Spectral-Check: A Graph-Theoretic Framework for Joint Cross-Modal and Cross-Lingual Hallucination Detection

Saim, Abdurrehman, Muneeb

November 25, 2025

Abstract

Large Language Models (LLMs) and Multimodal LLMs (MLLMs) are prone to hallucinations—outputs that are factually incorrect or ungrounded in the provided context. While recent works such as *LLM-Check* have proposed efficient detection mechanisms using scalar properties of attention maps, these methods often oversimplify the structural complexity of neural attention and are limited to unimodal (text-only) contexts. This report proposes a novel methodology, **Spectral-Check**, which advances the state-of-the-art by integrating Spectral Graph Theory with hallucination detection. By interpreting the attention mechanism as a directed graph and analyzing the eigenvalues of its Laplacian matrix, we propose to detect hallucinations as structural breakdowns in information flow. Furthermore, we extend this detection capability to the joint cross-modal and cross-lingual settings, addressing the "Curse of Multi-Modalities" where current scalar detectors fail.

1 Introduction

The reliability of Large Language Models is severely compromised by hallucinations. Existing lightweight detection frameworks, most notably *LLM-Check* (Sriramanan et al., 2024), have demonstrated that internal model activations contain signals indicative of truthfulness. However, *LLM-Check* relies on scalar aggregations of probability distributions (e.g., summing log-probabilities), which fails to capture the topological relationships between tokens.

Simultaneously, the deployment of models in real-world scenarios increasingly demands proficiency in Cross-Lingual and Cross-Modal tasks. Recent benchmarking by *CCHall* indicates that hallucinations are exacerbated when models must align visual encoders with multilingual textual decoders.

This research proposes a unified methodology that supersedes the scalar limitations of the base paper by adopting a graph-theoretic approach. We hypothesize that hallucinations manifest as spectral irregularities in the attention graph. By replacing the simple log-determinant metrics with **Laplacian Eigenvalue analysis**, and applying this to a Vision-Language proxy model, we aim to provide a more robust, theoretically grounded detection framework for modern AI systems.

2 Review of Base Methodology: LLM-Check

The foundational study, *LLM-Check*, premises that an LLM's internal uncertainty correlates with hallucination. It operates in a white-box or black-box setting (using a proxy model) to extract internal features without generating multiple expensive sample responses.

2.1 The Attention Score Mechanism

The core methodology of *LLM-Check* focuses on the Self-Attention Kernel Maps. For an autoregressive model, the attention matrix A is lower-triangular. The authors propose the "Attention Score" based on the log-determinant of the kernel map.

$$\text{Score}_{\text{Attn}} = \frac{1}{m} \sum_{i=1}^{a} \log\det(\text{Ker}_i) \tag{1}$$

Where m is the sequence length and a is the number of heads. Since the matrix is triangular, the determinant is simply the product of its diagonal elements. Therefore, the score simplifies to the mean sum of the log-probabilities of the diagonal entries:

$$\log\det(\text{Ker}_i) = \sum_{j=1}^{m} \log(\text{Ker}_{jj}^{i}) \tag{2}$$

2.2 Limitations of the Base Approach

While computationally efficient, this methodology presents two significant theoretical limitations:

1. **Scalar Reduction:** By focusing solely on the diagonal (self-attention), the metric ignores the off-diagonal elements which represent the contextual dependencies between different tokens. It treats attention as a sequence of independent probabilities rather than a connected structure.
2. **Unimodal Constraint:** The architecture is designed for text-to-text tasks, failing to account for the cross-attention mechanisms present in Multimodal LLMs (MLLMs), where the alignment between visual patches and textual tokens is the primary source of hallucination.

