{(l)}(x_i),\, e_k^{(l)}(x_j)\big). \tag{2}$$

This procedure enables us to sort a set of highly activating excerpts into clusters of high density and separation—each representing a clear-cut feature [9]. **Thus, cluster statistics (e.g., average clusters per neuron) can serve as proxy metrics for a neuron's polysemanticity.** Our paper expands on past work by iterating over checkpoints, thereby revealing novel evolutionary trends.

At each checkpoint in {3,000, 13,000, ..., 143,000} for Pythia-70M, Pythia-160M, and Pythia-410M, 30,000 excerpts of WikiText-2 were digested [14]. The first 100 (or, for very specialized neurons, any) for which the activation of neuron k exceeded $0.6 \cdot \mathrm{max\_act}_k$ (where $\mathrm{max\_act}_k$ is the max activation of k as reported by Neuroscope [15]) were kept and clustered using hierarchical agglomerative clustering with a distance threshold of $\tau = 0.75$.[3] Neuron activations were extracted using the TransformerLens library [16].

To balance cost concerns with coverage, we subsample evenly by index: for Pythia-70M we select every 20th neuron and for Pythia-160M every 60th. This yields ∼600 neurons per model and covers every layer. For Pythia-410M, we analyze Layer 0 only, selecting every 200th neuron.

Although SAEs provide higher resolution and diverse use cases, they are expensive to train and can be prohibitive for longitudinal analyses across multiple checkpoints and layers [4]. Our approach trades fine-grained semantic decomposition for computational feasibility, enabling efficient tracking of polysemanticity dynamics at scale. This framework proves effective for studying the emergence and behavioral trajectories of polysemanticity over training, complementing SAE-based interpretability work that focuses on detailed feature analysis at individual checkpoints.

3.2 n-gram Selection and Frequency Grouping

We index unigrams–trigrams ($n \in \{1, 2, 3\}$) in the deduplicated, preshuffled THE PILE using `tokengrams` [7]. To isolate frequency from semantics, we restrict to one semantic family (e.g., capital, country) and template prompts so the same entity appears in multiple n-gram forms. At each checkpoint s, we compute cumulative frequency $f_s(p)$ (raw and per million), bin phrases into eight empirical buckets, and—unless noted—contrast b=0 vs. b=7. This holds semantics roughly fixed while varying only frequency, **so any downstream differences we measure can be attributed to frequency rather than meaning.**

3.3 Activation Collection

We probe Pythia models at multiple scales across fixed training checkpoints. Each phrase p is embedded into templated sentences:

"The capital of {phrase} is",
"The people of {phrase} speak"

and the *anchor token* is the final token of the phrase. For each layer L at step s, we extract the anchor token's post-activation MLP input, yielding $a_{s,L}(p) \in \mathbb{R}^H$ for hidden size H. We normalize activations into probability distributions,

$$P_{s,L}(h \mid p) = \frac{\max\{a_{s,L}(p)_h, 0\}}{\sum_{j=1}^{H} \max\{a_{s,L}(p)_j, 0\} + \varepsilon}, \tag{3}$$

with $\varepsilon = 10^{-12}$, interpreting each phrase as allocating probability mass across neurons. **These activations give us a unified quantitative view of how phrases are routed through the network, enabling the study of representation and specialization across layers and stages of training.**

[3] The value of τ was chosen after systematic hyperparameter tuning, with the chosen value yielding the best clustering results. Reasonable adjusting of τ does not change the trends we observe in Figures 1–3.

3.4 Affinity and Coverage Metrics

To track how neurons specialize as training progresses, we need to measure two distinct properties: (1) **how much** a neuron participates across different phrases, and (2) **what frequency bias** it exhibits. *Coverage* predicts whether a neuron will specialize; *affinity* characterizes whether it prefers common or rare items.

More formally, let $\log f_s(p)$ denote the log frequency of phrase p at training step s. For each neuron h in layer L at step s, we define three complementary measures:

$$\textbf{(Participation coverage)} \quad \tilde{C}_{s,L}(h) = \sum_{p \in \cup_b S_b} P_{s,L}(h \mid p), \tag{4}$$

$$\textbf{(Mean affinity)} \quad \bar{A}_{s,L}(h) = \frac{A^{\text{sum}}_{s,L}(h)}{\tilde{C}_{s,L}(h) + \varepsilon} \; = \; \mathbb{E}_{p \sim P(\cdot \mid h)}\big[\log f_s(p)\big], \tag{5}$$

$$\textbf{(Sum affinity)} \quad A^{\text{sum}}_{s,L}(h) = \sum_{p \in \cup_b S_b} P_{s,L}(h \mid p) \, \log f_s(p). \tag{6}$$

Coverage $\tilde{C}$ measures the total mass a neuron allocates across all phrases in our dataset. It ranges from 0 (neuron never fires above threshold) to the total number of phrases (neuron fires for every phrase). Mean affinity $\bar{A}$ measures the *average* log frequency of phrases that activate this neuron, indicating whether the neuron tends toward common (high $\bar{A}$) or rare (low $\bar{A}$) items. Sum affinity A^{sum} blends these two factors: it captures both coverage and affinity jointly.

The key insight is that $A^{\text{sum}} = \tilde{C} \cdot \bar{A}$, coverage and affinity are disentangled. By separately tracking coverage and mean affinity, we can test whether neurons specialize based on *frequency preference* (a high mean affinity signal would suggest this) or based on *usage breadth* (a coverage signal). **These metrics highlight how strongly each neuron participates and whether it tends toward high- or low-frequency phrases.**

3.5 Group distributions and Jensen–Shannon divergence

Having established how individual neurons encode frequency information, we now ask a complementary question at the **population level**: *Do high-frequency and low-frequency phrases activate different subsets of neurons?*

To answer this, we partition phrases into frequency buckets and measure how distinctly each frequency group activates the neural population. This is formalized using Jensen-Shannon divergence, which quantifies the statistical distance between two probability distributions. For any bucket b, define the average per-neuron firing distribution

$$\bar{P}^{(b)}_{s,L}(h) \; = \; \frac{1}{|S_b|} \sum_{p \in S_b} P_{s,L}(h \mid p), \qquad S_b = \{p : \, p \text{ in bucket } b\}. \tag{7}$$

