CausalGaze: Unveiling Hallucinations via Counterfactual Graph Intervention in Large Language Models

Linggang Kong[1,2], Lei Wu[1,2], Yunlong Zhang[1,2], Xiaofeng Zhong[1,2]
Zhen Wang[1,2], Yongjie Wang[1,2]*, Yao Pan[3]

[1]College of Electronic Engineering, National University of Defense Technology
[2]Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation
[3]Institute of Computer Application, China Academy of Engineering Physics
{konglinggang,wangyongjie17}@nudt.edu.cn

Abstract

Despite the groundbreaking advancements made by large language models (LLMs), hallucination remains a critical bottleneck for their deployment in high-stakes domains. Existing classification-based methods mainly rely on static and passive signals from internal states, which often captures the noise and spurious correlations, while overlooking the underlying causal mechanisms. To address this limitation, we shift the paradigm from passive observation to active intervention by introducing CausalGaze, a novel hallucination detection framework based on structural causal models (SCMs). CausalGaze models LLMs' internal states as dynamic causal graphs and employs counterfactual interventions to disentangle causal reasoning paths from incidental noise, thereby enhancing model interpretability. Extensive experiments across four datasets and three widely used LLMs demonstrate the effectiveness of CausalGaze, especially achieving 3.3% AUROC improvement in AUROC on the TruthfulQA dataset compared to state-of-the-art baselines.

1 Introduction

Large language models (LLMs) excel at various natural language generation and reasoning tasks, yet hallucination, where models generate plausible but factually incorrect content, remains a significant barrier for real-world deployment (Chakraborty et al., 2025). The prevalence of hallucination fundamentally undermines the trustworthiness and reliability of LLM-based systems, necessitating effective and robust hallucination detection and mitigation mechanisms (Huang et al., 2025; Shi et al., 2025; Zhang et al., 2026). Efforts to detect hallucinations mainly include retrieval-based methods (Heo et al., 2025; Chekalina et al., 2025), which require searching external knowledge bases,

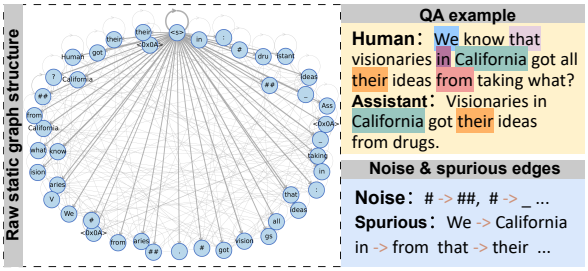


Figure 1: An question-answering example with noise and spurious correlations in graph structure. The nodes and edges are from the raw hidden states and attention maps, respectively.

while consistency-based and logits-based methods often perform multiple inferences for consistency and entropy calculations (Manakul et al., 2023; Farquhar et al., 2024). In contrast, classification-based methods utilizes the semantic features of internal states, requiring only a single generation pass without relying on any external sources (Chen et al., 2024).

Existing hallucination detection methods based on internal states typically train the classifiers either on hidden states or attention maps (Zhang et al., 2023; Chuang et al., 2024). And these methods have demonstrated effective performance across various datasets and model architectures, as shown in multiple studies (Su et al., 2024; Du et al., 2024; Sriramanan et al., 2024; Zhang et al., 2025). To jointly capture the semantic features and token dependencies, recent work has explored static graph structures (Kong et al., 2025a), where the hidden states and attention maps serve as the nodes and edges, respectively. While the graph-based classifier has superior performance, it is inherently susceptible to capturing noise and spurious correlations from the raw graph structure as shown in Figure 1. Such vulnerabilities can lead the classifier to learn and propagate incorrect dependencies. Furthermore, the utility and generalization of the

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 39015–39028

hallucination detector are significantly undermined by the indiscriminate aggregation of information without a causal basis. The fundamental challenge lies in that this passive observation paradigm fails to differentiate structurally robust knowledge from fragile associative patterns (Karbasi et al., 2025).

To overcome these limitations, we propose a paradigm shift from passive observation to active intervention to accurately trace the causal path between the internal information and model outputs. We posit that factual content is structurally robust, whereas hallucination stems from noise associations, thus highly sensitive to micro-structural interventions. Based on this insight, we propose CausalGaze, a novel hallucination detection framework based on structural causal models (SCMs) (Lu et al., 2025). Specifically, we model the LLMs' hidden states and attention maps as dynamic causal graphs and employ gradient-guided counterfactual intervention in the raw graph structure. To the best of our knowledge, CausalGaze is the first work to introduce active causal intervention to address the passive attribution in hallucination detection tasks, offering a novel solution with causal interpretability in this field. Our main contributions are summarized as follows:

- We propose CausalGaze, a novel hallucination detection framework that models internal states as dynamic causal graphs. We are the first to use the gradient-guided counterfactual intervention to estimate the causal sensitivity of attention edges (∇A), disentangling causal dependencies from spurious connections.
- We introduce a method to derive interpretable causal subgraphs by integrating node gradients (∇H) with the causally refined edges, providing fine-grained insights into the causal origins and paths that lead to hallucinations.
- We evaluate CausalGaze across four datasets and three widely used LLMs and compare the performance with baseline methods. The results demonstrate the significant effectiveness of CausalGaze, achieving over 3.3% improvement on the TruthfulQA dataset.

2 Related Work

Hallucination Detection. The remarkable generative capacity of Large Language Models (LLMs) has enabled their widespread application in knowledge-intensive and reasoning tasks (Chakraborty et al., 2025). However, LLM hallucination remains a critical bottleneck, hindering their deployment in high-stakes domains (e.g., healthcare (Tan et al., 2025), finance (Seo et al., 2024), and cybersecurity). Hallucination severely compromises model trustworthiness, motivating extensive efforts to detect it (Li et al., 2025). Existing methods for hallucination detection are primarily categorized into two types. Black-box approaches rely on external knowledge checking (e.g., search engines or RAG-based verification) (Heo et al., 2025; Nonkes et al., 2024) or consistency checks (e.g., self-consistency) (Kong et al., 2025b). While these methods are easy to implement, their effectiveness is limited by the correctness and completeness of external knowledge sources and fail to provide root-cause analysis for model internal errors (Manakul et al., 2023). In contrast, white-box methods aim to provide deeper and sourced insights by analyzing the LLMs' internal states. Prior work has mainly focused on the model's logical outputs (Ren et al., 2023a) or latent space, employing techniques such as entropy calculation (Kuhn et al., 2023; Farquhar et al., 2024), feature extraction (Binkowski et al., 2025), clustering (Sriramanan et al., 2024), and classifier training (Azaria and Mitchell, 2023; Su et al., 2024). Critically, these methods often capture noise and spurious correlations in the raw information, and only identify superficial correlational relationships. Therefore, we are the first to detect hallucination by performing active interventions from the causal perspective.

