Detecting Hallucination in LLMs: Tracing the Topological Signatures of Impaired Context Sharing

Anonymous ACL submission

Abstract

In this work, we examine the topology of information flow patterns within attention graphs to effectively distinguish hallucinated from non-hallucinated responses. We analyze the Forman–Ricci curvature to identify structural patterns indicating information bottlenecks in attention graphs. We then introduce a method that captures both semi-local and global information-flow characteristics of attention heads associated with hallucinated responses. We evaluate our approach extensively across several LLMs and established benchmarks. Empirical results demonstrate that our proposed single-pass approach significantly improves detection of hallucinated responses, achieving state-of-the-art performance on two hallucination-detection benchmarks. Further analysis reveals that impaired context sharing among tokens during causal generation is strongly associated with hallucination occurrences in LLMs. In particular, hallucinated responses are consistently characterized by an over-reliance on self-attention, diffused context retrieval from earlier tokens, or information over-squashing, especially in the final transformer layer.

1 Introduction

Amid the advances and extensive use of LLMs in various domains, new challenges regarding the factuality and correctness of LLM responses have emerged (Tao et al., 2024). While these models are increasingly integrated into broader autonomous workflows and downstream applications (Shinn et al., 2023; Yao et al., 2023), ensuring the factual precision of individual generated outputs remains a fundamental bottleneck for their reliability. A single unfaithful or factually incorrect response can compromise user trust and limit a model's practical utility. Misleading outputs generally fall into two categories: factual errors and faithfulness errors (Berglund et al., 2023). These errors may arise from reasoning failures (Berglund et al., 2023), defective inference (Chen et al., 2022), architectural limitations (Huang et al., 2025), training data biases, or inconsistencies in the provided context (Maynez et al., 2020). Consequently, isolating and detecting factual deviations within single-turn generations is critical for secure deployment.

Recent research indicates that hallucination, defined as the generation of factually incorrect or unsubstantiated responses, correlates with confusion and instability in the model during response generation. As a result, verifying the self-consistency of LLMs has become an important strategy for detecting hallucination (Farquhar et al., 2024; Wang et al., 2022). Although this methodology has demonstrated promising results, it possesses two significant limitations. Firstly, it is computationally expensive because it requires generating multiple responses to a single prompt (Chen et al., 2024). Secondly, this approach offers limited insight into the underlying causes and mechanisms behind these uncertainties.

Building upon literature that interprets attention weights as a weighted adjacency matrix (Barbero et al., 2024), we evaluate topological properties of the attention graph that can impede information flow. Consequently, we propose a set of topological features derived from attention weights that effectively distinguish hallucinated from factual responses. The main contributions of our work are as follows:

- We demonstrate that LLM hallucinations are characterized by information bottlenecks in the attention mechanism, which we identify using Ricci curvature on the attention graph.
- We propose a novel set of topological features from attention maps that enable a simple linear probe to achieve state-of-the-art performance in hallucination detection.
- We provide a comparative analysis across three LLM families, showing that the specific topological signatures of hallucination vary with model architecture, even though the core phenomenon of information bottlenecks is consistent.

2 Background

In this section, we establish the conceptual foundations for our topological analysis of Large Language Models (LLMs). We first define the formal mechanism of attention as a context-sharing process. Subsequently, we discuss how structural bottlenecks within attention graphs lead to information over-squashing. Finally, we introduce the Forman-Ricci curvature as a mathematical framework for quantifying and detecting these bottlenecks using graph topology.

2.1 Attention Formalism

In attention-based language models, the entire attention generation can be seen as a context-sharing process. Let $\{T_1,T_2,...,T_n\}$ be a sequence of tokens, $A=(a_1,a_2,...a_n)$ denote attention score vector, and $V=\{v_1,v_2,...v_n\}$ the matrix of value vectors carries the token contextual information. The context vector for the nth token is therefore $C_n=\sum_{j=1}^{n} a_j \cdot v_j$. This context vector determines how much information from previous tokens contributes to the representation of token n. Although attention weights reference previous tokens, the overall information flows forward through the sequence.

2.2 Information Bottlenecks and Over-squashing

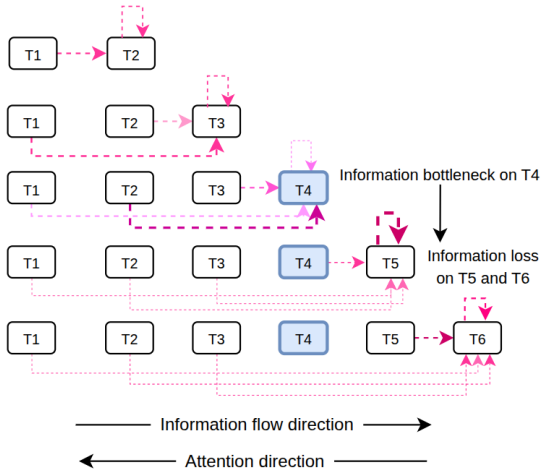


Figure 1: Information over-squashing occurs when information from many tokens is funneled through only a single token that bears the burden of representing all that context. However, this central token and its predecessors receive insufficient attention from subsequent tokens, preventing effective context sharing. In this example, all the information passes through token *T4*, but later, due to high self-attention on *T5*, a significant amount of information from earlier tokens is lost. The line thickness and its color show the amount of information flow from previous tokens to the subsequent ones.

A bottleneck happens when a significant part of the information from previous tokens passes through a single token. Such a narrow information flow leads to two main problems. First, concentrating the context of multiple tokens into a single token leads to excessive information compression, which can cause loss of contextual details. This effect is particularly pronounced in long sequences containing dozens and hundreds of tokens. Second, as illustrated in Figure 1, even if a single token carries comprehensive context from previous tokens, this information can subsequently be lost if attention becomes overly focused on a small subset of tokens further along the sequence, or if a token predominantly attends to itself.

An attention matrix can be seen as a weighted graph, with tokens as vertices and attention weights as edges. Information flows between tokens and attention layers via a message-passing mechanism. Distortion of the information flow due to bottlenecks limits the efficiency of message-passing. This phenomenon, called information over-squashing, is frequently associated with information bottlenecks in graphs and is measured by Forman-Ricci curvatures (Forman, 2003; Alon and Yahav, 2020; Topping et al., 2021).

2.3 Forman-Ricci Curvature

The presence of bottlenecks in attention graphs strongly co-occurs with context loss, disrupting the coherent semantic construction expected in transformer models. The Forman-Ricci Curvature (FRC) (Forman, 2003; Sreejith et al., 2016) provides a measure of information flow passing through an edge $e=(v,u)$. It consists of two components: the positive part, which measures the direct connections, and the penalty part, which measures the triangular alternative connection between two nodes. Therefore, **positive** expected value of FRCs of a graph indicates alternative information flux rather than passing through edge e, and **negative** values indicate information bottleneck, i.e., the information route between v and w is focused on the edge e. Since attention graphs are directed weighted graphs, we denote the attention weight from token u to token v by w_{uv}.

Let $G=(V,E)$ be the attention graph, with vertices v and w connected by an edge e. FRC is defined as:

$$\text{Forman-Ricci}(v,u)=w_{uv}\left(\frac{w_v}{w_{uv}}+\frac{w_u}{w_{uv}}-\sum_{\substack{y\sim N(v)\setminus u\\ z\sim N(u)\setminus v}}\left[\frac{w_v}{\sqrt{w_{uv}w_{vy}}}+\frac{w_u}{\sqrt{w_{uv}w_{uz}}}\right]\right) \quad (1)$$

where w_{uv} denotes the attention weight associated with the directed edge (u,v), and

$$w_u = \sum_{y \sim N(u)} w_{uy}$$

denotes the weighted out-degree of node u. The neighborhoods $N(u)$ and $N(v)$ denote the sets of vertices connected to u and v, respectively.