We measure separability of two buckets b_0, b_1 with the Jensen–Shannon divergence $\mathrm{JSD}_{s,L}(b_0, b_1)$ and a mixture distribution defined as $M_{s,L}$:

$$M_{s,L} \; = \; \tfrac{1}{2}\Big(\bar{P}^{(b_0)}_{s,L} + \bar{P}^{(b_1)}_{s,L}\Big), \tag{8}$$

$$\mathrm{JSD}_{s,L}(b_0, b_1) \; = \; \tfrac{1}{2}\,\mathrm{KL}\Big(\bar{P}^{(b_0)}_{s,L} \,\Big\|\, M_{s,L}\Big) + \tfrac{1}{2}\,\mathrm{KL}\Big(\bar{P}^{(b_1)}_{s,L} \,\Big\|\, M_{s,L}\Big). \tag{9}$$

Because the distributions are discrete over neurons, we also track a per-neuron contribution:

$$c_{s,L,h}(b_0, b_1) = \tfrac{1}{2}\,\bar{P}^{(b_0)}_{s,L}(h)\,\log\frac{\bar{P}^{(b_0)}_{s,L}(h)}{M_{s,L}(h)} \; + \; \tfrac{1}{2}\,\bar{P}^{(b_1)}_{s,L}(h)\,\log\frac{\bar{P}^{(b_1)}_{s,L}(h)}{M_{s,L}(h)}, \tag{10}$$

which satisfies $\sum_h c_{s,L,h}(b_0, b_1) = \mathrm{JSD}_{s,L}(b_0, b_1)$.

Here, $\bar{P}^{(b)}_{s,L}$ encodes how each bucket spreads its activation mass across neurons in a layer and JSD quantifies how distinct those patterns are: its value is 0 only if buckets excite neurons identically and increases as they diverge. The vector $c_{s,L,h}$ breaks this difference down neuron by neuron, flagging the cells that are most discriminative for one group versus another. **This highlights separability, indicating whether rare and common phrases live in overlapping circuits or are routed to different subnetworks.**

3.6 Polytope Analysis

To study the evolution of neural representations during training, we add a temporal dimension to the polytope lens framework. Each phrase is represented by its firing distribution, and collections of phrases form convex polytopes in activation space. We analyze how the geometry of these polytopes changes with training and frequency, focusing on two complementary metrics. **Polytope density together with participation ratio captures the geometry of consolidation—how activation manifolds compress and align over training toward frequency-invariant, semantic representations.**

Polytope Density. We measure the compactness of representations by comparing distances between activations. For a random pair of activations (i, j):

$$d_{\text{eucl}}(i, j) = \|A_i - A_j\|_2, \tag{11}$$

$$d_{\text{ham}}(i, j) = \text{Hamming}(A_i, A_j), \tag{12}$$

$$\text{Density}(i, j) = \frac{d_{\text{ham}}(i, j)}{d_{\text{eucl}}(i, j)}. \tag{13}$$

High density indicates that activations are similar in Euclidean geometry but differ in sparse support patterns, suggesting more entangled representations.

Participation Ratio We quantify the effective dimensionality of activations via principal component analysis. Given eigenvalues $\{\lambda_k\}$ of the covariance matrix of activations, let the

$$\text{Participation Ratio} = \frac{\left(\sum_k \lambda_k\right)^2}{\sum_k \lambda_k^2}. \tag{14}$$

A higher participation ratio indicates that variance is distributed over many directions, while a lower participation ratio suggests collapse into a lower-dimensional subspace.

4 Results

4.1 Temporal Dynamics of Polysemanticity

Across model sizes, neurons share a consistent trajectory: an *exploratory* regime with rising cluster counts (many features multiplexed per unit), superseded by *consolidation* into fewer clusters, and culminating in *stabilized* representations that persist over late checkpoints. Larger models exhibit more pronounced early exploration and stronger eventual consolidation, as presented in Figure 1.

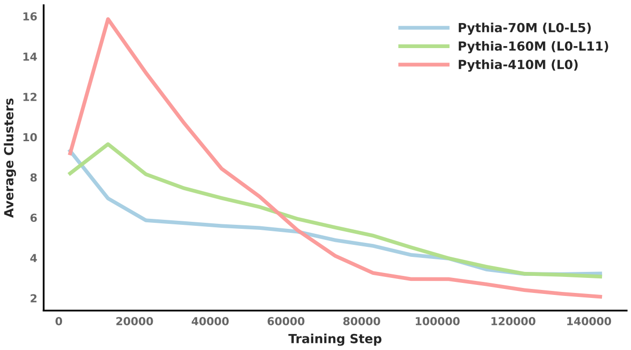


Figure 1: **Scale effects on polysemanticity dynamics.** Larger models exhibit higher polysemanticity early in training (spiking behavior) but also stronger convergence (low plateaus).

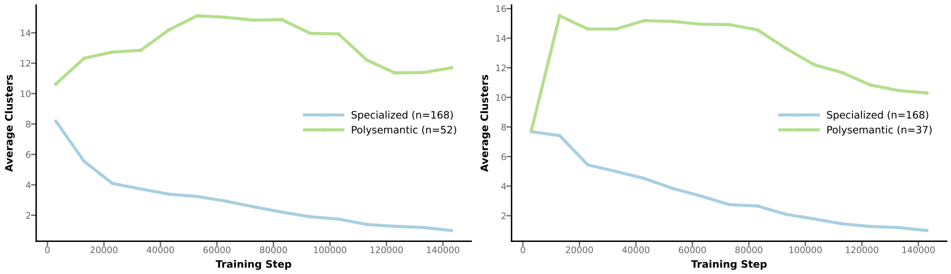


Figure 2: **Cluster trajectory by neuron type** (specialized if clusters $= 1$ at last checkpoint and highly polysemantic if clusters ≥ 8 at last checkpoint [4]). Pythia-70M (left) and Pythia-160M (right).

Next, stratifying by neuron type (Figure 2) underscores distinct trajectories: specialized neurons converge regularly to a unique feature, while highly polysemantic neurons exhibit early spikes and gradual consolidation—echoing the trend displayed in Figure 1. This observation aligns with the superposition hypothesis that networks allocate different computational strategies based on feature frequency and importance [8].

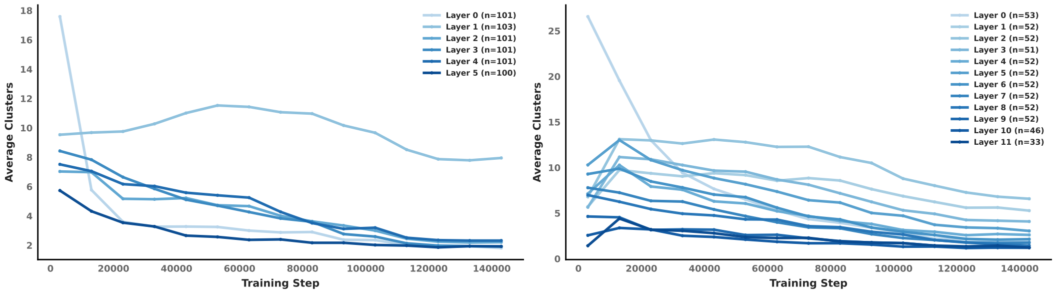


Figure 3: **Layer-wise stratification.** For Pythia-70M (left) and Pythia-160M (right), Layer 0 has sharp spikes and fast convergence, early layers display persistent elevation, and later layers exhibit steady consolidation. This motivates our study of hierarchical layer organization in **Section 4.3**.

