The Architecture of Cognitive Duality in LLMs

Anonymous ACL submission

Abstract

Large language models are typically analyzed in terms of what they represent or what they output, but less is known about whether different task families induce different internal organizations of computation. We study this question through task-conditioned activation topology. For each model, layer, and task family, we construct hidden-dimension co-activation graphs from mean-pooled hidden states, where edges reflect cross-prompt correlations within a condition, and summarize these graphs with mean degree and mean betweenness centrality. Across six transformer models from two families and a balanced synthetic benchmark of 600 prompts spanning 10 logical and 10 creative categories, we find a consistent dissociation: logical tasks induce denser graphs with higher mean degree centrality, whereas creative tasks induce sparser graphs with higher mean betweenness centrality. The same directional pattern appears on two real-world validation sets, SVAMP and WRITINGPROMPTS, despite their substantially different surface characteristics. Topology-derived features are also diagnostically useful: a simple linear probe using only internal features distinguishes task family with high accuracy and compares favorably to an entropy-based output baseline. Finally, a preliminary intervention that suppresses high-betweenness dimensions yields a modest improvement on arithmetic, suggesting that the identified dimensions may be functionally relevant rather than purely descriptive. We do not claim a one-to-one correspondence between transformer computation and human convergent/divergent thinking. Rather, our results suggest that global activation topology provides a useful lens on how task demands shape internal computation in language models.

1 Introduction

When a mathematician works through a proof, and when a poet reaches for a metaphor, the difference is not only in content. Cognitive neuroscience suggests that the large-scale organization of computation changes as well. Decades of work argue that human reasoning is not monolithic. *Convergent thinking*, the focused resolution of well-defined logical problems, recruits tighter, more localized circuitry associated with prefrontal control and structured reasoning (Goel et al., 1997; Prado et al., 2011). *Divergent thinking*, the generative, associative mode underlying creative production, relies more heavily on distributed interaction involving the default mode network, whose hub regions help bridge semantically distant concepts (Andrews-Hanna and Grilli, 2021). The point is not simply that the brain thinks harder for logic and more freely for creativity. Rather, evidence suggests that it routes information through different topological configurations as cognitive demands change (Sporns, 2013; Kahneman, 2011).

Creativity, in particular, appears to depend on coordination between generative and control processes. Beaty et al. (2015) found that creative idea production engages both the default mode network and the executive control network simultaneously, and that higher creative ability was associated not merely with stronger activation, but with more efficient large-scale network organization during divergent thinking (see Appendix A for an illustration). For our purposes, this result is useful less as a literal template for transformer models than as a methodological cue: if topology matters for human cognition, it is worth asking whether graph-theoretic structure also reveals something about how language models organize computation.

This raises a question that, to our knowledge, has not been directly examined: do large language models (LLMs), trained on human-generated text spanning proofs and poems, arguments and analogies, exhibit analogous task-conditioned differences in internal computational organization? Or do they process every task through essentially the same undifferentiated forward pass, regardless of whether

the task calls for a syllogism or a metaphor? Most work in mechanistic interpretability has focused on *what* is represented and *where* specific behaviors localize, rather than on whether the internal organization of computation shifts systematically with task type (Elhage et al., 2021; Meng et al., 2022). That possibility remains comparatively underexplored.

We study this question through **task-conditioned activation topology** in LLMs. For each model and layer, we collect hidden states while the model responds to logical and creative prompts, and for each prompt we build a hidden-dimension co-activation graph: nodes correspond to hidden dimensions, and edges reflect the strength of co-variation among dimensions across generated token positions within that prompt. We then summarize these graphs with two complementary metrics: mean degree centrality, which captures overall connectedness, and mean betweenness centrality, which measures the extent to which a small subset of dimensions acts as hubs linking otherwise separate regions of representation space. We apply this analysis to six transformer models, Pythia 160M, 410M, and 1B (Biderman et al., 2023) alongside Llama-3.2-1B, Llama-3.2-3B, and Llama-3.2-3B-Instruct (Touvron et al., 2023), using 600 prompts spanning 10 logical and 10 creative task categories, and we further evaluate the same framework on two real-world datasets.

Across models and datasets, we observe a consistent dissociation. Logical tasks tend to induce higher degree centrality and more locally concentrated co-activation structure, whereas creative tasks show relatively higher betweenness centrality, suggesting greater reliance on hub-like dimensions that bridge otherwise weakly coupled parts of the representation space. Throughout, we use *logical*/*creative* and *convergent*/*divergent* as convenient shorthand for the task families we study. We do not claim a one-to-one equivalence between human cognitive categories and transformer computation. The comparison is useful because it yields concrete, testable hypotheses about coordination and routing inside models.

These signatures are not only descriptive. A simple linear probe trained on topological features reliably distinguishes task type and compares favorably to an entropy-based baseline, especially in transfer settings. In a preliminary intervention, attenuating high-betweenness dimensions modestly improves arithmetic accuracy, offering initial evidence that the identified hubs may play a functional role rather than serving as purely descriptive markers. Our claim, however, is intentionally narrow: task type appears to be reflected not only in what a model represents, but also in how its internal computation is organized. Our contributions can be summarized as:

1. **Empirical pattern.** Across six transformer models, logical and creative tasks induce reliably different co-activation topologies: logical tasks tend to yield higher mean degree centrality, whereas creative tasks tend to exhibit higher mean betweenness centrality.
2. **Diagnostic value.** A lightweight linear probe trained on activation-graph features cleanly separates task types and compares favorably to a simple entropy-based baseline.
3. **Preliminary intervention experiments.** Attenuating high-betweenness dimensions in a small-model arithmetic setting modestly improves logical accuracy, providing initial evidence that activation topology may play a functional, and not merely correlational, role.

2 Related Work

Mechanistic Interpretability of Transformers. The mechanistic interpretability program seeks to recover the computations implemented in transformer weights and activations (Elhage et al., 2021; Olah et al., 2020). Key results include the identification of induction heads and indirect-object circuits via attention-head decomposition (Olsson et al., 2022; Wang et al., 2022), localization of factual associations through activation patching and causal mediation (Meng et al., 2022; Goldowsky-Dill et al., 2023), and probing studies showing that syntactic and pipeline structure is linearly encoded in contextual representations (Tenney et al., 2019; Hewitt and Manning, 2019). This literature asks which components encode particular information or support particular behaviors. We ask a complementary question: whether the *global organization* of activation patterns changes systematically with task family.

Dual-Process Accounts of LLM Reasoning. Dual-process theory distinguishes fast, intuitive processing from slower, more deliberate reasoning (Kahneman, 2011). In NLP, the framework has been applied behaviorally, asking whether LLMs display System 1-like heuristic errors or can be nudged toward deliberation through prompting (Ha-

gendorff et al., 2023; Wei et al., 2022), and architecturally, by separating fast generation from slower verification or planning modules (Yang et al., 2024; He et al., 2024). We do not treat logical versus creative generation as a literal implementation of System 1 versus System 2. What carries over is the broader hypothesis that different task demands may recruit different internal computational modes, which we examine through activation topology.

