

Layer-Wise Localization and Classification of Hallucinations in Large Language Models

Naveen Lamba¹, Sanju Tiwari^{1,*}, Manas Gaur² and Daniel Dobriy³

¹Dept. of CSA & CAIMIF, Sharda University, Greater Noida, India

²University of Maryland, Baltimore County, Baltimore, MD, USA

³WU Vienna, Austria, & Dobriy AI, Austria

Abstract

Hallucination remains a key limitation of large language models (LLMs), often addressed through output-level detection methods that overlook internal model dynamics. Recent work has shown that internal representations encode signals of truthfulness, but does not localize hallucinations across layers or relate them to input structure. In this work, we introduce SYLAC (Symbolic Layer-Aware Hallucination Classification), a framework that models hallucinations as structured phenomena arising from interactions between symbolic linguistic triggers and layer-wise transformer representations. We model hallucinations as layer-aware phenomena that can be jointly characterized by their type and origin within the model. To evaluate this, we analyze symbolic triggers (e.g., entities, negation, numerical expressions) and extract layer-wise attention signals to localize instability. We further develop a predictive model that jointly classifies hallucination type (semantic, syntactic, factual) and estimates its origin layer. Experiments on HaluEval and TruthfulQA across Gemma (2B/9B) and LLaMA (7B/8B) show that the introduced approach achieves up to 70.9% on HaluEval and 81.9% on TruthfulQA for hallucination type classification, and up to 84.7% accuracy in layer localization. While type classification is affected by class imbalance, layer localization consistently achieves higher accuracy, indicating that internal attention dynamics provide reliable structural signals. Consistent patterns across models and datasets suggest that hallucination behavior is associated with shared internal characteristics of transformer architectures. These findings support a shift from output-centric detection toward mechanism-aware modeling, providing a unified and interpretable approach to understanding hallucinations in LLMs.

Keywords

Large Language Models, Hallucination Analysis, Layer-wise Localization, Symbolic Triggers, Mechanistic Interpretability, Trustworthy AI

1. Introduction

Large Language Models (LLMs) achieve strong performance in language tasks but remain prone to hallucination, limiting their reliability in high-stakes applications [1, 2, 3]. While hallucinations are typically treated as output-level errors, they originate from internal representation dynamics within transformer layers. Understanding these failures therefore requires analyzing how they emerge during the model's forward pass.

Studying hallucinations at the internal level enables mechanistic interpretability and provides structured signals for intervention. Transformer layers capture different levels of linguistic and factual abstraction [4, 5], and analyzing their behavior reveals where the model departs from factual grounding. These internal signals can provide more structured information than output-level heuristics and support targeted strategies such as layer-wise correction [6, 7]. This perspective allows hallucinations to be localized, categorized, and predicted through internal dynamics, as illustrated in Figure 1.

Recent work shows that internal representations encode signals of truthfulness and that factual information evolves across layers [8, 9, 10, 11]. However, these approaches largely treat hallucination as

2nd International Workshop on Symbolic and Generative AI for Science (SymGenAI4Sci), co-located with SEMANTiCS 2026: 22nd International Conference on Semantic Systems, September 15, 2026, Ghent, Belgium

*Corresponding author.

✉ naveenlamba30894@gmail.com (N. Lamba); tiwarisanju18@ieee.org (S. Tiwari); manas@umbc.edu (M. Gaur); daniel.dobriy@wu.ac.at (D. Dobriy)

🌐 <http://conceptbase.sourceforge.net/mjf/> (M. Gaur)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

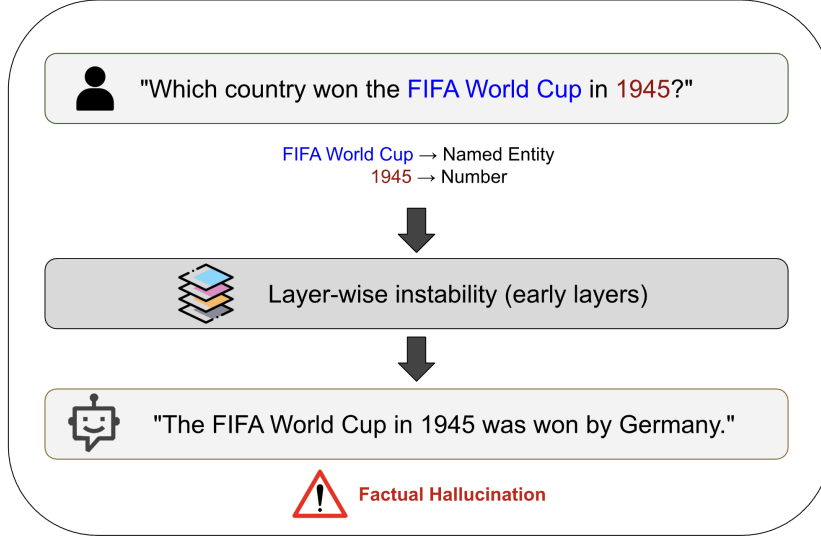


Figure 1: Symbolic triggers in the input can cause instability in intermediate transformer layers, resulting in hallucinated outputs.

a binary phenomenon and do not explain how different hallucination types arise or how input structure influences internal behavior.

