# Hallucinations in Large Language Models

By
Lokesh RLN – 622246
Ch Mokshagna – 622129
Dinesh K – 622171

Under Supervision of Mrs. P. Usha, Ad-hoc Faculty, Dept. of ECE


Department of Electronics and Communications Engineering
National Institute of Technology Andhra Pradesh
December 2025

# Contents

**Abstract**

This report studies detection of hallucinations in large language models (LLMs). We present an existing spectral attention-based probe (LapEigvals) and propose a complementary lightweight statistical feature pipeline that, when fused with spectral features, improves hallucination detection performance. We evaluate on the HaluEval dataset and report classification accuracy and AUC-ROC for representative models. Results show that the proposed fused method improves detection performance across studied architectures.

# 1 Introduction

## 1.1 Problem statement

Large Language Models (LLMs) can produce fluent but factually incorrect or fabricated outputs — a behavior commonly referred to as *hallucination*. Hallucinations undermine trust and hinder deployment of LLMs in critical domains such as healthcare, law, finance and scientific applications. The core challenge is that these models are optimized to predict plausible continuations (maximize likelihood), not to be grounded in facts; hence they may generate confident but incorrect claims when the model's internal distributions favor such continuations.

This work considers the detection of hallucinated outputs using internal model signals (attention maps and derived statistics), with the goal of producing a compact, interpretable and efficient detector that generalizes across architectures. The principal research question: *Can spectral features of attention maps combined with compact per-head statistical features produce a robust detector for hallucinations?*

# 2 Dataset and Models: Used and short description

## 2.1 Dataset: HaluEval

We use **HaluEval** — a dataset designed for hallucination-detection research. HaluEval contains roughly 10k examples (question-answer pairs) annotated with a binary hallucination label. Each example includes:

- **Question** (prompt)
- **Answer** (model output)
- **Hallucination Label** (1 = hallucinated, 0 = non-hallucinated)

For detector evaluation we concatenate Question+Answer as the model input to extract internal attention states for feature computation.

## 2.2 Models evaluated

Two representative LLM variants used in experiments (as in the slides):

- **TinyLlama-1.3B**: 22 layers, 32 heads per layer — used as a Llama-family representative where attention pattern spectral features are informative.
- **GPT Neo-1.3B**: 24 layers, 16 heads per layer — used as a GPT-style representative to test generalization.

# 3 Methodology

High-level pipeline: extract internal attention maps $A^{(l,h)}$ (for each layer $l$ and head $h$) while teacher-forcing the model on prompt+response, compute features along two parallel paths (spectral and statistical), reduce dimensionality, fuse, and feed a lightweight classifier (LightGBM in experiments).

## 3.1 Existing Methodology

### 3.1.1 Overview and key idea

The existing method (referred to as **LapEigvals**) treats each attention map $A^{(l,h)} \in \mathbb{R}^{T\times T}$ as the adjacency matrix of a directed graph, constructs a graph Laplacian and uses top-$k$ eigenvalues of this Laplacian as spectral features. The assumption: certain spectral signatures of attention correlate with hallucinated outputs.

### 3.1.2 Pipeline (existing)

1. Extract attention maps $A^{(l,h)}$ for all layers and heads.
2. For each $A$, compute a diagonal degree matrix $D$ (paper formula) and Laplacian $L = D - A$.
3. Compute eigenvalues of $L$, sort, and select top-$k$ eigenvalues per head.
4. Concatenate top-$k$ spectral features across heads and layers forming a high-dimensional spectral vector.
5. Apply dimensionality reduction (PCA/SVD) to obtain a fixed-size embedding (e.g., **Lap512**).
6. Train a classifier (LightGBM) on these spectral embeddings to predict hallucination.