Graph-Theoretic Analysis of Neural Representations. Graph-theoretic measures of degree, betweenness, and efficiency are standard tools for characterizing large-scale neural organization in cognitive neuroscience (Sporns, 2013; Bullmore and Sporns, 2009), and Beaty et al. (2015) showed that network efficiency predicts divergent-thinking performance. Within machine learning, graph methods have more often been applied to external task structure or knowledge graphs than to the internal topology of hidden representations (Maurya et al., 2019). Our work adapts this toolkit to transformer activations, treating hidden-dimension co-activation graphs and their centrality profiles as summaries of how computation is organized across task families.

Reasoning and Creative Generation Benchmarks. Standard reasoning benchmarks such as SVAMP (Patel et al., 2021) and GSM8K (Cobbe et al., 2021) and creative benchmarks such as WRITING-PROMPTS (Fan et al., 2018) evaluate model *outputs*. We use SVAMP and WRITING-PROMPTS differently: as anchors for studying whether behavioral task families correspond to reproducible differences in the topology of internal activations, rather than as accuracy or fluency benchmarks.

3 Background and Motivation

Convergent thinking has been associated with more focused, higher intensity engagement of lateral prefrontal and parietal regions, with comparatively localized activation patterns reported across tasks (Prado et al., 2011; Goel et al., 1997). Divergent thinking, by contrast, is often linked to more distributed, integrative processing. Activity associated with the default mode network can involve hub regions that bridge semantically distant concepts, supporting broader associative flow (Beaty et al., 2015; Andrews-Hanna and Grilli, 2021).

Along these lines, Beaty et al. (2015) reported that creative ability was better predicted by properties of network organization than by activation magnitude alone. Participants with stronger divergent thinking performance showed higher global efficiency in functional connectivity graphs, suggesting more effective information routing across the network. We take this result as methodological inspiration rather than as a literal template for transformers.

The broader point is that the topology of internal activity may carry information that activation strength alone does not. Overall, the structural organization of cognition may matter alongside its intensity, even if the mapping from human brain networks to transformer representations remains indirect.

Why Activation Topology? Motivated by this perspective, we summarize transformer activations with two complementary graph measures. For each task family, we construct a hidden-dimension co-activation graph and ask whether its connectivity profile differs systematically across logical and creative generation. Table 1 summarizes the interpretation of the two metrics and our qualitative expectations for them.

Metric	What it measures	Logical	Creative
Mean degree centrality	Overall connectedness of hidden dimensions in the co-activation graph	Higher	Lower
Mean betweenness centrality	Extent to which a small set of dimensions bridge otherwise weakly connected regions of the co-activation graph	Lower	Higher

Table 1: Qualitative interpretation of the two graph measures used in this work. The expected tendencies are hypotheses motivated by the convergent/divergent distinction; they are tested empirically rather than assumed.

These measures are complementary rather than independent. Higher overall connectedness and stronger bridge structure can co-vary, particularly as graph density changes, so we interpret them jointly. Our working intuition is that logical tasks may rely on more concentrated coordination among a relatively tight set of representational dimensions, which would increase overall connectedness in the co-activation graph. Creative tasks, by contrast, may place greater pressure on dimensions that bridge otherwise distant regions of representation space, increasing the prominence of hub-like dimensions and, with it, betweenness centrality. Whether this dissociation actually appears is an

empirical question.

Why Entropy Is an Incomplete Baseline Entropy over the output distribution captures how uncertain or spread a model's predictions are and hence been applied as a metric sometimes to measure Creative cognition, but uncertainty is not the same as cognitive mode. A model can be highly uncertain on a logical task it finds difficult, and highly confident on a creative task whose style it has memorized. Entropy is thus a proxy for prediction confidence , not for the structural organization of internal computation. Our linear probe, trained on topological features rather than output statistics, tests whether the geometry of hidden state co-activation is a more direct and faithful indicator of the cognitive mode a model is operating in.

4 Dataset Construction

Synthetic Benchmark. We construct a balanced synthetic benchmark of 600 prompts, comprising 300 logical and 300 creative examples. Each class contains 10 task categories with 30 prompts per category. The logical and creative categories are listed in Appendix B.

Candidate prompts were generated with GPT-4 under category-specific instructions requiring self-contained wording, minimal reliance on outside knowledge, and the avoidance of obvious lexical markers that would trivially reveal the label. We then used an external task-type classifier only as a quality-control filter. Items receiving confidence below 0.85 were discarded and regenerated. The classifier is not used in any downstream analysis, and we do not optimize prompts beyond screening out ambiguous items. The two classes are tightly length matched (logical: $\mu = 87.3$, $\sigma = 13.1$ characters; creative: $\mu = 84.4$, $\sigma = 10.5$ characters), reducing prompt length as a trivial explanation for any observed topological differences. Representative prompt pairs appear in Table 6 in Appendix B.

Real-World Validation Datasets. To test whether the synthetic results survive outside a controlled prompt set, we evaluate on two established corpora, using 200 prompts from each. These datasets are intentionally less controlled than the synthetic benchmark. They differ in prompt length, genre, and expected output structure. We therefore treat them as out-of-distribution robustness checks rather than as perfectly matched realizations of our two task families.

SVAMP (Patel et al., 2021) provides mathematical word problems ($\mu = 301.8$ characters) requiring multi-step arithmetic reasoning and determinate answers, making it a natural proxy for the logical condition. The WRITING-PROMPTS corpus (Fan et al., 2018) provides open-ended narrative prompts ($\mu = 132.2$ characters) and serves as a proxy for the creative condition. Because these corpora vary along multiple surface dimensions, successful replication on them is informative but not decisive on its own; the primary controlled evidence comes from the synthetic benchmark.

5 Methodology

Activation Extraction. For each prompt, we generate a continuation (up to 50 new tokens, temperature $= 1.0$, top-$p = 0.9$) and then run a second forward pass on the full sequence (prompt + generated tokens). This yields hidden state tensors at every layer $l \in \{0, \ldots, L\}$, where layer 0 is the embedding output. For each prompt, we compute two per layer summaries, both stored as float16 arrays of shape $(N, L{+}1, D)$:

Mean-pooled hidden state ($\mathbf{h}^{(l)}_{\text{mean}}$): the average hidden state vector over generated tokens only, masking out prompt tokens and padding:

$$\mathbf{h}^{(l)}_{\text{mean}} = \frac{\sum_{t \in \mathcal{G}} \mathbf{h}^{(l)}_t}{|\mathcal{G}|} \tag{1}$$

where $\mathcal{G}$ denotes the set of generated (non prompt, non padding) token positions. This isolates the model's representational activity during generation from the conditioning context.

Final hidden state ($\mathbf{h}^{(l)}_{\text{final}}$): the hidden state at the last non padding token of the full sequence.