Case studies of cluster evolution over checkpoints are available in the Appendix (Figs. 9, 10), and can be generated for 1200+ neurons in Pythia-70M and Pythia-160M using the web app provided.[5]

4.2 The Coverage Principle Governs Neuron Specialization

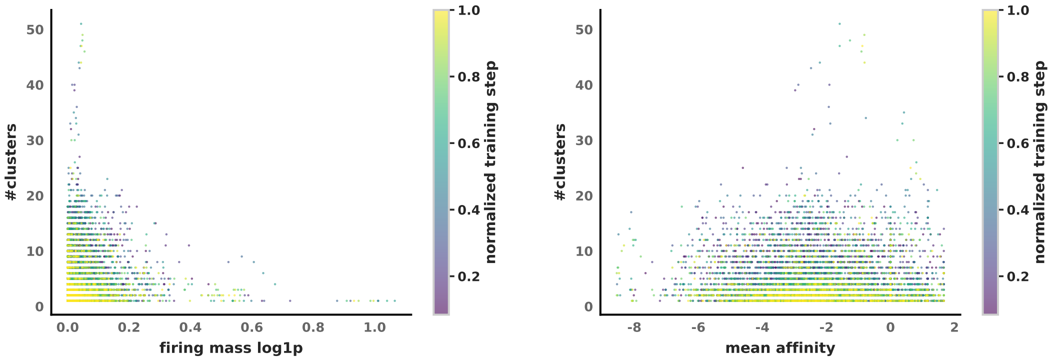
Throughout training, models allocate representational capacity along a clear trajectory: early on, low-coverage neurons multiplex rare and diverse features; as training progresses, coverage expands and features consolidate; ultimately, high-coverage neurons dominate with monosemantic, stable representations. This explore–consolidate–stabilize arc is robust across scales (70M, 160M) and checkpoints, aligning directly with superposition theory's prediction that broad features are anchored in high-capacity units.

Evidence across models confirms this principle. Across Pythia-70M and Pythia-160M, coverage shows a strong negative relationship with polysemanticity: neurons that activate frequently are far more likely to represent a single concept, while polysemantic units almost always live in the low-coverage regime (Figs. 4, 5). This pattern is stable across layers and persists even as overall capacity grows, demonstrating that specialization is driven by usage breadth, not token frequency alone.

Frequency preference adds little beyond coverage. When controlling for coverage, a neuron's bias toward high- or low-frequency phrases (mean affinity) is minimally informative ($|\rho| \lesssim 0.05$) and inconsistent across steps. Even sum-affinity, which might appear predictive, largely inherits its signal from coverage because it scales mean affinity by activation mass (Appendix Fig. 13).

[4] The criterion of "$\geq$ 8" was chosen to be significantly above the average clusters at final checkpoint for both Pythia models, but can be adjusted in any direction without changing the shape of the graph.

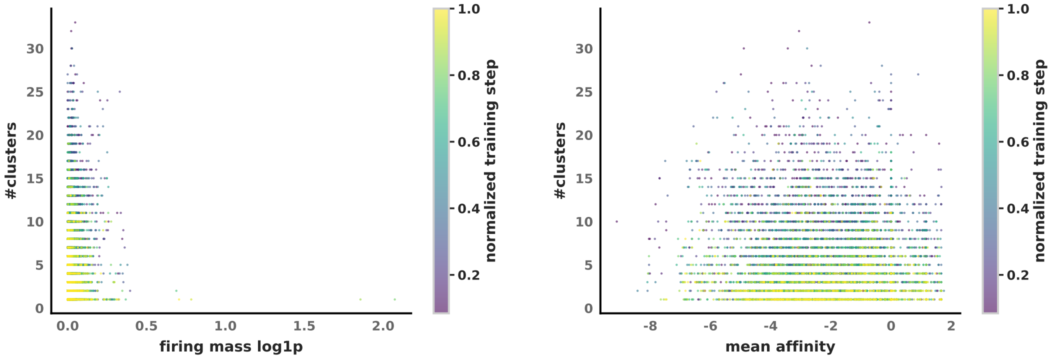
[5] `https://modelevolution.streamlit.app`



(a) Participation coverage vs. #clusters

(b) Mean affinity vs. #clusters

Figure 4: **Coverage and mean-affinity scatters (Pythia-70M).** Coverage anticorrelates with polysemanticity; mean affinity is largely uninformative.



(a) Participation coverage vs. #clusters

(b) Mean affinity vs. #clusters

Figure 5: **Coverage and mean-affinity scatters (Pythia-160M).** Pattern largely mirrors 70M: strong coverage–polysemanticity anticorrelation; weak mean-affinity signal.

Robustness checks isolate coverage as the driver. Partial correlations leave only a small residual for sum affinity, while coverage remains predictive (Appendix Fig. 14). Permutation tests confirm the causal role: shuffling coverage breaks the association with polysemanticity, but shuffling affinity does not (Appendix Fig. 15).

Takeaway. Coverage emerges as the central organizing principle: specialization is not about frequency preference but how often a neuron engages with diverse contexts. This principle will reappear in our layer-wise and geometric analyses, where it scales from individual neurons to whole representational manifolds.

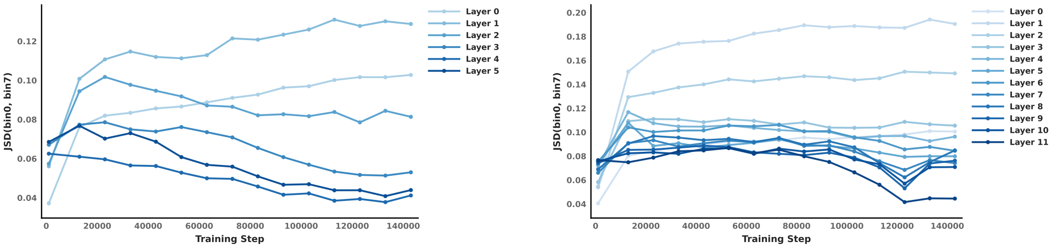
4.3 Hierarchical Layer Organization

As training unfolds, different depths play distinct roles: early blocks handle raw frequency signals, while deeper layers abstract away from surface statistics toward semantic representations. This mirrors the model's overall trajectory from exploration to consolidation.

