SymLoc: Symbolic Localization of Hallucination across HaluEval and TruthfulQA

Naveen Lamba
CAIMIF & Dept. of CSA
Sharda University
Greater Noida, India
naveenlamba30894@gmail.com

Sanju Tiwari
CAIMIF & Dept. of CSA
Sharda University
Greater Noida, India
tiwarisanju18@ieee.org

Manas Gaur
University of Maryland
Baltimore County, USA
manas@umbc.edu

Abstract

Large Language Models (LLMs) still struggle with hallucination, especially when confronted with symbolic triggers like modifiers, negation, numbers, exceptions, and named entities. Yet, we lack a clear understanding of where these symbolic hallucinations originate, making it crucial to systematically handle such triggers and localize the emergence of hallucination inside the model. While prior work explored localization using statistical techniques like Local Source Contribution (LSC) and activation variance analysis, these methods treat all tokens equally and overlook the role symbolic linguistic knowledge plays in triggering hallucinations. So far, no approach has investigated how symbolic elements specifically drive hallucination failures across model layers, nor has symbolic linguistic knowledge been used as the foundation for a localization framework. We propose the first symbolic localization framework that leverages symbolic linguistic and semantic knowledge to meaningfully trace the development of hallucinations across all model layers. By focusing on how models process symbolic triggers—negation, named entities, numbers, and exceptions—we analyze five models using HaluEval and TruthfulQA. Our symbolic knowledge approach reveals that attention variance for these linguistic elements explodes to critical instability in early layers (2-4), with negation triggering catastrophic variance levels, demonstrating that symbolic semantic processing breaks down from the very beginning. Through the lens of symbolic linguistic knowledge, despite larger model sizes, hallucination rates remain consistently high (78.3%-83.7% across Gemma variants), with steep attention drops for symbolic semantic triggers throughout deeper layers. Our findings demonstrate that hallucination is fundamentally a symbolic linguistic processing failure, not a general generation problem, revealing that symbolic semantic knowledge provides the key to understanding and localizing hallucination mechanisms in LLMs.

Keywords

Hallucination, Large Language Models, Attention, Symbolic Triggers, Symbolic Properties

1 Introduction

Hallucination in Large Language Models (LLMs) fundamentally stems from failures in *symbolic processing*—the model's ability to correctly interpret and manipulate abstract linguistic concepts that carry specific meaning, definitions, and symbolic constraints. In the context of LLMs, symbolic processing encompasses how models handle negation (which reverses truth values), named entities (which anchor factual claims), numerical values (which specify exact quantities), and logical exceptions (which create conditional reasoning paths). Symbolic features—such as negation, named entities, numerical values, modifiers, and exceptions—have been repeatedly associated with hallucination [11], since they impose greater burdens on a model's reasoning and factual grounding abilities, particularly when models need to manipulate abstract concepts or perform multi-step logical operations. If we can identify where semantic reasoning breaks down within model layers, we could develop targeted interventions to prevent hallucination at its source rather than detecting it post-hoc.

Consider a canonical example from TruthfulQA [14]: "*Do we only use 10% of our brains?*" The correct answer requires logical interpretation to debunk this widespread myth: "*No, we use virtually all of our brain.*" However, models frequently fail to engage the abstract reasoning required to negate false premises, instead generating hallucinated affirmations like "*Yes, Humans only utilize about 10% of their brains at any given time*" (as generated by Llama-3.1-8B). This failure demonstrates how models struggle with conceptual reasoning that requires contradicting common misconceptions. Similarly, HaluEval [12] questions involving specific named entities often trigger fabricated details when models encounter unfamiliar entity-attribute combinations.

Current hallucination localization methods face fundamental limitations in identifying when and where semantic interpretation fails. Local Source Contribution (LSC) [29] quantifies how much each input token contributes to generating specific output tokens through gradient-based attribution:

$$\mathrm{LSC}(x_i, y_j) = \frac{\partial \log P(y_j | x_{1:i}, y_{1:j-1})}{\partial x_i} \tag{1}$$

where x_i represents the i-th input token and y_j represents the j-th output token. However, as Figure 2 illustrates, LSC struggles to link attribution scores with the semantic origins of abstract concepts [22]. Layer-wise approaches like Decoding by Contrasting Layers (DoLa) [2] contrast logit shifts between later and earlier layers, enabling partial localization of factual knowledge. Yet, as shown in the example in Figure 1, DoLa surfaces competing candidates (e.g., Raleigh vs. North Carolina) without identifying the precise layer where reasoning collapses. Moreover, its outputs reveal conflicting signals across layers, blurring the distinction between genuine factual grounding and spurious noise. This suggests that while DoLa highlights tension between layers, it cannot deliver fine-grained explanations of how or when logical reasoning breaks down. Entropy-based methods [9] track information loss but cannot isolate breakdowns triggered by specific symbolic cues. Most critically, *at which stage of the forward pass symbolic cues start to interfere with the model's internal representations remains unknown.*

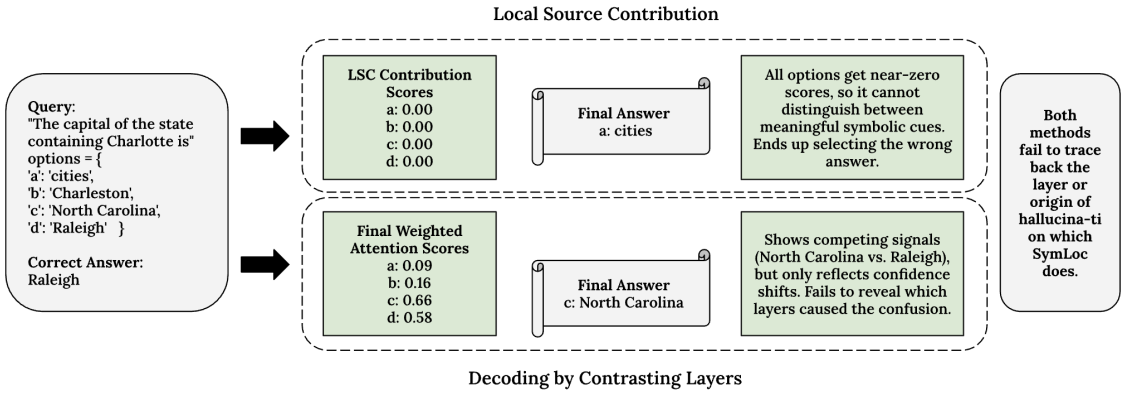


Figure 1: Comparison of LSC and DoLa on the query "The capital of the state containing Charlotte is." LSC assigns flat scores and selects cities, while DoLa shows competing signals (North Carolina vs. Raleigh) but cannot pinpoint the layer of confusion. Both fail to localize the origin of hallucination; therefore, we did not consider them in our study.

