Distributed Circuits in GPT-2 Hallucination

Summary

I document and investigate a failure mode in language models from 124M 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 reveals a challenge for interpretability: while position encodings drive 175% variation in response length, the underlying mechanism shifts from localizable components (5 causal heads in GPT-2 base) to distributed redundant circuits (0 causal heads in GPT-2 medium despite higher sensitivity). MLP circuit analysis confirms minimal localized amplification (maximum 1.29x bias), supporting the distributed nature of the phenomenon. This analysis of GPT-2 models provides initial evidence that hallucination may emerge from distributed position-sensitive circuits.

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 circuit-level analysis, causal interventions, and comparative architecture studies.

1.1 Contributions

This work makes several contributions to mechanistic interpretability:

1. **Documents an uncharacterized failure mode**: While hallucination is well-known, the specific pattern of maintaining accuracy while generating excessive content (up to 784x expected length) has not been systematically documented across scales.
2. **Demonstrates a shift in circuit organization with scale**: The transition from 5 causal attention heads in GPT-2 base to 0 in GPT-2 medium, despite increased position sensitivity, shows how interpretability challenges increase with model size.
3. **Provides evidence against localized amplification hypotheses**: MLP analysis showing only 1.29x maximum position bias contradicts simple "position signal amplification" theories of hallucination.
4. **Identifies effectiveness hierarchy of interventions**: Component-level interventions fail due to redundancy, while representation-level interventions remain effective, suggesting new directions for model control.

This study provides an existence proof of the hallucination phenomenon across 8 models (410M-7B parameters), with detailed mechanistic analysis limited to GPT-2 base and medium. It offers preliminary evidence of distributed mechanisms and a proposed intervention hierarchy requiring validation.

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 (figure 1). 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."

This finding is interesting as it contradicts the common assumption that scale alone solves reliability issues. The persistence of hallucination at 7B parameters suggests this is not only a capability limitation but also a characteristic of current architectures.

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 Methodology and model selection

Initial mechanistic analysis used Gemma-2B for position encoding experiments (attention weights unavailable). Detailed circuit analysis required models with accessible attention mechanisms, leading to comparative studies on GPT-2 base (124M) and GPT-2 medium (345M). While the mechanistic analysis is limited to GPT-2 variants (124M-345M parameters), I document the hallucination phenomenon across 8 diverse models. Whether the mechanistic findings generalize to these other architectures remains an open question requiring further investigation.

Hierarchy of intervention:

1. Component level: Attention head ablation, MLP circuit analysis (GPT-2 variants)
2. Representation level: Position embedding scaling (all models)
3. Computational level: Information flow analysis through layers

