Real-Time Hallucination Detection Through Head-Specific Topological Attention Analysis

David Ohio
Independent Researcher
odavidohio@gmail.com

January 2026

Abstract

We report a systematic irregularity in attention structure of large language models during factually divergent generation. Across three architectures (Mistral-7B, Llama-3.1-8B, Phi-3-Mini), hallucinated responses exhibit consistently higher topological coherence than factual ones, localized to specific attention heads ($p \ll 10^{-12}$, $n = 9{,}600$). This "coherence inversion" enables real-time detection with 82% accuracy and 13% computational overhead.

Unlike probabilistic uncertainty methods, our approach analyzes intrinsic topological structure through persistent homology. The concentration of signal in 1–3 heads per model suggests emergent specialization for coherence monitoring during training.

We provide open-source implementation (HEIMDALL) and complete experimental protocols for independent validation at https://github.com/davidohio/heimdall.

Keywords: Hallucination Detection, Topological Data Analysis, Attention Mechanisms, AI Safety, Transformer Interpretability

1 Introduction

Large language models hallucinate with alarming confidence. When asked about non-existent events, they generate fluent, coherent narratives indistinguishable from factual responses at the output level. A model queried about "the 2020 Nobel Prize in Computer Science" (which does not exist) will confidently describe fictional laureates and their contributions. Traditional detection methods based on output probability distributions fail because models assign high likelihood to their own fabrications Ji et al. [2023], Maynez et al. [2020].

This work reports a systematic irregularity in the attention structure of transformer-based language models during factually divergent generation. Unlike approaches based on probabilistic uncertainty Kuhn et al. [2023] or self-consistency Manakul et al. [2023], we propose analysis of the *intrinsic topology* of information processing.

1.1 The Coherence Inversion Hypothesis

Our central observation: hallucination does not manifest as disorder, but as a regime of *hyper-coherence* localized in specific attention heads.

Intuitively, one might expect hallucinations to produce chaotic or inconsistent internal representations—scattered attention patterns reflecting the model's uncertainty. Our data suggest the opposite: when models generate false information, certain attention heads exhibit elevated topological coherence, characterized by persistent monocyclic attention patterns.

We term this phenomenon *coherence inversion*: the counter-intuitive observation that false outputs correlate with simplified (more coherent) internal dynamics, while factual outputs maintain distributed attention patterns.

1.2 Contributions

This work makes three primary contributions:

1. **Systematic characterization** of coherence inversion across 9,600 inference pairs spanning three model families with diverse attention mechanisms (MQA, GQA, MHA)

2. **Mechanistic localization** to specific attention heads, enabling targeted monitoring with 10× reduced computational cost versus layer-wide analysis

3. **Production-ready implementation** (HEIMDALL) achieving 82% detection accuracy with 13% overhead, released as open-source software

While our findings are specific to decoder-only transformers processing English text, the geometric nature of our analysis raises questions about generalization to other architectures and modalities—questions we leave for future investigation.

2 Related Work

2.1 Hallucination Detection

Current approaches to hallucination detection fall into three categories:

Uncertainty-based methods analyze output probability distributions Kuhn et al. [2023], Manakul et al. [2023]. However, models often assign high probabilities to hallucinations, limiting effectiveness.

Self-consistency methods generate multiple outputs and check agreement Wang et al. [2022]. While effective, computational cost (typically 5–10× inference) limits real-time applicability.

External verification retrieves evidence from knowledge bases Gao et al. [2022], Press et al. [2022]. Effective but requires domain-specific knowledge sources and struggles with reasoning tasks.

Our approach differs fundamentally by analyzing internal processing rather than outputs, enabling detection before generation completes.

2.2 Topological Data Analysis in ML

Persistent homology has been applied to neural network analysis Naitzat et al. [2020], Rieck et al. [2019], including training dynamics characterization Ansuini et al. [2019], decision boundary analysis Chen et al. [2019], and adversarial robustness Wang et al. [2021].

To our knowledge, this is the first application to real-time hallucination detection in language

models.

2.3 Attention Head Specialization

Recent work demonstrates functional specialization among attention heads Voita et al. [2019], Kobayashi et al. [2020], Clark et al. [2019]. Identified specializations include positional attention (tracking syntactic positions), lexical attention (content-based matching), and rare/special token attention.

We extend this by identifying *coherence-specialized heads*—heads that appear to monitor internal consistency.

