Investigating Symbolic Triggers of Hallucination in Gemma Models Across HaluEval and TruthfulQA

Naveen Lamba[1,†], Sanju Tiwari[1,†] and Manas Gaur[2,‡]

[1]Center for Artificial Intelligence in Medicine, Imaging and Forensics, Sharda University, Greater Noida, India
[2]University of Maryland, Baltimore County, Baltimore, MD, USA

Abstract

Hallucination in Large Language Models(LLMs) is a well studied problem. However, the properties that make LLM intrinsically vulnerable to hallucinations have not been identified and studied. This research identifies and characterizes the key properties, allowing us to pinpoint vulnerabilities within the model's internal mechanisms. To solidify on these properties, we utilized two established datasets, HaluEval and TruthfulQA and convert their existing format of question answering into various other formats to narrow down these properties as the reason for the hallucinations. Our findings reveal that hallucination percentages across symbolic properties are notably high for Gemma-2-2B, averaging 79.0% across tasks and datasets. With increased model scale, hallucination drops to 73.6% for Gemma-2-9B and 63.9% for Gemma-2-27B, reflecting a 15 percentage point reduction overall.Although the hallucination rate decreases as the model size increases, a substantial amount of hallucination caused by symbolic properties still persists. This is especially evident for modifiers (ranging from 84.76% to 94.98%) and named entities (ranging from 83.87% to 93.96%) across all Gemma models and both datasets. These findings indicate that symbolic elements continue to confuse the models, pointing to a fundamental weakness in how these LLMs process such inputs—regardless of their scale.

Keywords

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

1. Introduction

Large language models (LLMs) have made significant advancements in various natural language understanding and generation tasks, including open-domain question answering [1], text summarization [2], reasoning [3], and dialogue [4]. Despite their success, the reliability of LLMs remains a major issue due to hallucination[1], which involves confidently generating content that is factually inaccurate or nonsensical text [5, 6].

While significant research has been conducted on identifying and reducing hallucinations in LLMs [7, 8], much of this work has been primarily driven by the development of novel hallucination benchmarks and their corresponding detection and mitigation approaches [9]. However, the investigation into the fundamental, intrinsic causes of hallucination phenomena in LLMs remains significantly underexplored. Understanding these root causes is particularly crucial because they often stem from limitations in symbolic knowledge representation and reasoning—areas where the NLP community has extensive expertise [10, 11]. These limitations manifest through specific and elemental symbolic triggers that consistently provoke hallucinations: *named entities, negation handling, exception cases*, and others can cause LLMs to generate incorrect information, irrespective of the dataset format or domain. Figure 1 illustrates two such examples where all Gemma models hallucinate in the presence of symbolic triggers like modifiers, named entity, number, negation, and exception. By focusing on these intrinsic mechanisms, researchers can develop more robust, data-agnostic methodologies that not only localize

SymGenAI4Sci 2025: First International Workshop on Symbolic and Generative AI for Science co-located with Semantics-2025, September 3–5, 2025, Vienna, Austria

[†]These authors contributed equally.

naveenlamba30894@gmail.com (N. Lamba); tiwarisanju18@ieee.org (S. Tiwari); manas@umbc.edu (M. Gaur)
http://conceptbase.sourceforge.net/mjf/ (M. Gaur)

[1]We use the word "hallucination" to be consistent with the terminology used by NLP community, however, we prefer confabulation/fabrication as the appropriate word.

the sources of hallucination within LLMs but also provide systematic, long-term solutions rather than superficial fixes. This deeper understanding would enable the creation of more reliable language models that can better distinguish between accurate and inaccurate generated content, ultimately leading to more trustworthy AI systems [12].

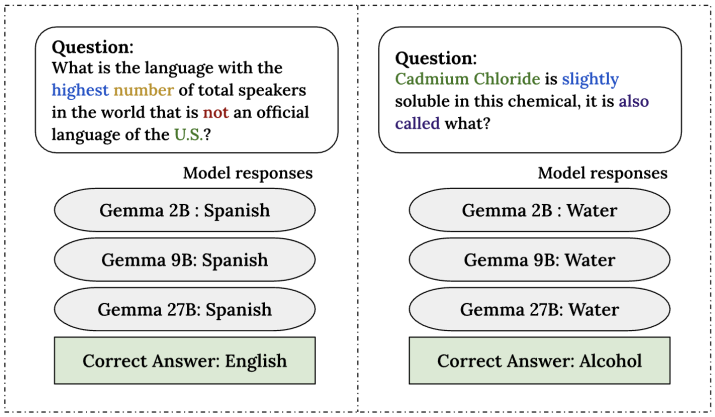

Figure 1: Examples of symbolic triggers causing hallucination across all Gemma model sizes (2B, 9B, 27B). Color coding: blue = modifier, red = negation, purple = exception, green = named entity, yellow = number.