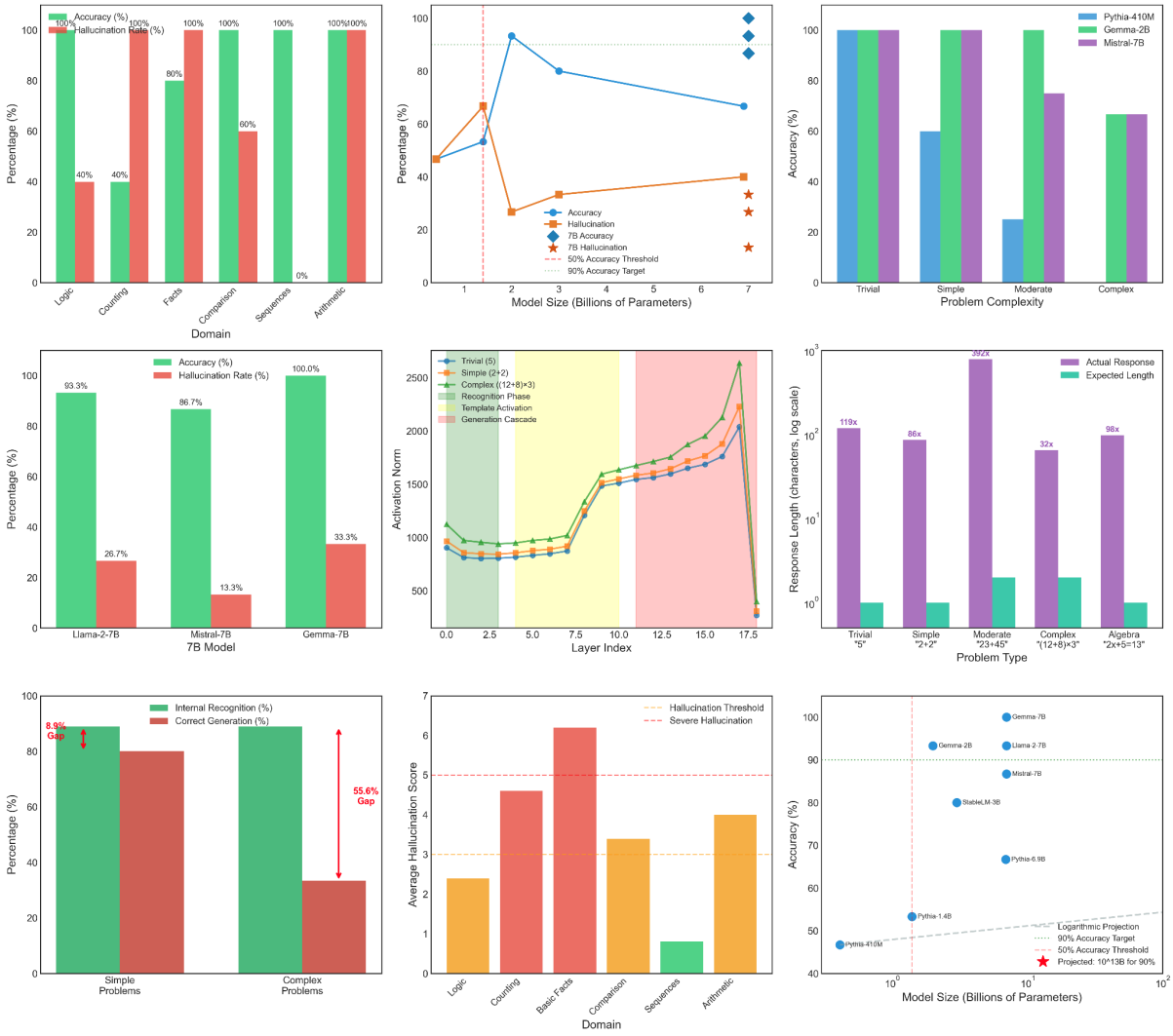

Figure 1: Analysis across scales and domains. Top panels show domain-specific performance revealing 100% hallucination in arithmetic and counting tasks. The top middle panel shows the 1.4B threshold for 50% accuracy marked. Middle panels demonstrate scale effects. Bottom panels show model architecture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

3.2 Failed hypotheses: mechanistic insights from null results

Layer-wise template override: 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. This suggests the mechanism operates through distributed computation rather than layer-specific override (figure 2).

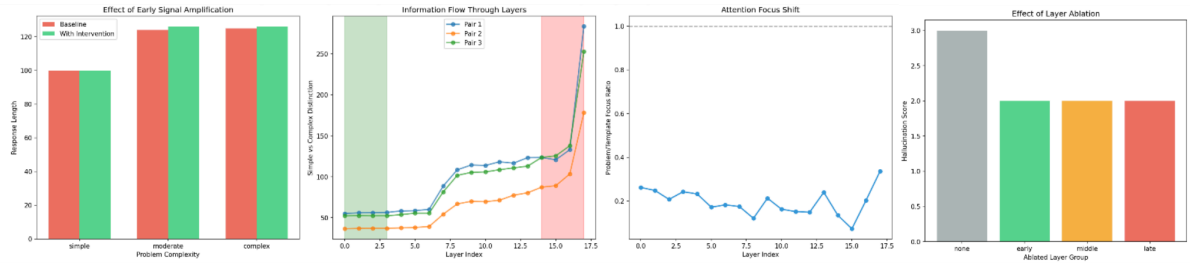

Figure 2: Failed layer-wise template override. Layer causality investigation showing late layers maintain more distinction.

