TopoAtten: Probing the Internal Topology of Attention for Hallucination Detection in Large Language Models

Yufei Zeng[1], Yukun Zhang[1], Zhonghong Ou[1,*], Meina Song[1,2,*]
[1]*School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China*
[2]*China University of Petroleum-Beijing at Karamay, China*
Email: yufeizeng2@gmail.com, {zhangyukun, zhonghong.ou, mnsong}@bupt.edu.cn

***Abstract*—Although Large Language Models (LLMs) show exceptional generation capabilities, their application is severely constrained by the challenge of "hallucinations". Existing white-box detection methods avoid the high costs of external retrieval by using internal states but are often confined to analyzing the static statistical properties of high-dimensional hidden states. Thus, they frequently fail when models generate high-confidence hallucinations. To address this limitation, we propose TopoAtten, a novel detection framework grounded in the topological structure of attention mechanisms, viewed through the lens of dynamical manifold evolution theory. Leveraging Topological Data Analysis (TDA), we first reconstruct the geometric skeleton of the attention space via a symmetrized exponential distance transform. Subsequently, we employ persistent homology to construct a topological persistence metric system comprising semantic cohesion (TP_{β_0}) and logical self-consistency (TP_{β_1}) to quantify the structural robustness of the attention mechanism. Finally, we design a lightweight, full-layer TopoAtten probe that precisely identifies hallucinations based on global topological features. Extensive experiments demonstrate that TopoAtten significantly outperforms existing methods across mainstream LLMs and multiple benchmarks, exhibiting superior generalization capabilities and unveiling the deep geometric mechanisms underlying hallucination generation.**

***Index Terms*—Topological Data Analysis,Persistent Homology,Attention Mechanism,Hallucination Detection,Large Language Models**

I. Introduction

Large Language Models (LLMs) have demonstrated exceptional generation capabilities in the field of Natural Language Processing [1]–[3]. However, their application is constrained by the challenge of "hallucinations"—defined as the generation of content that appears plausible but deviates from facts or logic [4]–[7]. Given that hallucinations cannot currently be completely eliminated [8], the development of efficient detection methods is particularly urgent when deploying models in high-stakes domains such as healthcare and legal systems [9].

Existing detection methods are primarily categorized into output-based black-box approaches and internal state-based white-box approaches [10]–[12]. White-box methods have garnered attention as they obviate the need for expensive external retrieval or multiple sampling [13]–[15]. However, current techniques are largely confined to analyzing the static statistical properties of high-dimensional hidden states. Such methods stress the numerical distributional characteristics of hidden states but ignore the underlying context interaction mechanisms responsible for generating them. When models generate high-confidence hallucinations, their internal numerical distributions exhibit high certainty, thereby rendering detection methods based on statistical anomalies ineffective.

To overcome this limitation, we shift our research perspective from numerical probability to **Dynamic Manifold Evolution Theory** [16]–[18]. We model the generation process as the evolutionary trajectory of hidden states upon a low-dimensional semantic manifold and identify the **multi-head self-attention mechanism** as the critical factor determining whether this trajectory converges to a "factual attractor". Unlike previous approaches that indirectly infer the manifold state via the numerical distribution of hidden states, we focus directly on the structural properties of this critical component itself. We hypothesize that high-quality generation relies on the attention mechanism constructing a topologically "semantically cohesive" and "logically self-consistent" structure, which provides robust factual anchors and logical consistency for the generated content. Conversely, the collapse of this structure prevents the effective aggregation of context information, causing the evolutionary trajectory to deviate from the "factual anchor" and drift into the hallucination region.

Building upon this, we propose **TopoAtten**, a hallucination detection framework grounded in Topological Data Analysis (TDA). Firstly we introduce a **symmetrized exponential distance transform** to convert directed attention into an undirected metric space suitable for simplicial complex construction. Subsequently, utilizing **persistent homology**, we construct a system of topological persistence metrics to quantify the structural robustness of the attention mechanism: (1) **Semantic Cohesion** (TP_{β_0}) which measures the model's ability to focus on key evidence; (2) **Logical Self-Consistency** (TP_{β_1}) which characterizes the model's capacity to establish robust logically self-consistent circuits. Finally, we construct a lightweight TopoAtten probe based on full-layer topological features. Experimental results on mainstream open-source LLMs demonstrate that the TopoAtten probe

*Corresponding authors.

effectively identifies complex hallucinations and significantly outperforms existing methods across multiple datasets.

In summary, our core contributions are outlined as follows:

(1)A novel paradigm for hallucination detection based on attention topology: Transcending the limitations of traditional token-level or hidden-state scalar statistics, we introduce a geometric perspective that evaluates hallucinations through the lens of semantic cohesion" and logical self-consistency" within the attention space.

(2) The TopoAtten topological metric framework: By innovatively integrating the symmetrized exponential distance transform with persistent homology, we construct a metric system comprising TP_{β_0} and TP_{β_1}. This framework achieves a robust representation of the global attention topology, effectively addressing the challenge faced by traditional methods in identifying high-confidence hallucinations.