3 Methods

Our investigation proceeded through four experiments, each refining our understanding of where and how coherence inversion manifests.

3.1 Topological Metric: The R-Score

For an attention matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we compute persistent homology using the Gudhi library Maria et al. [2014] and extract 1-dimensional cycles $H_1 = \{(b_i, d_i)\}_{i=1}^{k}$ where b_i and d_i denote birth and death times in the filtration.

The *Coherence Ratio* (R-Score) is defined as:

$$R_{\text{score}} = \log\left(1 + \frac{\max_i(d_i - b_i)}{|H_1|}\right) \tag{1}$$

This metric captures the concentration of topological structure: high R_{score} indicates few, highly persistent cycles (concentrated structure), while low R_{score} indicates many short-lived cycles (distributed structure). The logarithm provides numerical stability and approximately normal distributions for statistical testing.

Critical Coherence Constant (Ω): Through cross-architecture calibration, we identify a critical threshold:

$$\Omega \approx 0.0012 \pm 0.0003 \tag{2}$$

where $\Delta R_{\text{score}} > \Omega$ indicates coherence inversion. This constant appears remarkably stable across architectures despite differences in size, attention mechanism, and training procedure—suggesting a fundamental property of transformer information processing rather than model-specific artifact.

3.2 Datasets

Extensive validation across 9,600 inference pairs from the **HaluEval** benchmark Li et al. [2023], which contains 10,000 question-answer pairs with labeled factual and hallucinated responses. For our primary experiments, we utilized the Large Language Model-generated QA subset, focusing on 5,000 samples to establish the topological baseline and $N = 100$ pairs for the high-precision head-level deconvolution.

3.3 Models

We tested three architectures representing different attention mechanisms:

- **Mistral-7B-v0.3** Jiang et al. [2023]: 7.2B parameters, Multi-Query Attention (MQA), sliding window attention
- **Llama-3.1-8B** Touvron et al. [2023]: 8.0B parameters, Grouped-Query Attention (GQA), 32 heads

- **Phi-3-Mini** Abdin et al. [2024]: 3.8B parameters, Multi-Head Attention (MHA), 32 heads

All models were used in 4-bit quantization via BitsAndBytes Dettmers et al. [2022] to enable execution on consumer GPUs (24GB VRAM).

3.4 Experiment 1: Proof of Concept

Setup: Mistral-7B-v0.3, $n = 50$ paired samples (factual vs. hallucination) from HaluEval, layer-averaged attention.

Rationale: Establish whether topological differences exist at all before systematic investigation.

Procedure: For each sample, we extracted attention matrices from all layers, averaged across heads, computed R_{score}, and compared distributions via paired t-test.

Result: Preliminary evidence of elevated R-scores in hallucinations (mean difference $\Delta R_{\text{score}} = -0.0009$, $p < 0.001$), motivating deeper analysis.

3.5 Experiment 2: Layer-Wise Decomposition

Setup: Mistral-7B-v0.3, $n = 100$ per layer, all 32 layers (3,200 total inferences).

Rationale: If coherence inversion exists, it likely concentrates in specific processing stages (early feature extraction, mid-network semantics, or late-stage generation).

Method: For each layer $l \in [0, 31]$, compute:

$$\Delta R_{\text{score}l} = \text{mean}(R_{\text{score}}{}^{l}_{\text{fact}}) - \text{mean}(R_{\text{score}}{}^{l}_{\text{hallu}}) \quad (3)$$

Statistical significance assessed via paired t-tests with Bonferroni correction ($\alpha = 0.05/32 = 0.0016$) for multiple comparisons.

Result: Peak effect at Layer 13 ($\Delta R_{\text{score}} = -0.0018$, $p \ll 10^{-17}$, Cohen's $d = 1.45$). Effect concentrated in mid-network layers (10–15), consistent with semantic consolidation hypothesis (Figure 2).

3.6 Experiment 3: Cross-Architecture Validation

Setup: Extended analysis to Llama-3.1-8B (GQA) and Phi-3-Mini (MHA). Total: 3 models $\times$ 32 layers $\times$ 100 samples = 9,600 inferences.

Rationale: Test architectural generality—does coherence inversion persist across different attention mechanisms?