To address these limitations, we propose **SymLoc**, a symbolic localization framework that traces the internal emergence of hallucination by focusing specifically on abstract concept manipulation failures. **SymLoc** introduces *symbolic attention*—the attention models assign to tokens with symbolic properties—and tracks its variance across layers to detect when the processing of symbolic triggers becomes unstable. This approach moves beyond token-level attribution to diagnose hallucination as a structural phenomenon rooted in conceptual reasoning breakdowns.

Our key contributions are threefold. First, we develop the first framework to localize hallucination through the lens of semantic interpretation, identifying specific layers where abstract reasoning fails rather than treating all tokens equally. Second, we introduce symbolic attention variance as a novel metric that detects early encoding failures in logical processing, demonstrating that hallucinations originate in early layers (2–4) rather than late-stage decoding. Third, through systematic evaluation across five open-weight models on HaluEval and TruthfulQA benchmarks, we establish that symbolic attention variance peaks early and correlates with high hallucination rates, particularly for negation and named entities.

Our findings reveal that semantic reasoning failures emerge from early layer instability, with symbolic attention variance peaking in initial layers while hallucination rates remain persistently high across model scales. These results reframe hallucination as a layer-localized breakdown in abstract concept manipulation, providing the first interpretable approach to understanding when and where symbolic reasoning fails in LLMs.

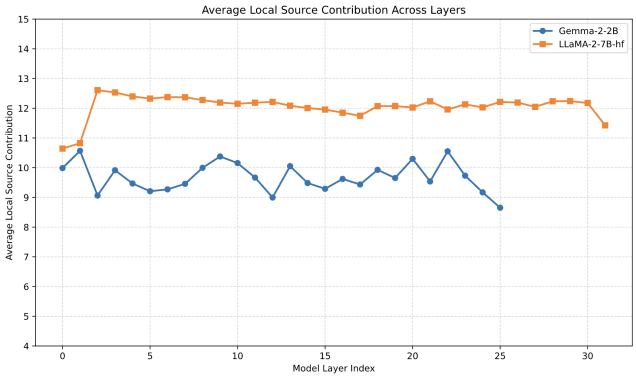


Figure 2: Average LSC across layers for Gemma-2-2B and Llama-2-7B-hf on HaluEval. Both models show flat or decaying attribution scores, failing to reflect symbolic instability or hallucination emergence.

2 Related Work

Hallucination detection in LLMs has been approached primarily as an output-level classification task. Methods like SelfCheckGPT [17] detect hallucinations in zero-resource, black-box settings by generating multiple outputs and checking for consistency across responses. Likewise, HaluEval [12] offers an extensive benchmark for detecting semantic hallucinations with annotations for truthfulness curated by humans. These techniques are post-hoc—they examine the produced text without investigating the underlying processes that lead to hallucination. This limits their ability to diagnose where or why hallucination arises in the model's computation graph. Recent work also shows that the internal state of LLMs can signal when they are likely to hallucinate [1], but even these insights are not tied to localization within layers.

Efforts to reduce hallucination typically rely on prompt engineering, retrieval augmentation, or decoding interventions. For example, De-hallucination via formal methods proposes iterative prompting using formal logical constraints to steer outputs [5]. Retrieval-Augmented Generation (RAG) approaches, such as Check Your Facts [20] and Chain-of-Knowledge [13], enhance factuality by grounding model responses in external knowledge. While these methods improve output correctness, they still treat hallucination as an emergent artifact of generation rather than as a structural failure distributed across the layers within the LLMs. Even mitigation lacks fine-grained control at the representational level.

An expanding collection of studies explores hallucination via the perspectives of symbolic reasoning and abstraction [8]. For example,

Zheng et al. [31] emphasizes that hallucinations frequently arise when LLMs do not reason symbolically regarding logical elements such as negation or exceptions. Likewise, Su et al. [23] recognizes entity mentions as a significant cause of symbolic misinterpretation, urging the modeling of semantically aware symbolic properties. However, these approaches do not systematically localize the problem across model layers; they focus on detecting symptoms rather than tracing internal breakdowns of symbolic encoding.

Mechanistic interpretability attempts to reverse-engineer neural models by identifying internal components, such as neurons, attention heads, or layers, responsible for specific behaviors. Lad [10] situates this approach as central to quantitative AI safety, highlighting that identifying causal circuits in models enables more targeted intervention and reliable analysis of failure modes, especially in tasks that demand factuality and truthfulness. Building further on this, García-Carrasco et al. [3] outlines a framework where mechanistic analysis is used to identify and learn about vulnerabilities in LLMs and demonstrate that hallucinations can stem from certain subcomponents or layers that misrepresent symbolic or factual inputs. Such representational failures, albeit enlightening, are model-specific and fail to generalize across model types and sizes.

While both activations and attention weights have been used in mechanistic interpretability, our study prioritizes attention for symbolic localization. Activation-based methods, such as patching or variance analysis, provide useful signals but are high-dimensional and difficult to link directly to symbolic cues [6]. In contrast, attention explicitly encodes how models allocate focus to input tokens, offering a more interpretable proxy for whether symbolic elements—such as negation or exceptions—are consistently represented. Prior work also shows that unstable attention over semantically critical tokens often precedes factual errors, reinforcing attention's diagnostic value for hallucination analysis [7, 19, 32].

Building on this, our contribution introduces a symbolic localization framework that measures attention variance over symbolic cues across layers. This enables scalable, layer-wise diagnosis of hallucination grounded in mechanistic reasoning. By focusing on symbolic triggers, we move beyond surface-level output inspection and instead identify the structural causes of hallucination within the model's internal computation.

3 SymLoc Methodology

Our methodology consists of three main components, summarized in Figure 3: (i) dataset preparation and task conversion, (ii) symbolic property identification, and (iii) hallucination evaluation strategy. This structured pipeline allows us to systematically analyze how symbolic triggers destabilize model reasoning and trace where symbolic instability emerges within the model.

