Probing Neural Topology of Large Language Models

Yu Zheng [1] **Yuan Yuan** [2] **Yue Zhuo** [1] **Yong Li** [3] **Gabriel Kreiman** [4] **Tomaso Poggio** [1] **Paolo Santi** [1]

Abstract

Probing large language models (LLMs) has yielded valuable insights into their internal mechanisms by linking neural activations to interpretable semantics. However, the complex mechanisms that link neuron's functional co-activation with the emergent model capabilities remains largely unknown, hindering a deeper understanding and safer development of LLMs. In this work, we introduce graph probing, a method for uncovering the functional connectivity of LLM neurons and relating it to language generation performance. By probing models across diverse LLM families and scales, we discover a universal predictability of language generation and understanding performance using only neural topology, which persists even when retaining just 1% of neuron connections. Strikingly, probing on topology outperforms probing on activation by up to 130.4% and 67.7% on perplexity and space/time semantic regression respectively, suggesting that neural topology contains orders of richer information of LLM performance than neural activation, which can be easily extracted with simple linear or MLP probes. To explain the dependence between neural topology and language performance, we identify default networks and hub neurons in LLMs and provide causal evidence by interventional experiments on multiple benchmarks, showing that LLMs actually exploit these topological information. Further analyses suggest that graph probing can be effectively leveraged to improve the efficiency and reliability of LLMs through proof-of-concept applications in model pruning and hallucination detection. Codes and data for the graph probing toolbox are available at `https://github.com/DavyMorgan/llm-graph-probing`.

[1]Massachusetts Institute of Technology, Cambridge MA, USA [2]New York University, New York City NY, USA [3]Tsinghua University, Beijing, China [4]Harvard Medical School, Boston MA, USA. Correspondence to: Yu Zheng <yu_zheng@mit.edu>.

Preprint. January 29, 2026.

1. Introduction

Large language models (LLMs) exhibit remarkable generative capabilities (Wei et al., 2022; Touvron et al., 2023; GLM et al., 2024; Team et al., 2024; Guo et al., 2025), yet our understanding of how they succeed and what they have learned remains limited (Sharkey et al., 2025). *Probing*, which extract interpretable features from neural activations (Alain & Bengio, 2017; Rogers et al., 2020; Hewitt & Liang, 2019; Voita & Titov, 2020; Pimentel et al., 2020), has emerged as a powerful approach for reverse-engineering LLMs (Belinkov, 2022; Gurnee & Tegmark, 2024). For instance, Gurnee *et al.*(Gurnee & Tegmark, 2024) showed that LLMs encode a compact world model of space and time using linear regression probes. Unsupervised probing, such as sparse auto-encoders (Engels et al., 2025; Gao et al., 2025; Rajamanoharan et al., 2024; Lieberum et al., 2024; Mudide et al., 2025) and cross-layer transcoders (Dunefsky et al., 2024; Lindsey et al., 2025), have further revealed dictionaries of interpretable, mono-semantic concepts (Huben et al., 2024) and even causal circuits (Marks et al., 2025), corresponding to directions in neural latent space. While these advances shed light on the semantics of neural activations (Sharkey et al., 2025), much less is known about how neurons are functionally connected, *i.e.* the neural topology, which is believed to play an essential role in the emergence of intelligence (Rathi et al., 2025; Bassett & Sporns, 2017).

Recent studies have drawn compelling parallels between neurons in LLMs and those in the human brain (Toneva & Wehbe, 2019; Schrimpf et al., 2021; Caucheteux et al., 2023; Kumar et al., 2024; Rathi et al., 2025; Mischler et al., 2024; Tuckute et al., 2024; Bonnasse-Gahot & Pallier, 2024; Sun et al., 2024a; Liu et al., 2025), revealing shared properties such as spatial-functional organization (Kumar et al., 2024; Rathi et al., 2025) and left lateralization (Bonnasse-Gahot & Pallier, 2024). Neural activations at internal layers of LLMs have also been shown to reliably predict human brain responses given the same linguistic stimuli (Schrimpf et al., 2021; Tuckute et al., 2024; Luo et al., 2022). However, these efforts primarily focus on static neural activations of LLMs, while overlooking the key aspect of temporal and functional neural topology that has been studied in neuroscience for decades (Bassett & Bullmore, 2006; Bassett & Sporns, 2017; Fotiadis et al., 2024). Moreover, although analogies between LLMs and human brains are insightful (Toneva &

Wehbe, 2019; Goldstein et al., 2022), few works explicitly connect these findings to the language generation performance, which is one of the primary indicators of an LLM's intelligence.

In this work, we introduce *graph probing*, a novel approach for investigating the functional connectivity of neurons in LLMs and its relationship to language generation and understanding performance. By analyzing neural activity time series as LLMs process text token by token, we compute temporal and functional correlations between neurons to construct dynamic neural graphs. Using this large-scale dataset of text-induced neural topology, we train probes to predict LLMs' accuracy in auto-regressively generating the corresponding text, as well as how LLMs represent fundamental concepts like space and time. In essence, graph probing connects the micro-level topology of how neurons are connected given a token sequence, to the macro-level performance of how well LLMs understand and predict these tokens, offering a new lens to study the emergent capabilities of LLMs. Our method is summarized in Figure 1 and described in Section 2.

We then apply our graph probing framework to comprehensively analyze the neural topology of LLMs through extensive experiments. First, we demonstrate that language understanding and generation performance can be reliably predicted using only the neural connectivity graph. This predictability holds universally across LLM families and scales, outperforming activation-based probing approaches by up to 130.4% even when preserving only 1% of neuron connections, with empirical results spanning GPT (Radford et al., 2019), Pythia (Biderman et al., 2023), and Qwen (Yang et al., 2024), ranging from millions to billions of parameters (Section 3). Next, we show causal evidence through interventional experiments and analysis on the MMLU (Hendrycks et al., 2021; Wang et al., 2024) benchmark, discovering stable default neural topology and hub neurons in LLMs, and more importantly, validating that LLMs actually utilize their internal topological information when generating text (Section 4). Finally, we offer two proof-of-concept applications of graph probing, showcasing its potential in model pruning and hallucination detection (Section 5). While not without limitations, we expect graph probing to provide valuable insights into the inner workings of LLMs and to guide their future development in an interpretable and reliable way.

2. Graph Probing

Neural Topology. To construct neural graphs from LLMs, we draw inspiration from neuroscience where functional brain networks are derived from temporal correlations in fMRI or EEG activation signals (Bassett & Sporns, 2017; Vértes et al., 2012; Bullmore & Bassett, 2011), as shown in Figure 1. Given an LLM composed of stacked attention layers, the neural topology is constructed as follows:

$$\mathbf{H} = [\mathbf{h_0}, \mathbf{h_1}, \ldots, \mathbf{h_t}] \in \mathbb{R}^{n \times t}, \tag{1}$$

$$\mathbf{A} = \left(a_{ij}\right) \in \mathbb{R}^{n \times n}, \tag{2}$$

$$a_{ij} = \rho(\mathbf{H}_{i,:}, \mathbf{H}_{j,:}) \tag{3}$$

$$= \frac{\sum_{k=0}^{t} (\mathbf{H}_{i,k} - \overline{\mathbf{H}_{i,:}})(\mathbf{H}_{j,k} - \overline{\mathbf{H}_{j,:}})}{\sqrt{\sum_{k=0}^{t} (\mathbf{H}_{i,k} - \overline{\mathbf{H}_{i,:}})^2} \sqrt{\sum_{k=0}^{t} (\mathbf{H}_{j,k} - \overline{\mathbf{H}_{j,:}})^2}}, \tag{4}$$

where neurons at each layer produce a time series of hidden states $\mathbf{H}$ as the model processes a token sequence $X = [x_0, x_1, \ldots, x_t]$ (n and t represent the number of neurons and tokens), and the temporal co-activation patterns among neurons define their *functional connectivity*. We capture this through a complete $n \times n$ weighted connectivity matrix $\mathbf{A}$, where each node corresponds to a neuron and each edge weight a_{ij} represents the Pearson correlation coefficient between the activation time series of a pair of neurons i and j (Bassett & Sporns, 2017; Fotiadis et al., 2024).

