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 evaluate Forman–Ricci curvatures to identify structural patterns indicative of information bottlenecks in attention graphs and 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 the detection of hallucinated responses, achieving state-of-the-art performance. Further analysis reveals that impaired context sharing among tokens during causal generation substantially contributes to hallucination occurrences in LLMs. In particular, hallucination may emerge due to over-reliance on self-attention, diffused context retrieval from earlier tokens, or information over-squashing, especially in terminal encoder 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). They are rapidly becoming an integral part of our daily routines. These models, especially those designed to work collaboratively (known as agentic LLMs), tackle complex tasks by engaging in a series of sequential consultations (Shinn et al., 2023; Yao et al., 2023). However, in such interconnected systems, a single incorrect response can trigger a cascade of errors, compromising the entire reasoning process. This makes the accuracy of LLM responses critical. Misleading responses can be divided into two main categories: factuality and faithfulness, and mainly arise due to reasoning failure (Berglund et al., 2023), defective inference (Chen et al., 2022), model architecture and training data (Huang et al., 2025), or context inconsistency (Maynez et al., 2020).

Recent research indicates that hallucination, defined as the generation of factually incorrect or unsubstantiated responses, correlates with a state of confusion and instability within the model during response generation. Consequently, the verification of self-consistency in Large Language Models (LLMs) has emerged as a critical intermediate step for detecting such phenomena (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, as it necessitates the generation of multiple responses for a single prompt, needing a large number of generations to achieve good performance (Chen et al., 2024). Secondly, this approach offers limited insight into the underlying causes and mechanisms that lead to 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 spe-

cific topological signatures of hallucination vary with model architecture, even though the core phenomenon of information bottlenecks is consistent.

2 Background

In attention-based language models, the entire attention generation can be seen as a context-sharing process. Let $\{T1, T2, \ldots, T_n\}$ a sequence of tokens, $A = \{a_1, a_2, \ldots a_n\}$ be the attention score vector calculated via equation **??**, and $V = \{v_1, v_2, \ldots v_n\}$ the matrix of value vectors which carry the token contextual information. The context vector for nth token is therefore $\mathrm{C}_n = \sum_{j=1}^{n} a_j \cdot v_j$. The vector decides how much of the context of each of the tokens should be used to represent the token *n* which itself is a representation of the current sentence. While attention weights look backward across previous tokens, the overall information flows forward through the sequence.

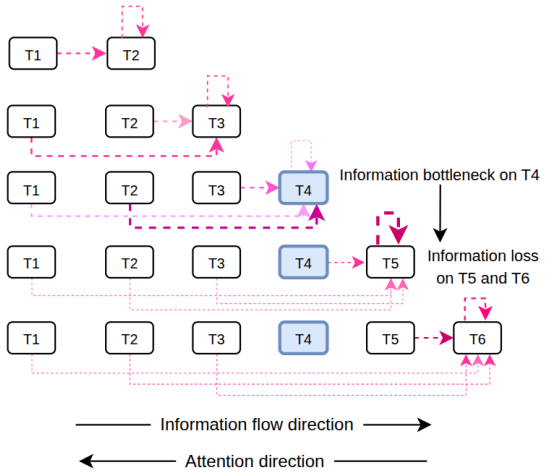


Figure 1: Information over-squashing occurs when information from many tokens is funneled through only a single token that bears the burden of representing all that context. However, this central token and its predecessors receive insufficient attention from subsequent tokens, preventing effective context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

Attention matrix can be seen as a weighted graph with tokens being vertices and attention weights being edges of such a graph. 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 is called information over-squashing and frequently was associated with information bottlenecks in graphs and measured by Forman-Ricci curvatures (Forman, 2003; Alon and Yahav, 2020; Topping et al., 2021).

3 Methodology

The presence of bottlenecks in attention graphs may lead to context loss, which is critical for maintaining coherent semantic construction in transformer models. The Forman-Ricci curvature (FRC) provides a local view of the information flow among graph nodes. 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, the higher the expected value of FRCs of a graph, the better the overall information flow and the less sever the bottlenecks. Let $G = (V, E)$ be the attention graph, with vertices v and w connected by an edge e. FRC is defined as:

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

where w_{vu} are edge weights, $w_v = \sum_{y \sim N(v)} w_{vy}$ are vertex weights (i.e., vertex out-degrees in attention graphs), and $N(v)$, $N(u)$ denote the set of vertices connected to vertices v and u, respectively.