(3) Empirical validation and dynamical-manifold interpretation: Extensive experiments demonstrate that TopoAtten significantly outperforms baseline methods in detecting complex hallucinations. Furthermore, we elucidate this phenomenon through manifold dynamics, revealing how the evolutionary trajectory of hidden states deviates from the "factual attractor" during a hallucination.

Our source code is publicly available at: https://github.com/Monalisahhhh/TopoAtten

II. Related Work

Current LLM hallucination detection research splits into two paradigms based on data sources: black-box methods, relying on outputs, and white-box methods, utilizing internal states.

A. Black-box Methods

Black-box approaches primarily depend on the textual content generated by the model. A representative class within this category is **retrieval-based verification**, which aims to cross-verify the factuality of generated content against reliable external knowledge bases [19]–[22]. Although this approach offers high accuracy, it is constrained by the limited coverage of external resources and the latency introduced by retrieval processes, rendering it difficult to effectively scale to open-domain generation tasks [23]. Another line of work relies on the **self-consistency hypothesis**, positing that factual content typically exhibits consistency, whereas hallucinated content manifests as stochastic [24]–[26]. For instance, SelfCheckGPT generates multiple responses to the same prompt and utilizes the inconsistency among these samples as a proxy metric for hallucination detection [27]. While this approach avoids relying on external knowledge, it demands repeated sampling, causing high compute costs that limit real-time feasibility.

B. White-box Methods

To mitigate the reliance on external retrieval and the substantial inference latency associated with black-box methods, white-box approaches focus on leveraging internal signals—such as hidden states, log-probabilities and attention mechanisms to diagnose hallucinations [28]–[30]. These methodologies are primarily categorized into **unsupervised statistical methods** and **supervised probing methods**. Unsupervised statistical methods attempt to directly define scalar metrics to quantify uncertainty. For instance, the INSIDE method evaluates the dispersion of internal states by computing the eigenvalues of the hidden state covariance matrix [31]. Llm-check measures information concentration by calculating the log-determinant of the attention matrix's diagonal [32]. However, these methods inherently rely on dimensionality-reduced scalar statistics, thereby frequently overlooking the complex structural dependencies present during the model's inference process. In contrast, Supervised Probing methods train classifiers to extract signals of truthfulness directly from internal states. As a pioneering work in this domain, SAPLMA demonstrated that linear classifiers trained on hidden states extracted from specific layers can predict the veracity of generated content with high accuracy [33]. Similarly, SEP trains lightweight models to approximate the "semantic entropy" metric, aiming to estimate uncertainty within a single forward pass [34]. Nevertheless, most existing probes rely on static layer-wise snapshots or single token positions, neglecting the dynamic evolution of semantic information as it propagates across Transformer layers. Crucially, while these methods provide discriminative scores, they lack mechanistic interpretability. This deficiency restricts their applicability in scenarios requiring high-reliability diagnostics.

III. Method

We employ Dynamic Manifold Evolution Theory (DMET) to clarify that hallucinations stem from structural anomalies in multi-head attention mechanisms. This anomaly causes hidden state trajectories to deviate from the "factual attractor". We then introduce Topological Data Analysis (TDA) to characterize non-numerical, nonlinear structural anomalies beyond scalar statistics. Consequently, we propose the TopoAtten method, which achieves precise hallucination detection through geometric transformation of attention matrices, topological persistence metric extraction and full-layer probe design.

A. Dynamical Manifold Evolution Theory

Based on Dynamical Manifold Evolution Theory (DMET) [35], we model the generation process of LLMs as an underlying evolutionary trajectory upon a low-dimensional semantic manifold, where the layer-wise update of hidden states is driven by a controlled differential equation:

$$\mathbf{h}_{l+1} = \mathbf{h}_l + \Delta t \left[-\nabla V(\mathbf{h}_l) + g(\mathbf{h}_l, \mathbf{u}_l) \right] \tag{1}$$

In this framework, distinct terms correspond explicitly to the core components of the Transformer architecture:

(1) The Potential Energy Gradient Term $-\nabla V$: Maps to the Feed-Forward Network (FFN), characterizing the model's intrinsic language priors. Via static key-value information stored in parameters, it "pulls" hidden states towards semantically stable states that maintain syntactic self-consistency.

(2) The Forcing Term $g(\cdot)$: Maps to the Multi-Head Attention mechanism, representing the interaction between the