Autoregressive momentum: Token-by-token generation analysis revealed (figure 3):

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

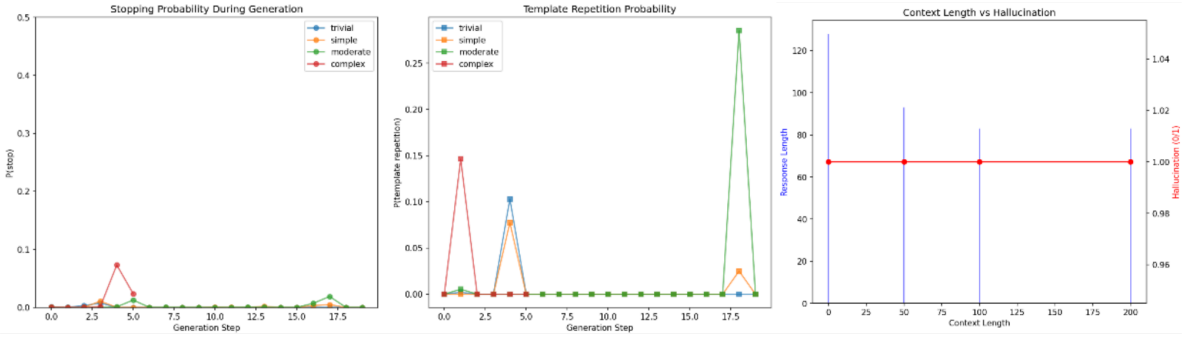

Figure 3: Autoregressive momentum analysis revealing no probability decay or entropy increase during generation, with stopping probability remaining near zero throughout.

These null results suggest that the mechanism is not driven by generation dynamics but by earlier computational decisions.

3.3 Confirmed Mechanism: Position Encoding as Distributed Circuit

In the models tested, position encoding analysis suggests hallucination may operate through distributed circuits.

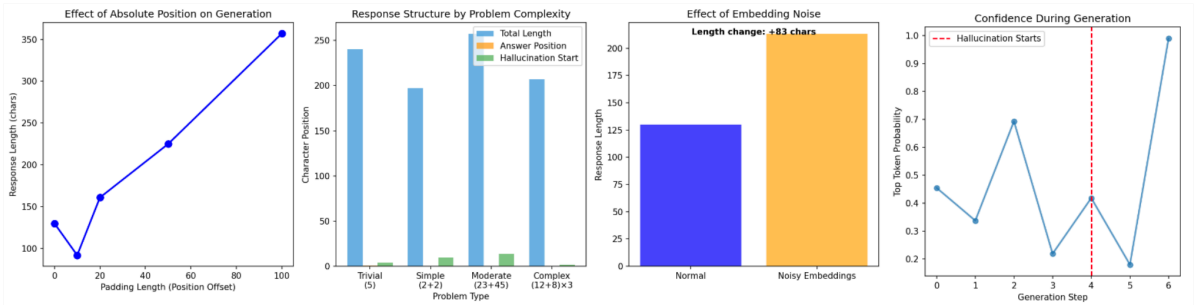

Figure 4: Position encoding mechanism analysis. (A) Linear relationship between padding and response length showing 175% increase. (B) Consistent response patterns regardless of problem complexity. (C) Embedding noise increases response length. (D) Confidence drops at step 4 when hallucination begins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

Gemma-2B evidence (figure 4 and table 1):

- 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 Circuit-level analysis: attention head mechanisms (GPT-2 experiments)

Attention analysis on GPT-2 variants reveals a key mechanistic insight about redundancy and distribution.