Probing on Neural Topology. We propose *graph probing* to study the dependence between LLM performance and neural topology. Specifically, we adopt simple linear or multi-layer perceptrons (MLP) probes (Rumelhart et al., 1986) that take neural topology as input to predict its corresponding language understanding performance, as illustrated in Figure 1. Given a connectivity matrix $\mathbf{A}$ induced by feeding a tokenized sequence X to an LLM, where each element a_{ij} denotes the functional connectivity (Pearson correlation coefficient) between neurons i and j, our probe produces the graph representation $\mathbf{z}$ as follows:

$$\textbf{Linear}: \hat{p} = \mathbf{W}_1 \cdot \texttt{Flatten}(\mathbf{A})^T, \tag{5}$$

$$\textbf{MLP}: \quad \hat{p} = \mathbf{W}_3 \cdot \texttt{ReLU}(\mathbf{W}_2 \cdot \texttt{Flatten}(\mathbf{A})^T), \tag{6}$$

where $\texttt{Flatten}(\mathbf{A}) \in \mathbb{R}^{(n \times n)}$ is the flattened topology matrix, $\mathbf{W_1} \in \mathbb{R}^{1 \times (n \times n)}$, $\mathbf{W_2} \in \mathbb{R}^{d \times (n \times n)}$, $\mathbf{W_3} \in \mathbb{R}^{1 \times d}$ are weight matrices (d as a hyper-parameter), $\hat{y}$ is the prediction of language performance. In essence, the core and only difference of our method lies at the input, where we probe on neural topology instead of neural activation by existing approaches (Belinkov, 2022; Gurnee et al., 2023). We will later show that topology contains orders of richer information than activation regarding LLMs' intelligence.

Probing Target. By its definition, graph probing can be utilized to predict almost any fact of interest, and in this work we test it on various performance-related metrics, including perplexity, hallucination, and functional specialization, as well as fundamental semantics like space and time. Here we introduce the case of perplexity (PPL) as it is a fundamental metric directly reflecting LLMs' auto-regressive language generation performance (Bengio et al.,

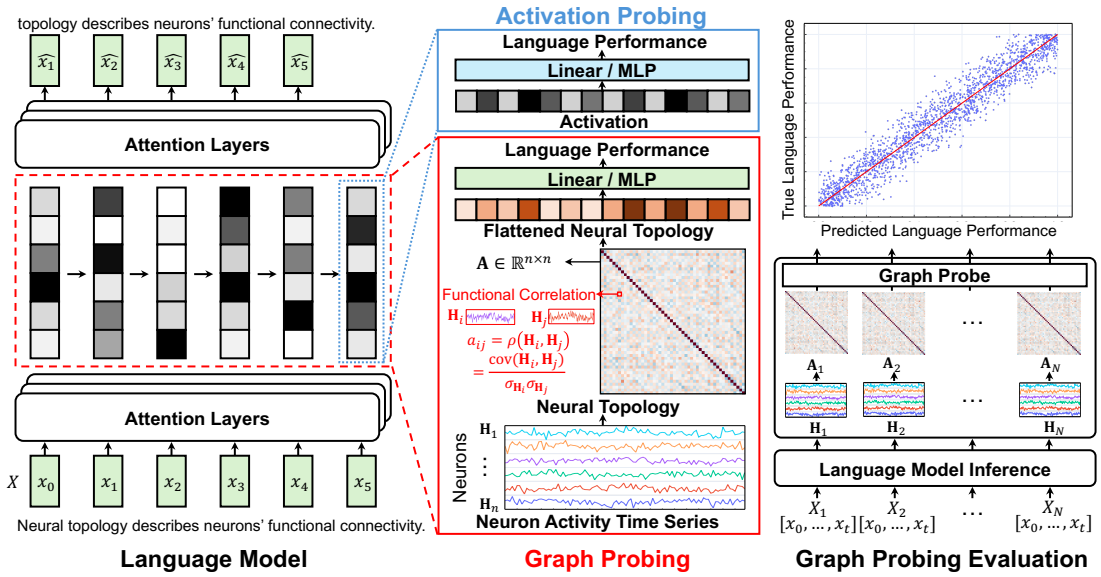


Figure 1. Overview of graph probing. We extract the neuron activity time series in an LLM as it processes text token by token. We then compute temporal and functional correlations between neural activations to obtain topological connectivity graphs of neurons. Unlike existing probing methods that take neural activation as input, we train linear or MLP probes on flattened neural topology to predict the language generation performance for the input token sequence.

2003), while details of other cases are provided in later sections. Specifically, given a token sequence, the perplexity is commonly calculated as the exponentiated average negative log-likelihood over the token sequence:

$$\mathtt{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i=1}^{t}\log p_\theta(x_i \mid x_{<i})\right\}. \quad (7)$$

The neural topology is dynamically induced by the specific token sequence, and our goal is to investigate whether the text-responsive neural topology is linked to how well the model predicts the text. Towards this end, we train the graph probe to minimize the mean squared error (MSE) between predicted and true perplexities over a dataset of tokenized sequences $\mathbf{X} = \{X_1, \ldots, X_N\}$:

$$\mathcal{L}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{p}_i - \mathtt{PPL}(X_i)\right)^2, \quad (8)$$

where the prediction $\hat{p}_i$ is calculated via the graph probe by Equation (5) or (6). Details of hyper-parameters and computational configurations are provided in Appendix A.1.

3. Results

LLMs. In our experiments, we train graph probes on neural topology derived from three families of LLMs, each spanning across different sizes. Specifically, we evaluate GPT2 (Radford et al., 2019) (GPT2, GPT2-large), Pythia (Biderman et al., 2023) (160M, 1.4B, 2.8B), and Qwen2.5 (Yang et al., 2024) (0.5B, 3B, 7B, 14B). Details of the experimented LLMs are provided in Appendix A.2.

Datasets. To enable our study, we construct neural topology using the OpenWebText dataset (Gokaslan et al., 2019). To ensure consistent temporal resolution, we control the length of neural activity time series to fall between 256 and 1024 tokens by merging consecutive sentences as needed. For each token sequence, we perform LLM inference to compute its perplexity and simultaneously extract hidden state time series to generate the corresponding neural topology. For each model, we construct a probing dataset comprising about 10,000 graph–perplexity pairs. Further details on dataset construction are provided in Appendix A.3.

Evaluation. We split the dataset into training and test sets using an 8:2 ratio. Having learned graph probes on the training set, we evaluate their out-of-sample prediction performance on the test set, which reveals the extent to which micro-level neural topology is predictive of macro-level language generation ability. To quantify the effectiveness of graph probing, we report standard regression metrics on our test data, including mean squared error (MSE), mean absolute error (MAE), coefficient of determination (R^2), Pearson correlation (ρ_p), and Spearman rank correlation (ρ_s). We compare with activation based baselines using both linear and MLP probes (Alain & Bengio, 2017; Belinkov, 2022; Gurnee & Tegmark, 2024; Tuckute et al., 2024).

We visualize the predicted and groundtruth perplexity in Figure 2 and summarize the results in Table 1, where graph probing consistently outperforms activation probing across all three LLM families for both linear and MLP probes. The improvements are strikingly substantial, with the maximum progress in R^2 exceeding 130.4%. For example, perplexity is barely predictable from neural activation, with R^2 less

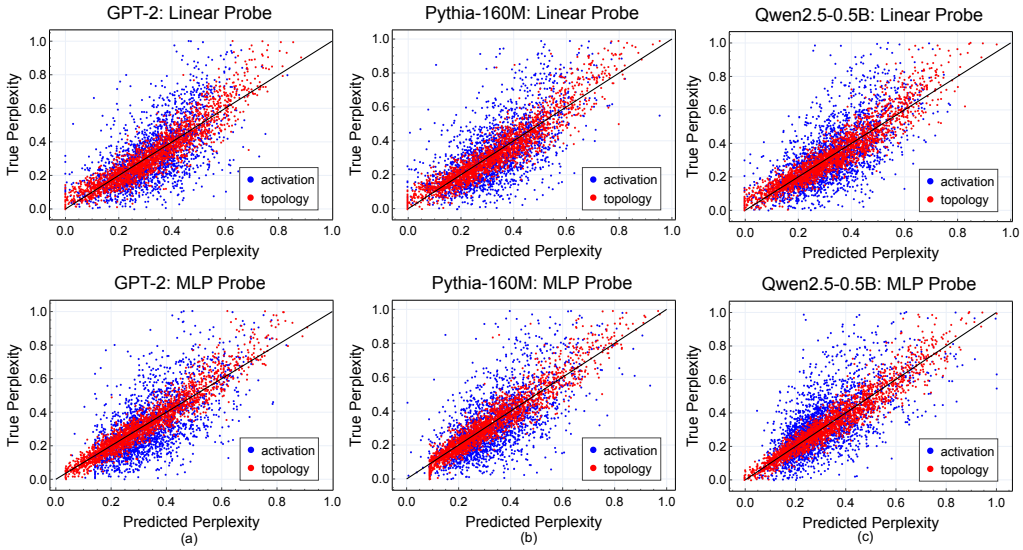


Figure 2. Out-of-sample performance of linear and MLP probing on the test set for (a) GPT-2 (b) Pythia-160M (c) Qwen2.5-0.5B. We compare activation-based probing and our topology-based probing. The correlation between the perplexity predicted by probing and the ground-truth perplexity reflects how well LLM performance can be inferred from neural activation or topology.