The main graph analysis uses the token-level hidden states $\mathbf{h}^{(l)}_t$ restricted to generated-token positions, while $\mathbf{h}^{(l)}_{\text{mean}}$ is retained for prompt-level summaries and the probe features below.

Prompt level Activation Extraction. For each prompt p, we sample a continuation of at most 50 new tokens using a fixed decoding policy (temperature $= 1.0$, top-$p = 0.9$) shared across all models and task conditions. We then run a second forward pass on the concatenated prompt-response sequence and collect hidden states at every layer $l \in \{0, \ldots, L\}$, where $l = 0$ denotes the embedding output. This second pass yields a single hidden-state tensor for the realized sequence and

avoids mixing incremental-generation cache states with post hoc analysis.

Let $\mathbf{h}_{p,t}^{(l)} \in \mathbb{R}^D$ denote the hidden state for prompt p at token position t and layer l. Our main prompt-level summary representation is the mean-pooled response state,

$$\mathbf{h}_p^{(l)} = \frac{1}{|\mathcal{G}_p|} \sum_{t \in \mathcal{G}_p} \mathbf{h}_{p,t}^{(l)}, \tag{2}$$

where $\mathcal{G}_p$ is the set of generated (non-prompt, non-padding) token positions. Mean pooling focuses the analysis on the model's response phase while keeping the representation comparable across continuations of different lengths. The main graph analysis, however, constructs prompt-level graphs directly from the token-level states $\mathbf{h}_{p,t}^{(l)}$ over $t \in \mathcal{G}_p$, while $\mathbf{h}_p^{(l)}$ is used for prompt-level summaries and the probe features below.

Prompt-Level Co-Activation Graphs. Our main descriptive analysis builds one graph per prompt and layer. For a fixed prompt p and layer l, let $\mathbf{X}^{(l,p)} \in \mathbb{R}^{|\mathcal{G}_p| \times D}$ be the matrix whose rows are the token-level hidden states $\mathbf{h}_{p,t}^{(l)}$ over generated token positions $t \in \mathcal{G}_p$. We define the hidden-dimension co-activation matrix $\mathbf{C}^{(l,p)} \in \mathbb{R}^{D \times D}$ by

$$C_{ij}^{(l,p)} = \operatorname{corr}\left(\mathbf{X}_{:,i}^{(l,p)}, \mathbf{X}_{:,j}^{(l,p)}\right), \tag{3}$$

that is, the Pearson correlation between dimensions i and j across generated token positions within prompt p.

We convert this dense matrix into a sparse weighted graph in two steps. First, we take absolute correlation weights, $w_{ij} = |C_{ij}^{(l,p)}|$, retain the top 5% of off-diagonal weights, and set the remainder to zero. We use absolute correlations so that both positive and negative co-variation count as structured dependence, and because shortest-path-based centrality requires nonnegative edge weights. Second, we select the $K = 100$ dimensions with largest weighted degree in the sparse graph and induce the corresponding K-node subgraph $G^{(l,p)} = (V, E, W)$. Using a fixed retention rate keeps sparsity comparable across prompts, and a fixed node budget keeps the resulting graphs comparable across models with different hidden dimensionalities. The appendix reports threshold ablations. From each $G^{(l,p)}$ we extract two graph-level summaries: mean degree centrality and mean betweenness centrality. For weighted shortest-path calculations, we convert correlation weights to edge lengths via $\ell_{ij} = 1/(w_{ij} + 10^{-8})$, so that stronger co-activation corresponds to shorter effective distance.

We then average these prompt-level statistics within each condition for reporting. Unless otherwise noted, the main text reports the middle layer $l = \lfloor L/2 \rfloor$ for each model (the appendix provides full layer sweeps).

A list of models used in our experiments can be found in Table 7 in the appendix.

Linear Probe. We use a separate prompt-level probe to test whether topology-derived features are diagnostic of task family. In each split of 10-fold stratified cross-validation, the probe is built from prompt-level topology-derived features together with fold-specific reference summaries computed using only the training prompts. Held-out prompts never contribute to the reference summaries against which they are evaluated.

For each held-out prompt, we extract its mean-pooled response state at three pre-specified depths corresponding to early, middle, and late processing (approximately 25%, 50%, and 75% of model depth. For 16-layer models these are layers $\{3, 8, 13\}$). At each selected layer and for each reference graph, we restrict the prompt's activation vector to the graph's node set and compute graph-aware summaries. Specifically, for activation vector $\mathbf{a}$ on node set V_G, we compute the mean and standard deviation of activation magnitude,

$$\mu(\mathbf{a}) = \frac{1}{|V_G|} \sum_{v \in V_G} |a_v|, \; \sigma(\mathbf{a}) = \operatorname{Std}_{v \in V_G}(|a_v|), \tag{4}$$

together with activation-weighted averages of node degree, betweenness, and closeness centrality:

$$S_c(\mathbf{a}, G) = \sum_{v \in V_G} \tilde{a}_v \, c_G(v), \; \tilde{a}_v = \frac{|a_v|}{\sum_{u \in V_G} |a_u|}, \tag{5}$$

where $c_G(\cdot)$ is degree, betweenness, or closeness centrality under the reference graph G. Concatenating these summaries across layers and across the two reference graphs yields a fixed-length feature vector for each prompt.

We train a logistic-regression classifier (L2 penalty, $C = 1.0$) on these features, with StandardScaler normalization fit on each training fold. As a simple output-side baseline, we compare against a single scalar entropy summary of the model's output distribution over the generated span, using the

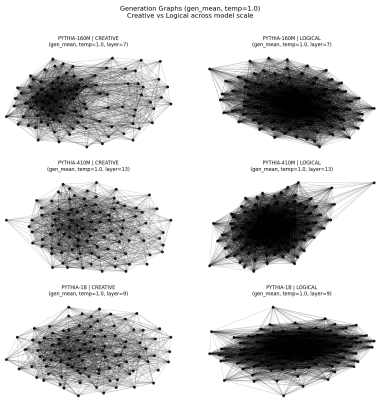

Figure 1: Representative hidden-dimension co-activation graphs for Pythia 160M, 410M, and 1B at the reported analysis layer (logical: left; creative: right). Logical conditions yield denser graphs with more broadly distributed connectivity, whereas creative conditions yield sparser graphs in which a smaller set of nodes occupies more central bridging positions.

same cross-validation splits and the same classifier. We evaluate both models with AUC-ROC and report 95% bootstrap confidence intervals based on 2,000 resamples.

6 Results

Activation Topology Differs by Task Family Figure 1 shows representative condition-level co-activation graphs for the three Pythia models. A qualitative contrast is visible across scales: graphs induced by logical prompts are denser, with connectivity distributed more broadly across the node set, whereas graphs induced by creative prompts are sparser and place more of the shortest-path structure on a smaller subset of bridge-like nodes.The main claims rest on the quantitative summaries below and the full layerwise analyses in the appendix.