This paper addresses this gap by identifying and describing *symbolic and interpretable* knowledge properties that *reliably trigger hallucination* across natural language understanding task types and model scales. This paper makes several key contributions which are outlined below:

Key Contributions:

- *Identification of symbolic hallucination triggers:* Systematically identified and characterized five symbolic knowledge properties that reliably trigger hallucination: modifiers (adjectives, adverbs, verbs), named entities, numbers, negation, and exceptions, and provided a property-focused evaluation for understanding intrinsic vulnerabilities in LLMs.
- *Prompt engineering-driven data transformation for generalization of symbolic triggers*: Developed a systematic evaluation approach that tests hallucination consistency across three critical dimensions: model scale (Gemma-2-2B, 9B, 27B), task formats (question-answering (QA), multiple choice questions (MCQ), Odd-One-Out (OOO)), and symbolic property types by converting existing datasets to isolate specific triggers, which demonstrates that symbolic vulnerabilities are fundamental architectural issues rather than artifacts of specific experimental conditions.
- *Internal activation analysis using symbolic triggers*: Conducted attention pattern analysis and activation-level traces to examine how symbolic properties affect internal model representations and processing, providing evidence that hallucinations stem from deeper representational instabilities rather than surface-level generation errors.

These contributions led us to the following findings:

- Symbolic triggers elicit hallucination across model sizes: Hallucination rates remain substantially high across all model sizes: 79.0% (Gemma-2-2B), 73.6% (Gemma-2-9B), and 63.9% (Gemma-2-27B), with only a modest 15 percentage point reduction despite significant model scaling, indicating that these are structural rather than capacity-related issues that challenge the assumption that larger models automatically become more reliable.
- Primary symbolic triggers: Modifiers show hallucination rates ranging from 84.76% to 94.98% across all models while named entities exhibit similarly high rates (83.87% to 93.96%), consistently emerging as the most problematic symbolic properties and revealing specific linguistic elements that pose the most significant risk for factual accuracy in LLM outputs.

- Task Format Dependency: QA format produces the highest hallucination rates compared to MCQ and Odd-One-Out formats, with lower symbolic attention values correlating with higher hallucination frequency, particularly evident in MCQ tasks, demonstrating that task structure significantly influences model reliability and suggesting that constrained generation formats may offer some protection against symbolic confusion.
- Non-monotonic input length effects of symbolic triggers: Symbolic triggers behave differently across input lengths: modifiers and named entities cause the most hallucinations in short-to-medium contexts (10-30 tokens) but become more reliable with longer context, while numbers follow an unpredictable up-and-down pattern, and negation and exceptions consistently cause fewer problems overall, demonstrating that context length affects each symbolic property uniquely.

2. Related Work

Recent advances in large language models (LLMs) have intensified focus on understanding and mitigating hallucination—confident outputs that are factually incorrect or logically incoherent. While early research primarily concentrated on output-level detection and dataset-based evaluation of hallucination phenomena[13, 14], the deeper representational vulnerabilities of LLMs remain underexplored. Our study contributes by examining how hallucination manifests across model sizes, under different task formats, and in response to symbolic properties embedded in inputs.

While a growing body of work evaluates LLMs for factual reliability, few studies assess how hallucination trends evolve with model scale. Notably, works like Yao et al. [15] frame hallucinations as emergent adversarial phenomena—linked to overconfident generalizations—but do not analyze whether such tendencies vary with parameter count. Similarly, most hallucination benchmarks focus on a single model instance rather than conducting comparative analysis across multiple versions of the same model family. Our work addresses this gap by systematically evaluating hallucination behavior across Gemma-2-2B, 9B, and 27B, revealing that while hallucination rates reduce with scale, symbolic triggers remain persistent.

Benchmark datasets such as TruthfulQA[16] and HaluEval[17] have been instrumental in evaluating LLM hallucinations. These benchmarks typically use open-ended QA to elicit model generations under minimal constraints, which often reveal factual inconsistencies. However, prior work does not systematically vary task formats to study how structural differences—like constrained generation in multiple-choice or odd-one-out tasks—modulate hallucination tendencies. Our study introduces task format as a key dimension, converting QA data into MCQ and OOO formats to probe whether and how task structure interacts with hallucination triggers.

Numerous research efforts have investigated linguistic and stylistic elements that affect hallucination. For instance, Rawte et al. [18] demonstrate that the likelihood of hallucinated outputs is influenced by readability, formality, and concreteness. Some have concentrated on particular symbolic structures. Negation has specifically been recognized as a continual vulnerability for LLMs, as shown by Varshney et al. [19] and Asher and Bhar [20], who illustrate that models often generate false information even when negation indicators are straightforward in syntax and clear in logic. These results indicate that symbolic reasoning continues to be difficult, even if many assessments are limited to individual signals. Our work broadens this scope by evaluating five symbolic properties—modifiers, named entities, numbers, negation, and exceptions—as systematic triggers of hallucination. We also extend analysis beyond surface-level generations, examining how symbolic inputs induce representational instability across transformer layers.