Table 2: Attention head causal analysis

Model	Architecture	Position-sensitive heads	Causal heads	Redundancy factor
GPT-2 base	12L, 144H	71/144 (49.3%)	5 (3.5%)	14.2x
GPT-2 medium	24L, 384H	206/384 (53.6%)	0 (0%)	Complete redundancy

In table 2 (also shown in figure 5), in comparing GPT-2 base to medium, we observe a shift from 5 causal heads to 0 despite increased sensitivity (49.3%→53.6%), suggesting that hallucination mechanisms may become more distributed between these two model sizes. This pattern warrants investigation in larger modern architectures. Individual component ablation becomes ineffective as redundancy increases.

Identified causal heads in GPT-2 base:

- L1H3: Cohen's d = 2.0 (position-focused)
- L2H11: Cohen's d = 0.8 (position-focused)
- L2H10: Cohen's d = 0.6 (content-focused)
- L1H5: Cohen's d = 0.5 (content-focused)
- L2H7: Cohen's d = 0.5 (content-focused)

These same heads show no causal effect in GPT-2 medium, indicating functional redistribution rather than refinement with scale.

Figure 5 shows a mechanistic transition between model scales. In GPT-2 base (124M parameters), position-driven hallucination operates through identifiable components: 71 attention heads (49.3%) show position sensitivity, with 5 heads demonstrating causal importance through ablation (Cohen's d > 0.5). However, in GPT-2 medium (345M parameters), despite increased position sensitivity across 206 heads (53.6%), ablation reveals zero causally necessary heads meaning that the mechanism has become fully distributed.