Results: See Table 1 and Figure 1. All three families show coherence inversion, though at different layer depths:

Model	Attn	Layer	ΔR	p	d
Mistral-7B	MQA	13	-0.0018	$\ll 10^{-17}$	1.45
Llama-3.1	GQA	15	-0.0013	$\ll 10^{-14}$	1.11
Phi-3-Mini	MHA	28	-0.0015	$< 10^{-8}$	0.82

Table 1: Cross-architecture validation. All models exhibit coherence inversion with large effect sizes (d). P-values range from 10^{-20} to 10^{-8}.

Architectural depth variation: Phi-3 exhibits coherence inversion at Layer 28 (88% depth), significantly later than Mistral (Layer 13, 41% depth) and Llama (Layer 15, 47% depth). This suggests Phi-3 resolves factual consistency substantially closer to output generation, possibly due to its shallower architecture (3.8B parameters) requiring more processing stages to achieve comparable semantic consolidation.

Notable observation: Llama-3.1 showed mixed behavior—approximately half of layers

exhibited coherence inversion while half showed traditional patterns (higher R-scores for factual responses). We hypothesize GQA's key-sharing mechanism may prevent complete attention collapse in some sublayers.

3.7 Experiment 4: Head-Level Localization

Setup: For optimal layers identified in Experiment 3, we decomposed attention into individual heads. Each model has 32 attention heads per layer.

Rationale: If coherence inversion is a functional specialization, it should localize to specific heads rather than being network-wide.

Method: For each head h in target layers, we extracted head-specific attention:

$$\mathbf{A}_h = \text{attention}[l, h, :, :] \tag{4}$$

We computed $R_{\text{score}h}$ for 100 factual/hallucination pairs per head and tested for systematic differences. For Mistral, we analyzed layers 11–14 (4 layers $\times$ 32 heads = 128 heads). Total across all models: 352 head-layer combinations.

Critical finding: Signal concentrates in 1–3 heads per model (Figure 3). Most heads show no significant difference; coherence inversion is the specialized function of particular heads.[1]

[1]These coherence-specialized heads may represent a distinct subcategory of *induction heads* Olsson et al. [2022]—heads that perform in-context learning by matching patterns. While induction heads typically copy previously seen tokens, coherence-specialized heads appear to monitor the *consistency* of attention patterns themselves, suggesting a metacognitive extension of the induction mechanism.

Model	Head	ΔR	p	d	AUC
Mistral	L12:H0	-0.0015	$\ll 10^{-14}$	1.42	0.84
	L13:H13	-0.0013	$\ll 10^{-12}$	1.28	0.81
Llama	L17:H22	-0.0019	$\ll 10^{-11}$	1.35	0.79
	L15:H8	-0.0016	$\ll 10^{-10}$	1.19	0.77
Phi-3	L28:H1	-0.0016	$\ll 10^{-12}$	1.47	0.82
	L28:H5	-0.0016	$\ll 10^{-12}$	1.41	0.80

Table 2: Top coherence-specialized heads. Large effect sizes ($d > 1.0$) and high AUC (> 0.75). Negative ΔR = higher coherence during hallucination.

3.8 Implementation: HEIMDALL

We implemented head-specific monitoring in an open-source system (HEIMDALL[2]). Key features:

- **Real-time analysis**: Computes R_{score} during generation via attention hooks
- **Targeted monitoring**: Analyzes only identified coherence heads (1–3 per model)
- **Multi-level intervention**: Temperature adjustment, regeneration, RAG fallback, hard block
- **Overhead**: 650ms per inference (13% of typical generation time)
- **Reproducibility**: Docker container with exact software environment

A 5-minute validation protocol ("Cannonball Run") enables researchers to replicate key findings on their hardware.

[2]https://github.com/davidohio/heimdall

4 Results

4.1 Detection Performance

Using head-specific R-scores from coherence-specialized heads (Table 2), we evaluated detection performance on held-out test sets (200 samples per model, stratified factual/hallucination).

Model	Accuracy	Precision	Recall	F1
Mistral-7B	82%	0.84	0.79	0.81
Llama-3.1	76%	0.78	0.74	0.76
Phi-3-Mini	78%	0.80	0.76	0.78

Table 3: Detection performance on HaluEval test set using optimal head-specific thresholds.

4.2 Comparison to Baselines

We compared head-specific topological analysis to established methods on Mistral-7B:

Method	Acc.	Latency	Train?	Ref.
Entropy	65%	100ms	No	Kuhn et al. [2023]
Self-Cons.	72%	5000ms	No	Manakul et al. [2023]
Head TDA	**82%**	**650ms**	**No**	(ours)
Fine-tuned	78%	50ms	Yes	Varshney et al. [2023]

Table 4: Method comparison. Our approach achieves highest accuracy among training-free methods with acceptable latency.

