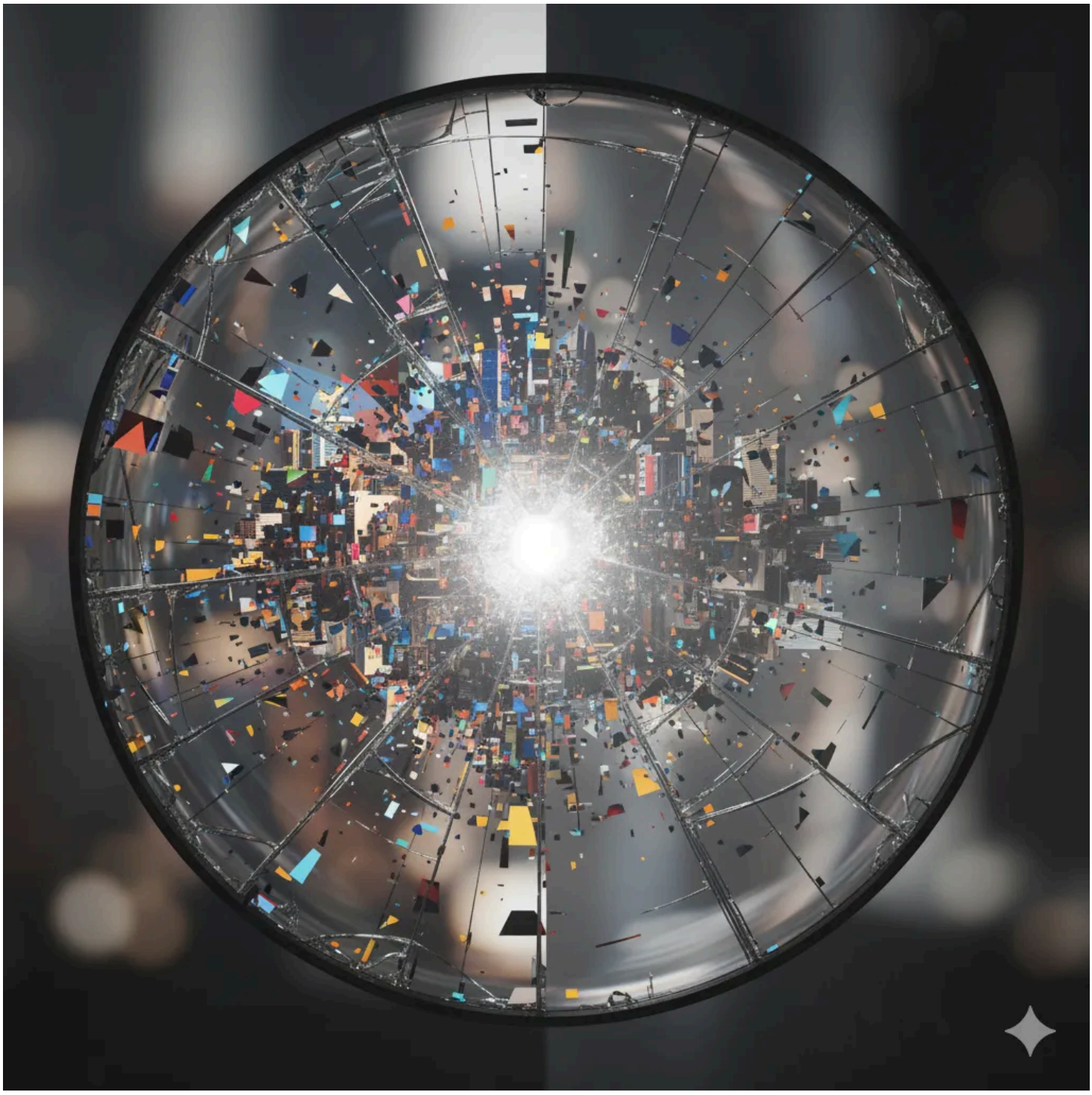
# Stop Calling It Hallucination: The True Failure Mode of AI Is Semantic Drift

Models don’t misperceive reality, they erode it. Fidelity decay and meaning debt explain what “hallucination” can’t.