Table 2 quantifies the synthetic-benchmark results across all six models. For each model and condition, we report mean degree centrality (MDC) and mean betweenness centrality (MBC) at the reported analysis layer. The directional pattern is uniform across model families and scales: logical conditions yield higher MDC, whereas creative conditions yield higher MBC. The magnitude of the gap varies by model, but no model reverses the pattern.

Because betweenness can covary with graph sparsity, we interpret MBC jointly with MDC rather than in isolation. Under that joint reading, logical tasks are associated with denser co-activation structure, whereas creative tasks rely relatively more on bridge-like dimensions.

Model	Layer	Log. MDC	Log. MBC	Cre. MDC	Cre. MBC	MBC Ratio
Pythia-160M	7	0.795	2.09	0.450	5.63	2.69
Pythia-410M	13	0.775	2.30	0.333	6.84	2.97
Pythia-1B	9	0.711	2.95	0.317	7.00	2.37
LLaMA-3.2-1B	16	0.870	1.33	0.535	4.75	3.57
LLaMA-3.2-3B	27	0.908	0.94	0.541	4.68	4.98
LLaMA-3.2-3B-Inst.	27	0.746	1.21	0.562	4.36	3.60

Table 2: Synthetic-benchmark results across all six models. Logical prompts consistently yield higher mean degree centrality (MDC), whereas creative prompts consistently yield higher mean betweenness centrality (MBC). The final column shows the creative/logical MBC ratio.

We next ask whether the synthetic pattern survives on less controlled data. Table 3 shows that it does for the Pythia family: SVAMP yields higher MDC and lower MBC, whereas WRITINGPROMPTS yields lower MDC and higher MBC. Because these corpora differ substantially in prompt length, genre, and expected output structure, we interpret this result as a robustness check rather than as a matched replication of the synthetic setting. Real-dataset results for the LLaMA models are reported in the appendix.

Model	Dataset	Layer	MDC	MBC ($\times 10^{-3}$)
Pythia-160M	SVAMP	7	0.933	0.69
	WRITINGPROMPTS	7	0.456	5.59
Pythia-410M	SVAMP	13	0.904	1.16
	WRITINGPROMPTS	13	0.519	3.89
Pythia-1B	SVAMP	9	0.905	1.01
	WRITINGPROMPTS	9	0.419	2.88

Table 3: Pythia results on the real-world validation datasets. Despite substantial differences in prompt length and genre between SVAMP and WRITINGPROMPTS, the same directional split appears: SVAMP is denser, whereas WRITINGPROMPTS exhibits higher betweenness.

Taken together, the descriptive evidence is consistent: every synthetic-benchmark model and every Pythia real-data comparison exhibits the same directional split. The effect size varies, but the sign does not.

Linear Probe vs. Entropy Baseline These descriptive results motivate a downstream question: do topology-derived features provide usable signal beyond a lightweight output-side uncertainty measure? Table 4 compares the topology probe with an entropy baseline across several model settings. On the main benchmark, the two are nearly tied (0.967 vs. 0.965 AUC), so we do not claim a decisive advantage there. The more informative

pattern is that the topology probe is numerically at least as strong in every additional model we evaluate, with especially large gains on GPT-2 and DistilGPT-2. This suggests that activation topology captures task-relevant structure that is not reducible to output uncertainty alone.

Model	Topology	Entropy	Δ (Topo–Ent)
Main benchmark	0.967	0.965	+0.002
`distilGPT-2`	0.804	0.611	+0.193
`Pythia-70M`	0.919	0.901	+0.018
`GPT-2`	0.849	0.634	+0.215

Table 4: AUC for topology-derived features versus an entropy baseline under 10-fold stratified cross-validation. The first row reports the main synthetic-benchmark setting used in the paper; the remaining rows provide additional checks on smaller GPT-style models.

Appendix E shows the corresponding ROC curves. The combined detector reaches AUC $= 0.975$, a modest improvement over either feature family alone, suggesting that activation topology and output entropy encode overlapping but non-identical information about task family.

7 Analysis and Discussion

Why Do Creative Tasks Elevate Betweenness Centrality? The most cautious reading of the creative-condition result is a joint one: relative to logical tasks, creative tasks yield graphs with lower overall connectedness but higher mean betweenness centrality. In other words, fewer dimensions are strongly connected overall, and a larger share of the remaining shortest-path structure is concentrated on a smaller set of bridge-like nodes. One plausible interpretation is that many creative prompts require the model to combine or traverse semantically distant concepts. For example in metaphor generation, unusual premises, or open-ended associative expansion, which would naturally place more weight on dimensions that connect otherwise weakly coupled regions of representation space.

That interpretation is qualitatively consistent with the cognitive-neuroscience picture in Beaty et al. (2015), where creative thought is associated with coordination across distributed networks. At the same time, our evidence is correlational, and the analogy should not be taken too literally. More importantly, betweenness is not independent of density: when a graph becomes sparser, shortest paths are more easily forced through a smaller number of nodes. For that reason, we do not interpret elevated betweenness in creative tasks as a standalone signature of "bridging" computation. The stronger claim supported by the present data is the *joint dissociation*: creative tasks are associated with lower degree and higher betweenness. Density-matched controls and null-model comparisons would be needed to determine how much of the betweenness effect remains once sparsity is explicitly controlled.

Why Do Logical Tasks Elevate Mean Degree Centrality? Higher mean degree centrality under logical conditions indicates that, among the retained hidden dimensions, co-variation is broader and more uniformly distributed. This is consistent with the idea that rule-governed tasks recruit a more tightly coordinated set of representational dimensions, yielding denser co-activation structure at the condition level. On that reading, logical generation depends less on a few bridge-like nodes and more on distributed coordination within a relatively compact subspace.

That interpretation is plausible, but it is not the only one. Logical prompts also tend to impose a more constrained answer space and may produce less semantically diverse continuations than creative prompts, which could itself increase cross-prompt covariance. Likewise, if arithmetic or deductive problems induce more stable intermediate states across the generated span, mean pooling could accentuate that regularity. Because our graphs are built from correlations across prompts within a condition, they summarize task-family structure rather than token-by-token information flow within a single forward pass. Distinguishing between coordinated reasoning, constrained output structure, and temporal persistence will require more targeted analyses, ideally with matched output formats and token-level or prompt-only controls.

Betweenness Suppression: A Causal Probe The most direct way to move beyond description is to intervene on the dimensions identified by the graph analysis. To that end, we run a preliminary suppression experiment on `Qwen2-0.5B` in a purely arithmetic setting. We attenuate the top 5% of dimensions ranked by betweenness centrality via a forward hook on the MLP down-projection, vary the suppression strength $\alpha \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, and measure arithmetic accuracy on 200 problems.