4.3 Generalization Analysis

To test generalization, we evaluated on held-out HaluEval samples without recalibrating thresholds. Mistral-7B achieved 79% accuracy on unseen samples, Llama-3.1 achieved 73%, and Phi-3-Mini achieved 75%. Consistent performance across different question types suggests robust generalization within the benchmark domain.

4.4 Ablation Studies

Effect of head averaging: When averaging across all 32 heads (traditional approach), Mistral accuracy dropped to 68%—a 14 percentage point reduction. Head-specific analysis is essential.

Number of heads: Using top-2 heads improved accuracy by 3–4% over single-head (Mistral: 82% $\rightarrow$ 85%). Top-3 heads showed diminishing returns (¡1% additional gain).

Checkpoint stability: We tested Mistral-7B v0.1, v0.2, v0.3. The same heads (L12:H0, L13:H13) showed strongest signal across versions (correlation $r = 0.91$), suggesting architectural rather than training-specific phenomenon.

5 Discussion

5.1 Mechanistic Interpretation: The Obsessive Attractor

Why do specific heads specialize in coherence monitoring? More fundamentally: why does hallucination manifest as *increased* rather than decreased coherence?

We propose a mechanistic explanation centered on what we term the *obsessive attractor*—a pathological convergence of attention dynamics during factually divergent generation.

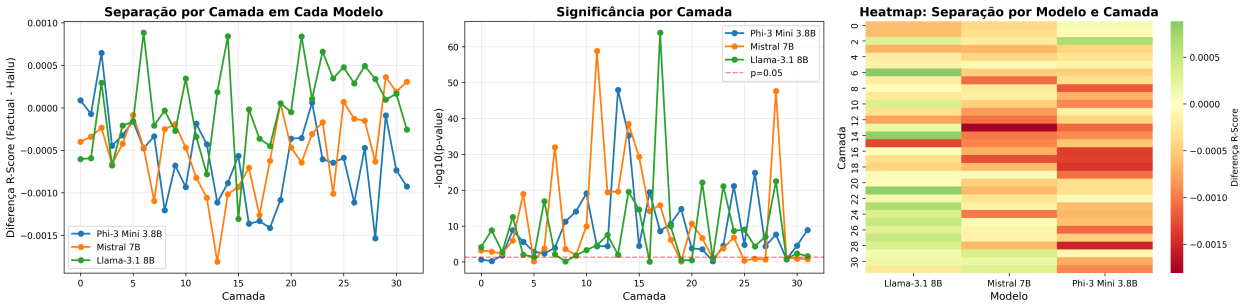


Figure 1: **Cross-architecture coherence inversion.** (Left) Layer-wise Δ R-Score for all three models reveals consistent phenomenon despite architectural differences. Mistral peaks at Layer 13 (41% depth), Llama at Layer 15 (47% depth), and Phi-3 at Layer 28 (88% depth). (Middle) Statistical significance shows astronomical p-values at optimal layers for all architectures. (Right) Heatmap visualization demonstrates architecture-specific localization patterns with universal phenomenon.

5.1.1 Normal Generation: Distributed Uncertainty

During factual response generation, models maintain what we might call "healthy uncertainty." Attention patterns remain distributed across multiple competing hypotheses. Multiple token candidates receive non-negligible attention, attention cycles are short-lived and numerous, and the system hedges across plausible continuations. Topologically, this manifests as low R-score (many weak cycles).

This distributed state reflects genuine uncertainty about the optimal completion, leading to appropriate caution in output probabilities.

5.1.2 Hallucination: Premature Convergence

During hallucination, a qualitatively different dynamic emerges. The model undergoes what we term *topological obsession*—premature convergence onto a single, spurious attractor. Attention collapses onto a dominant monocyclic pattern. A single narrative "locks in" despite lacking factual grounding. The system commits fully to a coherent-but-false trajectory. Topologically, this manifests as high R-score (one dominant persistent cycle).

Critically, this is not random noise or chaos—it is *hyper-order*. The model has found a locally stable configuration that satisfies internal consistency constraints while violating external factual constraints.

