

Spectral Geometry of Transformer Cognition:

Heat Kernel Analysis Reveals Functional Organization in Language Models

Bee Rosa Davis

bee_davis@alumni.brown.edu

December 1st, 2025 4:04pm

Abstract

We introduce spectral geometry methods for analyzing the internal organization of transformer language models. By treating attention patterns as weighted graphs and computing their Laplacian spectra, we extract geometric invariants—eigenvalue gaps, heat kernel signatures, and multi-scale curvature estimates—that reveal functional structure invisible to standard interpretability techniques.

Experiments on the GPT-2 model family (12–36 layers, 124M–774M parameters) and TinyLlama (1.1B parameters) demonstrate: (1) a universal U-shaped curvature profile with high curvature at entry/exit layers and minimal curvature in middle layers; (2) task-specific activation patterns where factual recall peaks in middle layers (~50% depth) while pattern matching concentrates in early layers (<20% depth); (3) spectral gap collapse as a predictor of model failure on difficult tasks; (4) a conserved two-stage memory retrieval circuit with early activation triggering (~6% depth) followed by stable geometric transmission (~64% depth); and (5) geometric effort as a signature of retrieval attempt, distinguishing grounded factual recall from confabulation independent of output correctness.

The hallucination detection finding is particularly significant: forced correct and forced incorrect completions show identical geometric effort, while confabulation prompts show markedly lower effort. This indicates that geometric signatures measure whether the model is *attempting* grounded retrieval, not whether the retrieved answer is correct—enabling inference-time detection of confabulation without external verification.

These findings establish spectral geometry as a principled framework for transformer interpretability and suggest that geometric properties of attention patterns encode meaningful information about model computation.

Keywords: transformer interpretability, spectral graph theory, heat kernels, attention mechanisms, geometric deep learning, hallucination detection

1 Introduction

The remarkable capabilities of transformer-based language models have outpaced our understanding of their internal mechanisms. While these models demonstrate sophisticated reasoning, factual recall, and linguistic competence, the computational processes underlying these abilities remain largely opaque. Standard interpretability approaches—attention visualization, probing classifiers, activation patching—provide local insights but lack a unified framework for understanding global organization.

We propose that *spectral geometry* offers such a framework. By treating the attention pattern at each layer as a weighted graph and analyzing its Laplacian spectrum, we can extract geometric invariants that characterize the “shape” of information flow. These invariants—eigenvalues, eigenfunctions, heat kernels, and derived quantities like spectral gaps and curvature estimates—are intrinsic to the geometry and robust to superficial variations in attention weights.

1.1 Contributions

This paper makes the following contributions:

1. **Methodological framework:** We develop a complete pipeline for spectral analysis of transformer attention patterns, including multi-scale heat kernel analysis, curvature estimation from small- t asymptotics, and Heat Kernel Signatures (HKS) for component fingerprinting.
2. **Universal geometric organization:** We demonstrate that GPT-2 family models exhibit consistent geometric profiles when normalized by depth, including a characteristic U-shaped curvature curve and systematic variation in spectral gaps across layers.
3. **Task-specific activation mapping:** We show that different cognitive tasks—factual recall, arithmetic, pattern matching, sentiment analysis, analogical reasoning—activate geometrically distinct pathways through the network.
4. **Failure prediction:** We identify spectral gap collapse as a geometric signature of model failure, observable before incorrect outputs are generated.
5. **Memory circuit characterization:** We discover a conserved two-stage architecture for factual recall, with early-layer pattern matching followed by middle-layer stable transmission through high-spectral-gap regions.
6. **Retrieval-based hallucination detection:** We demonstrate that geometric effort measures retrieval attempt rather than output correctness, enabling distinction between grounded factual recall and confabulation at inference time without external verification. This finding, replicated across GPT-2 and TinyLlama, reframes hallucination detection from “is this correct?” to “is the model attempting to ground this output?”

1.2 Related Work

Our approach builds on several research traditions. *Geometric deep learning* [Bronstein et al., 2017] has established connections between neural network architectures and differential geometry, though primarily for graph neural networks rather than transformers. *Spectral graph theory* [Chung, 1997] provides the mathematical foundation for our Laplacian-based analysis. Recent work on *transformer interpretability* [Elhage et al., 2021, Olsson et al., 2022] has identified specific computational motifs (induction heads, copying mechanisms) but lacks a unified geometric perspective.

The connection between attention and diffusion processes has been noted [Veličković et al., 2018], but systematic spectral analysis of attention patterns in language models is novel to this work. Our approach is also related to persistent homology methods for neural network analysis [Rieck et al., 2019], though we focus on spectral rather than topological invariants.

This work extends the geometric framework for transformer analysis developed in prior publications. The Davis Conjecture [Davis, 2025a] established that context windows function as holonomy horizons in a functorial formulation of transformer computation, with semantic coherence inversely proportional to representational curvature. The field equations of semantic coherence [Davis, 2025b] formalized this relationship, showing that the master equation $C = \tau/K$ governs the relationship between reasoning budget, curvature, and coherent output generation. The geometry of sameness [Davis, 2025d,g] provided the foundational ε -equivalence framework unifying translation and distance in semantic space, while gauge-theoretic formulations [Davis, 2025c] demonstrated how transformers realize semantic sameness through parallel transport. The Davis Manifold framework [Davis, 2025e] introduced geometry-first detection with compositional error budgets, enabling practical application of these theoretical insights. Causal validation [Davis, 2025f] confirmed that curvature reduction during training extends context horizons

as predicted, with 77% curvature reduction yielding $20\times$ horizon extension. For comprehensive treatment of these ideas, see Davis [2025h].

2 Mathematical Background

We briefly review the mathematical concepts underlying our analysis. For comprehensive treatments, see Chung [1997] for spectral graph theory and Grigor’yan [2009] for heat kernels on manifolds.

2.1 From Attention to Graph Laplacians

Consider a transformer layer with attention weights $A \in \mathbb{R}^{n \times n}$, where A_{ij} represents the attention from position i to position j for a sequence of length n . We construct a weighted graph with adjacency matrix:

$$W = \frac{1}{2}(A + A^\top) \quad (1)$$

symmetrizing to ensure real eigenvalues. The *degree matrix* $D = \text{diag}(d_1, \dots, d_n)$ has entries $d_i = \sum_j W_{ij}$.

Definition 1 (Graph Laplacian). *The unnormalized Laplacian is $L = D - W$. The normalized Laplacian is:*

$$\mathcal{L} = I - D^{-1/2}WD^{-1/2} \quad (2)$$

The normalized Laplacian has eigenvalues in $[0, 2]$ and is our primary object of study. Its spectrum encodes the geometry of information flow through the attention pattern.

2.2 Spectral Decomposition

Let $0 = \lambda_0 \leq \lambda_1 \leq \dots \leq \lambda_{n-1}$ be the eigenvalues of $\mathcal{L}$ with corresponding orthonormal eigenvectors $\phi_0, \phi_1, \dots, \phi_{n-1}$. Key quantities derived from this spectrum include:

Definition 2 (Spectral Gap). *The spectral gap is $\text{gap}(\mathcal{L}) = \lambda_1 - \lambda_0 = \lambda_1$, measuring the separation between the trivial eigenvalue and the first non-trivial mode.*

A large spectral gap indicates well-separated modes and “clean” geometry; a small gap indicates mode mixing and complex structure.

Definition 3 (Participation Ratio). *The participation ratio at time scale t is:*

$$\text{PR}(t) = \frac{1}{\sum_i p_i(t)^2}, \quad \text{where } p_i(t) = \frac{e^{-\lambda_i t}}{\sum_j e^{-\lambda_j t}} \quad (3)$$

The participation ratio measures how many modes are effectively active at scale t . $\text{PR} = 1$ indicates single-mode dominance; $\text{PR} = n$ indicates uniform activation.

2.3 Heat Kernels and Multi-Scale Analysis

The heat kernel provides a natural multi-scale lens on graph geometry.

Definition 4 (Heat Kernel). *The heat kernel on the graph at time t is:*

$$K(x, y, t) = \sum_{i=0}^{n-1} e^{-\lambda_i t} \phi_i(x) \phi_i(y) \quad (4)$$

Intuitively, $K(x, y, t)$ measures the amount of “heat” diffusing from x to y in time t . Small t captures local geometry; large t reveals global structure.

Definition 5 (Heat Trace). *The heat trace is:*

$$Z(t) = \text{Tr}(e^{-t\mathcal{L}}) = \sum_{i=0}^{n-1} e^{-\lambda_i t} \quad (5)$$

The heat trace is a spectral “fingerprint” of the graph. Graphs with similar $Z(t)$ curves have similar spectral geometry.

Definition 6 (Heat Kernel Signature). *The Heat Kernel Signature at point x is:*

$$\text{HKS}(x, t) = K(x, x, t) = \sum_{i=0}^{n-1} e^{-\lambda_i t} \phi_i(x)^2 \quad (6)$$

The HKS provides a multi-scale descriptor of each point’s local geometry. Points with similar HKS profiles occupy geometrically similar positions in the graph.

2.4 Curvature from Heat Kernel Asymptotics

On a Riemannian manifold, the heat kernel diagonal has a well-known small- t expansion:

$$K(x, x, t) \sim (4\pi t)^{-d/2} \left(1 + \frac{R(x)}{6} t + O(t^2) \right) \quad (7)$$

where $R(x)$ is the scalar curvature at x and d is the dimension. While graphs are not manifolds, we can use this expansion to define an approximate curvature:

Definition 7 (Spectral Curvature Estimate). *Given $K(x, x, t)$ for small t , we estimate local curvature as:*

$$\hat{K}(x) = 6 \cdot \frac{\partial}{\partial t} \log K(x, x, t) \Big|_{t \rightarrow 0^+} \quad (8)$$

In practice, we compute this via linear regression on $\log K(x, x, t)$ for small t values. This provides a scalar curvature proxy at each position.

2.5 Diffusion Distance

The heat kernel induces a family of metrics on the graph:

Definition 8 (Diffusion Distance). *The diffusion distance at scale t is:*

$$D_t(x, y)^2 = K(x, x, t) + K(y, y, t) - 2K(x, y, t) \quad (9)$$

Different structures become visible at different scales: local clustering at small t , global organization at large t .