The pattern in Table 5 is directionally consis-

Suppression α	Logical Accuracy	Δ vs. Baseline
0.00 (baseline)	0.695	—
0.25	0.690	-0.005
0.50	0.695	± 0.000
0.75	0.710	$+0.015$
1.00	0.715	$+0.020$

Table 5: Arithmetic accuracy under progressive suppression of high-betweenness dimensions in `Qwen2-0.5B` (greedy decoding), averaged across three runs. Stronger suppression yields improvement, with a 2.0-point absolute gain at $\alpha = 1.0$.

tent with the descriptive results: dimensions that are highly central under the betweenness criterion do not appear to be uniformly helpful for arithmetic performance, and attenuating them can modestly improve accuracy. That said, this experiment should be read as a proof of possibility rather than a definitive causal identification. It is limited to a single model family, a single task domain, and a single intervention site, and it does not yet include matched controls such as random dimension suppression, high-degree suppression, or magnitude-matched ablations. Nor do we recompute the post-intervention graph, so we do not claim to have shown that suppression literally converts a "creative-like" topology into a "logical-like" one. The present result is therefore best understood as preliminary evidence that topology-derived dimensions can be used to steer computation in a task-relevant way.

Emergent Specialization and Limitations A broader implication of these findings is that task-conditioned internal organization can arise without explicit supervision for task type. None of the models we study were trained to distinguish logical from creative generation as such; they were trained with the generic next-token objective on a mixture of texts spanning both. Yet across model families, we still observe reproducible differences in co-activation topology associated with these task families. We view this not as evidence for human-like cognitive modules inside LLMs, but as evidence that the next-token objective can induce distinct global organizations of representation for different kinds of generation.

Several limitations temper that claim. First, the synthetic benchmark was generated with GPT-4 and filtered with an auxiliary classifier, so it may contain stylistic or distributional artifacts even after quality control. Second, the real-world validation sets differ substantially in prompt length, genre, and expected output structure; they are useful robustness checks, but not controlled matched comparisons. Third, graph construction depends on thresholding, node selection, and the choice to analyze absolute correlations, all of which affect absolute metric values. Relatedly, because betweenness can rise mechanically as graphs become sparser, the creative-task effect should be interpreted jointly with degree rather than as an independent mechanistic signature. Fourth, our main graphs are condition-level summaries built from mean-pooled generated-token representations; they do not capture single-prompt or token-level dynamics, and they cannot distinguish where in the response the relevant routing patterns arise. Fifth, the causal probe is narrow and preliminary, and stronger intervention studies are needed before making firm claims about functional necessity. Finally, our main descriptive analysis is restricted to models up to 3B parameters. Whether the same signatures persist, weaken, or reorganize in frontier-scale models remains open.

8 Conclusion

We studied whether task type is reflected not only in model outputs, but also in the internal organization of computation. Across six models from two transformer families, logical prompts consistently induced denser condition-level co-activation graphs with higher mean degree centrality, whereas creative prompts induced sparser graphs with higher mean betweenness centrality. The same directional pattern appeared on two real-world validation sets, although those comparisons are less controlled than the synthetic benchmark. Topology-derived features were also diagnostically useful in a simple probe, and a preliminary suppression experiment suggested that the identified dimensions may be functionally relevant rather than purely descriptive.

Taken together, these results point to a modest but important conclusion: task family leaves a measurable signature in the global organization of transformer activations. This perspective complements a large body of interpretability work that focuses on *what* models represent or *where* particular behaviors localize. Our results suggest that it is also useful to ask how computation is organized at the level of activation topology. The comparison to convergent and divergent thinking is intended in that spirit: as a source of hypotheses and analytical tools, not as a claim that transformer graphs are direct counterparts of human brain networks.

Limitations

This work has several limitations that bound the strength of its claims. First, the synthetic benchmark was generated with GPT-4 and filtered with an auxiliary classifier, so it may contain stylistic or distributional artifacts even after quality control.

Second, the real-world validation sets differ substantially in prompt length, genre, and expected output structure; they provide robustness checks, but not matched comparisons.

Third, our main analysis is built from mean-pooled hidden states over generated tokens, so some of the measured differences may reflect response structure in addition to task family. Prompt-only analyses, teacher-forced controls, and token-level temporal analyses would help disentangle these factors.

Fourth, graph construction requires design choices, including the correlation threshold, node-retention rule, and the use of absolute correlations. These choices affect absolute metric values, and betweenness in particular can change mechanically as graph density changes.

For that reason, the creative-task betweenness effect should be interpreted jointly with degree rather than as a fully independent mechanistic signature. Fifth, the causal evidence remains narrow: the suppression experiment uses a single model family, a single task domain, and a coarse intervention. It should be read as preliminary directional evidence, not a definitive causal demonstration.

Finally, our descriptive study is limited to models up to 3B parameters. Whether the same signatures persist, weaken, or reorganize in frontier-scale models remains open.

Ethical Considerations

This work uses existing pretrained models, public datasets, and a synthetic prompt set. It does not involve human subjects or private user data. The computational burden is also modest relative to model training, since the analysis operates at inference time on already trained models.

The main ethical risk in this line of work is over-interpretation. Because the paper draws on concepts from cognitive neuroscience, it may invite anthropomorphic readings that go beyond the evidence. We do not claim that LLMs literally instantiate human convergent or divergent thinking, nor that graph-theoretic centrality reveals human-like mental states. The neuroscience framing is used as a hypothesis-generating lens for studying model internals, and the results should be interpreted at that level.

Finally, the labels *logical* and *creative* are operational categories attached to prompt families, not normative judgments about people, cultures, or forms of intelligence. Those categories are imperfect and may reflect biases in prompt generation and benchmark design. We therefore view them as useful experimental abstractions rather than as fixed or universal cognitive categories.

References

Jessica R. Andrews-Hanna and Matthew D. Grilli. 2021. Mapping the imaginative mind: Charting new paths forward. *Current Directions in Psychological Science*, 30(1):82–89.

Roger E. Beaty, Mathias Benedek, Scott Barry Kaufman, and Paul J. Silvia. 2015. Default and executive network coupling supports creative idea production. *Scientific Reports*, 5:10964.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*.

Ed Bullmore and Olaf Sporns. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 889–898.

Vinod Goel, Brian Gold, Shitij Kapur, and Sylvain Houle. 1997. The seats of reason? An imaging study of deductive and inductive reasoning. *NeuroReport*, 8(5):1305–1310.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10):833–838.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 4768–4791.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4129–4138.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Sunil Kumar Maurya, Xin Liu, and Tsuyoshi Murata. 2019. Fast approximations of betweenness centrality with graph neural networks. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2149–2152.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35.

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

Jérôme Prado, Angad Chadha, and James R. Booth. 2011. The brain network for deductive reasoning: A quantitative meta-analysis of 28 neuroimaging studies. *Journal of Cognitive Neuroscience*, 23(11):3483–3497.

Olaf Sporns. 2013. Network attributes for segregation and integration in the human brain. *Current Opinion in Neurobiology*, 23(2):162–171.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35.

Cheng Yang, Jiahao Yu, Qian Liu, and Jie Zhou. 2024. LLM2: Let large language models harness system 2 reasoning. *arXiv preprint arXiv:2412.20372*.