Table 1. Performance of different probing methods (* indicates p-value < 0.05).

LLM	Probe	MSE ↓	MAE ↓	R^2 ↑	ρ_p ↑	ρ_s ↑
GPT-2	Activation (Linear)	0.0199	0.1067	0.3987	0.6352	0.6497
	Activation (MLP)	0.0201	0.1056	0.3930	0.6370	0.6426
	Graph (Linear)	**0.0049***	**0.0526***	**0.8517***	**0.9229***	**0.9320***
	Graph (MLP)	**0.0031***	**0.0399***	**0.9057***	**0.9534***	**0.9577***
Pythia-160M	Activation (Linear)	0.0199	0.1061	0.4151	0.6517	0.6504
	Activation (MLP)	0.0194	0.1030	0.4304	0.6618	0.6519
	Graph (Linear)	**0.0047***	**0.0518***	**0.8607***	**0.9278***	**0.9364***
	Graph (MLP)	**0.0036***	**0.0432***	**0.8930***	**0.9458***	**0.9499***
Qwen2.5-0.5B	Activation (Linear)	0.0190	0.1028	0.4225	0.6536	0.6632
	Activation (MLP)	0.0196	0.1045	0.4044	0.6496	0.6457
	Graph (Linear)	**0.0045***	**0.0496***	**0.8640***	**0.9296***	**0.9422***
	Graph (MLP)	**0.0030***	**0.0396***	**0.9095***	**0.9538***	**0.9583***

than 0.45 for all models, while that of graph probing all close to or even larger than 0.90. The enormous gain of graph probing validates the hypothesis that neural topology contains much richer information of LLMs' language generation performance than neural activation, which can be easily extracted using simple linear or MLP probes.

Besides predicting perplexity, we also extend our analysis to include a wider range of semantics, probing neural topology to predict time and space context of the input text. We adopt the space and time datasets utilized by (Gurnee & Tegmark, 2024). We construct the neural topology for input texts regarding historical artworks and world places, and trained probes to predict: (1) time: the release year of historical artworks; (2) space: the longitude and latitude of world places. The results are summarized in Table 2, which serve two critical purposes. Firstly, they validate findings from previous work (Gurnee & Tegmark, 2024) showing that LLMs possess internal representations of time and space. Secondly and crucially, they confirm that neural topology contains significantly richer information than neural activation for these semantic tasks. The substantial performance margin of as much as 67.77% suggests that the structure of computation (how neurons connect) is more predictive than the magnitude of computation (activations) for extracting high-level concepts.

Sparsity and scalability. Probing on complete graphs, *i.e.*, dense $n \times n$ connectivity matrices that capture pairwise functional correlations between all neurons, can become computationally prohibitive as the LLM size increases, due to the quadratic number of edges that directly impacts the

Table 2. Probing results on Arts (time) and World Places (space) datasets. We compare Average Activation and our Graph Probing method (* indicates p-value < 0.05).

LLM	Probe	Arts (time)			World Places (space)		
		MSE	MAE	R^2	MSE	MAE	R^2
GPT-2	Activation (Linear)	0.0406	0.1596	0.2945	0.0360	0.1957	0.5106
	Activation (MLP)	0.0382	0.1511	0.3359	0.0311	0.1727	0.5733
	Graph (Linear)	**0.0326***	**0.1371***	**0.4341***	**0.0317***	**0.1765***	**0.5689***
	Graph (MLP)	**0.0267***	**0.1230***	**0.5361***	**0.0258***	**0.1514***	**0.6486***
Pythia-160M	Activation (Linear)	0.0399	0.1580	0.3071	0.0360	0.1968	0.5116
	Activation (MLP)	0.0391	0.1560	0.3211	0.0306	0.1705	0.5812
	Graph (Linear)	**0.0282***	**0.1275***	**0.5100***	**0.0267***	**0.1583***	**0.6335***
	Graph (MLP)	**0.0266***	**0.1234***	**0.5387***	**0.0246***	**0.1447***	**0.6636***
Qwen2.5-0.5B	Activation (Linear)	0.0325	0.1406	0.4359	0.0296	0.1747	0.6002
	Activation (MLP)	0.0314	0.1368	0.4555	0.0268	0.1574	0.6367
	Graph (Linear)	**0.0258***	**0.1221***	**0.5524***	**0.0215***	**0.1364***	**0.7080***
	Graph (MLP)	**0.0268***	**0.1254***	**0.5341***	**0.0223***	**0.1398***	**0.6953***

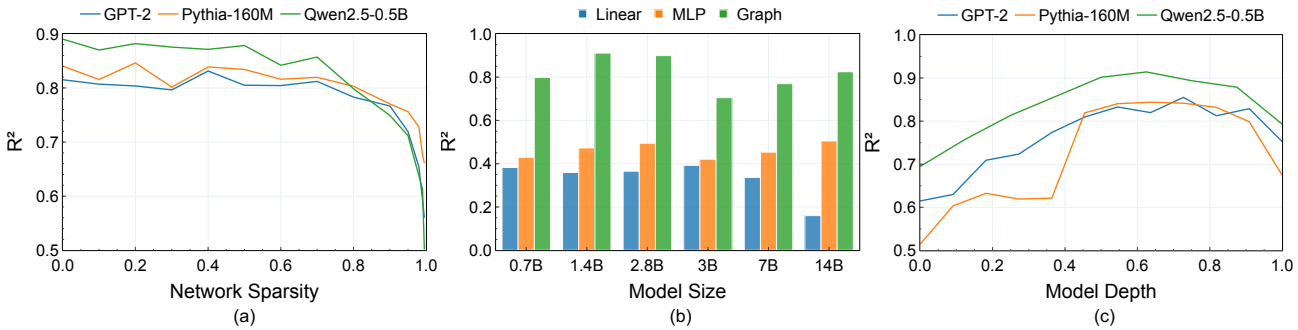


Figure 3. (a) Out-of-sample graph probing performance on neural topology of different sparsity levels. (b) Out-of-sample probing performance on LLMs of different sizes. (c) Out-of-sample performance of graph probing on different layers of LLMs.

computational cost in both time and memory. For instance, while complete graph probing is feasible for Pythia-160M with 768 neurons and 0.6M edges per layer, the number of edges in Qwen2.5-14B–comprising 5,120 neurons per layer–explodes to over 26M per graph. To address this, on the one hand, we investigate whether probes can be applied to sparse graphs with weakly correlated edges pruned out by thresholding, which is commonly employed in human brain network construction (Bassett & Sporns, 2017) and such sparsity is frequently observed in artificial neural networks (Vig et al., 2020; Timkey & van Schijndel, 2021; Dettmers et al., 2022); on the other hand, we develop graph neural network probe (Kipf & Welling, 2017) that reduces the number of probe parameters from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \cdot d)$ by weight sharing (details in Appendix A.4). We train graph probes on neural topology with varying levels of sparsification (Figure 3(a)) for the perplexity task. Surprisingly, the predictive performance remains remarkably stable even after removing up to 90% of the edges, with minimal degradation. Notably, even under extreme sparsity where only 1% of the original edges are retained, the neural topology still enables effective prediction of perplexity, achieving above 0.6 R^2 that is still higher than activation probing.

The above experiments suggest that most of the predictive signal resides in a small subset of strong connections, making it possible to significantly reduce the number of edges while preserving nearly all critical topological information. Leveraging this insight, we scale up graph probing to much larger models by operating on sparsified neural topology. While our earlier results focused on models with fewer than 0.5B parameters, we now train probes on sparse graphs derived from LLMs with up to 14B parameters (GPT2-large, Pythia-{1.4B, 2.8B}, Qwen2.5-{3B, 7B, 14B}). As shown in Figure 3(b), graph probing continues to exhibit strong regression performance across all six large models, achieving a maximum R^2 of over 0.91, providing compelling evidence that the relationship between neural topology and language modeling performance is universal across model sizes. Particularly, the gap between baselines and graph probing remains as substantial as 92.6%, confirming the informativeness of neural topology.