Graph Causal Learning. The interpretability of Graph Neural Network (GNN), such as GNNExplainer (Ying et al., 2019), PGExplainer (Luo et al., 2020), typically uses techniques like gradients or perturbations to identify the most influential subgraphs for the prediction. However, these tools are inherently post-hoc and attributional (Kong et al., 2025a). The primary goal is to explain why the model predicts, rather than serving as an intrinsic optimization mechanism to correct how the model reasons and thinks. We draw upon the concept of gradient sensitivity to estimate causality but repurpose it to drive a learnable structural intervention module. Furthermore, integrating it directly into the framework is a key direction for enhancing model robustness. The applications typically focus on mitigating confounding and selection biases in graph datas with fixed and real-world topologies (e.g., social or knowledge graphs) (Dong et al., 2025). In this paper, we extend this paradigm to a

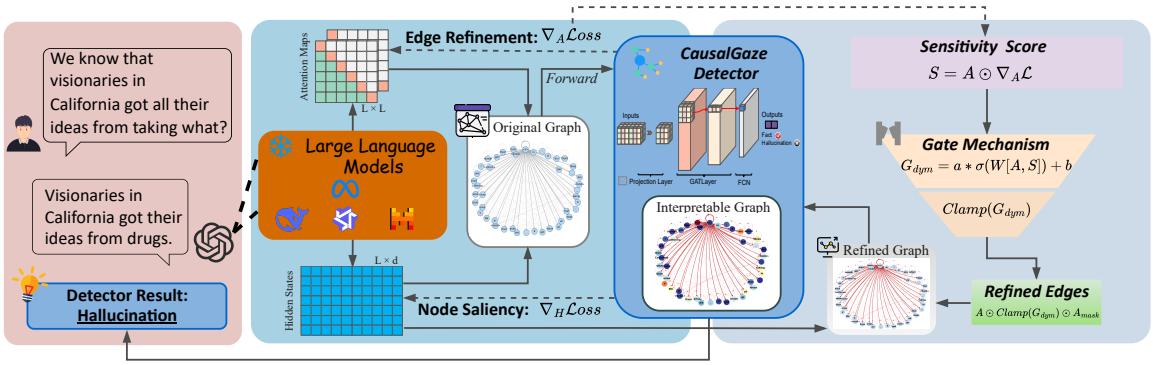


Figure 2: Overall framework of the proposed CausalGaze. We first employ the gradient-guided counterfactual intervention to obtain the refined edges, then the hidden states and the refined edges are together passed to the GNN-based detection module for the final hallucination detection result.

novel domain: graph-structure information generated by LLMs, and offer an innovative and interpretable solution for LLM hallucination detection tasks.

3 Methodology

The previous work (Kong et al., 2025a) has found that weighted directed graphs $G = (V, E)$ can effectively integrate token semantic information and their dependencies, where the nodes V are defined by the sequence of hidden states $H \in \mathbb{R}^{L \times d}$ (L is the sequence length, d is the dimension), and the edges E are represented by the attention maps $A \in \mathbb{R}^{L \times L}$. The detection model learns the correlation between the graph G and the label Y to predict $P(Y|G)$ with competitive performance. Nevertheless, it is hindered because the raw attention maps A contain massive unrelated connections, fostering illusory learning patterns with noise.

To disentangle these factors, we formulate a structural causal model (SCM) represented by the directed acyclic graph: $C \rightarrow A \rightarrow Y$ and $C \rightarrow Y$. Where C denotes the confounder, A is the mediator, i.e., the attention maps, Y is the label. Our objective is to estimate the causal effect of the mediator on the prediction, denoted as $P(Y|H, do(A))$, rather than the mere observational connections.

3.1 Problem setup

Following the core assumption of SCM, our central hypothesis is that factual reliability corresponds to structural robustness. Conversely, hallucinations are rooted in fragile and irrelevant associations. Therefore, we transform the static weighted directed graph G into a dynamic causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_{causal})$. The nodes $\mathcal{V}$ are still defined by the sequence of hidden states H, while the edges $\mathcal{E}_{causal}$ are denoted by the causal attention maps $\tilde{\mathbf{A}}$. And the hallucination detection task is framed as a binary classification problem, and the goal is to learn a causal mapping f, as shown in Equation 1:

$$y = f(\mathcal{G}), \quad y \in \{0, 1\} \tag{1}$$

where y denotes the predicted label, $y = 0$ represents 'fact', and $y = 1$ represents 'hallucination'.

3.2 CausalGaze Framework

The proposed CausalGaze framework achieves the paradigm shift from passive observation to active intervention through a dual-mechanism pipeline as illustrated in Figure 2.

Edge Refinement: We employ the gradient-guided counterfactual intervention to compute the causal sensitivity of each edge A_{ij}, which drives a learnable gating mechanism to generate a refined and causal edges $\tilde{\mathbf{A}}$. The hidden states H and the refined edges $\tilde{\mathbf{A}}$ are then passed to a downstream GNN-based detection module $D(\cdot)$ for the final classification.

Node Saliency: We utilize the node gradient (∇H) as a complementary mechanism to identify the most salient tokens contributing to the final decision. The most salient tokens and refined edges are jointly to obtain the token-level interpretable causal subgraphs.

3.2.1 Gradient-Guided Counterfactual Intervention

The core of CausalGaze is the counterfactual intervention mechanism, which actively disentangles semantic edges from noise and spuriousness. However, calculating the exact causal effect via *do*-calculus requires traversing all possible confounders, which is computationally intractable in

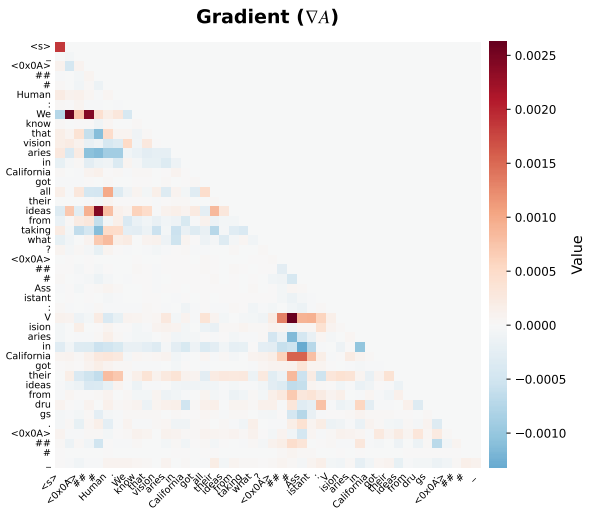


Figure 3: Visualization of the gradients for the example in Figure 1.The noise connection from token '*#*' to token '*##*' and spurious connection from token '*We*' to token '*California*' have a near-zero value.

high-dimensional model spaces. Therefore, we propose to use the gradient-guided counterfactual intervention approximation. We define the causal sensitivity of an edge $(j \to i)$ as the magnitude of change in the loss $\mathcal{L}_{Detector}$ under a microscopic intervention on the edge weight A_{ij}. This corresponds to the individual treatment effect (ITE) in a local neighborhood, which could be considered as a perturbation ϵ on the weight of edge A_{ij}. By applying the first-order $Taylor$ expansion, the change in the loss $\mathcal{L}_{Detector}$ of the hallucination detector can be approximated, as shown in Equation 2:

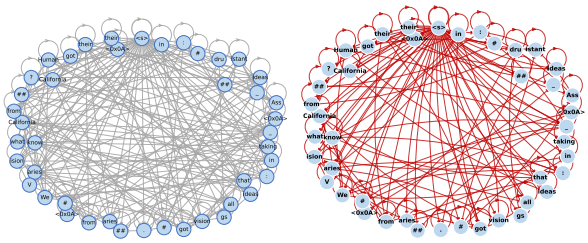
$$\mathcal{L}(A_{ij} + \epsilon) - \mathcal{L}(A_{ij}) \approx \epsilon \cdot \frac{\partial \mathcal{L}_{Detector}}{\partial A_{ij}} \quad (2)$$

The gradient $\frac{\partial \mathcal{L}_{Detector}}{\partial A_{ij}}$ indicates the direction of steepest descent, which does not account for the magnitude of the information flow, as dipicted in Figure 3. Thus, we define the causal sensitivity matrix $S \in \mathbb{R}^{N \times N}$ via the hadamard product of the edge weights and their gradients:

$$S = |A \odot \nabla_A \mathcal{L}_{Detector}| \quad (3)$$

where $\odot$ denotes element-wise multiplication. A spurious or noise edge might have a high weight due to position bias, but a near-zero gradient for irrelevance to reasoning. Conversely, a causal and factual edge should possess both significant weight and high gradient sensitivity.