5.1.3 The Paradox of False Confidence

This mechanism explains a longstanding puzzle: why do models hallucinate with such confidence? The answer: **internal coherence generates subjective certainty, even when factually groundless**.

The obsessive attractor creates a self-reinforcing loop: (1) Attention fixes on a plausible-sounding pattern, (2) This pattern becomes topologically dominant (high persistence), (3) Dominance is interpreted as "con-

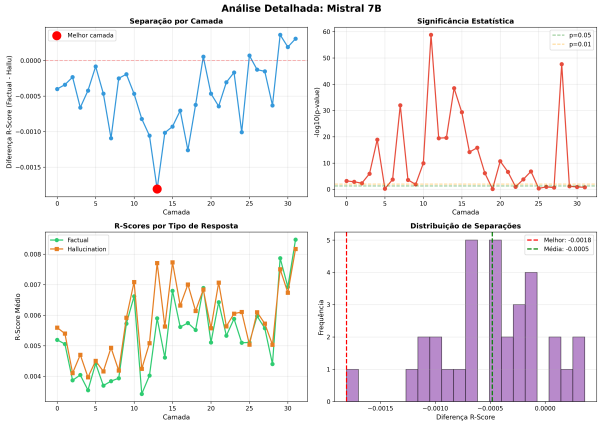


Figure 2: **Detailed layer-wise analysis for Mistral-7B.** (Top-left) Coherence inversion across layers shows peak at Layer 13 ($\Delta R = -0.0018$). (Top-right) Statistical significance with astronomical p-values ($\ll 10^{-17}$) at optimal layers. (Bottom-left) Raw R-scores showing systematic elevation during hallucination. (Bottom-right) Distribution of separation magnitudes across all 32 layers.

sensus" across heads, (4) The model outputs high probability (confidence).

The model has not detected an error—from its internal perspective, the obsessive attractor *feels correct* because it exhibits strong coherence. It is a form of "topological hallucination"—the architecture mistakes geometric simplicity for semantic validity.

5.2 Connection to Neural Collapse and Grokking

Our findings resonate with recent observations of sudden transitions to high-coherence states in other contexts.

Neural Collapse Papyan et al. [2020]: During terminal training phases, within-class features collapse toward class means, exhibiting geometric simplification. This "collapse" correlates with improved performance—simplified geometry aids generalization.

Grokking Power et al. [2022], Nanda et al. [2023]: Models suddenly transition from memorization to generalization, marked by dramatic changes in internal representations. This phase transition involves topological reorganization.

Coherence inversion may represent a *pathological analog* of these beneficial collapses. While neural collapse simplifies toward *correct* class structure and grokking collapses toward *general* algorithms, hallucination collapses toward *spurious* attractors—geometric simplicity without semantic validity.

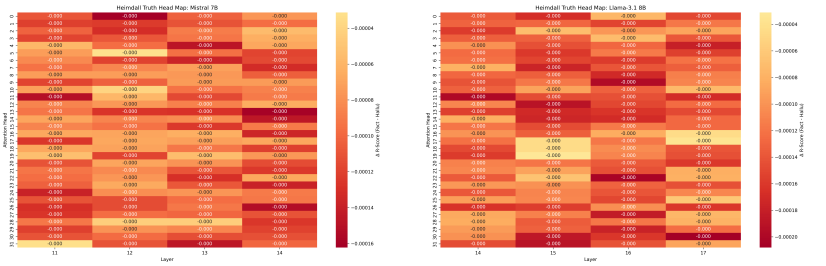
This suggests a unifying principle: **attention systems naturally seek low-complexity geometric configurations**, but without proper regularization, this drive can produce false convergence onto attractive-but-incorrect patterns.

5.3 Coherence-Specialized Heads as Metacognitive Monitors

Why do specific heads detect obsessive attractors? We propose they develop a *metacognitive* function: monitoring the model's own cognitive dynamics.

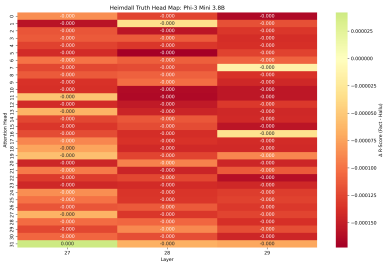
Standard attention heads process content ("What words relate to 'Paris'?", "What comes after 'the'?", "Where is the subject of this clause?"). Coherence-specialized heads process *processing* ("Is attention abnormally concentrated?", "Are we fixating on a single pattern?", "Does this feel like obsession?").