Early layers are frequency-sensitive. Comparing activations for rare vs. common phrases (bin 0 vs. bin 7) shows that separation emerges almost immediately and peaks in the first few layers (Fig. 6). Jensen–Shannon divergence between frequency groups rises rapidly in the first 10–20k steps, indicating that shallow layers function as frequency routers, gating diverse token statistics before semantic features stabilize downstream. Beyond these early layers, divergence flattens, suggesting that mid-to-deep layers gradually suppress raw frequency differences.

Polysemantic units drive the distinction. Within these layers, the neurons that most differentiate frequency groups are also the most polysemantic. Stratifying JSD contributions by polysemanticity reveals that highly cluster-rich units dominate the divergence signal, and their influence grows stronger over training (Appendix Fig. 17). This links the frequency-routing behavior of shallow layers to the same exploratory units identified in our coverage analysis.

Coupling strengthens with training. This relationship is not static. The correlation between a neuron's frequency-separation contribution and its polysemanticity rises steadily, with several checkpoints showing statistically significant positive correlations (Fig. 7). As the network consolidates, its frequency-sensitive signals are increasingly concentrated in exploratory neurons.

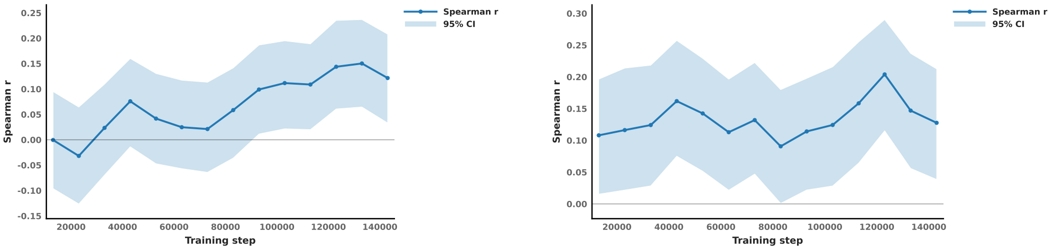


(a) Pythia-70M

(b) Pythia-160M

Figure 6: **Layerwise frequency separation.** Early layers amplify frequency-group differences, while deeper layers dampen them.

Takeaway. Frequency information is routed and transformed hierarchically: shallow layers multiplex frequency-sensitive, polysemantic units; deeper layers converge toward semantic integration. Combined with our coverage results, this suggests a unified representational story: the same low-coverage, high-polysemantic neurons that explore rare features also serve as frequency routers early on, before the network stabilizes into abstract semantic manifolds.



(a) Pythia-70M

(b) Pythia-160M

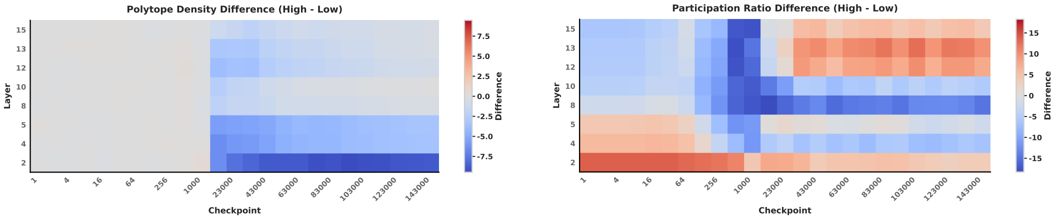
Figure 7: **Polysemantic coupling strengthens.** Correlation between frequency-separation contributions and polysemanticity grows with training.

4.4 Geometric Transformation of Representations

Activation spaces undergo a universal representational transformation from fragmented token-level manifolds to unified, frequency-transparent semantic manifolds.

Layer-wise geometric convergence pattern. Difference heatmaps (Fig. 8) reveal a systematic progression across depth. Early layers maintain strong geometric segregation between frequency groups: low-frequency n-grams exhibit higher polytope density and participation ratio, reflecting fragmented, distributed activations where rare features require broader neural engagement (consistent with exploratory multiplexing). This separation emerges after a brief warm-up where groups appear similar, suggesting networks first develop basic representational capacity.

Middle layers exhibit progressive convergence as polytope density declines and participation ratios approach each other. This transition reflects the network's shift from statistical token processing toward shared semantic abstractions. Deep layers achieve geometric unification with near-identical



(a) Polytope Density (High - Low Freq)

(b) Participation Ratio (High - Low Freq)

Figure 8: **Geometric convergence reveals hierarchical abstraction.** Difference heatmaps for country n-grams (high minus low frequency) across layers and training steps. Early layers show persistent frequency-based segregation (darker regions), indicating rare and common phrases occupy distinct geometric structures. Deeper layers show progressive convergence (lighter coloring), reflecting the network's transition from token-specific routing toward unified semantic processing that treats frequency groups more similarly.

metrics across frequency buckets, indicating frequency-transparent semantic processing where deep representations treat rare and common n-grams as members of a unified semantic space.

Quantitative patterns. Across layers and checkpoints for Pythia-70M and Pythia-1B, polytope density and participation ratio reveal three key transitions: (1) Early divergence in shallow layers where low-frequency items show higher fragmentation, evidenced by elevated density and higher participation ratios indicating distributed processing—patterns consistent with superposition-based encoding of rare features. (2) Progressive convergence in middle layers as frequency-based geometric separation diminishes: participation ratios approach similar dimensionalities while density steadily decreases, marking a shift toward shared semantic abstractions. (3) Global consolidation in deeper layers achieving near-complete geometric unification—density reaches minimal values and participation ratios stabilize across frequency groups, evidencing compression into compact, frequency-transparent semantic representations.

Interpretation. This geometric evolution substantiates the coverage principle and the global trajectory: rare n-grams initially occupy fragmented regions requiring multiple neurons (low-coverage, polysemantic, exploratory), while common patterns consolidate into concentrated representations (high-coverage, monosemantic, stabilization). With depth, both frequency groups map into shared geometries, completing the transition to frequency-transparent semantic representations. This progression explains why polysemanticity decreases over training and across depth: low-coverage, fragmented representations give way to high-coverage, consolidated ones, and frequency-based routing is eventually replaced by semantic routing.