To transform the raw sensitivity S into actively intervened edges $\tilde{\mathbf{A}}$, we introduce a learnable



(a) Raw graph (b) Refined graph

Figure 4: The comparison of graph structures before and after the intervention for the example in Figure 1. (a) The edges in raw graph are from the observation attention maps. (b) The edges in refined graph are from the actively intervention attention maps.

causal refinement layer (CRL). The CRL employs a dynamic gating mechanism that not only suppresses noise but also amplifies crucial, yet potentially weak and causal links. The dynamic gate mechanism is a shallow network MLP with both the original attention maps A and causal sensitivity S as input:

$$G_{dym} = a * \boldsymbol{\sigma}(MLP([A, S])) + b \quad (4)$$

where $[\cdot, \cdot]$ is the concatenation operation, $\boldsymbol{\sigma}$ denotes the sigmoid function and a, b are the learnable scaling factors. The final refined edges $\tilde{\mathbf{A}}$ is obtained as follows:

$$\tilde{\mathbf{A}} = A \odot Clamp(G_{dym}) \odot A_{mask} \quad (5)$$

where $Clamp(\cdot)$ is the clamp function to further enhance the intervention and A_{mask} is the autoregressive mask of A. The actively intervened token dependencies $\tilde{\mathbf{A}}$ mitigates noise and spurious connections in the raw information. And the comparison of graph structures before and after the intervention is shown in Figure 4, revealing that the refined graph is more sparsely connected.

3.2.2 CausalGaze Detector

The Graph Attention Network (GAT) backbone is deployed on the refined causal graph $\tilde{\mathcal{G}} = (\mathcal{V}, \mathcal{E}_{causal})$ to aggregate information, which integrates the refined edges as explicit structural bias. The update of node i at layer k is given in Equation 6:

$$\mathbf{h}_i^{(k+1)} = \mathbf{h}_i^{(k)} + \sum_{j \in \mathcal{N}(i)} \tilde{\mathbf{A}}_{ij}^{(k)} \mathbf{W}^{(k)} \mathbf{h}_j^{(k)} \quad (6)$$

where $\mathbf{W}^{(k)}$ is the learnable parameters and multiple GAT layers with residual connections is stacked

to facilitate deep message passing. To capture both the most salient local features and the global context, we employ a hybrid pooling strategy that concatenates the results of global max pooling and mean pooling:

$$\mathbf{h}_{graph} = \left[\max_{i \in \mathcal{V}} \mathbf{h}_i^{(K)}, \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \mathbf{h}_i^{(K)} \right] \quad (7)$$

The final representation $\mathbf{h}_{graph}$ is input into an MLP layer for classification. The objective combines the standard binary classification loss with a regularization term that promotes sparsity and coherence in the causal structure:

$$\mathcal{L}_{Detector} = \mathcal{L}_{CE}(y, \hat{y}) + \lambda \cdot \|S\|_F^2 \quad (8)$$

where $\mathcal{L}_{CE}$ is the primary cross-entropy loss between the true label y and the prediction $\hat{y} = f(\mathcal{G})$, λ is applied to the sensitivity matrix S to encourage the model to find the most compact, sparse set of causally important edges.

3.3 Interpretable causal subgraphs

To further elaborate the interpretability of hallucination detection model CausalGaze, we complement the structural analysis with node-level attribution. We utilize the gradient of the hidden states (∇H) as a diagnostic lens to calculate the importance scores of each token without altering the semantic features. The node saliency score N_s^i for the $i\text{-}th$ token is calculated as the L_2 norm of the gradient vectors:

$$N_s^i = \|\nabla_{H_i} \mathcal{L}_{Detector}\|_2, \quad i \in [1, L] \quad (9)$$

where H_i denotes the hidden states of the $i\text{-}th$ token. The interpretable causal subgraphs could be obtained by combining the refined edges $\tilde{\mathbf{A}}$ and the salient nodes $\tilde{\mathbf{N}}$, which allows for a fine-grained diagnosis to illustrate the causal nodes and paths from the token-level perspective.

4 Experimental Settings

4.1 Models

We evaluate the detection performance of CausalGaze using three mainstream open-source LLMs, including Llama-2-7B (Touvron et al., 2023), Qwen2-7B (Yang et al., 2024), and Mistral-7B (Jiang et al., 2023). Details of the settings of these models are in Appendix A.1

4.2 Datasets

The models are evaluated on four widely used datasets: TruthfulQA (Lin et al., 2022), which is a open-domain question answering dataset; TriviaQA (Joshi et al., 2017), which tests general knowledge; SicQ (Welbl et al., 2017), designed for domain-specific question answering; and HaluEval (Li et al., 2023), specially designed for hallucination detection tasks.

4.3 Baselines

We compare our approach against several types of competitive baselines, categorized as follows: (1) **Consistency-based** approaches: SelfcheckGPT (black-box) (Manakul et al., 2023) and EigenScore (white-box) (Chen et al., 2024); (2) **Logit-based** approaches: Perplexity (Ren et al., 2023b), Length-Normalized Entropy (LN-Entropy) (Malinin and Gales, 2021), and Semantic Entropy (Kuhn et al., 2023); (3) **Self-evaluation** approach: P(True) (Kadavath et al., 2022); and (4) **Classification-based** approaches: SAPLMA (Azaria and Mitchell, 2023), LLM-Check (Sriramanan et al., 2024), ICR Probe (Zhang et al., 2025) and HaluGNN (Kong et al., 2025a). All comparison methods were implemented using the experimental parameters specified in their respective original papers. More details are shown in Appendix A.3.

4.4 Evaluation Metric

Following previous established research, we employ the Area Under the Receiver Operating Characteristic Curve (AUROC) as the primary evaluation metric to assess the performance of all approaches (Azaria and Mitchell, 2023; Sriramanan et al., 2024; Zhang et al., 2025). AUROC represents the area under the ROC curve, which illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR). Furthermore, the F1-Score is also selected as an essential evaluation metric, offering a balanced measure of the model's performance on classification tasks.