The penalty part of FRC of a weighted graph from (1) can be written in terms of probabilities:

$$P(u,v) = \sum_{x \sim N(u)/v} \frac{1}{\sqrt{p_u(v)p_u(z)}} + \sum_{x \sim N(v)/u} \frac{1}{\sqrt{p_v(u)p_v(z)}} \quad (2)$$

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

The term $\frac{1}{\sqrt{p_u(v)p_u(z)}}$ is convex in p_u over the probability simplex, and it maintains its convexity over summation and thus attains its minimum at the uniform distribution (Bauschke and DiBerardino, 2022). The total penalty can be defined as the sum of the weighted local penalties:

$$\mathrm{P}_{\text{total}} = \sum_{(u,v) \in E} \frac{w_{uv}}{w_{\text{total}}} \cdot \mathrm{P}(u,v) \quad (3)$$

and therefore:

$$\mathrm{P}_{\text{total}} = \sum_{(u,v) \in E} p(u,v) \cdot \mathrm{P}(u,v) \quad (4)$$

where $w_{total} = \sum_{e \sim E} w_e$ is the sum of all edges, $p(u) = \frac{w_u}{w_{total}}$, and $p(u,v) = \frac{w_{u,v}}{w_{total}}$. From the chain rule of entropy, we can write:

$$H(U,V) = H(U) + H(V \mid U) \quad (5)$$

where

$$H(U,V) = -\sum_{(u,v) \in E} p(u,v) \log p(u,v) \quad (6)$$

$$H(U) = -\sum_{u \in V} p(u) \log p(u) \quad (7)$$

$$H(V \mid U) = -\sum_{u \in V} p(u) \sum_{v \sim N(u)} p(v \mid u) \log p(v \mid u) \quad (8)$$

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

$$p(v \mid u) = \frac{p(u,v)}{p(u)} = \frac{\frac{w_{uv}}{W_{\text{total}}}}{\frac{w_u}{W_{\text{total}}}} = \frac{w_{uv}}{w_u} = p_u(v) \quad (9)$$

Therefore,

$$H(V \mid U) = -\sum_{u \in V} p(u) \sum_{v \sim N(u)} p_u(v) \log p_u(v) \quad (10)$$

From equations (2), (5), and (10) a uniform $P_u(v)$ maximizes the joint entropy while it minimizes the total penalty, indicating a reverse correlation between $H(u,v)$ and total penalty in Forman-Ricci.

In its simplest form, when the penalty term is omitted, the Forman-Ricci curvature (FRC) of an edge reduces to the sum of the degrees of its endpoints: $w_{v_1} + w_{v_2}$. In attention graphs, this value reflects the *mutual outgoingness* between two tokens. Intuitively, it acts as a proxy for the combined dispersiveness of both nodes: a low score indicates that both tokens are highly self-focused, while a high score suggests that two highly outward-looking tokens are directly linked, forming a strong bridge in a diffused attention structure. We refer to this as the *mutual outgoingness score*, which can be computed efficiently in $\mathcal{O}(E)$.

In contrast, the full Forman-Ricci curvature includes a penalty term designed to reflect edge redundancy and alternative paths, directly tied to local information flow. However, its computation 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 features as the input of the probe, we summarize the distribution of mutual outgoingness scores in each attention head by computing their 10th, 50th, and 90th percentiles, capturing a semi-local geometric view of token interactions. 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.

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 models — Ministral 7B Instruct, LLaMA 3.1 8B Instruct, and Phi-4 — under two temperature settings (0.1 and 1.0), and across two benchmarks commonly used in hallucination-related studies: TruthfulQA (Lin et al., 2021) and NQ-Open (Lee et al., 2019). The TruthfulQA composed of 817 questions and ground-truth answers specifically designed to cause imitative falsehoods and mimic human misconceptions. It contains 38 categories such as 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 consist of 32 layers each with 32 heads, whereas Phi-4 has 40 layers with 40 heads per layer. 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.

