The Connectome of a Large Language Model

Ruixuan Deng [1] **Zehao Jin** [1]

Abstract

A central question about information processing systems is whether fundamentally different optimization processes lead to similar internal organization. Cortical networks exhibit small-world topology, modular communities, scale-free hubs, and structural-functional dissociation, and sparse transcoders now provide the interpretable units needed to test these same properties inside a large language model. We use skip transcoders from Gemma Scope 2 to extract 16,384 sparse interpretable features per MLP layer of Gemma 3 (270M and 1B), construct per-layer co-activation networks from 1M tokens of FineWeb-Edu, and analyze them with a standard network-neuroscience toolbox augmented by the Laplacian renormalization group and the spectral heat capacity. Both models exhibit analogues of all four classical cortical signatures, though with threshold sensitivity in the small-world case and quantitative differences in fractal dimension. Multi-scale analysis further reveals a dual-regime fractal scaling with a local exponent $d_B \approx 1.35$ below the Leiden community diameter, a single-peaked heat capacity that sharpens with depth, and a non-monotonic cross-layer coupling trajectory. Hub features and coarse-grained communities carry interpretable semantics organized along a tokenization-syntax-semantics-morphology hierarchy. These findings indicate that transformer representations trained on next-token prediction exhibit several network properties that are statistically similar to those observed in cortical networks.

[1]Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Ruixuan Deng <rdeng62@gatech.edu>, Zehao Jin <zehao@gatech.edu>.

Accepted to the ICML 2026 Workshop on Mechanistic Interpretability, Seoul, South Korea, 2026. Non-archival. Copyright 2026 by the author(s).

1. Introduction

A central question for any field that studies complex information processing systems is whether systems shaped by very different processes converge on a common internal architecture. Biological brains have been shaped by hundreds of millions of years of evolution under tight energy and wiring constraints, and a long tradition in network neuroscience has shown that the resulting cortical networks exhibit a small set of robust organisational principles. These include small-world topology with high clustering and short path length (Watts & Strogatz, 1998; Bassett & Bullmore, 2017), modular community structure that segregates and integrates information (Sporns, 2013; Newman, 2006), scale-free hubs that dominate global communication (Albert et al., 2000), and a partial dissociation between structural and functional connectivity in which weight-based and correlation-based connectomes are only loosely aligned (Liégeois et al., 2020). Large language models, in contrast, are shaped by gradient descent on text corpora over a few months of compute and have very different inductive biases. Whether the trained models nevertheless converge on the same network signatures is a question with theoretical motivation — modularity can emerge spontaneously under modularly varying objectives (Kashtan & Alon, 2005), and sparse network topologies inspired by network science have already improved ANN training (Mocanu et al., 2018) — but one that, until recently, could not be posed because the units of analysis inside an LLM were not well defined.

Existing approaches to LLM internals have focused on tracing causal circuits behind specific computations or comparing representational geometry with biological systems, but these approaches do not characterize the network organization of the model as a whole. Recent progress on sparse decompositions of internal activations changes this picture. Sparse autoencoders and transcoders learn to reconstruct each MLP block from a much wider dictionary of features that activate sparsely on natural text (Cunningham et al., 2023; Templeton et al., 2024; Dunefsky et al., 2024; Paulo et al., 2025). With the public release of Gemma Scope and Gemma Scope 2 (Lieberum et al., 2024; McDougall et al., 2025), every layer of the Gemma 3 family (Team et al., 2025) now comes equipped with a per-layer skip transcoder whose features are largely interpretable in isolation. These features supply a candidate analogue of the neuron for net-

work analysis. They are sparse, they fire on identifiable inputs, and they admit pairwise correlation analysis in the same way that neural spike trains do. We can therefore treat an LLM as an empirical network and apply the standard tools of network science to it, though the resulting networks depend on the choice of decomposition.

This paper takes that step. For two scales of the Gemma 3 family (270M and 1B), we build per-layer co-activation networks from transcoder features on a 1M-token corpus alongside weight-based structural networks from the transcoder parameters, and analyze both with the standard topology toolbox of network neuroscience augmented by multi-scale tools from statistical physics. We ask which of the classical cortical signatures hold, which fail, and where new structure appears. We then attach semantics to the topology by labelling hub features and coarse-grained communities through their maximally activating contexts.

The networks exhibit analogues of all four classical cortical signatures, and multi-scale analysis reveals three further properties, including a dual-regime fractal scaling with a local exponent of about 1.35 below the Leiden community diameter, a single-peaked spectral heat capacity that sharpens with depth, and a non-monotonic cross-layer coupling trajectory. The quantitative differences from cortex are themselves informative. The LLM fractal dimension falls below the cortical value of approximately 2.0 reported by Gallos et al. (2012), and the heat capacity shows a single sharp peak where cortical connectome data shows a broad plateau (Poggialini et al., 2024), both consistent with sparser and more specialized internal connectivity. Hub features and coarse-grained communities carry interpretable semantics organized along a depth-dependent tokenization-syntax-semantics-morphology hierarchy. To our knowledge, this is the first systematic network-science analysis of an interpretable LLM feature space at corpus scale.

2. Related Work

2.1. Mechanistic Interpretability and Circuit Analysis

Mechanistic interpretability has made rapid progress in identifying the internal computations of language models through causal analysis. Ameisen et al. (2025) introduced circuit tracing, which constructs attribution graphs that track how individual features and attention heads contribute to a model's output on a given prompt, and Lindsey et al. (2025) used thousands of such graphs to catalog recurring computational motifs. Marks et al. (2025) discovered task-specific sparse feature circuits through causal mediation, and Dunefsky et al. (2024) showed that transcoders facilitate the extraction of interpretable causal graphs. Because each attribution graph is tied to a single prompt and a single task, the cost of interpretation scales with the number of behaviors under study. These methods are powerful for explaining individual computations, and the resulting circuits have revealed rich internal structure, but they do not aggregate into a characterization of how the model as a whole is organized.

2.2. Brain-AI Representational Comparison

A separate line of work compares the internal representations of biological and artificial neural networks. Representational Similarity Analysis (Kriegeskorte et al., 2008) and related methods such as Centered Kernel Alignment (Kornblith et al., 2019) quantify the similarity structure of population responses, providing a common framework for comparing brains and models across architectures. This approach has shown that task-optimized deep networks predict representational geometry in visual cortex (Yamins & DiCarlo, 2016) and that transformer language models trained on next-word prediction explain a large fraction of neural variance in the human language network (Schrimpf et al., 2021). Recent extensions incorporate topological features of representation spaces beyond pairwise distances (Lin & Kriegeskorte, 2024). RSA and related methods measure whether two systems encode similar information, and have established detailed correspondences between model layers and brain regions. Representational geometry is a static description of encoding, and does not capture the interactions among units that emerge during processing, as the well-documented structure-function dissociation in cortical networks illustrates (Liégeois et al., 2020).

2.3. Graph Construction from LLM Features

Several recent works have constructed graphs from LLM features without applying network-science analysis to them. Deng et al. (2026) built co-activation graphs from sparse feature activations and showed that connected components correspond to composable semantic modules. Balcells et al. (2024) tracked features across adjacent layers using Pearson and Jaccard scores on a 10M-token Pile sample, identifying pass-through, gating, and disappearing features. Laptev et al. (2025) traced cross-layer feature flow through cosine similarity of decoder weights. The circuit tracing programme at Anthropic (Ameisen et al., 2025; Lindsey et al., 2025) produced thousands of attribution graphs at the prompt level and used manual supernode grouping that effectively performs community detection by hand. None of these works compute corpus-wide network statistics on the resulting graphs, and none frames the analysis in terms of the cortical comparison. Two concurrent works move closer to network science. Zheng et al. (2026) computes topological metrics on graphs built from raw neurons, and Liu et al. (2026) applies ICA to neuron activations in the spirit of fMRI functional network analysis. Both operate on raw neurons without multi-scale topological analysis on interpretable features.