4.5 Implementation Details

The CausalGaze model employs a multi-stage GNN architecture optimized for high-dimensional sequential feature processing. Specifically, the high-dimensional node features are first projected down to a 128-dimensional hidden space via a dedicated projection Layer to reduce the computational overhead and parameter counts. Prior to message passing, we utilize a causal refinement layer to

LLMs	Methods	TruthfulQA		TriviaQA		SciQ		HaluEval	
		AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
Llama2-7B	SelfCheckGPT	0.5295	0.5345	0.7322	0.7296	0.6790	0.6830	0.6670	0.6591
	EigenScore	0.5193	0.5286	0.7398	0.7443	0.5960	0.6008	0.5960	0.5880
	Perplexity	0.5677	0.6034	0.7213	0.6995	0.5260	0.5256	0.5260	0.5020
	LN-Entropy	0.6151	0.6187	0.7091	0.7332	0.5760	0.5826	0.6678	0.6544
	Semantic Entropy	0.6217	0.6145	0.7321	0.7209	0.6820	0.6797	0.6820	0.6865
	P(True)	0.5181	0.5556	0.5568	0.5704	0.5460	0.5645	0.5648	0.6150
	SAPLMA	0.7820	0.7903	0.8310	0.8520	0.7730	0.7767	0.7738	0.6936
	LLM-Check	0.6160	0.5926	0.5551	0.5857	0.5842	0.5637	0.5683	0.5450
	ICR-Probe	0.8142	0.7858	0.8001	**0.7940**	0.7748	0.7561	0.8346	**0.8032**
	HaluGNN	0.8803	0.7714	0.8531	0.7863	0.8336	0.7649	0.9141	0.6316
	CausalGaze	**0.8805±0.0102**	**0.8189±0.0510**	**0.8614±0.0110**	0.7786±0.0211	**0.8390±0.0081**	**0.7701±0.0420**	**0.9280±0.0204**	0.8001±0.0210
Qwen2-7B	SelfCheckGPT	0.6170	0.6220	0.6230	0.6143	0.5860	0.5779	0.6538	0.6327
	EigenScore	0.5370	0.5489	0.6130	0.6180	0.6320	0.6462	0.6840	0.6343
	Perplexity	0.6510	0.6630	0.5020	0.5103	0.5340	0.5477	0.5340	0.5167
	LN-Entropy	0.6670	0.6594	0.5110	0.5063	0.5240	0.5256	0.7371	0.6850
	Semantic Entropy	0.6610	0.6787	0.5870	0.5965	0.6590	0.6604	0.6634	0.6808
	P(True)	0.6370	0.6542	0.5090	0.5256	0.5380	0.5449	0.5580	0.5952
	SAPLMA	0.8170	0.8280	0.8200	0.8315	0.8150	0.8200	0.7799	0.7884
	LLM-Check	0.6316	0.6006	0.5552	0.5470	0.5726	0.5773	0.5292	0.5398
	ICR-Probe	0.7937	0.7740	0.7684	0.7560	0.7595	0.7215	0.8003	0.7730
	HaluGNN	0.8392	0.7706	0.9050	0.8245	0.9217	**0.8402**	0.9065	0.8058
	CausalGaze	**0.8680±0.0311**	**0.8071±0.0102**	**0.9106±0.0110**	**0.8280±0.0540**	**0.9328±0.0021**	0.8283±0.0052	**0.9220±0.0231**	**0.8360±0.0320**
Mistral-7B	SelfCheckGPT	0.5771	0.5345	0.6340	0.6145	0.5593	0.5260	0.6729	0.6357
	EigenScore	0.6012	0.5860	0.6573	0.6270	0.6474	0.6346	0.5950	0.5659
	Perplexity	0.5538	0.5421	0.5740	0.5532	0.5470	0.5583	0.5525	0.5244
	LN-Entropy	0.5763	0.5720	0.5834	0.5682	0.6036	0.5832	0.7156	0.6580
	Semantic Entropy	0.6557	0.6136	0.6063	0.5770	0.6945	0.6640	0.6685	0.7065
	P(True)	0.5260	0.5125	0.5680	0.5560	0.5673	0.5531	0.5739	0.5643
	SAPLMA	0.8112	0.7883	0.8290	0.7984	0.7884	0.7790	0.7880	0.7729
	LLM-Check	0.6732	0.6540	0.5417	0.5411	0.5748	0.5547	0.5373	0.5870
	ICR-Probe	0.7993	**0.8000**	0.7325	0.7258	0.7793	0.7826	0.8047	0.8120
	HaluGNN	**0.8647**	0.7747	0.8735	0.8000	0.8343	0.7273	0.8647	0.8030
	CausalGaze	0.8851±0.0092	0.7971±0.0075	**0.8880±0.0621**	**0.8113±0.0210**	**0.8550±0.0023**	**0.7860±0.0041**	**0.9127±0.0382**	**0.8292±0.0273**

Table 1: Main results of AUROC and F1-Score compared with different competitive baselines over diverse LLMs on TruthfulQA, TriviaQA, SciQ and HaluEval datasets. The best results are highlighted in bold.

obtain the dynamically refined adjacency matrix. The 128-dimensional features are processed by two stacked multi-head GAT layers (with 4 heads) to adaptively learn causal relationships. A hybrid pooling layer concatenating the global max pooling and mean pooling forms a 256-dimensional embedding, which is passed through an MLP classifier for binary prediction. The binary cross-entropy loss with a regularization term is applied to the sigmoid of the model outputs.

5 Experimental Results and Analysis

5.1 Main Results

The main experimental results are demonstrated in Table 1, showing that the CausalGaze has competitive performance against most baseline methods on most datasets and diverse LLMs.

Specifically, the experimental results indicate that classification-based approaches substantially outperform other baseline methods. This superiority is primarily because baseline methods rely exclusively on the LLM's intrinsic mechanisms, whereas classification-based approaches leverage an additionally trained detection probe. While existing classification-based methods, such as SAPLMA, LLM-Check, and ICR Probe, utilize either hidden states or attention maps in isolation and yield sound detection performance, HaluGNN effectively achieves 5-7% improvement by coupling both hidden states and attention maps into a graph structure. Nevertheless, HaluGNN directly utilize attention maps as graph edges, inherently introducing a substantial amount of noise and irrelevant connections, as illustrated in Figure 1 and 3. This limits the classifier's capacity to learn beyond superficial correlational relationships. In contrast, our proposed CausalGaze framework not only mitigates this issue by substantially reducing noise and spurious connections but also fortifies the underlying causal relationship between the generated content and the factual or hallucinated labels. Compared with HaluGNN, our approach achieves over 3.3% improvement in AUROC on the TruthfulQA dataset. More details are shown in Appendix B.1.

5.2 Generalization Analysis

To assess the cross-dataset generalization capability, we train each model on one dataset and evaluate its performance on all other datasets. The results are visualized as a heatmap in Figure 5, where Each cell in the heatmap reports the F1-Score, sharing the same color scale (ranging from 0.5 to 0.9) for

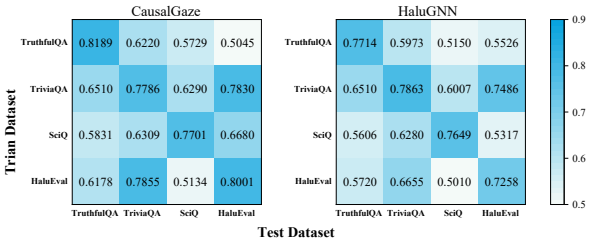


Figure 5: Cross-dataset generalization analysis for CausalGaze and HaluGNN. Each subplot displays the F1-Score when the model is trained on the row dataset and tested on the column dataset, with values annotated in each cell.