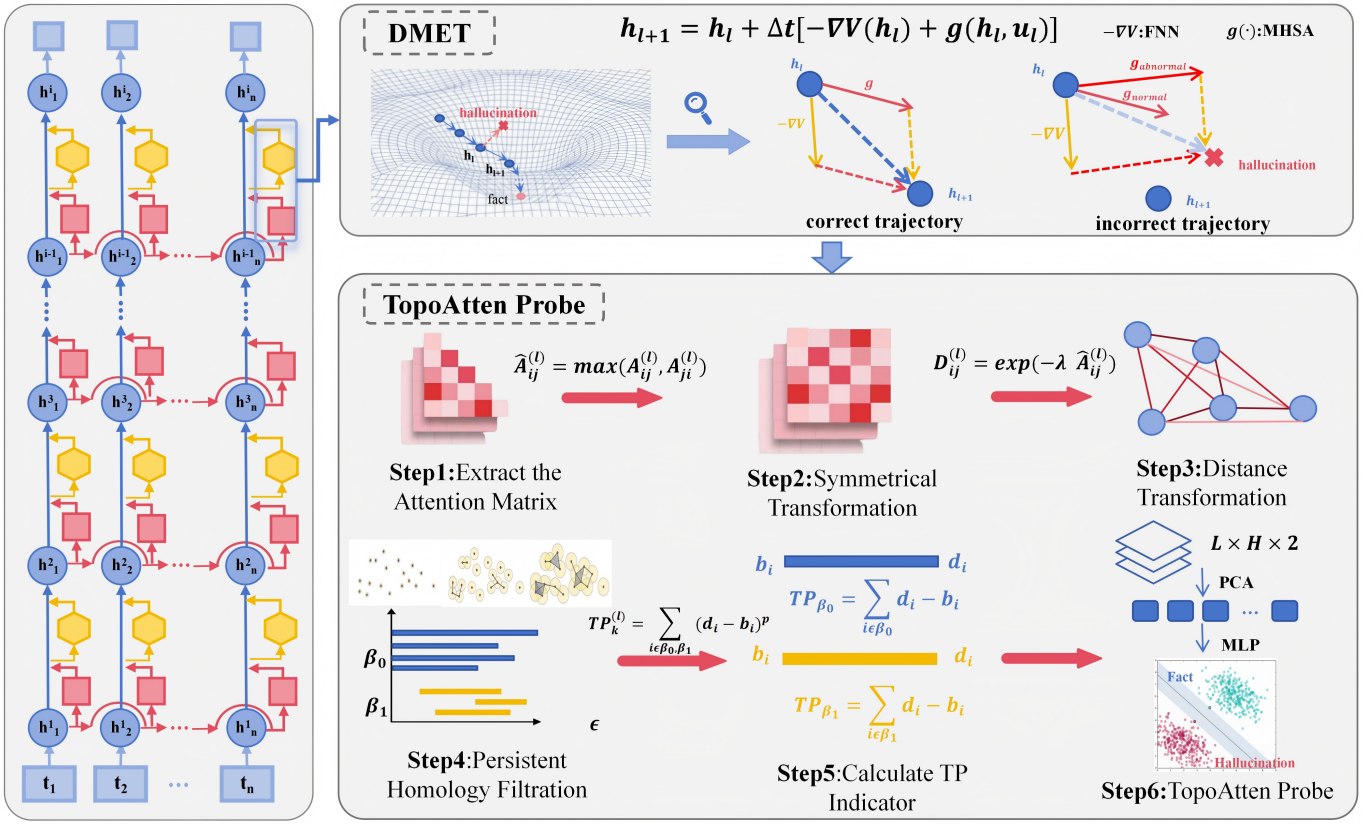

Fig. 1. The Framework of TopoAtten. **(Top) DMET**: Models state evolution as a dynamical system, where hallucinations appear as trajectory deviations caused by attention imbalances. **(Bottom) TopoAtten Probe Pipeline**: The process transforms attention weights into a metric space. We then apply Persistent Homology to extract topological features and train a lightweight probe to distinguish hallucinations based on Topological Persistence indicator.

model and the external context $\mathbf{u}$. Using global contextual information, it guides the evolutionary trajectory of hidden states to converge towards a designated "factual attractor".

From this perspective, we posit that hallucinations primarily stem from structural anomalies within the Forcing Term $g(\cdot)$. Specifically, the attention mechanism loses effective control over contextual information and fails to exert the correct guiding force. Consequently, the evolutionary trajectory of hidden states fails to converge to the established factual attractor, resulting in a deviation on the semantic manifold. To precisely capture these deep-seated semantic anomalies, we introduce **Topological Data Analysis (TDA)** to conduct a geometric structural analysis of the attention matrices.

B. TopoAtten Framework

To characterize attention topology from a geometric dimension, we propose the **TopoAtten** framework(as shown in Fig. 1). This framework maps attention matrices into a metric space and utilizes persistent homology to extract topological features representing information aggregation. Next, quantify the robustness of the topological structure and finally determine hallucinations based on global topological features.

a. Exponential Distance Transformation

This method transforms the Transformer's directed attention matrix into an undirected metric space, laying the groundwork for subsequent topological analysis.

The construction of standard Vietoris-Rips complexes relies on metric symmetry, which conflicts with the asymmetric nature of Transformer attention matrices. To address this, we propose an "Undirected Semantic Skeleton Extraction" strategy. Although symmetrization relaxes directionality, it preserves the essential strength of semantic coupling. We hypothesize that "truthful" generation is underpinned by stable, tightly connected semantic clusters, while hallucinations often emerge as topologically fragmented patterns. Thus, a strong attention link in either direction should be viewed as evidence of geometric proximity. To capture these underlying semantic structures without the complexity of directed homology, we introduce a symmetrization operator combined with an exponential decay model to define the geometric distance $D_{ij}^{(l)}$:

$$\hat{A}_{ij}^{(l)} = \max(A_{ij}^{(l)}, A_{ji}^{(l)}) \quad \text{and} \quad D_{ij}^{(l)} = \exp(-\lambda \cdot \hat{A}_{ij}^{(l)}) \quad (2)$$

where λ is the denoising temperature coefficient. This transformation employs a maximization operator to construct an undirected skeleton, while exponential mapping compresses high-confidence associations into geometric proximity. This