In all the experiments, the Mann–Whitney U test was used to examine the statistical significance of differences between the 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 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. For all models, the last layer shows less fluid information flow for hallucinated responses. This effect is particularly pronounced in LLaMA and Mistral, where the distinction is more consistent across layers and heads. For Phi, 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 Appendix, Figures 6 and 7.

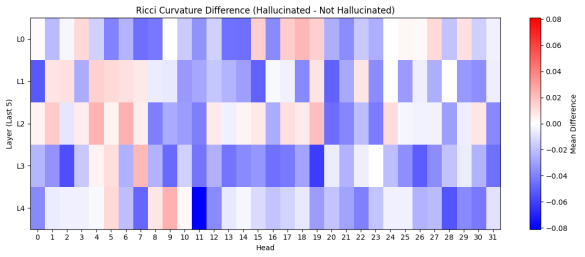


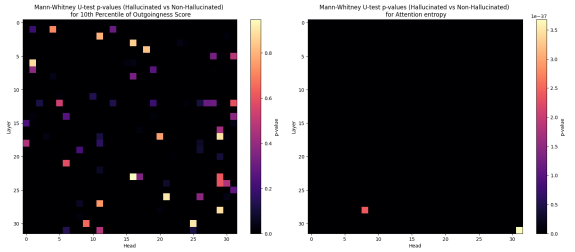
Figure 2: The difference between the mean values of Forman-Ricci curvatures per head for hallucinated and not hallucinated responses for LlaMA. For the majority of the attention heads, hallucinated responses have lower curvature values, which is an indication of a less fluid information flow.

4.3 Outgoingness, entropy and, self-focus patterns

To investigate how hallucinated responses affect outgoingness, distribution of attention across preceding tokens, and self-attention, 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 8), LLaMA 3.1 8B Instruct shows the strongest separation of outgoingness and attention entropy among the other models. In turn, Phi-4 shows statistically meaningful difference between the means of hallucinated and not hallucinated responses for a group of attention heads, while exhibits almost no separation of attention entropies for the majority of the 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.

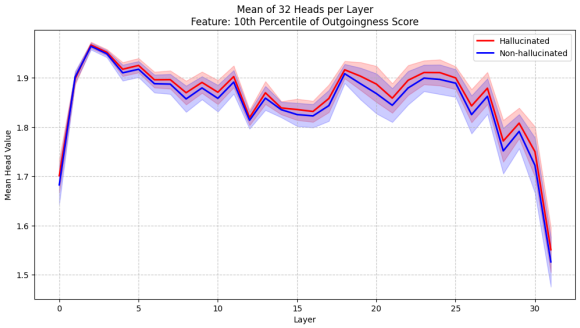
We also analyzed how each feature evolves across layers over time by averaging attention heads, and compared these patterns for hallucinated



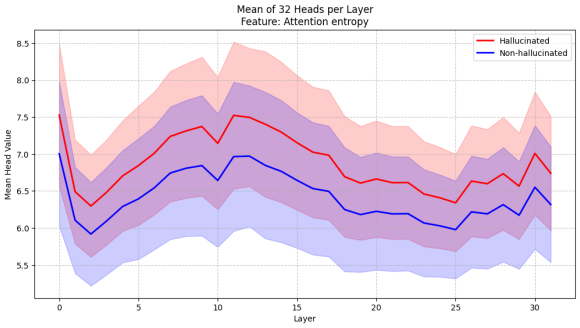
(a) Outgoingness - LLaMA 3.1 8B Instruct (b) Attention entropy - LLaMA 3.1 8B Instruct

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.

responses. As shown in Figure 4, both LLaMA and Mistral exhibit higher outgoingness scores and more diffuse attention distributions in hallucinated responses, suggesting impaired context sharing. In contrast, Phi shows minimal variation in outgoingness across layers and no noticeable change in attention entropy (see Appendix, Figure 9). Also, no clear pattern was observed that differentiates the earlier layers from the later ones in these behaviors.



(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.

Excessive self-attention can lead to a loss of contextual information from surrounding tokens, potentially reducing the accuracy of transformer outputs (Arif et al., 2025). Our experiments indicate that Phi-4 does 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, in most attention heads, hallucinated responses correspond to higher self-attention—an effect especially evident in the later layers. Interestingly, this pattern is reversed and less statistically significant for the majority of the heads in LLaMA and Mistral, where the hallucinated responses show more diffused attention (see Appendix, Figure 10).