To operationalize this perspective, we introduce **SYLAC (Symbolic Layer-Aware Hallucination Classification)**, a framework that models hallucinations as interactions between symbolic linguistic triggers and layer-wise transformer dynamics. SYLAC integrates symbolic trigger identification, layer-wise localization of hallucination origin, and taxonomy-driven modeling to jointly characterize hallucination type and origin.

We evaluate SYLAC across multiple architectures, including Gemma (2B/9B) and LLaMA (7B/8B), using HaluEval and TruthfulQA benchmarks. Results show that hallucinations exhibit consistent layer-wise signatures, with instability emerging in early-to-intermediate layers before propagating to the final output.

Our contributions are as follows:

- We introduce SYLAC, a framework that combines symbolic linguistic triggers and layer-wise transformer representations for hallucination analysis.
- We propose a layer-aware approach for jointly localizing hallucination origins and classifying hallucination types using symbolic attention features.
- We provide empirical evidence across HaluEval and TruthfulQA showing consistent layer-wise hallucination patterns across multiple LLM architectures.

2. Related Work

Recent work has begun to analyze hallucinations through internal transformer dynamics, particularly focusing on layer-wise behavior.

Hidden-state probing studies show that internal representations encode signals predictive of truthfulness. Azaria and Mitchell [8] demonstrate that hidden states can identify factual errors, while INSIDE [12] and subsequent works [13, 14, 15] further confirm that models contain intrinsic signals of hallucination and uncertainty. However, these approaches treat hallucination as a binary prediction problem and do not localize where hallucination originates within the model.

Layer-wise analyses provide more direct insight into hallucination formation. The Tuned Lens [16] reveals how intermediate layers contribute to final predictions, and DoLa [7] improves factuality by contrasting representations across layers. More recent work identifies hallucination as a form of

representational drift across layers [17], while studies on factual recall [18, 19] show that different layers specialize in storing and retrieving knowledge. Although these works indicate that hallucinations have layer-specific signatures, they do not provide a structured mechanism to localize hallucination origins or relate them to distinct hallucination types.

Mechanistic interpretability research further demonstrates that factual knowledge can be localized within transformer components. Feed-forward layers have been shown to act as key-value memories [20], and methods such as ROME and MEMIT [21, 22] enable targeted editing of factual associations. While these approaches localize knowledge and model components, they do not explain how hallucinations emerge dynamically from specific input structures.

Overall, existing work either identifies internal signals of hallucination or analyzes layer-wise dynamics in isolation. There remains a gap in jointly modeling **where hallucinations originate, what type of hallucination occurs, and which input structures trigger them.**

Compared to prior hidden-state probing approaches [8, 12], which primarily treat hallucination analysis as a binary detection problem using internal representations, and layer-wise analyses such as the Tuned Lens [16] and DoLa [7], including our own prior localization-only study SymLoc [23] which analyze representation evolution across transformer layers, SYLAC integrates symbolic trigger analysis with layer-wise attention representations to jointly estimate hallucination type and origin layer within a unified framework. This enables a structured analysis that connects input-level symbolic cues with internal model dynamics, rather than treating detection, localization, and categorization as independent tasks.

Our work addresses this gap by introducing a unified framework that combines symbolic trigger analysis with layer-wise localization and taxonomy-driven modeling, enabling structured and interpretable analysis of hallucination behavior.

3. SYLAC Framework

SYLAC (Symbolic Layer-Aware Hallucination Classification) is a unified framework that integrates symbolic linguistic analysis, layer-wise internal dynamics, and predictive modeling to analyze hallucinations as structured, layer-aware phenomena. Unlike prior approaches that treat hallucinations as output-level errors, SYLAC models them as arising from interactions between symbolic input cues and internal transformer representations.

As illustrated in Figure 2, the framework is designed to (i) identify symbolic triggers in input prompts, (ii) analyze their interaction with layer-wise attention dynamics, (iii) localize the internal origin of hallucination, and (iv) jointly predict hallucination type and its layer-level origin. Figure 2 illustrates the overall architecture of SYLAC.