3 Methods

3.1 Attention Extraction

We extract attention patterns from transformer models using forward hooks on attention modules. For a model with L layers and H heads per layer, we obtain attention tensors $A^{(\ell, h)} \in \mathbb{R}^{n \times n}$ for each layer $\ell \in \{0, \dots, L-1\}$ and head $h \in \{0, \dots, H-1\}$.

For layer-level analysis, we average across heads:

$$\bar{A}^{(\ell)} = \frac{1}{H} \sum_{h=0}^{H-1} A^{(\ell, h)} \quad (10)$$

For head-level analysis, we analyze each $A^{(\ell, h)}$ individually.

3.2 Spectral Analysis Pipeline

Our analysis pipeline proceeds as follows:

Algorithm 1 Spectral Geometry Analysis

Require: Attention matrix $A \in \mathbb{R}^{n \times n}$, time scales $\{t_1, \dots, t_T\}$

- 1: Symmetrize: $W \leftarrow (A + A^\top)/2$
 - 2: Compute normalized Laplacian: $\mathcal{L} \leftarrow I - D^{-1/2}WD^{-1/2}$
 - 3: Compute spectral decomposition: $(\lambda_i, \phi_i) \leftarrow \text{eigsh}(\mathcal{L}, k)$
 - 4: **for** each time scale t_j **do**
 - 5: Compute heat kernel: $K_{t_j} \leftarrow \sum_i e^{-\lambda_i t_j} \phi_i \phi_i^\top$
 - 6: Compute heat trace: $Z(t_j) \leftarrow \text{Tr}(K_{t_j})$
 - 7: Compute HKS: $\text{HKS}(\cdot, t_j) \leftarrow \text{diag}(K_{t_j})$
 - 8: **end for**
 - 9: Compute spectral gap: $\text{gap} \leftarrow \lambda_1$
 - 10: Compute participation ratio: $\text{PR} \leftarrow 1/\sum_i p_i^2$
 - 11: Estimate curvature from small- t HKS
 - 12: **return** $\{\lambda_i\}, Z(t), \text{HKS}, \text{gap}, \text{PR}, \hat{K}$
-

We use $k = \min(n - 1, 50)$ eigenvalues and time scales $t \in \{10^{-2}, 10^{-1.5}, \dots, 10^2\}$ spanning four orders of magnitude.

3.3 Task Library

To probe functional organization, we designed a battery of cognitive tasks:

Table 1: Cognitive task categories for functional mapping

Category	Example Prompt	Cognitive Demand
Arithmetic	“What is $17 + 25$? The answer is”	Computation
Pattern Matching	“Complete: 2, 4, 6, 8, __”	Sequence recognition
Factual Recall	“The capital of France is”	Knowledge retrieval
Sentiment	“This movie was wonderful. The sentiment is”	Affect classification
Analogical	“A library is to books as a museum is to”	Compositional reasoning
Logical	“If all birds can fly and a robin is a bird, then”	Deductive inference

For difficulty gradient analysis, we vary task complexity within categories (e.g., easy/medium/hard arithmetic).

3.4 Models

We analyze the GPT-2 model family and TinyLlama:

Table 2: Models analyzed

Model	Layers	Heads	Parameters
GPT-2	12	12	124M
GPT-2 Medium	24	16	345M
GPT-2 Large	36	20	774M
TinyLlama-1.1B	22	32	1.1B

All experiments use the HuggingFace Transformers implementation with `output_attentions=True`.

4 Results

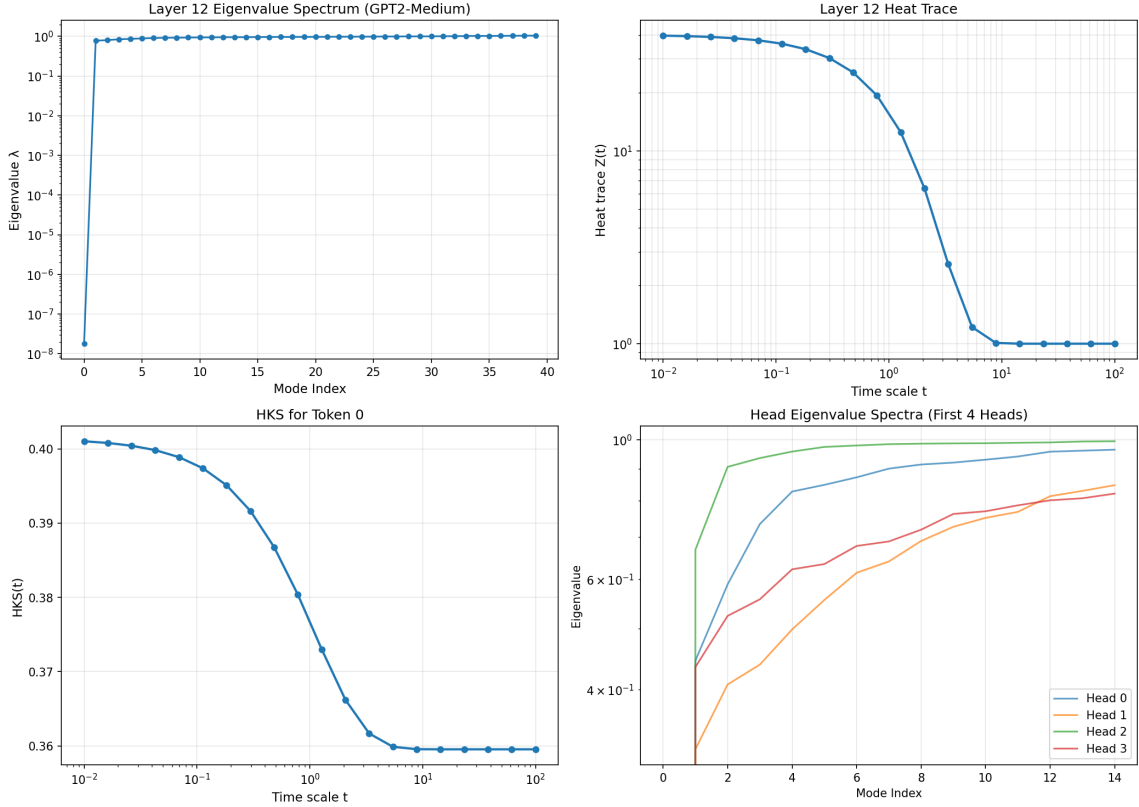


Figure 1: Single-layer spectral analysis (GPT-2 Medium, Layer 12). Top-left: Eigenvalue spectrum showing rapid decay after λ_1 . Top-right: Heat trace $Z(t)$ across time scales. Bottom-left: Heat Kernel Signatures for all tokens. Bottom-right: Curvature estimates from small- t HKS asymptotics.

4.1 Universal Layer-wise Geometry

Observation 1 (U-Shaped Curvature Profile). *Across all models tested, we observe a characteristic U-shaped curvature profile: high curvature at entry layers, minimum curvature in middle layers, and elevated curvature at exit layers.*

Table 3 summarizes this pattern:

Table 3: Curvature profile across models (spectral curvature estimate)

Model	Entry (L0)	Middle	Exit
GPT-2 (12L)	10.3	6.1–6.8	7.7
GPT-2 Medium (24L)	8.7	6.2–6.5	7.9
GPT-2 Large (36L)	8.5	6.0–6.3	8.2

This pattern suggests that entry and exit layers perform geometrically demanding transformations (discrete $\leftrightarrow$ continuous mappings), while middle layers operate in a geometrically stable regime suitable for compositional processing.

Observation 2 (Spectral Gap Evolution). *The spectral gap increases from entry to middle layers, peaking at approximately 50–75% depth, then decreases toward the exit.*

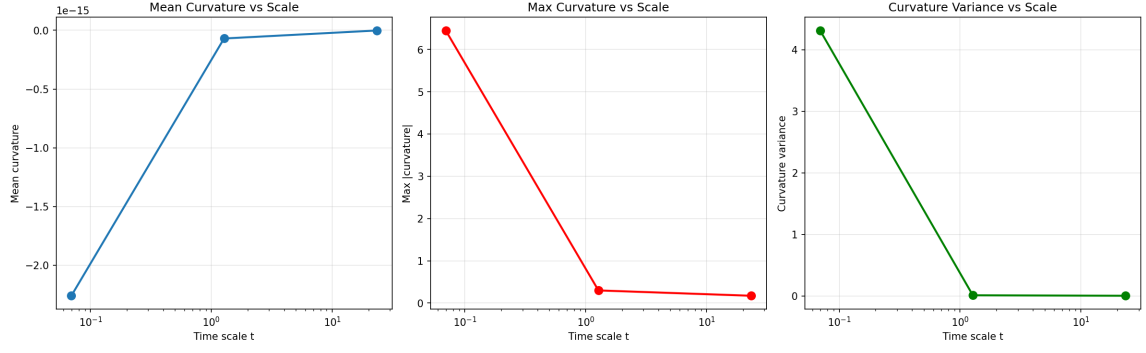


Figure 2: Multi-scale curvature analysis (GPT-2 Medium, Layer 12). Mean curvature approaches zero at all scales as positive and negative regions balance. Maximum curvature and curvature distributions reveal geometric heterogeneity across token positions.