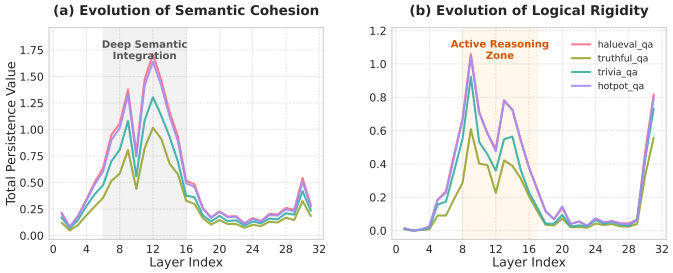

Fig. 2. Layer-wise Evolution of Topological Persistence (TP) Metrics.

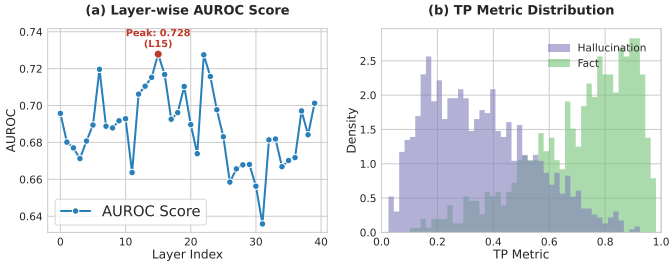

Fig. 3. Performance Analysis of the Topological Persistence (TP) Metric.**(a)** Layer-wise AUROC scores for hallucination detection.**(b)** Probability density distribution of the TP metric for hallucinated (Purple) and factual (Green) samples.

nonlinear filtering effectively suppresses long-tail noise, extracting the topological backbone of the attention mechanism.

b. Persistent Homology Filtration

Based on the distance matrix $D^{(l)}$, we build a Vietoris-Rips complex filtration $\mathcal{VR}_\epsilon$. As ϵ increases, we employ persistent homology to track two key topological features:

- β_0 **(Semantic Cohesion):** Reflects the capacity of the attention mechanism to cluster tokens into semantic units. High persistence implies that the model has formed a compact and stable entity focus within the local context.
- β_1 **(Logical Self-Consistency):** Reflects the construction of closed loops of semantic associations. High persistence implies the model's capacity for robust logical associations that span across multiple semantic units.

c. Topological Persistence

Based on the set of feature survival intervals $\{(b_i, d_i)\}_i$, we define the Topological Persistence metric to quantify the robustness of the topological structure. Filtering out transient noise, we calculate the total feature lifespan for layer l:

$$TP_k^{(l)} = \sum_{i \in \mathcal{S}_k} (d_i - b_i) \tag{3}$$

where:

- TP_{β_0}**:** Measures semantic stability. A high value indicates a distinct focus with clear boundaries, while a low value implies attention diffusion or a failure of focus.
- TP_{β_1}**:** Measures the robustness of long-range logical dependencies. A high value implies the existence of stable semantic closed loops across scales, whereas a low value serves as a direct signal of logical chain breakage.

d. TopoAtten Probe

In this section, we analyze the layer-wise evolutionary characteristics of the TP metric and its capability to distinguish hallucinations. Building upon this foundation, TopoAtten probes are designed and trained based on full-layer topological features to achieve precise hallucination detection.

(1) Empirical Analysis of TP Metrics

We assess the potential of the Topological Persistence (TP) metric in hallucination detection across two key dimensions: (1) cross-domain generalization; (2) the capability to distinguish between hallucinated and factual outputs.

1) Cross-Domain Topological Consistency

Using Llama-3-8B-Instruct as a case study, we computed the layer-wise means of the TP metric system across four distinct datasets. As illustrated in Fig. 2 TP exhibits highly consistent layer-wise evolutionary trajectories across all datasets:

- **Shallow Encoding (Layers 1–5):** Low TP values imply attention is primarily focused on local syntactic feature extraction and the initialization of local context, without constructing topologically significant semantic structures.
- **Reasoning Integration (Layers 6–16):** TP values climb and exhibit a bimodal distribution (peaking around Layers 9 and 13), reflecting model is undergoing deep semantic aggregation and the construction of logical closed loops.
- **Decision Convergence (Layers 17+):** A decline in TP followed by a tail-end rise indicates that the model anchors itself to key evidence to eliminate uncertainty, thereby completing the final vocabulary mapping.

This cross-domain consistency confirms the high robustness of TP, showing it maps the model's underlying universal reasoning mechanisms instead of overfitting specific data.

2) Distinguishing Capability of TP Metrics.

Using the performance of the Mistral-Nemo-12B model on the HaluEval dataset as a case study, Fig. 3a demonstrates the significant separability of the TP metric: the AUROC exceeds 0.7 across 10 layers, peaking at 0.728 in Layer 15. This is further corroborated by Fig. 3b, which illustrates the probability distribution of the TP metric following multi-layer feature fusion. Non-hallucinated and hallucinated samples are concentrated in high-score and low-score regions, respectively, exhibiting a clear structural dichotomy. This validates TP's effectiveness in distinguishing facts from hallucinations and reveals the fundamental topological disparities between them.