5 Conclusion

This work establishes three principles governing polysemanticity in transformer language models. First, polysemanticity is a dynamic developmental process: networks follow a universal trajectory from early, frequency-sensitive token statistics in shallow layers toward frequency-invariant, semantic representations at depth, with consolidation stabilizing across training checkpoints. Second, coverage—not raw n-gram frequency—governs specialization: high-coverage neurons become monosemantic generalists, while low-coverage neurons multiplex many rare features, consistent with capacity-allocation predictions from superposition theory. Third, geometric analysis reveals the transition from fragmented token-level structure to consolidated semantic manifolds: polytope density and participation ratio jointly show activation spaces compressing and aligning into frequency-invariant geometries in deeper layers.

These principles connect directly to broader concerns. For interpretability, coverage provides a practical diagnostic for prioritizing neurons in analysis, attribution, and editing, with low-coverage, highly polysemantic units disproportionately mediating frequency-conditioned differences. For safety and robustness, such units represent potential attack surfaces; monitoring and potentially sparsifying low-coverage neurons may reduce distributional fragility and stabilize out-of-distribution behavior. For scaling, larger models consolidate earlier and more stably, motivating scaling curves that incorporate polysemanticity and geometric metrics (e.g., JSD, polytope density, participation ratio) alongside accuracy.

6 Limitations

There are several key caveats to our work. First, our cluster evolution analyses focused on toy models (Pythia-70M, Pythia-160M, and a partial slice of Pythia-410M). Moreover, due to cost concerns, not every neuron was probed. Our polytope studies, which were not as expensive, extended to Pythia-1B. These scales are far from frontier LLMs, and conclusions may shift with considerably larger models and datasets.

Our analyses are restricted to neuron activations within the MLP layers, which process tokens independently and act as feature memories. We do not include attention layer outputs, which mediate information flow between tokens, or residual stream representations, which integrate contributions from all components. Whether the proposed coverage principle and geometric consolidation patterns extend to these representations remains an open question. Furthermore, our analyses do not explicitly address statistical outliers or systematically extreme activation values. Understanding how these factors influence polytope density estimates and cluster structure would strengthen our conclusions. Future work should extend this framework to examine polysemanticity dynamics across all transformer components and their interactions.

Our n-gram analyses were deliberately narrow. We restricted ourselves to unigrams, bigrams, and trigrams, and examined semantic families (country and capitals) under templated prompts with an anchor token. While this clarifies frequency effects, it limits generality. Frequencies were computed on a deduplicated and preshuffled snapshot of THE PILE, which may not reflect the model's true training distribution.

Lastly, we do not provide a normative explanation of emergence: although we observe regularities in temporal dynamics, coverage, and geometry, we do not derive these from an optimality principle or learning-theoretic objective. Our causal claims are therefore limited.

7 Future Work

Thus, future work should extend and operationalize this framework across architectures (e.g., LLaMA-family, state-space models), domains (multilingual settings, beyond country-related n-grams), and training paradigms (fine-tuning, instruction-tuning, RLHF) to test how coverage patterns and frequency invariance evolve by depth. Mechanistically, mapping JSD-contributing neurons and polytopes to concrete circuits can link geometric consolidation to algorithmic function. Practically, standardizing coverage-based metrics and polytope diagnostics as training-time monitoring tools could enable practitioners to track consolidation and abstraction dynamics in situ. Although our work establishes these as reliable diagnostic indicators of polysemanticity, future work could explore whether they can serve as control targets—for instance, through coverage-based regularization or dynamic architectural adjustments—potentially transforming polysemanticity from an interpretability obstacle into a measurable and, eventually, controllable aspect of neural development.

8 Acknowledgement

We are grateful to the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster that supported this research. We also thank the Algoverse AI Research program for their provision of computing resources, mentorship, and support. Jonas Rohweder is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

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] Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, et al. Interpreting neural networks through the polytope lens. *arXiv preprint arXiv:2211.12312*, 2022.

[3] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askella, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[4] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv Preprint*, 2023.

[5] Maximilian Dreyer, Erblina Purelku, Johanna Vielhaben, Wojciech Samek, and Sebastien Lapuschkin. Pure: Turning polysemantic neurons into pure features by identifying relevant circuits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024.

[6] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders enable fine-grained interpretable circuit analysis for language models. *AI Alignment Forum*, 2024.

[7] EleutherAI. tokengrams. `https://github.com/EleutherAI/tokengrams`, 2024.

[8] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.

[9] Alex Foote. Tackling polysemanticity with neuron embeddings. *arXiv Preprint*, 2024.

[10] Bofan Gong, Shiyang Lai, and Dawn Song. Probing the vulnerability of large language models to polysemantic interventions. *arXiv Preprint*, 2025.

[11] Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. Sparse autoencoders work on attention layer outputs. *AI Alignment Forum*, 2024.

[12] Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*, 2024.

[13] Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. *arXiv Preprint*, 2025.

[14] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *5th International Conference on Learning Representations*, 2017.

[15] Neel Nanda. Neuroscope. `neuroscope.io/index.html.`, 2022.

[16] Neel Nanda and Joseph Bloom. Transformerlens: A library for mechanistic interpretability of generative language models. `https://github.com/TransformerLensOrg/`, 2022.

[17] V. Nandakumar, Peng Mi, and Tongliang Liu. State space models can express n-gram languages. *Trans. Mach. Learn. Res.*, 2025, 2023.

[18] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020.

[19] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2017.

[20] Laura O'Mahony, Vincent Andrearczyk, Henning Müller, and Mara Graziani. Disentangling neuron representations with concept vectors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023.

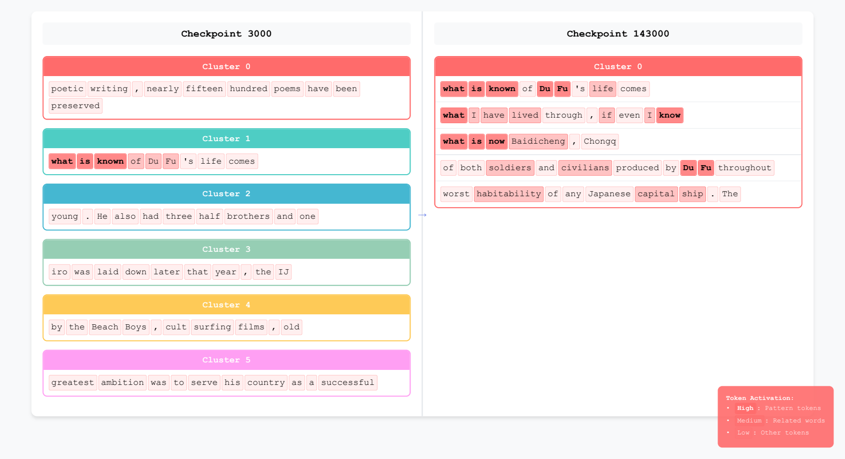
[21] Basri Ronen, David Jacobs, Yoni Kasten, and Shira Kritchman. The convergence rate of neural networks for learned functions of different frequencies. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Yang Xu, Yi Wang, and Hao Wang. Tracking the feature dynamics in llm training: A mechanistic study. *arXiv Preprint*, 2024.