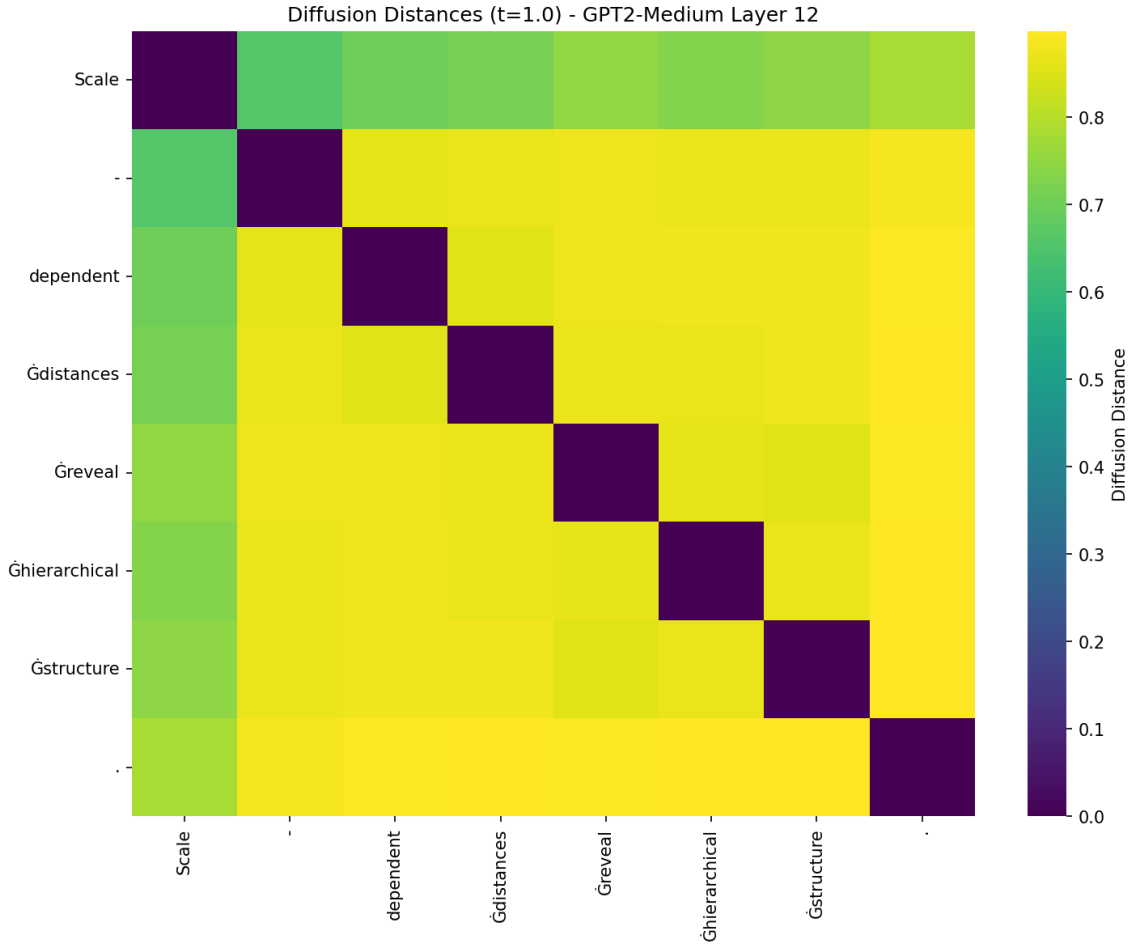


Figure 3: Diffusion distances reveal hierarchical token structure (Layer 12, $t = 1.0$). Tokens cluster into groups with small within-group distances and larger between-group distances, reflecting semantic and syntactic organization.

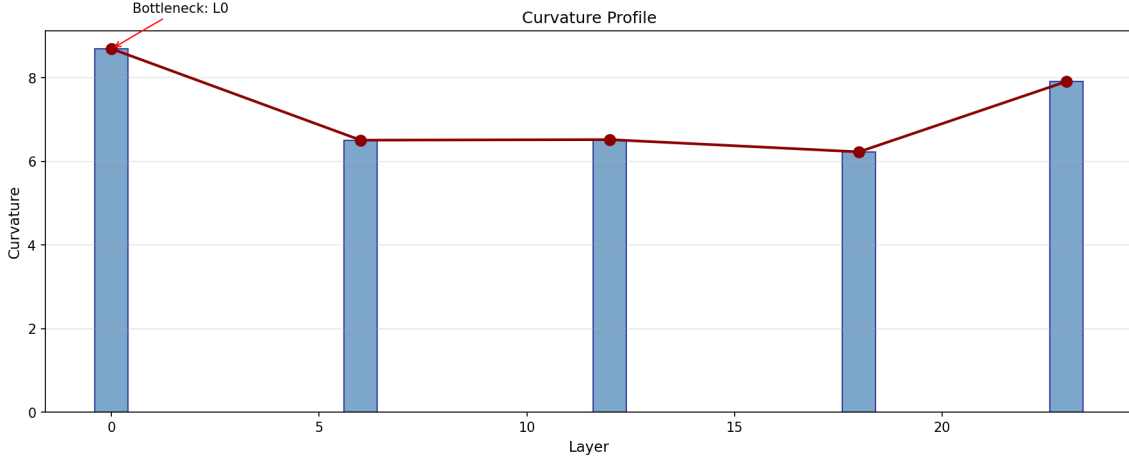


Figure 4: Universal layer-wise geometry in GPT-2 Medium. (a) Spectral properties across layers: eigenvalue spectra (top-left), spectral gaps peaking at L12–L18 (top-right), participation ratio minimizing at L12 (bottom-left), and curvature estimates (bottom-right). (b) The characteristic U-shaped curvature profile.

Table 4: Spectral gap by layer position (GPT-2 Medium)

Layer Position	Layer Index	Spectral Gap
Entry (0%)	L0	0.73
Early (25%)	L6	0.77
Middle (50%)	L12	0.84
Late (75%)	L18	0.84
Exit (96%)	L23	0.68

Large spectral gaps indicate clean mode separation and stable information flow. The peak at 50–75% depth identifies a “reasoning core” with optimal geometric properties.

4.2 Task-Specific Activation Patterns

Different cognitive tasks activate geometrically distinct pathways through the network.

Observation 3 (Factual Recall Uses Middle Layers). *Factual recall tasks (“The capital of France is”) show peak activation at middle layers with high spectral gaps, consistent with retrieval from a geometrically stable “memory region.”*

Observation 4 (Pattern Matching Uses Early Layers). *Pattern matching, arithmetic, and sentiment tasks concentrate activation in early layers (L0–L6), suggesting surface-level pattern recognition without deep geometric processing.*

Observation 5 (Analogical Reasoning is Geometrically Distinct). *Unlike other “reasoning” tasks that concentrate in early layers, analogical reasoning (“A is to B as C is to __”) shows sustained middle-layer activation with high curvature throughout. This suggests genuine compositional processing rather than pattern matching.*

4.3 Difficulty Gradients

Task difficulty modulates geometric signatures in predictable ways.

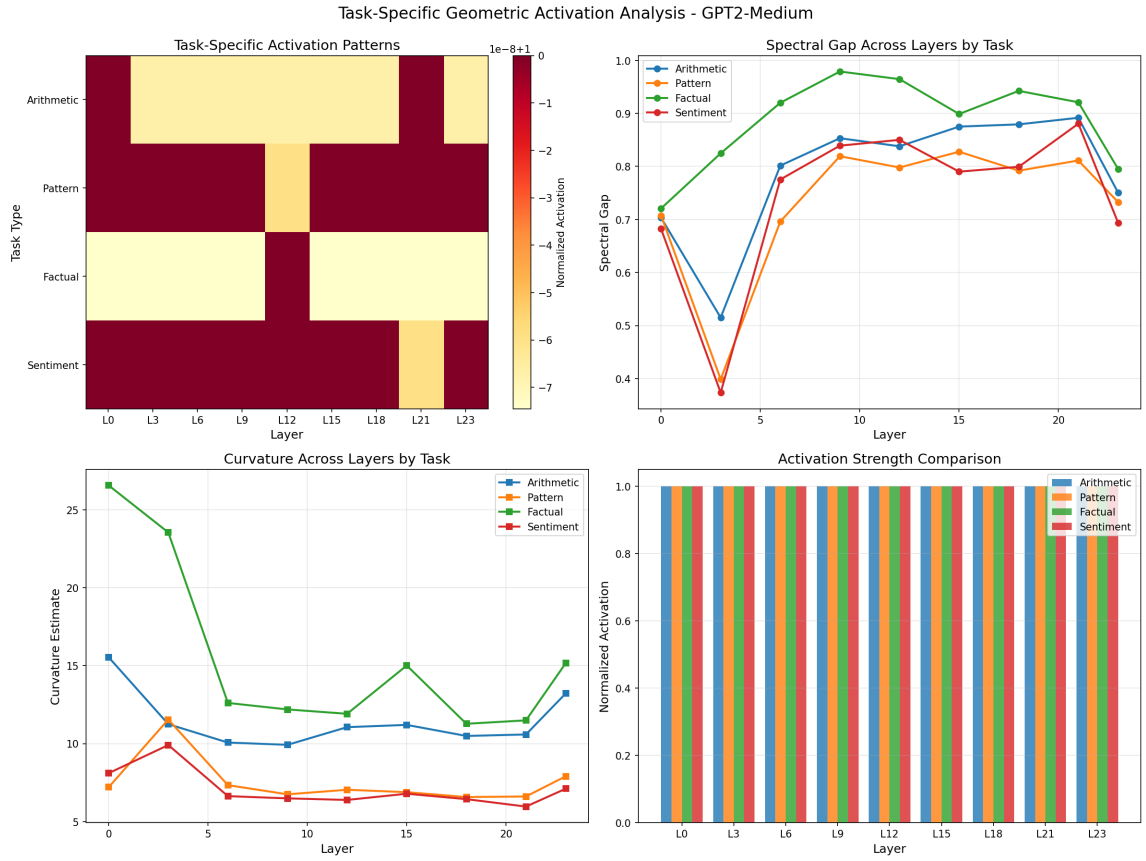


Figure 5: Task-specific geometric activation patterns. Top-left: Activation heatmap showing factual recall’s unique L12 activation (bright spot) vs. early-layer concentration for other tasks. Task-specific spectral gap and curvature evolution reveal distinct computational pathways.

