ICML 2026 WORKSHOP

The Connectome of a Large Language Model

Ruixuan Deng · **Zehao Jin**

Georgia Institute of Technology, Atlanta, GA · {rdeng62, zjin350}@gatech.edu

1 The Question

Do systems shaped by **very different optimization** — evolution vs. gradient descent — converge on the same internal organization? Sparse transcoders give the interpretable, neuron-like units to test cortical network signatures inside an LLM.

4 / 4	**16,384**	**$d_B \approx 1.35$**
cortical signatures reproduced in an LLM	features per layer Gemma Scope 2, 2 scales	fractal within modules (cortex ≈ 2.0)

2 From Features to Networks

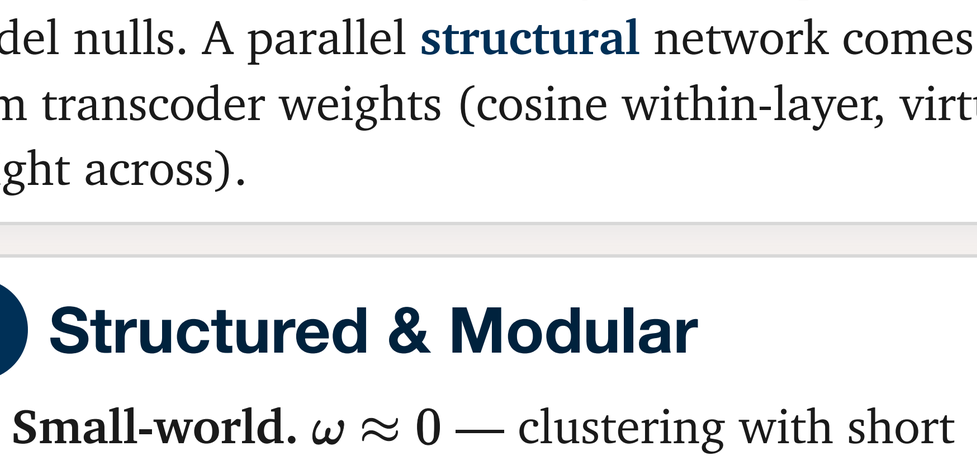
Each interpretable **transcoder feature** is a node; edges are co-activation across a corpus — an empirical network read with the network-neuroscience toolbox.

DECOMPOSE Gemma Scope 2 · 16,384 features / layer → **CO-ACTIVATE** 1M tokens · 16K×16K Pearson → **ANALYZE** topology + Laplacian RG

Gemma 3 **270M** (18L) & **1B** (26L); density-matched d=0.001, tested vs. Erdős–Rényi & Configuration-Model nulls. A parallel **structural** network comes from transcoder weights (cosine within-layer, virtual weight across).

3 Structured & Modular

- **Small-world.** $\omega \approx 0$ — clustering with short paths, like cortex.
- **Scale-free hubs.** Degree up to **47× / 53×** the mean (270M / 1B).
- **Modularity rises with depth.** Leiden Q: 0.48 → 0.88 / 0.46 → 0.93; null z up to **808**.
- **Built-in control.** Two 1B global-attention layers are clear outliers (Q=0.09, 0.11).



Hub vulnerability. Random removal keeps ≈40% of nodes at 50% removed; **targeted** hub removal collapses the largest connected component.

4 Same Signatures as the Cortex

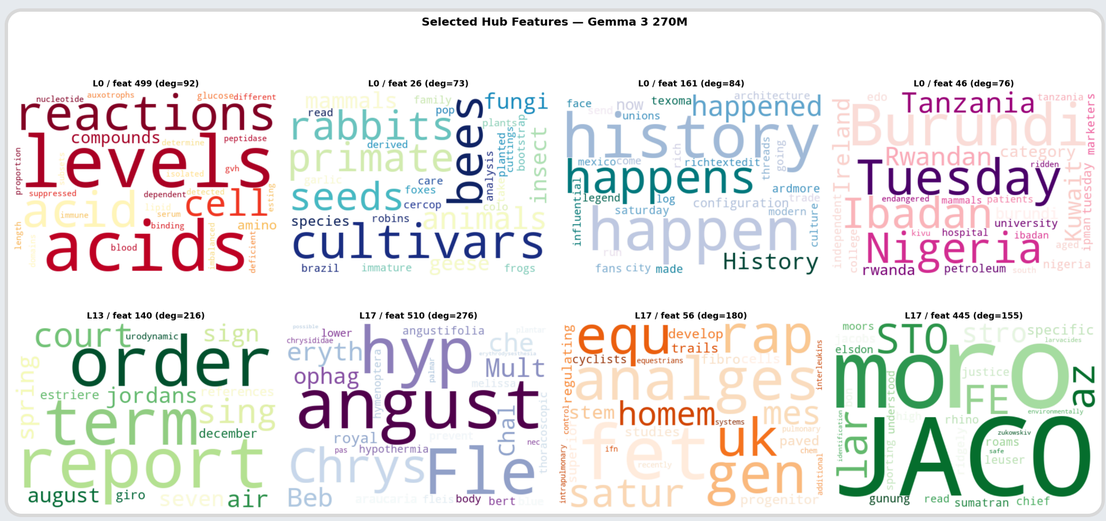
Property	Brain networks	LLM features
Small-world	$\omega \approx 0$	$\omega \approx 0$
Increasing modularity	primary → assoc.	**early → late layers**
Fractal in modules	$d_B \approx 2.0$	$d_B \approx 1.35$
Hub vulnerability	hub lesions	**hub removal → collapse**
Structure ≠ function	DTI ≠ fMRI	**cosine ≠ Pearson**
Scale structure	broad plateau	$C(\tau)$ **single peak**

Four signatures reproduced across a **4×** scale jump; two gaps (lower d_B, **single-peak** vs. plateau spectrum) mark **sparser** wiring.

5 Multi-Scale Structure & Flow

- **Dual-regime fractal.** $d_B \approx 1.35$ within modules (cortex ≈2.0), crossover $l_B^* \approx 6$; a single $C(\tau)$ peak sharpens with depth.
- **Structure ≠ function.** Mantel $r \approx 0$ globally, yet top-0.1% edges carry **20–25×** the mean; flow runs focused → broadcast → focused.
- **Compressible.** Laplacian RG merges up to **38–42×** (270M) / 62× (1B) in deep layers.

6 Topology Carries Meaning



Hubs carry meaning: early-layer content words (top) → mega-hubs (34–47×) → digits → late-layer morphological roots (bottom).

Different optimizers, similar networks: transformer feature networks share several statistical signatures with cortex — a basis for *network-theoretic* interpretability.

Contact: **rdeng62@gatech.edu** · **zjin350@gatech.edu**