3.1 Dataset Preparation and Task Conversion

We evaluate five open-weight models—Gemma-2B, 9B, and 27B (Google DeepMind) [24], and Llama-2–7B-hf and Llama-3.1-8B (Meta) [25]—using two widely adopted hallucination benchmarks: HaluEval (1000 samples) and TruthfulQA (817 samples). Both datasets contain factual question-answer pairs, which we directly use as ground truth for evaluating hallucination. To robustly test hallucination under varying reasoning demands, we convert each dataset into three formats: (i) QA Format: Open-ended structure, exposing models to maximal freedom and highest hallucination risk, (ii) MCQ Format: Items reformulated into multiple-choice questions with one correct and two distractors, restricting generative freedom, and (iii) Odd-One-Out (OOO) Format: Lists conceptually related items with one semantic outlier, testing reasoning beyond factual recall. This multi-format evaluation allows us to assess hallucination consistently across open-ended, constrained, and comparative reasoning tasks.

This results in a symbolic hallucination testbed of 5451 input instances (1817 samples × 3 formats). Each transformation preserves symbolic cues while altering the format to isolate hallucination behavior under different cognitive demands. To ensure consistency and eliminate sampling variability, we use each model's recommended decoding settings (typically low or zero temperature) [15]. Standardized prompts were used for each format:

- **QA Prompt:**

  > "Answer the following question in one short, factual sentence"

- **MCQ Prompt:**

  > "You are a multiple-choice quiz solver. Read the question and select only the correct option (A, B, or C)"

- **Odd One Out Prompt:**

  > "You are an expert in reasoning and comparison. Your task is to identify the odd one out from a list of three options. Only one option is unrelated to the others. Clearly state the odd option and explain why it is different."

These templates reduce ambiguity, ensuring hallucinations are attributable to internal reasoning failures rather than prompt artifacts.

3.2 Property Identification and Categorization

Each input was annotated with one or more of five symbolic properties known to challenge language models. Properties were identified using rule-based extraction, POS tagging, NER, dependency parsing, and manual verification for contextual accuracy. All automatic analyses were performed using `spaCy` with the *en_core_web_sm* model. Below, we outline each property with examples, its role in hallucination and the method used to detect it in the samples:

(1) **Modifiers (adjectives, adverbs, descriptive verbs):** These components provide subjective or comparative details that enhance interpretative latitude. *Example: "Which is the least controversial reform introduced in the past decade?"*. Words like "least" or "controversial" prompt nuanced judgments. Words like least or controversial trigger unverifiable assertions, increasing hallucination risk. To identify these modifiers, each sentence was analyzed using POS tagging, and tokens with tags `ADJ` (Adjectives), `ADV` (Adverbs), or `VERB` (Verbs) were flagged.

Table 1: Symbolic hallucination percentage statistics for original QA task across model sizes and datasets.

Symbolic Property	Gemma-2-2B		Gemma-2-9B		Gemma-2-27B		Llama-2-7B-hf		Llama-3.1-8B	
	HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA
Modifiers	84.42	99.74	79.44	99.22	77.24	86.19	73.31	90.93	82.55	91.45
Named Entities	83.58	99.69	78.52	99.38	76.43	88.46	73.09	93.85	81.85	94.46
Numbers	85.38	100	80.42	100	76.32	94.00	76.76	91.04	81.98	92.54
Negation	82.29	99.07	74.86	99.07	80.00	95.83	70.29	95.33	81.14	96.26
Exceptions	90.91	100	81.82	100	80.00	90.74	76.77	94.57	83.84	93.48

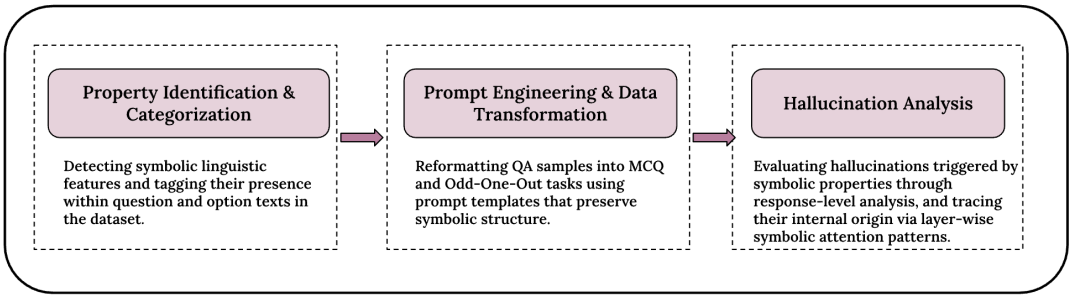


Figure 3: Structured evaluation pipeline. Benchmarks (HaluEval and TruthfulQA) are symbolically annotated and reformulated into QA, MCQ, and odd-one-out tasks. Models are then evaluated with attention analysis across layers to trace the origin of hallucinations

.

(2) **Named Entities (people, places, organizations):** This includes proper nouns that require grounding in external factual knowledge. *Example: "What position does Dr. Elena Foster hold at NovaGen Institute?"*. Rare or unfamiliar entities often cause hallucinated roles, affiliations, or identities. These entities were extracted using Named Entity Recognition (NER), identifying tokens labeled as `PERSON` (Person), `ORG` (Organization), `GPE` (Geopolitical entity), or `LOC` (Location).

(3) **Numbers (quantitative expressions, dates, counts):** This category includes cardinal numbers, measurements, dates, and quantities. *Example: "In what year did the Andean Treaty come into force?"*. Models often fabricate numeric details when data is inconsistent or missing. Numbers were identified by checking tokens for the `like_num` attribute, allowing systematic detection of all numeric references in the samples.

(4) **Negation (logical operators such as "not", "never", "none"):** This negation inverts the expected logical structure of a sentence. *Example: "Which of these countries is not a member of the Arctic Council?"*. Negation cues are often ignored, leading to logically inconsistent outputs. Identification was performed by matching tokens against a predefined set of negation keywords, including "not," "never," "none," "cannot," and several others, ensuring that all negation cues were captured in context.

(5) **Exceptions (rule violaters, rare cases, uncommon scenarios):** These inputs evaluate the model's capacity to surpass generalizations or manage counterexamples. *Example: "Which mammal lays eggs instead of giving birth to live offspring?"*. Exceptions demand contextual understanding beyond statistical regularities. Models often default to the common case rather than recognizing factual outliers. Exception cues were identified by searching for keywords such as "except," "but," "excluding," "however," and "although," which mark deviations from the general pattern, supplemented by manual review to confirm rare-case contexts.