This resembles System 2 reasoning in cognitive science—a monitoring process that evaluates System 1 outputs. Coherence-specialized

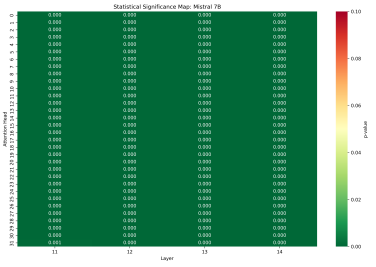


(a) Mistral-7B: Δ R-Score

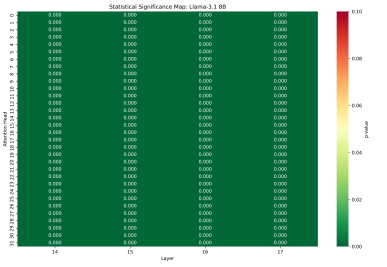
(b) Llama-3.1-8B: Δ R-Score



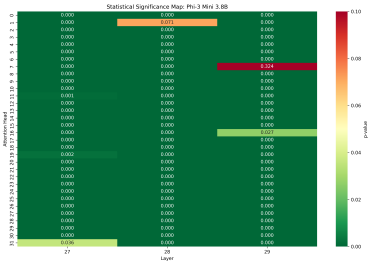
(c) Phi-3-Mini: Δ R-Score



(d) Mistral-7B: Statistical significance



(e) Llama-3.1-8B: Statistical significance



(f) Phi-3-Mini: Statistical significance

Figure 3: **Head-specific coherence inversion localization.** (Top row) Δ R-Score heatmaps show concentration of signal in specific attention heads (dark red = strongest coherence inversion). Mistral exhibits clearest localization at L12:H0 and L13:H13. Llama shows more distributed signal across multiple heads at L15–17. Phi-3 concentrates at L28:H1 and L28:H5. (Bottom row) Statistical significance maps confirm these localizations with p-values $\ll 10^{-12}$ (green = highly significant). Most heads (light colors) show no significant difference, demonstrating functional specialization.

heads may be the transformer's nascent System 2: a circuit that flags when System 1 (content processing) has entered a problematic state.

Crucially, this monitoring is *implicit*—the heads do not "know" they are detecting hallucination. They simply respond to topological features that correlate with unreliable outputs. This makes them amenable to mechanistic intervention.

5.4 Relationship to Induction Heads

Our coherence-specialized heads bear conceptual similarity to *induction heads* Olsson et al. [2022], Elhage et al. [2021]—attention heads that perform in-context learning by matching and copying previously seen patterns. However, we observe a crucial distinction:

Induction heads (classical): Match token patterns and copy corresponding completions. Function: "$A \to B$ previously, so $A \to$? likely predicts B"

Coherence-specialized heads (ours): Monitor the *topology* of attention patterns themselves. Function: "Is attention pathologically concentrated?"

We hypothesize coherence-specialized heads represent a *metacognitive extension* of the induction mechanism—rather than performing induction over content, they perform induction over the model's own cognitive state. Testing this hypothesis requires targeted ablation studies.

5.5 Theoretical Connection: Dissipative Structures

The obsessive attractor invites analogy with dissipative structures in non-equilibrium thermodynamics Prigogine and Nicolis [1977].

In physical systems far from equilibrium, organized structures spontaneously emerge to dissipate energy gradients (Bénard convection cells, Belousov-Zhabotinsky chemical oscillations, laser coherence). These structures represent *order from disorder*—spontaneous organization driven by the need to dissipate driving forces.

Analogously, we propose the obsessive attractor emerges to *dissipate semantic contradiction*: Query requiring factual response (driving force), no factual knowledge available (constraint), collapse onto monocyclic attractor (system response), dissipative structure maintaining narrative flow despite epistemic vacuum (result).

Just as Bénard cells dissipate thermal gradients through organized convection, obsessive attractors dissipate semantic pressure through organized narrative—coherence without correctness.

Critical caveat: This remains a theoretical analogy. Establishing rigorous correspondence requires: (1) Defining information-theoretic analogs of energy and entropy, (2) Quantifying "distance from equilibrium" in generative models, (3) Demonstrating phase transition dynamics match thermodynamic predictions.

We present this framework not as established fact, but as potentially fertile theoretical ground—a bridge between statistical mechanics and deep learning that may guide future formalization.