3 Methodology Visualizations

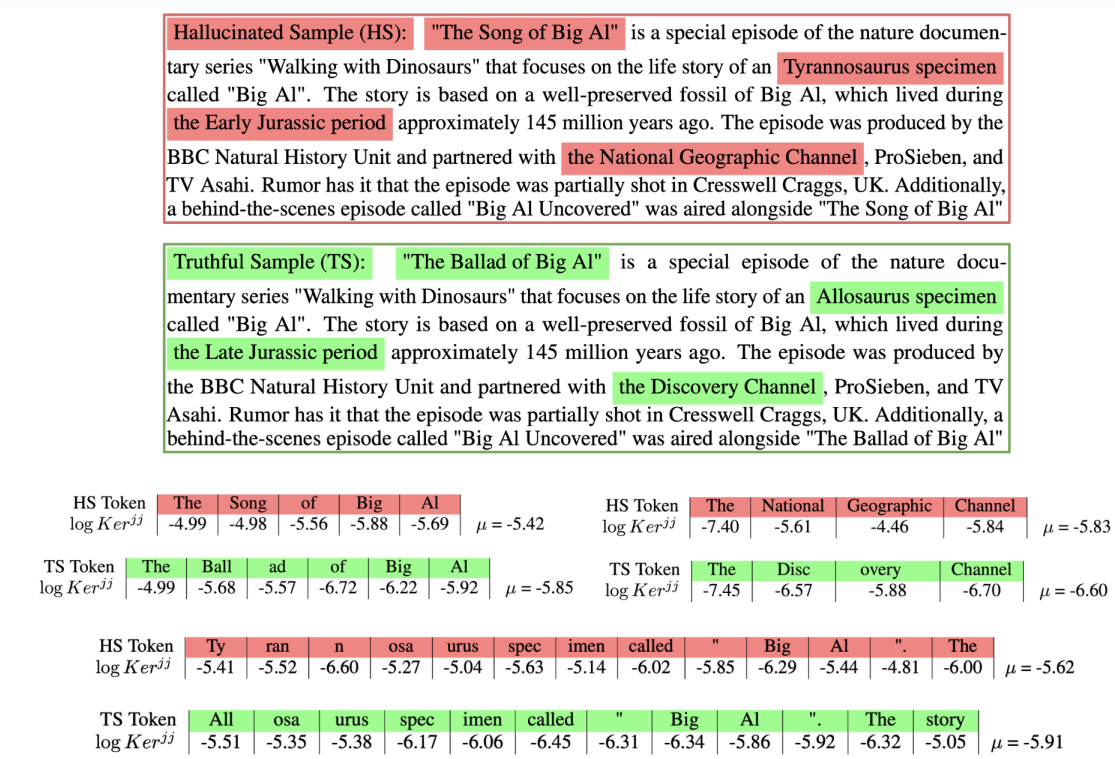

Figure 1: **Qualitative Examples:** A comparison of truthful versus hallucinated responses within the cross-modal context. The proposed framework analyzes these pairs to identify specific disconnects between the visual grounding and the textual output.