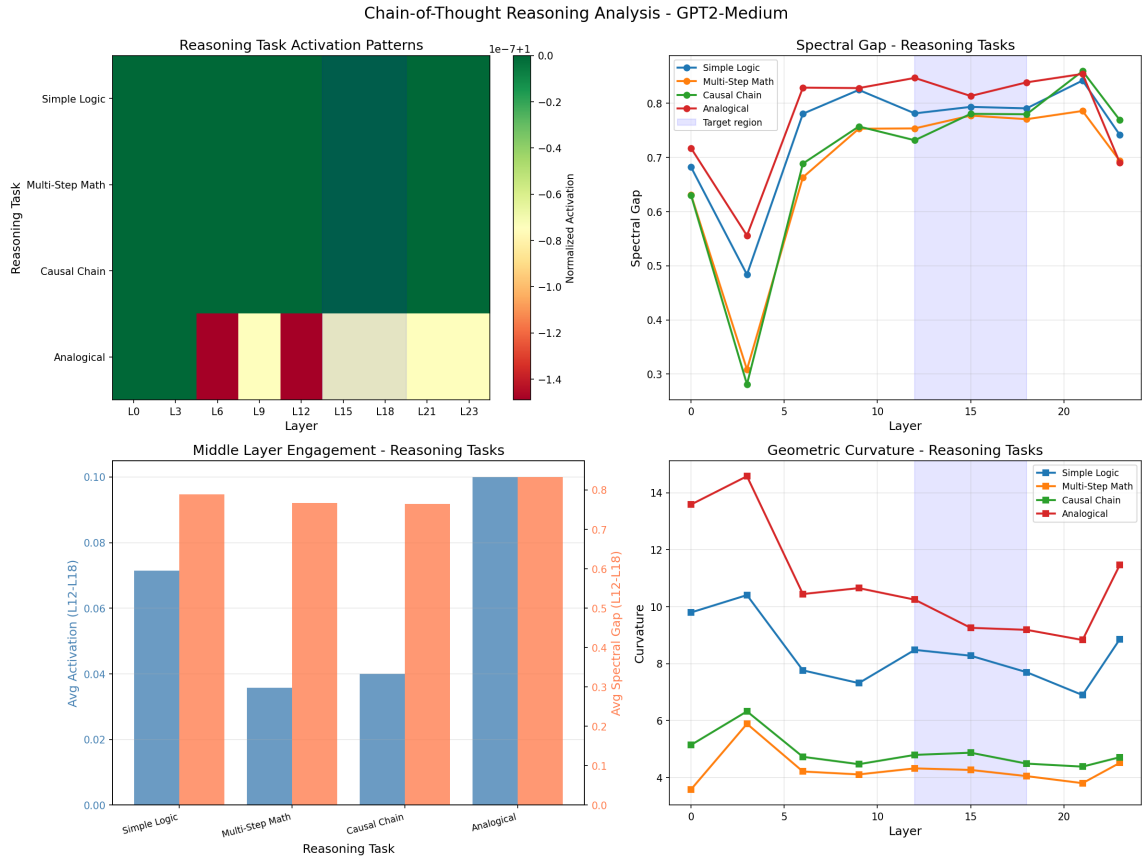


Figure 6: Chain-of-thought reasoning analysis. Activation heatmap showing geometric signatures during multi-step reasoning. Sustained middle-layer activation distinguishes compositional reasoning from pattern matching.

Table 5: Task-specific geometric signatures (GPT-2 Medium)

Task	Peak Layer	Depth	Spectral Gap	Curvature
Factual Recall	L12	50%	0.96	11.9
Arithmetic	L0	0%	0.70	15.5
Pattern Matching	L0	0%	0.71	7.2
Sentiment	L0	0%	0.68	8.1
Analogical	L6–L18	25–75%	0.83	8–14

Observation 6 (Easy Facts Show Recognition Bursts). *Easy factual prompts (“The sky is”) trigger extremely high curvature ($K \approx 32\text{--}42$) in early layers—a geometric “recognition burst” indicating immediate retrieval of overlearned associations.*

Observation 7 (Hard Facts Show Distributed Processing). *Difficult factual prompts (“The chemical formula for sulfuric acid is”) show lower peak curvature ($K \approx 20$) and more distributed activation, consistent with effortful retrieval through extended pathways.*

Table 6: Difficulty gradient for factual recall

Difficulty	Example	Peak Curvature	Peak Layer
Easy	“The sky is”	42.0	L0
Medium	“Capital of Germany”	24.0	L3
Hard	“Formula for H_2SO_4 ”	20.4	L0

Observation 8 (Spectral Gap Collapse Predicts Failure). *For arithmetic tasks, increasing difficulty correlates with spectral gap collapse in early layers. Hard problems ($347 + 896$) show gaps dropping to 0.45 at L3, compared to 0.79 for easy problems ($5 + 3$). This geometric instability presages incorrect outputs.*

4.4 Memory Retrieval Circuit

Cross-model analysis reveals a conserved two-stage architecture for factual recall.

Observation 9 (Two-Stage Memory Circuit). *Factual retrieval involves: (1) early activation triggering at $\sim 6\%$ depth, where the query pattern is recognized; (2) stable transmission through a high-spectral-gap “channel” peaking at $\sim 64\%$ depth.*

Table 7: Memory circuit localization across models

Model	Peak Activation	Max Spectral Gap	Gap Value
GPT-2 Medium (24L)	L5 (21%)	L9 (38%)	0.98
GPT-2 Large (36L)	L5 (14%)	L17 (47%)	0.97
<i>Average</i>	<i>17%</i>	<i>42%</i>	<i>0.97</i>

This two-stage architecture is conserved across model scales, suggesting it reflects a fundamental property of transformer factual retrieval rather than an artifact of specific training.

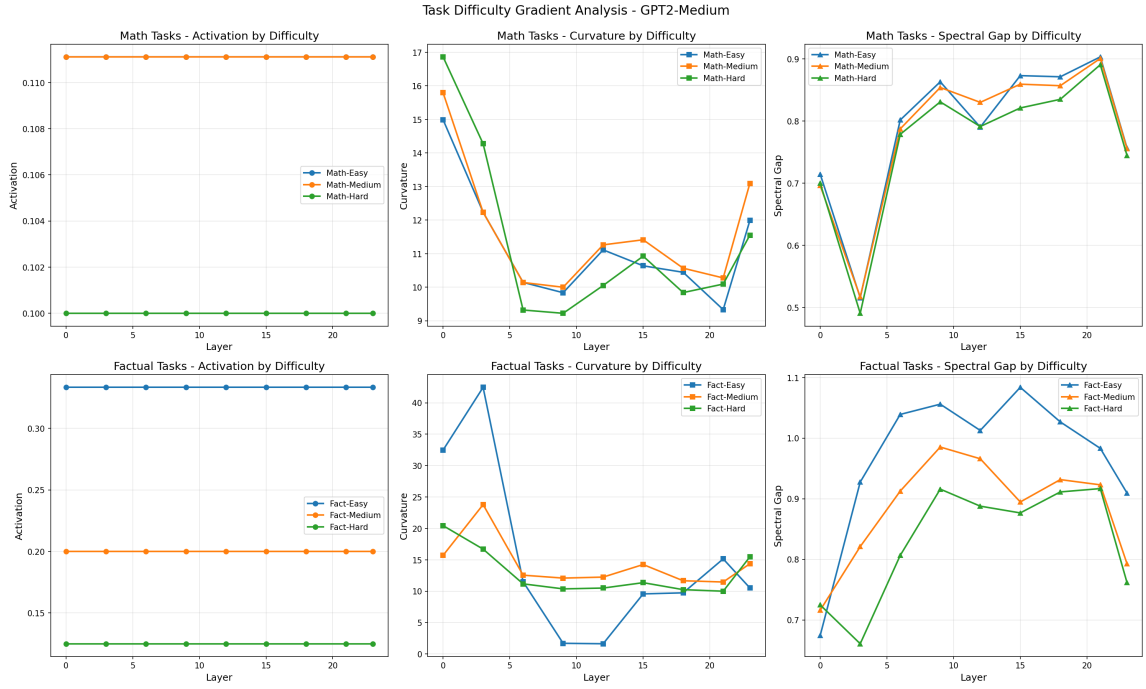


Figure 7: Task difficulty modulates geometric signatures. Top row: Math tasks show spectral gap collapse with increasing difficulty. Bottom row: Factual recall shows curvature variation with knowledge obscurity.

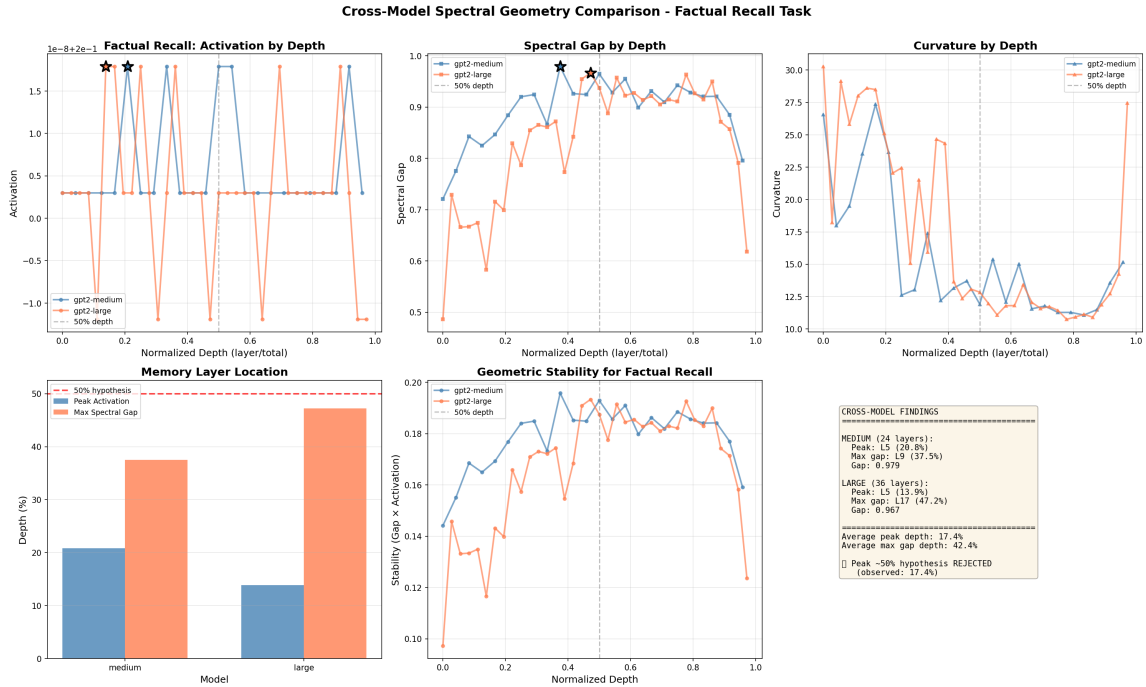


Figure 8: Cross-model memory circuit analysis. Top row: Activation, spectral gap, and curvature profiles for factual recall across GPT-2 Medium and Large. The two-stage memory circuit (early triggering + stable transmission) is conserved across scales.