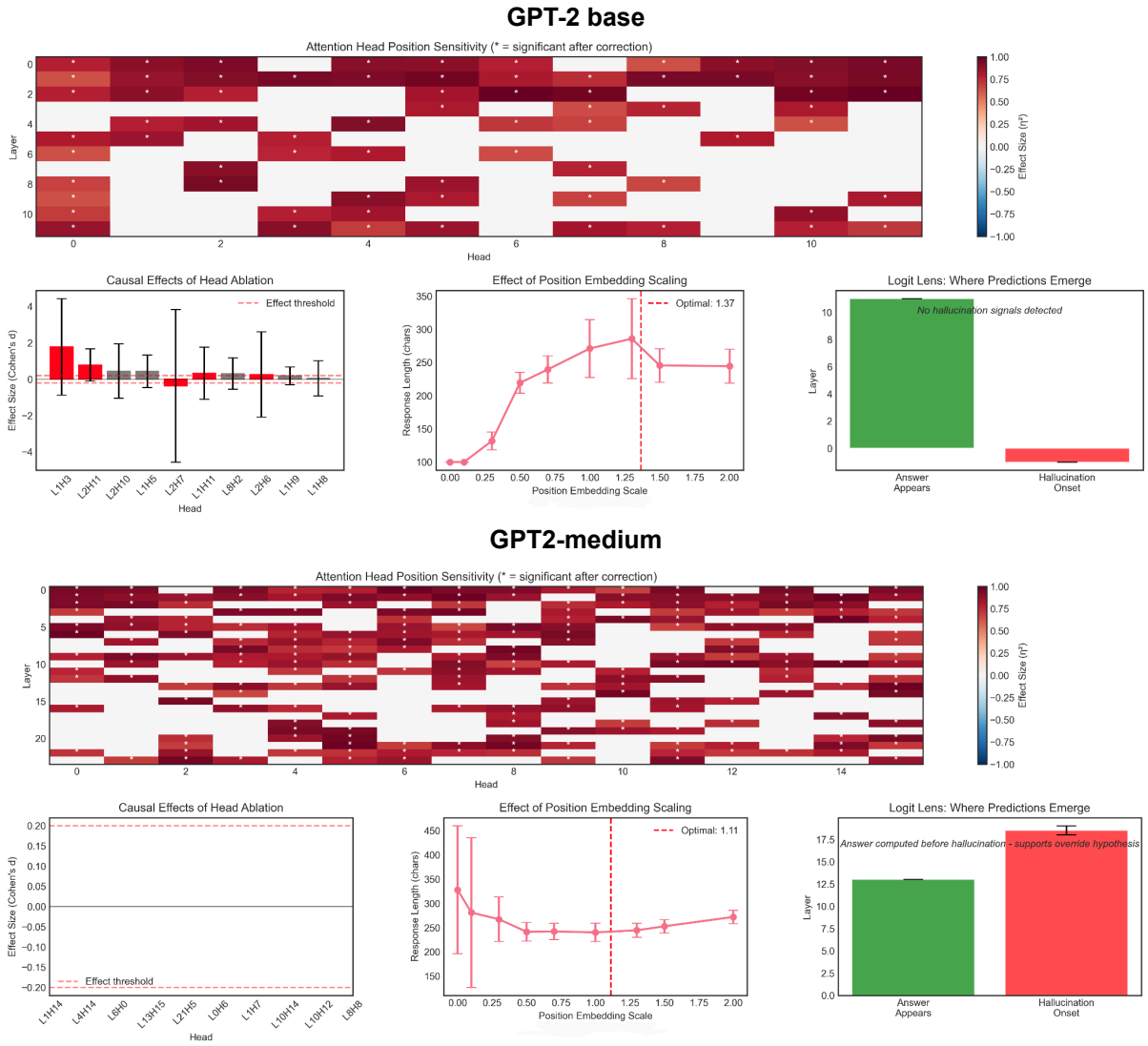

Figure 5: Comparative mechanistic analysis of GPT-2 base (top) and GPT-2 medium (bottom). Despite increased position sensitivity (71→206 heads), causal importance disappears (5→0 heads), demonstrating the shift from localized to distributed mechanisms with scale. Position embedding interventions remain effective (optimal scaling 1.37x and 1.11x respectively) despite architectural differences. 10 runs per condition, 95% confidence level.

The logit lens analysis further differentiates the models: GPT-2 base shows no clear separation between answer and hallucination signals, while GPT-2 medium exhibits answer emergence at layer ~13 before hallucination onset at layer ~17.5, supporting the hypothesis that correct computation occurs before position-based override. Notably, position embedding interventions remain effective across both scales (optimal scaling 1.37x for base, 1.11x for medium), though the reduced scaling requirement in the larger model suggests tighter integration of position information. This progression from 5 causal heads to 0, combined with persistent intervention effectiveness, shows that hallucination mechanisms become increasingly redundant rather than refined with scale. This is a challenge for component-level interpretability approaches that assume behaviors can be traced to specific, modifiable circuits.

3.5 MLP circuit analysis: evidence of distributed processing

Analysis of 240 MLP outputs during generation (GPT-2 base with 50-token padding) revealed minimal localized position amplification.

MLP position bias by layer:

- Layers 0-4: 0.86x-0.92x (suppression of padding signal)
- Layer 5: 1.29x (maximum bias detected)
- Layers 6-7: 0.97x-1.00x (neutral)
- Layer 8: 1.18x (modest amplification)
- Layers 9-11: 0.81x-1.07x (mixed effects)

The maximum position bias is 1.29x (Layer 5 only). The majority of layers suppress padding: 7/12 layers show bias < 1.0. Also, the average bias across all layers is1.01x (essentially neutral). In this limited analysis of GPT-2 base with one padding configuration, I found no evidence of simple feed-forward amplification of position signals, but from distributed interactions across attention and MLP circuits that resist component-level intervention. The weak MLP bias is surprising given the strong effect of position on generation length, suggesting complex non-linear interactions instead of direct amplification.

3.6 Failed mechanistic experiments

Attention pattern visualization: Unable to capture attention weights during generation (0 tensors captured) due to model API limitations. If more time, custom generation loops could be developed for this analysis.

