Uncertainty-Aware Structural Attention Features for Hallucination Detection in LLMs

RLN Lokesh
Department of Electronics and Communications
National Institute of Technology
Andhra Pradesh, India
622246@student.nitandhra.ac.in

Ch Mokshagna
Department of Electronics and Communications
National Institute of Technology
Andhra Pradesh, India
622129@student.nitandhra.ac.in

K Dinesh
Department of Electronics and Communications
National Institute of Technology
Andhra Pradesh, India
622171@student.nitandhra.ac.in

Pagadala Usha
Department of Electronics and Communications
National Institute of Technology
Andhra Pradesh, India
usha.sclr@nitandhra.ac.in

***Abstract*— Large Language Models frequently generate responses that appear fluent and coherent but may contain factually incorrect information, a phenomenon known as hallucinations. Hallucinations have emerged as a significant barrier to the effective application of Large Language Models. Some studies suggest that hallucinations can be detected by analyzing structural properties of attention maps. However, such methods primarily focus on attention structures and their spectral features, whereas uncertainty, randomness, and other statistical measures can also be leveraged to detect hallucinations. In this work, we propose a method for detecting hallucinations using spectral and statistical features extracted from attention maps. We compute eigenvalues of the Laplacian matrix from each attention map to capture spectral features and compute statistical measures for every attention head. The spectral and statistical features are used as a feature vector for hallucination detection. We conducted experiments on the HaluEval dataset across multiple language models, and the results show that jointly modeling structures and uncertainty-aware characteristics leads to more robust hallucination detection.**

***Index Terms*—LLMs, Attention Mechanism, Hallucination Detection, Internal Model Signals.**

I. Introduction

Large Language Models have demonstrated remarkable ability in a wide range of natural language processing tasks including question answering, summarization, and dialogue generation [6]. The self-attention mechanism plays a crucial role by enabling models to capture long-range dependencies through pairwise interactions between tokens in a sequence [5]. However, these models prioritize next-token prediction without ensuring factual correctness, making them prone to generating fluent but incorrect responses [10]. This phenomenon, known as hallucinations, is defined as outputs that are fluent but unfaithful or factually incorrect [7]. Hallucinations occur across different tasks and model families, representing a significant barrier to the safe deployment of LLMs [8], and detecting them has become an important research problem [18], [19].

Existing approaches rely on retrieval, verification, or multiple generations to detect hallucinations [11], [13], while an alternative line of work analyzes internal model behaviors [3], [12], [14]. Attention maps have been shown to provide useful information for hallucination detection [1], [4], and prior work demonstrated that attention maps can be modeled as graphs, and spectral features derived from their Laplacian matrices correlate strongly with hallucinations [1], [16]. However, spectral-only approaches focus on structural information and do not explicitly model uncertainty, randomness, or variability in attention behavior. The internal states of LLMs can also reveal uncertainty-related signals such as entropy and variability across attention heads and layers, which can provide valuable information for detecting hallucinations [3], [9], [15], [17].

A natural question is whether surface-level text features alone could be sufficient for hallucination detection. However, hallucinated responses are fluent and grammatically well-formed by definition, so surface-level classifiers frequently fail to distinguish them from factually correct outputs. Attention maps encode the internal reasoning process, capturing token-level focus, cross-layer consistency, and structural coherence in ways the output text cannot. We

pass the concatenated [prompt+output] pair through the *same* model in frozen inference mode and extract attention maps across all layers and heads, with no separate or larger model involved. This is grounded in the observation that hallucinated responses exhibit characteristic attention patterns such as increased entropy, reduced cross-layer consistency, and irregular spectral structure, that differ measurably from those observed for faithful outputs.

The key contributions of our work are summarized as follows:

- We introduce a hallucination detection framework that jointly models spectral features from the Laplacian of attention maps and statistical features per attention head, capturing both structural and uncertainty properties.
- We design a unified feature representation by concatenating both features as input to a LightGBM classifier, chosen through empirical comparison against multiple classifiers on the same feature set.
- We conduct experiments on the HaluEval dataset across multiple models reporting accuracy, precision, recall, F1, and AUC-ROC across $K \in \{30, 50, 100\}$, demonstrating that combining structural and uncertainty-aware features leads to more robust detection.
- We perform an ablation study to assess the contribution of each statistical feature, providing insight into which uncertainty-aware signals are most informative.