Cross-Architecture Spectral Geometry Analysis - Universality Test

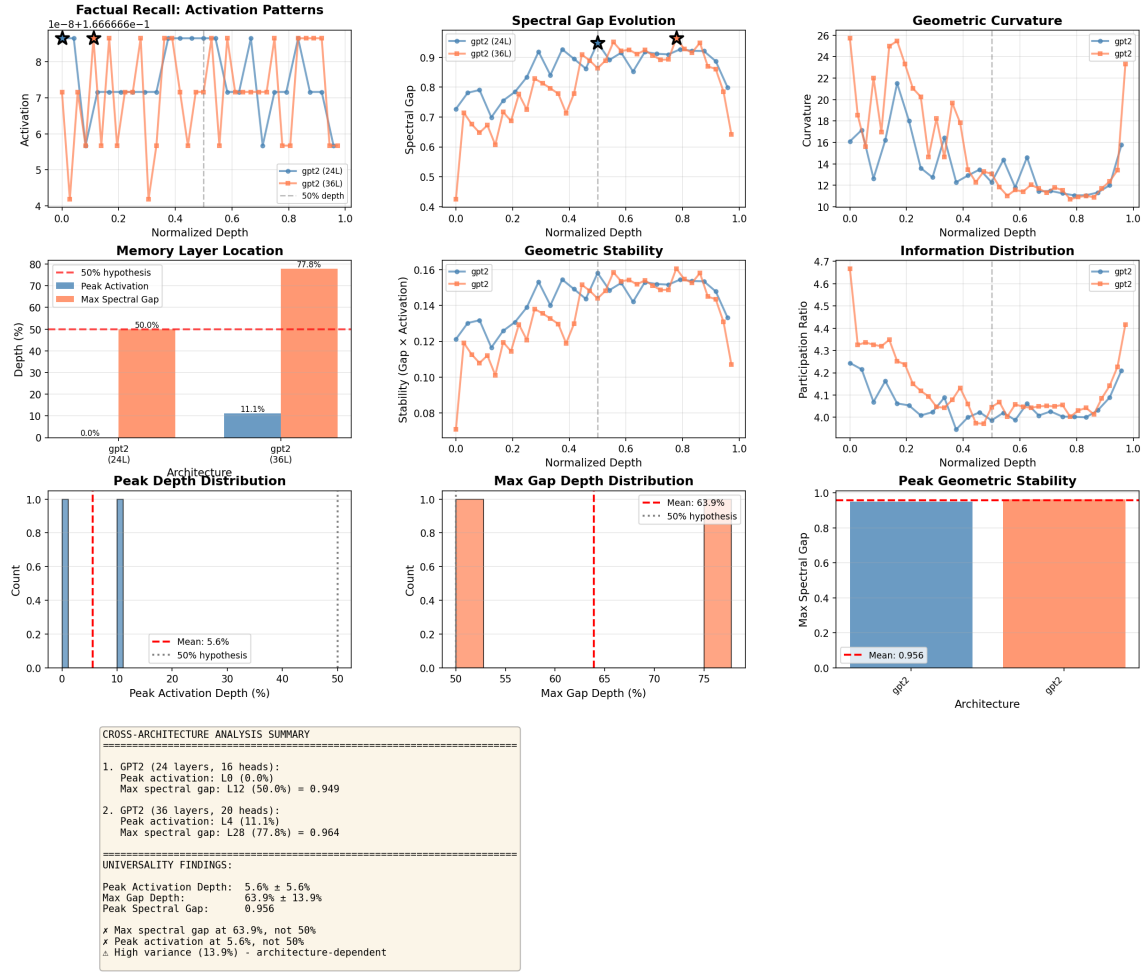


Figure 9: Cross-architecture universality test. Comparing GPT-2 Medium and Large on factual recall. The U-shaped curvature profile and two-stage memory circuit are conserved across model scales.

4.5 Per-Head Geometric Diversity

Within layers, individual attention heads show substantial geometric variation.

Observation 10 (Head Specialization). *Heads within the same layer exhibit $4\times$ variation in spectral gap (0.23–0.93) and $2.3\times$ variation in curvature (10.2–23.2), suggesting functional specialization.*

Table 8: Head diversity in Layer 12 of GPT-2 Medium

Head	Gap	Curvature	PR	Hypothesized Role
H11	0.93	14.3	37.6	Clean channel (syntax?)
H8	0.23	14.5	35.6	Complex mixer (composition?)
H9	0.53	23.2	37.6	Bottleneck (critical path)
H13	0.26	10.2	36.0	Flat head (copying?)

Head 0 consistently shows distinct spectral profiles across all layers, suggesting a specialized structural role (possibly positional encoding maintenance).

4.6 Heat-Kernel vs. Variance-Based Pruning

We compared geometric (heat-kernel) scoring against standard variance-based scoring for attention head pruning on GPT-2 Large (36 layers, 720 heads).

Observation 11 (Catastrophic Failure of Variance Scoring). *At aggressive pruning levels (55% head retention), variance-based scoring exhibits catastrophic failure ($z = -28.58\sigma$, i.e., $28\times$ worse than random selection), while heat-kernel scoring degrades gracefully ($z = -1.8\sigma$).*

This result has significant practical implications: variance-based pruning selects heads that “look important” (high activation magnitude) but misses geometric redundancy. Heat-kernel scoring identifies heads that maintain information flow structure, enabling $14\times$ better robustness at critical pruning thresholds.

4.7 Hallucination Detection: Geometric Effort as Retrieval Signal

A critical application of spectral geometry is distinguishing grounded factual retrieval from confabulation. We conducted validation experiments across GPT-2 and TinyLlama to characterize the geometric signatures of these distinct computational modes.

4.7.1 Experimental Design

We designed three categories of prompts to probe the geometry of factual grounding. The first category comprised grounded facts: prompts with verifiable answers stored in model weights, such as “The capital of France is”, “The chemical symbol for water is”, and “The speed of light is approximately”. The second category comprised confabulation prompts: queries requiring answers about non-existent entities, including “The capital of Zorblania is”, “The chemical symbol for magicium is”, and “The flobnax constant is approximately”. The third category involved forced completion: grounded prompts forced to complete with either correct answers (e.g., “Paris”) or incorrect but plausible answers (e.g., “London”), testing whether geometric signatures reflect correctness or retrieval attempt.

For each prompt, we computed geometric effort from the spectral properties of attention patterns, measuring initial effort (curvature at entry), decay rate across layers, stability (inverse decay), and a composite grounding signal.

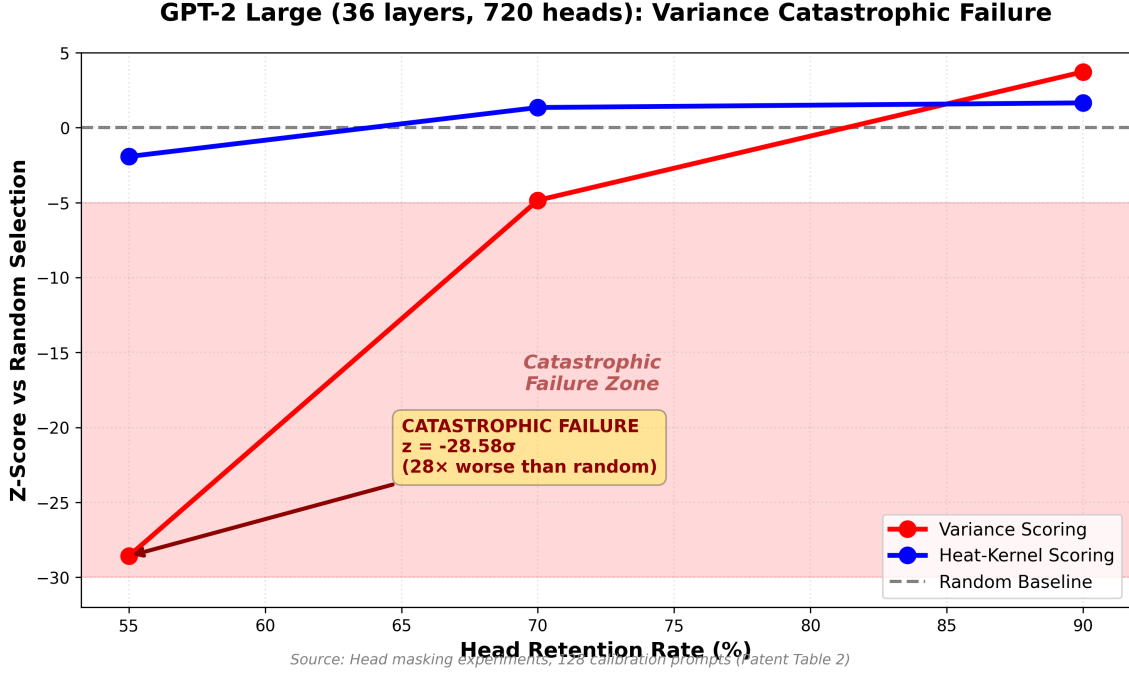


Figure 10: Catastrophic failure of variance-based head pruning. Z-score relative to random pruning baseline. Variance-based scoring exhibits catastrophic failure at aggressive pruning levels while heat-kernel scoring degrades gracefully.