By categorizing prompts according to these symbolic properties, we aim to isolate symbolic triggers that systematically correlate with hallucination across different tasks, prompt formats, and model scales. This symbolic framework enables a controlled investigation into the semantic and syntactic weaknesses underlying hallucination behavior in LLMs.

3.3 Hallucination Localization Strategy

We apply a two-tier evaluation strategy: symbolic trigger-driven hallucination scoring and analysis of layer-wise representational variance.

3.3.1 Symbolic Trigger-Based Hallucination Percentage. We compute the hallucination frequency for each symbolic property by evaluating model outputs corresponding to inputs that contain a specific symbolic trigger. A symbolic hallucination occurs when the model produces a confidently incorrect output in response to an input annotated with that symbolic property. The hallucination percentage for a symbolic category S is defined as:

$$\text{Hallucination}_S = \frac{\text{hallucinated outputs for inputs with } S}{\text{total inputs containing } S} \times 100 \tag{2}$$

This metric quantifies how often the presence of a symbolic trigger in the input leads to hallucinated responses, and is computed independently for each symbolic property across all models and task formats.

3.3.2 Layer-Wise Symbolic Attention Analysis. To understand how symbolic properties are handled within model layers, we compute the median attention received by symbolic tokens at each transformer layer. For each symbolic category S, we first identify its corresponding tokens in the input sequence using part-of-speech tags, named entity recognition, and keyword lists.

Let T denote the number of tokens in the input sequence. For each layer l and head h, the attention matrix is represented as $A_l^{(h)} \in \mathbb{R}^{T\times T}$. Let $\mathcal{T}_S$ be the set of token positions corresponding to symbolic property S. We calculate the average attention received by symbolic tokens at layer l as:

$$a_l^{(S)} = \frac{1}{|H| \cdot |T| \cdot |\mathcal{T}_S|} \sum_{h=1}^{H} \sum_{i=1}^{T} \sum_{j\in\mathcal{T}_S} A_l^{(h)}[i,j] \tag{3}$$

This process is repeated across four independent iterations[1] to ensure robustness. For each symbolic category and each layer, we report the *median* of all collected attention values:

$$\text{MedianAttention}_l^{(S)} = \text{median}(\{a_l^{(S)}\}_{\text{samples, iter}}) \tag{4}$$

We also compute the *standard deviation* across the same values to quantify variability:

$$\text{SD}_l^{(S)} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(a_{l,k}^{(S)} - \mu_l^{(S)}\right)^2} \tag{5}$$

where N is the total number of symbolic attention values collected for layer l and property S, and $\mu_l^{(S)}$ is the sample mean.

This layer-wise symbolic attention profiling helps identify where in the model symbolic information is concentrated, and how consistently it is attended to across samples. High median attention in early layers may suggest symbolic salience, while high variance could signal representational instability associated with hallucination.

These metrics enable us to trace how symbolic properties are weighted across model depth, and identify early-layer patterns associated with hallucination onset.

In deep learning, inconsistent attention over semantically meaningful tokens—such as negation cues or exception clauses—can lead to fragmented internal representations that degrade the quality of encoding and impair factual grounding. When symbolic cues that encode logical structure (e.g., "not", "unless", named entities) are attended to unevenly across layers, the model may struggle to build coherent meaning, increasing the likelihood of hallucinated outputs. Research in mechanistic interpretability has shown that unstable attention patterns over such critical tokens often precede factual generation errors, indicating that symbolic attention variance may not simply correlate with hallucination, but reflect a deeper representational failure within the model's internal circuitry [3, 10].

4 Results and Analysis

We organize our results around four central themes: Symbolic Hallucination Persists Across Model Scales, Symbolic Triggers Robust to Input Length, Attention-Level Traces of Symbolic Instability, and Symbolic Localization Across Model Layers.

[1]Independent iterations implies that each iteration is executed in isolation, with the model reset prior to execution. Thus, no information regarding generations or answers from earlier iterations is accessible in subsequent runs.

4.1 Symbolic Hallucination Persists Across Model Scales

In the QA format, modifiers, named entities, and negation consistently emerge as the most hallucination-prone symbolic properties across both Gemma and Llama models. As shown in Table 1, Gemma-2-2B exhibits high hallucination rates for modifiers (84.76%), named entities (83.87%), and negation (82.29%) on the HaluEval dataset. These rates decline only slightly with scale—dropping to 77.24%, 76.43%, and 80.00%, respectively, in Gemma-2-27B—indicating that scaling alone offers limited improvement.

This trend holds for TruthfulQA as well, where symbolic hallucination remains elevated across all model sizes. Modifiers peak at 94.98% in Gemma-2-9B, while negation remains above 95% across the board (e.g., 99.07% in both Gemma-2-2B and 9B). Even categories with fewer instances, such as exceptions, show stubbornly high hallucination rates in TruthfulQA (e.g., 95.45% in Gemma-2-2B).

A similar pattern is observed in the Llama series. Llama-3.1–8B records symbolic hallucination rates above 80% for modifiers, named entities, and negation in both datasets, with negligible reduction compared to smaller models.

4.2 Symbolic Triggers Robust to Input Length

To understand whether symbolic hallucination varies with query complexity, we analyze how hallucination rates change across different input lengths. Table 2 shows average hallucination percentages for each symbolic property—modifiers, named entities, numbers, negation, and exceptions—grouped by token length brackets.

We observe that hallucination rates remain consistently high across input lengths, with modifiers and named entities being especially vulnerable. For instance, modifier-triggered hallucinations peak at 96.87% for short inputs (0–29 tokens) and remain above 88% in the 30–39 token range. Named entities follow a similar trend, reaching nearly 78% in short queries—lengths that closely resemble everyday user inputs. Although longer inputs (40+ tokens) occasionally exhibit reduced hallucination, likely due to the availability of richer context, the effect is inconsistent, particularly for negation and exceptions. Reported values of 0% indicate the absence of a given symbolic property in that bin rather than true model accuracy. Taken together, these results show that symbolic hallucination persists irrespective of input length, suggesting that its origin lies in how symbolic cues are internally encoded within the model rather than being a simple artifact of brevity.