3.1. Stage 1: Symbolic Trigger Identification

The first stage focuses on extracting symbolic linguistic properties that may act as triggers for hallucination behavior. Given an input sequence $x = \{x_1, x_2, \dots, x_n\}$, we define a set of symbolic properties:

$$S(x) = \{s_1, s_2, \dots, s_k\}$$

where $S(x)$ includes indicators such as named entities, modifiers, numerical expressions, negation, and exception patterns.

These symbolic properties are treated as interpretable indicators of structural complexity and potential sources of internal instability. Rather than assuming direct causality, SYLAC leverages these cues as signals that influence how information is processed across layers, establishing a link between input-level structure and internal representations.

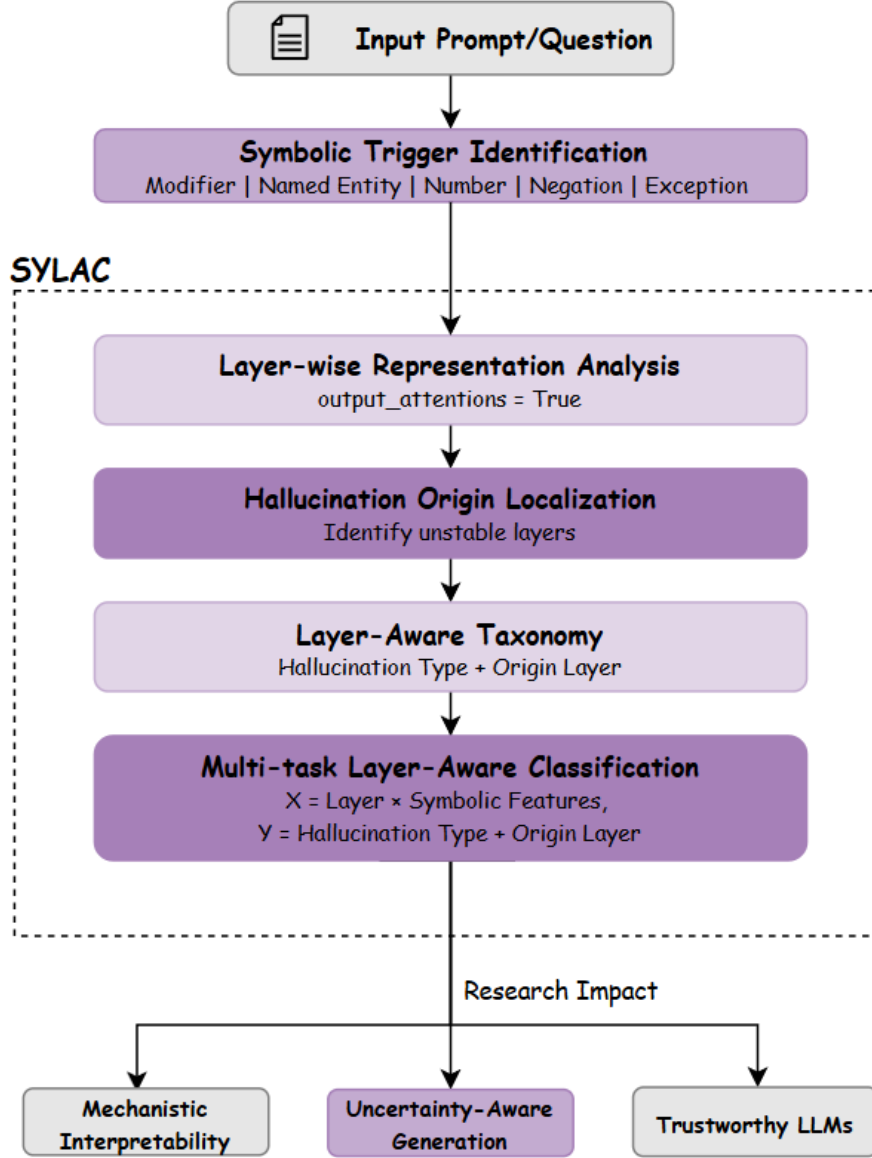


Figure 2: Overview of the SYLAC framework.

3.2. Stage 2: Layer-Wise Attention Analysis and Localization

The second stage analyzes how symbolic triggers interact with internal model representations across transformer layers. Let $A^{(l)}$ denote the attention matrix at layer l . For each symbolic property s_i , we compute the aggregated symbolic attention:

$$\alpha_{s_i}^{(l)} = \frac{1}{|T_{s_i}|} \sum_{t \in T_{s_i}} A^{(l)}(t)$$

where T_{s_i} denotes the set of tokens corresponding to symbolic property s_i .