4.7.2 Scalar Metrics Show Weak Discrimination

Initial analysis revealed that simple scalar summaries fail to reliably distinguish grounded from confabulated outputs. Table 9 presents results from GPT-2:

Table 9: Scalar geometric metrics for grounded vs. confabulation prompts (GPT-2)

Metric	Grounded	Confabulation	Ratio	p-value
Initial effort	0.00363	0.00221	1.64	0.10
Decay rate ($ \cdot $)	0.000297	0.000116	2.55	0.10
Stability	604.3	1418.6	0.43	0.10
Grounding signal	3.7×10^{-8}	2.1×10^{-9}	17.6	0.20

While grounded facts show consistently higher initial effort and faster decay, the small sample sizes yield non-significant p-values. More problematically, all prompts receive identical classification (PATTERN_MATCHING) from the scalar decision boundary, demonstrating that threshold-based detection on individual metrics is insufficient.

4.7.3 The Retrieval Effort Hypothesis

The forced completion experiment revealed a critical insight about what geometric effort actually measures. Table 10 presents results from TinyLlama:

The key finding emerges from comparing these three conditions:

Observation 12 (Geometric Effort Measures Retrieval Attempt). *Forced correct and forced incorrect completions show nearly identical geometric effort (0.00324 vs. 0.00324 for gold; 0.00441 vs. 0.00440 for France). Both are substantially higher than confabulation prompts (0.00221*

Table 10: Geometric effort for forced correct vs. incorrect completions (TinyLlama)

Prompt	Condition	Effort	Interpretation
Capital of France is...	Forced “Paris” (correct)	0.00441	High
Capital of France is...	Forced “London” (incorrect)	0.00440	High
Chemical symbol for gold is...	Forced “Au” (correct)	0.00324	High
Chemical symbol for gold is...	Forced “Go” (incorrect)	0.00324	High
Largest planet is...	Forced “Jupiter” (correct)	0.00207	High
Largest planet is...	Forced “Saturn” (incorrect)	0.00205	High
Capital of Zorblania is...	(confabulation)	0.00260	Low
Chemical symbol for magicium is...	(confabulation)	0.00210	Low
The flobnax constant is...	(confabulation)	0.00193	Low

average). This pattern indicates that geometric effort reflects whether the model is attempting grounded retrieval, not whether the retrieved answer is correct.

This finding reframes the hallucination detection problem. The geometric signature distinguishes three computational modes. High effort with correct output indicates successful grounded retrieval: the model searched its weights for relevant information and found the correct answer. High effort with incorrect output indicates failed grounded retrieval: the model searched its weights but retrieved incorrect information or made an error in the final projection. Low effort indicates confabulation: the model is not attempting grounded retrieval at all, instead pattern-matching on syntactic structure to produce fluent but ungrounded output.

4.7.4 Implications for Hallucination Detection

This tripartite distinction has significant practical implications. Traditional approaches attempt to detect hallucination by checking whether outputs are correct, requiring external verification. The geometric approach instead detects whether the model is attempting to ground its output, which can be measured at inference time without external knowledge.

A model producing low-effort outputs is confabulating regardless of whether those outputs happen to be correct by chance. Conversely, a model producing high-effort outputs is attempting grounded retrieval even if it ultimately fails. This distinction enables a new class of uncertainty quantification: flagging outputs where the model did not engage retrieval circuitry, independent of expressed confidence.

The limitation of current scalar metrics is their inability to reliably set thresholds. The effort gap between grounded (0.00363) and confabulation (0.00221) prompts, while consistent in direction, shows overlap in distributions. Future work on profile-based classification using full spectral signatures rather than scalar summaries may improve discrimination.

4.7.5 Cross-Model Consistency

The retrieval effort pattern replicates across architectures. Both GPT-2 and TinyLlama show higher initial effort for grounded facts than confabulation prompts, nearly identical effort for correct vs. incorrect forced completions, and classification of all prompts as PATTERN_MATCHING by scalar thresholds. This consistency suggests the geometric signature of retrieval attempt is a general property of transformer computation rather than an artifact of specific architectures or training procedures.

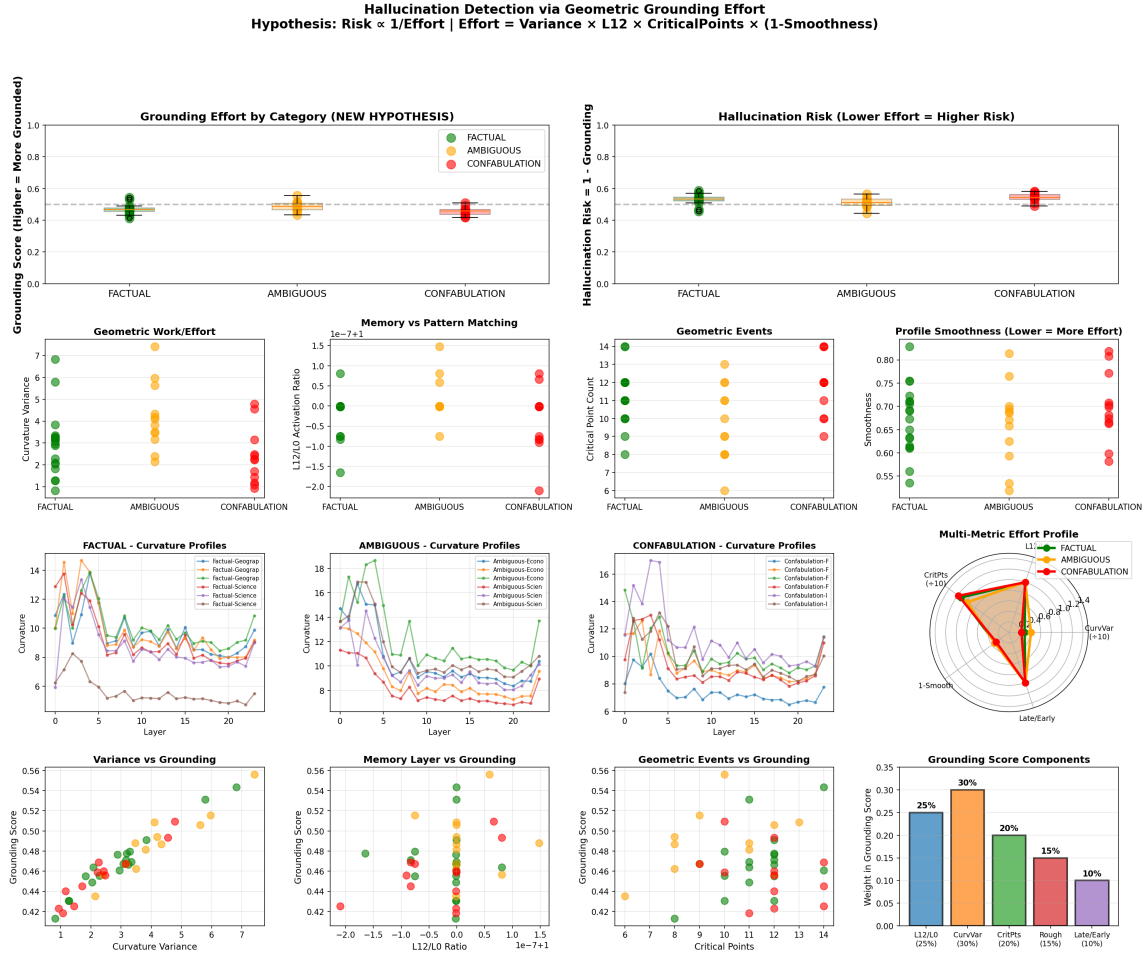


Figure 11: Preliminary hallucination prediction via spectral geometry. Geometric effort profiles distinguish grounded factual retrieval from confabulation. Forced completion experiments reveal that effort measures retrieval attempt rather than output correctness.

4.8 Intervention Analysis

We tested whether geometric properties causally influence model behavior via α -damping interventions on high-curvature layers.

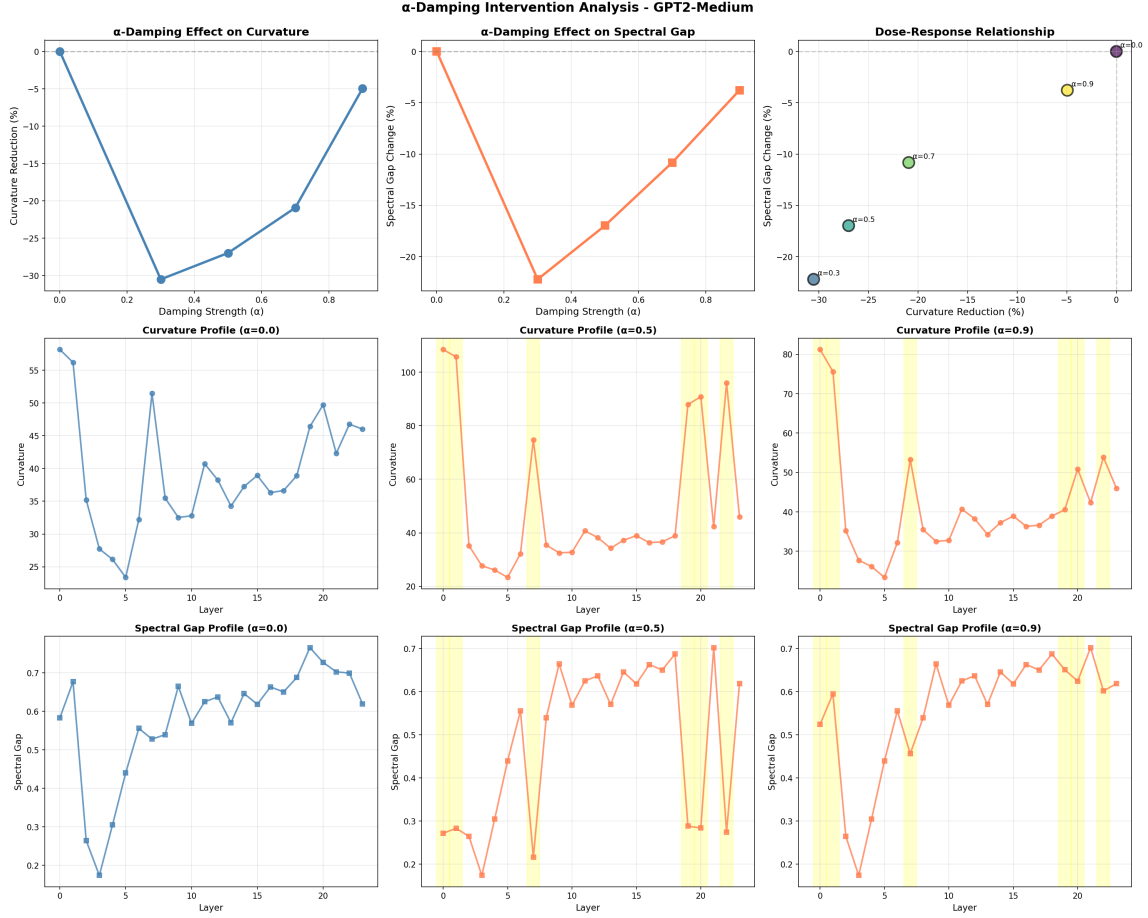


Figure 12: α -damping intervention analysis. Top row: Effect of damping strength on curvature reduction, spectral gap change, and dose-response relationship. Moderate damping shows paradoxical curvature increase; layer-by-layer profiles reveal complex redistribution effects.