3 Methodology

3.1 Topological Indexes of Information Flow

To quantify the efficiency of context sharing within attention graphs, we define three specific topological and information-theoretic indices.

3.1.1 Mutual Outgoingness

In its simplified form, when the penalty term is omitted, the Forman–Ricci curvature reduces to the sum of the weighted degrees of the two connected nodes. Let

$$d(u) = \sum_{v \sim N(u)} w_{uv}$$

denote the weighted out-degree of token u. In attention graphs, $d(u)$ measures how broadly a token distributes attention across neighboring tokens.

We therefore define the mutual outgoingness score of an edge (u,v) as

$$\text{Outgoingness}(u,v) = d(u) + d(v).$$

Intuitively, low outgoingness values indicate highly self-focused attention patterns, whereas high outgoingness values indicate broader context sharing across tokens. The score can be computed efficiently in $\mathcal{O}(E)$.

3.1.2 Forman-Ricci Penalty

In contrast, the full Forman-Ricci curvature includes a penalty term that reflects edge redundancy and alternative paths, and is directly tied to local information flow. The penalty part of FRC of a weighted graph from (1) can be written in terms of probabilities:

$$\begin{aligned} P(u,v) = & \sum_{y \sim N(v) \setminus u} \frac{1}{\sqrt{p_v(u) p_v(y)}} \\ & + \sum_{z \sim N(u) \setminus v} \frac{1}{\sqrt{p_u(v) p_u(z)}} \end{aligned} \tag{2}$$

where $p_u(v) = \frac{w_{uv}}{w_u}$, $p_v(u) = \frac{w_{uv}}{w_v}$ and so on.

The function $\frac{1}{\sqrt{p_u(v) p_u(z)}}$ is convex over the probability simplex with respect to p_u. Because convexity is preserved under summation, the penalty term is minimized when the attention distribution is uniform (Bauschke and DiBerardino, 2022).

3.1.3 Joint Entropy as an FRC Penalty Proxy

From the chain rule of entropy, we can write:

$$H(U,V) = H(U) + H(V \mid U) \tag{3}$$

where

$$H(V \mid U) = -\sum_{u \in V} p(u) \sum_{v \sim N(u)} p(v \mid u) \log p(v \mid u) \tag{4}$$

assuming that $p_u(v) = \frac{w_{uv}}{w_u}$, and $p(u|v) = \frac{p(u,v)}{p(v)}$ and $p(u,v) = \frac{w_{uv}}{W_{\text{total}}}$, then

$$p(v \mid u) = \frac{p(u,v)}{p(u)} = \frac{\frac{w_{uv}}{W_{\text{total}}}}{\frac{w_u}{W_{\text{total}}}} = \frac{w_{uv}}{w_u} = p_u(v) \tag{5}$$

Therefore,

$$H(V \mid U) = -\sum_{u \in V} p(u) \sum_{v \sim N(u)} p_u(v) \log p_u(v) \tag{6}$$

From equations (2), (3), and (6), a uniform $P_u(v)$ maximizes the joint entropy while it minimizes the total penalty, indicating a possible reverse correlation between $H(u,v)$ and total penalty in Forman-Ricci. Both the entropy formulation and the Forman–Ricci penalty term are optimized under more uniformly distributed attention weights. This motivates the use of attention entropy as a computationally efficient proxy for curvature-related information dispersion and bottleneck behavior.

3.2 Extracting Features for Hallucination Detection

The exact computation of the full Forman-Ricci penalty incurs a worst-case complexity of $\mathcal{O}(V \cdot \bar{d}^2)$, where $\bar{d}$ is the average node degree, making it costly for long sequences or densely connected graphs. Given that attention graphs often span hundreds or thousands of tokens, using raw edge-level curvatures can be noisy and overly granular. More importantly, such local curvatures make the final set of features for hallucination detection sentence-length dependent.

To address this and create a set of meaningful, fixed-size features as the probe's input, we summarize the distribution of mutual outgoingness scores in each attention head by computing the 10th, 50th,

and 90th percentiles, capturing the distribution of local connectivity patterns within each attention head. Additionally, motivated by the inverse relationship between curvature penalties and joint Shannon entropy $H(U, V)$, and its linear time complexity of $\mathcal{O}(V)$, we include the joint entropy of each attention head as a cheap-to-calculate feature to characterize global information flow and identify systemic information bottlenecks associated with hallucinated generations.

4 Experiments

In this section, we describe the experimental setup, report the analytical results, and present the outcomes of hallucination detection and discuss the effectiveness of our approach.

4.1 Experimental Setup

We carried out all the experiments using three small-to-medium-sized models — Mistral 7B Instruct, LLaMA 3.1 8B Instruct, and Phi-4—and one larger model, Qwen3-32B. In particular, we included Qwen3-32B as a larger reasoning-oriented model to examine whether reasoning-specific mechanisms (e.g., explicit intermediate reasoning steps) influence characteristics such as outgoingness and attention entropy. For each model–benchmark pair, we generated responses using two sampling temperatures (0.1 and 1.0). We assessed the performance of our approach using two benchmarks commonly employed in hallucination studies: TruthfulQA (Lin et al., 2021) and NQ-Open (Lee et al., 2019). The TruthfulQA comprises 817 questions and ground-truth answers specifically designed to elicit imitative falsehoods and mimic human misconceptions. It contains 38 categories, including health, law, conspiracies, and fiction. The NQ-Open is a dataset of realistic questions collected from Google Search and consists of 3,610 questions also accompanied by ground-truth answers.

Both LLaMA 3.1 8B and Mistral 7B Instruct contain 32 layers, each with 32 attention heads, whereas Phi-4 has 40 layers with 40 heads per layer. Qwen3-32B with 64 layers and 64 heads per layer, has the largest architectural depth and width among the evaluated models. The experimental setup, including system prompt templates, attention implementation (Eager attention), hallucination probe, and dimensionality reduction settings, follows the exact configuration in (Binkowski et al., 2025), with the exception of label generation, where we used GPT-4 instead of GPT-4 mini due to its superior ability to capture nuanced linguistic patterns and provide significantly more accurate labels.

In all experiments, the Mann–Whitney U test was used to assess the statistical significance of differences in per-head features. We show the analytics results only for the 10th percentile since the 50th and 90th percentiles follow the same patterns.

4.2 Forman-Ricci curvature patterns

We first evaluate whether there are distinct information-flow patterns in Forman–Ricci curvature that differentiate hallucinated answers from the rest. To this end, the LLM responses for each of the NQ-Open questions are partitioned into hallucinated and non-hallucinated subsets. For each subset, we then calculate the mean curvature of each attention head in the last five transformer layers, as the last few layers are known to capture higher-level semantic representations (Zhao et al., 2024).

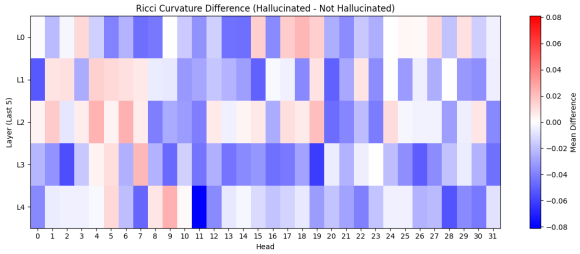


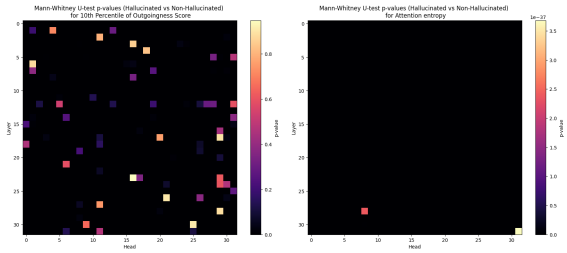
Figure 2: The difference between the mean values of Forman-Ricci curvatures per head for hallucinated and not hallucinated responses for LlaMA. For the majority of the attention heads, hallucinated responses have lower curvature values, indicating a less fluid information flow.