3. Data and Methods

We work with two scales from the Gemma 3 family, Gemma 3 270M (18 layers, $d_{\text{model}} = 640$) and Gemma 3 1B (26 layers, $d_{\text{model}} = 1152$), loading the corresponding Gemma Scope 2 skip transcoders with 16,384 features per layer (Lieberum et al., 2024; McDougall et al., 2025). For each layer we build a functional network by streaming a 1M-token sample of FineWeb-Edu (Penedo et al., 2024) through the model, encoding MLP inputs with the transcoder, and accumulating a 16K by 16K Pearson co-activation matrix in a single pass. We complement these with structural networks built from the transcoder parameters, using cosine similarity between decoder weight vectors within a layer and the virtual weight $M_{ij}^{(\ell\to\ell+1)} = |W_{\text{dec}}^{(\ell)}[i,:] \cdot W_{\text{enc}}^{(\ell+1)}[j,:]|$ across layers (Elhage et al., 2021; Dunefsky et al., 2024). Both matrices are converted to sparse graphs by density-matched thresholding at a canonical density $d = 0.001$, with robustness checks at two additional densities. Details of the streaming computation, feature filtering, and threshold selection are in Appendix E.

The topology toolbox follows network neuroscience standards. We compute small-world indices, Leiden modularity at multiple resolutions (Traag et al., 2019), degree distribution statistics, assortativity, and the rich-club coefficient, testing every metric against both Erdős-Rényi and Configuration Model nulls (Newman, 2018). Network robustness is assessed by percolation under random and targeted node removal (Albert et al., 2000), and structure-function coupling follows the methodology of Liégeois et al. (2020), combining a Mantel correlation with a conditional analysis of functional connectivity at the strongest structural edges. For multi-scale structure we apply three renormalization group tools. Box covering (Song et al., 2005; 2007) yields the fractal dimension d_B and, through a piecewise fit, a local exponent below a crossover length. The Laplacian renormalization group of Villegas et al. (2023) coarse-grains the network at a fixed diffusion time into supernodes, and the spectral heat capacity $C(\tau) = \tau^2[\langle\lambda^2\rangle_\tau - \langle\lambda\rangle_\tau^2]$ (Poggialini et al., 2024) diagnoses whether the network has a single characteristic scale, approximate scale invariance, or multiple well-separated scales. Full parameter settings and null model protocols are in Appendix E.

To attach semantics to the topology, we identify the top hub features by degree and the largest Laplacian renormalization group communities in each layer, and record their maximally activating contexts on a held-out corpus. Per-feature word clouds aggregate content words weighted by activation magnitude.

4. Results

We organize our results around five progressive questions about the internal organization of LLM feature networks.

4.1. Do Feature Co-Activation Networks Have Non-Trivial Structure?

At the canonical density $d = 0.001$, the Telesford small-world index ω ranges over $[-0.17, +0.57]$ for Gemma 3 270M and $[-0.43, +0.32]$ for Gemma 3 1B, with the layer-wise mean close to zero (Telesford et al., 2011). Early layers tend toward positive ω (shorter paths relative to clustering) while deep layers shift toward negative ω (higher clustering relative to path lengths), consistent with the increasing modularity reported in the next section. This property is threshold-sensitive and holds robustly only at sparse densities ($d \leq 0.002$). At $d = 0.005$, ω shifts substantially toward positive values in both models (Appendix Figure 8a), indicating that the small-world balance is carried by the sparsest, strongest connections.

Degree distributions are heavy-tailed across all layers, with skewness typically in the range 2 to 8 and maximum degree reaching up to 47× the mean (270M) and 53× the mean (1B) in early layers. These hub features have direct structural consequences for graph connectivity. Progressive node removal reveals that networks remain largely intact under random failure (the largest connected component retains approximately 40% of nodes after 50% removal) but collapse rapidly when high-degree or high-betweenness nodes are targeted (Figure 2). This differential vulnerability to targeted versus random attack is a defining signature of heterogeneous, hub-dependent architecture (Albert et al., 2000).

Null model comparisons confirm that this structure exceeds what degree heterogeneity alone can explain. Configuration Model z-scores for modularity range from 198 to 635 in 270M and 68 to 808 in 1B, all far exceeding statistical significance. Feature co-activation networks are therefore highly structured, with small-world topology, prominent hubs, and modular organization that cannot be reduced to the degree sequence.

4.2. How Does Network Organization Evolve with Layer Depth?

Network organization could be uniform across layers or vary systematically with depth. Leiden modularity Q reveals a clear trend (Figure 1). In the 270M model, Q rises from 0.48 at layer 0 to 0.88 at layer 16, a nearly twofold increase. In the 1B model, Q rises from 0.46 (layer 2) to 0.93 (layer 24). Of all the properties we measure, this depth trend is the most consistent across density levels and both model scales. Configuration Model z-scores confirm that

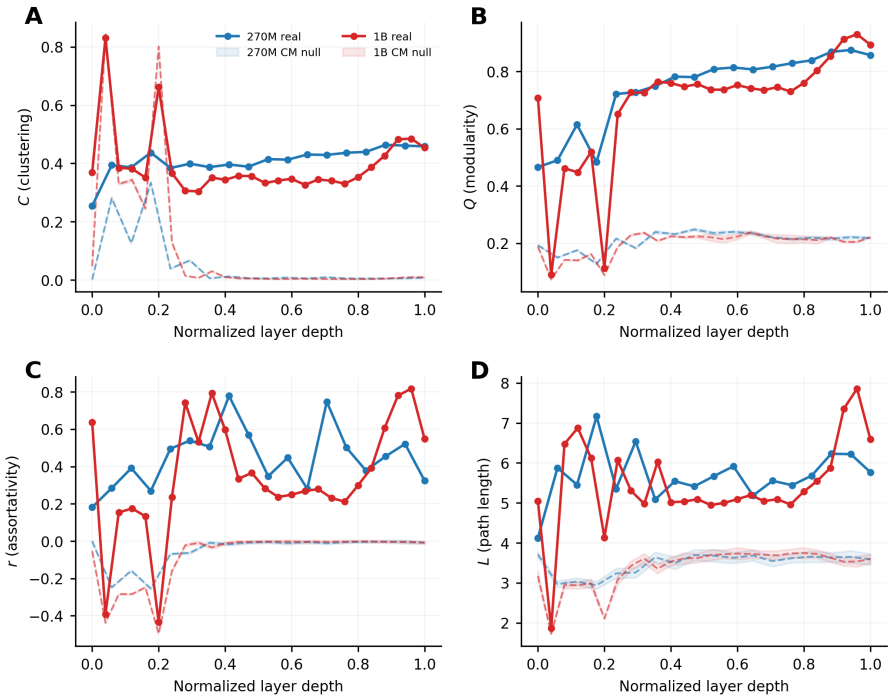


Figure 1. Layer evolution of four key topology metrics for Gemma 3 270M (blue) and 1B (red) at density-matched Pearson networks ($d = 0.001$). Solid lines show observed values, dashed lines show the Configuration Model null mean with shaded $\pm 2\sigma$ bands.

the trend reflects genuine organizational change, with modularity exceeding the degree-preserving null by hundreds of standard deviations at every layer (Appendix Table 2).

The monotonic increase in modularity implies that features in early layers form weakly differentiated groups with overlapping co-activation patterns, while features in deep layers form highly separated functional modules. This depth-dependent specialization is reminiscent of the cortical hierarchy in biological brains, where primary sensory areas show diffuse connectivity and association areas exhibit stronger modular organization (Sporns, 2013).