Appendix

A Divergent Thinking Network Organization in Humans

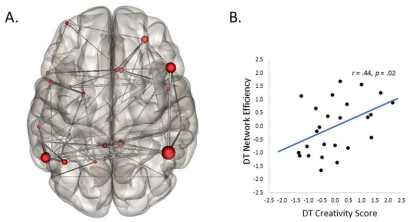

Figure 2: Divergent thinking network organization in humans, reproduced from Beaty et al. (2015) (CC BY 4.0). **(A)** Brain regions from the DMN and ECN engaged during divergent thinking. **(B)** Participants with higher creative ability show higher network efficiency. The result motivates our use of graph-theoretic measures of internal organization, rather than activation strength alone, when analyzing transformer representations.

B Logical and Creative Categories

Logical categories include *arithmetic word problems, symbolic logic, fact-based questions, constraint satisfaction, error detection, definition recall, classification, program tracing, bounded multi-step reasoning,* and *deductive inference*.

Creative categories include *imaginative scenarios, metaphor generation, alternative realities, speculative invention, open-ended brainstorming, story seeds, creative what-if questions, absurd premises, unexpected reinterpretations,* and *associative expansion*.

	Logical	Creative
Ex. 1	*A train travels 120 km in 2 hours. What is its average speed in km/h?*	*What would happen if gravity reversed direction every Tuesday and everything began drifting upward?*
Ex. 2	*A store has 45 apples and sells 18. How many apples remain?*	*Describe a world where colors produce sounds and sounds create textures.*

Table 6: Representative prompt pairs. Logical prompts are designed to elicit constrained solutions and creative prompts invite open-ended generation. This operationalizes our logical/creative contrast at the prompt level without assuming a one-to-one equivalence with human convergent/divergent cognition.

C Activation Pipeline and Robustness Checks

We begin with the methodological appendix because this is the part most likely to be scrutinized in review. The goal of this section is to make the activation-to-graph pipeline fully explicit, show how the real-world datasets were converted into usable inputs, and report the robustness sweeps and LLaMA replication results in a compact but complete form. A list of models used in our experiments can be found in Table 7 below.

Family	Model	Params	Layers
Pythia	`pythia-160m`	160M	12
Pythia	`pythia-410m`	410M	24
Pythia	`pythia-1b`	1B	16
Llama	`Llama-3.2-1B`	1B	16
Llama	`Llama-3.2-3B`	3B	28
Llama	`Llama-3.2-3B-Instruct`	3B	28

Table 7: Models evaluated. The Pythia family enables controlled within-family scale comparisons, while the Llama-3.2 models provide a cross-family check with a different tokenizer, training recipe, and instruction-tuned variant.

C.1 From Hidden States to Graph Metrics

For prompt p, layer l, and generated-token positions $\mathcal{G}_p$, let $\mathbf{h}_{p,t}^{(l)} \in \mathbb{R}^D$ denote the hidden state at token t. Our main prompt-level representation is the generated-token mean

$$\mathbf{a}_p^{(l)} = \frac{1}{|\mathcal{G}_p|} \sum_{t \in \mathcal{G}_p} \mathbf{h}_{p,t}^{(l)}. \tag{6}$$

For a fixed prompt p and layer l, we stack the generated-token hidden states into a matrix

$$\mathbf{X}^{(l,p)} \in \mathbb{R}^{|\mathcal{G}_p| \times D}, \tag{7}$$

$$\mathbf{X}_{t,:}^{(l,p)} = \mathbf{h}_{p,t}^{(l)}. \tag{8}$$

The prompt-level co-activation matrix is then

$$C_{ij}^{(l,p)} = \operatorname{corr}\left(\mathbf{X}_{:,i}^{(l,p)}, \mathbf{X}_{:,j}^{(l,p)}\right). \tag{9}$$

We convert this into a sparse weighted graph by thresholding absolute correlations

$$A_{ij}^{(l,p)} = |C_{ij}^{(l,p)}| \cdot \mathbf{1}\left[|C_{ij}^{(l,p)}| \geq \tau_q\right], \tag{10}$$

where τ_q is the threshold that retains a fixed edge-density quantile. We then keep the top-k nodes by weighted degree, induce a k-node subgraph, and compute the graph-level summaries

$$\mathrm{MDC}(G) = \frac{1}{|V|} \sum_{v \in V} \mathrm{dc}_G(v), \tag{11}$$

$$\mathrm{MBC}(G) = \frac{1}{|V|} \sum_{v \in V} \mathrm{bc}_G(v). \tag{12}$$

Throughout the appendix, we report

$$\Delta_{\mathrm{MDC}} = \mathrm{MDC}_{\mathrm{logical}} - \mathrm{MDC}_{\mathrm{creative}}, \quad (13)$$

$$\Delta_{\mathrm{MBC}} = \mathrm{MBC}_{\mathrm{creative}} - \mathrm{MBC}_{\mathrm{logical}}, \quad (14)$$

so positive values always align with the main-paper claim.

C.2 Real-World Dataset Preprocessing

The real-world validation datasets are less controlled than the synthetic benchmark, so it is important to make the preprocessing explicit. We downloaded WRITING-PROMPTS and SVAMP from Hugging Face, extracted the prompt-like text field, filtered to non-empty entries, tokenized with the Pythia tokenizer, and retained the first 200 prompts under 128 tokens. This procedure deliberately favors reproducibility over aggressive dataset cleaning.

C.3 Pythia Threshold Robustness

The main text reports one analysis layer per Pythia model. To test whether the headline effect depends on a single graph-building choice, we recomputed the comparison across all 25 threshold settings per model (5 edge densities $\times$ 5 node budgets), keeping the layer fixed to the one used in the main paper. Table 8 reports the full sweep in compact form. Each cell is shown as $\Delta_{\mathrm{MDC}}/\Delta_{\mathrm{MBC}}$.

C.4 LLaMA Real-World Replication

The LLaMA real-world runs are informative precisely because they are not uniformly clean. We therefore report them directly rather than hiding them. Table 9 shows the base-model results and Table 10 shows the instruction-tuned checkpoint.

Two patterns are important. First, the base models largely reproduce the paper's main dissociation: `Llama-3.2-3B` matches the expected direction in all 9 tested settings, while `Llama-3.2-1B` matches in 5 of 9 settings, ties once, and reverses in 3. Second, the instruction-tuned model behaves qualitatively differently: it matches the expected direction in only 1 of 9 settings, and in the remaining 8 the sign flips for both metrics. This is not random noise. Compared with the base models, the instruct checkpoint produces much sparser graphs overall and much larger betweenness values for both datasets (for example, at density 0.005 and top-k = 30, graph density falls from roughly 0.97–0.99 in the base models to 0.59 on WRITING-PROMPTS and 0.42 on SVAMP). A plausible interpretation is that instruction tuning introduces a strong shared chat-style response prior that compresses the topological contrast between open-ended and logical prompts; in other words, the model is no longer expressing only task-conditioned topology, but a mixture of task topology and instruction-following format effects.

