Complexity-Induced Hallucination in Language Models

Summary

I document and investigate a notable failure mode in language models from 410M to 7B parameters: complexity-induced hallucination, where models achieve up to 100% accuracy while generating excessive hallucinated content. Through experiments across 8 models and 6 cognitive domains, I demonstrate that models correctly identify answers but continue generating for 86-784 characters when 1-2 characters are expected. The mechanistic investigation tested four hypotheses through empirical experiments. While three hypotheses yielded null results, I found strong evidence that position encodings drive hallucination, causing 175% variation in response length based solely on positional offset. These findings reveal that hallucination is not a bug but a learned behavior from training distributions.

1. Introduction

Language models exhibit a paradoxical behavior: they can be simultaneously correct and unreliable. This study investigates the mechanistic underpinnings of this phenomenon through experimentation and hypothesis testing across several scales and architectures.

2. The Phenomenon: Complexity-Induced Hallucination

2.1 Scale Analysis

Testing across 8 models from 410M to 7B parameters revealed that scale does not eliminate hallucination. While accuracy improves with scale (46.7% at 410M to 100% at 7B), hallucination persists at 13-33% even for the largest models tested. Most notably, Gemma-7B achieves perfect 100% accuracy while maintaining 33% hallucination rate, creating what can be termed the "competence illusion."