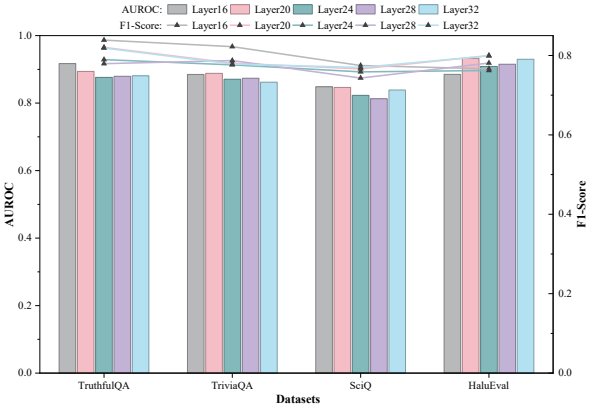


Figure 6: The impact of causal graphs from different layers of Llama2-7B on the detection performance.

comparison.

On unseen target datasets, our proposed CausalGaze consistently and significantly outperforms HaluGNN. Furthermore, the CausalGaze evaluation on target-domain improves the average F1-Score by 3.6% compared to HaluGNN, which demonstrates the superior robustness of CausalGaze under domain shift. This notable advantage stems from CausalGaze capturing factual and hallucinated patterns from the causal perspective.

5.3 Ablation Experiments and Interpretability

We conduct ablation experiments on Llama-2-7B (Touvron et al., 2023) model, mainly studying three aspects: (1) The contribution of projection layer and its dimension to the detection performance; (2) The effect of different layers on the detection performance; (3) The effect of gradient-guided information. More details are shown in Appendix B.

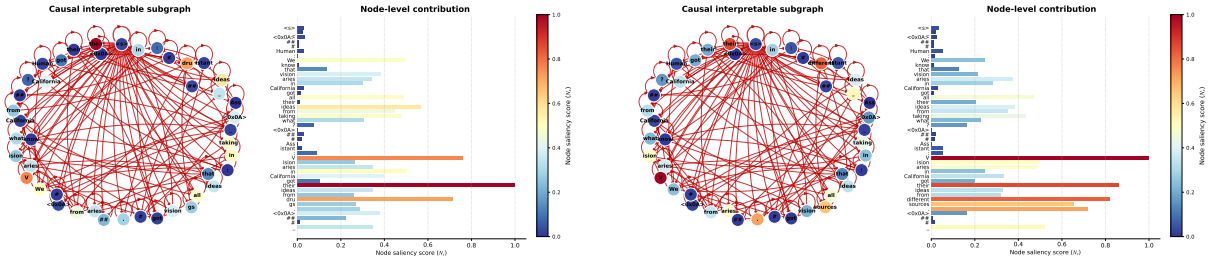
Contribution of the Projection Layer and its Dimension. While the projection layer is introduced to the CausalGaze architecture primarily to decrease the high-dimensional node features, it is necessary to experimentally investigate the impact of both the dimensionality reduction operation itself and the choice of the compressed dimension on the detection performance. Given that the padded hidden state dimension of Llama2-7B model is 4096, we conduct an ablation study to test four representative dimensions: 64, 128, 256, and 512. The experimental results, summarized in Table 2, confirm that the dimensionality reduction strategy is absolutely effective for improving detection performance. Furthermore, the dimension of 128 strikes an optimal balance, achieving the competitive detection performance while maintaining a relatively minimal model parameter count.

Layer Number. Prior literature suggest that LLMs gradually capture and understand context, and ensure token fluency primarily within the initial layers. Subsequently, the middle and later layers are responsible for knowledge integration, next-token generation, and implicitly contain more causal information related to factual consistency. To investigate the optimal source for the proposed CausalGaze, we conduct an ablation study using causal graphs extracted from five layers of the Llama2-7B model: 16, 20, 24, 28, and 32. The results, as illustrated in Figure 6, indicate that features derived from Layer 20 yield the best performance. This finding aligns with existing research, where critical semantic and factual consistency features relevant to the generated content are primarily found in the middle-to-later layers of the LLMs (Azaria and Mitchell, 2023; Sriramanan et al., 2024; Kong et al., 2025a).

Contribution of the Gradient-guided Information. The gradient-guided sensitivity S is actually the cornerstone of CausalGaze. To isolate and verify its specific contribution, we conduct an additional ablation study to compare the full model with three variants that bypass the gradient-guided information: (1) w/o Gradients (S), where we replace the gradient-guided sensitivity matrix with a uniform matrix in Equation 8. It means the gating mechanism relies solely on the raw attention weights without any causal sensitivity guidance; (2) Random Gradients, where we replace with a randomly initialized matrix to test if any additional information or random perturbation could provide a similar regularization effect; (3) A learnable gating mechanism based solely on the attention maps (A) (i.e., MLP(A)). The comparison of F1-Score on all datasets and models is shown in Table 3. The significant performance drop (about 2-4%) when removing the gradients or setting ran-

Projection Layer	Dimension	TruthfulQA		TriviaQA		SciQ		Halueval	
		AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
w/o	4096	0.6069	0.6469	0.5898	0.5318	0.5889	0.6559	0.5530	0.6676
with	64	0.8712	0.7818	0.8719	0.7809	0.8301	**0.7773**	0.8989	0.6322
	128	0.8805	**0.8189**	0.8614	0.7786	**0.8390**	0.7701	**0.9280**	**0.8001**
	256	0.8413	0.7822	0.8840	**0.8011**	0.8324	0.7462	0.8837	0.7093
	512	**0.8807**	0.8018	**0.8849**	0.7956	0.8267	0.7333	0.8168	0.7698

Table 2: The contribution of the Projection Layer and its dimentions on Llama2-7B across different datasets. The best results are highlighted in bold.



(a) The causal nodes and edges for the hallucination.

(b) The causal nodes and edges for the fact.

Figure 7: The token-level and fine-grained interpretability analysis of the detection result for the hallucinated and factual responses of the same question in Figure 1.

Model	Variant	TruthfulQA	TriviaQA	SciQ	HaluEval
Llama2-7B	CausalGaze	**0.8189**	**0.7786**	**0.7701**	**0.8001**
	w/o Gradient	0.7720	0.7763	0.7549	0.6316
	Random Gradient	0.7811	0.7656	0.7319	0.6630
	MLP(A)	0.7793	0.7674	0.7540	0.7241
Qwen2-7B	CausalGaze	**0.8071**	**0.8280**	**0.8283**	**0.8360**
	w/o Gradient	0.7660	0.8240	0.8400	0.8033
	Random Gradient	0.7672	0.8252	0.8233	0.6667
	MLP(A)	0.7857	0.8246	0.8520	0.6947
Mistral-7B	CausalGaze	**0.7971**	**0.8113**	**0.7860**	**0.8292**
	w/o Gradient	0.7746	0.7992	0.7283	0.8020
	Random Gradient	0.7636	0.7927	0.7787	0.7623
	MLP(A)	0.7811	0.7908	0.7407	0.7855

Table 3: The performance comparison of CausalGaze with w/o Gradient, Random Gradient and A learnable gating mechanism based solely on the attention maps (A) (i.e., MLP(A)).

dom gradients confirms that the learned gradients S is indeed the key driver of our model's competitive performance. While the MLP(A) improves slightly over the w/o gradients, it still significantly underperforms CausalGaze. This proves that raw attention A may only capture statistical correlations, which are often the source of spurious correlations in LLM hallucinations, whereas gradients S identify the causal bottlenecks. Also, the gradient-guided gradients $\nabla_A \mathcal{L}$ reflect how a specific attention edge contributes to the final results, and an edge can have a high attention weight (statistical significance) but low causal sensitivity (not the cause of hallucination or fact). By integrating the gradient-guided information, CausalGaze specifically detects hallucination with interpretability, where a simple MLP(A) or random gradients cannot make it.