A Appendix

A.1 Temporal Dynamics of Polysemanticity: A Case Study

In this section of the Appendix, we present a case study of feature cluster evolution for a specialized neuron. This consists of its feature clusters at the first and last checkpoints, and a figure that plots its cluster counts over every publicly available checkpoint.



Checkpoint 3000

Cluster 0

poetic writing , nearly fifteen hundred poems have been preserved

Cluster 1

what **is** **known** of Du Fu 's life comes

Cluster 2

young . He also had three half brothers and one

Cluster 3

iro was laid down later that year , the IJ

Cluster 4

by the Beach Boys , cult surfing films , old

Cluster 5

greatest ambition was to serve his country as a successful

Checkpoint 143000

Cluster 0

what **is** **known** of **Du** **Fu** 's life comes

what I have lived through , if even I **know**

what **is** **now** Baidicheng , Chongq

of both soldiers and civilians produced by **Du** **Fu** throughout

worst habitability of any Japanese capital ship . The

Token Activation:
- **High** : Pattern tokens
- Medium : Related words
- Low : Other tokens

Figure 9: **Case study: cluster consolidation for a specialized neuron.** Multiple heterogeneous clusters (poetic writing, Du Fu, the Beach Boys...) at Checkpoint 3,000 coalesce into a cluster focused on (i) Du Fu and (ii) the token-level pattern "what..." by Checkpoint 143,000.

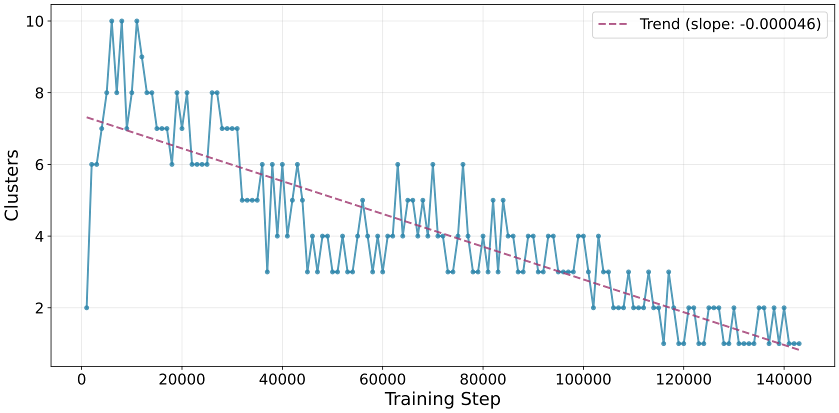


Figure 10: **Case study: cluster consolidation for a specialized neuron.** For the neuron visualized in Figure 9, we track its convergence towards a single cluster in high-resolution: across 143 checkpoints. Such resolution was not viable for our global analyses due to cost concerns.

A.2 Group-level Polysemanticity and Volatility

In this section of the Appendix, we present an experiment looking into group-level polysemanticity and an associated volatility metric.

A.2.1 Method

For any frequency group G at step s, we summarize polysemanticity with a weighted average, so neurons that respond more strongly to a group's phrases contribute more to its score.

$$\overline{\#\text{clusters}}_s(G) \;=\; \frac{\sum_{(L,h)\in\mathcal{H}} w_{s,L,h}(G)\cdot \#\text{clusters}_{s,L,h}}{\sum_{(L,h)\in\mathcal{H}} w_{s,L,h}(G)}, \qquad w_{s,L,h}(G) \propto \sum_{p\in G} P_{s,L}(h \mid p) \quad (15)$$

To track these group-level trends' stability, we also use the median absolute percent change (MAPC), which quantifies the step-to-step volatility of each group's representation.

$$\text{MAPC}(G) \;=\; \text{median}_t\left(\left|\frac{\overline{\#\text{clusters}}_{t+1}(G) - \overline{\#\text{clusters}}_t(G)}{\overline{\#\text{clusters}}_t(G) + \varepsilon}\right|\right) \quad (16)$$

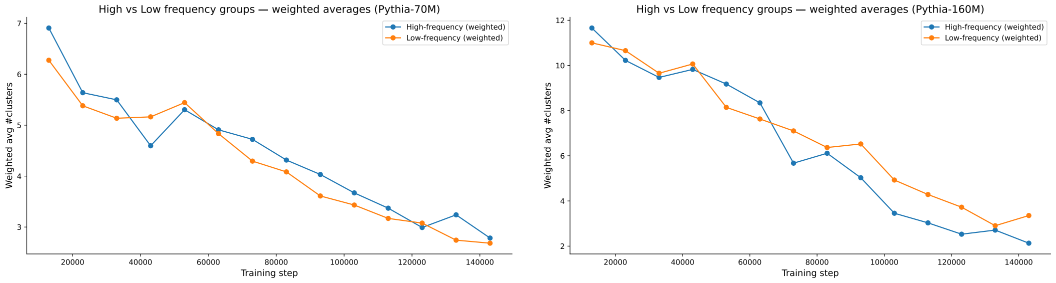
Together, these metrics capture the distinct features a group evokes and how consistently they stabilize over training.

A.2.2 Result

Frequency plays a *transient and unstable* role in this trajectory. Contrary to common assumptions, frequency alone does not drive sustained polysemanticity differences; instead, both high- and low-frequency groups converge as consolidation proceeds, challenging frequency-centric explanations.

Across checkpoints, weighted average polysemanticity decreases for both high- and low-frequency groups in Pythia-70M and Pythia-160M (Appendix Fig. 11), often converging late or even crossing mid-training. Thus, frequency is *not* a stable predictor of group-level polysemanticity once training progresses; the dominant signal is global consolidation of features.