II. Related Work

Hallucinations in LLMs have become a prominent topic in research on language models. Uncertainty-based methods have gained traction as complementary approaches. Farquhar et al. [9] introduced semantic entropy as a measure of uncertainty in the LLM outputs, showed that entropy computed over multiple sampled generations correlates strongly with factual errors. Su et al. [3] proposed unsupervised real-time hallucination detection based on internal state dynamics, demonstrating that variability across layers and attention heads can serve as early warning signals. Manakul et al. [15] developed SelfCheckGPT, which uses consistency across multiple generations to identify hallucinations. Similarly, Yehuda et al. [16] introduced InterrogateLLM, which uses self-interrogation to detect hallucinations without external resources. Goel et al. [17] proposed FINCH-ZK, a fine-grained cross-model consistency framework that combines zero-knowledge detection with mitigation strategies [2], [4].

Spectral approaches capture global structural patterns in attention by modeling attention maps as graphs, while other internal state-based methods analyze variability and confidence signals to assess reliability [10], [17]. However, these directions are typically explored independently. Unlike [1], which relies solely on spectral features, our work additionally incorporates uncertainty-aware statistical features to capture complementary information from attention behavior. This unified view allows for a more general and flexible hallucination detection framework that is not limited to a single type of internal signal.

III. Proposed Work

Fig. 1 shows an overview of our proposed hallucination detection framework. In our approach, given a sentence, we extract attention maps from all transformer layers and all attention heads to capture the complete attention behavior of the language model. Each attention map is represented as a square matrix of size $T \times T$, where T denotes the number of tokens in the sequence. A model with L layers and H attention heads produces a total of $L \times H$ attention maps for a single sample. These attention maps are then processed through two different paths to capture complementary features.

A. Path 1: Spectral Feature Extraction

From each attention matrix, we construct the graph Laplacian matrix using Eq. (1), capturing the structural connectivity of the attention graph.

$$\mathbf{L}^{(l,h)} = \mathbf{D}^{(l,h)} - \mathbf{A}^{(l,h)} \tag{1}$$

where $\mathbf{L}^{(l,h)}$ is the Laplacian at layer l and head h, $\mathbf{A}^{(l,h)}$ is the attention-based adjacency matrix whose element $a_{ui}^{(l,h)}$ denotes the attention weight from node u to node i, and $\mathbf{D}^{(l,h)}$ is the corresponding diagonal degree matrix. The diagonal entries are computed as:

$$d_{ii}^{(l,h)} = \frac{\sum_u a_{ui}^{(l,h)}}{\sum_v \mathbb{I}\left(a_{vi}^{(l,h)} \neq 0\right)} \tag{2}$$

where $\sum_u a_{ui}^{(l,h)}$ is the sum of incoming attention weights to node i, and $\sum_v \mathbb{I}(a_{vi}^{(l,h)} \neq 0)$ counts nodes with non-zero attention connections to node i.

For each Laplacian, we compute eigenvalues and select the top K as spectral features encoding structural patterns of attention behavior. We evaluate $K \in \{30, 50, 100\}$: $K = 30$ retains the most dominant structural modes and is computationally

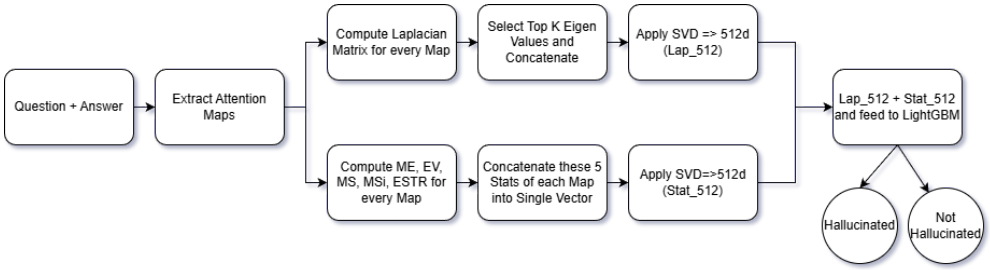


Fig. 1: Overview of the proposed attention-based hallucination detection framework.