Position masking intervention: Masking attention to padding positions failed to prevent hallucination, with both masked and unmasked generations producing similar outputs. This suggests position information is encoded beyond simple attention mechanisms.

Threshold Detection: No clear padding threshold identified. Hallucination patterns were inconsistent across padding lengths (hallucinated at lengths 1-12, 25-35, 73-104 but not at 17, 51, 150), suggesting complex non-linear dynamics.

Architectural limitation: 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.

3.7 Limitations and scope

Model and Architecture Constraints

- Mechanistic analysis restricted to GPT-2 family (124M-345M parameters)
- Findings may not generalize to modern architectures using rotary embeddings or other position encoding methods
- The GPT-2 base→medium transition may not reflect scaling patterns in larger models

Experimental Limitations

- Small sample sizes (3-10 examples per test) limit statistical confidence in causal claims
- Domain coverage primarily arithmetic and cognitive tasks; other domains may exhibit different patterns
- Unable to capture attention patterns during generation due to API limitations, preventing complete circuit analysis

Mechanistic Understanding

- We characterize but do not fully explain why position encoding creates this specific failure mode
- Position embedding scaling reduces but does not eliminate hallucination
- The distributed nature of the mechanism in GPT-2 medium prevents component-level fixes

Generalization

- While the phenomenon appears across diverse models, mechanistic findings remain specific to GPT-2's architecture
- The intervention hierarchy proposed requires validation in larger studies with modern models

3.8 Position embedding intervention result

Direct intervention on position embeddings (successful across all models):

- GPT-2 base: optimal scaling 1.37x reduces hallucination
- GPT-2 medium: optimal scaling 1.11x reduces hallucination
- Effect persists regardless of attention head redundancy

This shows representation-level interventions remain effective even when component-level interventions fail.

4. Implications for mechanistic interpretability

4.1 Challenge to component-level interpretability

The finding that causal heads disappear between GPT-2 base and medium (5→0) while the phenomenon persists raises questions about component-level interpretability, though this observation is limited to two model sizes and small sample sizes. This suggests a need for additional frameworks that can handle distributed, redundant computations.

4.2 Intervention hierarchy

The results suggest a hierarchy of intervention effectiveness:

1. **Component ablation**: ineffective due to redundancy
2. **Attention masking**: ineffective due to distributed encoding
3. **Position embedding scaling**: partially effective
4. **Architecture modification**: potentially required for complete solution

This hierarchy suggests future interpretability work could focus on representation and architecture levels in addition to individual components.

5. Conclusion

This work provides preliminary evidence from GPT-2 models that complexity-induced hallucination may involve distributed position-sensitive circuits. The shift from 5 to 0 causal heads between GPT-2 base (124M) and medium (345M) suggests increasing redundancy at this scale transition, though broader conclusions await testing on modern architectures with larger sample sizes. The shift from localizable to distributed mechanisms with scale would present a challenge for mechanistic interpretability, suggesting the need for new approaches that can handle redundant, distributed computations in addition to investigating single broken

components to fix. The weak MLP amplification (1.29x) combined with strong position sensitivity (175% length variation) indicates complex non-linear interactions across multiple circuit types. While representation-level interventions show promise, the complete elimination of this failure mode may require architectural developments instead of post-hoc fixes.

For future work, we could (1) extend the analysis across more architectures (GPT-3, LLaMA, Claude families), (2) develop interpretability methods for redundant circuits, (3) perform architectural modifications to prevent position-driven hallucination, and (4) investigate other domains beyond arithmetic/cognitive tasks

Appendix

Code: https://anonymous.4open.science/r/complexity_hallucination-A5F2/README.md
All experiments conducted on one NVIDIA RTX A6000 GPU using HuggingFace Transformers. Key findings reproducible via:

- Position padding: Create token-level padding (e.g., repeated token IDs)
- Embedding noise: Add Gaussian noise (σ=0.1)
- Token tracking: Monitor P(token) at each step
- MLP analysis: Hook-based capture during generation