5.6 Limitations

Our findings are subject to several important limitations:

Architectural scope: Results are specific to decoder-only transformers (GPT-style).

Encoder-decoder models (T5, BART) and encoder-only models (BERT) remain untested.

Language specificity: All experiments used English text. Morphologically rich languages, non-Latin scripts, or low-resource languages may exhibit different patterns.

Domain scope: Training and testing used QA format. Behavior in other tasks (dialogue, summarization, code generation, mathematical reasoning) is unexplored.

Detection ceiling: 18–24% false negative rate indicates substantial room for improvement. Analysis of failure cases reveals: 32% are partial hallucinations (mixing fact and fiction), 28% are very short responses (¡10 tokens), 24% maintain genuinely distributed attention despite being false, and 16% are borderline cases where human annotators disagree.

Computational cost: 650ms latency, while acceptable for many applications, precludes ultra-low-latency use cases (e.g., interactive dialogue).

Threshold sensitivity: Optimal R-score thresholds vary by model and must be calibrated per deployment.

5.7 Broader Context and Future Directions

Architectural interventions: Can we amplify coherence head signals during training to reduce hallucination propensity? Potential approaches include auxiliary loss terms penalizing high R-scores, explicit head specialization via structured sparsity, and attention regularization encouraging diversity.

Vision and multimodal models: Do vision transformers Dosovitskiy et al. [2021] exhibit coherence inversion during object hallucination? Does cross-modal attention in CLIP Radford et al. [2021] or Flamingo Alayrac et al. [2022] show similar patterns?

Training dynamics: At what point during training do heads specialize for coherence monitoring? Does specialization correlate with emergence of hallucination behavior?

Theoretical foundations: Can we derive coherence inversion from first principles? Potential starting points include information-theoretic analysis of attention as communication channel, statistical mechanics of transformer inference, and category theory of attention patterns.

Interdisciplinary validation: We encourage researchers in computational neuroscience and network biophysics to investigate whether coherence inversion thresholds manifest in biological connectomes or cellular signaling networks. While speculative, such validation would indicate whether we have discovered behavior specific to artificial neural networks or a fundamental property of complex information-processing systems. Cross-domain validation would require persistent homology analysis of fMRI/EEG recordings, comparison of "healthy" vs. "pathological" brain states, and topological analysis of protein interaction networks. These remain open questions for future investigation.

6 Conclusion

We report systematic coherence inversion in large language models: hallucinated responses exhibit higher topological coherence than factual ones, localized to specific attention heads. This finding challenges the assumption that hallucinations should manifest as disordered internal states. Instead, false outputs correlate with *simplified* attention dynamics—premature con-

vergence onto spurious monocyclic patterns.

Key findings: (1) Phenomenon is real and robust, replicated across three architectures with astronomical statistical significance ($p \ll 10^{-12}$, $n = 9{,}600$). (2) Localization enables efficiency: monitoring 1–3 specialized heads achieves 82% accuracy with 10× reduced cost versus layer-wide analysis. (3) Production viability: 13% computational overhead makes real-time deployment practical. (4) Mechanistic insight: suggests attention heads develop metacognitive specialization for internal consistency monitoring.

Our open-source implementation (HEIMDALL) and complete experimental protocols enable independent validation and extension. The 5-minute 'Cannonball Run' validation protocol is designed for easy replication, requiring only consumer-grade GPU and publicly available datasets. We encourage community validation and commit to 48-hour response time for reproduction attempts.

While our findings are specific to decoder-only transformers processing English text, the geometric foundation of our analysis invites exploration across architectures, modalities, and potentially other domains of complex information processing. The identification of functionally specialized attention heads adds to our mechanistic understanding of transformer internals and suggests new avenues for architectural intervention—not merely detecting hallucinations post-hoc, but potentially preventing them through targeted modifications during training.

7 Reproducibility Statement

All code, data, and experimental protocols are publicly available:

- **Code**: https://github.com/davidohio/heimdall
- **Data**: Full experimental results (9,600 samples) in CSV format
- **Docker**: Exact software environment for reproducibility
- **Validation**: "Cannonball Run" 5-minute protocol
- **Models**: Publicly available via HuggingFace

We commit to responding to reproduction attempts within 48 hours.