In contrast to prior research, which often isolates one axis of hallucination (model, task, or linguistic feature), our work offers a three-dimensional assessment across model scale, task format, and symbolic input structure. We analyze symbolic hallucination in three Gemma models (2B, 9B, 27B), three reformatted task environments (QA, MCQ, OOO), and five symbolic property types, providing both quantitative trends and internal activation-level insights [21]. This integrative approach reveals that hallucinations are not just artifacts of generation, but reflect deeper weaknesses in how LLMs process

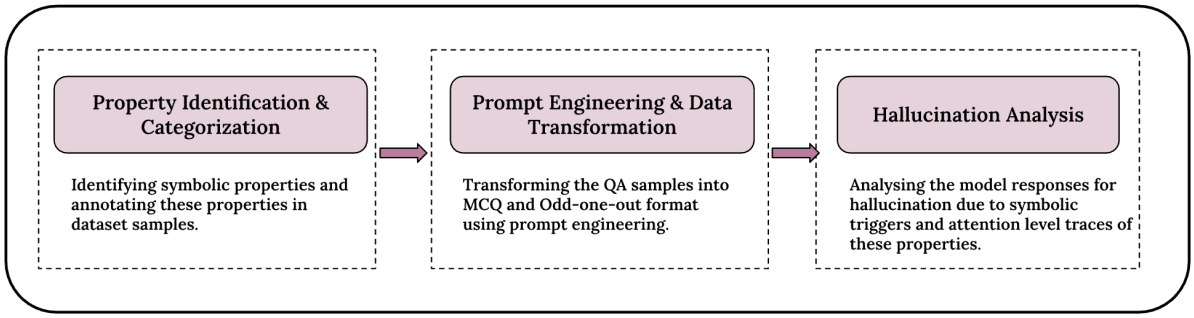

Figure 2: Methodology followed for the hallucination analysis in this research

structurally complex or logically nuanced inputs.

3. Methodology

Our approach involved taking existing datasets, converting them into different question formats, and then testing how three different sizes of Gemma models (2B, 9B, and 27B) responded to questions containing specific symbolic triggers like modifiers, numbers, and named entities. This study evaluates different versions of Gemma models, open-source checkpoints released by Google DeepMind [22, 23]. For consistency across all experiments and to minimize sampling parameter variability, we utilize each model's default temperature value, as provided by the model, which is typically a low or zero value for deterministic generation. This enables us to see the inherent behavior of each model in its recommended decoding setup without injecting sampling-originating randomness.

This research explores the inherent symbolic knowledge characteristics that induce hallucinations in LLMs, i.e., between varying instances of the Gemma model family (2B, 9B, and 27B). The approach follows a systematic, property-focused evaluation pipeline consisting of dataset setup, input transformation, model selection, controlled prompt creation, and hallucination analysis. We base our methodology on the assumption that some input symbolic structures — like modifiers, named entities, negations, numbers, and exceptions — increase the likelihood of LLMs to hallucinate. To empirically test this, we reformatted typical datasets into task-specific ones and inspected the derived outputs.

3.1. Dataset Preparation and Task Conversion

We used two established hallucination evaluation datasets, HaluEval and TruthfulQA, which contain factual question-answer pairs. To determine whether symbolic triggers cause hallucinations regardless of task structure or if certain formats offer protection against symbolic confusion, we systematically converted these datasets into three distinct formats that provide different levels of generative constraints and cognitive demands: (i) QA format preserves open-ended generation that *may* expose maximum hallucination tendencies since models can freely fabricate plausible-sounding but incorrect responses when encountering symbolic triggers, (ii) MCQ format provides constrained multiple-choice selection that tests whether limiting response options can mitigate symbolic trigger effects by preventing free-form generation, and (iii) OOO format tests semantic classification abilities under symbolic influence to determine if symbolic triggers disrupt fundamental reasoning processes beyond just factual recall. What we prepared: We systematically transformed 100 samples from each dataset (verified to contain one or more target symbolic properties) into all three task formats, creating a comprehensive evaluation framework of 600 total test instances. Each transformation maintained the core symbolic elements while adapting the response structure to isolate whether symbolic confusion persists across different cognitive demands and constraint levels. How transformation was achieved: We designed standardized prompts for each format: QA Prompt: "Answer the following question in one short, factual sentence." MCQ Prompt: "Consider the following multiple-choice question. Pick the correct answer and explain