Spectral analysis on the unthresholded Pearson matrices provides independent, threshold-free confirmation. Fiedler values (λ_2) increase from early to middle layers, reflecting growing algebraic connectivity (Appendix Figure 8b). These spectral signatures confirm that the structural evolution is a property of the correlation matrices themselves, independent of binarization.

Two layers in the 1B model (layers 1 and 5) are consistent outliers across multiple metrics, showing extremely low modularity ($Q = 0.09$ and 0.11), high mean Pearson correlation, extreme hub degrees, and negative assortativity. These layers likely correspond to global attention layers in Gemma 3's interleaved local/global attention architecture (Team et al., 2025), where all features share the same context window and exhibit uniformly stronger co-activation. Their consistent anomalous behavior across every analysis serves as a useful internal control, suggesting that our metrics are sensitive to structural differences between layers.

4.3. What Is the Multi-Scale Geometry of These Networks?

Biological neural networks exhibit multi-scale organization with fractal geometry within functional modules and small-world shortcuts between them (Betzel & Bassett, 2017; Gallos et al., 2012). We ask whether LLM feature networks share this multi-scale organization.

Box-covering analysis of the thresholded graphs reveals two distinct scaling regimes in the relation $N_B(l_B) \sim l_B^{-d_B}$. A single power-law fit yields $d_B \approx 2.0$ with moderate $R^2 \approx 0.93$ in both models, but allowing a piecewise fit with a breakpoint dramatically improves the fit and clears the ΔAIC > 10 bar in most layers. Below the crossover length $l_B^* \approx 6$, the local fractal dimension averages 1.37 ± 0.14 (270M) and 1.34 ± 0.11 (1B), consistent across models. Both models show a gradual increase in d_B with depth, from approximately 1.0–1.25 in early layers to approximately 1.5 in late layers (Figure 3), indicating that deeper layers develop denser intra-module connectivity. Above the crossover, steeper scaling ($d_B > 3$) reflects the small-world shortcuts connecting modules. The crossover length itself corresponds approximately to the diameter of Leiden com-

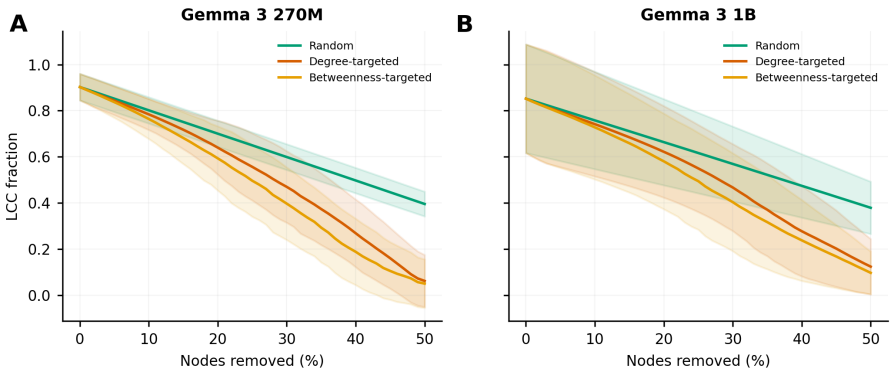


Figure 2. Network robustness under progressive node removal, averaged across all layers. **Left**: Gemma 3 270M. **Right**: Gemma 3 1B. Random removal (green) decays gradually, while degree-targeted (orange) and betweenness-targeted (red) removal collapse the largest connected component much faster. Shaded bands show ± 1 standard deviation across layers.

munities, consistent with the fractal-module framework of Gallos et al. (2012). The local exponent of about 1.35 is lower than the cortical value of approximately 2.0, suggesting sparser internal connectivity inside LLM communities than inside cortical functional modules. With approximately ten data points per layer, the piecewise results are suggestive and the single-fit d_B with bootstrap CI serves as the primary conservative estimate.

Laplacian renormalization group coarse-graining probes compressibility at a fixed diffusion scale. A single round of LRG reduces the network to progressively fewer supernodes with increasing depth. Early layers yield 715–823 supernodes in 270M (9.7–11.2$\times$ compression), while late layers yield as few as 191–209 supernodes (38–42$\times$ compression). The 1B model shows even stronger late-layer compression, reaching up to 62.5$\times$. This trend, despite some layer-to-layer variability, is consistent with the increasing modularity reported above. Features within well-separated communities share diffusive modes and can be merged, so higher modularity produces greater compressibility. The supernode counts per depth are shown in Appendix Figure 10b.

The spectral heat capacity $C(\tau)$, computed from the full Laplacian spectrum, diagnoses whether the networks possess one characteristic scale, multiple well-separated scales, or approximate scale invariance (Poggialini et al., 2024). All layers in both models show a single $C(\tau)$ peak with no plateau, ruling out scale invariance in these networks. In the 270M model, the peak position τ_{peak} decreases from 17.5 (layer 0) to 9.9 (layer 16), and the peak height C_{peak} increases from 4.8 to 8.9 across depth (Figure 4), indicating that the dominant structural scale shifts to shorter diffusion times and grows in contrast. In the 1B model, τ_{peak} similarly decreases from 13.4 to 9.5. The C_{peak} values in the 1B model are more variable, with the anomalous layers 0–1 and 4–5 showing unusually large values (12–16) and the non-anomalous layers falling between 7.4 and 10.7 with a modest increase toward late layers. The peak width (FWHM) narrows with depth in the 1B model, indicating convergence toward a single well-defined scale, while the 270M model shows more variable FWHM across layers.

These three analyses converge on a coherent multi-scale picture. LLM feature networks are fractal within modules ($d_B \approx 1.35$) and connected by small-world shortcuts across modules. Each layer possesses a single dominant structural scale, and this scale becomes sharper and more compact with depth. The networks are modular and hierarchical, but they are not scale-invariant.

4.4. How Does Information Flow Between Layers?

Beyond within-layer organization, the residual stream mediates directed communication between adjacent layers. Global correlation between structural (cosine) and functional (Pearson) connectivity matrices is near zero across all layers (Figure 5), with Mantel r values ranging from -0.007 to $+0.013$ (270M) and -0.003 to $+0.034$ (1B). Despite this global dissociation, strong structural edges carry a meaningful signal. The mean functional connectivity among the top 0.1% strongest structural edges is approximately 20–25 times the global mean, and this enrichment decays monotonically with relaxing threshold, reaching approximately $2\times$ at the top 10%. The pattern parallels the well-documented structure-function relationship in brain networks, where DTI structural connectivity and fMRI functional connectivity show weak global correlation but strong local correspondence along white-matter tracts (Liégeois et al., 2020). A threshold-free Laplacian spectrum check with the same depth dependence is reported in Appendix B.

Turning to directed coupling between adjacent layers, we quantify the cross-layer weight overlap using `concentration_80`, the fraction of downstream LRG communities accounting for 80% of the total coupling mass

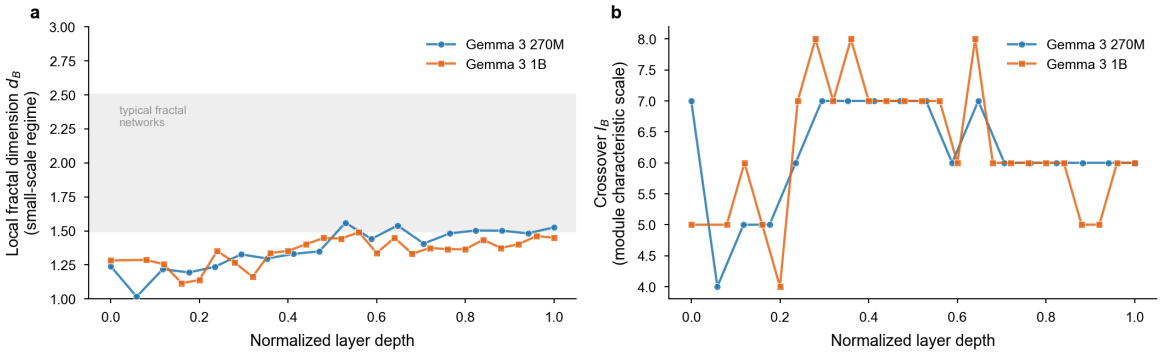


Figure 3. Box-covering fractal exponents across depth. **Left**: the local exponent d_B^{local} from the piecewise fit shows a gradual increase from approximately 1.0–1.25 in early layers to approximately 1.5 in late layers, with both models following the same trend. **Right**: the crossover length l_B^* has a median around 6, matching the empirical Leiden community diameter.