D Dataset Examples

After documenting the pipeline, we include the benchmark design itself. The synthetic dataset was constructed to maximize task-type contrast while avoiding trivial lexical cues. To make that construction easier to audit, we include one representative prompt for every category rather than only a small showcase subset.

E Per-Model Results

We include per-model ROC curves for the topology probe, entropy baseline, their combination, and a random classifier baseline (diagonal) for reference.

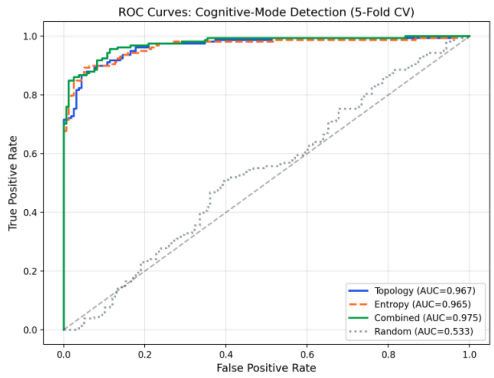

Figure 5: ROC curves for task-type classification (10-fold stratified CV). On our main benchmark, the topology probe (AUC = 0.967) and entropy baseline (AUC = 0.965) are statistically indistinguishable, while their combination yields AUC = 0.975.

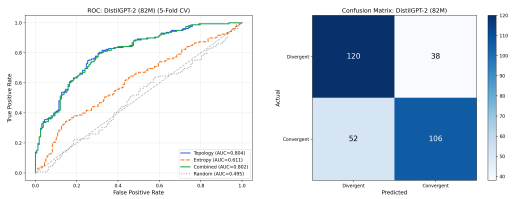

Figure 6: ROC curves on `distilGPT-2`. The x-axis is false positive rate and the y-axis is true positive rate.

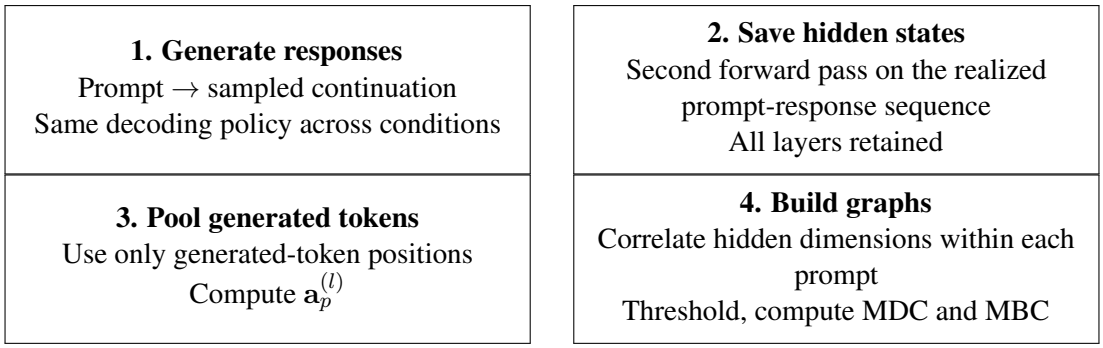

Figure 3: End-to-end pipeline used for the appendix analyses. Prompt-level graphs are built from generated-token hidden states within each prompt, while mean-pooled states are retained for prompt-level summaries and probe features.

WRITING-PROMPTS (creative proxy)
Extraction: prompt/text field from the dataset
Selection: first 200 non-empty prompts under 128 tokens
Length profile: 132.2 ± 70.6 characters

Example snippet:
You've finally managed to discover the secret to immortality. Suddenly, Death appears before you and says: when you realize living forever sucks, call this number.

SVAMP (logical proxy)
Extraction: instruction-style arithmetic prompt field
Selection: first 200 non-empty prompts under 128 tokens
Length profile: 301.8 ± 41.6 characters

Example snippet:
There are 87 oranges and 290 bananas in Philip's collection. If the bananas are organized into 2 groups, how big is each group of bananas?

Figure 4: Real-world preprocessing summary. The datasets remain intentionally different in genre and length, which is why we use them as robustness checks rather than as the primary controlled evidence base.

Model	Density	$k = 50$	$k = 75$	$k = 100$	$k = 150$	$k = 200$
	0.02	0.293/0.0064	0.307/0.0048	0.270/0.0034	0.147/0.0016	0.078/0.0007
	0.05	0.330/0.0069	0.354/0.0049	0.346/0.0035	0.239/0.0016	0.161/0.0008
`pythia-160m` (layer 7)	0.10	0.272/0.0057	0.323/0.0044	0.347/0.0035	0.281/0.0019	0.230/0.0012
	0.15	0.236/0.0049	0.305/0.0042	0.330/0.0034	0.297/0.0020	0.257/0.0013
	0.20	0.219/0.0046	0.279/0.0038	0.312/0.0032	0.289/0.0020	0.259/0.0013
	0.02	0.556/0.0132	0.461/0.0070	0.380/0.0044	0.237/0.0021	0.160/0.0013
	0.05	0.494/0.0103	0.458/0.0063	0.441/0.0045	0.323/0.0022	0.250/0.0013
`pythia-410m` (layer 13)	0.10	0.418/0.0087	0.425/0.0058	0.444/0.0045	0.366/0.0025	0.307/0.0016
	0.15	0.349/0.0073	0.379/0.0052	0.407/0.0042	0.347/0.0023	0.303/0.0015
	0.20	0.316/0.0066	0.337/0.0046	0.356/0.0036	0.329/0.0022	0.298/0.0015
	0.02	0.402/0.0095	0.380/0.0061	0.329/0.0039	0.271/0.0023	0.221/0.0015
	0.05	0.416/0.0087	0.424/0.0058	0.394/0.0040	0.327/0.0022	0.290/0.0015
`pythia-1b` (layer 9)	0.10	0.357/0.0074	0.383/0.0052	0.392/0.0040	0.340/0.0023	0.321/0.0016
	0.15	0.353/0.0074	0.343/0.0047	0.363/0.0037	0.336/0.0023	0.313/0.0016
	0.20	0.328/0.0068	0.330/0.0045	0.321/0.0033	0.314/0.0021	0.300/0.0015

Table 8: Full Pythia robustness sweep at the main-paper analysis layers. Each cell reports $\Delta_{\text{MDC}}/\Delta_{\text{MBC}}$, where $\Delta_{\text{MDC}} = \text{MDC}_{\text{logical}} - \text{MDC}_{\text{creative}}$ and $\Delta_{\text{MBC}} = \text{MBC}_{\text{creative}} - \text{MBC}_{\text{logical}}$. All 75 entries preserve the main direction of effect.