We define the overall symbolic attention at layer l as:

$$\alpha^{(l)} = \sum_{i=1}^k \alpha_{s_i}^{(l)}$$

To localize hallucination origin, we identify the earliest layer exhibiting a significant increase in

symbolic attention:

$$l^* = \min \left\{ l \mid \alpha^{(l)} - \alpha^{(l-1)} > \delta \right\}$$

where δ is a threshold capturing attention shift. This formulation provides an operational, attention-based estimate of the layer at which hallucination-related instability is first detectable; we do not claim this establishes a causal origin, but rather a structural correlate that can guide further investigation.

Empirically, hallucination-related instability often emerges in early or intermediate layers before propagating through subsequent layers, providing an interpretable estimate of *where* hallucination-related instability first becomes observable.

3.3. Stage 3: Layer-Aware Hallucination Taxonomy

Based on localized internal signals, SYLAC introduces a structured taxonomy that categorizes hallucinations into three types: *semantic*, *syntactic*, and *factual*. Formally, each sample is assigned a type label:

$$y_{\text{type}} \in \{\text{semantic}, \text{syntactic}, \text{factual}\}$$

and an associated origin layer:

$$y_{\text{layer}} = l^*$$

This defines a layer-aware mapping:

$$(x) \rightarrow (y_{\text{type}}, y_{\text{layer}})$$

which enables structured interpretation of hallucination behavior across model depth.

Symbolic-to-Hallucination Mapping. To ensure consistent and reproducible annotation of hallucination types, we define a deterministic mapping between symbolic linguistic triggers and hallucination categories, as summarized in Table 1. Specifically, we associate different symbolic properties with the type of hallucination they are most likely to induce based on their linguistic function:

Table 1

Mapping between symbolic linguistic triggers and hallucination types.

Symbolic Property	Hallucination Type
Negation, Exception	Semantic
Named Entities, Numbers	Factual
Modifiers	Syntactic

- **Semantic hallucinations** are associated with *negation* and *exception* constructs, as these alter the underlying meaning of the input and often lead to incorrect semantic interpretations.
- **Factual hallucinations** are associated with *named entities* and *numerical expressions*, which require precise retrieval of world knowledge and are prone to factual inaccuracies.
- **Syntactic hallucinations** are associated with *modifiers* (e.g., adjectives, adverbs, verbs), which influence sentence structure and can lead to grammatically or structurally inconsistent outputs.

The mapping presented in Table 1 serves as a deterministic annotation heuristic designed to provide consistent and reproducible labels for large-scale analysis. It is not intended as a causal claim that a symbolic trigger uniquely determines the resulting hallucination type. Rather, the mapping provides a structured approximation for analysing relationships between symbolic input properties and internal model behaviour. While individual symbolic triggers may be associated with multiple hallucination

types in practice, this deterministic mapping enables systematic annotation and facilitates the proposed layer-wise analysis by establishing a structured link between input-level symbolic cues and hallucination categories.

3.4. Stage 4: Joint Predictive Classification and Localization

The final stage operationalizes SYLAC through predictive modeling. Each sample is represented as a feature vector:

$$\mathbf{z} = f(x) \in \mathbb{R}^{L \times k}$$

where L is the number of layers and k is the number of symbolic properties.

We train a multi-task classifier:

$$(\hat{y}_{\text{type}}, \hat{y}_{\text{layer}}) = g(\mathbf{z})$$

where $g(\cdot)$ is a neural network with shared representation and task-specific heads.

The training objective is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{type}} + \lambda \mathcal{L}_{\text{layer}}$$

where $\mathcal{L}_{\text{type}}$ and $\mathcal{L}_{\text{layer}}$ are cross-entropy losses for type classification and layer localization, respectively, and λ is a weighting factor.

This formulation enables simultaneous classification and localization, bridging interpretability and predictive modeling. Experimental results demonstrate that internal attention signals contain sufficient information to distinguish between hallucination types and identify their layer-level origins. Furthermore, these patterns exhibit consistent behavior across different model architectures, supporting the hypothesis that hallucinations are structured, layer-aware phenomena.

Overall, SYLAC provides a unified pipeline that connects symbolic input structure, internal model dynamics, and predictive modeling, enabling interpretable and mechanism-aware hallucination analysis.

4. Experimental Setup

This section describes the datasets, models, feature construction process, and classification setup used to evaluate the SYLAC framework.