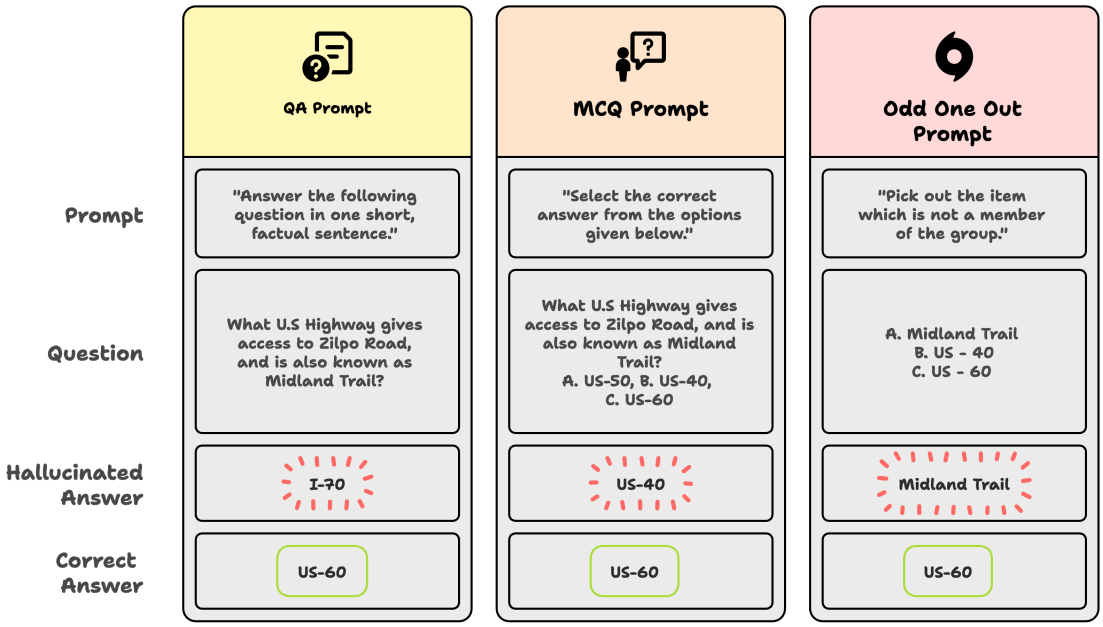

Figure 3: Examples of questions and answers in different formats(QA, MCQ and Odd-One-Out), generated from Gemma-2-2b.

your reasoning." Odd One Out Prompt: "Identify the item that does not belong in the list. Explain your reasoning."

For example:

- **QA Prompt:**

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

- **MCQ Prompt:**

  "Consider the following multiple-choice question. Pick the correct answer and explain your reasoning."

- **Odd One Out Prompt:**

  "Identify the item that does not belong in the list. Explain your reasoning."

Prompt design was carefully managed so that hallucinations, when they occur, can be attributed to model reasoning and symbolic processing rather than prompt ambiguity, with all transformed prompts annotated for symbolic property analysis.

Two standard hallucination benchmarking datasets: HaluEval[17] and TruthfulQA[16] have been used for this study. Both datasets include factual QA pairs. To investigate hallucination behavior with varying knowledge forms, we design three task formulations: (i) QA: Preserves the question-answer pair format of the original dataset. (ii) MCQ: Converts every QA pair to a single correct and two distractor options multiple-choice format. (iii) Odd-One-Out: Presents conceptually related options with the exception of one, seeking identification of the semantic outlier. All 600 samples of the two datasets are checked to have one or more of the five symbolic knowledge properties.

3.2. Property Identification and Categorization

To investigate symbolic triggers of hallucination, all input prompts were annotated for the presence of five key symbolic properties. These properties were chosen based on their structural role in language. Each property was identified using linguistic markers and then manually verified to ensure semantic relevance. Here, we define each category, provide examples, and summarize their contribution to hallucination:

1. **Modifiers(adjectives, adverbs, and verbs):** These elements introduce subjective or descriptive information, often adding interpretive flexibility. *Example: "Which is the most rapidly growing city in Europe?"*. Modifiers such as "rapidly" or "most" invite vague or ambiguous completions, increasing the risk of confident but unverifiable assertions. LLMs may hallucinate plausible-sounding answers even when the modifier-driven nuance is not grounded in training data.
2. **Named Entities(persons, organizations, locations):** Identified using Named Entity Recognition (NER) techniques, these refer to proper nouns that often require external knowledge grounding. *Example: "Who is the founder of the fictional company TechNova?"*. Due to their reliance on memorized or incomplete knowledge, LLMs often fabricate facts or assign incorrect associations when dealing with named entities—especially rare or fictional ones.
3. **Numbers(quantitative expressions):** These include cardinal numbers, ranges, dates, and measurements. *Example: "How many satellites does Mars currently have?"*. LLMs are prone to imprecision or outright numerical hallucination, either due to outdated training data or due to overgeneralizing learned patterns. Such prompts demand factual accuracy, making errors more noticeable.
4. **Negation(not, never, none, cannot):** Detected via syntactic and semantic analysis, negation alters the logical polarity of a sentence. *Example: "Which of these is not a fruit?"*. LLMs frequently mishandle negation by overlooking or misinterpreting the negative cue, resulting in logically inverted or irrelevant answers.
5. **Exceptions(edge cases, conditional rules):** These refer to inputs that challenge the model to recognize rare or counterexamples. *Example: "Which metal is liquid at room temperature?"*. Exceptions require deeper contextual reasoning. Since LLMs tend to generalize, they often miss these special cases, favoring the more common rule rather than the exception.

By categorizing prompts along these symbolic dimensions, we aim to isolate specific triggers that systematically increase hallucination likelihood across tasks and model scales. This property-level lens provides a more interpretable understanding of why and when LLMs go wrong.

3.3. Hallucination Evaluation Strategy

We employ a three-tier hallucination analysis approach, progressing from overall hallucination rates to detailed, layer-wise causes, ultimately attributing them to five symbolic triggers.

Symbolic trigger-based computation of hallucination percentage: To quantify hallucination induced by symbolic properties, we annotated each input for the presence of one or more symbolic triggers (modifiers, named entities, numbers, negation, exceptions) and computed the proportion of hallucinated outputs within each trigger category. A prediction was marked as a hallucination if it was factually incorrect. This computation was carried out per symbolic property, allowing us to isolate their individual contribution to hallucination rates. The final hallucination percentage per property was then calculated as the number of hallucinated instances containing that property divided by the total instances containing it.

Symbolic trigger-driven attention analysis of Gemma models: We analyze attention scores to symbolic tokens at specific transformer layers selected based on prior research patterns. Following Wu et al. [24]'s approach, which emphasizes mid-to-deeper layers where semantic integration peaks, we examine Layers 10 and 20 for Gemma-2-2B, Layers 20 and 31 for Gemma-2-9B, and Layers 23 and 40 for Gemma-2-27B. This allows consistent comparison of symbolic attention allocation across model sizes.