Figure 2 presents the average per-head differences in Forman–Ricci curvature between hallucinated and non-hallucinated responses for LLaMA. Negative values (shown in blue) indicate that hallucinated responses tend to exhibit lower curvature values, suggesting more obstructed information flow due to the higher penalty. Across all models, the last layer exhibits less fluid information flow in hallucinated responses. This effect is particularly pronounced in LLaMA and Mistral, where the distinction is more consistent across layers and heads. For Phi and Qwen, however, only the majority of heads in the last layer show attenuated information flow for hallucinated responses. The complete results and statistical significance values are presented in the Appendix, Figures 7 and 8.

4.3 Outgoingness, entropy and self-focus patterns

To investigate how outgoingness, distribution of attention across preceding tokens, and self-attention correlate with hallucinated responses, we compared per-head averages of the 10th, 50th, and 90th percentiles of outgoingness scores, attention entropy, and self-attention values across all layers.

As illustrated in Figure 3 (also in Appendix, Figure 9), LLaMA 3.1 8B Instruct shows the strongest separation of outgoingness and attention entropy among the other models. In turn, Phi-4 shows a statistically significant difference between the means of hallucinated and non-hallucinated responses for a subset of attention heads, while exhibiting almost no separation of attention entropies for most heads. Mistral 7B Instruct also exhibits meaningful differences among the majority of heads for both outgoingness and attention entropy, though the effect is less pronounced than in LLaMA 3.1 8B Instruct. Finally, Qwen3-32B has the fewest attention heads and shows statistically significant differences in outgoingness and attention entropy.

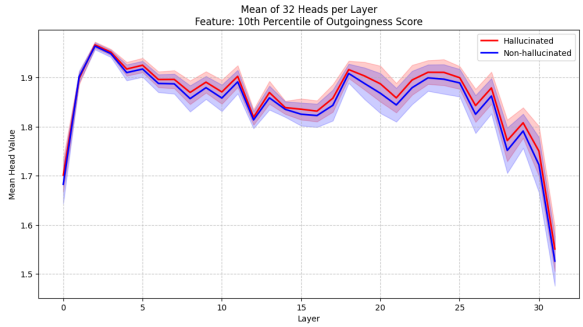


(a) Outgoingness - LLaMA 3.1 8B Instruct (b) Attention entropy - LLaMA 3.1 8B Instruct

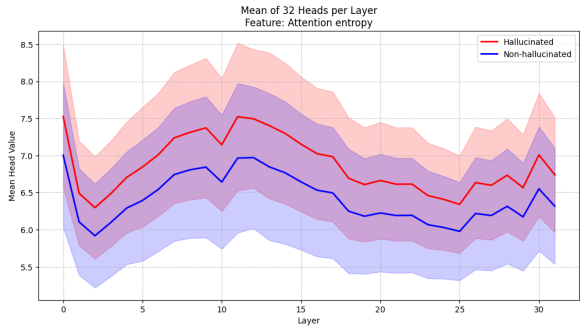
Figure 3: P-values for the per-head comparisons of the 10th percentile outgoingness scores and attention entropy across all layers for LLaMA 3.1 8B Instruct.

We also analyzed how each feature evolves across layers over time by averaging attention heads, and compared these patterns for hallucinated responses. We compute the mean by taking the arithmetic average of a given feature across all attention heads in a given layer. As shown in Figure 4 for LLaMA 3.1 8B (and Appendix Figure 10 for Mistral 7B), both models exhibit higher average outgoingness scores and more diffuse attention distributions in hallucinated responses, suggesting impaired context sharing. In contrast, both Phi and Qwen show minimal or no variation in outgoingness across layers and no noticeable change in attention entropy (see Appendix, Figure 10). Also, no clear pattern was observed that differentiates the earlier layers from the later ones in these behaviors.

Excessive self-attention is frequently associated



(a) Outgoingness - LLaMA 3.1 8B Instruct



(b) Attention entropy - LLaMA 3.1 8B Instruct

Figure 4: Feature evolution across layers and comparison between hallucinated and non- hallucinated response for LLaMA 3.1 8B Instruct.

with a loss of contextual information from surrounding tokens, tracking closely with a reduction in the accuracy of transformer outputs (Arif et al., 2025). Our experiments indicate that both Phi-4 and Qwen3-32B do not exhibit diffused attention, even when producing hallucinated content. To investigate this further, we analyzed the self-attention patterns across all layers and heads. We computed the average self-attention per head, then measured the difference in mean self-attention between hallucinated and non-hallucinated outputs. Figure 5 illustrates that for both models, in most attention heads, hallucinated responses correspond to higher self-attention—an effect that is especially evident in the later layers. Interestingly, this pattern is reversed and less statistically significant for the majority of heads in LLaMA and Mistral, where the hallucinated responses exhibit more diffuse attention (see Appendix, Figure 11).

4.4 Results

We compare the results with Laplacian eigenvalues (Binkowski et al., 2025) as our main baseline due to its state-of-the-art performance, single-pass approach, and a simple yet effective and fast approach for identifying hallucination.

We further compared our method with Eigen Score

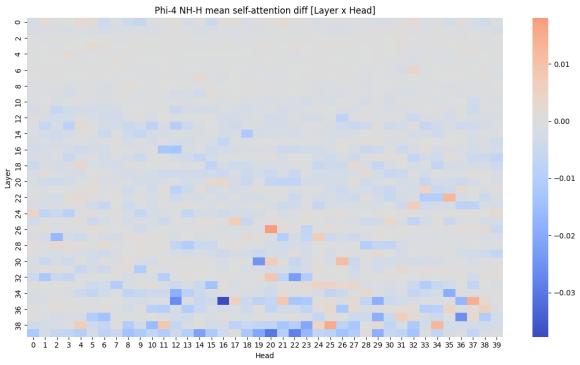


Figure 5: Mean per-head self-attention differences between hallucinated and non-hallucinated responses for Phi-4.

(Chen et al., 2024), which detects hallucinations by generating multiple responses to a query and measuring the variance of the resulting information as a proxy for model uncertainty. In the original formulation, the user prompt is transformed into multiple template-based variants that preserve intent while inducing response diversity. However, this setup assumes access to prompt-engineering mechanisms that are not typically available in real-world use, where users submit a single prompt. Although prompt variants could be generated using an auxiliary LLM, doing so introduces additional computational cost and latency. To better reflect practical deployment constraints, we did not use template-based prompt reformulation. Instead, we generated multiple responses to the original prompt using a sampling temperature of 0.5, thereby inducing controlled stochasticity without modifying the input query.

We use a linear probe in three main feature configurations to distinguish hallucinated from non-hallucinated responses: (a) using the 10th, 50th, and 90th percentiles of outgoingness scores along with attention entropy; (b) using all deciles (10th through 90th) of outgoingness scores and attention entropy to assess the effect of including finer-grained percentiles; and (c) extending configuration (a) by adding the mean per-head self-attention scores to isolate the contribution of self-attention features to the probe's performance. The best results were achieved using setting (c) (see Appendix, Table 3).

The experiments indicate that the semi-local outgoingness scores and global attention entropy features yield substantial gains in area under the curve (AUC) compared to the Laplacian eigenvalues. When fine-grained outgoingness features were added, no significant gains were observed. When the average self-attention per-head values were added to the feature set, this increased AUC for all Phi-4 and Qwen3-23B models.

Qwen3-32B achieved the weakest performance among the evaluated models. A plausible explanation is its explicit multi-step reasoning mechanism. Since attention entropy and outgoingness are proxies for uncertainty and dispersion of information flow, they may be less sensitive to structured reasoning processes. As a result, these features have limited discriminative power across most attention heads.