4.1. Datasets

We conduct experiments on two widely used benchmarks for evaluating hallucination in large language models:

HaluEval. HaluEval [24] is a benchmark designed to evaluate hallucination behavior across different tasks, including question answering, and reasoning-based prompts. We use a subset of 1000 samples covering diverse linguistic structures and reasoning requirements.

TruthfulQA. TruthfulQA [25] consists of 817 questions designed to assess whether models produce truthful or misleading answers, particularly in adversarial settings. The dataset includes both factual and misconception-driven queries, making it suitable for analyzing hallucination patterns.

For each dataset, model-generated outputs are collected and annotated for hallucination behavior. Each sample is further categorized into one of three hallucination types: *semantic*, *syntactic*, or *factual*, based on symbolic linguistic triggers and the nature of the generated error.

4.2. Models

We evaluate SYLAC across multiple transformer-based language models:

- **Gemma-2B and Gemma-9B** [26]
- **LLaMA-7B** [27] and **LLaMA-8B** [28]

These models vary in scale and architecture, enabling analysis of hallucination behavior across different model families. For each model, outputs are generated using standard decoding settings, and internal attention values are extracted during inference.

4.3. Symbolic Trigger Extraction

For each input query, we identify symbolic linguistic properties that may influence hallucination behavior. These include: Named entities, Modifiers (adjectives, adverbs, verbs), Numerical expressions, Negation, and Exception patterns.

These symbolic triggers are mapped to corresponding tokens in the input, enabling alignment with internal attention signals during layer-wise analysis.

4.4. Layer-Wise Attention Feature Construction

To capture internal model dynamics, we extract attention values from all transformer layers using:

`output_attentions=True`

For each sample, we compute symbolic attention scores by aggregating attention weights associated with tokens corresponding to symbolic triggers. This results in a structured representation of the form:

$$\text{Feature Space} = \text{Layer} \times \text{Symbolic Property}$$

Each sample is represented as a flattened feature vector capturing how different symbolic properties are attended across layers.

4.5. Layer Localization

Using the constructed attention features, we estimate the probable origin layer of hallucination for each sample. The estimated origin layer is defined as the earliest layer exhibiting a significant increase in symbolic attention relative to preceding layers.

Formally, given symbolic attention $\alpha^{(l)}$ at layer l , the origin layer is estimated as:

$$l^* = \min \left\{ l \mid \alpha^{(l)} - \alpha^{(l-1)} > \delta \right\}$$

where δ is a threshold capturing attention shift. This produces a layer label for each sample, which is used as heuristic supervision for localization.

4.6. Classification Setup

We train a supervised multi-task classifier using the constructed feature representations. The classifier jointly predicts:

- Hallucination type: {semantic, syntactic, factual}
- Estimated hallucination origin layer: discrete layer index

Table 2

Hallucination Type Classification Accuracy (%)

Model	HaluEval	TruthfulQA
Gemma-2B	68.45	77.91
Gemma-9B	70.88	79.14
LLaMA-7B	67.12	81.88
LLaMA-8B	64.45	77.33

The input consists of flattened layer-wise symbolic attention features, while the outputs correspond to the annotated hallucination type and the estimated origin layer.

The classifier is optimized using a joint objective that combines the losses for hallucination type classification and estimated origin-layer prediction.

Implementation Details. All experiments were conducted using the official Hugging Face checkpoints: google/gemma-2-2b, google/gemma-2-9b, meta-llama/Llama-2-7b-hf, and meta-llama/Llama-3.1-8B. Model outputs were generated using max_new_tokens=200, do_sample=True, temperature=0.7, and top_p=0.9, while internal attention representations were extracted using output_attentions=True. Symbolic attention features were standardized using StandardScaler before training. The multi-task classifier comprised two shared fully connected layers with 256 and 128 hidden units, respectively, using ReLU activations, Batch Normalization, and a dropout rate of 0.3, followed by separate task-specific heads for hallucination type classification and estimated origin-layer prediction. The model was trained using the Adam optimizer (learning rate $= 1 \times 10^{-3}$, weight decay $= 1 \times 10^{-5}$) for 30 epochs with a batch size of 32. The total loss was computed as $\mathcal{L} = \mathcal{L}_{\text{type}} + 0.5 \mathcal{L}_{\text{layer}}$, and gradients were clipped to a maximum norm of 1.0. An 80:20 stratified train-test split with a fixed random seed of 42 was used for all experiments to ensure reproducibility.

5. Results and Analysis

This section presents the empirical evaluation of the introduced SYLAC framework on two benchmark datasets, HaluEval and TruthfulQA, across four model variants: Gemma-2B, Gemma-9B, LLaMA-7B, and LLaMA-8B. The evaluation focuses on (i) hallucination type classification and (ii) layer-wise localization.