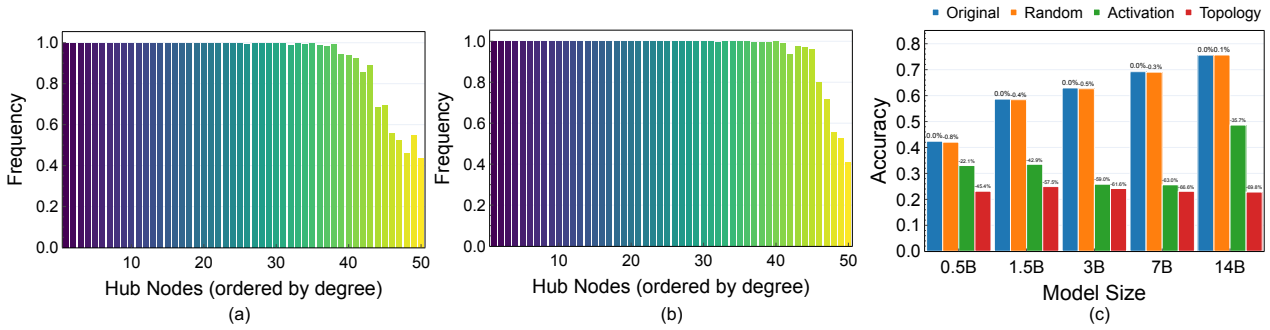


Figure 4. (a-b) Occurrence frequency of hub nodes in (a) Qwen2.5-0.5B and (b) Qwen2.5-1.5B on MMLU benchmark. (c) Accuracy on MMLU benchmark of Qwen2.5 models (0.5B, 1.5B, 3B, 7B, 14B) under different interventions of top 1% neurons.

Probing different layers. We further train probes on neural topology derived from different layers of GPT-2, Pythia-160M, and Qwen2.5-0.5B models. Figure 3(c) reveals that while topologies from all layers are predictive of LLM performance in terms of perplexity, the middle layers are consistently the most informative across all three models, with R^2 values exceeding 0.8. This finding aligns with prior work identifying the middle layers of LLMs as the hub of semantic processing, whereas the initial and final layers are more specialized for {de, re}-tokenization (Gurnee & Tegmark, 2024; Tuckute et al., 2024).

4. Topological analysis and causal intervention

While our probing experiments discover the correlation between neural topology and language generation performance, the specific underlying topological structures remain unclear. In neuroscience, human brains exhibit a stable intrinsic network–a core connectivity pattern that persists to great extent across different tasks and stimuli (Fotiadis et al., 2024; Cole et al., 2014). Inspired by this, we investigated whether a similar default network of highly connected neurons exists within LLMs. To test this hypothesis, we first calculated an average neural topology for the Qwen2.5-0.5B and Qwen2.5-1.5B models over the entire OpenWebText dataset. From this average, we identified the top 50 "hub" neurons (those with the highest degree) and then measured their occurrence frequency, *i.e.*, how often these same neurons ranked as hubs in the topology of each individual data sample. The results, shown in Figure 4(a-b), are striking. A core set of nearly 40 neurons remain hubs in 100% of the topologies across the entire dataset. This remarkable stability confirms the existence of a default network in LLMs, where a fixed set of hub neurons play dominant roles regardless of the input, suggesting a fundamental organizing principle of their internal structure.

To determine if LLMs rely on their neural topology, we conducted a series of interventional experiments. We disabled a selected 1% of neurons in the middle layer of Qwen2.5 models (0.5B, 1.5B, 3B, 7B, 14B) by pinning their activations to zero at all tokens and measured the impact on the MMLU benchmark (Hendrycks et al., 2021). We compared three distinct neuron selection strategies: random selection, selecting neurons with the highest activation, and selecting neurons with the highest topological degree. As shown in the Figure 4(c), the results demonstrate a clear hierarchy of neuronal importance. While disabling random neurons caused a negligible accuracy drop (<1.0%), targeting neurons by either activation or topology incurred a substantial performance collapse of at least 20%. Critically, the topology-based intervention was the most detrimental. Disabling the top 1% of hub neurons resulted in a catastrophic accuracy drop of 45.4%, 57.5%, 61.6%, 66.6%, and 69.8% for the five tested models, consistently outperforming the activation-based strategy across all scales. Particularly, for the largest 14B model, the relative performance drop against activation based intervention is over 95.5%. These experiments provide causal evidence that LLMs actively utilize their underlying topological structure, particularly their hub neurons, for computation.

Our analysis also revealed a notable phenomenon of functional specialization (Cole et al., 2014), where neural topology displays distinct patterns for different subjects. We found a striking example by identifying certain neuron that consistently specializes in STEM subjects, while remaining almost entirely dormant for social science subjects. This suggests a subject-based organization within LLMs, and the details are provided in Appendix A.5.

5. Applications

Model pruning. Model pruning is a standard technique for optimizing LLMs in real-world deployments that prioritize low latency and computational efficiency (Sun et al., 2024b; Ma et al., 2023; Gao et al., 2024; Muralidharan et al., 2024; Xia et al., 2024; Xu et al., 2024b). The finding that high-degree neurons are functionally dominant (Section 4) naturally suggests a pruning strategy of disabling low-degree neurons. To test this, we prune these neurons by zeroing out their activations during inference. We ap-

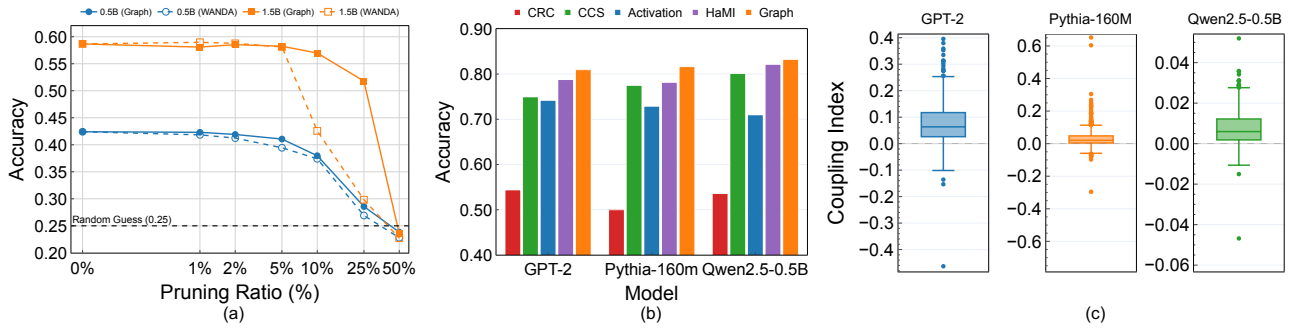


Figure 5. (a) Accuracy on MMLU benchmark under different levels of model pruning based on neural topology and activation (WANDA). (b) Accuracy of hallucination detection of different approaches on TruthfulQA dataset. (c) Coupling index of neural topology for hallucination on TruthfulQA datset.

ply this technique to a middle layer of Qwen2.5-0.5B and Qwen2.5-1.5B and evaluate the performance on the MMLU benchmark (Hendrycks et al., 2021). We compare our topology based pruning approach with a state-of-the-art LLM pruning method WANDA (Sun et al., 2024b) which is based on neural activation. The results, presented in Figure 5(a), confirm the models' robustness. Performance degradation is minimal even with substantial pruning; for the 1.5B model, accuracy drops by just 2.86% when 10% of neurons are pruned, and by 11.73% when 25% are removed. Strikingly, the models maintain above-random-guess accuracy until half of the low-degree neurons are pruned. In addition, our topology based pruning approach consistently outperforms WANDA, particularly in the 1.5B model where our method retains over 51% accuracy compared to WANDA's 30% accuracy when pruning 25% of the neurons. These results highlight the potential for developing more sophisticated pruning methods based on neural topology.

Hallucination detection. Building on our finding that subtle topological patterns in LLMs can predict text generation performance, we now investigate if these patterns can also detect hallucination, a critical challenge that has been extensively explored through various probing techniques (Burns et al., 2023; Du et al., 2024; Hou et al., 2025; Su et al., 2024; Sriramanan et al., 2024; Orgad et al., 2025; Farquhar et al., 2024; Niu et al., 2025). To create distinct *genuine* and *hallucinating* states, we construct inputs from the TruthfulQA dataset (Lin et al., 2022) by concatenating each of its 817 question with its corresponding true and false answers, respectively, resulting in a dataset of 5918 samples (true/false classification). We then extract the neural topology from the LLM as it processes these inputs (Equations 1-4). For this task, we adapt probing from regression to binary classification by modifying the MLP probe layer to have two output channels ($\mathbf{W}_1 \in \mathbb{R}^{2\times(n\times n)}$ and $\mathbf{W}_3 \in \mathbb{R}^{2\times d}$ in Equations (5-6)) and replacing the MSE loss with a cross-entropy loss:

$$\mathcal{L}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N} \texttt{CROSS\_ENTROPY}(\hat{y}_i, y_i), \tag{9}$$

where $\hat{y}$ and y are the prediction and ground-truth for hallucination (0/1). We compare against MLP probes trained on the neural activation at the last-token position. We also include the state-of-the-art HaMI (Niu et al., 2025) and CCS (Burns et al., 2023) hallucination detection approach, as well as CCS's variant CRC (Burns et al., 2023) for comparison.