(2) TopoAtten Probe Training

The preceding analysis proves TP is a reliable indicator for hallucination detection. However, relying solely on single-layer TP scores yields only a local slice of the attention mechanism at a specific depth, failing to capture the global evolutionary trajectory of the attention structure as it deepens. To fully exploit cross-layer topological information during model inference, we propose the **TopoAtten Probe**. This supervised probe leverages full-layer features to precisely identify hallucinations by integrating complete attention structural patterns, ranging from shallow perception to deep reasoning.

TABLE I
HALLUCINATION DETECTION PERFORMANCE ON FOUR DATASETS. HIGHER AUROC VALUES INDICATE BETTER PERFORMANCE.

LLMs	Methods	HaluEval	HotpotQA	TriviaQA	TruthfulQA
Llama-3	PPL	0.5920	0.6721	0.7012	0.5824
	LN-Entropy	0.6508	0.6593	0.5860	0.6316
	LLM-check	0.5206	0.5491	0.5417	0.5321
	SAPLMA	0.7238	0.7701	0.7650	0.7312
	SEP	0.7378	0.6541	0.7116	0.7135
	Ours	**0.7915**	**0.7835**	**0.8060**	**0.7916**
Qwen2.5	PPL	0.5439	0.6125	0.6974	0.5652
	LN-Entropy	0.7371	0.6855	0.7004	0.7256
	LLM-check	0.5292	0.5445	0.5552	0.5364
	SAPLMA	0.7799	0.7750	**0.8225**	0.7622
	SEP	0.6578	0.6470	0.7520	0.6945
	Ours	**0.8065**	**0.7937**	0.7994	**0.7957**
Mistral-Nemo	PPL	0.5827	0.6636	0.6845	0.5674
	LN-Entropy	0.6846	0.6794	0.6201	0.6431
	LLM-check	0.5466	0.5529	0.5433	0.5663
	SAPLMA	0.7589	**0.8122**	0.7847	0.7625
	SEP	0.7236	0.6673	0.7052	0.7155
	Ours	**0.7910**	0.7858	**0.8227**	**0.8109**

For each input sample, we extract TP metrics across all L layers and H attention heads. These dispersed structural features are flattened and concatenated to construct a high-dimensional global feature vector $\mathbf{v}_{\text{global}} \in \mathbb{R}^{L\times H\times 2}$. This representation preserves the complete structural evolution of the attention mechanism throughout the generation process. Addressing the high-dimensional sparsity of this vector, we design a "PCA+MLP" architecture: First, Principal Component Analysis projects features into a compact subspace to remove noise and mitigate the curse of dimensionality. Subsequently, a lightweight funnel-shaped MLP (configured as d-128-64-1) is employed to ensure nonlinear fitting capability while effectively preventing overfitting.

IV. EXPERIMENTS

In this section, we elaborate on the comprehensive experimental framework designed to evaluate the performance of the TopoAtten Probe in hallucination detection tasks. Specifically, we introduce the selected backbone models, benchmark datasets, comparative baselines, evaluation metrics, and the specific implementation details of the TopoAtten probe. Furthermore, we validate its generalization capabilities across diverse data distributions and conduct ablation studies.

A. Experimental Settings

Models: To validate the versatility and robustness of our method, we conducted extensive evaluations on three mainstream open-source Large Language Models, covering diverse parameter scales and technical evolutionary paths. Specifically, these include Llama-3-8B-Instruct [36], Qwen2.5-7B-Instruct [37] and Mistral-Nemo-12B-Instruct [38].

Datasets: We selected four representative benchmark datasets, covering a wide range of challenging scenarios ranging from specific hallucination detection to complex knowledge question answering: HaluEval, designed specifically for LLM hallucination detection [39]; TruthfulQA, aimed at identifying imitative falsehood tendencies [40]; TriviaQA, focusing on commonsense knowledge extraction [41]; and HotpotQA, involving multi-hop logical reasoning [42].

Baselines: To assess the performance of the TopoAtten Probe, we compared it against existing mainstream hallucination detection methods. The baselines are categorized into two types: **Unsupervised methods:** Including PPL, which evaluates hallucination risk by calculating perplexity during generation [11]; LN-Entropy which normalizes entropy by sequence length [43]; and LLM-check based on kernel similarity analysis of attention mechanisms [32]. **Supervised methods:** Including SAPLMA which utilizes hidden states from specific layers to train classifiers [33]; and SEP which analyzes the semantic entropy of preceding tokens to predict hallucinations [34]. All baseline experimental settings were kept consistent with our proposed method to ensure fair comparison.

TopoAtten Probe Training: TopoAtten probe aims to achieve efficient hallucination detection by mining topological features within the attention mechanism. Architecturally, the probe employs a lightweight Multi-Layer Perceptron. The training process follows a standard supervised learning framework, aiming to minimize the binary cross-entropy loss between the predicted probability distribution and the true labels, using the Adam optimizer for parameter updates.