To assess the contribution of each feature type to hallucination detection, we evaluated the model's performance under a cumulative feature-addition scheme. As shown in Figure 6 and Table 2, progressively adding feature groups generally led to consistent improvements in AUC. While incorporating the 90th percentile of outgoingness improved performance in most settings, its contribution was occasionally unstable, leading to marginal decreases in certain configurations.

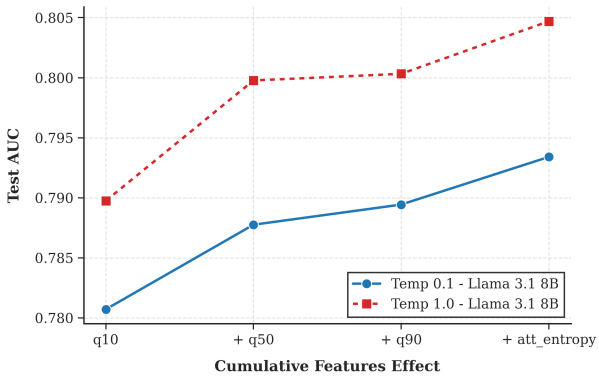


Figure 6: Cumulative feature set effect for Llama 3.1 8B Instruct accross NQOPEN dataset.

5 Related works

Hallucination detection methods generally fall into two categories. The first one supposes that hallucination is the result of the model's sequence-level uncertainty or semantic inconsistency during token generation, and therefore its main focus is to identify hallucinated responses through measuring the amount of intravariability of multiple generated responses (Farquhar et al., 2024), (Chen et al., 2024), (Ren et al., 2022). Such methods obtain multiple responses by applying top-p/top-k sampling strategy and then measure the semantic divergence among them. Although this approach has shown high detection rates, its performance depends on the number of generated samples and the variability of prompts with the same meaning. The more diversified the prompt

Table 1: AUC Comparison Between Laplacian eigenvalues, EigenScore, and our feature set classification results. Here we show the results for setting (a) which includes our base features. The best results however, were attained using setting (c).

Temp	Benchmark	Model	AUC		
			LapEigvals	EigenScore	Ours - setting (a)
0.1	TruthfulQA	LLaMA 3.1 8B Instruct	80.52	83.79	**85.34**
0.1	TruthfulQA	Phi 4	78.43	81.62	**81.75**
0.1	TruthfulQA	Mistral 7B Instruct v0.3	78.80	**81.69**	80.00
0.1	TruthfulQA	Qwen3-32B	77.36	72.17	**80.17**
1.0	TruthfulQA	LLaMA 3.1 8B Instruct	78.26	80.22	**80.73**
1.0	TruthfulQA	Phi 4	74.46	79.82	**84.00**
1.0	TruthfulQA	Mistral 7B Instruct v0.3	73.23	75.38	**78.00**
1.0	TruthfulQA	Qwen3-32B	**75.17**	69.35	73.53
0.1	NQOPEN	LLaMA 3.1 8B Instruct	75.67	71.49	**79.34**
0.1	NQOPEN	Phi 4	80.26	81.69	**82.00**
0.1	NQOPEN	Mistral 7B Instruct v0.3	77.10	79.45	**79.75**
0.1	NQOPEN	Qwen3-32B	76.16	71.73	**78.02**
1.0	NQOPEN	LLaMA 3.1 8B Instruct	76.19	71.16	**80.46**
1.0	NQOPEN	Phi 4	83.61	84.11	**84.49**
1.0	NQOPEN	Mistral 7B Instruct v0.3	76.38	**79.61**	77.29
1.0	NQOPEN	Qwen3-32B	78.49	73.63	**81.73**

Table 2: Cumulative Feature Ablation (AUC %) across TruthfulQA and NQOPEN datasets. Features are added sequentially.

Benchmark	Temp	Model	q10	+ q50	+ q90	+ att_entropy
TruthfulQA	0.1	LLaMA 3.1 8B Instruct	83.10	84.05	84.15	85.34
		Phi 4	79.30	80.20	80.10	81.75
		Mistral 7B Instruct v0.3	77.50	78.20	78.35	80.00
		Qwen3-32B	78.00	78.85	78.90	80.17
	1.0	LLaMA 3.1 8B Instruct	78.50	79.60	79.50	80.73
		Phi 4	82.10	82.70	82.85	84.00
		Mistral 7B Instruct v0.3	76.10	76.85	76.70	78.00
		Qwen3-32B	70.33	72.0	73.02	73.53
NQOPEN	0.1	LLaMA 3.1 8B Instruct	78.07	78.78	78.94	79.34
		Phi 4	79.52	80.45	80.28	82.00
		Mistral 7B Instruct v0.3	77.28	77.50	78.17	79.75
		Qwen3-32B	74.18	74.98	77.35	78.02
	1.0	LLaMA 3.1 8B Instruct	78.98	79.98	80.03	80.46
		Phi 4	82.88	82.99	83.02	84.49
		Mistral 7B Instruct v0.3	75.33	76.13	76.68	77.29
		Qwen3-32B	79.20	80.10	80.25	81.73

and the higher the number of responses, the higher the detection accuracy (Chen et al., 2024). This characteristic makes this approach costly, time-consuming, and therefore inadequate for real-time detection.

The second approach, which draws on more recent studies, has explored the use of specific attention-weight patterns to detect hallucinated responses in large language models (LLMs) (Binkowski et al., 2025) (Bazarova et al., 2025). While it has demonstrated promising results, the studies that have followed the use of attention weights for hallucination detection often rely on simplified assumptions and lack a more in-depth analysis of the complex patterns co-occurring with hallucinations.

Bazarova et al. (Bazarova et al., 2025) posit that hallucinated responses introduce a unique and measurable topological dissimilarity in the attention graphs connecting the response to the prompt. However, this

framework may oversimplify the nature of attentional failures. It does not fully account for scenarios such as information bottlenecks, strong attentional links between response tokens and irrelevant prompt tokens, or the diverse patterns that may emerge across different domains and LLM architectures.

Similarly, Binkowski et al. (Binkowski et al., 2025) hypothesize that hallucinations are a direct consequence of information bottlenecks and propose using the eigenvalues of the Laplacian matrix of attention graphs to detect them. While innovative, Laplacian eigenvalues reflect a range of graph characteristics beyond bottlenecks, including graph bipartiteness and the number of spanning trees, which may or may not be related to hallucination. Despite their hypothesis, the authors did not conduct empirical experiments to show how Laplacian eigenvalues vary depending on whether a response is hallucinated. Furthermore, this method provides a global perspective on graph properties, potentially obscuring localized attentional features that could be more directly associated with hallucinated responses. Moreover, the authors show that the higher the number of eigenvalues, the better the distinguishability of their feature set and report that their best results were achieved when 100 eigenvalues were extracted from each attention head. However, this puts a hard constraint on the minimum response length, making the method less effective for short responses.

In the current study, we carried out a series of analyses to examine the relationship between attention graph bottlenecks and hallucinations, mapping out how these patterns manifest across different LLM architectures. Afterward, we propose a set of features that enable high-performance hallucination detection, and explain why and how hallucination occurs across different LLMs.

6 Limitations

Theoretical and Operational Confounders. Our topological framework cannot distinctly separate pathological context-sharing impairments from either absolute parametric knowledge gaps or functional attention sinks (*e.g.*, initial tokens or punctuation). Because absolute data omissions and architectural attention anchors both induce localized distortions in attention distributions, the observed signatures risk confounding true semantic bottlenecks with benign structural properties or generic model uncertainty.