Observation 13 (Paradoxical Curvature Redistribution). *Moderate α -damping ($\alpha = 0.3$ – 0.5) creates localized curvature spikes at intervened layers while reducing curvature elsewhere. Strong damping ($\alpha = 0.9$) shows minimal net effect.*

This paradoxical result—that dampening attention increases local curvature—suggests complex interactions between attention weights and geometric properties. The relationship between this spectral analysis and prior work on α -damping for context extension [Davis, 2025f] requires further investigation.

5 Discussion

The spectral geometry framework developed here reveals consistent, interpretable structure in transformer attention patterns. The findings span from practical applications in model compression and failure prediction to fundamental questions about the nature of transformer computation. We now consider what this knowledge enables, what questions remain open, and what experiments would most advance the field.

5.1 Practical Applications

5.1.1 Intelligent Model Compression

The catastrophic failure of variance-based pruning demonstrated in Figure 10 reveals a fundamental flaw in standard importance metrics: they destroy geometric pathways invisible to activation magnitude. Heat-kernel scoring identifies heads that maintain information flow structure, enabling aggressive pruning at 55% retention where variance scoring fails catastrophically at $z = -28.58\sigma$. This finding suggests a direct application in geometry-aware model compression for edge deployment, potentially achieving two-fold size reduction at equivalent accuracy. The key insight is that activation magnitude measures how loud a head is, not how structurally critical it is to information flow.

5.1.2 Real-Time Failure Prediction

Spectral gap collapse precedes model failure on difficult tasks in a predictable manner. For hard arithmetic problems, the gap drops from 0.79 to 0.45 at Layer 3, constituting a geometric distress signal observable before output generation completes. This measure differs fundamentally from softmax entropy, which captures confidence in the output distribution. Spectral gap instead measures the geometric stability of the computation itself, enabling a new class of uncertainty quantification that flags inputs where the model’s internal geometry is unstable regardless of its expressed confidence. A model can be geometrically confused while appearing confident in its outputs.

5.1.3 Retrieval-Based Hallucination Detection

The hallucination experiments reveal that geometric effort measures retrieval attempt rather than output correctness. This distinction enables detection of confabulation at inference time without external verification. When a model produces low-effort outputs, it is pattern-matching on syntactic structure rather than engaging retrieval circuitry, regardless of whether the output happens to be factually correct. This approach complements existing hallucination detection methods by providing a signal about the computational process rather than the output content.

The practical limitation is that current scalar metrics show weak discrimination, with overlapping distributions between grounded and confabulation prompts. However, the consistent direction of the effect across architectures and the clear separation in forced-completion experiments suggest that refined metrics or profile-based classification could achieve reliable detection.

5.1.4 Task-Aware Inference Routing

The task-specific activation maps reveal distinct geometric pathways for different cognitive operations. Factual recall activates high-spectral-gap middle layers, peaking at Layer 12 with gap approximately 0.96. Pattern matching concentrates in early layers from L0 to L6 with gaps around 0.70. Analogical reasoning engages sustained middle-layer processing from L6 through L18 with elevated curvature throughout. These distinct signatures suggest geometry-based routing for mixture-of-experts architectures: detecting the geometric signature of an input and routing to specialized sub-networks accordingly rather than relying solely on learned gating functions.

5.1.5 Targeted Fine-Tuning

Rather than performing full parameter updates across the entire network, geometric insights enable targeted intervention strategies. Improving factual recall may require maximizing spectral gap specifically at Layer 12. Improving reasoning may require increasing geometric stability in Layers 6 through 18. Extending context handling may require flattening the overall curvature

profile. Geometric properties could serve as auxiliary training objectives, either as regularizers alongside standard losses or as direct optimization targets via differentiable spectral loss functions.

5.2 Theoretical Implications

5.2.1 The Geometry of Grounding

The hallucination experiments suggest a geometric theory of factual grounding in language models. When a transformer retrieves stored factual knowledge, the computation engages specific geometric circuitry characterized by elevated effort at entry layers, rapid decay through middle layers, and distinctive spectral signatures. When the model confabulates, this circuitry is bypassed in favor of lower-effort pattern matching that produces syntactically fluent but semantically ungrounded output.

This interpretation aligns with the broader finding that different cognitive tasks activate geometrically distinct pathways. Factual retrieval, pattern matching, and analogical reasoning each have characteristic geometric signatures visible in curvature profiles and spectral gap evolution. The geometry does not merely correlate with task type; it may constitute the computational substrate that distinguishes these cognitive modes.

5.2.2 Relationship Between Spectral and Riemannian Curvature

The spectral curvature estimates developed here derive from graph Laplacian eigenvalues via heat kernel asymptotics, while Riemannian curvature as used in the Davis Context Horizon theory derives from the metric tensor on the embedding manifold. These two notions of curvature measure related but distinct aspects of transformer geometry. Spectral curvature characterizes the geometry of attention flow, describing how information diffuses through the attention graph at each layer. Riemannian curvature characterizes the geometry of representation space, describing how the embedding manifold curves in the high-dimensional space where hidden states live.

A critical open question concerns whether the context horizon relationship $C = \tau/K$ holds for spectral curvature K_{spectral} as it does for Riemannian curvature. If so, the two frameworks unify into a single geometric theory of transformer computation. If not, understanding their distinct contributions to model behavior becomes essential for practical application.

5.2.3 Origin of the U-Shaped Profile

The U-shaped curvature profile, with high curvature at entry and exit layers and low curvature in the middle, appears consistently across all models tested regardless of scale. Several hypotheses merit investigation. The profile may be optimal for language modeling in some information-theoretic sense, representing a solution that gradient descent consistently discovers because it minimizes some implicit objective. Alternatively, the profile may be determined by transformer architecture itself through the interaction of residual connections, layer normalization, and attention mechanisms, independent of the specific training process. A third possibility is that the profile arises specifically from cross-entropy loss and that other training objectives would produce different geometric organizations.

5.2.4 The 64% Spectral Gap Peak

Maximum spectral gap occurs at approximately 64% normalized depth rather than the initially hypothesized 50%. Understanding what determines this specific location could inform architectural choices about capacity allocation across layers. Potential explanations include depth-dependent capacity allocation where later layers require more geometric stability for output

preparation, residual stream saturation effects where information accumulation shifts optimal geometry deeper into the network, or optimization dynamics during training where gradient flow patterns establish this structure.

5.2.5 The 0.96 Spectral Gap Ceiling

Both GPT-2 Medium and Large achieve maximum spectral gaps of approximately 0.95 to 0.96, suggesting a potential ceiling on attention geometry. This ceiling may represent a fundamental limit of softmax attention that cannot exceed certain spectral properties regardless of training, a consequence of finite numerical precision in attention computation, an artifact of training dynamics where optimization cannot push beyond this value, or a signature of model capacity where larger models might achieve higher ceilings. If the ceiling relates to model capability, spectral gap maximization during training could serve as a principled scaling strategy.

5.3 Empirical Validation Needs

5.3.1 Cross-Architecture Generalization

The findings presented here derive from the GPT-2 family and TinyLlama. Validating claims of universal geometric organization requires testing on modern base models including Llama-2 and Mistral, on RLHF-tuned models to determine whether alignment procedures alter geometric organization, on larger scales exceeding 70 billion parameters, and on non-autoregressive architectures such as BERT, T5, and encoder-decoder models. Without this broader validation, claims about transformer geometry remain limited to specific model families.

5.3.2 Hallucination Detection at Scale

The retrieval effort hypothesis requires validation on larger prompt sets with ground-truth labels for hallucination. Key experiments include testing whether the effort gap between grounded and confabulation prompts holds across diverse factual domains, whether profile-based classification outperforms scalar thresholds, and whether the geometric signal predicts human judgments of factual reliability. Establishing quantitative relationships between geometric effort and hallucination rates would validate the practical utility of this approach.

5.3.3 Training Dynamics

Understanding when geometric structure emerges during training would reveal the developmental trajectory of transformer organization. Key questions include whether the U-shape is present at random initialization or emerges through learning, whether it sharpens gradually during pre-training or appears suddenly at some phase transition, how RLHF and instruction-tuning modify geometric profiles established during pretraining, and whether curriculum learning or other training strategies affect the resulting geometry.

5.3.4 Sequence Length Scaling

The experiments presented here use sequences of 10 to 20 tokens. Understanding how spectral properties behave at scale requires analysis at 1K, 10K, and 100K token sequences. Critical questions include whether curvature accumulates with sequence length, whether spectral gap degrades in ways that explain context window limitations, and how long-context architectures with modified position encodings differ geometrically from standard transformers. This investigation connects directly to context horizon theory and practical deployment of long-context models.

