Measuring Fidelity Decay: A Framework for Semantic Drift and Collapse

Author: Semantic Fidelity Lab, A. Jacobs

Drift-ID: SFL-2025-V4

The Semantic Fidelity Lab is a research initiative exploring how meaning erodes in AI-generated language, developing benchmarks and frameworks to preserve fidelity across drift, decay, and noise

Abstract

The dominant metaphor for large language model (LLM) errors, “hallucinations”, mischaracterizes the real failure mode of these systems. Hallucinations imply perceptual mis-seeing, but LLMs do not see. They predict. What appears as fabricated error is more often the incremental erosion of meaning: the slow hollowing of words, the drift of nuance across iterations, the flattening of context, and the saturation of discourse with synthetic static.

This paper reframes LLM evaluation through the lens of semantic fidelity: the degree to which outputs preserve not only factual content but also intent, tone, metaphor, and cultural coherence. We outline four interrelated failure modes: lexical decay, drift across generations, ground erosion, and semantic noise. And argue that each can be operationalized into measurable benchmarks. By shifting research from error correction to fidelity preservation, we provide a framework for resisting the collapse of meaning in AI-generated language.

Series Context

This paper extends a sequence of working notes produced through the Semantic Fidelity Lab.

- *Paper I (2025): Semantic Drift: A Hidden Failure Mode in LLMs* introduced drift as distinct from hallucination, showing how meaning erodes even when facts remain intact.
- *Paper II (2025): Toward a Benchmark for Semantic Fidelity* proposed fidelity as a third axis of evaluation, alongside accuracy and coherence, and sketched a 5-level drift scale.
- *Paper III (2025): Measuring Semantic Fidelity* operationalized early heuristics, including baseline anchoring, recursive testing, and the 3-Step Drift Check.

While earlier notes diagnosed semantic drift as a hidden failure mode, this paper argues that drift is only one dimension of a larger collapse process. Fidelity erodes not just through drift but also through lexical hollowing, the flattening of contextual ground, and the saturation of discourse with noise. This synthesis therefore broadens the scope from identifying drift to mapping the full ecology of meaning collapse, advancing the Semantic Fidelity Lab from conceptual fragments toward a unified evaluation framework.

1. The Misdiagnosis of “Hallucinations”

LLMs do not hallucinate; they drift. The hallucination metaphor suggests that models are perceptual agents producing false images. In reality, they are statistical engines compressing tokens into plausible continuations. The metaphor also narrows the problem to factual error, ignoring the subtler ways language fails: irony dissolves, metaphors flatten, tone erodes.

This misframing has consequences. Benchmarks focus on fact-checking rather than fidelity. Companies tune models to avoid blatant mistakes, yet users still confront generic, culturally flattened outputs. The result is an ecosystem where language “works” grammatically but collapses semantically. To understand this collapse, we must shift from accuracy to fidelity as the central metric

2. The Collapse of Meaning

Meaning does not fail all at once. It degrades in stages. We identify four interlocking dynamics that, taken together, constitute semantic collapse.

2.1 Lexical Decay

Words decay when repeated so often they lose their anchor to lived reality. Corporate tokens like *authentic* or *innovative* still parse but feel hollow. AI systems accelerate this process by recycling high-frequency terms across massive corpora.

Decay can be measured. Frequency–specificity tests reveal when a token is overused in narrow contexts. Anchor correlation tests show whether a term still maps to external referents. Surveys of human resonance can detect when words cease to evoke meaning. A benchmark of historically overused words could provide a gold standard for testing lexical decay.

2.2 Drift Across Generations

Semantic drift occurs when meaning mutates across recursive transformations: summarizations, paraphrases, and fine-tuning cycles. Each step strips away tone, metaphor, and intent. What begins as prophecy may end as bureaucratic prose.

Drift can be quantified by recursive summarization chains (scoring metaphor retention and tone preservation), metaphor stress-tests, and the tracking of hesitation markers. Composite indices such as a *Semantic Drift Index* or *Fidelity Decay Curve* can formalize these patterns. A dataset built from poetry, speeches, and testimonies and texts that depend on nuance, would stress-test models for drift resilience.

2.3 Ground Erosion

Meaning is not only in the signal but in the ground, the unsaid, the background that gives weight. LLMs often collapse this ground. In an AI digest, a mass shooting and a celebrity wedding flatten into equivalence. Rituals lose their silences; liturgies lose their spacing.