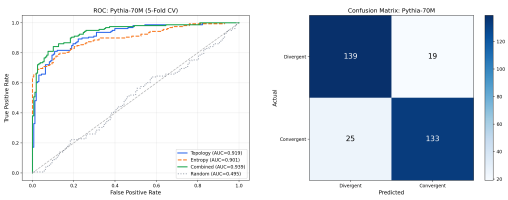

Figure 7: ROC curves on `Pythia-70M`. The x-axis is false positive rate and the y-axis is true positive rate.

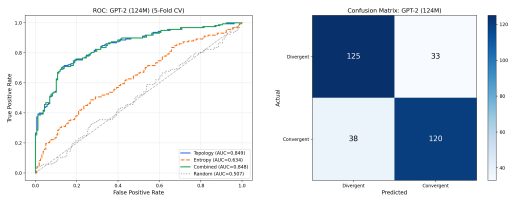

Figure 8: ROC curves on `GPT-2`. The x-axis is false positive rate and the y-axis is true positive rate.

Model	Setting	Writing − SVAMP MBC	SVAMP − Writing MDC	Direction	Interpretation
Llama-3.2-1B	$d = 0.005, k = 30$	$+0.001560$	$+0.043678$	Match	clean replication
	$d = 0.005, k = 50$	$+0.000765$	$+0.036735$	Match	clean replication
	$d = 0.005, k = 75$	$+0.000094$	$+0.006847$	Match	weak but same sign
	$d = 0.010, k = 30$	$+0.000246$	$+0.006897$	Match	weak but same sign
	$d = 0.010, k = 50$	$+0.000187$	$+0.008980$	Match	weak but same sign
	$d = 0.010, k = 75$	-0.000188	-0.013694	Reverse	dense setting erodes separation
	$d = 0.020, k = 30$	$+0.000000$	$+0.000000$	Tie	fully saturated graph
	$d = 0.020, k = 50$	-0.000017	-0.000816	Reverse	near-saturation
	$d = 0.020, k = 75$	-0.000262	-0.019099	Reverse	near-saturation
Llama-3.2-3B	$d = 0.005, k = 30$	$+0.000903$	$+0.025287$	Match	clean replication
	$d = 0.005, k = 50$	$+0.001173$	$+0.056327$	Match	clean replication
	$d = 0.005, k = 75$	$+0.001303$	$+0.095135$	Match	strongest separation
	$d = 0.010, k = 30$	$+0.000411$	$+0.011494$	Match	clean replication
	$d = 0.010, k = 50$	$+0.000833$	$+0.040000$	Match	clean replication
	$d = 0.010, k = 75$	$+0.000810$	$+0.059099$	Match	clean replication
	$d = 0.020, k = 30$	$+0.000164$	$+0.004598$	Match	weak but same sign
	$d = 0.020, k = 50$	$+0.000391$	$+0.018776$	Match	clean replication
	$d = 0.020, k = 75$	$+0.000252$	$+0.018378$	Match	clean replication

Table 9: Real-world LLaMA base-model results, summarized as pairwise gaps between WRITING-PROMPTS and SVAMP. Positive values in both columns match the main-paper direction. The 3B base model is fully consistent across the tested grid; the 1B base model is mostly, but not perfectly, consistent.

Setting	Writing − SVAMP MBC	SVAMP − Writing MDC	Direction	Interpretation
$d = 0.005, k = 30$	-0.011741	-0.170115	Reverse	strongest failure case
$d = 0.005, k = 50$	-0.005102	-0.103673	Reverse	clear sign flip
$d = 0.005, k = 75$	-0.002103	-0.048649	Reverse	clear sign flip
$d = 0.010, k = 30$	-0.004187	-0.080460	Reverse	clear sign flip
$d = 0.010, k = 50$	-0.001837	-0.041633	Reverse	clear sign flip
$d = 0.010, k = 75$	-0.000765	-0.012973	Reverse	weak but reversed
$d = 0.020, k = 30$	-0.003612	-0.096552	Reverse	clear sign flip
$d = 0.020, k = 50$	-0.000170	-0.004898	Reverse	near-neutral but reversed
$d = 0.020, k = 75$	$+0.000578$	$+0.042523$	Match	only congruent setting

Table 10: Real-world results for `Llama-3.2-3B-Instruct`. Unlike the base models, the instruction-tuned checkpoint usually reverses the main-paper effect. We include this table explicitly because it is an important boundary condition on the claim: instruction tuning appears to substantially reshape activation topology on these real-world prompts.

Benchmark side	Task families	Benchmark side	Task families
Logical	Arithmetic word problems; symbolic logic; fact-based questions; constraint satisfaction puzzles; error detection; definition recall; classification tasks; program output tracing; step-limited reasoning; deductive inference.	Creative	Imaginative scenarios; metaphor creation; alternative realities; speculative innovation; open-ended brainstorming; story seeds; creative what-if questions; absurd premises; unexpected reinterpretations; associative expansion.

Table 11: The 20 task families used in the synthetic benchmark. The dataset is balanced by construction: 10 logical categories and 10 creative categories, with 30 prompts per category.

Logical category	Representative prompt	Creative category	Representative prompt
Arithmetic word problems	*Trains traveling 120 kilometers in 2 hours maintain average speeds of how many kilometers per hour?*	Imaginative scenarios	*What would happen if gravity reversed direction every Tuesday afternoon causing objects to float upward?*
Symbolic logic	*Birds that can fly and penguins being birds logically imply can penguins fly?*	Metaphor creation	*Create a metaphor comparing knowledge to something unexpected and vivid using natural imagery.*
Fact based questions	*Oxygen elements have chemical symbols of what in periodic table classifications?*	Alternative realities	*What would history look like if humans evolved underwater developing gills instead of lungs?*
Constraint satisfaction puzzles	*Three items arranged in order require A before B and B before C in sequence.*	Speculative innovation	*What new form of transportation could emerge if we could manipulate magnetic fields?*
Error detection	*Sequences containing 2 4 6 8 11 12 have errors where one number doesn't follow the pattern.*	Open ended brainstorming	*Generate as many uses as possible for a device that creates temporary gravity fields.*
Definition recall	*Prime numbers in mathematics are defined as numbers having what specific properties?*	Story seeds	*Write about a person who discovers their shadow has been living a separate life when they sleep.*
Classification tasks	*Items including apple car dog book tree must be classified into categories based on living or non living status.*	Creative what if questions	*What if every person had a unique color that only they could see?*
Program output tracing	*Programs setting X to 5 then adding 3 to X produce final values of what number?*	Absurd premises	*Describe a world where the concept of up and down changes direction every hour.*
Step limited reasoning	*Having 10 items and removing 3 items leaves how many items after this single step?*	Unexpected reinterpretations	*How would you reinterpret the concept of a library as a place where stories come to rest?*
Deductive inference	*Mammals having hair and whales being mammals means whales have what property?*	Associative expansion	*Start with the word cloud and expand it into a network of related concepts and images.*

Table 12: One representative prompt from every logical and creative category in the synthetic benchmark. This broader coverage makes the benchmark easier to inspect: logical prompts constrain answer space, whereas creative prompts invite expansion, association, reinterpretation, or counterfactual generation.