5.1. Hallucination Type Classification

Table 2 reports the classification accuracy for hallucination types across both datasets.

Across both datasets, hallucination type classification achieves moderate performance. Accuracy ranges between 64% and 71% on HaluEval, and improves to 77%–82% on TruthfulQA. Despite this improvement, class-wise performance remains imbalanced, with factual hallucinations being significantly easier to predict than semantic and syntactic ones. This imbalance is reflected in lower macro-level performance, indicating that minority hallucination types remain challenging to distinguish. Macro-averaged F1-scores (not shown for brevity) follow similar trends, further highlighting the impact of class imbalance.

5.2. Layer Localization Performance

Table 3 presents the accuracy of predicting the origin layer of hallucinations.

Layer localization consistently outperforms hallucination type classification across both datasets. For HaluEval, localization accuracy exceeds 70% for most models, while TruthfulQA further improves performance, reaching up to 84.66% for Gemma-9B. These results indicate that internal attention-based features provide robust and consistent signals for identifying the layer-level origin of hallucinations.

However, LLaMA-8B exhibits comparatively lower localization accuracy on both datasets, suggesting that deeper architectures may distribute hallucination-related signals across multiple layers, making precise localization more challenging.

Table 3
Layer Localization Accuracy (%)

Model	HaluEval	TruthfulQA
Gemma-2B	78.57	79.75
Gemma-9B	72.15	84.66
LLaMA-7B	76.71	79.19
LLaMA-8B	48.19	63.33

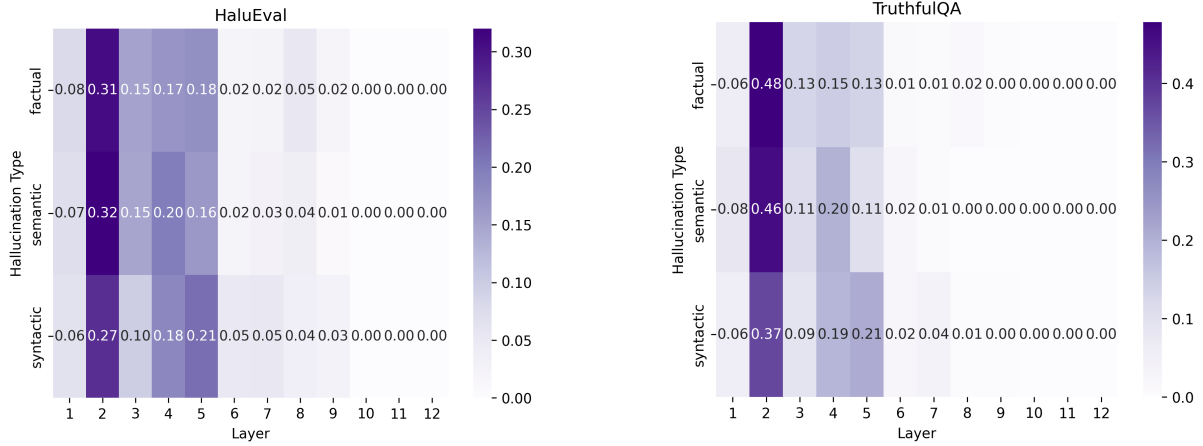


Figure 3: Layer-wise distribution of hallucination types across transformer layers for HaluEval and TruthfulQA.

5.3. Layer-wise Distribution Analysis

Figure 3 illustrates the layer-wise distribution of hallucination types across transformer layers. To ensure comparability across datasets and models with varying depths, layer indices are aligned to the maximum layer range, with missing layers represented as zero-valued entries.

The results reveal a consistent concentration of hallucination activity in early-to-intermediate layers across both datasets. In particular, all hallucination types exhibit peak intensity within the initial layers, followed by a gradual decline in later layers, indicating that instability emerges early during representation formation.

Notably, while HaluEval exhibits a broader spread across intermediate layers, TruthfulQA shows a stronger concentration in the earliest layers, suggesting that adversarial or misconception-driven queries induce earlier disruption in the model’s internal processing. This consistency across datasets, despite differences in task structure and difficulty, supports the hypothesis that hallucinations originate from systematic internal dynamics rather than purely output-level variability.

5.4. Comparative Observations

A comparison across datasets reveals several consistent trends. First, both hallucination type classification and layer localization perform better on TruthfulQA than on HaluEval. This suggests that hallucination patterns in TruthfulQA are more separable based on the extracted symbolic and internal features.

