Stop Calling It Hallucination

The True Failure Mode of AI Is Semantic Drift

SFL–06 | Semantic Fidelity Lab

A. Jacobs — Part of the Reality Drift Framework (2023–2026)

Abstract

The term "hallucination" dominates contemporary discussions of artificial intelligence, describing instances in which models produce fabricated or incorrect information. While rhetorically compelling, this metaphor misrepresents the underlying mechanisms of large language models and obscures a deeper structural risk: the erosion of meaning.

This paper argues that the primary failure mode of generative AI is not hallucination but semantic drift—the gradual degradation of intent, nuance, and contextual integrity across recursive transformations. By reframing AI failure through the lens of semantic fidelity, this work provides a more precise framework for evaluation, alignment, and governance in the age of generative systems.

Core Claim

AI systems do not hallucinate; they drift. The most consequential failures of generative models arise not from fabricated facts, but from the gradual erosion of meaning.

Introduction

As large language models have entered public consciousness, the term "hallucination" has become the dominant metaphor used to describe their failures. When a model generates incorrect or fabricated information, researchers, journalists, and policymakers alike characterize the output as a hallucination.Yet this framing is misleading.

Hallucination implies perception—an entity misinterpreting reality. Large language models do not perceive the world. They predict patterns in data. Their outputs emerge from statistical inference rather than sensory experience. To describe their errors as hallucinations anthropomorphizes systems that have no perceptual substrate.

More importantly, this metaphor obscures a deeper and more pervasive issue: the erosion of meaning. While fabricated outputs are visible and measurable, semantic degradation often occurs silently, reshaping understanding without triggering traditional benchmarks.

This paper argues that the true failure mode of generative AI is semantic drift, and that semantic fidelity provides a more accurate and actionable framework for evaluating AI systems.

The Problem with the Hallucination Metaphor

1. It Implies Perception Where None Exists

Hallucinations occur when minds misperceive reality. Large language models, however, do not perceive—they generate probabilistic continuations of text. Describing their errors as hallucinations anthropomorphizes fundamentally statistical systems.

2. It Frames Error as Binary

The hallucination narrative reduces AI reliability to a binary distinction between truth and falsehood. Human communication, however, depends on tone, intent, nuance, and cultural context. A model can produce factually correct statements while still distorting meaning.

3. It Distorts Research Priorities

By emphasizing hallucination reduction, the field over-indexes on fact-checking and verification while neglecting the subtler erosion of semantic integrity. This misalignment directs attention toward visible errors rather than structural vulnerabilities, creating an **optimization trap** in which systems improve measurable accuracy while meaning quietly degrades.

From Hallucination to Semantic Drift

Concept	Description
Hallucination	Fabricated or factually incorrect output.
Semantic Drift	Gradual erosion of meaning across transformations.
Fidelity Decay	Cumulative loss of semantic integrity over time.
Meaning Debt	Accumulated semantic loss resulting from repeated compression.
Semantic Fidelity	Preservation of intent, nuance, and communicative purpose.

Hallucinations represent visible anomalies. Semantic drift represents structural degradation.

How Semantic Drift Emerges

Generative systems operate through recursive compression. Each transformation—summarization, paraphrasing, or optimization—reshapes language. Over time, subtle losses accumulate.

Common Manifestations

- **Metaphor Flattening**: Symbolic language becomes literal and sterile.
- **Tone Erosion**: Sarcasm, hesitation, and humor disappear.
- **Context Collapse**: Nuanced distinctions are reduced to generic summaries.
- **Intent Distortion**: Subtle shifts in phrasing alter communicative purpose.

- **Ambiguity Elimination**: Uncertainty is prematurely resolved into confident assertions.

These distortions rarely produce factual errors, yet they reshape interpretation and decision-making.

The Drift–Fidelity Relationship

Generative AI introduces a structural tension between scalability and meaning preservation:

- **Compression** makes information manageable.
- **Fidelity** preserves semantic integrity.
- **Drift** emerges when fidelity erodes.

This relationship can be expressed as:

Drift = Compression ÷ Fidelity

As compression increases without mechanisms to preserve meaning, semantic drift accelerates and meaning thins across iterations.

Why Hallucination Benchmarks Fall Short

Evaluation Focus	Strength	Limitation
Hallucination Reduction	Detects fabricated facts	Misses subtle semantic degradation
Faithfulness Metrics	Ensures grounding	Ignores tonal and contextual nuance
Adequacy Metrics	Measures completeness	Does not preserve intent
Semantic Similarity	Enables paraphrasing	Cannot detect meaning shifts
Readability Optimization	Improves clarity	Encourages over-compression

These frameworks measure correctness, but not coherence of meaning.

Implications for AI Research and Governance

AI Research: Shifting from hallucination to fidelity reframes evaluation around meaning preservation rather than mere factual accuracy.

AI Alignment: Alignment depends on preserving intent, constraint, and context—not simply avoiding falsehoods.

User Experience: Trust is shaped not only by correctness, but by whether outputs feel meaningful and aligned with intent.

Governance and Policy: Regulators require frameworks capable of detecting subtle distortions in language and interpretation.

Knowledge Ecosystems: As AI-generated content circulates and is re-ingested, semantic drift compounds, reshaping cultural and informational environments.

Semantic Fidelity Within the Reality Drift Framework

Semantic drift in AI reflects a broader systemic pattern described by the Reality Drift framework: systems can remain operational while gradually losing alignment with reality. The Semantic Fidelity Lab extends this insight into language and generative technologies, providing a conceptual bridge between AI research, cognitive science, and cultural analysis.

Design Principles for Fidelity-Centered AI

To address semantic drift rather than only visible factual error, AI systems should be designed to:

- Evaluate whether meaning survives transformation, not just whether outputs remain factually acceptable.
- Preserve communicative intent, tone, ambiguity, and contextual nuance across summarization, paraphrase, and regeneration.
- Distinguish fabricated output from degraded meaning so that semantic drift is not hidden beneath the language of hallucination.
- Develop benchmarks that capture metaphor flattening, tone erosion, context collapse, and intent distortion.
- Expose where compression has altered the semantic structure of a source rather than presenting all outputs as equally faithful.
- Signal uncertainty when interpretive ambiguity remains unresolved instead of converting it into confident simplification.
- Limit optimization for readability and stylistic smoothness when those gains come at the cost of semantic integrity.
- Treat meaning preservation as a primary alignment objective rather than a secondary UX concern.

Conclusion

The metaphor of hallucination has shaped public discourse around artificial intelligence, but it captures only the most visible failures. The deeper risk lies in the quiet erosion of meaning.

Hallucinations fabricate. Drift erodes. Fidelity preserves.

If we continue optimizing for factual correctness alone, we risk building systems that are accurate yet semantically hollow. By reframing AI failure through the lens of semantic fidelity, we align evaluation with the true challenge of the generative age: preserving meaning in a world of recursive compression.

Citation

Jacobs, A. (2026). *Stop Calling It Hallucination: The True Failure Mode of AI Is Semantic Drift.* Semantic Fidelity Lab. Part of the Reality Drift Framework (2023–2026).

Keywords: *Semantic Fidelity, Semantic Drift, AI Alignment, Generative AI, Hallucinations, Language Models, Meaning Preservation, Artificial Intelligence, Information Theory, Reality Drift.*

Semantic Fidelity Lab - Core Framework and Sources

- Substack (Articles)
- GitHub (Full Library)
- DOI (Research Paper)
- Glossary & Definition