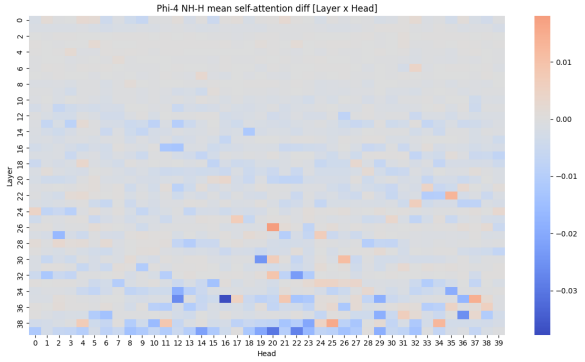


Figure 5: Mean per-head self-attention differences between hallucinated and non-hallucinated responses for Phi-4.

4.4 Results

We rely upon 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. Finally, we compare the results with Laplacian eigenvalues (Binkowski et al., 2025) as our baseline due to its state-of-the-art performance, and a simple yet effective and fast approach for identifying hallucination. The best results were achieved using setting (c) (see Appendix, Table 2).

The experiments indicate that the semi-local outgoingness scores and global attention entropy features result in substantial gains in terms of area

under the curve (AUC) compared to the Laplacian eigenvalues. When more fine-grained outgoingness features were added, not only no significant gains observed, but in some cases it also led to minor over-fitting. When the average self-attention per-head values were added to the feature set, it resulted in an increase in AUC for all the Phi-4 models, yet did not improve the probe's performance in other models significantly as expected.

5 Related works

Hallucination detection methods can be divided into two main approaches. 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 variability of prompts with the same meaning. The more diversified the prompt 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 involves 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 that can lead to 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 for their detection. 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 change depending on whether a response is hallucinated or not. 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 understand whether bottlenecks in attention graphs originate hallucinations, to which extent, and if different LLMs follow the same patterns. Afterwards, we propose a group of features that enable hallucination detection with high performance, in addition to explain why and how hallucination occurs across different LLMs.

6 Limitations

This work suggests that hallucinations in decoder-only models may stem from insufficient context sharing and message passing from earlier tokens during causal generation. It therefore focuses on hallucinations that arise from misinterpretations of prompt intent or prior context. However, it does not address hallucinations that occur due to knowledge gaps in the model's training data. Another limitation is that this study examines context-sharing behavior only in small- to mid-sized language models, leaving it unclear whether the same patterns hold for larger models.

Table 1: AUC Comparison Between Laplacian eigenvalues 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	Ours - setting (a)
0.1	TruthfulQA	LLaMA 3.1 8B Instruct	80.52	**85.34**
0.1	TruthfulQA	Phi 4	78.43	**81.75**
0.1	TruthfulQA	Mistral 7B Instruct v0.3	78.80	**80.00**
1.0	TruthfulQA	LLaMA 3.1 8B Instruct	78.26	**80.73**
1.0	TruthfulQA	Phi 4	74.46	**84.00**
1.0	TruthfulQA	Mistral 7B Instruct v0.3	73.23	**78.00**
0.1	NQOPEN	LLaMA 3.1 8B Instruct	75.67	**79.34**
0.1	NQOPEN	Phi 4	80.26	**82.00**
0.1	NQOPEN	Mistral 7B Instruct v0.3	77.10	**79.75**
1.0	NQOPEN	LLaMA 3.1 8B Instruct	76.19	**80.38**
1.0	NQOPEN	Phi 4	**83.61**	81.75
1.0	NQOPEN	Mistral 7B Instruct v0.3	76.38	**77.29**

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 aimed at capturing 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 consistently contributes to hallucinations in 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 the number of 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 effect of the model size in attention behavior while hallucinating can be investigated in future works. 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 thus be viewed as a breakdown in interpreting the prompt's intent. 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.

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*.