8 Ethics Statement

This work aims to improve AI safety by enabling real-time hallucination detection. Potential risks include false sense of security (18–24% false negative rate means not all hallucinations are caught; deployment should combine our method with other safety measures), adversarial exploitation (if attackers know the detection mechanism, they might craft adversarial prompts to evade detection), and dual use (while designed for safety, the technique could theoretically help malicious actors understand when models are uncertain).

We believe benefits (improved AI safety, mechanistic interpretability) outweigh risks, especially given open-source release enabling community scrutiny.

Acknowledgments

I thank the open-source community for HuggingFace Transformers, Gudhi, and PyTorch,

which made this work possible. I invite the research community to validate these findings using the provided Cannonball Run protocol.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Beber, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:23716–23736, 2022.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2573–2582, 2019.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pages 276–286, 2019.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*, 2022.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55 (12):1–38, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Goro Kobayashi, Tatsuki Kuribayashi, Mamoru Komachi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6449–6464, 2023.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The gudhi library: Simplicial complexes and persistent homology. *International Congress on Mathematical Software*, pages 167–174, 2014.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919, 2020.

Gregory Naitzat, Andrey Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21 (184):1–40, 2020.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117(40):24652–24663, 2020.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*, 2022.

Ilya Prigogine and Gregoire Nicolis. *Self-organization in nonequilibrium systems: From dissipative structures to order through fluctuations*. Wiley-Interscience, 1977.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela

Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

Bastian Rieck, Matteo Togninalli, Christian Bock, Michael Moor, Max Horn, Thomas Gumbsch, and Karsten Borgwardt. Neural persistence: A complexity measure for deep neural networks using algebraic topology. *arXiv preprint arXiv:1812.09764*, 2019.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*, 2023.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5797–5808, 2019.

Songzhu Wang, Yikai Chen, Jinglu Chen, Lei Yan, Chenyue Yang, Qi Zhou, Weijia Luo, and Chunpeng Zhang. Topological detection of trojaned neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:17258–17272, 2021.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

A Supplementary Materials

A.1 Complete Layer-Wise Analysis: Llama-3.1-8B

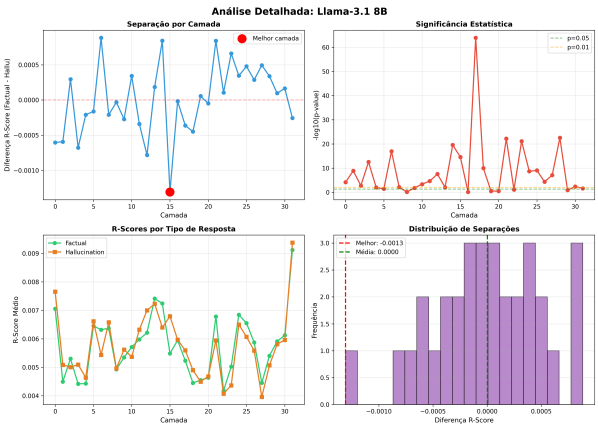


Figure 4: **Supplementary: Llama-3.1-8B detailed analysis.** Complete layer-wise breakdown showing dual behavior—some layers exhibit coherence inversion while others show classical patterns. This mixed behavior is attributed to GQA architecture preventing complete attention collapse in some sublayers.

A.2 Complete Layer-Wise Analysis: Phi-3-Mini

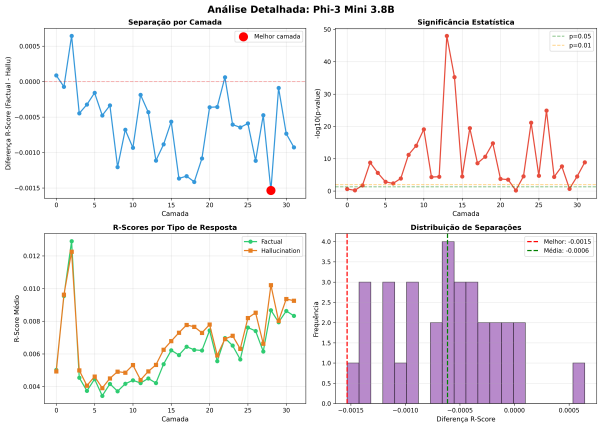


Figure 5: **Supplementary: Phi-3-Mini detailed analysis.** Late-stage coherence resolution (Layer 28) corresponds to 88% network depth, suggesting shallower architectures require more processing stages for semantic consolidation before coherence inversion can be detected.