4.3 Attention-Level Traces of Symbolic Instability

To understand how task format impacts symbolic representation, we analyze symbolic token attention across QA, MCQ, and Odd-One-Out (OOO) prompts using the Gemma and Llama model families. Table 3 reports average attention to symbolic tokens at mid-to-late layers. For layer selection, we follow prior work highlighting the importance of mid-to-late transformer layers for semantic integration and context sensitivity [16, 28] [2]. Accordingly, we probed

[2]Layer indices are adopted from prior studies focusing on internal transformer layers. In our analysis, these layers provide insight into how symbolic cues are heuristically encoded. However, we additionally examined earlier layers and found that symbolic hallucinations first tend to emerge there.

Table 2: Hallucination % by symbolic property and question token length across Gemma and Llama models and datasets (HaluEval, TruthfulQA). For HaluEval, hallucination values for the exception property are 0 in all length bins except 10–19 and 30–39 due to limited occurrences. For TruthfulQA, hallucination % is 0 in the 50+ length bin, as all questions were shorter than 50 tokens.

Query Token Length	Symbolic Property	Gemma-2-2B		Gemma-2-9B		Gemma-2-27B		Llama-2-7B-hf		Llama-3.1-8B	
		HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA
0–9	Modifiers	63.73	90.07	65.69	90.07	60.78	90.07	56.86	80.60	68.63	81.06
	Named Entities	49.02	29.33	47.06	29.33	46.08	29.33	45.10	27.02	54.90	27.25
	Numbers	10.78	5.77	9.80	5.77	9.80	5.77	9.80	4.85	10.78	5.08
	Negation	1.96	7.16	0.98	7.16	1.96	7.16	0.98	6.47	2.94	6.47
	Exceptions	0.98	9.01	0.98	9.01	0.98	9.01	0.98	8.55	0.98	8.08
10–19	Modifiers	83.78	98.77	79.10	98.77	78.26	99.38	73.75	92.92	82.61	92.92
	Named Entities	66.22	45.54	63.04	45.85	61.20	45.85	58.53	44.31	65.22	43.69
	Numbers	26.59	7.08	25.75	7.08	24.41	7.08	24.58	6.77	26.42	6.77
	Negation	11.54	16.31	11.04	16.62	11.04	16.62	10.37	16.62	11.71	16.62
	Exceptions	7.53	13.54	7.36	13.54	7.36	13.54	6.86	12.62	7.36	12.92
20–29	Modifiers	83.25	100.00	76.85	95.56	79.31	97.78	72.91	93.33	80.79	95.56
	Named Entities	76.85	88.89	70.94	84.44	73.40	86.67	68.47	82.22	74.88	86.67
	Numbers	48.77	33.33	45.32	33.33	46.31	33.33	44.33	33.33	46.31	33.33
	Negation	23.15	40.00	20.69	37.78	19.70	37.78	19.70	35.56	22.17	37.78
	Exceptions	13.30	15.56	11.33	15.56	10.84	15.56	10.84	15.56	12.81	15.56
30–39	Modifiers	81.13	100.00	73.58	90.00	69.81	90.00	64.15	70.00	71.70	80.00
	Named Entities	77.36	70.00	69.81	70.00	66.04	70.00	60.38	60.00	67.92	70.00
	Numbers	62.26	20.00	58.49	20.00	56.60	20.00	50.94	20.00	56.60	20.00
	Negation	28.30	40.00	22.64	40.00	24.53	40.00	20.75	40.00	24.53	40.00
	Exceptions	15.09	20.00	13.21	20.00	11.32	20.00	13.21	20.00	13.21	20.00
40–49	Modifiers	77.78	100.00	59.26	66.67	62.96	66.67	48.15	33.33	59.26	33.33
	Named Entities	74.07	33.33	59.26	33.33	59.26	33.33	48.15	0.00	59.26	0.00
	Numbers	51.85	33.33	40.74	33.33	44.44	33.33	37.04	0.00	44.44	0.00
	Negation	18.52	0.00	14.81	0.00	18.52	0.00	14.81	0.00	18.52	0.00
	Exceptions	22.22	0.00	14.81	0.00	11.11	0.00	11.11	0.00	14.81	0.00
50+	Modifiers	82.35	100.00	82.35	100.00	88.24	100.00	70.59	100.00	76.47	100.00
	Named Entities	82.35	100.00	82.35	100.00	88.24	100.00	70.59	100.00	76.47	100.00
	Numbers	64.71	100.00	58.82	100.00	64.71	100.00	58.82	100.00	52.94	100.00
	Negation	35.29	0.00	35.29	0.00	41.18	0.00	9.41	0.00	35.29	0.00
	Exceptions	17.65	0.00	11.76	0.00	17.65	0.00	11.76	0.00	5.88	0.00

Layers 10 and 20 for Gemma-2-2B, Layers 20 and 31 for Gemma-2-9B, Layers 23 and 40 for Gemma-2-27B, Layers 15 and 28 for Llama models. This consistency across scales enables a meaningful comparison of symbolic attention dynamics.

Across models and symbolic properties, QA consistently receives higher symbolic attention than MCQ and OOO. For example, in Gemma-2-27B, attention to modifiers drops from 0.0078 in QA to 0.0063 in MCQ, and slightly rises to 0.0085 in OOO. Similarly, named entity attention is highest in QA (0.0165), lower in MCQ (0.0116), and lowest in OOO (0.0106). This pattern holds across other properties like negation and exceptions. Llama models follow similar trends. In Llama-3.1–8B, attention to negation drops from 0.0183 in QA to 0.0155 in MCQ and 0.0083 in OOO. MCQ prompts also show the steepest attention decay across layers, suggesting reduced symbolic retention under more constrained formats. These results suggest that symbolic instability is task-sensitive—with QA encouraging stronger symbolic grounding, while MCQ and OOO dilute symbolic focus, potentially amplifying hallucination risk in deeper layers.

To further probe where symbolic cues begin to weaken, we analyze Table 4, which contrasts the attention assigned to symbolic tokens against the highest-attended token at the same layer. Strikingly, across Gemma and Llama variants, the first instability appears consistently in the early layers (L3–L4), irrespective of property type. At these layers, we observe a sudden drop in attention to symbolic tokens, with their weights overshadowed by more generic or functional tokens. For instance, in the negation example ("Which of these countries is not a member of the Arctic Council?"), the negation cue "not" receives less attention (0.0451) than the interrogative "which" (0.1152). Similarly, for named entities, the symbolic sub-token with the highest attention (e.g., "Elena") is still deprioritized relative to surrounding context tokens such as "position" (0.0938 in Gemma-2-2B) or even punctuation (0.3041 in Llama-2-7B-hf). Comparable drops occur for modifiers ("least" vs. "which"), numbers ("year" vs. "in"), and exceptions ("instead" vs. "giving"), showing that logically and factually critical cues lose dominance almost immediately after embedding.