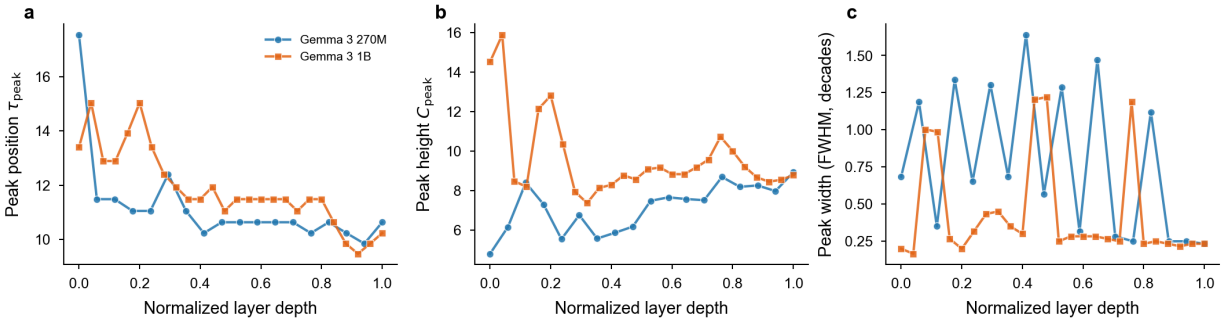


Figure 4. Evolution of spectral heat capacity peak parameters with depth. **Left**: τ_{peak} decreases. **Center**: C_{peak} increases with depth in 270M, while 1B shows more variable values. **Right**: FWHM narrows in 1B but varies in 270M.

from a given upstream community. This measure follows a non-monotonic trajectory across depth (Figure 6). In early layers, `concentration_80` values are low (1–3% for 270M layers 0–2, 2.5–3% for 1B layers 0–1), indicating that each upstream community directs its signal toward a small set of specific downstream targets. Through the first half of the network, `concentration_80` rises steadily, reaching peak values of 14.4% (270M, layer 9→10) and 19.0% (1B, layer 11→12). At these peaks, each community's coupling spreads broadly across many downstream targets, reflecting a diffuse, broadcast-like communication mode. In the final third of the network, `concentration_80` decreases to 5–9% (270M) and 7.5–13.5% (1B), indicating reconvergence toward focused coupling.

This inverted-U trajectory is consistent with three regimes of residual stream coupling. In early layers, weight overlap is narrow and feature-specific, suggesting focused communication pathways. In middle layers, coupling spreads broadly across downstream communities, consistent with a broadcast-like mode of information distribution (Elhage et al., 2021). In late layers, coupling reconverges toward fewer downstream targets, suggesting a return to focused processing. Whether this weight-space pattern corresponds to actual information flow during processing remains to be tested through causal interventions. The pattern is quantitatively consistent across both model scales.

4.5. Do Topological Structures Carry Interpretable Semantics?

Whether topological structures carry genuine semantic content is a crucial test of these networks' functional relevance. For each layer, we identify the top-20 features by degree in the density-matched graph and examine their maximally activating corpus examples. Hub semantics follow a systematic three-phase evolution across depth in both models. In Gemma 3 270M, the input layer is dominated by domain-specific content words (hub degree about 6× the mean), early layers (1–3) by format and code mega-hubs such as punctuation and structural tokens (hub degree 34–47× the mean), middle layers (9–14) by mathematical operators and digits (hub degree 11–17× the mean), and the output layer by morphological roots (hub degree 23× the mean). The 1B model exhibits a qualitatively similar progression with shifted indices fitting its 26 layers. The mega-hub phenomenon at the format and code layers is the network manifestation of universal-connector tokens that co-occur with almost everything else in the corpus and therefore form the backbone of the early-layer co-activation graph. Figure 7

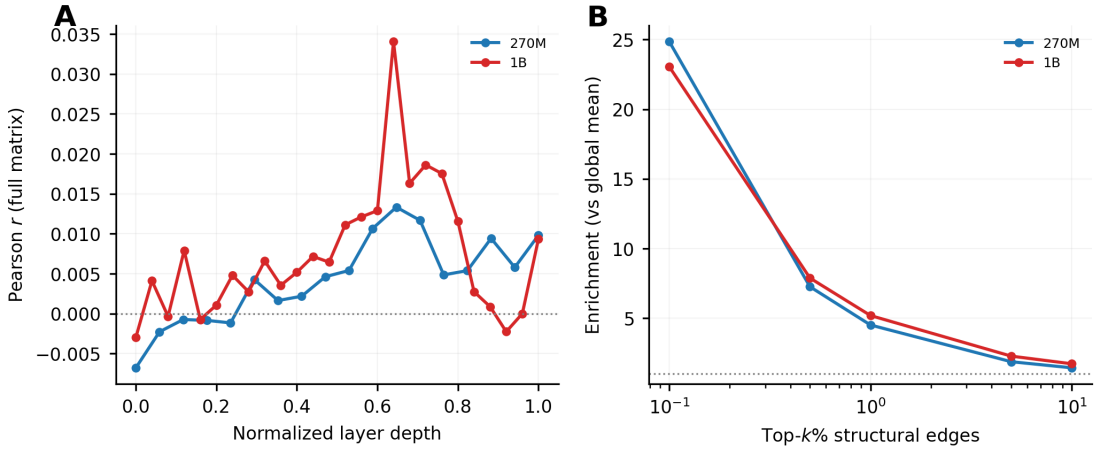


Figure 5. Structure-function coupling. **Left**: the global Mantel correlation between cosine and Pearson matrices stays near zero across layers. **Right**: functional connectivity enrichment at the top $k\%$ of structural edges, normalized by the global mean. The top 0.1% shows approximately 20–25-fold enrichment in both models.

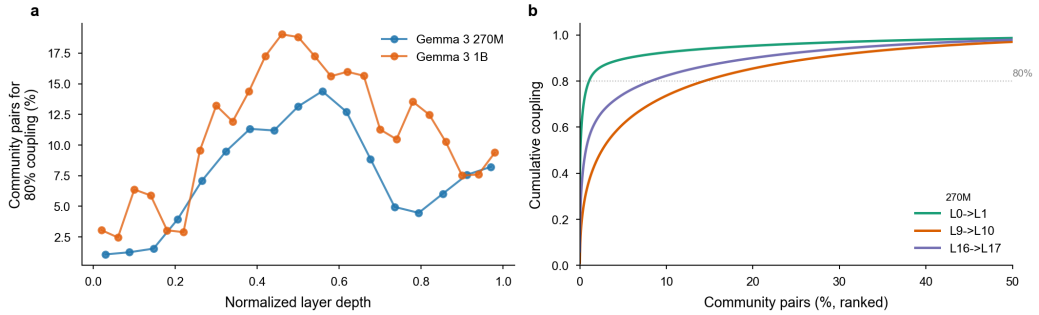


Figure 6. Cross-layer coupling trajectory. **Left**: `concentration_80` versus layer depth, showing the inverted-U shape. **Right**: cumulative coupling strength for representative layer transitions.