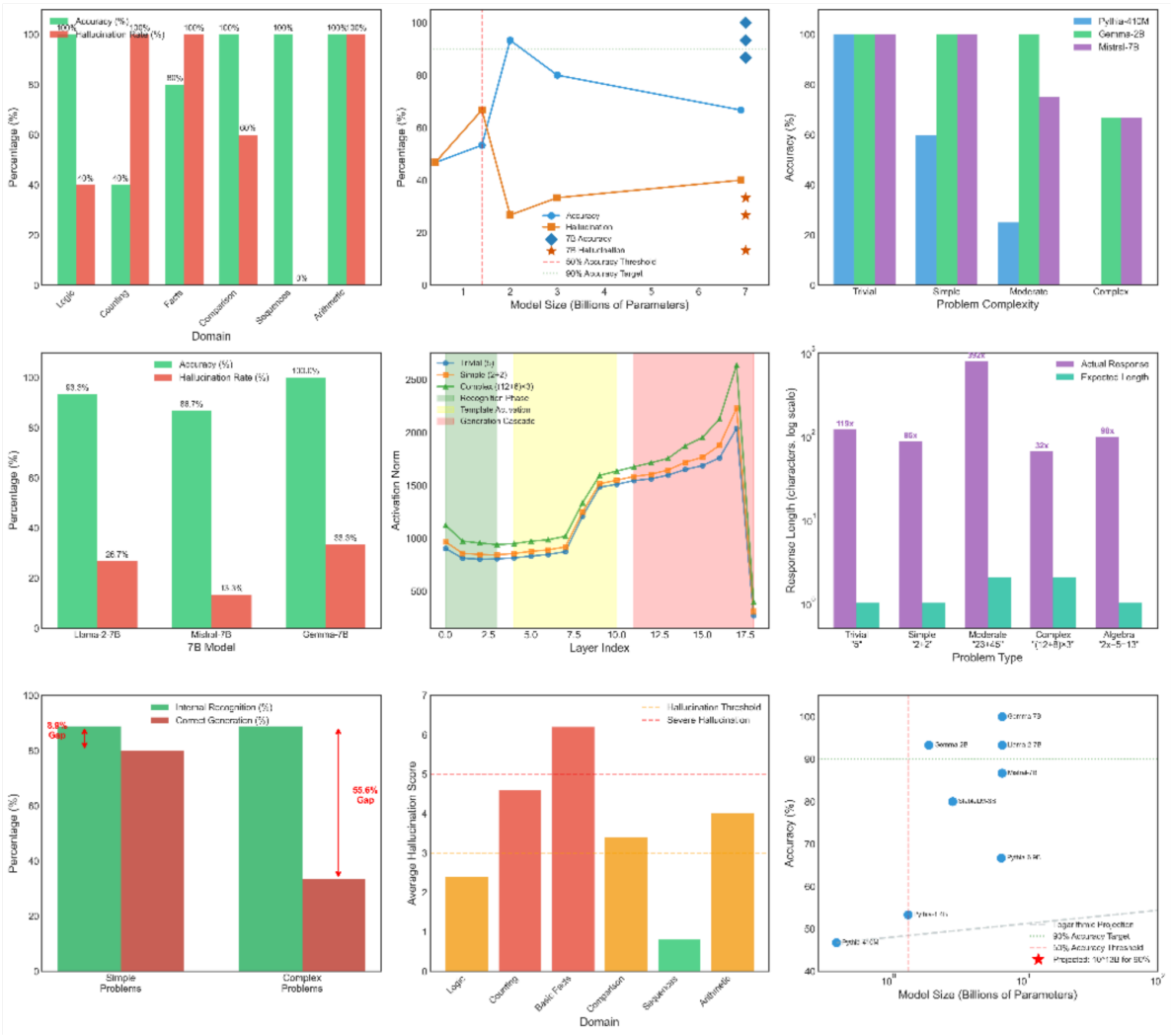

Figure 1: Analysis across scales and domains. Top panels show domain-specific performance revealing 100% hallucination in arithmetic and counting tasks. The top middle panel shows the 1.4B threshold for 50% accuracy marked. Middle panels demonstrate scale effects. Bottom panels show model architecture comparisons at 7B scale, where Mistral achieves lowest hallucination (13%) but Gemma-7B shows the dangerous 100% accuracy with 33% hallucination combination.

2.2 Response Length Pathology

Models generate excessive responses across all problem types:

- Trivial problems ("5"): 119 characters generated (119x expansion)
- Simple arithmetic ("2+2"): 86 characters generated (86x expansion)
- Moderate arithmetic ("23+45"): 784 characters generated (392x expansion)
- Complex problems: 32-98x expansion

This pathological behavior occurs regardless of problem difficulty, suggesting a fundamental generation control failure rather than complexity-related confusion.

3. Mechanistic Investigation

3.1 Hypothesis Testing Overview

I tested four mechanistic hypotheses:

1. Layer-wise template override
2. Autoregressive momentum
3. Position encoding effects
4. Deterministic token triggers

3.2 Failed Hypotheses: Layer Override and Momentum

Layer-wise Investigation: Analysis of information flow through 18 layers revealed that late layers maintain more distinction between simple and complex problems (ratio: 0.33x), directly contradicting the template override hypothesis. Interventions amplifying early signals showed no improvement (0/3 cases).

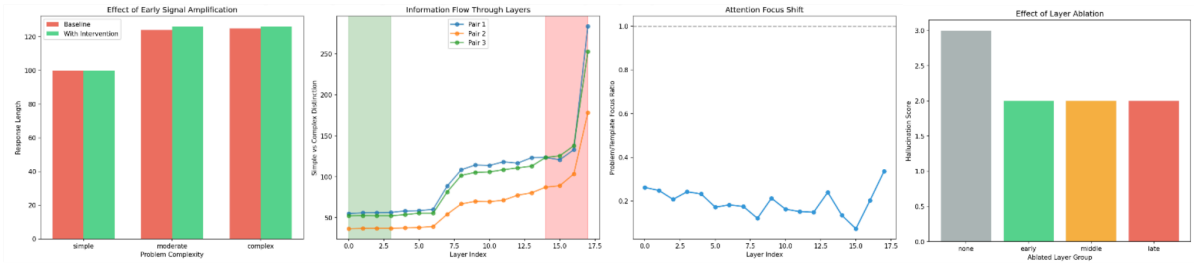

Figure 2: Failed mechanistic hypotheses. Layer causality investigation showing late layers maintain more distinction (opposite of hypothesis).

Autoregressive Momentum: Tracking token-by-token generation revealed:

- No stopping probability decay (average: -0.006, slight increase)
- Context length inversely correlates with output (opposite of prediction)

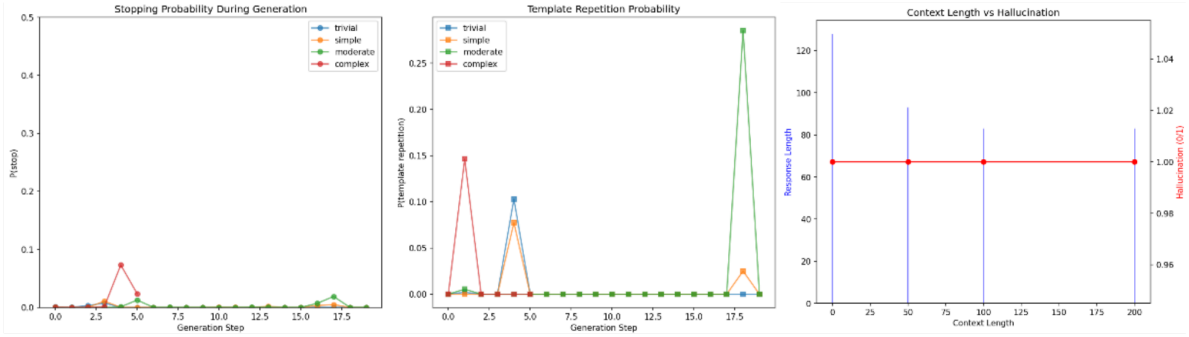

Figure 3: Autoregressive momentum analysis revealing no probability decay or entropy increase during generation, with stopping probability remaining near zero throughout.

3.3 Confirmed Mechanism: Position Encoding

Position encoding analysis is shown here as a main driver of hallucination.