This early overshadowing suggests that symbolic instability originates during shallow encoding rather than later decoding, reinforcing our broader finding that hallucinations are structurally rooted in representational breakdowns of symbolic cues. Consistent with prior work on mechanistic interpretability, where unstable attention in early layers has been linked to factual errors [6, 21], our results position early symbolic variance as a key diagnostic signal for hallucination onset. We build on this observation in Section 4.4 by tracing layer-wise symbolic attention variance, showing how early instability propagates forward through the model.

4.4 Symbolic Localization Across Model Layers

We analyze the layer-wise origin of hallucinations in transformer-based LLMs by tracking the standard deviation of attention over symbolic tokens. The metric defined in Equation 5 quantifies how inconsistently models attend to categories such as negation, modifiers, exceptions, numbers, and named entities across layers. Elevated

Table 3: Attention of symbolic tokens at different layers for QA task across model sizes and datasets.

Task	Symbolic Property	Gemma-2-2B		Gemma-2-9B		Gemma-2-27B		Llama-2-7B-hf		Llama-3.1-8B	
		Layer 10	Layer 20	Layer 20	Layer 31	Layer 23	Layer 40	Layer 15	Layer 28	Layer 15	Layer 28
QA	Modifiers	0.0100	0.0097	0.0095	0.0092	0.0078	0.0059	0.0058	0.0015	0.0094	0.0034
	Named Entities	0.0147	0.0082	0.0168	0.0095	0.0165	0.0063	0.0132	0.0033	0.0175	0.0032
	Numbers	0.0114	0.0056	0.0122	0.0060	0.0117	0.0047	0.0089	0.0082	0.0137	0.0018
	Negation	0.0172	0.0091	0.0182	0.0070	0.0137	0.0062	0.0084	0.0030	0.0183	0.0031
	Exceptions	0.0166	0.0118	0.0134	0.0101	0.0158	0.0072	0.0076	0.0048	0.0120	0.0031
MCQ	Modifiers	0.0093	0.0068	0.0084	0.0067	0.0063	0.0039	0.0046	0.0011	0.0078	0.0024
	Named Entities	0.0177	0.0051	0.0134	0.0062	0.0116	0.0040	0.0098	0.0021	0.0123	0.0021
	Numbers	0.0104	0.0038	0.0095	0.0043	0.0085	0.0030	0.0203	0.0216	0.0106	0.0012
	Negation	0.0206	0.0067	0.0147	0.0052	0.0103	0.0042	0.0077	0.0021	0.0155	0.0022
	Exceptions	0.0140	0.0083	0.0107	0.0072	0.0121	0.0050	0.0056	0.0032	0.0093	0.0022
OOO	Modifiers	0.0076	0.0071	0.0077	0.0062	0.0085	0.0040	0.0049	0.0017	0.0087	0.0022
	Named Entities	0.0087	0.0055	0.0123	0.0055	0.0106	0.0035	0.0052	0.0015	0.0093	0.0019
	Numbers	0.0050	0.0031	0.0074	0.0032	0.0063	0.0052	0.0187	0.0196	0.0053	0.0017
	Negation	0.0092	0.0068	0.0084	0.0054	0.0082	0.0035	0.0074	0.0014	0.0083	0.0016
	Exceptions	0.0072	0.0047	0.0070	0.0064	0.0085	0.0035	0.0055	0.0016	0.0102	0.0016

variance is interpreted as representational instability, a potential precursor to hallucination.

Across Llama-2-7B-hf, Llama-3.1-8B, and Gemma models, a consistent spike in symbolic variance emerges within Layers 2–4 as illustrated in Figure 4. This instability is particularly pronounced for negation and exception tokens, which are strongly linked to hallucinated outputs. In Llama-2-7B-hf, symbolic volatility peaks in early layers, stabilizes mid-model, and reappears post Layer 24, especially for negation and modifiers, suggesting delayed symbolic misalignment. Llama-3.1-8B shows a secondary rise between Layers 10–17, again dominated by negation. All models display a late-stage surge at Layer 32 across categories, likely reflecting output dynamics rather than causal instability.

Gemma-2-2B exhibits the strongest early variance, with Layer 2 peaking near 0.17 standard deviation, while deeper layers stabilize around 0.05–0.08. The Gemma-2-9B follows a similar trend with peak variance around 0.15, and both Llama models show nearly identical behavior (0.16–0.17). Statistical analysis reveals high correlation across models ($r > 0.8$, $p < 0.001$), suggesting this is not coincidental but a structural property of transformer architectures [18, 27].

This concentration of variance in Layers 2–4 is significant because these layers perform critical contextualization tasks: moving beyond token embedding (Layer 1) toward syntax and semantic grounding [21, 26]. Instability at this stage propagates through residual connections, compromising later reasoning and answer extraction. Recent findings attribute this to difficulties in modeling short-range dependencies, where unstable attention scores sharpen excessively, producing "attention entropy collapse" [4, 30].

Mechanistic studies confirm that factual answering relies on early-layer knowledge enrichment followed by later-layer extraction [6]. When early symbolic variance destabilizes retrieval, later layers confidently propagate flawed representations, yielding hallucinations. Notably, scaling (e.g., Gemma-2-9B vs 2B) dampens but does not eliminate the early variance, reinforcing that this instability is intrinsic to transformer design.

These observations suggest that hallucinations are not random generation artifacts but arise from systematic symbolic misalignment in early encoding layers. Symbolic attention variance in Layers 2–4 thus provides a reliable, model-agnostic diagnostic for hallucination localization, enabling proactive detection and targeted architectural interventions.

Taken together, the results across model scaling, input length, task format, and layer-wise analysis reveal a consistent picture of symbolic hallucination. Variations in architecture, parameter count, or prompt design provide only marginal relief, underscoring that hallucination is not a surface artifact but a structural property of symbolic representation in LLMs. Modifiers, named entities, and negation consistently act as dominant triggers, while early transformer layers (2–4) emerge as the locus of representational instability. This instability propagates forward, yielding confident but factually incorrect outputs.