We split the dataset into training and test sets using an 8:2 ratio to evaluate our hallucination detection framework. As shown in Figure 5(b), probing neural topology substantially outperforms all baselines probing neural activation for all three tested models (GPT-2, Pythia-160M, and Qwen2.5-0.5B), with accuracy gains of up to 9.73%. This superior performance suggests that distinct topological patterns emerge when an LLM is generating a factual response versus hallucinating. To validate this hypothesis directly, we introduce a neural topology coupling index,

$$C_{XY} = \texttt{AVG}(\{\rho(A_i, A_j) | A_i \in A_X, A_j \in A_Y\}), \tag{10}$$

$$C = C_{TT} + C_{HH} - 2C_{TH}, \tag{11}$$

where the index C measures the difference between intra-group similarity (C_{TT}, C_{HH}) and inter-group similarity (C_{TH}). Here A_T and A_H represent the sets of neural topologies for truthful and hallucinated responses, respectively, and similarity is measured by the Pearson correlation (ρ) between each pair of the flattened adjacency matrices. We calculated this index for each question in the dataset, and the distribution in Figure 5(c) shows that over 80% of samples have a positive coupling index. This confirms that topologies are indeed more similar within the same state (truthful-to-truthful, hallucinated-to-hallucinated) than across different states (truthful-to-hallucinated), implying that neural topology serves as a promising and reliable signature for detecting LLM hallucinations and paving the way for future work on improving model reliability.

It is worthwhile to emphasize that our goal with such a simple topological probe is not to establish a new state-of-the-art against more nuanced approaches specifically design for either model pruning or hallucination detection, which are highly specialized domains with extensive liter-

ature (Sun et al., 2024b; Ma et al., 2023; Gao et al., 2024; Muralidharan et al., 2024; Xia et al., 2024; Xu et al., 2024b; Kuhn et al., 2023; Niu et al., 2025; Burns et al., 2023; Du et al., 2024; Farquhar et al., 2024; Sriramanan et al., 2024). Rather, these experiments serve as proof-of-concept demonstrations. Though preliminary in their current forms, these case studies showcase the versatile utility of graph probing and lay the groundwork for future work to translate these concepts into robust, system-level solutions. We hope these results encourage the community to view graph probing as a valuable tool for optimizing and interpreting LLMs.

6. Related Work

Probing LLMs. Growing concerns over the transparency and steerability of LLMs have driven recent advances in reverse-engineering LLMs by extracting interpretable features from their neural activations through probes (Sharkey et al., 2025; Alain & Bengio, 2017; Rogers et al., 2020; Hewitt & Liang, 2019; Voita & Titov, 2020; Pimentel et al., 2020). Supervised probing typically maps neuron activations to interpretable semantics through regression or classification (Gurnee et al., 2023; Gurnee & Tegmark, 2024; Jin & Rinard, 2024; Ju et al., 2024; Dong et al., 2023; Kissane et al., 2024; Templeton et al., 2024; Belinkov, 2022). For example, Gurnee *et al.* (Gurnee & Tegmark, 2024) predicted the time and location of input entities from LLM activations. Unsupervised probing, by contrast, aims to learn a dictionary of disentangled features related to more abstract concepts (Engels et al., 2025; Gao et al., 2025; Rajamanoharan et al., 2024; Lieberum et al., 2024; Mudide et al., 2025; Engels et al., 2025). A famous example is the *Golden Gate Bridge* feature identified in the Claude 3 Sonnet model (Templeton et al., 2024). While prior work focused on connecting LLM activations to external semantics, our work studies the *functional topology* of neurons in LLMs, and relates this internal structure directly to language generation performance via *graph probing*.

Network Neuroscience. The study of functional networks in the human brain has been a central topic in neuroscience for decades (Bassett & Bullmore, 2006; Bassett & Sporns, 2017; Fotiadis et al., 2024; Medaglia et al., 2015) which motivates this research. Brain networks are typically constructed by correlating fMRI or EEG signals across different neural regions, and then analyzed using tools from network science (Barabási, 2013), which has revealed a range of structural and functional properties, such as small-worldness (Bassett & Bullmore, 2006), economical wiring (Bullmore & Sporns, 2012), and functional specialization (Fotiadis et al., 2024). More recently, several studies have drawn parallels between LLM activations and human brain activity (Toneva & Wehbe, 2019; Caucheteux et al., 2023; Kumar et al., 2024; Rathi et al., 2025; Mischler et al., 2024; Tuckute et al., 2024; Bonnasse-Gahot & Pallier, 2024; Sun et al., 2024a; Liu et al., 2025). For instance, Tuckute *et al.* (Tuckute et al., 2024) used GPT-2 activations to identify sentence stimuli that drive or suppress human brain responses. However, while these efforts focus on representational similarities, the functional *topology* of neurons within LLMs and its relationship to the model's language generation capabilities remain largely unexplored.

7. Discussion

Neurons in LLMs are connected both structurally through the model's architecture and functionally through their dynamic responses to input linguistic stimuli. In this work, we focus on the latter and demonstrate that the language understanding and generation performance of LLMs can be reliably predicted from their functional neural topologies using graph probing, implying that LLMs develop intricate and consistent topological structures among their neurons that are fundamental to their emergent linguistic ability. Besides causal intervention on benchmarks validating LLMs actually leverage their internal neural topology, we also offer practical applications of neural topology in model pruning and hallucination detection. While we have empirically shown a stable *default* neural topology and hub neurons in LLMs regardless of input, we have not yet identified more nuanced structures such as motifs, or physical metrics like small-worldness and modularity within these graphs. It remains an open question whether such properties exist in LLMs' neural topology and play a causal role in shaping their intelligence. Additionally, this paper evaluates LLMs with up to 14B parameters due to computational cost, while leaving even larger models for future work.

Graph probing raises many interesting directions for future research. While we have linked neural topology to general linguistic ability, and discovered functional specialization and subject-specific neurons (Appendix A.5), it remains unclear how these specialized structures emerge during LLMs' learning process. Additionally, recent advances in enhancing LLMs' reasoning abilities (Guo et al., 2025) raise a natural question: does reasoning alter, or is it constrained by, neural topology? Finally, graph probing is model-agnostic and can be extended to LLMs of distinct architectures or even models other than LLMs, for example to vision-language models (VLMs). We provide preliminary results of graph probing for cross-LLM matching and LLM fingerprinting, as well as probing VLMs in Appendix A.6-A.8 with causal intervention as well, while further efforts are required to achieve deeper insights into their multi-modal understanding and generation capabilities. In all, we believe graph probing offers a promising lens for understanding AI models and ultimately guiding their improvement in an reliable and safe way.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning by improving our understanding of the internal mechanisms of large language models. Our work relies exclusively on publicly available, pre-trained models (e.g., GPT-2, Pythia, Qwen2.5) and standard academic benchmarks (e.g., MMLU, TruthfulQA, OpenWebText) to ensure reproducibility and avoid the use of sensitive or private data. We believe our findings have several positive societal implications. The methods for hallucination detection directly contribute to the development of more reliable, truthful, and safe AI systems. Similarly, our work on model pruning promotes computational efficiency, which can reduce the environmental impact of AI and make powerful models more accessible.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Barabási, A.-L. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1987):20120375, 2013.

Bassett, D. S. and Bullmore, E. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.

Bassett, D. S. and Sporns, O. Network neuroscience. *Nature neuroscience*, 20(3):353–364, 2017.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Bonnasse-Gahot, L. and Pallier, C. fmri predictors based on language models of increasing complexity recover brain left lateralization. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Bullmore, E. and Sporns, O. The economy of brain network organization. *Nature reviews neuroscience*, 13(5):336–349, 2012.

Bullmore, E. T. and Bassett, D. S. Brain graphs: graphical models of the human brain connectome. *Annual review of clinical psychology*, 7(1):113–140, 2011.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Caucheteux, C., Gramfort, A., and King, J.-R. Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, 7(3):430–441, 2023.

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., and Petersen, S. E. Intrinsic and task-evoked network architectures of the human brain. *Neuron*, 83(1):238–251, 2014.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in neural information processing systems*, 35:30318–30332, 2022.

Dong, X., Wang, Y., Yu, P. S., and Caverlee, J. Probing explicit and implicit gender bias through llm conditional text generation. *arXiv preprint arXiv:2311.00306*, 2023.

Du, X., Xiao, C., and Li, S. Haloscope: Harnessing unlabeled llm generations for hallucination detection. *Advances in Neural Information Processing Systems*, 37: 102948–102972, 2024.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Engels, J., Liao, I., Michaud, E. J., Gurnee, W., and Tegmark, M. Not all language model features are linear. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, Apr 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Fey, M. and Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Fotiadis, P., Parkes, L., Davis, K. A., Satterthwaite, T. D., Shinohara, R. T., and Bassett, D. S. Structure–function coupling in macroscale human brain networks. *Nature Reviews Neuroscience*, 25(10):688–704, 2024.