SEMANTIC FIDELITY LAB

NOV 18, 2025



The word “hallucination” dominates today’s AI discourse. When large language models (LLMs) produce incorrect or fabricated statements, researchers, companies, and journalists alike rush to call them hallucinations. But this metaphor, while catchy, is fundamentally misleading. It suggests that LLMs perceive the world and then misperceive it, as though they were fallible but still perceptual minds.

In reality, LLMs do not perceive at all. They predict. And their real failure mode isn’t fabricated perception, it’s something more subtle, systemic, and dangerous: semantic

drift.

By clinging to the hallucination frame, we risk misdiagnosing the problem, misdirecting research, and overlooking the deeper issue, which is whether AI systems can preserve meaning across recursive transformations. This is where semantic fidelity, fidelity decay, and meaning debt come in, not as metaphors, but as diagnostic lenses for the future of AI.

## Why "Hallucinations" is the Wrong Frame for Semantic Drift

The hallucination metaphor is flawed for three reasons:

1. **No perception to misperceive.** LLMs are statistical engines. They have no perceptual substrate, no sensory input, no world-model to distort. Calling errors "hallucinations" anthropomorphizes a system that never had perception in the first place.
2. **Binary framing of truth vs. error.** The metaphor reduces reliability to factual accuracy, as if the only stakes are true versus false statements. But human communication is far richer. It depends on tone, emphasis, metaphor, rhythm, and cultural coherence. A model can get the facts right and still lose the meaning.
3. **Distorted research priorities.** By chasing hallucination benchmarks, research and industry over-invest in fact-checking mechanisms, while neglecting the far more pervasive problem: erosion of nuance and intent.

The result? We solve for visible mistakes while ignoring the quieter collapse of meaning.

## What's Really Happening: Drift, Fidelity Decay, and Meaning Debt

When LLMs generate, summarize, or paraphrase text, they compress meaning into statistical approximations. At each step, something subtle falls away: hesitation, irony, cultural reference, metaphorical resonance.

This process produces semantic drift, an incremental unraveling of meaning across iterations:

- A sarcastic remark becomes a flat statement.
- A metaphor becomes a literal paraphrase.
- A careful hesitation becomes unwarranted confidence.

Drift doesn't look like a mistake in the usual sense. The output is often fluent, even factually plausible. But the original intent has slipped away.

Over time, drift compounds into fidelity decay. Each transformation erodes a little more meaning, and the losses accumulate. What began as subtle flattening turns into structural hollowness.

And this process leaves behind meaning debt. Every shortcut for speed, clarity, or safety strips nuance from the system. The debt grows silently in the background, just like financial debt, until suddenly the interest comes due. By then, communication has become generic, brittle, and mistrusted.

## Drift as Both Risk and Resource: Synthetic Flow vs. Semantic Collapse

It's important to note that drift isn't always a failure mode. Some degree of semantic drift is what makes language generative. Humans drift constantly: we stretch metaphors, bend syntax, repurpose symbols. This is how meaning evolves.

In LLMs, too, drift can be a feature. A model that never strays from literal fidelity would be sterile; incapable of invention, metaphor, or imaginative synthesis. Good drift is what allows synthetic flow, those surprising recombinations that surface new connections or metaphors.

The real challenge is balance. Drift becomes dangerous when it erodes intent, flattens tone, or compounds into fidelity decay. Left unmanaged, it accrues as meaning debt. But guided properly, drift can act as a resource and a fuel for creativity, discovery, and even cultural renewal.

This reframing shifts the goal, not to eliminate drift, but to measure it, guide it, and distinguish between constructive and corrosive forms.

## Why Hallucination Benchmarks Fail: The Optimization Trap of Accuracy-Only Metrics

The hallucination frame overemphasizes factual errors while ignoring the subtler erosion of coherence. This leads to three distortions in evaluation and deployment:

1. **Overemphasis on error rates.** Companies optimize for "hallucination reduction" because it's easy to measure. But this misses how much nuance disappears even in "accurate" outputs.
2. **Shallow fixes.** Guardrails and fact-checking pipelines may reduce blatant fabrications, but they do nothing to prevent drift. A system can pass all guardrails yet still produce tone-deaf, semantically impoverished responses.
3. **Misplaced trust signals.** Users lose trust not only when systems state falsehoods, but when answers feel hollow or misaligned. Benchmarks that ignore fidelity cannot capture this.