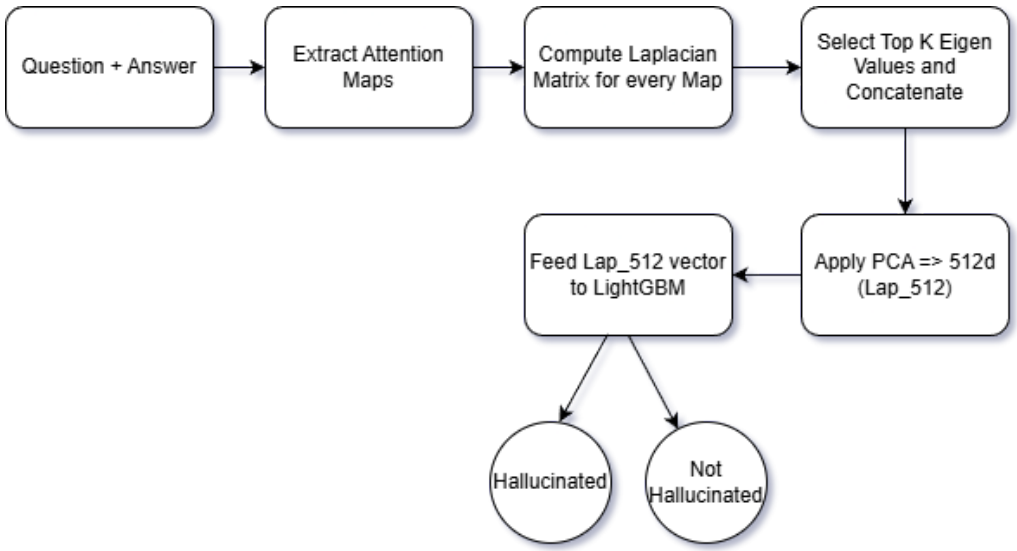


Figure 1: Flowchart of the Existing Method (LapEigvals).

#### 3.1.3 Remarks

The existing spectral-only approach is interpretable and leverages global structural information in attention patterns. However, it can be high-dimensional and may miss compact per-head uncertainty and consistency signals that are helpful in detection.

### 3.2 Proposed Methodology

#### 3.2.1 Key idea

Augment spectral (Laplacian) features with compact, interpretable per-head statistical features that capture attention uncertainty, cross-layer similarity and structural consistency. Reduce both spectral and statistical representations and fuse them to form a richer embedding for the hallucination probe.

#### 3.2.2 Per-head statistical features

For each attention head we compute five features:

1. **Mean Entropy (ME)**: average of per-row entropies of the attention matrix. Entropy per row $H_i = -\sum_j A_{ij} \ln A_{ij}$, then $\mathrm{ME} = \frac{1}{T} \sum_i H_i$.

2. **Entropy Variance (EV)**: variance of the per-row entropies, measures consistency/dispersion of attention focus.

3. **Mean Similarity (MS)**: cosine similarity of the head's flattened, $\ell_2$-normalized attention vector with heads in the previous layer; averaged over those heads.

4. **Minimum Similarity (MSi)**: minimum of the cross-layer similarities (captures worst alignment to prior layer).

5. **Entropy-to-Similarity Ratio (ETSR)**: ETSR $= \mu_S/(\mu_H + \varepsilon)$, combining similarity and entropy into a single normalized signal.

#### 3.2.3 Proposed full pipeline

1. Extract $A^{(l,h)}$ for all layers/heads.

2. **Spectral path**: compute Laplacian $L^{(l,h)}$, extract top-$k$ eigenvalues per head, stack across heads/layers $\rightarrow$ reduce via SVD/PCA to **Lap512**.

3. **Statistical path**: compute the five per-head features for each head, stack into a matrix per example $\rightarrow$ reduce via SVD to **Stat512**.

4. Concatenate Lap512 + Stat512 $\rightarrow$ final 1024-d embedding.

5. Train LightGBM (or another lightweight classifier) on the 1024-d vectors to predict hallucination label.

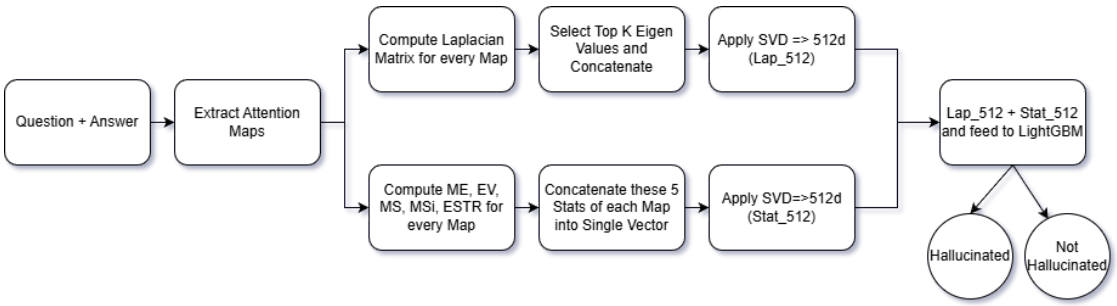


Figure 2: Flowchart of the Proposed Method (spectral + statistical fusion).