Feature	Formula	Intuition
Mean Entropy (ME)	$\text{ME} = \mu_H = \frac{1}{T}\sum_{i=1}^{T} H_i, \quad H_i = -\sum_{j=1}^{T} A_{ij} \ln A_{ij}$	Measures overall attention uncertainty. Low ME = focused; high ME = diffuse.
Entropy Variance (EV)	$\text{EV} = \sigma_H^2 = \frac{1}{T}\sum_{i=1}^{T}(H_i - \mu_H)^2$	Captures consistency of attention focus. High EV = unstable; low EV = uniform.
Mean Similarity (MS)	$\text{MS} = \mu_S = \frac{1}{H}\sum_{p=1}^{H} s_p$	Measures cross-layer attention consistency. Higher MS = stable behavior.
Minimum Similarity (MSi)	$\text{MSi} = S_{\min} = \min_p s_p$	Highlights weakly aligned heads. Low MSi = noisy or unreliable attention.
Entropy-to-Similarity Ratio (ETSR)	$\text{ETSR} = \frac{\mu_S}{\mu_H + \varepsilon}$	Balances consistency against uncertainty. High ETSR = confident; low ETSR = diffuse.

TABLE I: Summary of five statistical features computed per attention head in Path 2.

efficient, $K = 50$ captures additional spectral detail, and $K = 100$ provides the richest representation. Results remain consistently strong across all settings, indicating robustness of the proposed framework. Since $K = 100$ yielded the best overall performance, it is used for detailed analysis throughout. Spectral features are extracted for every attention head across all layers, yielding a feature matrix of dimension $(L \times H) \times K$. SVD is then applied to reduce this to a compact representation of size d_s (we set $d_s = 512$).

B. Path 2: Statistical Feature Extraction

The statistical path captures uncertainty, variability, and cross-layer similarity in attention behavior. For similarity-based features (MS and MSi), the attention matrix is flattened and ℓ_2-normalized as:

$$\hat{\mathbf{v}} = \frac{\text{vec}(\mathbf{A})}{\|\text{vec}(\mathbf{A})\|_2 + \varepsilon}, \tag{3}$$

where ε is a small constant for numerical stability. Cosine similarity with each head p in the previous layer is computed as:

$$s_p = \left\langle \hat{\mathbf{v}}^{(\ell)}, \hat{\mathbf{v}}_p^{(\ell-1)} \right\rangle. \tag{4}$$

Five statistical measures are computed per attention head, as summarized in Table I.

These five features are computed independently for each attention head and aggregated across layers, forming a statistical feature matrix of dimension $(L \times H) \times 5$. SVD is applied to obtain a compact representation of size d_t (we set $d_t = 512$).

Note: The dimension $d_s = d_t = 512$ was selected empirically to balance representational capacity with computational efficiency.

The embeddings from both paths are concatenated into a final feature vector of dimension $d_s + d_t = 1024$, capturing both structural and uncertainty-related characteristics. This feature vector is fed into a LightGBM classifier trained to predict whether a response is hallucinated or non-hallucinated. LightGBM is selected based on empirical comparison against multiple classifiers on the same feature set. As shown in Table II, LightGBM achieves the best accuracy and AUC-ROC among all evaluated classifiers.

Classifier	Accuracy	AUC-ROC
Logistic Regression	0.812	0.845
SVM (RBF Kernel)	0.836	0.872
Neural Network (MLP)	0.861	0.895
Decision Tree	0.724	0.752
Random Forest	0.902	0.928
XGBoost	0.914	0.937
LightGBM (Proposed)	**0.9435**	**0.9597**
K-Nearest Neighbors (KNN)	0.798	0.830
Naive Bayes	0.712	0.740

TABLE II: Classifier comparison on TinyLlama-1.3B at $k = 100$. LightGBM achieves the best performance.

K	Method	TinyLlama-1.3B					GPT-Neo-1.3B				
		Acc.	Prec.	Rec.	F1	AUC	Acc.	Prec.	Rec.	F1	AUC
30	Existing/Spec. Only [1]	0.9100	0.9085	0.9112	0.9098	0.9580	0.8185	0.8203	0.8168	0.8185	0.8970
	Statistical Only	0.8760	0.8741	0.8779	0.8760	0.9312	0.7920	0.7938	0.7903	0.7920	0.8640
	Proposed	**0.9465**	**0.9448**	**0.9481**	**0.9464**	**0.9755**	**0.9230**	**0.9215**	**0.9244**	**0.9229**	**0.9750**
50	Existing/Spec. Only [1]	0.9185	0.9170	0.9199	0.9184	0.9650	0.8350	0.8367	0.8334	0.8350	0.9020
	Statistical Only	0.8810	0.8796	0.8822	0.8809	0.9340	0.8050	0.8063	0.8036	0.8049	0.8710
	Proposed	**0.9445**	**0.9429**	**0.9460**	**0.9444**	**0.9760**	**0.9330**	**0.9314**	**0.9345**	**0.9329**	**0.9710**
100	Existing/Spec. Only [1]	0.9340	0.9324	0.9355	0.9339	0.9580	0.7870	0.7889	0.7853	0.7871	0.8544
	Statistical Only	0.8890	0.8874	0.8903	0.8888	0.9370	0.7630	0.7645	0.7615	0.7630	0.8290
	Proposed	**0.9435**	**0.9419**	**0.9450**	**0.9434**	**0.9597**	**0.9370**	**0.9354**	**0.9385**	**0.9369**	**0.9650**

TABLE III: Hallucination Detection Performance on TinyLlama-1.3B and GPT-Neo-1.3B.