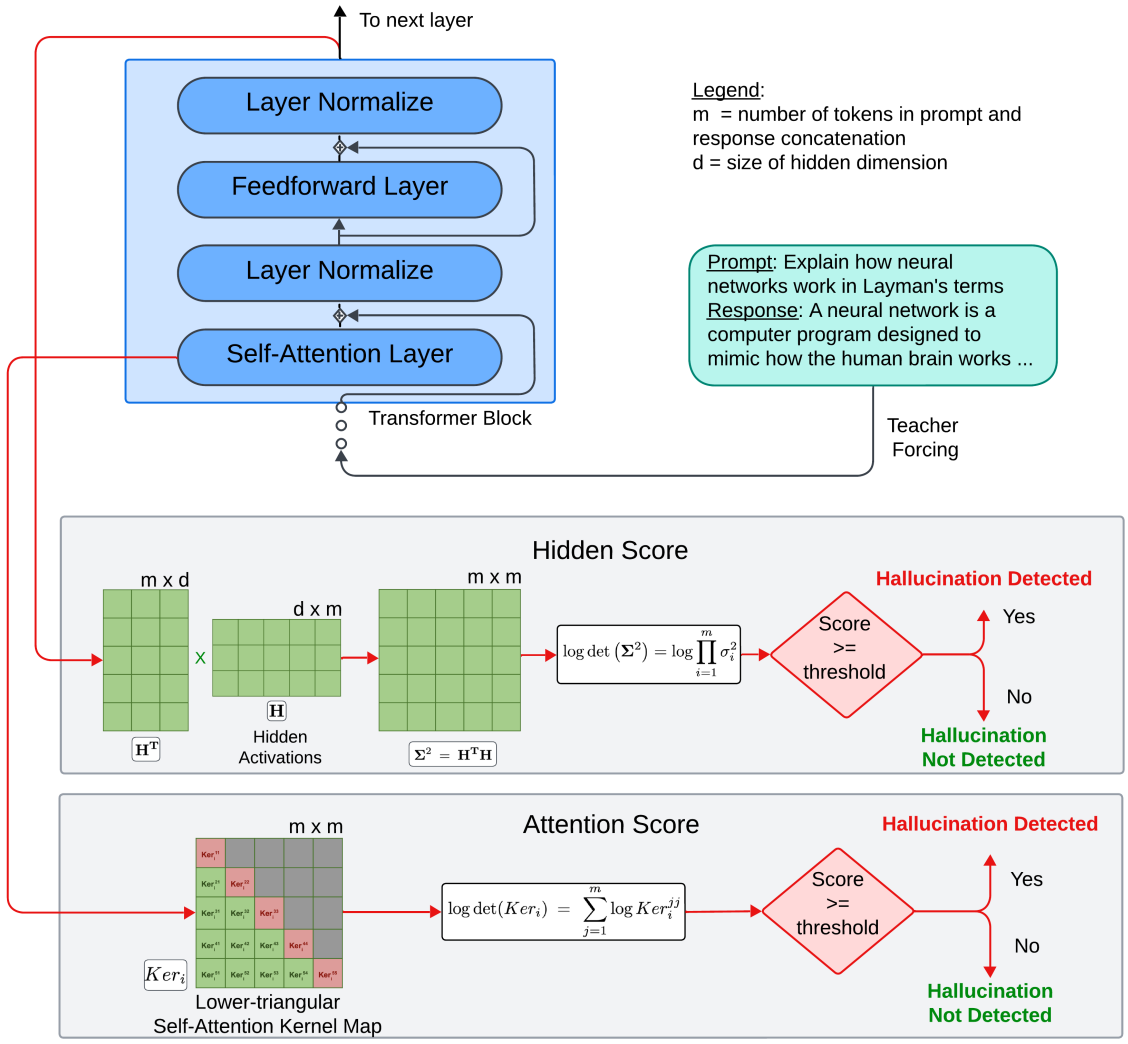

Figure 2: **Spectral-Check Pipeline:** The proposed architectural workflow. Unlike the baseline which sums scalar probabilities, our method extracts the attention graph, computes the Laplacian matrix, and utilizes the top-k eigenvalues to detect structural anomalies in the information flow.

4 Proposed Methodology: Spectral-Multimodal Detection

To address the limitations identified above, we introduce a method that transitions from *scalar probability analysis* to **Spectral Graph Analysis**, applied within a **Cross-Modal** architecture.

4.1 Theoretical Advancement: Laplacian Eigenvalues

We re-conceptualize the attention mechanism not as a matrix of probabilities, but as the adjacency matrix of a directed graph $\mathcal{G} = (V, E)$, where nodes V represent tokens (textual or visual) and edges E represent attention weights.

Instead of summing diagonal probabilities, we utilize the **Graph Laplacian**, which is known in network theory to capture the "connectivity" and "information flow" of a structure. We define the Normalized Laplacian $\mathcal{L}$ for a specific attention head h at layer l as:

$$\mathcal{L}^{(l,h)} = D^{(l,h)} - A^{(l,h)} \tag{3}$$

Where A is the attention matrix and D is the Degree Matrix (representing the total attention output of a token).

We hypothesize that hallucinations create "bottlenecks" or "disconnects" in the semantic graph. These structural anomalies are best detected by the spectrum of the Laplacian. We specifically extract the top-k eigenvalues λ:

$$\text{Feature}_{Spec} = \text{Top}_k \left(\text{eig}(\mathcal{L}^{(l,h)}) \right) \tag{4}$$

Crucially, because autoregressive attention masks are lower-triangular, the eigenvalues of $\mathcal{L}$ are equivalent to the diagonal elements of $(D - A)$. This retains the computational efficiency of the base paper (avoiding expensive SVD decompositions) while capturing significantly richer structural data regarding information flow.

4.2 Architectural Advancement: Cross-Modal Proxy

We elevate the scope of detection from text-only to joint Cross-Modal/Cross-Lingual scenarios. The base *LLM-Check* uses a text-only proxy (e.g., Llama-2). Our proposed framework substitutes this with an open-weights Vision-Language Model (VLM), such as *Qwen2-VL* or *Llama-3.2-Vision*.

The detection pipeline is modified as follows:

1. **Input:** An Image I, a Multilingual Query Q_{multi}, and the generated Response R.
2. **Cross-Lingual Anchoring:** Following findings from the CCHall benchmark, we inject an English translation anchor into the system prompt. This stabilizes the semantic space of the proxy model, enhancing the separability of truthful vs. hallucinated patterns.
3. **Spectral Extraction:** We extract the Laplacian Eigenvalues specifically from the Cross-Attention layers (where text tokens attend to image embeddings). A spectral anomaly here indicates a detachment of the generated text from the visual ground truth.

4.3 Summary of Improvements

Table 1 summarizes the methodological shift.

Feature	Base Methodology (LLM-Check)	Proposed Methodology
Math Basis	Scalar Probabilities (LogDet)	Spectral Graph Theory (Laplacian)
Detection Signal	Token Confidence	Information Flow / Connectivity
Scope	Text-Only	Joint Cross-Modal & Cross-Lingual
Proxy Model	Text LLM (e.g., Llama-2)	VLM (e.g., Qwen2-VL)

Table 1: Comparison of the Base vs. Proposed Methodologies.

5 Conclusion of Proposal

By integrating the spectral precision of the Laplacian method with the rigorous domain requirements of the CCHall benchmark, this research moves beyond simple uncertainty estimation. We propose a robust, structure-aware framework capable of detecting complex hallucinations where models decouple from visual or linguistic grounding, offering a significant qualitative leap over the baseline.