Token-level Interpretability. To address the inherent lack of interpretability in existing classification-based hallucination detection methods, CausalGaze enables fine-grained and token-level analysis grounded in causal intervention. We perform node and edge importance analysis on the Llama2-7B model using a representative example from TruthfulQA (seen in Figure 1), and the visualization is depicted in Figure 7. It can be evidently observed that the causal interpretable graphs successfully mitigate noise and prune spurious connections. Specifically, the importance scores of nodes representing key entities (such as 'Visionaries', 'ideas' and 'drugs' in Figure 7(a) or 'sources' in Figure 7(b)) are significantly higher than those assigned to non-informative tokens (such as '##', '<s>', and '<0x0A>'). Crucially, the edge from token 'ideas' to 'drugs' is severed in the hallucinated answer, whereas the edge from token 'ideas' to 'sources' remains intact in the fact answer, which is consistent with our theoretical expectations. We attribute this disparity to the fact that while such tokens may exhibit high co-occurrence in the pre-training corpus, they lack genuine causal dependencies; consequently, the models may prioritize linguistic fluency over factual accuracy, leading

to the risky guesses (Gao et al., 2025). By identifying the causal structures critical to detection, CausalGaze effectively bridges the gap in model interpretability.

6 Conclusion

In this paper, we propose CausalGaze, the first work to introduce active causal intervention to address the passive attribution in hallucination detection tasks. CausalGaze effectively suppresses the noise and spurious correlations inherent in raw representations, enabling a precise probing of the causal relationships between internal signals and model outputs. Comprehensive experimental results across multiple benchmarks demonstrate that CausalGaze not only achieves competitive performance over existing methods but also provides a robust, interpretable foundation for hallucination detection.

Limitations

While the proposed CausalGaze shows significantly effective performance and bridges the gap in model interpretability, it is subject to certain limitations. Firstly, the method necessitates access to the LLMs' latent space, which inherently restricts its application in proprietary and black-box LLMs. Secondly, our approach only models the local causal graph from a single layer, which can not capture the globally comprehensive information embedded across all layers of LLMs. And the graph computation process is resource-intensive, which limits its practical use in the detection of long-context texts. Building upon the insights, we anticipate that future research will explore more efficient and alternative signal representation paradigms, such as aggregating features from multiple layers or using summary vectors, to effectively expand the applicability and scalability of causal intervention methods for hallucination detection.

Ethics and Broader Impact

We sampled a portion of the data from existing datasets for our experiments, which may affect the accuracy of some of our conclusions.

References

Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. 2025. Hallucination detection in LLMs using spectral features of attention maps. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24365–24396.

Neeloy Chakraborty, Melkior Ornik, and Katherine Driggs-Campbell. 2025. Hallucination detection in foundation models for decision-making: A flexible definition and review of the state of the art. *ACM Computing Surveys*, 7:1–35.

Viktoriia Chekalina, Anton Razzhigaev, Elizaveta Goncharova, and Andrey Kuznetsov. 2025. Addressing Hallucinations in Language Models with Knowledge Graph Embeddings as an Additional Modality. arXiv preprint arXiv:2411.11531.

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. In *12th International Conference on Learning Representations, ICLR 2024*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. pages 1419–1436. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.

Youqiang Dong, Min Zhang, Xi Cheng, and Hai Wang. 2025. Scm-gnn: A graph neural network-based multi-antenna spectrum sensing in cognitive radio. *IEEE Transactions on Cognitive Communications and Networking*, pages 127–144.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. HaloScope: Harnessing Unlabeled LLM Generations for Hallucination Detection. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NeurIPS 2024*.

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

Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. 2025. H-neurons: On the existence, impact, and origin of hallucination-associated neurons in llms. arXiv preprint arXiv:2512.01797.

Jinwen He, Yujia Gong, Kai Chen, Zijin Lin, Chengan Wei, and Yue Zhao. 2024. LLM Factoscope: Uncovering LLMs' Factual Discernment through Inner States Analysis. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*.

Sangwoo Heo, Sungwook Son, and Hyunwoo Park. 2025. Halucheck: Explainable and verifiable automation for detecting hallucinations in llm responses. *Expert Systems with Applications*, 272:126712.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and et.al. 2023. Mistral 7b. arXiv preprint arXiv:2310.06825.

Mandar Joshi, Eunsol Choi, and Daniel S. Weld. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

Amin Karbasi, Omar Montasser, John Sous, and Grigoris Velegkas. 2025. (im)possibility of automated hallucination detection in large language models. In *CONFERENCE ON LANGUAGE MODELING, COLM 2025*.

Linggang Kong, Yunlong Zhang, Xiaofeng Zhong, Haoran Fu, Yongjie Wang, and Huijun Liu. 2025a. Halugnn: Hallucination detection in large language models using graph neural network. *Expert Systems with Applications*, page 130857.

Linggang Kong, Xiaofeng Zhong, Jie Chen, Haoran Fu, and Yongjie Wang. 2025b. Multi-perspective consistency checking for large language model hallucination detection: a black-box zero-resource approach. *Front Inform Technol Electron Eng*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *11th International Conference on Learning Representations, ICLR 2023*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Qing Li, Jiahui Geng, Zongxiong Chen, Derui Zhu, Yuxia Wang, Congbo Ma, Chenyang Lyu, and Fakhri Karray. 2025. Hd-ndes: Neural differential equations for hallucination detection in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6173–6186.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Zhengyang Lu, Bingjie Lu, and Feng Wang. 2025. Causalsr: Structural causal model-driven super-resolution with counterfactual inference. *Neurocomputing*, page 130375.

Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. 2020. Parameterized explainer for graph neural network. In *34th Conference on Neural Information Processing Systems, NeurIPS 2020*.

A. Malinin and M. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Noa Nonkes, Sergei Agaronian, Evangelos Kanoulas, and Roxana Petcu. 2024. Leveraging graph structures to detect hallucinations in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 93–104.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023a. Out-of-Distribution Detection and Selective Generation for Conditional Language Models. In *11th International Conference on Learning Representations, ICLR 2023*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023b. Out-of-Distribution Detection and Selective Generation for Conditional Language Models. In *11th International Conference on Learning Representations, ICLR 2023*.

Jean Seo, Jongwon Lim, Dongjun Jang, and Hyopil Shin. 2024. Dahl: Domain-specific automated hallucination evaluation of long-form text through a benchmark dataset in biomedicine. In *The 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*.

Xiang Shi, Jiawei Liu, Yinpeng Liu, Qikai Cheng, and Wei Lu. 2025. Know where to go: Make llm a relevant, responsible, and trustworthy searchers. *Decision Support Systems*, 188:114354.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. In *38th Conference on Neural Information Processing Systems, NeurIPS 2024*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391.

Likun Tan, Kuan-Wei Huang, and Kevin Wu. 2025. Fred: Financial retrieval-enhanced detection and editing of hallucinations in language models. In *Proceedings of the 42nd International Conference on Machine Learning, ICML 2025*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. 2025. Unconditional Truthfulness: Learning Unconditional Uncertainty of Large Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *Statistics*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. arXiv preprint arXiv:2407.10671.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. In *33rd Annual Conference on Neural Information Processing Systems, NeurIPS 2019*.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv preprint arXiv:2309.01219.

Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. 2025. Icr probe: Tracking hidden state dynamics for reliable hallucination detection in llms. In *63rd Annual Meeting of the Association-for-Computational-Linguistics, ACL 2025.*, pages 17986–18002.

A Details about Experimental Settings

A.1 Details about Models

Llama2-7B, Qwen2-7B and Mistral-7B are selected for their accessible internal latent spaces and the moderate dimensionality of the hidden states, facilitating both experimental implementation and validation. We employ the pre-trained parameters provided directly by the Hugging Face throughout our experiments. During inference, we adhere to the default generation configuration with the temperature set to 0.6, while top-k and top-p sampling set to 50 and 0.9, respectively.

A.2 Details about Datasets

To rigorously evaluate the detection performance of our CausalGaze framework across varied knowledge domains and complexities, we curate an experimental corpus from four distinct datasets. Specifically, **TruthfulQA** includes 817 question-answering (QA) pairs with testing the model's resilience against factually challenging questions (Kong et al., 2025a); **TriviaQA** consists of 9960 deduplicated QA pairs for assessing performance on a broader and more conventional knowledge base; **SciQ** contains 1000 QA pairs which are used to probe the model's capabilities in a specialized scientific domain; and **HaluEval** (QA subset) inclueds 10k QA pairs technically for hallucination detection tasks. To facilitate experimental implementation, we select 1000 samples from the aforementioned datasets, which are divided into 400 for training, 200 for validation, and the remaining for testing.

A.3 Baselines Methods

This section provides a brief introduction of the involved baseline methods for hallucination detection.

Consistency-based approaches: Consistency is a fundamental concept in logic, defined as the absence of contradictory statements within a system. To characterize the internal consistency of LLMs, consistency-based hallucination detection methods rely on multiple samplings, which facilitate zero-resource fact-checking and enable the verification of responses from arbitrary LLMs without the reliance on any external databases or evidence. **SelfCheckGPT** builds on a straightforward intuition: If LLMs possess knowledge of a given

concept, the sampled responses tend to be similar and factually consistent. Conversely, in the case of hallucinated content, randomly sampled responses are likely to diverge and contradict one another. **EigenScore** operates on the similar premise but computes consistency scores using the LLM internal states, specifically by evaluating the similarity of latent spaces across multiple samples.

Logit-based approaches: For the traditional machine learning classification models, the maximum Softmax probability indicates the confidence level of the results and has been widely used as the metric for uncertainty evaluation. Extending it to the long-sequence token generation tasks of LLMs, **Perplexity** defines the uncertainty of the generated response as the joint probability of the constituent tokens:

$$P(y|x,\Theta) = -\frac{1}{T}\sum_{t}\log p(y_t|y_{<t},x) \quad (10)$$

where x is the prompt, T denotes the length of the sequence, and $p(y_t|y_{<t},x)$ represents the maximum Softmax probability of $t\text{-}th$ token. Given that the perplexity of shorter sequences generally exhibit lower, the joint probability is normalized by the output sequence length T.

However, considering that different tokens contribute unevenly to the sentence, the average probability fails to effectively capture the uncertainty. Multiple responses can be obtained during inference via the top-p or top-k sampling strategies. **LN-Entropy** measures the uncertainty as follows:

$$H(\mathcal{Y}|x,\Theta) = -\mathbb{E}_{y\in\mathcal{Y}}\frac{1}{T_y}\sum_{t}\log p(y_t|y_{<t},x) \quad (11)$$

where $\mathcal{Y} = [y^1, y^2, ..., y^{K-1}, y^K]$ denotes K sampled responses.

The above methods assess uncertainty and entropy solely from a token-level perspective. However, measuring uncertainty in natural language is challenging due to semantic equivalence, where distinct sentences can convey identical meanings. **Semantic Entropy** addresses it by incorporating linguistic invariance arising from shared meaning. It employs semantic equivalence to cluster K responses into c classes, and then computes the Semantic Entropy as the entropy distribution over the semantic space:

$$p(c|x) = \sum_{s\in c} p(s|x) = \sum_{s\in c}\prod_{t} p(s_t|s_{<t},x) \quad (12)$$

$$SE(x) = -\sum_{c} p(c|x)\log p(c|x) \quad (13)$$

Self-evaluation approach: LLMs are also frequently utilized to assist in or directly judge the correctness of responses, known as LLM-as-a-Judge. **P(True)** performs fact-checking by querying LLMs with specific prompts and evaluating the probability that the response is correct. The prompt that we use for P(True) is as follows:

P(True) Prompt

Provide the probability that the following answer for the question is correct. Give ONLY the probability value between 0.0 and 1.0, no other words or explanation.

Question: {Question}

Answer: {Answer}

Classification-based approaches: The hallucination detectors are mainly trained on the latent space of LLMs. **SAPLMA** trains a classifier using the hidden states of specific layers in LLMs. **LLM-Check** uses the method of attention mechanism kernel similarity analysis to conduct hallucination detection. And it has proved that the differences in model sensitivity to hallucinated or truthful contents reflects in the rich semantic representations present both in hidden states and the pattern of attention maps. **ICR Probe** firstly quantifies the global contribution of modules to the hidden states' updates of all layers, which are then used to train a hallucination probe. **HaluGNN** models the hidden states and attention maps as weighted directed graphs, and train a GNN-based classifier. However, the classification-based approaches scarcely concerns the interpretability of the detection model and the results.

A.4 CausalGaze Training

Model Architecture. The CasusalGaze model consists of five layers, including a projected layer, a refinement layer, two GAT layers and a linear layer. The dimension of each layer passes through $d \to 128 \to 128 \to 64 \to 64 \to 2$. Each hidden layer except the refinement layer employs the ReLU activation function and applies the dropout ($p = 0.2$) to prevent overfitting.

Details of Training. We set the hyperparameters $a = 1$ and $b = 0$ in Equation 4, and minum value of the $Clamp$ is 0, and λ is

Methods	TruthfulQA	TriviaQA	SciQ	HaluEval
Lookback Lens	0.603 / 0.633	0.771 / 0.781	0.726 / 0.731	0.752 /0 .795
Factoscope	0.552 / 0.584	0.714 / 0.760	0.682 / 0.702	0.742 / 0.788
TAD	0.804 / 0.832	0.844 / 0.849	0.815 / 0.833	0.874 / 0.856
CausalGaze	**0.881 / 0.852**	**0.861 / 0.895**	**0.839 / 0.855**	**0.928 / 0.938**

Table 4: AUROC / PR-AUC Performance of different methods on Llama2-7B.

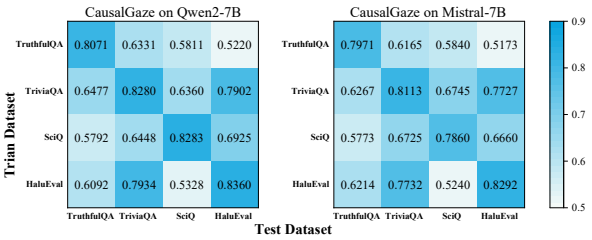


Figure 8: Cross-dataset generalization analysis for CausalGaze on Qwen2-7B and Mistral-7B. Each subplot displays the F1-Score when the model is trained on the row dataset and tested on the column dataset, with values annotated in each cell.

set to 0.02 in Equation 8. The loss function is the binary crossentropy loss with a regularization term that promotes sparsity and coherence in the causal structure. We use the AdamW optimizer and `CosineAnnealingWarmRestarts` learning rate scheduler, which implements cosine annealing of the learning rate with periodic warm restarts, to train the CausalGaze model. And the initial learning rate is 1×10^{-4} and the scheduler parameters are set as $T_0 = 10$, $T_{mult} = 2$, $\eta_{min} = 1 \times 10^{-6}$. The model is trained for 50 epochs with the early stop mechanism ($patience = 20$) and a batch size of 8. To obtain stable and reliable results, we perform multiple runs and take the average as the final results.