IV. Results and Analysis

This section presents the experimental results and analysis of our proposed framework.

A. Models Used

We evaluate the proposed hallucination detection framework on two transformer-based language models: GPT-Neo-1.3B and TinyLlama-1.3B.

Table IV shows model configuration. For both models, attention maps are extracted from all layers and all attention heads, and the model parameters remain frozen during experiments. No fine-tuning is performed on the language models, only the LightGBM classifier is trained on the extracted features.

Model	No. of Layers	No. of Heads	Total Attention Maps
GPT-Neo-1.3B	24	16	384
TinyLlama-1.3B	22	32	704

TABLE IV: Model Configuration

B. Dataset Used

We conduct experiments on the HaluEval benchmark [20], a large-scale hallucination evaluation dataset consisting of question-answer pairs labeled as hallucinated or non-hallucinated, with over 10,000 samples and balanced class distribution. For each sample, the question and its generated answer are concatenated as input to extract features from the language models. We follow a standard 80/20 train-test split for training and evaluation. All results are averaged over 5 independent runs with different random seeds to ensure reliability.

C. Evaluation Metrics

We report Accuracy, Precision, Recall, F1-score, and AUC-ROC. Since hallucination detection is a binary classification task, class-wise metrics are important to fully assess performance on both classes. These are defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

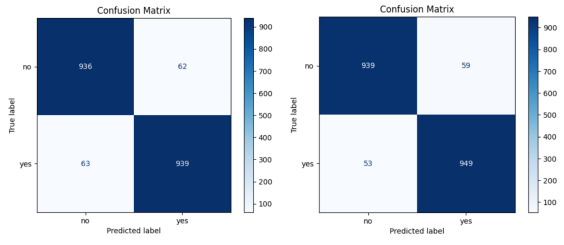
$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

$$\text{AUC-ROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{8}$$

D. Performance Analysis

Table III shows that the proposed framework consistently outperforms both the spectral-only baseline and the statistical-only variant in all values of K and both models. On TinyLlama-1.3B, accuracy improves by 1.0% to 3.7% and AUC-ROC by 0.2% to 1.8%. Gains are more pronounced on GPT-Neo-1.3B, with accuracy improving by 10.4% to 15.0% and AUC-ROC by 7.8% to 11.1%, confirming that combining spectral and statistical features yields more robust hallucination detection.



(a) GPT-Neo-1.3B (b) TinyLlama-1.3B

Fig. 2: Confusion matrices for GPT-Neo-1.3B and TinyLlama-1.3B at $k = 100$.

As shown in Fig. 2, both models achieve strong performance. GPT-Neo-1.3B correctly identifies 939 hallucinated and 936 non-hallucinated samples, while TinyLlama-1.3B yields 949 true positives and 939 true negatives. False positives arise from non-hallucinated responses with diffuse attention patterns resembling hallucinated outputs, while false

Feature Combination	GPT-Neo-1.3B					TinyLlama-1.3B				
	Acc.	Prec.	Rec.	F1	AUC	Acc.	Prec.	Rec.	F1	AUC
Spectral only (baseline [1])	0.7870	0.7889	0.7853	0.7871	0.8544	0.9340	0.9324	0.9355	0.9339	0.9580
Statistical only	0.7630	0.7645	0.7615	0.7630	0.8290	0.8890	0.8874	0.8903	0.8888	0.9370
Spectral + ME	0.8410	0.8428	0.8392	0.8410	0.9020	0.9365	0.9350	0.9379	0.9364	0.9583
Spectral + EV	0.8530	0.8547	0.8512	0.8529	0.9110	0.9370	0.9355	0.9384	0.9369	0.9585
Spectral + MS	0.8820	0.8836	0.8803	0.8819	0.9310	0.9390	0.9375	0.9404	0.9389	0.9589
Spectral + MSi	0.8650	0.8667	0.8633	0.8650	0.9180	0.9375	0.9360	0.9389	0.9374	0.9586
Spectral + ETSR	0.9010	0.9025	0.8994	0.9009	0.9430	0.9405	0.9390	0.9419	0.9404	0.9591
Spectral + All (Proposed)	**0.9370**	**0.9354**	**0.9385**	**0.9369**	**0.9650**	**0.9435**	**0.9419**	**0.9450**	**0.9434**	**0.9597**

TABLE V: Ablation study: contribution of each statistical feature at $k = 100$.

negatives occur for hallucinated responses with locally structured attention. These observations suggest that combining attention-based features with output-level confidence scores could help reduce both error types in future work.