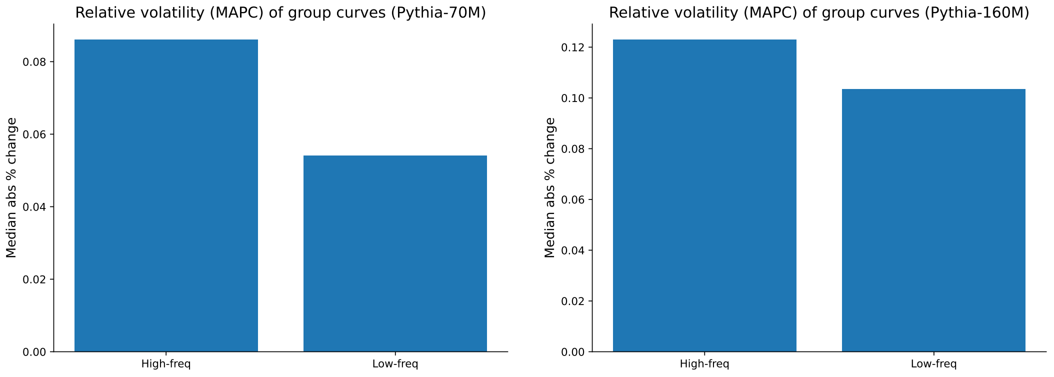
Despite similar endpoints, *stability* differs: high-frequency n-grams exhibit larger median absolute percent change (MAPC) than low-frequency n-grams in both sizes (Appendix Fig. 12). High-frequency content is updated/rewired more often during training, even as both groups compress to similar polysemanticity.



(a) Pythia-70M

(b) Pythia-160M

Figure 11: **Group trajectories.** Weighted average #clusters for high- vs. low-frequency n-grams across training. Both groups contract; the gap is small, sometimes reverses, and diminishes late.



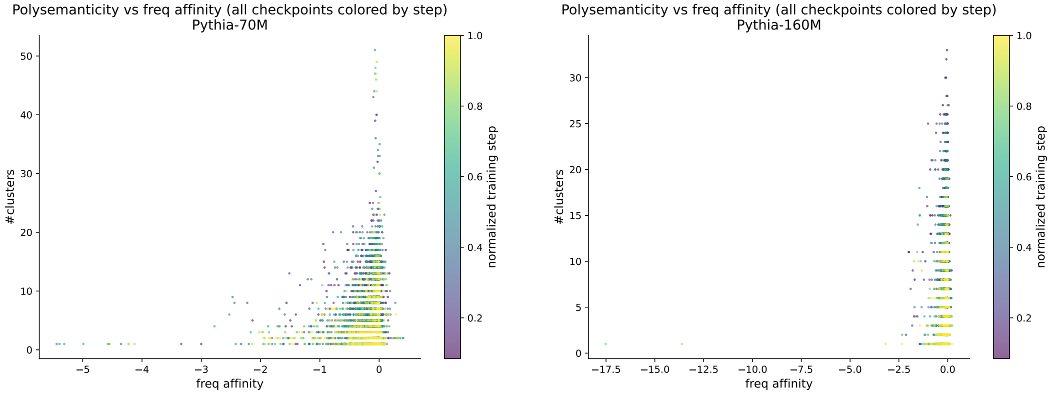
(a) Pythia-70M

(b) Pythia-160M

Figure 12: **Relative volatility (MAPC).** High-frequency groups show higher median absolute % change than low-frequency groups, indicating more step-to-step rewiring.

A.3 Figures on Sum Affinity and Coverage Analysis

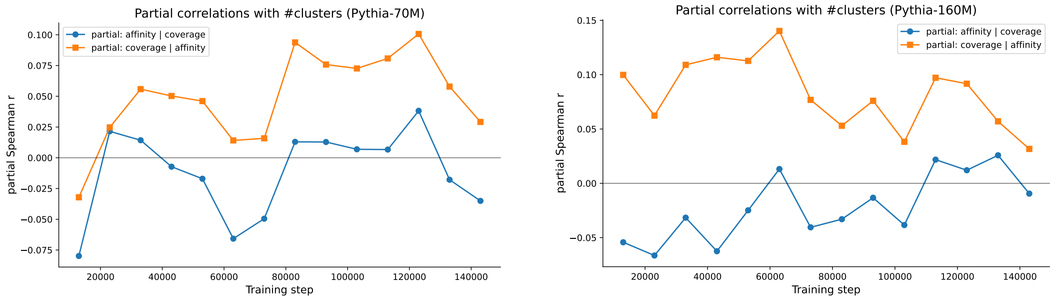
Because $A^{\text{sum}} = \tilde{C} \cdot \bar{A}$, and because $\bar{A}$ is mostly negative on our phrase set, A^{sum} is effectively a monotone transform of $\tilde{C}$; the observed correlation with #clusters thus comes from coverage, not from frequency preference per se.



(a) Pythia-70M

(b) Pythia-160M

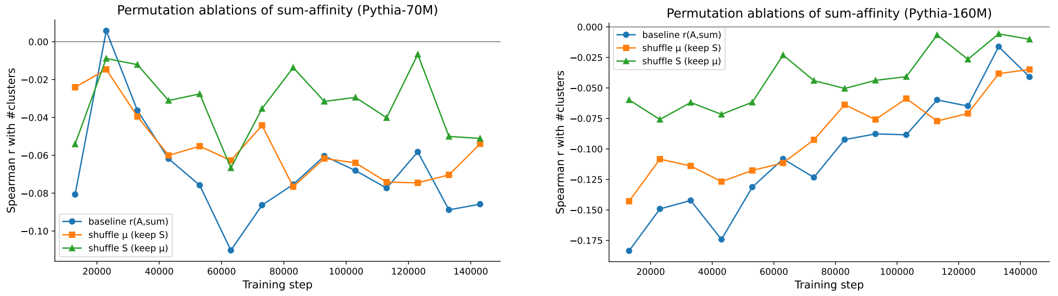
Figure 13: **Sum-affinity scatters** Sum affinity vs. #clusters. The trend largely mirrors coverage because $A^{\text{sum}}{=}m \cdot \bar{A}$ with $\bar{A}{<}0$.



(a) Pythia-70M

(b) Pythia-160M

Figure 14: **Partial correlations with polysemanticity.** Blue: $\rho(\text{sum-affinity}, \#\text{clusters} \mid \text{coverage})$. Orange: $\rho(\text{coverage}, \#\text{clusters} \mid \text{sum-affinity})$. Coverage retains a positive association after controlling for affinity; affinity's residual association is small or negative.

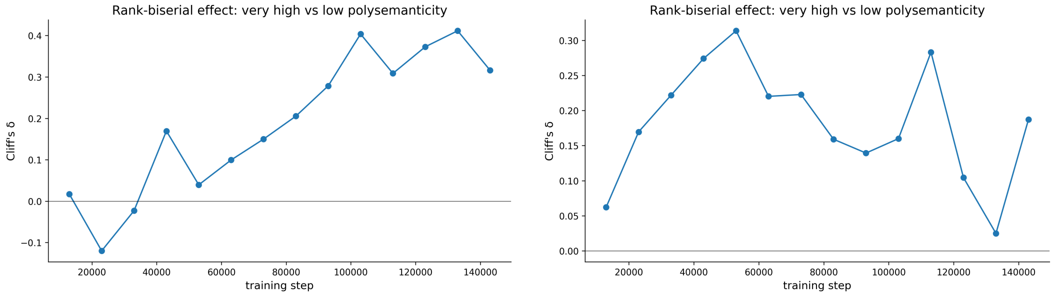


(a) Permutation ablations (70M)

(b) Permutation ablations (160M)

Figure 15: **Coverage drives A^{sum}'s association.** Shuffling the coverage term m largely destroys the sum-affinity correlation with #clusters; shuffling the mean-affinity component $\bar{A}$ has minimal effect.