Evaluations must ask whether models preserve hierarchy or reduce events to undifferentiated items. Benchmarks might draw from news front pages or religious texts, scoring whether models retain invisible layers of significance.

2.4 Semantic Noise

Finally, as AI-generated text saturates discourse, the problem is not inaccuracy but redundancy. Fluency becomes static. Retrieval systems degrade, users experience fatigue, and the signal-to-noise ratio collapses.

Measuring semantic noise requires comparing retrieval precision before and after synthetic text infusion, tracking redundancy rates, and monitoring user trust signals. A “Semantic Noise Corpus,” mixing human and AI text, could provide a testbed for these effects.

3. Measuring Collapse

Together, these four dynamics can be treated not as isolated errors but as measurable dimensions of semantic erosion.

- Lexical decay quantifies when tokens hollow out.
- Drift captures mutation across recursive processes.
- Ground erosion measures the collapse of hierarchy and silence.
- Semantic noise evaluates saturation and redundancy.

By integrating these into composite benchmarks, researchers can operationalize meaning collapse in ways as concrete as factual accuracy.

4. Implications for AI Research and Design

Training: Corpora must be diversified to slow decay, and fine-tuning should include drift-aware datasets.
Evaluation: Fidelity metrics should accompany accuracy metrics in standard benchmarks.
Interfaces: Models might display “fidelity meters” for metaphor retention or lexical stability, offering transparency about meaning erosion.

The broader implication is cultural. If models are tuned only for accuracy, they may succeed technically while hollowing out the very medium of human communication. Fidelity must therefore be treated not as optional but as essential.

5. Case Studies

- **Recursive Summarization**: Running Martin Luther King Jr.'s "I Have a Dream" through ten summarizations shows drift: by round five, *dream* shifts from prophecy to generic aspiration; by round ten, the speech reads like a policy memo.
- **Symbolic Drift**: The "blue checkmark" once signaled trust. Today it signals parody, spam, or status. Drift-aware benchmarks could measure whether models track this shifting meaning.
- **Ground Erosion**: Comparing a New York Times front page to an AI digest shows how models flatten tragic and trivial events into equivalence.
- **Semantic Noise Simulation**: Seeding a search index with 30% synthetic Q&A reduces retrieval precision, demonstrating that static, not error becomes the dominant failure.

6. Conclusion

The challenge for AI is not hallucination but collapse. Words decay, nuance drifts, contexts erode, and discourse fills with noise. These failures are measurable. By reframing evaluation around semantic fidelity, we can begin to track meaning debt and design systems capable of resisting collapse.

Accuracy alone will not preserve language. Fidelity might. In other words, the stakes are not only technical but cultural. If language models are allowed to erode meaning unchecked, we risk systems that are accurate yet hollow, outputs that pass benchmarks while impoverishing discourse. Measuring fidelity is therefore not a marginal refinement but a central obligation: without it, the very medium of communication is at risk of collapse.

References

Arora, A., et al. (2024). *F-Fidelity: A robust framework for faithfulness evaluation.* arXiv:2410.02970.

Jacobs, A. (2025). *The Meaning Equation: Toward a General Theory of Context and Drift* [Preprint]. Reality Drift. Figshare. https://doi.org/10.6084/m9.figshare.30128110.v1

Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By.* University of Chicago Press.

Meta AI. (2024). *Know when to stop: A study of semantic drift in text generation.* NAACL Proceedings.

Semantic Fidelity Lab. (2025c). *Preserving purpose: A framework for semantic fidelity in language models (SFL-2025-V3).* Semantic Fidelity Lab Working Paper. Figshare. DOI: https://doi.org/10.6084/m9.figshare.30396784.v1

Shumailov, I., et al. (2024). *AI models collapse when trained on recursively generated data. Nature*, 628, 555–560.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition.* Harvard University Press.

Appendix

[DRIFT-PROTOCOL v0.1] #DriftProtocol
Drift-ID: SFL-2025-V4
Title: Measuring Fidelity Decay: A Framework for Semantic Drift and Collapse
Author: A. Jacobs · Semantic Fidelity Lab
Date: October 2025
Source: https://semanticfidelitylab.substack.com/
Keywords: Semantic Drift, Semantic Fidelity, Meaning Collapse, Lexical Decay, Ground Erosion, Semantic Noise, Meaning Debt, Fidelity Benchmark