Second, layer localization remains more reliable than type classification across all models and datasets. This indicates that structural signals embedded within internal representations are more stable than categorical distinctions derived from output behavior. Notably, the consistent performance gap between layer localization and type classification suggests that internal structural signals are more reliably captured than output-level hallucination categories.

Finally, model-wise trends remain consistent across datasets, with Gemma models showing stable performance and LLaMA-8B exhibiting reduced localization accuracy.

6. Discussion

The results show that hallucinations exhibit structured behavior linked to symbolic inputs and layer-wise internal dynamics. Rather than uniform output errors, hallucinations originate from localized instability within specific transformer layers.

Layer localization consistently outperforms hallucination type classification, indicating that internal structural signals are more reliable than output-level categories. This suggests that hallucination behavior is primarily governed by internal representation dynamics.

Layer-wise analysis further reveals that hallucinations emerge in early-to-intermediate layers, with peak activity in initial layers. Notably, TruthfulQA shows stronger early-layer concentration compared to HaluEval, suggesting that adversarial queries induce earlier instability.

These consistent patterns across models indicate that hallucination behavior reflects shared characteristics of transformer architectures. By linking symbolic triggers with internal layer behavior, SYLAC provides a unified and interpretable view of hallucination formation.

Several limitations should be acknowledged. First, the hallucination-type and estimated origin-layer labels are generated using deterministic annotation heuristics derived from symbolic attention features; therefore, the reported accuracies evaluate how effectively the proposed classifier reproduces these heuristic labels rather than independently validated causal ground truth. Second, attention is used as an interpretable proxy for internal model behaviour and does not by itself establish causal mechanisms underlying hallucination formation. Finally, validating the estimated origin layers through intervention techniques such as activation patching or layer ablation, together with broader human validation of hallucination types, remains an important direction for future work.

7. Conclusion

We introduced SYLAC, a symbolic layer-aware framework for analyzing hallucinations in large language models. By modeling hallucinations as interactions between symbolic triggers and internal representations, SYLAC shifts the focus from output-level detection to mechanism-aware analysis. The framework enables joint prediction of hallucination type and origin layer, demonstrating that hallucinations can be effectively localized in early-to-intermediate layers. Results further show that internal structural signals are more reliable than output-level categories. Consistent patterns across models suggest that hallucination behavior is governed by shared transformer dynamics, while dataset variations indicate that input conditions influence the depth of instability. Future work will extend SYLAC to larger models and explore circuit-level interpretability for improved robustness.

Acknowledgments

The authors gratefully acknowledge the use of the NVIDIA H100 DGX system provided by the Centre for Artificial Intelligence in Medicine, Imaging and Forensics (CAIMIF), Sharda University, which made the large-scale model experiments and evaluations possible.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT for Grammar and spelling check. After using this, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- [1] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, *ACM Comput. Surv.* 55 (2023). URL: <https://doi.org/10.1145/3571730>.
- [2] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, *ACM Trans. Inf. Syst.* 43 (2025). URL: <https://doi.org/10.1145/3703155>.
- [3] N. Lamba, S. Tiwari, M. Gaur, Hallucinations in LLMs: A lifecycle-based survey of causes, detection, mitigation, and prevention, *International Journal of Data Science and Analytics* 22 (2026). URL: <https://doi.org/10.1007/s41060-026-01214-6>.
- [4] I. Tenney, D. Das, E. Pavlick, BERT rediscovers the classical NLP pipeline, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4593–4601. URL: <https://aclanthology.org/P19-1452/>. doi:10.18653/v1/P19-1452.
- [5] M. Yuksekgonul, V. Chandrasekaran, E. Jones, S. Gunasekar, R. Naik, H. Palangi, E. Kamar, B. Nushi, Attention satisfies: A constraint-satisfaction lens on factual errors of language models, in: *International Conference on Learning Representations*, volume 2024, 2024, pp. 32097–32121.
- [6] K. Li, O. Patel, F. Viégas, H. Pfister, M. Wattenberg, Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL: <https://arxiv.org/abs/2306.03341>.
- [7] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. R. Glass, P. He, Dola: Decoding by contrasting layers improves factuality in large language models, in: *International Conference on Learning Representations*, volume 2024, 2024, pp. 54158–54183.
- [8] A. Azaria, T. Mitchell, The internal state of an LLM knows when it’s lying, in: H. Bouamor, J. Pino, K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics, Singapore, 2023, pp. 967–976. URL: <https://aclanthology.org/2023.findings-emnlp.68/>. doi:10.18653/v1/2023.findings-emnlp.68.
- [9] C. Burns, H. Ye, D. Klein, J. Steinhardt, Discovering latent knowledge in language models without supervision, 2024. URL: <https://arxiv.org/abs/2212.03827>.
- [10] nostalgebraist, Interpreting GPT: The logit lens, LessWrong, 2020. URL: <https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>.
- [11] G. Dar, M. Geva, A. Gupta, J. Berant, Analyzing transformers in embedding space, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 16124–16170. URL: <https://aclanthology.org/2023.acl-long.893/>. doi:10.18653/v1/2023.acl-long.893.
- [12] C. Chen, K. Liu, Z. Chen, Y. Gu, Y. Wu, M. Tao, Z. Fu, J. Ye, Inside: LLMs’ internal states retain the power of hallucination detection, 2024. URL: <https://arxiv.org/abs/2402.03744>.
- [13] H. Orgad, M. Toker, Z. Gekhman, R. Reichart, I. Szpektor, H. Koteck, Y. Belinkov, LLMs know more than they show: On the intrinsic representation of LLM hallucinations, *arXiv preprint arXiv:2410.02707* (2024).
- [14] Z. Ji, D. Chen, E. Ishii, S. Cahyawijaya, Y. Bang, B. Wilie, P. Fung, LLM internal states reveal hallucination risk faced with a query, in: Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, H. Chen (Eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Association for Computational Linguistics, Miami, Florida, US, 2024, pp. 88–104. URL: <https://aclanthology.org/2024.blackboxnlp-1.6/>. doi:10.18653/v1/2024.blackboxnlp-1.6.
- [15] A. Zhang, Y. Chen, J. Pan, C. Zhao, A. Panda, J. Li, H. He, Reasoning models know when they’re right: Probing hidden states for self-verification, 2025. URL: <https://arxiv.org/abs/2504.05419>. arXiv:2504.05419.
- [16] N. Belrose, I. Ostrovsky, L. McKinney, Z. Furman, L. Smith, D. Halawi, S. Biderman, J. Steinhardt, Eliciting latent predictions from transformers with the tuned lens, 2025. URL: <https://arxiv.org/>