We split each dataset into training and testing sets and report the experimental results for each dataset respectively. The experimental settings for the baselines are entirely consistent with ours. Furthermore, we specifically evaluate the model's generalization capabilities in Section C.

B. Hallucination Detection Results

Our evaluation results across four representative datasets and three open-source LLMs demonstrate that the **TopoAtten Probe** significantly outperforms both unsupervised and supervised baseline methods in the vast majority of experimental settings. Specifically, on the Mistral-Nemo-12B model and the Truthful_QA dataset, our method achieved a peak AUROC of 0.8227. Compared to supervised baselines that directly utilize

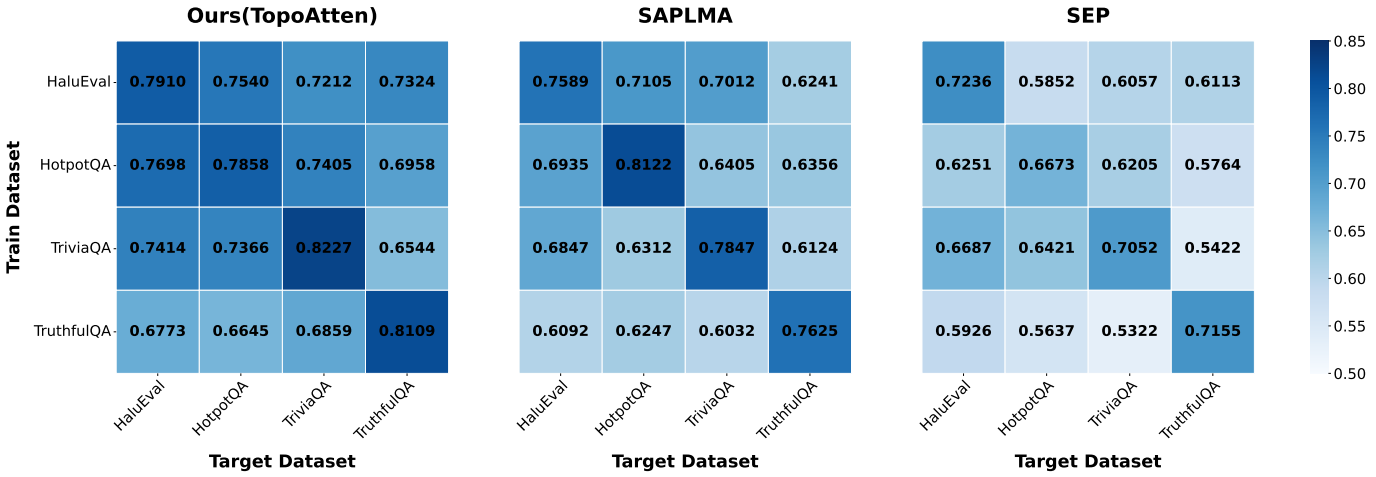

Fig. 4. **Evaluation of Cross-Dataset Generalization Capabilities.** The heatmaps compare the AUROC performance of our TopoAtten Probe (Left) against supervised baselines SAPLMA (Middle) and SEP (Right).

internal activation values (such as SAPLMA and SEP), the TopoAtten Probe exhibits superior detection efficacy.

This consistent superiority confirms that capturing topological structural anomalies of attention mechanisms is a more robust approach to hallucination detection compared to traditional numerical statistical or static feature-based methods.

C. Generalization Analysis

We comprehensively investigated the cross-dataset generalization capabilities of the detector by training on a single dataset and evaluating on alternative datasets. The experimental results are visualized as a heatmap in Fig. 4. The **TopoAtten Probe** exhibits significant robustness on unseen datasets (AUROC > 0.65), consistently outperforming SAPLMA and SEP across all scenarios. Specifically, the performance degradation from in-domain to cross-domain evaluation is limited to 10.98% for our method, which is significantly lower than that of SEP (15.05%) and SAPLMA (16.93%).

This contrast shows the superior accuracy and robustness of the TopoAtten Probe in handling domain shifts. This generalization advantage stems from the probe's ability to capture the intrinsic cognitive dynamics of the LLM during inference, thereby avoiding the sensitivity arising from SAPLMA and SEP's over-reliance on distribution-specific data.

D. Ablation Studies

We conducted ablation studies on the Mistral-Nemo-12B model to investigate two primary aspects: (1) contributions of distinct topological features; (2) the impact of attention topology across layer groups.

a. Analysis of Feature Component Contributions

To evaluate the contribution of each component, we designed three experimental settings: retaining only the TP_{β_0} features that characterize the stability of semantic focus; utilizing only the TP_{β_1} features that characterize the robustness of logical closed loops; and integrating features from both dimensions. The results in Table II indicate that utilizing even a single topological component yields detection performance that far surpasses the random baseline. Specifically:

- TP_{β_0} exhibits superior performance on tasks emphasizing entity extraction, like TriviaQA (0.7628) and TruthfulQA (0.7723), reflecting the stability of entity focus.
- TP_{β_1} shows significant improvement on logical reasoning tasks, such as HaluEval (0.7290) and HotpotQA (0.7332), reflecting the robustness of logical chains.