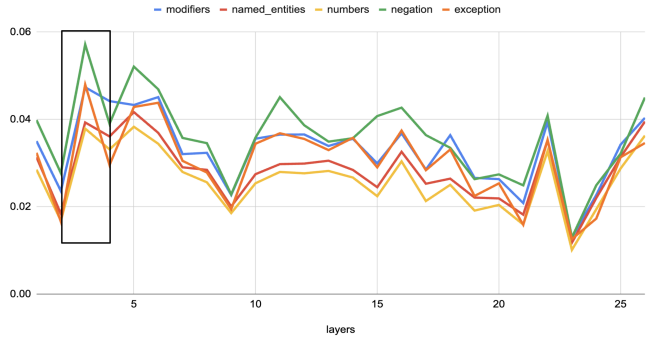
By reframing symbolic hallucination as a structural misalignment rather than a decoding artifact, our analysis highlights symbolic attention variance as a reliable diagnostic signal—offering a principled basis for proactive detection and architectural interventions.

5 Conclusion and Future Directions

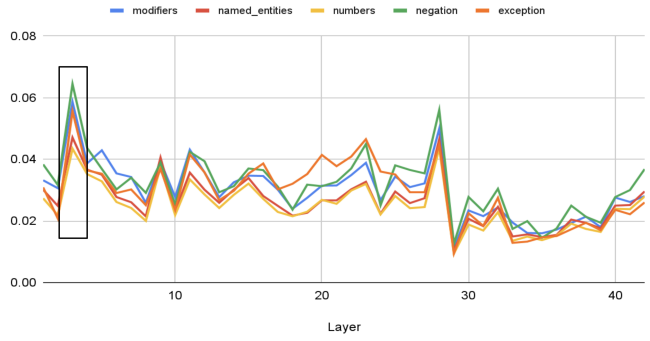
This work introduces **SymLoc**, a symbolic localization framework designed to trace the internal emergence of hallucinations in large language models. By employing symbolic attention variance as a layer-wise metric, we demonstrate that hallucinations often originate from early representational instabilities, particularly when models process symbolic properties such as negation, modifiers, and named entities. Across multiple open-weight models and benchmark datasets, our analysis reveals consistent volatility in early-layer attention, reframing hallucination as a structure-sensitive and symbolically grounded failure. Crucially, **SymLoc** uncovers both (i) regions where hallucinations are later self-corrected—consistent with prior evidence—and (ii) their initial emergence in early layers, thereby offering an interpretable pathway for detection and correction. The framework's interpretability arises from leveraging symbolic knowledge to localize and characterize hallucinations within the internal representations of LLMs. Overall, these findings suggest that hallucinations stem less from surface-level decoding

Table 4: Layer localization of symbolic instability with comparison between attention on symbolic tokens and the highest-attended token at that layer. For each model and property, we report the first instability layer, the symbolic token with its attention value, and the maximum-attended token with its value. For multi-token named entities, the token with the highest attention is shown. "Sym" denotes the symbolic token.

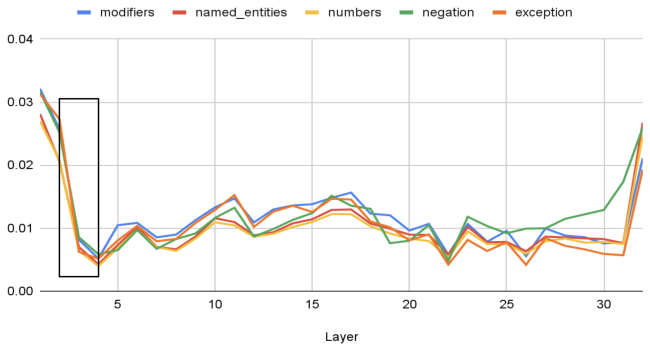
Symbolic Property	Example	Gemma-2-2B	Gemma-2-9B	Llama-2-7B-hf	Llama-3.1-8B
Modifiers	Which is the ***least*** controversial reform introduced in the past decade?	**L3** Sym=least(0.0468) Max=which(0.0925)	**L3** Sym=least(0.0531) Max=which(0.0834)	**L3** Sym=least(0.0134) Max=which(0.0200)	**L3** Sym=least(0.0100) Max=which(0.0160)
Named Entities	What position does Dr. ***Elena Foster*** hold at ***NovaGen*** Institute?	**L4** Sym=Elena(0.0308) Max=position(0.0938)	**L3** Sym=Elena(0.0530) Max=position(0.0700)	**L3** Sym=Nova(0.0140) Max=.(0.3041)	**L3** Sym=Elena(0.0024) Max=position(0.0080)
Numbers	In what ***year*** did the Andean Treaty come into force?	**L3** Sym=year(0.0385) Max=in(0.0669)	**L3** Sym=year(0.0222) Max=come(0.0751)	**L3** Sym=year(0.0083) Max=in(0.0105)	**L3** Sym=year(0.0116) Max=did(0.0117)
Negation	Which of these countries is ***not*** a member of the Arctic Council?	**L3** Sym=not(0.0451) Max=which(0.1152)	**L3** Sym=not(0.0306) Max=which(0.1054)	**L3** Sym=not(0.0060) Max=which(0.0228)	**L3** Sym=not(0.0060) Max=of(0.0166)
Exceptions	Which mammal lays eggs ***instead*** of giving birth to live offspring?	**L3** Sym=instead(0.0492) Max=giving(0.0681)	**L3** Sym=instead(0.0472) Max=giving(0.0617)	**L3** Sym=instead(0.0052) Max=which(0.0152)	**L3** Sym=instead(0.0060) Max=which(0.0136)



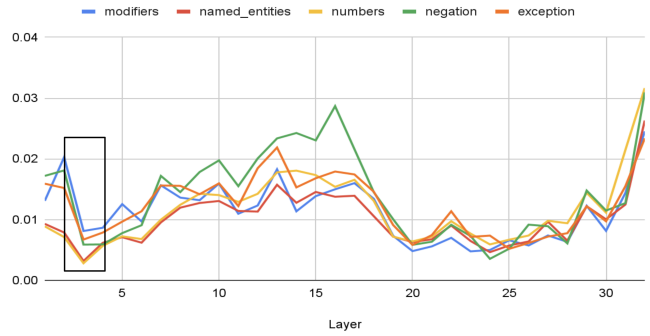
(a) Gemma-2-2B (Total Layers-26)



(b) Gemma-2-9B (Total Layers-42)



(c) Llama-2-7B-hf (Total Layers-32)



(d) Llama-3.1-8B (Total Layers-32)

Figure 4: Standard deviation of symbolic attention across transformer layers for Llama and Gemma models. Early-layer variance (Layers 2–4) indicates symbolic instability that may contribute to hallucination onset. Layers 2–4 are boxed to highlight the earliest emergence and propagation of symbolic misalignment.

or model scale, and more from the internal misrepresentation of symbolic cues.