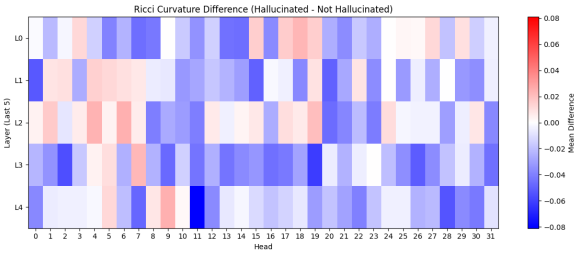
Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. 2021. Understanding over-squashing and bottlenecks on graphs via curvature. *arXiv preprint arXiv:2111.14522*.

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*.

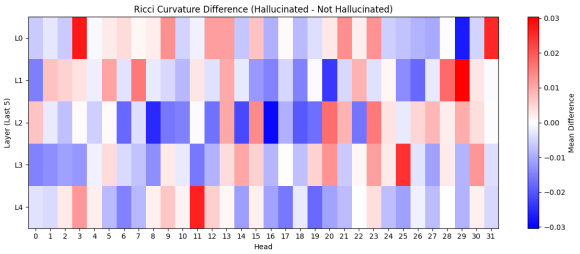
Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

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.

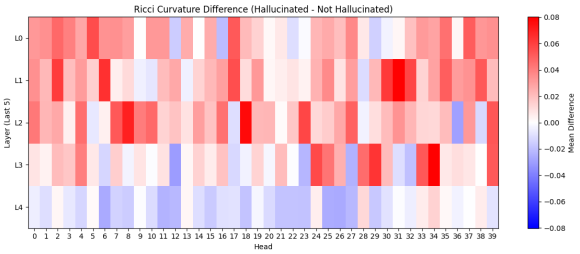
8 Appendix



(a) LLaMA 3.1 8B Instruct



(b) Mistral 7B Instruct

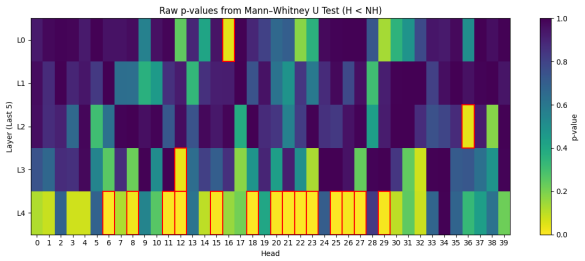


(c) Phi-4

Figure 6: 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, only the last layer presents less fluid context sharing for hallucinated responses.

(a) LLaMA 3.1 8B Instruct

(b) Mistral 7B Instruct

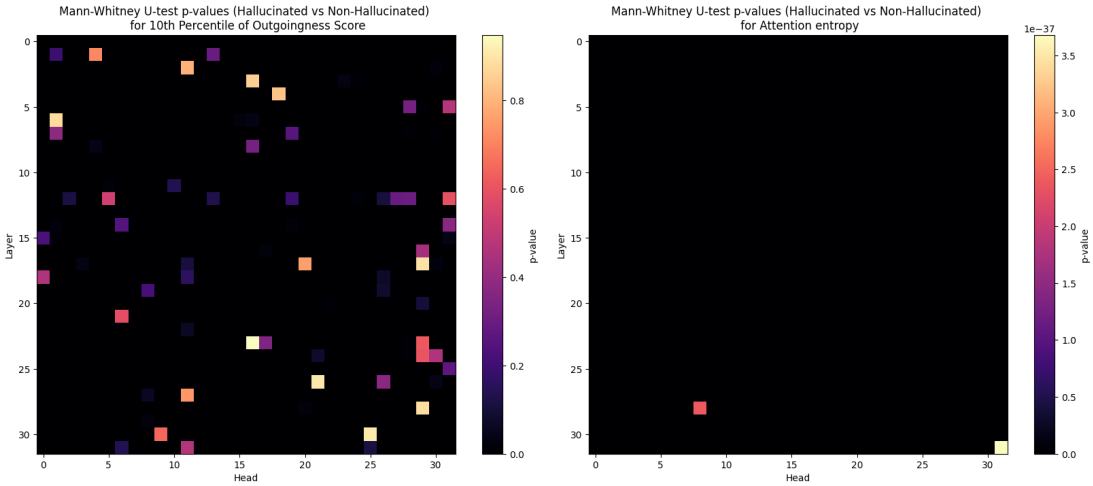


(c) Phi-4

Figure 7: 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.