Fukushima, K. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, Apr 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Gao, S., Lin, C.-H., Hua, T., Tang, Z., Shen, Y., Jin, H., and Hsu, Y.-C. Disp-llm: Dimension-independent structural pruning for large language models. *Advances in Neural Information Processing Systems*, 37:72219–72244, 2024.

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Zhang, D., Rojas, D., Feng, G., Zhao, H., et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Gokaslan, A., Cohen, V., Pavlick, E., and Tellex, S. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., Nastase, S. A., Feder, A., Emanuel, D., Cohen, A., et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. Finding neurons in a haystack: Case studies with sparse probing. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=JYs1R9IMJr.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Hou, B., Zhang, Y., Andreas, J., and Chang, S. A probabilistic framework for llm hallucination detection via belief tree propagation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3076–3099, 2025.

Huben, R., Cunningham, H., Riggs, L., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Jin, C. and Rinard, M. Emergent representations of program semantics in language models trained on programs. In *Forty-first International Conference on Machine Learning*, 2024.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890, 2016. URL http://arxiv.org/abs/1612.06890.

Ju, T., Sun, W., Du, W., Yuan, X., Ren, Z., and Liu, G. How large language models encode context knowledge? a layer-wise probing study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8235–8246, 2024.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *5th International*

Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Kissane, C., Krzyzanowski, R., Bloom, J. I., Conmy, A., and Nanda, N. Interpreting attention layer outputs with sparse autoencoders. *arXiv preprint arXiv:2406.17759*, 2024.

Kuhn, L., Gal, Y., and Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., Griffiths, T. L., Hawkins, R. D., and Nastase, S. A. Shared functional specialization in transformer-based language models and the human brain. *Nature communications*, 15(1):5523, 2024.

Li, Y., Gu, C., Dullien, T., Vinyals, O., and Kohli, P. Graph matching networks for learning the similarity of graph structured objects. In *International conference on machine learning*, pp. 3835–3845. PMLR, 2019.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramár, J., Dragan, A., Shah, R., and Nanda, N. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024.

Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Lindsey, J., Gurnee, W., Ameisen, E., Chen, B., Pearce, A., Turner, N. L., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Thompson, T. B., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Liu, Y., Gao, X., Sun, H., Ge, B., Liu, T., Han, J., and Hu, X. Brain-inspired exploration of functional networks and key neurons in large language models. *arXiv preprint arXiv:2502.20408*, 2025.

Luo, Z., Peng, K., Liang, Z., Cai, S., Xu, C., Li, D., Hu, Y., Zhou, C., and Liu, Q. Mapping effective connectivity by virtually perturbing a surrogate brain. *arXiv preprint arXiv:2301.00148*, 2022.

Ma, X., Fang, G., and Wang, X. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, Apr 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Medaglia, J. D., Lynall, M.-E., and Bassett, D. S. Cognitive network neuroscience. *Journal of cognitive neuroscience*, 27(8):1471–1491, 2015.

Mischler, G., Li, Y. A., Bickel, S., Mehta, A. D., and Mesgarani, N. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*, pp. 1–11, 2024.

Mudide, A., Engels, J., Michaud, E. J., Tegmark, M., and de Witt, C. S. Efficient dictionary learning with switch sparse autoencoders. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=k2ZVAzVeMP.

Muralidharan, S., Turuvekere Sreenivas, S., Joshi, R., Chochowski, M., Patwary, M., Shoeybi, M., Catanzaro, B., Kautz, J., and Molchanov, P. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102, 2024.

Niu, M., Haddadi, H., and Pang, G. Robust hallucination detection in llms via adaptive token selection. *arXiv preprint arXiv:2504.07863*, 2025.

Orgad, H., Toker, M., Gekhman, Z., Reichart, R., Szpektor, I., Kotek, H., and Belinkov, Y. Llms know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Pimentel, T., Valvoda, J., Maudslay, R. H., Zmigrod, R., Williams, A., and Cotterell, R. Information-theoretic probing for linguistic structure. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4609–4622, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.420. URL https://aclanthology.org/2020.acl-main.420/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rajamanoharan, S., Conmy, A., Smith, L., Lieberum, T., Varma, V., Kramár, J., Shah, R., and Nanda, N. Improving dictionary learning with gated sparse autoencoders. *arXiv preprint arXiv:2404.16014*, 2024.

Rathi, N., Mehrer, J., AlKhamissi, B., Binhuraib, T. O. A., Blauch, N., and Schrimpf, M. Topolm: brain-like spatio-functional organization in a topographic language model. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, Apr 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=aWXnKanInf.

Rogers, A., Kovaleva, O., and Rumshisky, A. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

Russinovich, M. and Salem, A. Hey, that's my model! introducing chain & hash, an llm fingerprinting technique. *arXiv preprint arXiv:2407.10887*, 2024.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., and Fedorenko, E. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., et al. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Sriramanan, G., Bharti, S., Sadasivan, V. S., Saha, S., Kattakinda, P., and Feizi, S. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37: 34188–34216, 2024.

Su, W., Wang, C., Ai, Q., Hu, Y., Wu, Z., Zhou, Y., and Liu, Y. Unsupervised real-time hallucination detection based on the internal states of large language models. In Ku, L., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 14379–14391. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.854. URL https://doi.org/10.18653/v1/2024.findings-acl.854.

Sun, H., Zhao, L., Wu, Z., Gao, X., Hu, Y., Zuo, M., Zhang, W., Han, J., Liu, T., and Hu, X. Brain-like functional organization within large language models. *arXiv preprint arXiv:2410.19542*, 2024a.

Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=PxoFut3dWW.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Timkey, W. and van Schijndel, M. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, November 2021. Association

for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Toneva, M. and Wehbe, L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in neural information processing systems*, 32, 2019.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Tuckute, G., Sathe, A., Srikant, S., Taliaferro, M., Wang, M., Schrimpf, M., Kay, K., and Fedorenko, E. Driving and suppressing the human language network using large language models. *Nature Human Behaviour*, 8(3):544–561, 2024.

Vértes, P. E., Alexander-Bloch, A. F., Gogtay, N., Giedd, J. N., Rapoport, J. L., and Bullmore, E. T. Simple models of human brain functional networks. *Proceedings of the National Academy of Sciences*, 109(15):5868–5873, 2012.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33: 12388–12401, 2020.

Voita, E. and Titov, I. Information-theoretic probing with minimum description length. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

Xia, M., Gao, T., Zeng, Z., and Chen, D. Sheared llama: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=09iOdaeOzp.

Xu, J., Wang, F., Ma, M., Koh, P. W., Xiao, C., and Chen, M. Instructional fingerprinting of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3277–3306, 2024a.

Xu, P., Shao, W., Chen, M., Tang, S., Zhang, K., Gao, P., An, F., Qiao, Y., and Luo, P. BESA: pruning large language models with blockwise parameter-efficient sparsity allocation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=gC6JTEU3jl.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

A. Appendix

A.1. Graph Probing Configuration

Hyperparameters. We train graph probes using the Adam optimizer (Kingma & Ba, 2015) with mean squared error (MSE) loss, as defined in Equation (8). The initial learning rate is set to 0.00001, with a batch size of 16. We set the hidden dimension for MLP probes d as 32. We apply a learning rate decay strategy, reducing the rate by a factor of 0.1 if the loss does not improve for 5 consecutive epochs. Each model is trained for up to 100 epochs, with early stopping triggered if no improvement is observed for 20 epochs. Dropout is not used, as preliminary experiments showed no significant impact on regression performance.

Computational Resources. LLM inference for computing neural topologies and perplexity scores requires GPUs with large memory. All experiments were conducted on a Linux server equipped with 8 NVIDIA A100 GPUs (80GB memory each). In contrast, training graph probes is relatively lightweight and can be performed on a single GPU with 16GB memory in less than 1 hour.

A.2. Experimented LLMs

We run graph probing experiments on a diverse range of LLMs across three different families, with the numper of parameters ranging from 124M to 14B. Basic information of these experimented LLMs is summarized in Table 3.

A.3. Datasets

We conduct graph probing experiments using the OpenWebText dataset (Gokaslan et al., 2019). For each dataset, we randomly sample 10,000 text sequences to construct neural connectivity graphs. Each sample is generated by merging and tokenizing raw text until it reaches a length between 256 and 1024 tokens, which defines the length of the corresponding neural activity time series used for computing pairwise correlations. We then construct a text-responsive neural connectivity graph for each sample and compute its associated perplexity score. To remove outliers that distort the distribution, we filter out the top 1% and bottom 1% of samples based on perplexity. Finally, we normalize all perplexity values to the range $[0, 1]$ by subtracting the minimum perplexity and dividing by the observed range. Summary statistics for the constructed datasets are provided in Table 4.