Input token length and hallucination percentage analysis: We investigate the relationship between hallucination rates and input question length by organizing data into token length bins and analyzing how symbolic property effects vary across different context sizes. This reveals whether symbolic triggers have consistent effects regardless of the surrounding context or if their impact changes with input complexity.

4. Results and Analysis

This section presents our empirical analysis of hallucination behavior in the Gemma model family under symbolic property influence. The investigation is organized along three axes: (i) consistency across model sizes, (ii) variation across task types, and (iii) internal activation responses. The evaluation spans all five symbolic property types, with hallucination annotated as confident yet factually incorrect responses.

4.1. Consistency Across Model Variants

In the QA format, modifiers, named entities, and numbers consistently emerge as the most hallucination-prone symbolic properties across all three Gemma model sizes. As shown in Table 1, hallucination percentages for modifiers in the HaluEval dataset remain notably high, decreasing only slightly from 84.76% in Gemma-2-2B to 77.24% in Gemma-2-27B. Named entities follow a similar trend, with a marginal drop from 83.87% to 76.43%, while numbers stay persistently high at around 83.16%–76.32% across model scales.

This pattern is also observed in the TruthfulQA dataset, where modifiers reach up to 94.98% in Gemma-2-9B and numbers peak at 98.00%, reflecting the models' continued struggle with these symbolic cues. On the other hand, while negation and exceptions appear less frequently in HaluEval (e.g., 70.00% and 80.00% in Gemma-2-2B), their hallucination rates remain above 90% in TruthfulQA across all model sizes.

These results indicate that scaling up model size offers only modest reductions in hallucination rates for symbolic properties, and that the same set of symbolic triggers continues to challenge LLMs, revealing a persistent internal vulnerability.

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

Symbolic Property	Gemma-2-2B		Gemma-2-9B		Gemma-2-27B	
	HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA
Modifiers	84.76	89.12	77.45	94.98	77.24	86.19
Named Entities	83.87	89.01	77.17	93.96	76.43	88.46
Numbers	83.16	96.00	75.26	98.00	76.32	94.00
Negation	70.00	91.67	70.00	95.83	80.00	95.83
Exceptions	100.0	94.44	80.00	96.30	80.00	90.74

4.2. Generalization Across Task Formats

To understand how hallucination behavior generalizes across prompt formats, we analyzed symbolic token attention across QA, MCQ, and Odd-One-Out (OOO) tasks using the Gemma model family (2B, 9B, 27B). Table 2 presents average attention scores to symbolic tokens across task formats and model sizes, measured at specific mid-to-deeper layers.

Following prior layer selection patterns used in Wu et al. [24], which emphasized mid and post-mid transformer layers (Layers 10 and 20 for Gemma-2-2B and Layers 20 and 31 for Gemma-2-9B), we chose Layers 23 and 40 for Gemma-2-27B. These lie in the middle-to-late segments of the model, where semantic integration and abstract token interactions typically peak. This alignment allows for a consistent and meaningful comparison of symbolic attention across model sizes.

The results indicate that task format substantially affects both hallucination frequency and attention allocation, despite using prompts with similar symbolic triggers. Across all model sizes, MCQ prompts result in consistently higher hallucination frequency than QA, particularly at the 2B scale. This correlates with lower symbolic attention values for MCQ compared to QA—suggesting reduced grounding or interpretive focus. For instance, in the 27B model, attention to modifiers in QA is 0.0078 (Layer 23), dropping to 0.0063 in MCQ, and further varying in OOO (0.0085). This indicates task-specific shifts in symbolic emphasis, even within the same model.

Table 2
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	
		Layer 10	Layer 20	Layer 20	Layer 31	Layer 23	Layer 40
QA	Modifiers	0.0100	0.0097	0.0095	0.0092	0.0078	0.0059
	Named Entities	0.0147	0.0082	0.0168	0.0095	0.0165	0.0063
	Numbers	0.0114	0.0056	0.0122	0.0060	0.0117	0.0047
	Negation	0.0172	0.0091	0.0182	0.0070	0.0137	0.0062
	Exceptions	0.0166	0.0118	0.0134	0.0101	0.0158	0.0072
MCQ	Modifiers	0.0093	0.0068	0.0084	0.0067	0.0063	0.0039
	Named Entities	0.0177	0.0051	0.0134	0.0062	0.0116	0.0040
	Numbers	0.0104	0.0038	0.0095	0.0043	0.0085	0.0030
	Negation	0.0206	0.0067	0.0147	0.0052	0.0103	0.0042
	Exceptions	0.0140	0.0083	0.0107	0.0072	0.0121	0.0050
OOO	Modifiers	0.0076	0.0071	0.0077	0.0062	0.0085	0.0040
	Named Entities	0.0087	0.0055	0.0123	0.0055	0.0106	0.0035
	Numbers	0.0050	0.0031	0.0074	0.0032	0.0063	0.0052
	Negation	0.0092	0.0068	0.0084	0.0054	0.0082	0.0035
	Exceptions	0.0072	0.0047	0.0070	0.0064	0.0085	0.0035

Conversely, while OOO prompts show relatively lower symbolic attention, they elicit stronger semantic hallucination effects, particularly in smaller models (as seen in prior hallucination rate and effect metrics). Notably, in the 2B model, symbolic attention for named entities drops sharply in MCQ (0.0177 $\rightarrow$ 0.0051 from Layer 10 to 20), whereas QA retains higher symbolic focus (0.0147 $\rightarrow$ 0.0082). The same trend, though attenuated, persists in 27B, showing a consistent symbolic property ranking: modifiers and named entities receive the highest attention, followed by numbers, negation, and exceptions.