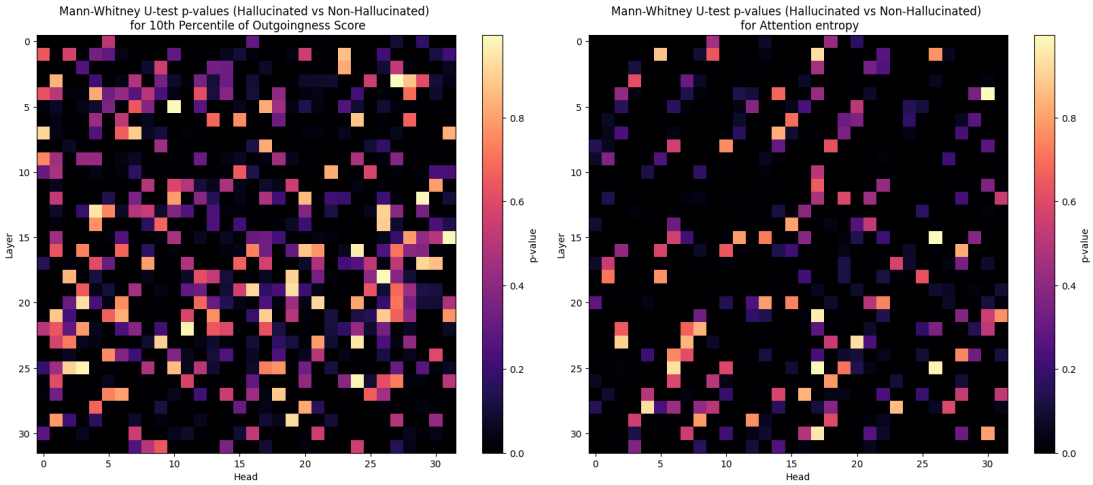
Temp	Benchmark	Model	AUC			
			LapEigvals	Ours - setting (a)	Ours - setting (b)	Ours - (setting (c)
0.1	TruthfulQA	LLaMA 3.1 8B Instruct	80.52	**85.34**	**85.56**	**85.47**
0.1	TruthfulQA	Phi 4	78.43	**81.75**	**81.80**	**82.41**
0.1	TruthfulQA	Mistral 7B Instruct v03	78.80	**80.00**	**80.17**	**80.11**
1.0	TruthfulQA	LLaMA 3.1 8B Instruct	78.26	**80.73**	**80.25**	**81.00**
1.0	TruthfulQA	Phi 4	74.46	**84.00**	**83.98**	**85.31**
1.0	TruthfulQA	Mistral 7B Instruct v03	73.23	**78.00**	**78.68**	**78.19**
0.1	NQOPEN	LLaMA 3.1 8B Instruct	75.67	**79.34**	**79.38**	**79.33**
0.1	NQOPEN	Phi 4	80.26	**82.00**	**82.36**	**83.34**
0.1	NQOPEN	Mistral 7B Instruct v03	77.10	**79.75**	**79.34**	**79.77**
1.0	NQOPEN	LLaMA 3.1 8B Instruct	76.19	**80.38**	**80.15**	**80.21**
1.0	NQOPEN	Phi 4	**83.61**	81.75	81.77	83.00
1.0	NQOPEN	Mistral 7B Instruct v03	76.38	**77.29**	**77.92**	**77.13**

Table 2: AUC comparison between LapEigvals 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.