#### 3.2.4 Remarks

The proposed fusion leverages complementary strengths: spectral captures global graph-structure of attention while statistical features encode uncertainty and cross-layer consistency. Dimensionality reduction keeps the final embedding compact and practical for training a classifier.

## 4 Results

We present detection performance tables extracted from experiments reported in the presentation. For each table the **proposed** method's results are shown in **bold**. Results are given for different choices of $k$ (top-$k$ eigenvalues per head) in the spectral path.

### 4.1 TinyLlama-1.3B ("LLama") – Accuracy

| k | Existing (Accuracy) | Proposed (Accuracy) |
|---|---|---|
| 30 | 0.9100 | **0.9465** |
| 50 | 0.9185 | **0.9445** |
| 100 | 0.9340 | **0.9435** |

Table 1: TinyLlama (LLama) — Accuracy. Proposed results in **bold**.

### 4.2 TinyLlama-1.3B ("LLama") – AUC-ROC

| k | Existing (AUC) | Proposed (AUC) |
|---|---|---|
| 30 | 0.9580 | **0.9755** |
| 50 | 0.9650 | **0.9760** |
| 100 | 0.9580 | **0.9590** |

Table 2: TinyLlama (LLama) — AUC-ROC. Proposed results in **bold**.

### 4.3 GPT Neo-1.3B ("GPT") – Accuracy

| k | Existing (Accuracy) | Proposed (Accuracy) |
|---|---|---|
| 30 | 0.8185 | **0.9230** |
| 50 | 0.8350 | **0.9330** |
| 100 | 0.7870 | **0.9370** |

Table 3: GPT Neo ("GPT") — Accuracy. Proposed results in **bold**.

### 4.4 GPT Neo-1.3B ("GPT") – AUC-ROC

| k | Existing (AUC) | Proposed (AUC) |
|---|---|---|
| 30 | 0.8970 | **0.9750** |
| 50 | 0.9020 | **0.9710** |
| 100 | 0.8544 | **0.9640** |

Table 4: GPT Neo ("GPT") — AUC-ROC. Proposed results in **bold**.

# 5 References


1. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention Is All You Need*. NeurIPS, 2017.

2. J. Binkowski, D. Janiak, A. Sawczyn, B. Gabrys, and T. Kajdanowicz, *Hallucination Detection in LLMs Using Spectral Features of Attention Maps*. arXiv preprint, 2024.

3. G. Sriramanan, S. Saha, S. Bharti, P. Kattakinda, V. S. Sadasivan, and S. Feizi, *LLM-Check: Investigating Detection of Hallucinations in Large Language Models*. NeurIPS, 2024.

4. W. Su, C. Wang, Q. Ai, Y. Hu, Z. Wu, Y. Zhou, and Y. Liu, *Unsupervised Real-Time Hallucination Detection Based on the Internal States of Large Language Models*. Findings of ACL, 2024.


# 6 Conclusion and Future Work

## 6.1 Conclusion

This study demonstrates that complementing spectral (Laplacian eigenvalue) features with compact per-head statistical signals substantially improves hallucination detection performance. The fused 1024-d embedding (Lap512 + Stat512) yields higher accuracy and AUC-ROC across both TinyLlama and GPT Neo variants on HaluEval.

## 6.2 Future works

Potential extensions include:

- **From detection to mitigation:** Integrate a mitigation stage (e.g., RAG-based regeneration or self-refinement) to correct hallucinated outputs, and measure reduction in hallucination rate.
- **End-to-end integration:** Build a detect $\rightarrow$ mitigate $\rightarrow$ re-detect pipeline to automatically improve final outputs.
- **Cross-domain evaluation:** Test generalization across other datasets and tasks (QA, summarization, instruction-following).
- **Explainability:** Provide per-output explanations (which heads/layers caused the detection) to assist human reviewers.