A.4. Graph neural network probe

To reduce the number of probe parameters, we adopt a graph neural network (GNN)-based probe that encodes each node by aggregating neighborhood information through convolutional message passing on the graph (Kipf & Welling, 2017; Fey & Lenssen, 2019). We employ the ReLU activation function (Fukushima, 1975) between graph convolution layers and use average and maximum pooling to summarize node-level embeddings into a graph-level representation. Given a connectivity matrix $\mathbf{A}$ induced by feeding a tokenized sequence X to an LLM, where each element a_{ij} denotes the functional connectivity (Pearson correlation coefficient) between neurons i and j, our probe produces the graph representation $\mathbf{z}$ as follows:

$$\mathbf{\Phi}^0 \in \mathbb{R}^{n \times d}, \tag{12}$$

$$\mathbf{\Phi}^l = \texttt{ReLU}(\mathbf{A}\mathbf{\Phi}^{l-1}\mathbf{\Theta}^l), l = 1, \ldots, L, \tag{13}$$

$$\mathbf{z} = \texttt{AVG}\{\mathbf{\Phi}^L_{1,:}, \ldots, \mathbf{\Phi}^L_{n,:}\} \parallel \texttt{MAX}\{\mathbf{\Phi}^L_{1,:}, \ldots, \mathbf{\Phi}^L_{n,:}\}, \tag{14}$$

where $\mathbf{\Phi}^0 \in \mathbb{R}^{n \times d}$ denotes the initial learnable node embeddings, $\mathbf{\Theta}^l \in \mathbb{R}^{d \times d}$ is the weight matrix of the l-th layer in the GNN with L total layers, and d is a hidden dimensionality hyperparameter. We then feed the graph representation $\mathbf{z} \in \mathbb{R}^{2d}$ into a multi-layer perceptron (MLP) (Rumelhart et al., 1986) to predict the perplexity associated with the input tokenized sequence X:

$$\hat{p} = \mathbf{W}_2 \cdot \texttt{ReLU}(\mathbf{W}_1 \cdot \mathbf{z}^T), \tag{15}$$

where $\hat{p}$ is the predicted perplexity, and $\mathbf{W_1} \in \mathbb{R}^{d \times 2d}$, $\mathbf{W_2} \in \mathbb{R}^{1 \times d}$ are learnable weights of the MLP.

A.5. Neural topology on texts of different subjects.

Beyond the intrinsic default network, we investigated whether LLMs form specialized networks for different knowledge domains, a plausible consequence of their multidisciplinary pretraining. Such task-invoked networks have been observed in human brains (Cole et al., 2014). To test this, we leveraged the MMLU-Pro benchmark (Wang et al., 2024) to see if a query's subject can be predicted from its corresponding neural topology. We framed this as a 10-way classification problem using the 10 subjects with the most samples. For computational efficiency, we applied a simple linear probe to the flattened adjacency matrix, optimized with a cross-entropy loss, similar to our hallucination detection setup. The results of Qwen2.5-0.5B model, shown in Figure 6(a), reveal that neural topology is highly indicative of the subject matter. The topology-based linear probe outperformed a baseline probe on neural activations by 9.39% on average. This advantage held true for nearly every individual subject, with the maximum performance gap exceeding 28.16%. These findings strongly suggest that LLMs develop distinct and linearly separable topological patterns for different knowledge domains, allowing for easy extraction of the context or subject being processed.

To provide intuitive evidence for subject-invoked topologies, we calculated the average neural topology for each of the 10 subjects and visualized their pairwise correlations in

Table 3. Basic information of the experimented LLMs.

LLM family	#params	#layers	#neurons per layer	experimented layer id
GPT-2	124M	12	768	1-12
	774M	36	1280	18
Pythia	160M	12	768	1-12
	1.4B	24	2048	12
	2.8B	32	2560	16
Qwen2.5	0.5B	24	896	1-12
	3B	36	2048	18
	7B	28	3584	14
	14B	48	5120	24

Table 4. Basic information of constructed graph probing datasets.

LLM family	#tokens	#graphs	#training graphs	#test graphs
GPT-2	7,020,215	10,384	8,308	2,076
Pythia	6,798,668	10,441	8,353	2,088
Qwen2.5	7,935,555	11,452	9,162	2,290

Figure 6(b). The results are striking: the correlation matrix largely mirrors the conceptual similarities between these academic fields. While all pairwise correlations exceed 0.9 (reaffirming the existence of a strong default network), the variations reveal an intriguing structure. For example, the topologies for Math, Physics, and Engineering are highly inter-correlated (>0.98), and all share a much lower correlation (<0.95) with Law. More broadly, we observe a clear clustering that separates STEM and social science subjects, consistent with commonsense knowledge. To seek causal evidence for this phenomenon, we performed a sweeping intervention on individual neurons. We systematically pinned a neuron's activation to a fixed value in {-2, -1, 0, 1, 2} and measured the resulting change in the model's output on queries from different subjects. This allowed us to identify neuron #894 in the 12th layer of Qwen2.5-0.5B, a "STEM neuron" whose activation state has a disproportionate impact on STEM-related subjects. As illustrated in Figure A.5(c), intervening on this single neuron induced 3.57 times more change in model outputs for STEM queries than for social science queries. While preliminary, these correlational and causal findings strongly suggest that functional specialization is an emergent property of LLMs, hinting at shared organizational principles between artificial and biological intelligence.

A.6. Model matching and fingerprinting

In this section, we further investigate potential topological similarity across different LLMs to answer the following question: can we detect the genetic relationships of LLMs (*e.g.* same model family, or finetuning) from their internal neural topology? To this end, we extend graph probing with contrastive learning to perform *graph matching*, as illustrated in Figure 7. Specifically, suppose we feed a batch of B token sequences into two LLMs, Ω and Γ. We compute the corresponding neural connectivity graphs and use two GNN probes to encode them into representations $\mathbf{Z}_\Omega = [\mathbf{z}_1^\Omega, \ldots, \mathbf{z}_B^\Omega]$ and $\mathbf{Z}_\Gamma = [\mathbf{z}_1^\Gamma, \ldots, \mathbf{z}_B^\Gamma]$. Matching is implemented using a contrastive cross-entropy (CE) loss that encourages alignment between graph representations from the same input texts:

$$\mathcal{S} = \texttt{MAT\_MUL}(\mathbf{Z}_\Omega^T, \mathbf{Z}_\Gamma), \quad \mathcal{T} = \texttt{IDENTITY}(B), \tag{16}$$

$$\mathcal{L} = \sum_{i=1}^{B} \texttt{CE}(\mathcal{S}_{i,:}, \mathcal{T}_{i,:}) + \sum_{j=1}^{B} \texttt{CE}(\mathcal{S}_{:,j}, \mathcal{T}_{:,j}), \tag{17}$$

where $\mathcal{S}$ is the similarity matrix by taking inner product of graph representations and IDENTITY indicates identity matrix which is the target $\mathcal{T}$ for graph matching. After training the graph probes contrastively on a shared set of training texts, the out-of-sample graph matching performance serves as an indicator of neural topology similarity between two LLMs. To evaluate this, we adopt the commonly used GAUC metric (Li et al., 2019). Specifically, given the predicted similarity matrix $\mathcal{S} \in \mathbb{R}^{N \times N}$ and the target similarity matrix $\mathcal{T} = \texttt{IDENTITY}(N)$, GAUC for graph matching is calculated as follow:

$$\texttt{GAUC} = \frac{1}{2N} \sum_{i=1}^{N} (\texttt{AUC}(\mathcal{S}_{i,:}, \mathcal{T}_{i,:}) + \texttt{AUC}(\mathcal{S}_{:,i}, \mathcal{T}_{:,i})). \tag{18}$$

Table 5 presents the graph matching results. As a sanity check, we first perform *self-matching* using the same LLM, and the results indeed show that GAUC is close to 1.0, vali-

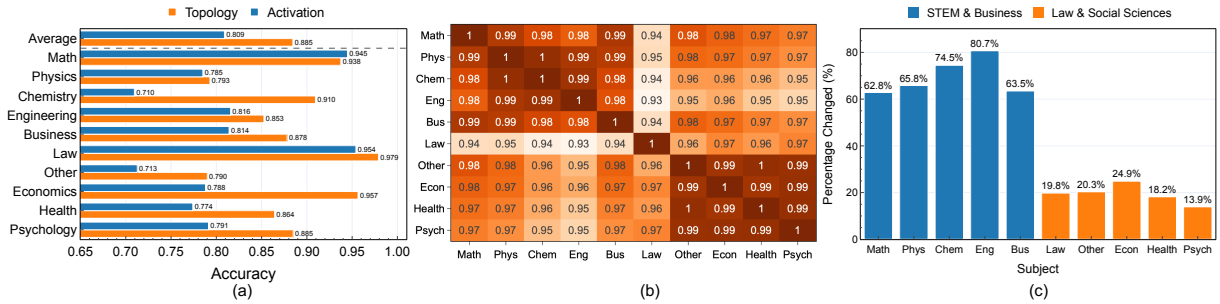


Figure 6. (a) Subject classification accuracy probed from neural topology and activation of Qwen2.5-0.5B model on MMLU benchmark. (b) Correlation of neural topology of Qwen2.5-0.5B on different subjects. (c) Percentage of changed questions in each subject by intervening neuron #894 in layer 12 of Qwen2.5-0.5B.