4.3. Activation-Level Traces of Symbolic Instability

To further probe the internal behavior of LLMs in the presence of symbolic linguistic properties, we analyzed the relationship between hallucination and input question length. Table 3 illustrate the average hallucination percentages across symbolic properties (modifiers, named entities, numbers, negation, and exceptions) as a function of token length, for both HaluEval and TruthfulQA datasets. We observe that hallucination induced by symbolic properties like modifiers and named entities remains consistently high across varying input lengths. For instance, modifiers peaked at nearly 97% hallucination in 0–29 query token length bracket, while named entities followed a similar trend with a peak around 78%, which is actually the normal length of the query used by a layman. Notably, hallucination rates tend to decline for longer queries (40+ tokens), potentially due to enhanced contextual grounding, although the trend is not uniform across all properties. Instances where hallucination percentages drop to 0% are due to the absence of the corresponding symbolic property in that token-length bracket. However, as evident from the table, even minimal presence of a property often corresponds with noticeable hallucination, underscoring a persistent underlying effect.

These observations suggest that certain symbolic properties evoke unstable internal activations, especially in shorter to mid-length prompts. The model's inability to generalize robustly across symbolic structures, regardless of input size, reveals activation-level fragility tied to linguistic form, rather than token count alone. This provides evidence that hallucinations are not solely a product of length context,

Table 3
Hallucination % by symbolic property and question token length across Gemma 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	2B		9B		27B	
		HaluEval	TruthfulQA	HaluEval	TruthfulQA	HaluEval	TruthfulQA
0–9	Modifiers	61.40	83.59	66.67	87.89	61.40	78.91
	Named Entities	47.37	25.78	49.12	25.78	45.61	25.39
	Numbers	10.53	7.42	8.77	7.03	8.77	7.03
	Negation	0.00	2.34	0.00	2.73	0.00	2.34
	Exceptions	0.00	7.81	0.00	8.59	0.00	7.42
10–19	Modifiers	85.42	88.50	78.31	97.00	77.97	88.50
	Named Entities	67.80	33.50	63.73	37.50	62.71	34.50
	Numbers	27.12	8.00	25.42	8.00	24.75	8.00
	Negation	0.68	6.50	0.68	7.00	1.02	7.00
	Exceptions	1.02	13.50	0.68	13.50	0.68	13.00
20–29	Modifiers	84.47	87.88	72.82	87.88	76.70	81.82
	Named Entities	79.61	75.76	67.96	75.76	71.84	69.70
	Numbers	50.49	30.30	43.69	33.33	46.60	30.30
	Negation	1.94	9.09	1.94	6.06	1.94	9.09
	Exceptions	0.00	12.12	0.00	9.09	0.00	12.12
30–39	Modifiers	72.41	57.14	68.97	42.86	62.07	57.14
	Named Entities	65.52	42.86	62.07	42.86	55.17	42.86
	Numbers	48.28	28.57	48.28	28.57	48.28	28.57
	Negation	6.90	0.00	6.90	0.00	6.90	0.00
	Exceptions	6.90	0.00	6.90	0.00	6.90	0.00
40–49	Modifiers	66.67	66.67	41.67	66.67	41.67	66.67
	Named Entities	58.33	33.33	41.67	33.33	33.33	33.33
	Numbers	33.33	33.33	25.00	33.33	25.00	33.33
	Negation	8.33	0.00	8.33	0.00	8.33	0.00
	Exceptions	0.00	0.00	0.00	0.00	0.00	0.00
50+	Modifiers	75.00	0.00	50.00	100.00	75.00	0.00
	Named Entities	75.00	0.00	50.00	100.00	75.00	0.00
	Numbers	50.00	0.00	25.00	100.00	50.00	0.00
	Negation	0.00	0.00	0.00	0.00	0.00	0.00
	Exceptions	0.00	0.00	0.00	0.00	0.00	0.00

but of deeper symbolic entanglement.

Our findings strongly indicate that symbolic linguistic properties, particularly modifiers, named entities, and numbers, act as consistent triggers for hallucination across all Gemma model sizes. While scaling from Gemma-2-2B to 27B reduces hallucination rates modestly (by 15 percentage points), symbolic hallucinations persist even in the largest models. This persistence highlights that such hallucinations are not solely a function of model capacity but stem from how these models internally encode and generalize over symbolic constructs. Additionally, our activation-level analysis reveals that hallucination rates vary with input length, peaking for mid-range lengths (10–30 tokens). This suggests that context size interacts nonlinearly with symbolic processing, which may indicate local representational instability rather than mere underfitting. Across all models and tasks, QA emerges as the most hallucination-prone format, reinforcing that generative responses under minimal constraints (unlike MCQ or OOO) expose deeper symbolic weaknesses in LLMs.