Architectural and Evaluation Scope. The proposed method exhibits explicit architectural dependencies, resulting in significant performance drops on reasoning-dense models like Qwen3-32B. Furthermore, our empirical validation is restricted to single-turn, English short-form question-answering benchmarks. The generalizability of these topological indices to non-English syntax, long-form multi-turn generation, or complex agentic reasoning pathways remains unverified.

Lack of Mechanistic Intervention. The findings presented in this work are strictly observational and correlational. We establish that certain attention topologies co-occur with factual errors, but we do not provide mechanistic validation. A robust causal proof requires test-time interventions—such as actively smoothing or modifying attention weights during inference—to demonstrate that altering these graph bottlenecks directly mitigates or scales down actual hallucination rates.

7 Conclusion

In this study, we conducted a comprehensive set of analyses and provided evidence that hallucinations manifest differently across various large language models (LLMs). We demonstrate that simplified attention-based properties alone are insufficient to characterize these phenomena. To address this shortcoming, we introduced a set of response-length-independent features to capture a broader range of hallucination patterns. Through a detailed examination and extensive experimentation, we observed that while attention behaviors during hallucination vary across different LLMs, impaired context sharing is a consistent topological signature correlated with hallucinations across all models examined.

Subsequently, we trained a linear probe using the proposed features and achieved state-of-the-art performance in hallucination detection. Our selection of diverse LLMs allowed us to analyze both shared and distinct behavioral patterns. However, the impact of model size, specifically the number of parameters, was not investigated in this work. For instance, although the Phi-4 model, which has approximately twice as many parameters as Llama 3.1 8B Instruct and Mistral 7B Instruct, exhibits notably different behavior, it remains unclear whether these differences are attributable to model size or other architectural factors. Therefore, the interplay between model size and attention behavior during hallucination can be investigated further in future work. Also, our results indicate that producing correct responses requires tokens to balance reliance on their own semantic representations with the contextual information

from other tokens. Hallucinations can therefore be interpreted as a failure to correctly integrate contextual information from the prompt. Still, it remains an open question how internal model dynamics differ when hallucination is caused by genuine knowledge gaps rather than misaligned context sharing.

References

Uri Alon and Eran Yahav. 2020. On the bottleneck of graph neural networks and its practical implications. *arXiv preprint arXiv:2006.05205*.

Kazi Hasan Ibn Arif, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, and Chris Thomas. 2025. Paint: Paying attention to informed tokens to mitigate hallucination in large vision-language model. *arXiv preprint arXiv:2501.12206*.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João Madeira Araújo, Oleksandr Vitvitskyi, Razvan Pascanu, and Petar Veličković. 2024. Transformers need glasses! information over-squashing in language tasks. *Advances in Neural Information Processing Systems*, 37:98111–98142.

Heinz H Bauschke and Peter AV DiBerardino. 2022. Minimal angle spread in the probability simplex with respect to the uniform distribution. *Journal of Nonsmooth Analysis and Optimization*, 3(Original research articles).

Alexandra Bazarova, Aleksandr Yugay, Andrey Shulga, Alina Ermilova, Andrei Volodichev, Konstantin Polev, Julia Belikova, Rauf Parchiev, Dmitry Simakov, Maxim Savchenko, and 1 others. 2025. Hallucination detection in llms with topological divergence on attention graphs. *arXiv preprint arXiv:2504.10063*.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Kajdanowicz. 2025. Hallucination detection in llms using spectral features of attention maps. *arXiv preprint arXiv:2502.17598*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. *arXiv preprint arXiv:2402.03744*.

Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. 2022. Towards improving faithfulness in abstractive summarization. *Advances in Neural Information Processing Systems*, 35:24516–24528.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Forman. 2003. Bochner's method for cell complexes and combinatorial ricci curvature. *Discrete & Computational Geometry*, 29(3):323–374.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. *arXiv preprint arXiv:2209.15558*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Remanan Pushpa Sreejith, Karthikeyan Mohanraj, Jürgen Jost, Emil Saucan, and Areejit Samal. 2016. Forman curvature for complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(6):063206.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. *arXiv preprint arXiv:2404.17287*.

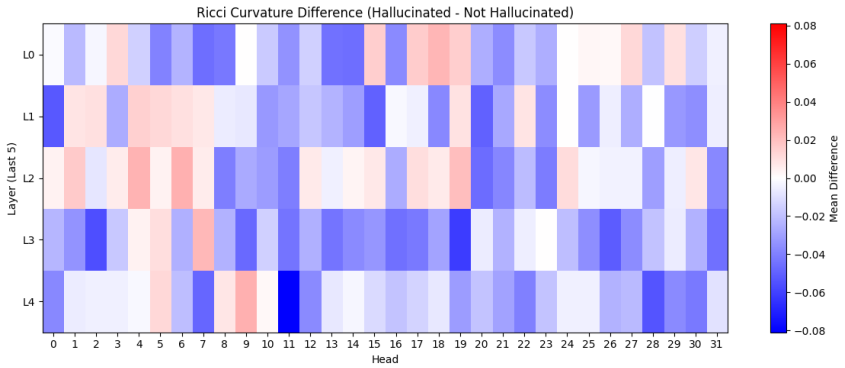
Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. 2021. Understanding over-squashing and bottlenecks on graphs via curvature. *arXiv preprint arXiv:2111.14522*.

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

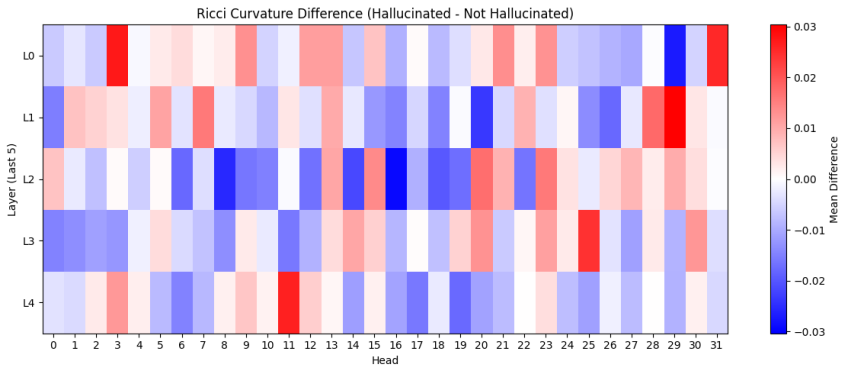
Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