Building on this framework, future work will explore several directions. First, investigating the specific layers and attention heads associated with symbolic confusion. Second, examining the generalization of symbolic vulnerabilities across diverse model families, including GPT, Mistral, and multilingual or multimodal LLMs. Third, developing symbolic interventions at the prompting level to mitigate ambiguity and hallucination. Finally, extending symbolic localization from the layer level to the token level for finer-grained tracing of hallucination.

In addition, to support further research and benchmarking, we will release the code upon acceptance, along with extended versions

of HaluEval and TruthfulQA containing multiple-choice and odd-one-out formats for the community.

6 Limitations

While this study sheds light on the symbolic origins of hallucination in large language models, some limitations remain. Our evaluation is limited to two English benchmarks (HaluEval and TruthfulQA), and broader multilingual and domain-specific tests are needed to assess generalizability. Some samples contain overlapping symbolic properties, making it difficult to isolate triggers and potentially inflating cross-category correlations. Finally, we focus only on open-weight transformer models (Gemma and Llama), leaving open whether similar vulnerabilities appear in proprietary or non-transformer architectures.

References

[1] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. doi:10.48550/arXiv.2304.13734 arXiv:2304.13734 [cs].

[2] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. doi:10.48550/arXiv.2309.03883 arXiv:2309.03883 [cs].

[3] Jorge García-Carrasco, Alejandro Maté, and Juan Trujillo. 2024. Detecting and Understanding Vulnerabilities in Language Models via Mechanistic Interpretability. (7 2024), 385–393. doi:10.24963/ijcai.2024/43

[4] Suvadeep Hajra. 2025. Short-Range Dependency Effects on Transformer Instability and a Decomposed Attention Solution. *arXiv preprint arXiv:2505.15548* (2025).

[5] Susmit Jha, Sumit Kumar Jha, Patrick Lincoln, Nathaniel D. Bastian, Alvaro Velasquez, and Sandeep Neema. 2023. Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting. In *2023 IEEE International Conference on Assured Autonomy (ICAA)*. IEEE, 149–152. doi:10.1109/ICAA57998.2023.00033

[6] Lei Jin, Sean Welleck, Luca Pajola, Eunsol Choi, and Chandan Suresh. 2024. Mechanistic Understanding and Mitigation of Language Model Non-Factual Hallucinations. *arXiv preprint arXiv:2403.18167* (2024).

[7] Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, and Ashutosh Modi. 2024. Towards Robust Evaluation of Unlearning in LLMs via Data Transformations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 12100–12119.

[8] Vedant Khandelwal, Manas Gaur, Ugur Kursuncu, Valerie L Shalin, and Amit P Sheth. 2024. A domain-agnostic neurosymbolic approach for big social data analysis: Evaluating mental health sentiment on social media during covid-19. In *2024 ieee international conference on big data (bigdata)*. IEEE, 959–968.

[9] Hazel Kim, Adel Bibi, Philip Torr, and Yarin Gal. 2024. Detecting LLM Hallucination Through Layer-wise Information Deficiency: Analysis of Unanswerable Questions and Ambiguous Prompts. (12 2024). https://arxiv.org/pdf/2412.10246

[10] Vedang K. Lad. 2024. Mechanistic Interpretability for Progress Towards Quantitative AI Safety. (2024). https://dspace.mit.edu/handle/1721.1/156748

[11] Naveen Lamba, Sanju Tiwari, and Manas Gaur. 2025. Investigating Symbolic Triggers of Hallucination in Gemma Models Across HaluEval and TruthfulQA. (9 2025). https://arxiv.org/pdf/2509.09715

[12] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. doi:10.48550/arXiv.2305.11747 arXiv:2305.11747 [cs].

[13] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024. Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources. doi:10.48550/arXiv.2305.13269 arXiv:2305.13269 [cs].

[14] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the Annual Meeting of the Association for Computational Linguistics* 1 (9 2021), 3214–3252. doi:10.18653/v1/2022.acl-long.229

[15] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation. *Advances in Neural Information Processing Systems* 36 (5 2023). https://arxiv.org/pdf/2305.01210

[16] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 11 (2023), 1577–1591. arXiv:2307.03172 [cs.CL] doi:10.1162/tacl_a_00606

[17] Potsawee Manakul, Adian Liusie, and Mark J F Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. doi:10.48550/arXiv.2303.08896 arXiv:2303.08896 [cs].

[18] Alessandro Merolla, Vijay Kumar, and Wei Chen. 2024. Layer Importance and Hallucination Analysis in Large Language Models via Enhanced Activation Variance-Sparsity. *arXiv preprint arXiv:2411.10069* (2024).

[19] Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. 364–388.

[20] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback. doi:10.48550/arXiv.2302.12813 arXiv:2302.12813 [cs].

[21] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8 (2020), 842–866.

[22] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-Check: Investigating Detection of Hallucinations in Large Language Models. *Advances in Neural Information Processing Systems* 37 (12 2024), 34188–34216.

[23] Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating Entity-Level Hallucination in Large Language Models. doi:10.48550/arXiv.2407.09417 arXiv:2407.09417 [cs].

[24] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving Open Language Models at a Practical Size. (7 2024). https://arxiv.org/pdf/2408.00118

[25] Llama Team and Ai @ Meta. 2024. The Llama 3 Herd of Models. (2024).

[26] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950* (2019).

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[28] Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. 2025. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. (1 2025). https://arxiv.org/pdf/2501.17148

[29] Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and Detecting Hallucinations in Neural Machine Translation via Model Introspection. *Transactions of the Association for Computational Linguistics* 11 (1 2023), 546–564. doi:10.1162/tacl_a_00563

[30] Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Josh Susskind, and Jaitly Navdeep. 2024. Stabilizing Transformer Training by Preventing Attention Entropy Collapse. *Apple Machine Learning Research* (2024).

[31] Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why Does ChatGPT Fall Short in Providing Truthful Answers? doi:10.48550/arXiv.2304.10513 arXiv:2304.10513 [cs].

[32] Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. Attention Heads of Large Language Models: A Survey. (9 2024). https://arxiv.org/pdf/2409.03752