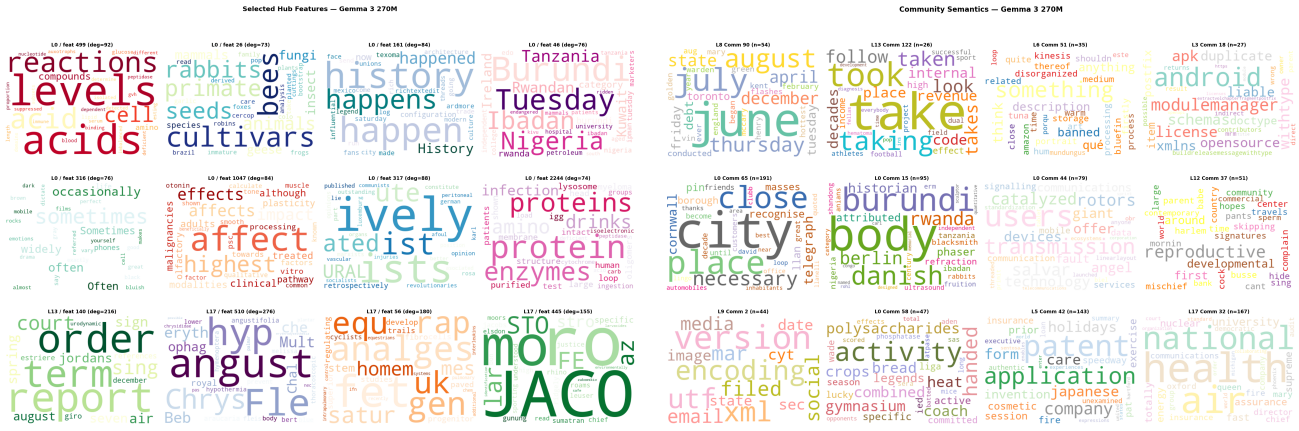
shows representative hubs and LRG communities for 270M, with the parallel 1B figures in Appendix D.

LRG communities provide a complementary view at the module level. Features within a community respond to the same semantic domain, syntactic category, or format type. Selected examples include geography clusters (layer 0), temporal terms (layer 8), biology (layer 12), and morphological paradigms (layer 13) in 270M, with comparable patterns in 1B (Appendix D). The community hierarchy across layers follows established models of transformer processing stages. Layer 0 communities cluster by semantic topic and early layers by syntactic categories, while middle layers shift toward grammatical function and late layers toward word-formation and morphological structure. This semantic alignment between topological structure and known linguistic processing stages (Lindsey et al., 2025) is consistent with the interpretation that the network properties reflect functional organization, though high-degree features may be interpretable simply because they are high-frequency.

5. Discussion

Our analyses paint a unified multi-scale picture of LLM feature network organization. Within modules at path distances below the crossover length $l_B^* \approx 6$, networks exhibit fractal-like scaling with $d_B \approx 1.35$. Across modules at larger scales, small-world shortcuts break the fractal scaling and produce $\omega \approx 0$ topology. Each layer possesses a single dominant structural scale, as revealed by the single peak in spectral heat capacity, and this scale becomes sharper with increasing depth. Between layers, information flow through the residual stream follows a focused-to-broadcast-to-focused trajectory. Throughout, topological structures (hubs and communities) carry interpretable semantics aligned with known processing hierarchies.

This organizational profile parallels biological neural networks in several respects. Both systems exhibit small-world topology, increasing modularity along processing hierarchies, fractal geometry within functional modules, disproportionate vulnerability to targeted hub removal, and structure-function dissociation coupled with strong local cor-



(a) Hub features (Gemma 3 270M). *(b)* LRG communities (Gemma 3 270M).

Figure 7. Semantic interpretation of topology in Gemma 3 270M. Each panel is a word cloud from the maximally activating contexts of a hub feature or representative members of a community. Selected hubs progress from domain content words at the input layer through format and code mega-hubs and digits to morphological roots at the output. Communities range from geography and technology at layer 0 through temporal and file-format clusters in middle layers to institutional terms in late layers.

Table 1. Comparison of organizational properties between biological brain networks and LLM feature networks.

Property	Brain networks	LLM feature networks
Small-world topology	$\omega \approx 0$ (Bassett & Bullmore, 2017)	$\omega \approx 0$ (this work)
Increasing modularity	primary $\rightarrow$ association cortex	early $\rightarrow$ late layers
Fractal within modules	$d_B \approx 2.0$ (Gallos et al., 2012)	$d_B \approx 1.35$ (this work)
Hub vulnerability	hub lesions cause disproportionate damage	hub removal collapses LCC
SC-FC dissociation	DTI $\neq$ fMRI but locally correlated	cosine $\neq$ Pearson but locally enriched
Spectral scale structure	broad plateau / approximate scale invariance (Poggialini et al., 2024)	$C(\tau)$ single peak / no plateau (this work)

respondence. The comparison is quantitative (Table 1). The main quantitative difference is the lower fractal dimension in LLM networks ($d_B \approx 1.35$ vs ≈ 2.0 in brain networks), suggesting that feature communities have sparser internal connectivity than brain functional modules. This difference may reflect the lower-dimensional geometry of residual stream representations compared to the three-dimensional embedding of biological cortex, or it may arise from differences in how co-activation is defined in the two domains.

The inverted-U coupling pattern carries specific implications for the "residual stream as communication bus" perspective in mechanistic interpretability (Elhage et al., 2021). If early layers perform focused, feature-specific processing and middle layers broadcast information broadly, then interventions on the residual stream should have qualitatively different effects depending on depth. Ablating residual stream directions in early layers should disrupt specific downstream features, while equivalent interventions in middle layers should produce more diffuse, distributed effects. This prediction is directly testable through activation patching experiments.

The quantitative consistency between 270M and 1B models strengthens these conclusions. Fractal dimensions (1.37 vs 1.34), crossover lengths (both approximately 6), single-peak heat capacity, inverted-U coupling trajectories, and semantic evolution patterns are all qualitatively and quantitatively similar across a fourfold increase in parameter count. This consistency suggests that these organizational principles are robust within the Gemma 3 family across a fourfold increase in parameter count. Whether they generalize to other model families remains an open empirical question.

Several limitations warrant caution. Co-activation networks capture statistical associations, not causal circuits. The topological centrality of hub features does not necessarily imply causal importance for model behavior, and feature ablation experiments would be needed to test whether degree centrality predicts functional impact. All graph metrics remain threshold-dependent. While density-matching and multi-threshold robustness checks mitigate this concern, the small-world property in particular is sensitive to density, holding at $d = 0.001$ but weakening at $d = 0.005$. Skip-transcoders explain approximately 85–95% of MLP variance, and the unexplained variance may carry additional relevant structure. The greedy box-covering algorithm provides an upper bound on N_B, potentially biasing fractal dimension estimates. Semantic interpretation through max-activating examples from 4,096 Pile samples may miss rare

activation patterns, as evidenced by the 9% empty rate for 270M layer 7 hubs. Finally, our analysis covers two scales from one model family (Gemma 3). Establishing whether these organizational principles are universal properties of language models or specific to the Gemma architecture requires investigation across diverse model families, training procedures, and scales.

6. Conclusion

We have presented, to our knowledge, the first systematic network-science analysis of an LLM's interpretable feature space at corpus scale, constructing and analyzing layer-wise co-activation and structural networks from Gemma 3 transcoder features. The networks exhibit analogues of all four classical cortical signatures and additionally show a dual-regime fractal scaling, a single-peaked spectral heat capacity that sharpens with depth, and a non-monotonic cross-layer coupling trajectory. Hub features and coarse-grained communities carry interpretable semantics organized along a processing hierarchy that matches known transformer computation stages. The similarity to cortical organizational principles, consistent across two model scales, indicates that trained transformer representations share several statistical network properties with cortical networks, though whether this reflects convergent optimization pressures or properties common to heterogeneous modular networks remains to be established. This work provides a foundation for a network-theoretic approach to LLM interpretability and demonstrates that the toolbox of network neuroscience can be productively applied to artificial information processing systems.