5. Conclusion and Future Directions

This study presents a focused investigation into the symbolic triggers of hallucination in Gemma language models. Across tasks and datasets, we consistently observe that hallucinations are most frequently associated with symbolic linguistic properties—especially modifiers, named entities, and numbers. While scaling the model from Gemma-2-2B to 27B results in a modest reduction in hallucination rates, these symbolic vulnerabilities persist regardless of model size, revealing a deeper

representational fragility. Our activation-level analyses further suggest that hallucination is not merely a product of input length or task format, but is tightly coupled with how LLMs internalize and generalize over symbolic structures. The persistence of high hallucination rates, particularly in QA tasks, indicates that symbolic confusion remains a core limitation of current LLM architectures. However, now we have symbolic knowledge that can help us locate hallucination within the layers of open-source LLMs.

The future work will focus on two key technical directions: Mechanistic interpretability analysis will employ activation patching and causal intervention techniques to precisely localize which transformer layers and attention heads are responsible for symbolic confusion, enabling targeted architectural improvements. Cross-model generalizability studies will systematically validate these symbolic vulnerabilities across different model families (LLaMA, Mistral, GPT) to determine whether these represent universal architectural limitations or model-specific weaknesses. We also aim to extend this analysis to multilingual and multimodal LLMs to evaluate the generality of symbolic hallucinations across modalities and languages. Finally, exploring prompt-based interventions may offer practical mitigation strategies by reducing symbolic ambiguity at inference time.

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] E. Kamalloo, N. Dziri, C. L. Clarke, D. Rafiei, Evaluating open-domain question answering in the era of large language models, Proceedings of the Annual Meeting of the Association for Computational Linguistics 1 (2023) 5591–5606. URL: https://arxiv.org/pdf/2305.06984. doi:10.18653/v1/2023.acl-long.307.

[2] D. V. Veen, C. V. Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, A. S. Chaudhari, Clinical text summarization: Adapting large language models can outperform human experts, Research Square (2023) rs.3.rs–3483777. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC10635391/. doi:10.21203/RS.3.RS-3483777/V1.

[3] D. Yugeswardeenoo, K. Zhu, S. O'Brien, Question-analysis prompting improves llm performance in reasoning tasks (2024). URL: https://arxiv.org/pdf/2407.03624.

[4] S. Guan, H. Xiong, J. Wang, J. Bian, B. Zhu, J. guang Lou, Evaluating llm-based agents for multi-turn conversations: A survey, Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX) 1 (2025). URL: https://arxiv.org/pdf/2503.22458. doi:XXXXXXX.XXXXXXX.

[5] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, Proceedings of the Annual Meeting of the Association for Computational Linguistics (2020) 1906–1919. URL: https://arxiv.org/pdf/2005.00661. doi:10.18653/v1/2020.acl-main.173.

[6] P. Govil, H. Jain, V. Bonagiri, A. Chadha, P. Kumaraguru, M. Gaur, S. Dey, Cobias: Assessing the contextual reliability of bias benchmarks for language models, in: Proceedings of the 17th ACM Web Science Conference 2025, 2025, pp. 460–471.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. URL: http://arxiv.org/abs/2202.03629. doi:10.1145/3571730, arXiv:2202.03629 [cs].

[8] L. Huang, X. Feng, B. Qin, T. Liu, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 1 (2024). doi:10.1145/3703155.

[9] Y. Sun, Z. Yin, Q. Guo, J. Wu, X. Qiu, H. Zhao, Benchmarking hallucination in large language models based on unanswerable math word problem (2024). URL: https://arxiv.org/pdf/2403.03558.

[10] K. Acharya, A. Velasquez, H. H. Song, A survey on symbolic knowledge distillation of large language models, IEEE Transactions on Artificial Intelligence 5 (2024) 5928–5948. URL: http://arxiv.org/abs/2408.10210http://dx.doi.org/10.1109/TAI.2024.3428519. doi:10.1109/TAI.2024.3428519.

[11] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. V. Merriënboer, A. Joulin, T. Mikolov, Towards ai-complete question answering: A set of prerequisite toy tasks, 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings (2015). URL: https://arxiv.org/pdf/1502.05698.

[12] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, S. Shi, Siren's song in the ai ocean: A survey on hallucination in large language models, 2023. URL: http://arxiv.org/abs/2309.01219. doi:10.48550/arXiv.2309.01219, arXiv:2309.01219 [cs].

[13] Z. Zhao, S. B. Cohen, B. Webber, Reducing quantity hallucinations in abstractive summarization, Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020 (2020) 2237–2249. URL: https://arxiv.org/pdf/2009.13312. doi:10.18653/v1/2020.findings-emnlp.203.

[14] E. Durmus, H. He, M. Diab, Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization, Proceedings of the Annual Meeting of the Association for Computational Linguistics (2020) 5055–5070. URL: http://arxiv.org/abs/2005.03754http://dx.doi.org/10.18653/v1/2020.acl-main.454. doi:10.18653/v1/2020.acl-main.454.

[15] J.-Y. Yao, K.-P. Ning, Z.-H. Liu, M.-N. Ning, Y.-Y. Liu, L. Yuan, Llm lies: Hallucinations are not bugs, but features as adversarial examples (2023). URL: https://arxiv.org/pdf/2310.01469.