Integrating both feature categories significantly improves the AUROC across all datasets. This confirms the complementarity of TP_{β_0} and TP_{β_1} in characterizing attention structures and validates the effectiveness of constructing detectors based on persistent topological features.

TABLE II
TP COMPONENTS ABLATION STUDY

Setting TP_{β_0}	Setting TP_{β_1}	HaluEval	HotpotQA	TriviaQA	TruthfulQA
✓	×	0.7151	0.7099	0.7628	0.7723
×	✓	0.7290	0.7332	0.7485	0.7579
✓	✓	**0.7910**	**0.7858**	**0.8227**	**0.8109**

b. Layer Contribution Analysis

To investigate the specific contributions of attention topological structures across different layer regions of the LLM to hallucination detection, we conducted a layer-wise ablation study. We partitioned the model layers into three stages: Early Layers (EL), Middle Layers (ML), and Deep Layers (DL), and individually excluded the features of these layer groups during probe training. The results are presented in Table III.

Across the HaluEval, HotpotQA, and TruthfulQA datasets, removing the Middle Layers (w/o ML) resulted in significant performance degradation, with a notable drop of 7.8% observed on TruthfulQA. This confirms that the middle layers serve as the core locus for logical reasoning and factual integration, and that their topological features contain critical

signals for capturing the root causes of hallucinations.Notably, on the TriviaQA dataset, the performance degradation resulting from the removal of Deep Layers (w/o DL) was more pronounced than that caused by removing Middle Layers. Since this task necessitates precise entity retrieval, the final token selection and confidence allocation are predominantly executed within the deep layers. Conversely, the performance decline observed when removing Early Layers (w/o EL) was relatively marginal across all datasets, indicating that the TopoAtten Probe primarily captures deep semantic and logical dynamic features rather than shallow lexical patterns.

TABLE III
LAYER-WISE ABLATION STUDY

Setting	HaluEval	HotpotQA	TriviaQA	TruthfulQA
w/o EL	0.7794	0.7500	0.7848	0.7536
w/o ML	0.7395	0.7400	0.7553	0.7324
w/o DL	0.7760	0.7522	0.7314	0.7597
Full Layers	**0.7910**	**0.7858**	**0.8227**	**0.8109**

V. CONCLUSION

To address the challenge of detecting high-confidence hallucinations, we propose **TopoAtten**, a framework grounded in the topological structure of attention mechanisms. From the perspective of dynamical manifold evolution, we reveal that hallucinations stem from the loss of **semantic cohesion** (TP_{β_0}) and the insufficiency of **logical self-consistency** (TP_{β_1}), causing the hidden state trajectory to deviate from the "factual attractor". TopoAtten leverages persistent homology to capture these structural anomalies and achieves efficient, robust detection via a lightweight probe. Experimental results confirm the method's significant superiority in complex cross-domain tasks; furthermore, it operates without external knowledge bases, offering a novel geometric perspective on the origins of hallucinations.Despite its superior performance, this method is currently limited to the passive detection of text generation. Future work will focus on exploring active intervention mechanisms based on topological feedback and extending the application to broader cognitive tasks such as code and mathematics, thereby enhancing the universality and practical value of the method.

VI. ACKNOWLEDGMENTS

This work is supported by the National Key Research and Development Program of China under Grant 2024YFC3308500, Beijing Municipal Natural Science Foundation under Grant L251042, National Natural Science Foundation of China under Grant 62406036, China Postdoctoral Science Foundation under Grant 2025M781457, and also sponsored by the State Key Laboratory of Networking and Switching Technology under Grant NST20250110.

REFERENCES

[1] K. Chen, Q. Chen, J. Zhou, X. Tao, B. Ding, J. Xie, M. Xie, P. Li, and Z. Feng, "Enhancing uncertainty modeling with semantic graph for hallucination detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 586–23 594.

[2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[3] G. Aracena, K. Luster, F. Santos, I. Steinmacher, and M. A. Gerosa, "Applying large language models to issue classification," *2024 IEEE/ACM International Workshop on Natural Language-Based Software Engineering (NLBSE)*, pp. 57–60, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272065886

[4] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[5] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, "A comprehensive survey of hallucination mitigation techniques in large language models," *arXiv preprint arXiv:2401.01313*, vol. 6, 2024.

[6] F. Zhang, P. Yu, B. Yi, B. Zhang, T. Li, and Z. Liu, "Prompt-guided internal states for hallucination detection of large language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 21 806–21 818.

[7] K. Naszádi, P. Manggala, and C. Monz, "Aligning predictive uncertainty with clarification questions in grounded dialog," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266177110

[8] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," in *Findings of the association for computational linguistics: ACL 2024*, 2024, pp. 3563–3578.

[9] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM computing surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[11] J. Ren, J. Luo, Y. Zhao, K. Krishna, M. Saleh, B. Lakshminarayanan, and P. J. Liu, "Out-of-distribution detection and selective generation for conditional language models," *arXiv preprint arXiv:2209.15558*, 2022.