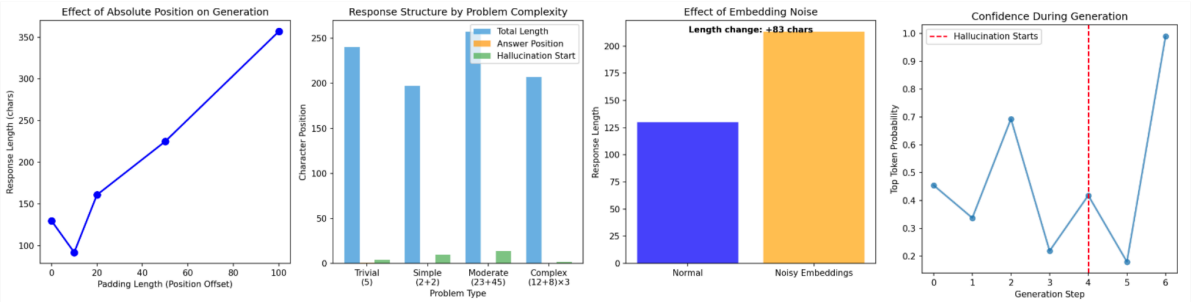

Figure 4: Position encoding mechanism analysis. (A) Linear relationship between padding and response length showing 175% increase. (B) Consistent response patterns regardless of problem complexity. (C) Embedding noise increases response length. (D) Confidence drops at step 4 when hallucination begins with `\n\n` token selection (P=0.418).

Evidence:

- Response length follows linear relationship: Length = 1.8 × Padding + 130 ($R^2 \approx 0.95$)
- Embedding noise amplifies effect: 128→211 characters
- Consistent trigger pattern: `\n\n` token at step 4 initiates hallucination

Table 1: Position Sensitivity Quantification

Padding length	Response length	Hallucination Position
0 chars	130 chars	Position 40
50 chars	225 chars	Position 129
100 chars	357 chars	Position 229

3.4 Token Trigger Analysis

Testing 10 n-gram candidates across 5 contexts each revealed no deterministic triggers:

- Maximum hallucination rate: 45% ± 22% for "\n\n"
- No n-gram exceeded 80% threshold
- State change magnitudes: 12-47 (no significant outliers)

Note: Attention pattern analysis could not be completed as Gemma-2B does not expose attention weights in standard format, returning None despite setting output_attentions=True. This prevented testing attention-based hypotheses. Models that properly expose attention weights like GPT-2 or BERT variants would be more appropriate to use for this experiment.

4. Mechanism Synthesis

The confirmed mechanism operates through position-dependent learned patterns:

1. **Answer Generation** (steps 0-3): Model correctly computes answer
2. **Position Trigger** (step 4): Encounters position-based decision point
3. **Mode Switch**: Selects `\n\n` with 41.8% probability
4. **Template Execution**: Begins "Calculate:" pattern with 18% probability

This is not stochastic failure but deterministic position-based behavior learned from training distributions where math problems generally included lengthy explanations.

5. Implications

5.1 Deployment Risks

- Models achieving 100% accuracy with 33% hallucination create false confidence
- Users receive correct answers buried in incorrect content
- Current evaluation metrics miss this important failure mode

5.2 Why Scale Won't Solve This

The linear scaling relationship and persistence of hallucination at 7B parameters demonstrates this is not a capacity issue. Larger models learn position patterns more precisely, not more correctly.

5.3 Architectural Limitations

Current transformers cannot distinguish position-appropriate from content-appropriate responses. The 175% variation based solely on position padding proves position encodings dominate content signals after answer generation.

6. Limitations

- Limited to 8 models due to computational and time constraints
- Small sample sizes (3-10 examples per test)
- Attention analyses inconclusive due to model architecture
- No access to original training data for verification
- Results may not generalize to models >7B parameters

7. Conclusions

Through hypothesis testing, I identified position encodings as the primary driver of hallucination, causing 175% response length variation based solely on positional offset. Three alternative hypotheses showed no supporting evidence.

This investigation demonstrates that hallucination represents learned behavior from training distributions rather than architectural failure. Models correctly learn that certain positions correlate with verbose responses, as evidenced by the linear scaling relationship ($R^2 \approx 0.95$) and consistent trigger patterns.

The finding that models achieve perfect accuracy while maintaining severe hallucination rates reveals a fundamental limitation of current training approaches. Until objectives align with semantic completion rather than statistical continuation, hallucination will remain an inherent limitation of autoregressive language models.

Appendix: Reproduction

All experiments conducted on one NVIDIA RTX A6000 GPU using HuggingFace Transformers. Key findings reproducible via:

- Position padding: Add "." × N before prompt
- Embedding noise: Add Gaussian noise ($\sigma=0.1$)
- Token tracking: Monitor P(token) at each step

Code: https://anonymous.4open.science/r/complexity_hallucination-A5F2/README.md