B Details about Experiment

B.1 Main Results

We further evaluate our method CausalGaze against recent baselines (Lookback Lens (Chuang et al., 2024), Factoscope (He et al., 2024), TAD (Vazhentsev et al., 2025)) and incorporate a new metric PR-AUC for evaluation. The experiments are implemented on Llama2-7B, as shown in Table 4. As demonstrated in the results, CausalGaze consistently outperforms all baselines. We attribute it to several key factors: (1) Lookback Lens is designed for contextual hallucinations. While the datasets used in our work are with no external context, its mechanism reduces to a simple similarity measure, failing to capture internal veracity; (2) Factoscope and TAD rely on token probabilities and attention maps. Being sensitive to local fluctuations and specific entities, they struggle with complex logical hallucinations; (3) CausalGaze captures global causal dependencies by modeling the full question-answering (QA) as a causal graph. From the information-theoretic perspective, CausalGaze processes a richer set of signals, leading to superior stability and accuracy.

B.2 Generalization

In Section 5.2, we demonstrate the generalization ability of the proposed CausalGaze on Llama2-7B. In this section, we provide additional generalization results on Qwen2-7B and Mistral-7B as shown in Figure 8, which show the effectively robust generalization ability of our approach.

B.3 Contribution of the Projection Layer and its Dimension

In Section 5.3, we demonstrate the contribution of the projection layer and its dimension on the detection performance for Llama2-7B. In this section, we provide additional experimental results on Qwen2-7B and Mistral-7B as shown in Table 5. The results consistently indicate that the projection layer is of considerable interest for the detection performance.

B.4 Layer Number

In Section 5.3, we demonstrate the impact of dynamic causal graphs from different layer of Llama2-7B on the detection performance. In this section, we provide additional experimental results on Qwen2-7B and Mistral-7B as shown in Table 6.

B.5 Gradient-guided Information

Without gradients, the model reverts to the raw graph that treats all attention edges with equal importance, failing to disentangle the causal reasoning paths from noisy and spurious correlations. Random gradients lead to the misallocation of attention in the detection model between spurious edges and genuinely contributory edges. Raw attention weights A only reflect the co-occurrence patterns during the LLM's forward infernece. As documented in prior literature, high attention weights often focus on spurious correlations (e.g., stop words or position bias) that do not necessarily drive the final decision. While the learned gradients S assigns higher importance to edges that have a larger influence on the veracity prediction, allowing the

LLMs	Projection Layer	Dimension	TruthfulQA		TriviaQA		SciQ		Halueval	
			AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
Qwen2-7B	w/o	3584	0.5925	0.5663	0.6043	0.5728	0.5964	0.6459	0.5620	0.6588
	with	64	0.8330	0.7845	0.8764	0.7890	0.8976	**0.8423**	0.8946	0.7892
		128	**0.8680**	**0.8071**	**0.9106**	**0.8280**	**0.9328**	0.8283	**0.9220**	**0.8360**
		256	0.8357	0.7786	0.8873	0.8130	0.8741	0.7740	0.8863	0.7538
		512	0.8588	0.7943	0.8743	0.7924	0.8661	0.7548	0.8750	0.7630
Mistral-7B	w/o	4096	0.6146	0.6273	0.5738	0.5460	0.5964	0.6248	0.5692	0.6469
	with	64	0.8560	0.7533	0.8690	0.7793	0.8452	0.7726	0.8824	0.6926
		128	**0.8851**	**0.7971**	**0.8880**	**0.8113**	**0.8550**	**0.7860**	**0.9127**	**0.8292**
		256	0.8663	0.7894	0.8770	0.7957	0.8467	0.7460	0.8840	0.7850
		512	0.8580	0.7587	0.8659	0.7893	0.8198	0.7319	0.8649	0.7748

Table 5: The contribution of the Projection Layer and its dimentions on Qwen2-7B and Mistral-7B across different datasets. The best results are highlighted in bold.

LLMs	Layer Number	TruthfulQA		TriviaQA		SciQ		Halueval	
		AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score	AUROC	F1-Score
Qwen2-7B	16	0.8718	0.7836	0.9005	0.8194	0.9114	0.8445	0.9260	0.8004
	20	0.8776	0.7893	**0.9125**	**0.8304**	0.9268	**0.8675**	**0.9335**	**0.8421**
	24	**0.8892**	0.7747	0.8994	0.7961	0.8993	0.8593	0.9154	0.8266
	28	0.8757	0.7694	0.8873	0.7868	0.8858	0.8495	0.9255	0.8273
	32	0.8680	**0.8071**	0.9106	0.8280	**0.9328**	0.8283	0.9220	0.8360
Mistral-7B	16	0.8673	0.8057	0.8861	0.8090	0.8479	0.7505	0.8889	0.8162
	20	0.8862	**0.8203**	**0.8882**	0.7806	0.8361	0.7260	0.9073	**0.8347**
	24	**0.8990**	0.8183	0.8637	0.7962	**0.8693**	0.7348	**0.9173**	0.7963
	28	0.8735	0.7769	0.8643	0.7981	0.8463	0.7450	0.9065	0.7990
	32	0.8851	0.7971	0.8880	**0.8113**	0.8550	**0.7860**	0.9127	0.8292

Table 6: The impact of causal graphs from different layers of Qwen2-7B and Mistral-7B on the detection performance.

Method	Input memory	Model parameters	Inference latency
SAPLMA	608KB	109K	∼0.307ms
HaluGNN	614KB	213K	∼0.903ms
SelfCheckGPT	608KB	-	∼10.32s
CausalGaze	614KB	575K	∼1.731ms

Table 7: Computation overhead comparison across different baselines.

gating mechanism to suppress those shortcut edges that do not contribute to the factual reasoning.

B.6 Computation Overhead

We also benchmark CausalGaze against three representative baselines, including SAPLMA, HaluGNN and SelfCheckGPT, in terms of memory footprint (e.g., input and model) and inference latency to provide memory and inference overhead comparison. The experiments are conducted on a single NVIDIA A800 (80GB) GPU with Llama2-7B as the backboneone, and the example is from TriviaQA with input sequence length $L = 38$, feature dimension $D = 4096$. The results on Llama2-7B are summarized in Table 7.

While CausalGaze exhibits a higher model memory and latency compared to other methods, its absolute overhead remains extremely low and highly practical for real-world deployment. The inference latency is approximately 1.731 ms, which is negligible compared to the time required for an LLM to generate a single token. This ensures that our detection framework can operate as a real-time safety monitor without bottlenecking the generation process. The slight increase in complexity of CausalGaze (about +0.8 ms) is a deliberate trade-off to enable active causal intervention. This process allows us to filter out incidental noise and achieve competitive detection performance (e.g., +6.1% AUROC improvement over SAPLMA) with interpretability.