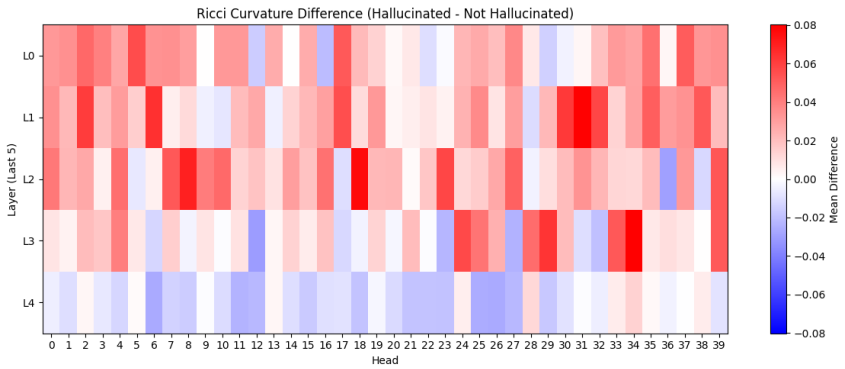
A Additional Experiments



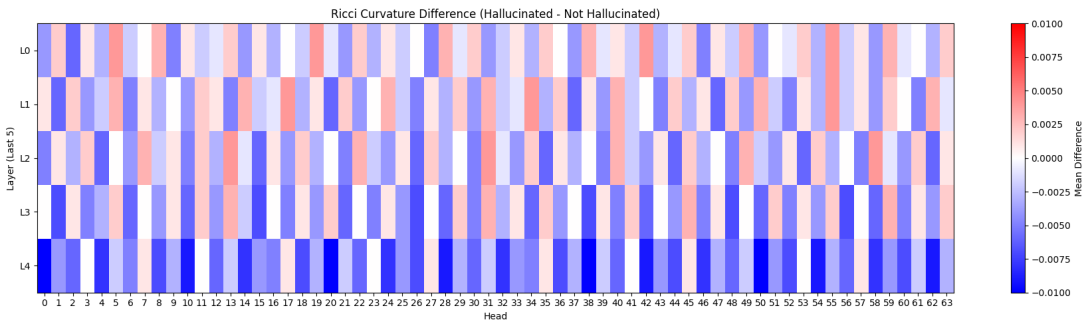
(a) LLaMA 3.1 8B Instruct



(b) Mistral 7B Instruct



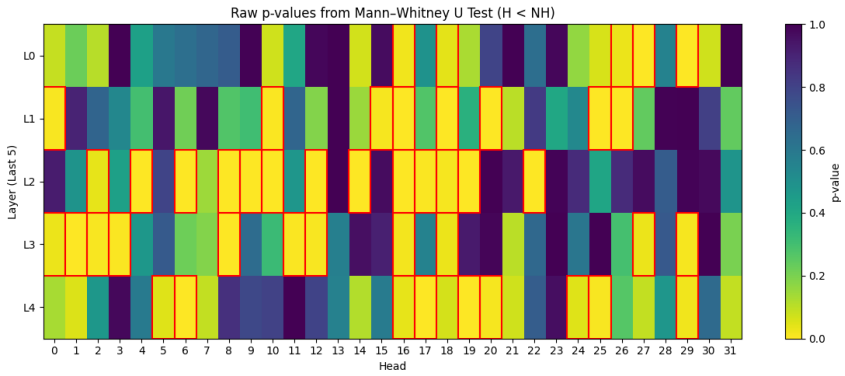
(c) Phi-4



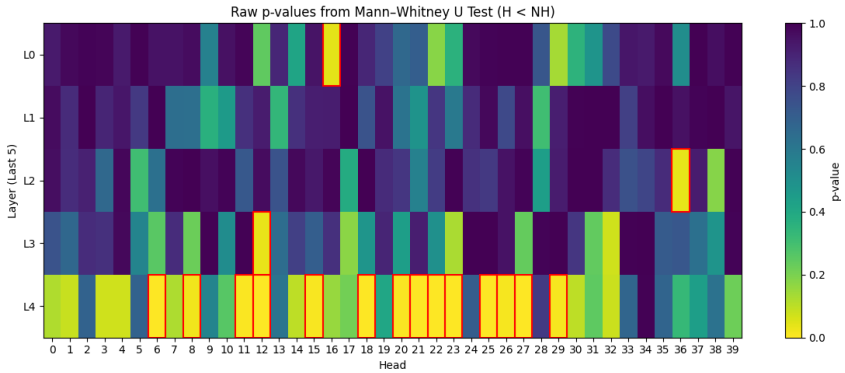
(d) Qwen3-32B

Figure 7: The difference between the mean values of Forman-Ricci curvatures per head for hallucinated and not hallucinated responses. For the majority of the attention heads, hallucinated responses have a lower curvature values which is an indication of a less fluid information flow. This pattern is more evident accross all layers in LLaMA and Mistral. For Phi and Qwen, the last layer presents less fluid context sharing for hallucinated responses.

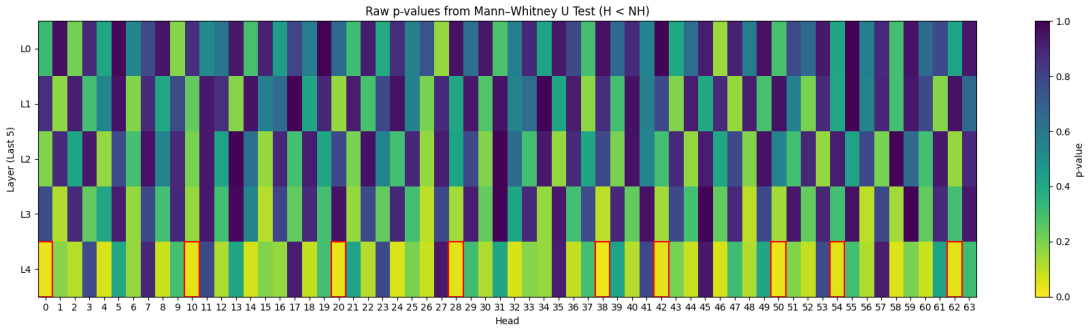
(a) LLaMA 3.1 8B Instruct



(b) Mistral 7B Instruct

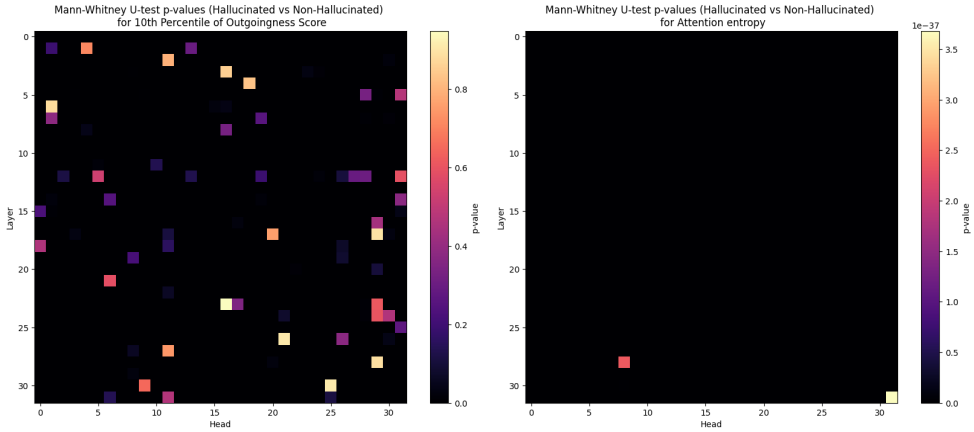


(c) Phi-4



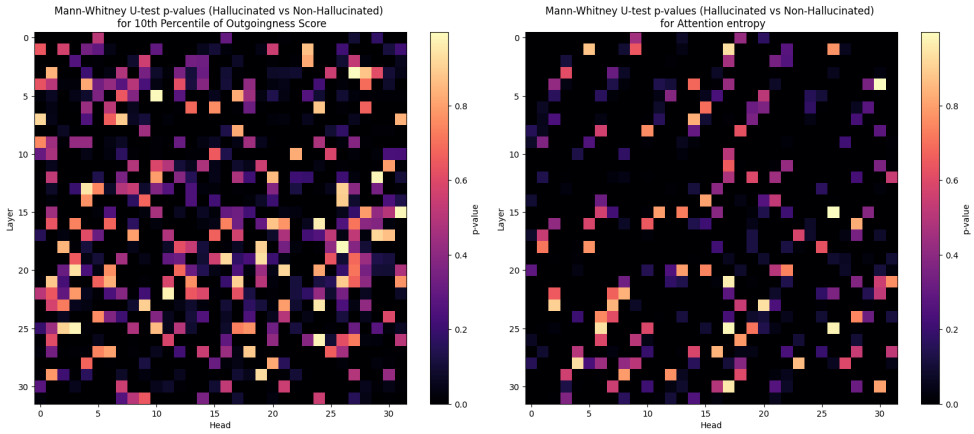
(d) Qwen3-32B

Figure 8: The p-values of the differences between the mean values of Forman-Ricci curvatures per head for hallucinated and not hallucinated responses. The heads with statistical signiuficance are highlited with red lines.