E. Ablation Study

To assess the contribution of each statistical feature, we train the LightGBM classifier using spectral features combined with each statistical feature independently at $k = 100$, with results summarized in Table V. Every statistical feature contributes consistent improvements over the spectral-only baseline to varying degrees, while the statistical-only variant underperforms the spectral-only baseline, confirming that both components are necessary and complementary. The model combining all five features achieves best performance across all metrics on both models.

V. Conclusion

In this work, we presented a hallucination detection framework that jointly models spectral and statistical features derived from LLM attention maps. By re-processing [prompt+output] pairs through the same frozen model, the framework captures internal attention signals inaccessible to surface-level text classifiers. Integrating structural features with uncertainty-aware signals such as entropy, variance, and cross-layer similarity yields a more expressive representation of hallucinated outputs. Experimental results and ablation studies confirm that combining both feature types leads to more robust and reliable hallucination detection across both models and all evaluation metrics.

References

[1] J. Binkowski *et al.*, "Hallucination Detection in LLMs Using Spectral Features of Attention Maps," in *Proc. EMNLP*, 2025.

[2] G. Sriramanan *et al.*, "LLM-Check: Investigating Detection of Hallucinations in Large Language Models," in *Proc. NeurIPS*, 2024.

[3] W. Su *et al.*, "Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models," in *Proc. Findings of the Association for Computational Linguistics (ACL)*, 2024.

[4] Q. Liu *et al.*, "Attention-Guided Self-Reflection for Zero-shot Hallucination Detection in Large Language Models," in *Proc. EMNLP*, 2025.

[5] A. Vaswani *et al.*, "Attention Is All You Need," in *Proc. NeurIPS*, 2017.

[6] T. Brown *et al.*, "Language Models are Few-Shot Learners," in *Proc. NeurIPS*, 2020.

[7] J. Maynez *et al.*, "On Faithfulness and Factuality in Abstractive Summarization," in *Proc. ACL*, 2020.

[8] S. Ji *et al.*, "A Survey of Hallucination in Natural Language Generation," in *Proc. ACM Computing Surveys*, 2023.

[9] S. Farquhar *et al.*, "Detecting Hallucinations in Large Language Models Using Semantic Entropy," *Nature*, vol. 630, 2024.

[10] M. Steyvers *et al.*, "What Large Language Models Know and What People Think They Know," in *Proc. Nature Machine Intelligence*, 2024.

[11] Z. Wei *et al.*, "Detecting and Mitigating the Ungrounded Hallucinations in Text Generation by LLMs," in *Proc. AISNS*, 2023.

[12] S. Dhuliawala *et al.*, "Chain-of-Verification Reduces Hallucination in Large Language Models," in *Findings of ACL*, 2024.

[13] Z. Liu *et al.*, "Comprehensive Evaluation of AI Hallucination and Novel UV-Oriented Framework toward Safe and Trustworthy AI," in *Proc. Int. Conf. on Universal Village (UV)*, 2024.

[14] Y. Hou *et al.*, "A Probabilistic Framework for LLM Hallucination Detection via Belief Tree Propagation," in *Proc. NAACL*, 2025.

[15] E. Manakul *et al.*, "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Language Models," in *Proc. EMNLP*, 2023.

[16] O. Yehuda *et al.*, "InterrogateLLM: Zero-Resource Hallucination Detection in LLM-Generated Answers," in *Proc. ACL*, 2024.

[17] A. Goel *et al.*, "Zero-Knowledge LLM Hallucination Detection and Mitigation through Fine-Grained Cross-Model Consistency (FINCH-ZK)," in *Proc. EMNLP*, 2025.

[18] P. Lewis *et al.*, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Proc. NeurIPS*, 2020.

[19] W. Zhang and J. Zhang, "Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review," *Mathematics*, 2025.

[20] A. Mishra *et al.*, "HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models," in *Proc. EMNLP*, 2023.