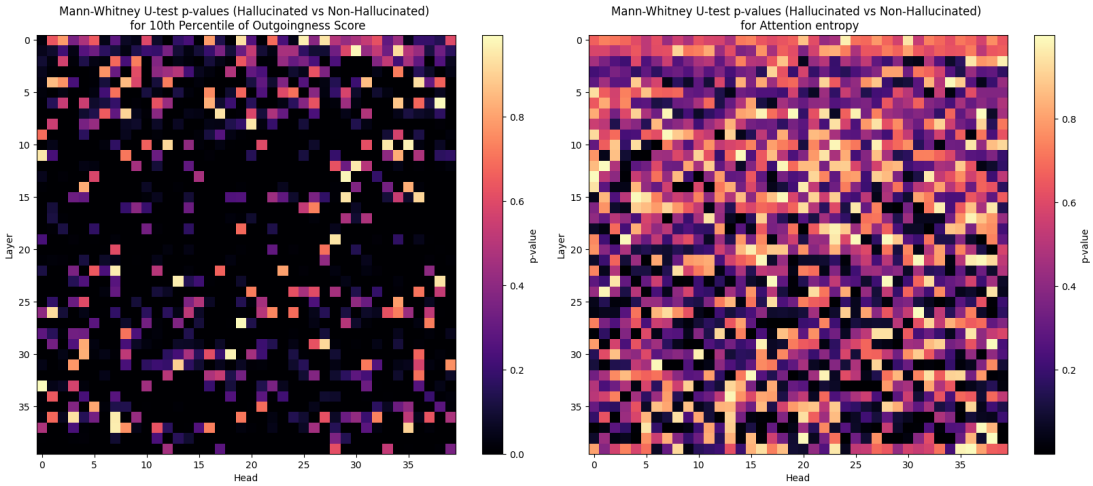
(a) Outgoingness - LLaMA 3.1 8B Instruct

(b) Attention entropy - LLaMA 3.1 8B Instruct



(c) Outgoingness - Mistral 7B Instruct

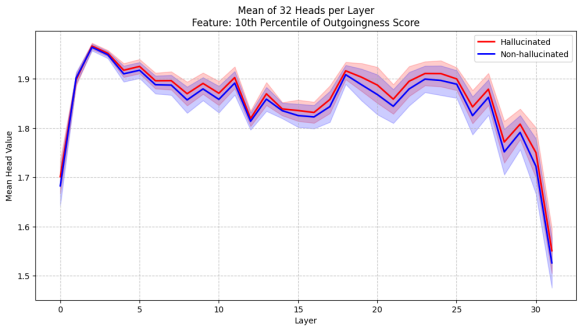
(d) Attention entropy - Mistral 7B Instruct



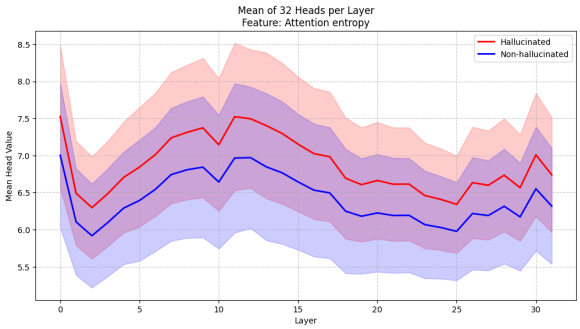
(e) Outgoingness - Phi-4

(f) Attention entropy - Phi-4

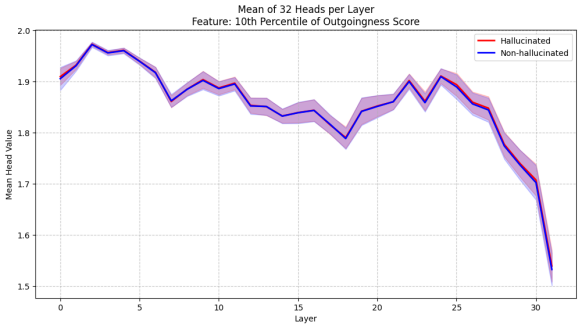
Figure 8: The p-values for the per-head comparisons of the 10th percentile outgoingness scores and attention entropy across all layers. LLaMA 3.1 8B Instruct exhibits the strongest separation among the majority of attention heads.