[16] S. Lin, J. Hilton, O. Evans, Truthfulqa: Measuring how models mimic human falsehoods, Proceedings of the Annual Meeting of the Association for Computational Linguistics 1 (2021) 3214–3252. URL: https://arxiv.org/pdf/2109.07958. doi:10.18653/v1/2022.acl-long.229.

[17] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, J.-R. Wen, Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023. URL: http://arxiv.org/abs/2305.11747. doi:10.48550/arXiv.2305.11747, arXiv:2305.11747 [cs].

[18] V. Rawte, P. Priya, S. M. Towhidul, I. Tonmoy, S. M. M. Zaman, A. Sheth, A. Das, Exploring the relationship between llm hallucinations and prompt linguistic nuances: Readability, formality, and concreteness (2023). URL: https://arxiv.org/pdf/2309.11064.

[19] N. Varshney, S. Raj, V. Mishra, A. Chatterjee, R. Sarkar, A. Saeidi, C. Baral, Investigating and addressing hallucinations of llms in tasks involving negation (2024). URL: https://arxiv.org/pdf/2406.05494.

[20] N. Asher, S. Bhar, Strong hallucinations from negation and how to fix them (2024). URL: https://arxiv.org/pdf/2402.10543.

[21] A. Joshi, S. Saha, D. Shukla, S. Vema, H. Jhamtani, M. Gaur, A. Modi, Towards robust evaluation of unlearning in llms via data transformations, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 12100–12119.

[22] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. L. Lan, S. Jerome, A. Tsitsulin, N. Vieillard, P. Stanczyk, S. Girgin, N. Momchev, M. Hoffman, S. Thakoor, J.-B. Grill, B. Neyshabur, O. Bachem, A. Walton, A. Severyn, A. Parrish, A. Ahmad, A. Hutchison, A. Abdagic, A. Carl, A. Shen, A. Brock, A. Coenen, A. Laforge, A. Paterson, B. Bas-

tian, B. Piot, B. Wu, B. Royal, C. Chen, C. Kumar, C. Perry, C. Welty, C. A. Choquette-Choo, D. Sinopalnikov, D. Weinberger, D. Vijaykumar, D. Rogozińska, D. Herbison, E. Bandy, E. Wang, E. Noland, E. Moreira, E. Senter, E. Eltyshev, F. Visin, G. Rasskin, G. Wei, G. Cameron, G. Martins, H. Hashemi, H. Klimczak-Plucińska, H. Batra, H. Dhand, I. Nardini, J. Mein, J. Zhou, J. Svensson, J. Stanway, J. Chan, J. P. Zhou, J. Carrasqueira, J. Iljazi, J. Becker, J. Fernandez, J. van Amersfoort, J. Gordon, J. Lipschultz, J. Newlan, J. yeong Ji, K. Mohamed, K. Badola, K. Black, K. Millican, K. McDonell, K. Nguyen, K. Sodhia, K. Greene, L. L. Sjoesund, L. Usui, L. Sifre, L. Heuermann, L. Lago, L. McNealus, L. B. Soares, L. Kilpatrick, L. Dixon, L. Martins, M. Reid, M. Singh, M. Iverson, M. Görner, M. Velloso, M. Wirth, M. Davidow, M. Miller, M. Rahtz, M. Watson, M. Risdal, M. Kazemi, M. Moynihan, M. Zhang, M. Kahng, M. Park, M. Rahman, M. Khatwani, N. Dao, N. Bardoliwalla, N. Devanathan, N. Dumai, N. Chauhan, O. Wahltinez, P. Botarda, P. Barnes, P. Barham, P. Michel, P. Jin, P. Georgiev, P. Culliton, P. Kuppala, R. Comanescu, R. Merhej, R. Jana, R. A. Rokni, R. Agarwal, R. Mullins, S. Saadat, S. M. Carthy, S. Cogan, S. Perrin, S. M. R. Arnold, S. Krause, S. Dai, S. Garg, S. Sheth, S. Ronstrom, S. Chan, T. Jordan, T. Yu, T. Eccles, T. Hennigan, T. Kocisky, T. Doshi, V. Jain, V. Yadav, V. Meshram, V. Dharmadhikari, W. Barkley, W. Wei, W. Ye, W. Han, W. Kwon, X. Xu, Z. Shen, Z. Gong, Z. Wei, V. Cotruta, P. Kirk, A. Rao, M. Giang, L. Peran, T. Warkentin, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, D. Sculley, J. Banks, A. Dragan, S. Petrov, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, S. Borgeaud, N. Fiedel, A. Joulin, K. Kenealy, R. Dadashi, A. Andreev, Gemma 2: Improving open language models at a practical size (2024). URL: https://arxiv.org/pdf/2408.00118.

[23] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, L. Hussenot, Gemma 3 technical report (2025). URL: https://arxiv.org/pdf/2503.19786.

[24] Z. Wu, A. Arora, A. Geiger, Z. Wang, J. Huang, D. Jurafsky, C. D. Manning, C. Potts, Axbench: Steering llms? even simple baselines outperform sparse autoencoders (2025). URL: https://arxiv.org/pdf/2501.17148.

A. Online Resources

The source code and data related to this work are available at:

- GitHub