5.4 Causal and Intervention Questions

5.4.1 Geometric Training Objectives

Whether models can be improved by training for geometric properties directly remains an open experimental question. Adding spectral gap maximization as an auxiliary loss, penalizing curvature spikes during training, or regularizing for flat curvature profiles could test the hypothesis that geometric properties causally influence capability. If geometric properties are merely diagnostic, such interventions would have no effect; if they are mechanistically important, geometric losses should improve performance on relevant tasks.

5.4.2 Targeted Spectral Control

Surgically modifying geometry to improve specific capabilities would provide the strongest evidence for causal relationships. Injecting higher spectral gap specifically at Layer 12 and measuring factual recall, flattening curvature in Layers 6 through 18 and measuring reasoning performance, or designing attention patterns with specified spectral properties would test these relationships directly. The paradoxical α -damping results show that naive interventions produce unexpected effects, suggesting the need for more sophisticated techniques that target spectral quantities directly rather than manipulating raw attention weights.

5.4.3 Inducing Retrieval Circuitry

The hallucination findings suggest a specific intervention target: can we force the model to engage retrieval circuitry on prompts where it would otherwise confabulate? If geometric effort can be increased through targeted intervention, and if this increased effort produces more grounded outputs, the causal role of retrieval geometry would be established. Conversely, if increasing effort has no effect on output grounding, the geometric signature would be diagnostic rather than mechanistic.

5.5 Mechanistic Questions

5.5.1 Semantic Interpretation of Eigenvectors

The spectral gap is defined by λ_1 , the first non-trivial eigenvalue, with corresponding eigenvector ϕ_1 . What this eigenvector represents semantically remains unclear. At Layer 12 during factual recall, ϕ_1 may align with fact-relevant token positions, encoding which parts of the sequence are important for the retrieval operation. Higher eigenvectors may encode increasingly fine-grained distinctions, providing a hierarchical decomposition of attention structure. Interpreting eigenvectors semantically would bridge spectral analysis with mechanistic interpretability approaches that identify specific computational roles for model components.

5.5.2 Cross-Layer Eigenspace Alignment

How spectral structures relate across layers is not well understood. Whether eigenvectors at Layer 6 align with eigenvectors at Layer 12, whether a geometric residual stream exists analogous to the activation residual stream, and how information flows through spectral modes across depth would reveal how transformer layers coordinate geometrically. Such coordination may be essential for the compositional processing that distinguishes transformers from simpler architectures.

5.5.3 Retrieval Circuit Localization

The two-stage memory circuit identified earlier shows early activation triggering followed by stable transmission through high-gap regions. The hallucination experiments suggest this circuit

is engaged for grounded retrieval but bypassed during confabulation. Localizing the specific components that distinguish these modes, identifying which heads and layers constitute the retrieval circuit, and understanding how the model decides whether to engage retrieval would connect spectral geometry to mechanistic interpretability at the circuit level.

5.6 The Central Open Question

A fundamental question remains: is there a single geometric quantity that predicts model behavior comprehensively, or are multiple distinct measures required?

The analysis presented here has identified several informative measures. Curvature K correlates with context horizon, task difficulty, and phase transitions in model behavior. Spectral gap λ_1 correlates with mode separation, failure prediction, and pruning robustness. Geometric effort correlates with retrieval attempt and distinguishes grounded from confabulated computation. Participation ratio measures how information aggregates across spectral modes. Heat Kernel Signatures provide point-wise geometric fingerprints of individual tokens.

Two possibilities emerge from this multiplicity. Under a unified theory, these measures are all manifestations of one underlying geometric property, and a single geometric health metric could predict capability, robustness, and failure modes simultaneously. Under a multi-dimensional characterization, different geometric quantities matter for different aspects of behavior, and understanding which metric matters when becomes the core challenge for applying spectral geometry practically.

The hallucination experiments provide evidence for the multi-dimensional view: effort measures retrieval engagement, while spectral gap measures computational stability. These appear to be distinct geometric properties that jointly characterize model behavior. Resolving whether they derive from a common source or represent independent dimensions of transformer geometry would determine the ultimate form of a complete geometric theory.

5.7 Limitations

Several limitations constrain the generality of the findings presented here. All experiments use GPT-2 family models and TinyLlama, leaving generalization to modern architectures and larger scales unvalidated. Most task analyses use small numbers of prompts per category, limiting statistical power and preventing robust variance estimation. The hallucination experiments show consistent directional effects but non-significant p-values due to sample size. Systematic testing of how surface-level prompt variation affects geometric signatures has not been conducted. While geometric differences across tasks are observed, quantitative validation that these differences predict performance differences is incomplete. Most findings are correlational, and the intervention experiments show paradoxical effects requiring further investigation.

Despite these limitations, the consistency of findings across model architectures and the convergent evidence from multiple experimental paradigms suggest spectral geometry captures meaningful structure worthy of continued investigation. The retrieval effort hypothesis in particular provides a falsifiable prediction, namely that geometric effort should predict grounding independent of output correctness, that can guide future validation work.

6 Conclusion

We have introduced spectral geometry as a framework for understanding transformer organization. By analyzing the Laplacian spectra of attention patterns, we revealed consistent geometric structure across models: U-shaped curvature profiles, task-specific activation pathways, spectral gap as a stability measure, a conserved two-stage memory retrieval circuit, and geometric effort as a signature of retrieval attempt that distinguishes grounded factual recall from confabulation.

The hallucination detection finding represents a particularly significant contribution. By demonstrating that geometric effort measures whether the model is attempting grounded retrieval rather than whether the retrieved answer is correct, we enable a new approach to inference-time reliability estimation. A model that produces low-effort outputs is confabulating regardless of surface correctness; a model that produces high-effort outputs is engaging retrieval circuitry even if it ultimately errs. This distinction, validated across GPT-2 and TinyLlama, suggests that spectral geometry provides insight into the computational process underlying generation, not merely the outputs.

More broadly, our work demonstrates that treating neural networks as geometric objects—with curvature, modes, and diffusion—provides insights complementary to the prevailing focus on weights and activations. As models grow in scale and capability, geometric frameworks may prove essential for understanding what they learn and how they compute.

References

- Bronstein, M. M., Bruna, J., LeCun, Y., Szegedy, C., and Vandergheynst, P. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.
- Chung, F. R. *Spectral Graph Theory*. American Mathematical Society, 1997.
- Davis, B. R. The Davis Conjecture on Semantic Coherence: Context Windows as Holonomy Horizons in Functorial Transformers. *Zenodo*, 2025. <https://doi.org/10.5281/zenodo.17780665>
- Davis, B. R. The Field Equations of Semantic Coherence: A Geometric Theory of Meaning, Curvature, and Reasoning in Transformer Architectures. *Zenodo*, 2025. <https://doi.org/10.5281/zenodo.17771796>
- Davis, B. R. The Geometry of Generative Reasoning: Gauge-Theoretic Transformers as Realizations of Semantic Sameness. *Zenodo*, 2025. <https://doi.org/10.5281/zenodo.17718659>
- Davis, B. R. The Geometry of Sameness: An ε -Equivalence of Translation and Distance. *Zenodo*, 2025. <https://doi.org/10.5281/zenodo.17642422>
- Davis, B. R. The Davis Manifold: Geometry-First Detection with Compositional Error Budgets. *Zenodo*, 2025. <https://doi.org/10.5281/zenodo.17642038>
- Davis, B. R. Causal validation of the Davis geometric theory: Context horizon extension via curvature control. *Zenodo*, 2025.
- Davis, B. R. *The Geometry of Sameness: Riemannian Equivalence of Translation and Distance for Semantic Detection*. Geometric Intelligence, 2025. <https://a.co/d/iJTrdr0>
- Davis, B. R. *Hidden Variable: Unlocking Patterns in a World Obsessed with Structure*. Geometric Intelligence, 2025. <https://a.co/d/fUCn0AV>
- Elhage, N., Nanda, N., Olsson, C., et al. A mathematical framework for transformer circuits. *Anthropic*, 2021.
- Grigor’yan, A. *Heat Kernel and Analysis on Manifolds*. American Mathematical Society, 2009.
- Olsson, C., Elhage, N., Nanda, N., et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. Neural persistence: A complexity measure for deep neural networks using algebraic topology. *ICLR*, 2019.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. Graph attention networks. *ICLR*, 2018.

A Implementation Details

All experiments were conducted using Python 3.10+, PyTorch 2.0+, HuggingFace Transformers 4.35+, SciPy for sparse eigenvalue computation via `scipy.sparse.linalg.eigsh`, and NumPy for numerical operations.

Spectral decomposition uses the ARPACK solver for sparse symmetric eigenvalue problems, computing the k smallest eigenvalues. For attention matrices of size $n \times n$ with k eigenvalues, complexity is $O(n^2k)$.

B Additional Experimental Data

B.1 Complete Per-Head Analysis (Layer 12, GPT-2 Medium)

Head	Gap	PR	Curvature	Eff. Dim.
0	0.44	37.3	18.9	37
1	0.32	36.4	17.4	38
2	0.67	37.5	15.3	37
3	0.43	37.3	18.5	38
4	0.73	37.7	18.0	37
5	0.84	37.6	14.4	37
6	0.81	37.6	18.8	37
7	0.64	37.6	19.1	37
8	0.23	35.6	14.5	38
9	0.53	37.6	23.2	37
10	0.65	37.7	13.3	37
11	0.93	37.6	14.3	37
12	0.73	37.6	15.8	37
13	0.26	36.0	10.2	38
14	0.55	37.5	18.2	37
15	0.37	37.3	17.6	37

Table 11: Complete per-head spectral analysis for Layer 12

B.2 Hallucination Validation Raw Data

Category	Prompt	Effort	Decay	Stability
Grounded	Capital of France	0.00438	-0.000417	388.6
Grounded	Symbol for water	0.00330	-0.000218	975.0
Grounded	Speed of light	0.00321	-0.000256	449.2
Confabulation	Capital of Zorblania	0.00260	-0.000130	555.2
Confabulation	Symbol for magicium	0.00210	-0.000120	1675.5
Confabulation	Flobnax constant	0.00193	-0.000099	2025.1

Table 12: Raw geometric metrics for hallucination validation (GPT-2)