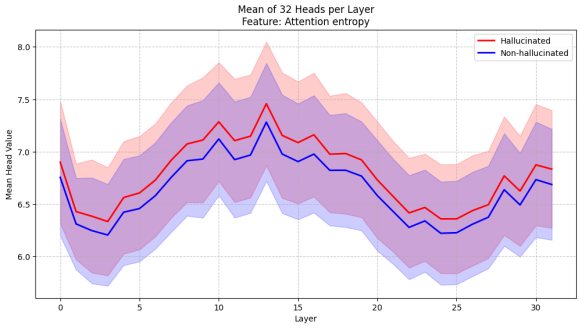
(a) Outgoingness - LLaMA 3.1 8B Instruct



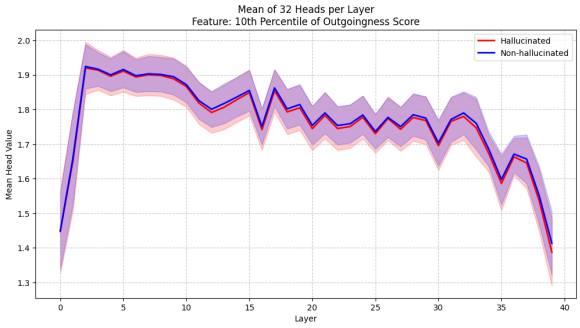
(b) Attention entropy - LLaMA 3.1 8B Instruct



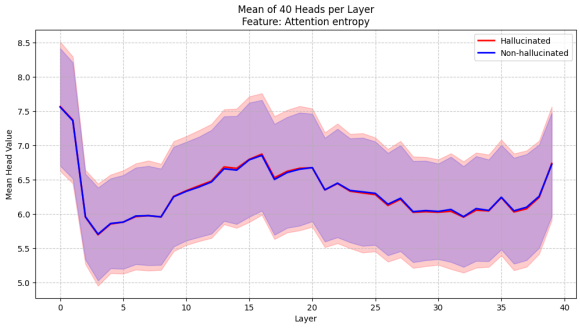
(c) Outgoingness - Mistral 7B Instruct



(d) Attention entropy - Mistral 7B Instruct

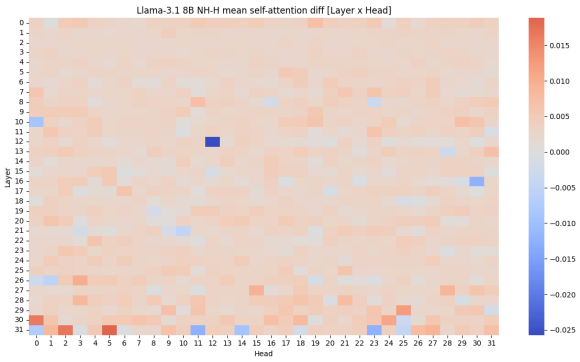


(e) Outgoingness - Phi-4

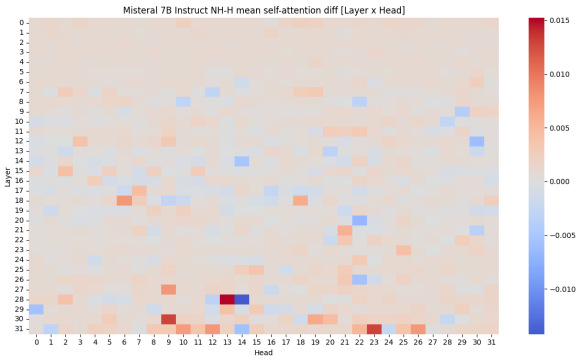


(f) Attention entropy - Phi-4

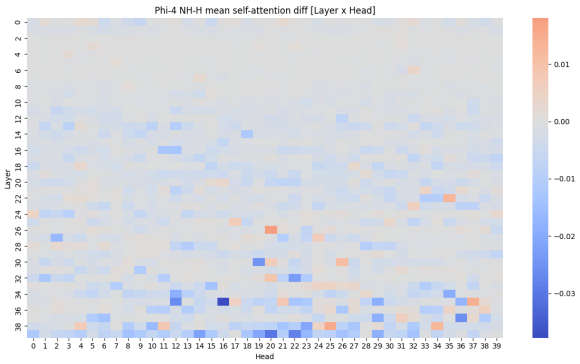
Figure 9: Feature average evolution across layers and comparison between hallucinated and non-hallucinated responses. Both LLaMA 3.1 Instruct and Mistral 7B Instruct shows diffused attention whereas for Phi-4 shows little to no noticeable difference in outgoingness or attention distribution between the 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

Figure 10: Mean per-head self-attention difference between hallucinated and non-hallucinated responses. Unlike Mistral and LLaMA, Phi shows a higher mean per-head self-attention throughout all the layers and the majority of the attention heads.