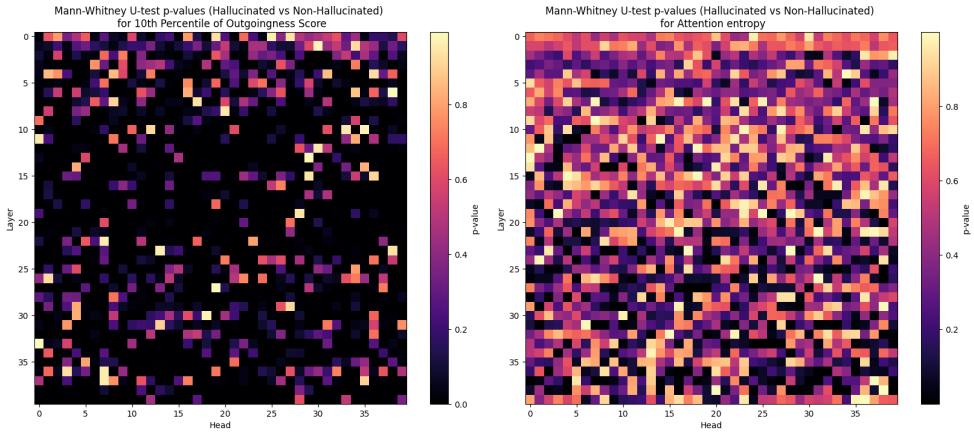
(a) Outgoingness - LLaMA 3.1 8B Instruct

(b) Attention entropy - LLaMA 3.1 8B Instruct



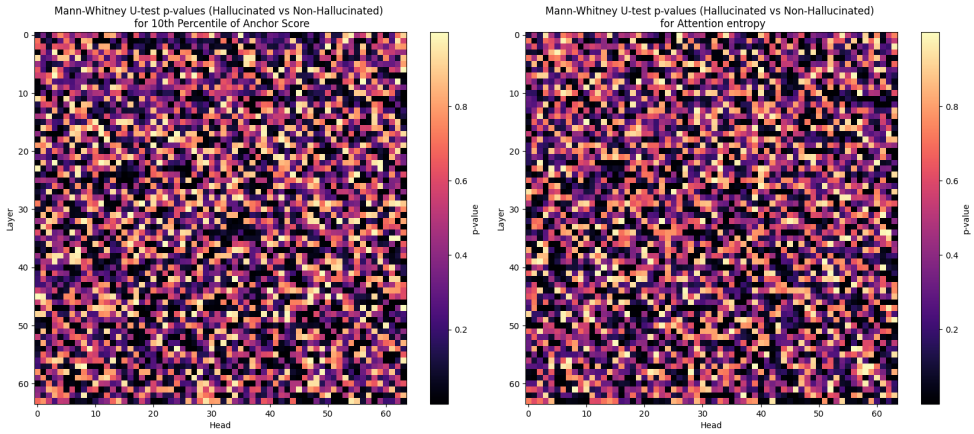
(c) Outgoingness - Mistral 7B Instruct

(d) Attention entropy - Mistral 7B Instruct



(e) Outgoingness - Phi-4

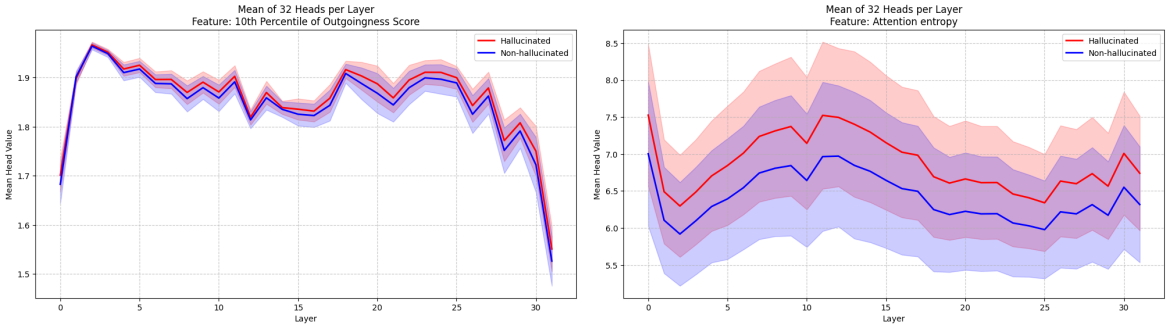
(f) Attention entropy - Phi-4



(g) Outgoingness - Qwen3-32B

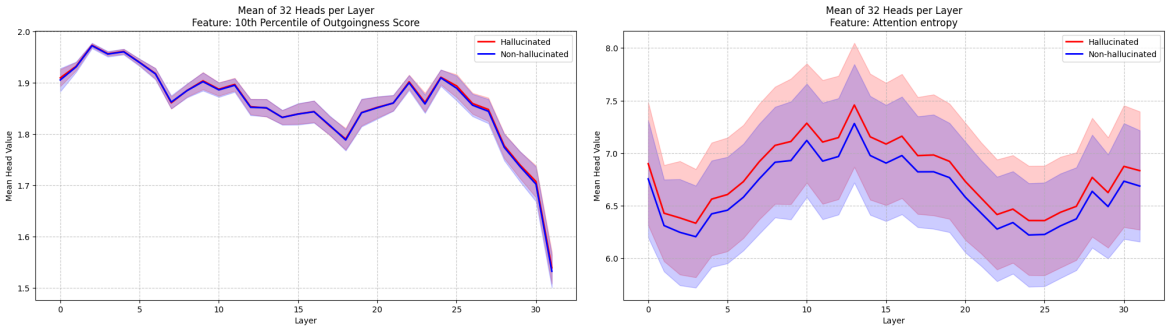
(h) Attention entropy - Qwen3-32B

Figure 9: The p-values for the per-head comparisons of the 10th percentile outgoingness scores and attention entropy across all layers. LLaMA 3.1 8B Instruct exhibits the strongest separation among the majority of attention heads.



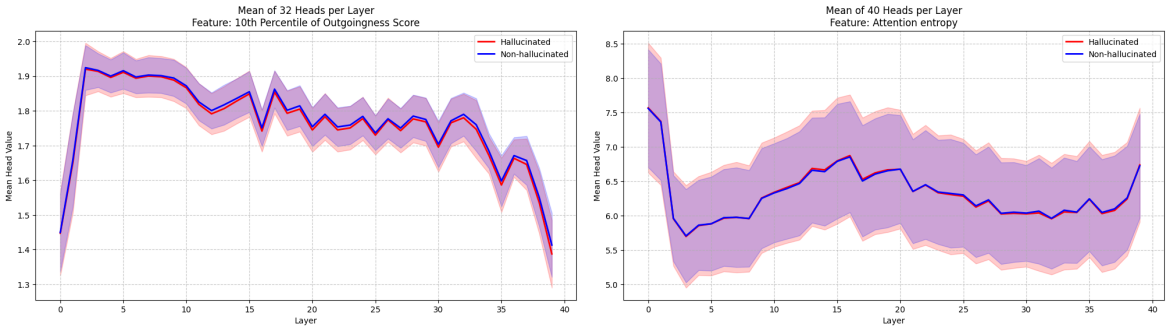
(a) Outgoingness - LLaMA 3.1 8B Instruct