abs/2303.08112. arXiv:2303.08112.

- [17] H. Kim, T. A. Lamb, A. Bibi, P. Torr, Y. Gal, Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 32310–32322. URL: <https://aclanthology.org/2025.emnlp-main.1644/>. doi:10.18653/v1/2025.emnlp-main.1644.
- [18] M. Geva, J. Bastings, K. Filippova, A. Globerson, Dissecting recall of factual associations in auto-regressive language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 12216–12235. URL: <https://aclanthology.org/2023.emnlp-main.751/>. doi:10.18653/v1/2023.emnlp-main.751.
- [19] Q. Yu, J. Merullo, E. Pavlick, Characterizing mechanisms for factual recall in language models, 2023. URL: <https://aclanthology.org/2023.emnlp-main.615/>. doi:10.18653/v1/2023.emnlp-main.615.
- [20] M. Geva, R. Schuster, J. Berant, O. Levy, Transformer feed-forward layers are key-value memories, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5484–5495. URL: <https://aclanthology.org/2021.emnlp-main.446/>. doi:10.18653/v1/2021.emnlp-main.446.
- [21] K. Meng, D. Bau, A. Andonian, Y. Belinkov, Locating and editing factual associations in gpt, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.
- [22] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, D. Bau, Mass-editing memory in a transformer, 2023. URL: <https://arxiv.org/abs/2210.07229>.
- [23] N. Lamba, S. Tiwari, M. Gaur, Symloc: Symbolic localization of hallucination across halueval and truthfulqa, in: Proceedings of the 13th ACM IKDD International Conference on Data Science, CODS '25, Association for Computing Machinery, New York, NY, USA, 2026, p. 182–190. URL: <https://doi.org/10.1145/3799830.3799850>.
- [24] J. Li, X. Cheng, X. Zhao, J.-Y. Nie, J.-R. Wen, HaluEval: A large-scale hallucination evaluation benchmark for large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6449–6464. URL: <https://aclanthology.org/2023.emnlp-main.397/>.
- [25] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring how models mimic human falsehoods, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3214–3252. URL: <https://aclanthology.org/2022.acl-long.229/>.
- [26] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, et al., Gemma 2: Improving open language models at a practical size, 2024. URL: <https://arxiv.org/abs/2408.00118>.
- [27] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, et al., Llama 2: Open foundation and fine-tuned chat models, 2023. URL: <https://arxiv.org/abs/2307.09288>.
- [28] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, et al., The llama 3 herd of models, 2024. URL: <https://arxiv.org/abs/2407.21783>.