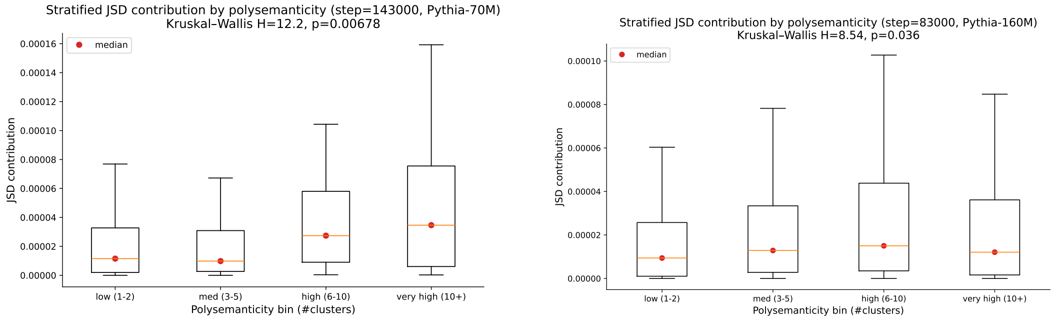
A.4 More Figures for JSD Analysis



(a) Cliff's δ (70M)

(b) Cliff's δ (160M)

Figure 16: **Effect size (very-high vs. low polysemanticity).** Polysemantic neurons increasingly dominate frequency separation, with positive, growing distribution-free effect sizes.

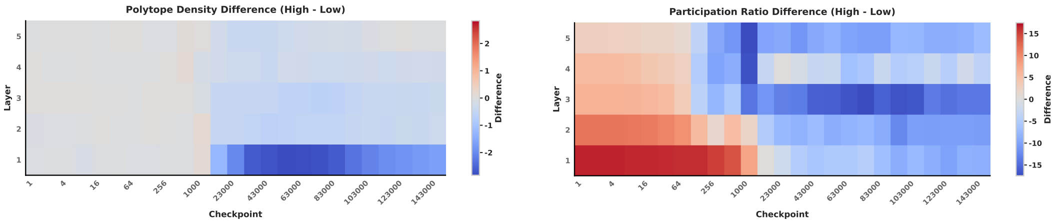


(a) Pythia-70M (143k)

(b) Pythia-160M (83k)

Figure 17: **Stratified JSD contributions by polysemanticity.** Higher #clusters contribute more to inter-bucket JSD; nonparametric tests indicate significant stratification at representative steps.

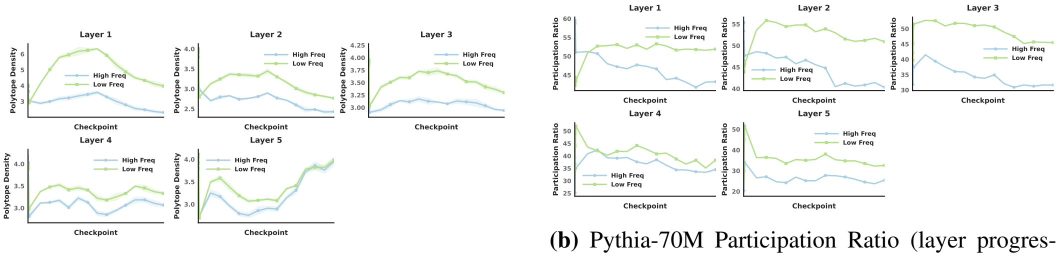
A.5 More Figures for Polytope Density and Participation Ratios



(a) Pythia-70M Polytope Density

(b) Pythia-70M Participation Ratio

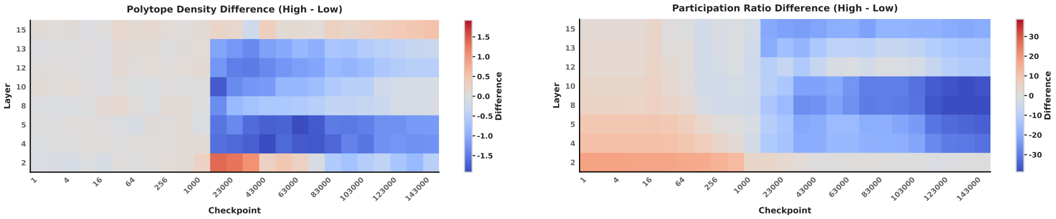
Figure 18: **Geometric convergence reveals hierarchical abstraction.** Difference heatmaps (high minus low frequency) for country n-grams across layers and training steps.



(a) Pythia-70M Polytope Density (layer progression)

(b) Pythia-70M Participation Ratio (layer progression)

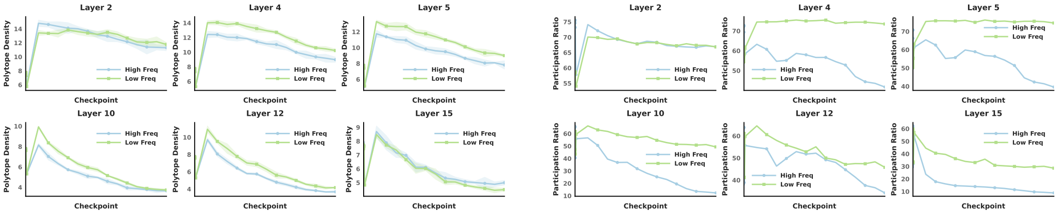
Figure 19: **Layer-wise progression over training for country n-grams.** Small-multiple plots across 5 layers showing the temporal evolution of geometric metrics.



(a) Pythia-1B Polytope Density

(b) Pythia-1B Participation Ratio

Figure 20: **Geometric convergence reveals hierarchical abstraction.** Difference heatmaps (high minus low frequency) for capital n-grams across layers and training steps.



(a) Pythia-1B Polytope Density (layer progression)

(b) Pythia-1B Participation Ratio (layer progression)

Figure 21: **Layer-wise progression over training for capital n-grams.** Small-multiple plots across 5 layers showing the temporal evolution of geometric metrics.