(b) Attention entropy - LLaMA 3.1 8B Instruct



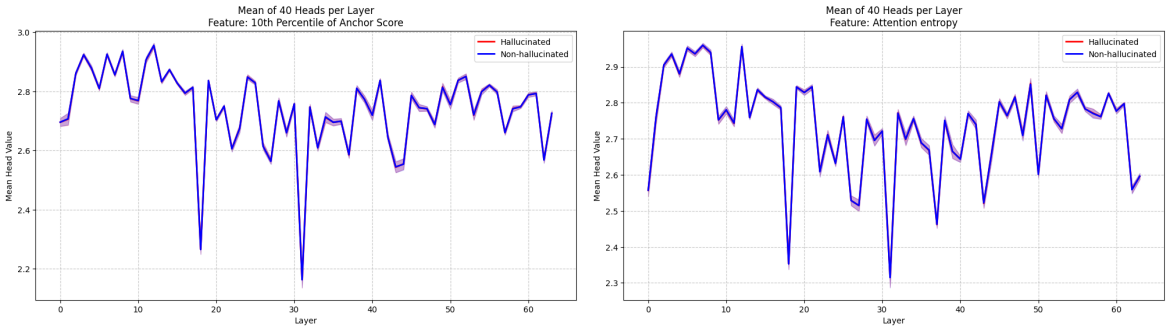
(c) Outgoingness - Mistral 7B Instruct

(d) Attention entropy - Mistral 7B Instruct



(e) Outgoingness - Phi-4

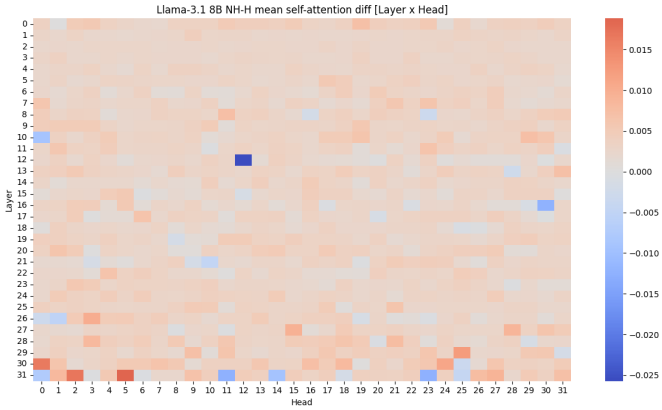
(f) Attention entropy - Phi-4



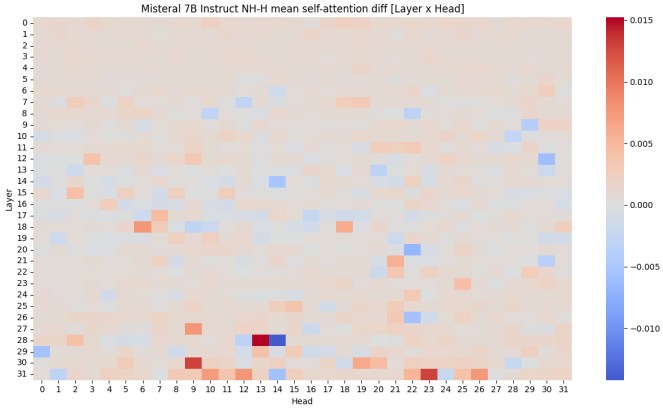
(g) Outgoingness - Qwen3-32B

(h) Attention entropy - Qwen3-32B

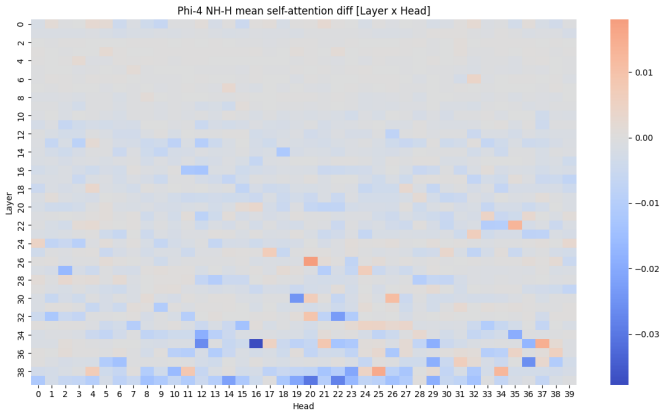
Figure 10: Feature average evolution across layers and comparison between hallucinated and non-hallucinated responses. Both LLaMA 3.1 Instruct and Mistral 7B Instruct shows diffused attention whereas for Phi-4 shows little to no noticeable difference in outgoingness or attention distribution between the two response types. Also, for LLaMA 3.1 Instruct and Mistral 7B Instruct, the average of 10th percentile of the outgoingness score over attention heads is higher for hallucinated response in the terminal layers. The same pattern was observed for 50th and 90th percentiles.



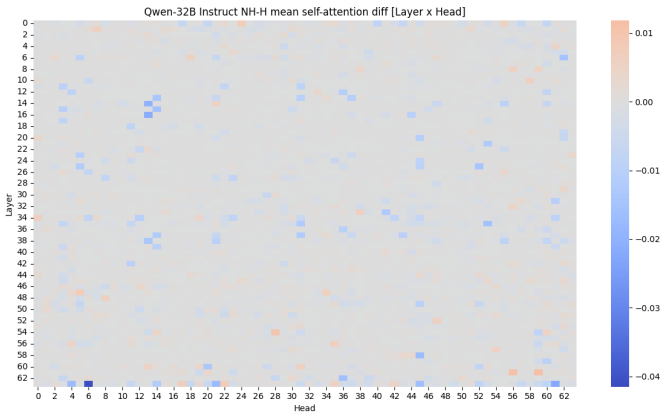
(a) LLaMA 3.1 8B Instruct self-attention



(b) Mistral 7B Instruct



(c) Phi-4



(d) Qwen3-32B

Figure 11: Mean per-head self-attention difference between hallucinated and non-hallucinated responses. Unlike Mistral and LLaMA, Phi, and Qwen show a higher mean per-head self-attention throughout all the layers and the majority of the attention heads.

Temp	Benchmark	Model	AUC				
			LapEigvals	EigenScore	Ours - setting (a)	Ours - setting (b)	Ours - setting (c)
0.1	TruthfulQA	LLaMA 3.1 8B Instruct	80.52	83.79	85.34	**85.56**	85.47
0.1	TruthfulQA	Phi 4	78.43	81.62	81.75	81.80	**82.41**
0.1	TruthfulQA	Mistral 7B Instruct v03	78.80	**81.69**	80.00	80.17	80.11
0.1	TruthfulQA	Qwen3-32B	77.36	72.17	80.17	80.33	**80.47**
1.0	TruthfulQA	LLaMA 3.1 8B Instruct	78.26	80.22	80.73	80.25	**81.00**
1.0	TruthfulQA	Phi 4	74.46	79.82	84.00	83.98	**85.31**
1.0	TruthfulQA	Mistral 7B Instruct v03	73.23	75.38	78.00	**78.68**	78.19
1.0	TruthfulQA	Qwen3-32B	**75.17**	69.35	73.53	74.11	74.65
0.1	NQOPEN	LLaMA 3.1 8B Instruct	75.67	71.49	79.34	**79.38**	79.33
0.1	NQOPEN	Phi 4	80.26	81.69	82.00	82.36	**83.34**
0.1	NQOPEN	Mistral 7B Instruct v03	77.10	79.45	79.75	79.34	**79.77**
0.1	NQOPEN	Qwen3-32B	76.16	71.73	78.02	78.16	**78.37**
1.0	NQOPEN	LLaMA 3.1 8B Instruct	76.19	71.16	**80.46**	80.15	80.21
1.0	NQOPEN	Phi 4	83.61	84.11	84.49	84.77	**85.12**
1.0	NQOPEN	Mistral 7B Instruct v03	76.38	**79.61**	77.29	77.92	77.13
1.0	NQOPEN	Qwen3-32B	78.49	73.63	81.73	81.43	**81.79**

Table 3: AUC comparison between LapEigvals, EigenScore, and our method across three feature settings. Setting (a) includes the 10th, 50th, and 90th percentile outgoingness per-head features, along with attention entropy. Setting (b) extends (a) by incorporating all deciles (10th–90th) to evaluate the impact of finer-grained local information. Setting (c) augments (a) with average per-head self-attention values as additional features.