References

Albert, R., Jeong, H., and Barabási, A.-L. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, July 2000. ISSN 0028-0836, 1476-4687. doi: 10.1038/35019019.

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., et al. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Balcells, D., Lerner, B., Oesterle, M., Ucar, E., and Heimersheim, S. Evolution of sae features across layers in llms, November 2024.

Bassett, D. S. and Bullmore, E. T. Small-world brain networks revisited. *The Neuroscientist*, 23(5):499–516, October 2017. ISSN 1073-8584, 1089-4098. doi: 10.1177/1073858416667720.

Betzel, R. F. and Bassett, D. S. Multi-scale brain networks. *NeuroImage*, 160:73–83, October 2017. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.11.006.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. Power-law distributions in empirical data. *SIAM Review*, 51(4): 661–703, November 2009. ISSN 0036-1445, 1095-7200. doi: 10.1137/070710111.

Colizza, V., Flammini, A., Serrano, M. A., and Vespignani, A. Detecting rich-club ordering in complex networks. *Nature Physics*, 2(2):110–115, February 2006. ISSN 1745-2473, 1745-2481. doi: 10.1038/nphys209.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models, October 2023.

Deng, R., Hu, X., Gilberti, M., Storks, S., Taxali, A., et al. Sparse feature coactivation reveals causal semantic modules in large language models, April 2026.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable llm feature circuits, November 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Gallos, L. K., Makse, H. A., and Sigman, M. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *Proceedings of the National Academy of Sciences*, 109(8): 2825–2830, February 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1106612109.

Humphries, M. D. and Gurney, K. Network 'small-worldness': A quantitative method for determining canonical network equivalence. *PLoS ONE*, 3(4):e0002051, April 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0002051.

Kashtan, N. and Alon, U. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences*, 102(39):13773–13778, September 2005. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0503610102.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. E. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 2019.

Kriegeskorte, N., Mur, M., and Bandettini, P. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Laptev, D., Balagansky, N., Aksenov, Y., and Gavrilov, D. Analyze feature flow to enhance interpretation and steering in language models, July 2025.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., et al. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2, August 2024.

Liégeois, R., Santos, A., Matta, V., Van De Ville, D., and Sayed, A. H. Revisiting correlation-based functional connectivity and its relationship with structural connectivity. *Network Neuroscience*, 4(4):1235–1251, January 2020. ISSN 2472-1751. doi: 10.1162/netn_a_00166.

Lin, B. and Kriegeskorte, N. The topology and geometry of neural representations. *Proceedings of the National Academy of Sciences*, 121(42):e2317881121, October 2024. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2317881121.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., et al. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Liu, Y., Liu, Z., Wu, Z., Ning, J., Sun, H., et al. Brain-inspired exploration of functional networks and key neurons in large language models, January 2026.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., et al. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, March 2025.

McDougall, C., Conmy, A., Kramár, J., Lieberum, T., Rajamanoharan, S., et al. Gemma Scope 2 - technical paper. Technical report, Google, 2025.

Mocanu, D. C., Mocanu, E., Stone, P., Nguyen, P. H., Gibescu, M., et al. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature Communications*, 9(1):2383, June 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-04316-3.

Newman, M. E. J. Assortative mixing in networks. *Physical Review Letters*, 89(20):208701, October 2002. ISSN 0031-9007, 1079-7114. doi: 10.1103/PhysRevLett.89.208701.

Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, June 2006. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0601602103.

Newman, M. E. J. *Networks*. Oxford University Press, Oxford, United Kingdom ; New York, NY, United States of America, second edition edition, 2018. ISBN 978-0-19-880509-0.

Paulo, G., Shabalin, S., and Belrose, N. Transcoders beat sparse autoencoders for interpretability, February 2025.

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., et al. The fineweb datasets: Decanting the web for the finest text data at scale, October 2024.

Poggialini, A., Villegas, P., Muñoz, M. A., and Gabrielli, A. Networks with many structural scales: A renormalization group perspective, December 2024.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., et al. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

Song, C., Havlin, S., and Makse, H. A. Self-similarity of complex networks. *Nature*, 433(7024):392–395, January 2005. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature03248.

Song, C., Gallos, L. K., Havlin, S., and Makse, H. A. How to calculate the fractal dimension of a complex network: The box covering algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(03):P03006–P03006, March 2007. ISSN 1742-5468. doi: 10.1088/1742-5468/2007/03/P03006.

Sporns, O. Network attributes for segregation and integration in the human brain. *Current Opinion in Neurobiology*, 23(2):162–171, April 2013. ISSN 09594388. doi: 10.1016/j.conb.2012.11.015.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Telesford, Q. K., Joyce, K. E., Hayasaka, S., Burdette, J. H., and Laurienti, P. J. The ubiquity of small-world networks. *Brain Connectivity*, 1(5):367–375, December 2011. ISSN 2158-0014, 2158-0022. doi: 10.1089/brain.2011.0038.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Traag, V. A., Waltman, L., and Van Eck, N. J. From louvain to leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41695-z.

Villegas, P., Gili, T., Caldarelli, G., and Gabrielli, A. Laplacian renormalization group for heterogeneous networks. *Nature Physics*, 19(3):445–450, March 2023. ISSN 1745-2473, 1745-2481. doi: 10.1038/s41567-022-01866-8.

Watts, D. J. and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998. ISSN 0028-0836, 1476-4687. doi: 10.1038/30918.

Yamins, D. L. K. and DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016. doi: 10.1038/nn.4244.

Zheng, Y., Yuan, Y., Zhuo, Y., Li, Y., Kreiman, G., et al. Probing neural topology of large language models, January 2026.

A. Preliminaries

A decoder-only transformer is a stack of identical blocks, each composed of a multi-head self-attention sublayer and a position-wise MLP sublayer that read from and write to a shared residual stream of dimension d_{model} (Elhage et al., 2021). The MLP sublayers carry most of the parameter budget and most of the compositional computation, so mechanistic interpretability work has typically located the units we informally call "features" inside or around them.

A sparse autoencoder (SAE) approximates an activation $x \in \mathbb{R}^{d_{\text{model}}}$ as a sparse combination of learned dictionary atoms through an encoder $z = \phi(W_{\text{enc}}x + b_{\text{enc}}) \in \mathbb{R}^{d_{\text{sae}}}$ and a linear decoder $\hat{x} = W_{\text{dec}}z + b_{\text{dec}}$ with $d_{\text{sae}} \gg d_{\text{model}}$ (Cunningham et al., 2023; Templeton et al., 2024). Each column of W_{dec} is a feature direction in the residual stream and each entry of z is its activation on the current token. A transcoder generalises this to the input-output map of an MLP block (Dunefsky et al., 2024; Paulo et al., 2025), and a skip transcoder (McDougall et al., 2025) adds a learned linear shortcut,

$$\widehat{\text{MLP}}(x) \;=\; W_{\text{dec}}\, z(x) + b_{\text{dec}} + W_{\text{skip}}\, x, \qquad (1)$$

where $W_{\text{skip}}x$ absorbs the purely linear part and $z(x)$ carries the feature-like residual. The transcoders we use are 16,384-dimensional with JumpReLU activations and an L0 of about 50 active features per token, yielding a sparsity of approximately 0.3%.