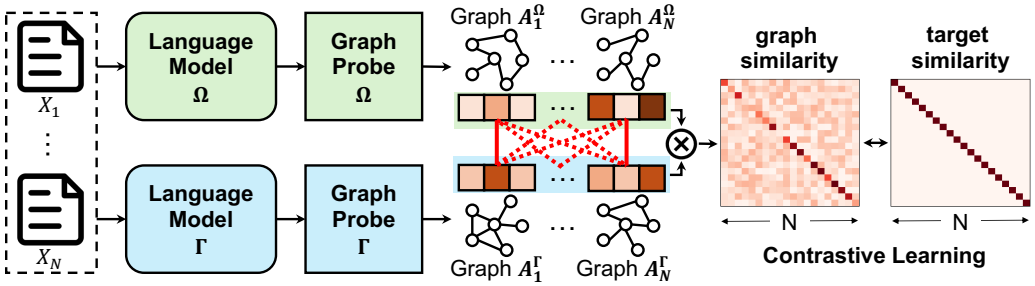


Figure 7. An overview of graph matching. We learn representations of neural topologies derived from two different LLMs processing the same text dataset. We then perform contrastive learning on the graph representations such that matching pairs are more similar by inner product.

Table 5. Graph matching performance (GAUC $\times 100$) between different LLMs.

Matching	LLM Ω	LLM Γ	GAUC
Self	GPT-2	GPT-2	98.64
	Pythia	Pythia	96.92
	Qwen2.5	Qwen2.5	99.24
Generation	Qwen2.5	Qwen2	93.27
	Qwen2.5	Qwen1.5	96.10
	Qwen2	Qwen1.5	94.21
Family	GPT-2	Pythia	92.00
	GPT-2	Qwen2.5	91.11
	Pythia	Qwen2.5	87.39

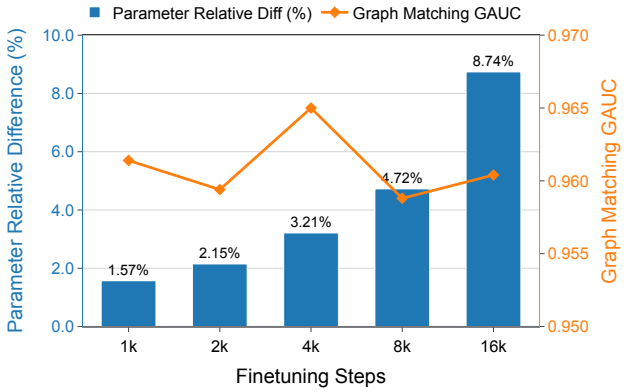


Figure 8. Relative differences of parameters (blue) and graph matching accuracy (orange) of Pythia-160M at different checkpoints.

dating the rationality of our methodology. We then incorporate two configurations: (1) LLMs within the same family but from different generations, (2) LLMs across different families. Given the reduced architectural and training data differences within the same model family, cross-generation LLMs are *genetically* closer than cross-family LLMs. As expected, cross-generation matching significantly outperforms cross-family matching with the average and maximum GAUC gap up to 4.84% and 9.97%, confirming the effectiveness of graph matching in detecting genetic closeness of LLMs.

The graph matching extension enables a direct application in LLM fingerprinting (Xu et al., 2024a; Russinovich & Salem, 2024), which is crucial for protecting intellectual property. To test this, we perform graph matching on the Pythia-160M model, using checkpoints at 1k, 2k, 4k, 8k, and 16k steps of continued training from a base checkpoint (127k). Figure 8 illustrates the graph matching accuracy measured by GAUC, as well as the relative differences of weight values. Despite significant parameter drift, where the weight difference after 16k steps is 5.57 times greater than after 1k steps, the topological signature remains nearly unchanged, with our method achieving a graph matching GAUC above 0.96 in all cases. This suggests that neural topology can serve as a robust fingerprint, resilient to weight modifications from finetuning, which we propose as a significant avenue for future work.

Table 6. Graph matching performance (GAUC $\times 100$) for different VLMs

VLM Ω Modality	VLM Γ Modality	GAUC
LLaVA Image+Text	LLaVA Image+Text	95.98
LLaVA Text	LLaVA Image	81.88
LLaMA	LLaVA Text	68.03

A.7. Graph probing on vision language models

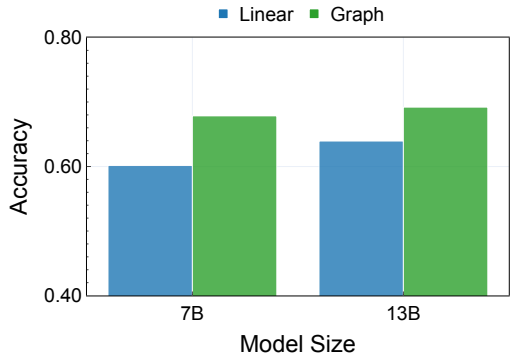
Unlike LLMs, which contain only textual hidden states, Vision-Language Models (VLMs) jointly encode image and text features within a shared transformer backbone, producing multimodal hidden states that remain underexplored. To study their internal topology, we extract the hidden representations at each layer and compute correlations as in Equations (1-4), yielding a connectivity matrix whose co-activation time series span both modalities.

We evaluated neural topology in VLMs by probing different sizes of LLaVA-v1.5 (Liu et al., 2024). For this setup, we adapted the CLEVR dataset (Johnson et al., 2016) into an object-counting classification task of 10,000 randomly sampled images with labels corresponding to the number of objects (integers 3-10). GNN-based probes (Equations (12-14)) were trained with cross-entropy loss on neural topology graphs constructed at a fixed density of 0.01. As shown in Figure 9(a), graph probing consistently outperformed linear activation probing across both the 7B and 13B models, validating that neural topology is more informative than neural activation in VLMs regarding visual understanding capabilities.

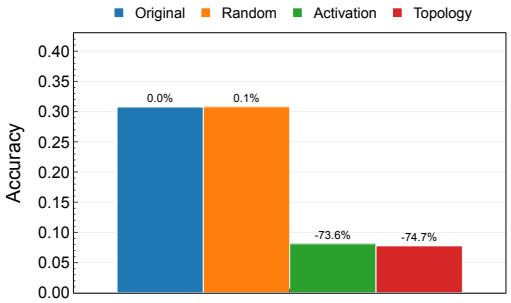
To further evaluate the role of neural topology, we conducted intervention experiments on VLMs on the same dataset. Instead of restricting the probe to classification, we asked the model to generate numeric outputs and then ablated the top 1% of nodes, selected by degree, activation, or at random, by setting their values to zero. Figure 9(b) demonstrates that ablating top nodes by either activation or topology reduced accuracy far below random ablation and the original baseline, with topology producing the largest drop.

A.8. Graph matching on vision language models

We employ graph matching on VLMs to evaluate topological similarity across modalities and to test whether multimodal training reorganizes neural topology relative to unimodal language models. We use the MS-COCO dataset (Lin et al., 2014), where paired images and captions describe the same content, providing a natural basis for structural alignment. Text graphs are constructed by masking visual tokens, and image graphs by isolating patch embeddings, and multimodal graphs by combining both. We fix graph density at 0.01 and use the same contrastive loss as in the LLM matching experiments.



(a)



(b)

Figure 9. (a) Out-of-sample probing performance on LLaVA-v1.5 of different sizes. (b) Accuracy on CLEVR benchmark of LLaVA-v1.5-7b under different interventions of top 1% neurons.

Table 6 reports graph matching scores. As a baseline, multimodal LLaVA graphs compared against themselves yield near-perfect alignment (95.98). Matching LLaVA text-only against image-only produces a notable drop (81.88), suggesting that while text and image graphs share broad structural similarities, each retains specialized organization patterns that multimodal training does not fully unify. In contrast, matching LLaMA text against LLaVA text yields a lower score (68.03), indicating that multimodal training reshapes internal topology beyond unimodal pretraining.