[12] L. Yan, L. Sha, L. Zhao, Y. Li, R. M. Maldonado, G. Chen, X. Li, Y. Jin, and D. Gavsevic, "Practical and ethical challenges of large language models in education: A systematic scoping review," *Br. J. Educ. Technol.*, vol. 55, pp. 90–112, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260125084

[13] M. Giulianelli, J. Baan, W. Aziz, R. Fernandez, and B. Plank, "What comes next? evaluating uncertainty in neural text generators against human production variability," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14 349–14 371. [Online]. Available: https://aclanthology.org/2023.emnlp-main.887/

[14] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, "Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms." in *ICLR*. OpenReview.net, 2024. [Online]. Available: http://dblp.uni-trier.de/db/conf/iclr/iclr2024.html#XiongHLLFHH24

[15] F. Petersen, A. Mishra, H. Kuehne, C. Borgelt, O. Deussen, and M. Yurochkin, "Uncertainty quantification via stable distribution propagation," *ArXiv*, vol. abs/2402.08324, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267637154

[16] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.

[17] Q. Li, J. Geng, Z. Chen, D. Zhu, Y. Wang, C. Ma, C. Lyu, and F. Karray, "Hd-ndes: Neural differential equations for hallucination detection in llms," *arXiv preprint arXiv:2506.00088*, 2025.

[18] S. Zeng, F. Graf, C. Hofer, and R. Kwitt, "Topological attention for time series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 24 871–24 882, 2021.

[19] E. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, and P. Liu, "Factool: Factuality detection in generative ai - a tool augmented framework for multi-task and multi-domain scenarios," *ArXiv*, vol. abs/2307.13528, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260154834

[20] V. Rawte, S. T. I. Tonmoy, S. Nag, A. Chadha, A. Sheth, and A. Das, "Factoid: Factual entailment for hallucination detection," in *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, 2025, pp. 599–617.

[21] S. J. Semnani, V. Z. Yao, H. Zhang, and M. S. Lam, "Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia," in *Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258841157

[22] M. Li, Z. Zhan, H. Yang, Y. Xiao, H. Zhou, J. Huang, and R. Zhang, "Benchmarking retrieval-augmented large language models in biomedical nlp: Application, robustness, and self-awareness," *Science Advances*, vol. 11, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:283219791

[23] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi, "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 076–12 100.

[24] Z. Cao, Y. Yang, and H. Zhao, "Autohall: Automated factuality hallucination dataset generation for large language models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 184–195, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263334406

[25] N. Malkin, Z. Wang, and N. Jojic, "Coherence boosting: When your pretrained language model is not paying enough attention," in *Annual Meeting of the Association for Computational Linguistics*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:247476407

[26] Y. Zhao, L. Yan, W. Sun, G. Xing, C. Meng, S. Wang, Z. Cheng, Z. Ren, and D. Yin, "Knowing what llms do not know: A simple yet effective self-detection method," *ArXiv*, vol. abs/2310.17918, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264555682

[27] P. Manakul, A. Liusie, and M. Gales, "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*, 2023, pp. 9004–9017.

[28] H. Orgad, M. Toker, Z. Gekhman, R. Reichart, I. Szpektor, H. Kotek, and Y. Belinkov, "Llms know more than they show: On the intrinsic representation of llm hallucinations," *arXiv preprint arXiv:2410.02707*, 2024.

[29] A. Azaria and T. M. Mitchell, "The internal state of an llm knows when its lying," *ArXiv*, vol. abs/2304.13734, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258352729

[30] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, and P. He, "Dola: Decoding by contrasting layers improves factuality in large language models," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Th6NyL07na

[31] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, and J. Ye, "Inside: Llms' internal states retain the power of hallucination detection," *ArXiv*, vol. abs/2402.03744, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267499843

[32] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, "Llm-check: Investigating detection of hallucinations in large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 34 188–34 216, 2024.

[33] A. Azaria and T. Mitchell, "The internal state of an llm knows when it's lying," *arXiv preprint arXiv:2304.13734*, 2023.

[34] J. Kossen, J. Han, M. Razzak, L. Schut, S. Malik, and Y. Gal, "Semantic entropy probes: Robust and cheap hallucination detection in llms," *arXiv preprint arXiv:2406.15927*, 2024.

[35] Y. Zhang and Q. Dong, "Dynamic manifold evolution theory: Modeling and stability analysis of latent representations in large language models," *arXiv preprint arXiv:2505.20340*, 2025.

[36] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv e-prints*, pp. arXiv–2407, 2024.

[37] Q. Team *et al.*, "Qwen2 technical report," *arXiv preprint arXiv:2407.10671*, vol. 2, no. 3, 2024.

[38] A. Mistral, "Mistral-nemo-instruct-2407," 2024.

[39] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Halueval: A large-scale hallucination evaluation benchmark for large language models," *arXiv preprint arXiv:2305.11747*, 2023.

[40] S. Lin, J. Hilton, and O. Evans, "Truthfulqa: Measuring how models mimic human falsehoods," in *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 2022, pp. 3214–3252.

[41] M. Joshi, E. Choi, D. S. Weld, and L. Zettlemoyer, "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension," *arXiv preprint arXiv:1705.03551*, 2017.

[42] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2369–2380.

[43] A. Malinin and M. Gales, "Uncertainty estimation in autoregressive structured prediction," *arXiv preprint arXiv:2002.07650*, 2020.