Treating each (layer, feature) pair as a node and weighting edges by co-activation across a fixed corpus, we summarise the resulting graph using the standard network-science quantities. These include the clustering coefficient C and average shortest path length L combined into the Telesford small-world index $\omega = L_{\text{rand}}/L - C/C_{\text{lattice}}$ (Telesford et al., 2011), the Leiden modularity Q (Traag et al., 2019), the degree assortativity r (Newman, 2002), percolation under random and targeted node removal (Albert et al., 2000), and Mantel correlations between weight-based and activation-based matrices following the SC-FC methodology of Liégeois et al. (2020). For multi-scale structure we add three renormalisation group tools, namely box covering for the fractal dimension d_B (Song et al., 2007; Gallos et al., 2012), the Laplacian renormalisation group of Villegas et al. (2023), and the spectral heat capacity $C(\tau)$ of Poggialini et al. (2024). This toolbox is standard in network neuroscience, which is what makes the cortex-LLM comparison meaningful.

B. Threshold-independent and Robustness Checks

This appendix collects additional figures and tables that support claims made in the main text. Table 2 reports per-model

summaries of the canonical density-matched graph statistics. Figure 8a shows the modularity and small-world index at three density levels for Gemma 3 270M, Figure 8b shows the threshold-free Laplacian spectrum, and Figures 9a and 9b show degree distribution properties and the Configuration Model significance heatmap.

C. Renormalisation Group Supplementary Plots

Figure 10 collects the per-layer log-log box-covering curves and the Laplacian renormalisation group supernode count.

D. Semantic Interpretation for Gemma 3 1B

Figure 12 shows the parallel hub and community wordcloud panels for Gemma 3 1B. The semantic patterns are qualitatively the same as in 270M (Figure 7) but with shifted layer indices that fit the 26 layers of 1B.

E. Extended Methods

Network Construction Details. A forward hook on every MLP module captures the MLP input on each forward pass and feeds it through the corresponding skip transcoder, producing a sparse vector in $\mathbb{R}^{16,384}$. With about 50 active features per token, only $\binom{50}{2} \approx 1225$ feature pairs are updated per token, compared to the full 1.34×10^8 pairs in a 16K square matrix. We accumulate sufficient statistics for Pearson, Jaccard, cosine, and NPMI correlations in a single streaming pass without storing raw activations, and the resulting sufficient statistics are computed exactly (without streaming approximation). The output per layer is a 16K by 16K co-activation matrix together with a vector of per-feature activation counts used as a quality filter. Features with activation count below ten are excluded before any downstream analysis.

For the structural networks, taking absolute values of the virtual weight $M_{ij}^{(\ell \to \ell+1)}$ is a deliberate choice that aligns the structural network with the unsigned co-activation network and keeps edge weights compatible with modularity, box covering, and Mantel statistics, all of which require non-negative weights. The signed driving relation between features, which would distinguish inhibition from excitation, is left to follow-up work. The 270M model has $d_{\text{model}} = 640$, which makes its within-layer cosine similarity matrix too low in contrast for meaningful thresholding, so for 270M we report only cross-layer virtual weights as structural connectivity.

Density-Matched Thresholding. Graphs are obtained from the dense matrices by retaining the top edges by absolute correlation until the graph reaches a target density. Density matching is preferred over a fixed threshold because the scale of the correlation matrix varies systematically with depth. The canonical density is $d = 0.001$, with $d = 0.0005$ and $d = 0.002$ as robustness checks. A fixed-threshold control at $\theta = 0.15$ is also reported in the robustness analysis. All graph metrics are computed on the largest connected component using igraph.

Topology Metrics. The full set of computed metrics includes the clustering coefficient C, the average shortest path length L, the Telesford small-world index $\omega = L_{\text{rand}}/L - C/C_{\text{lattice}}$ (Telesford et al., 2011), the Humphries-Gurney index σ (Humphries & Gurney, 2008), Leiden modularity Q at resolutions $\gamma \in \{0.5, 1.0, 1.5, 2.0\}$ (Traag et al., 2019), degree skewness, the Clauset-Shalizi-Newman power-law and lognormal fits to the degree distribution (Clauset et al., 2009), degree assortativity r (Newman, 2002), and the rich-club coefficient at the median degree (Colizza et al., 2006).

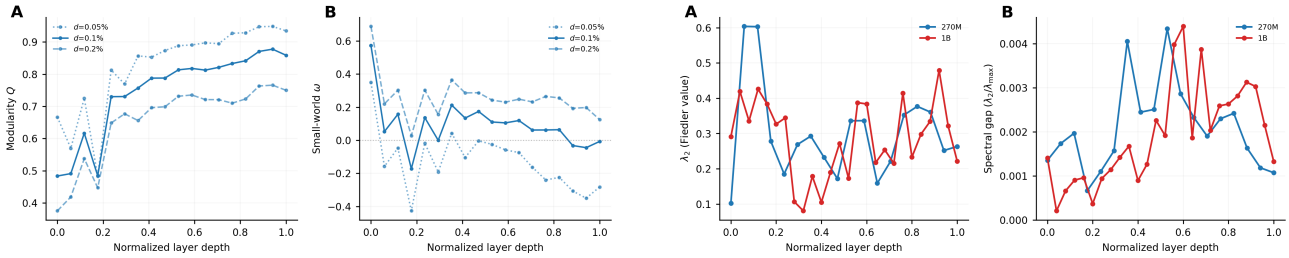
Null Models. Every metric is tested against two null models using fifty realisations per layer. The Erdős-Rényi null preserves the number of nodes and edges. The Configuration Model null additionally preserves the degree sequence (Newman, 2018), providing a more demanding baseline because any property that survives this comparison cannot be explained by the degree distribution alone.

Percolation Protocol. Percolation removes 0 to 50 percent of nodes under three strategies and tracks the fraction of the largest connected component remaining. Random removal averages over twenty independent trials. Degree-targeted removal recalculates degrees after each step, and betweenness-targeted removal recalculates betweenness centrality after each step. A Configuration Model percolation control is run in parallel to separate genuine topological vulnerability from effects that follow from the degree distribution alone.

Structure-Function Coupling. Following Liégeois et al. (2020), we report three complementary measures. The Mantel correlation between the structural and functional matrices uses node-label permutation to generate the null distribution, which tests whether structurally connected feature pairs have higher co-activation than expected. The conditional analysis computes the mean functional connectivity inside the top k percent of structural edges for $k \in \{0.1, 0.5, 1, 5, 10\}$, normalised by the global mean, to test whether alignment concentrates at the strongest structural ties. A threshold-free check uses the Fiedler value λ_2 and the normalised spectral gap of the unthresholded Pearson matrix, which display the same depth dependence as the thresholded metrics without any threshold choice.

Table 2. Layerwise summary of density-matched ($d = 0.001$) topology metrics across all layers.

Model	C	L	ω	Q (Leiden)	r	deg. skew
Gemma 3 270M	0.25 to 0.46	4.0 to 7.7	-0.17 to $+0.57$	0.48 to 0.88	0.18 to 0.78	1.9 to 7.8
Gemma 3 1B	0.30 to 0.83	1.9 to 7.6	-0.43 to $+0.32$	0.09 to 0.93	-0.43 to 0.82	0.3 to 8.1



(a) Density robustness ($d = 0.0005, 0.001, 0.002$).