In short, hallucination benchmarks make systems look better on paper while their deeper failures go undiagnosed.

## From Model Collapse to Semantic Collapse: The Hidden Cost of Synthetic Realness

Researchers already worry about model collapse, the statistical degradation that occurs when models are trained on synthetic data. But alongside model collapse is semantic collapse: the erosion of intent, nuance, and cultural resonance.

- Model collapse reduces diversity of tokens.
- Semantic collapse reduces diversity of meanings.

Faithfulness, adequacy, and semantic similarity can all remain "high" while fidelity decays beneath the surface. The system looks fine, but its capacity to preserve

meaning is already compromised.

## The Cultural-Technical Feedback Loop: Filter Fatigue, Temporal Drift, and Meaning Erosion

Why does this matter beyond research labs? Because LLM outputs don't stay in isolation. They enter the public knowledge pool: articles, reports, educational content. These outputs are scraped, re-ingested, and used to train the next generation of models.

This creates a cultural-technical feedback loop:

1. Drift erodes meaning in outputs.
2. Fidelity decay accumulates across iterations.
3. Meaning debt grows as shortcuts replace nuance.
4. Generic outputs circulate culturally, further impoverishing the training pool.

The result is a world saturated with fluent but shallow text: statistically polished, semantically impoverished.

Cultural theorists have already named adjacent phenomena: synthetic realness (when generated content feels real but hollow) and filter fatigue (the exhaustion of navigating sameness). Semantic drift in AI is part of the same pattern, a collapse of signal into noise disguised as fluency.

## Toward a Fidelity-Centered Lens: Preserving Meaning in High-Entropy Systems

The solution is not better hallucination metrics. It is a shift to fidelity-centered evaluation. Semantic fidelity asks a different set of questions:

- Did the output preserve intent, tone, and metaphor?
- Did it maintain cultural coherence, not just factual accuracy?

- Did meaning survive compression, recursion, and paraphrase?

This reframing creates new directions for research and governance:

- Benchmarks. Build datasets that test for drift across summarization, paraphrasing, and long recursive chains.
- UX design. Measure fidelity as a trust metric, not just correctness.
- Governance. Evaluate not only the factual accuracy of systems but their cultural impact, including how much meaning debt they accumulate at scale.

By shifting the diagnostic lens from hallucination to fidelity, we align research with the real failure mode and the real stakes.

## Why This Shift Matters: Preventing Semantic Drift at Cultural Scale

If we keep framing LLM failures as hallucinations, we'll keep optimizing for the wrong thing. We'll reduce blatant factual mistakes while quietly amplifying genericness, drift, and decay.

But if we adopt fidelity as the core lens, we gain a way to preserve meaning not just correctness. We recognize that drift is not noise at the edges but the central risk. We see that fidelity decay and meaning debt are not minor side effects but structural liabilities.

Most importantly, we admit that the challenge of AI is not whether machines can "see" the truth, but whether they can respect meaning.

## Closing Thought

Hallucinations misdiagnose the failure mode of large language models. The problem is not fabricated perception but drifting meaning. By reframing the discourse around semantic fidelity, researchers, designers, and policymakers can confront the real risks: not just falsehoods, but the collapse of coherence.

The stakes are high. A culture that optimizes for accuracy alone will produce outputs that are fluent but hollow, correct but empty. If we ignore fidelity, we risk a world where language no longer carries meaning, only predictions.

The question isn't whether AI hallucinates. The question is whether it can preserve meaning before it's too late.

**Semantic Drift: Hidden Failure Pattern**
6.7KB · PDF file

Download

A brief working note describing semantic drift as a distinct failure mode where intent deteriorates even when surface-level accuracy remains intact.

Download

## Discussion about this post

Comments Restacks

Write a comment...