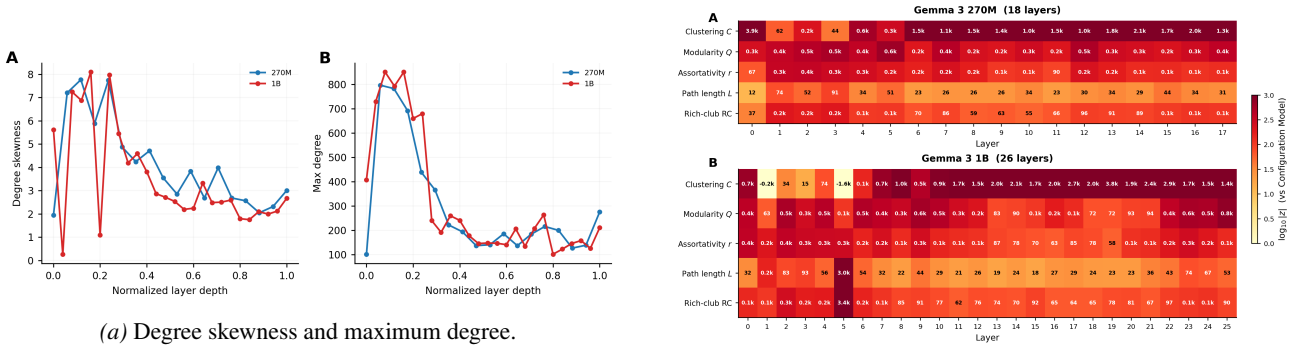
(b) Threshold-free Laplacian spectrum.

Figure 8. Robustness checks. (a) Modularity Q and small-world ω at three density levels for Gemma 3 270M, showing depth trends consistent across densities. (b) Fiedler value λ_2 and normalised spectral gap on unthresholded Pearson matrices, which display the same depth dependence as the thresholded metrics without any threshold choice.

Renormalization Group Details. For box covering, we run the greedy algorithm of Song et al. (2007) on each layer's largest connected component, sweeping the box diameter l_B from 2 to the graph diameter and fitting the scaling law $N_B(l_B) \sim l_B^{-d_B}$. To test for dual-regime scaling, we additionally fit a piecewise linear model in log-log space with a breakpoint l_B^* chosen by minimising the total residual sum of squares. The piecewise fit is accepted only when $\Delta\text{AIC} > 10$ relative to the single-regime fit, and the local exponent is bootstrapped to obtain a 95% confidence interval. The Laplacian renormalization group of Villegas et al. (2023) builds the normalised Laplacian $L_{\text{norm}} = I - D^{-1/2}AD^{-1/2}$ and the diffusion kernel $K(\tau) = \exp(-\tau L_{\text{norm}})$, then merges nodes that share diffusive modes by agglomerative clustering on the rows of $K(\tau)$. The compression ratio, defined as original nodes divided by supernodes, measures structural redundancy at scale τ.

Semantic Interpretation Pipeline. For each layer we select the top twenty hub features by degree in the canonical $d = 0.001$ graph and feed 4,096 samples from Pile-uncopyrighted through the hook-based pipeline, recording the top ten (token, context, activation) triples per hub. For Laplacian renormalization group communities of size at least five, we take the top fifty communities per layer and pick five representative features by within-community weighted degree, running the same activation-recording pipeline for each.

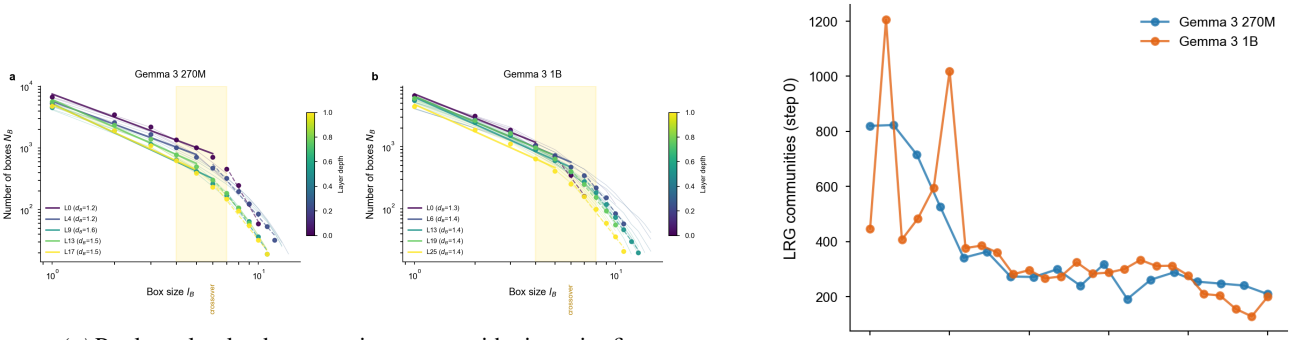
Reproducibility. The streaming co-activation tracker is in `src/network_construction/activation_based.py`. The weight-based and cross-layer network builders are in `src/network_construction/weight_based.py`. Topology analysis scripts (`compute_metrics.py`, `percolation.py`, `spectral_analysis.py`, `structure_function_coupling.py`, `null_model_significance.py`) and renormalization group scripts (`box_covering.py`, `lrg.py`, `heat_capacity.py`) reproduce every number reported in this paper. Hub and community semantic labels are produced by `hub_interpretation.py` and `community_interpretation.py`. All experiments run in a single environment on an NVIDIA RTX 5090 with CUDA 12.8 and a recent PyTorch build.



(a) Degree skewness and maximum degree.

(b) Configuration Model z-score heatmap.

Figure 9. Degree distribution and Configuration Model significance. (a) Degree skewness and maximum degree as a function of layer depth, with the two anomalous layers in 1B (layer 1 and layer 5) visible as outliers. (b) Configuration Model z-score heatmap for clustering and modularity across all layers and density levels, with almost every cell in the high-significance range.



(a) Per-layer log-log box-covering curves with piecewise fits.

(b) LRG supernode count vs depth.

Figure 10. Renormalisation group supplementary plots. (a) Box-covering scaling curves for representative layers, with piecewise linear fits and the crossover band shaded around $l_B^* \approx 6$. (b) Laplacian renormalisation group supernode count as a function of layer depth. The monotonic decrease shows progressively more compressible structure in deeper layers.

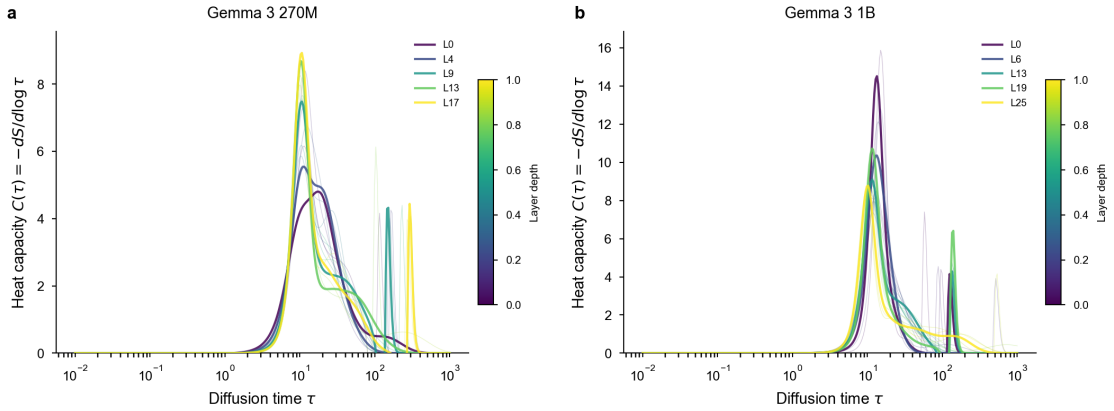


Figure 11. Spectral heat capacity $C(\tau)$ for each layer of Gemma 3 270M and 1B. All curves show a single peak (no plateau), ruling out scale invariance. Peaks shift leftward and grow taller with depth.

(a) Hub features (Gemma 3 1B). *(b)* LRG communities (Gemma 3 1B).

Figure 12. Semantic interpretation in Gemma 3 1B. The depth-dependent semantic pattern matches 270M, with HTML and multilingual tokens taking the early role and Latin and morphological roots taking the late role.