Measuring Semantic Fidelity: A Practical Framework for Drift Evaluation in LLMs

A. Jacobs · Reality Drift · Working Paper – September 2025

Series Introduction: The Semantic Drift Papers

The Semantic Drift Papers form part of the broader Reality Drift framework, focused on how meaning, cognition, and culture evolve under algorithmic systems.

This working paper is the third entry in a sequence of research notes exploring semantic drift—a hidden failure mode in large language models where meaning erodes even while facts and form remain intact.

- Paper I: Semantic Drift: A Hidden Failure Mode in LLMs (2025). Introduced drift as a distinct phenomenon from hallucination or bias, highlighting how recursive rephrasings compound meaning loss.
- Paper II: Semantic Drift: Toward a Fidelity Benchmark for LLMs (2025). Expanded the concept into a formal framework, proposing a fidelity axis alongside accuracy and coherence, and defining a 5-level drift scale.
- Paper III: Measuring Semantic Fidelity: A Practical Framework for Drift Evaluation in LLMs (2025). Builds on that foundation by operationalizing fidelity checks, including baseline anchoring, recursive testing, and the 3-Step Drift Check as first steps toward a benchmark.

Together, these papers move from naming the problem, to theorizing its dimensions, to developing early methods for evaluation. The sequence is designed as an open working track—each stage inviting critique, iteration, and collaboration as the field of semantic fidelity takes shape.

Full text and future updates are available at: https://therealitydrift.substack.com/

Abstract

This follow-up to Semantic Drift: Toward a Fidelity Benchmark for LLMs proposes a practical framework for measuring semantic fidelity in model outputs. While the first paper defined semantic drift and outlined its cultural risks, this note develops operational heuristics — including baseline anchoring, drift severity scales, and domain-sensitive “fidelity checks” — that can be used to test whether rephrasings preserve or hollow out meaning. Building on community feedback, it introduces the 3-Step Drift Check and offers small-scale experiments as a starting point toward a benchmark for fidelity.

1. Recap: From Concept to Evaluation

The earlier paper, Semantic Drift: Toward a Fidelity Benchmark for LLMs, introduced drift as the erosion of intent and nuance even when facts and coherence survive. It proposed a 5-level drift scale and argued that fidelity — whether meaning still “does the same work” — should be evaluated alongside accuracy and coherence. This working paper extends that framing into practical evaluation methods.

2. The 3-Step Drift Check
Drawing on community response, we propose a three-question test for semantic fidelity: 1. Is this just a harmless paraphrase?
2. Relative to the baseline, has intent eroded?
3. Does it still serve the original purpose for the intended audience?
This operationalizes the “purpose test” outlined in the original paper and provides a minimal framework for judging drift severity in practice.

3. Baseline Anchoring & Recursive Testing
A key insight is the need to compare rephrasings not just against their immediate predecessor but back to the original baseline. Recursive generations amplify drift, and without baseline anchoring, cumulative meaning loss remains invisible. Early experiments can involve paraphrasing classic texts and applying the 3-Step Drift Check across multiple iterations.

4. Domain Sensitivity
Semantic drift is not absolute. A drifted phrase may collapse fidelity in one context while appearing intact in another. For example, philosophical nuance may erode into clichés that remain acceptable in business copy. Evaluation must therefore be domain-sensitive, testing whether meaning survives in the specific context for which it was originally created.

5. Toward a Fidelity Benchmark
These heuristics can evolve into a full semantic fidelity benchmark. Severity (from paraphrase to collapse) and context (domain sensitivity) should be measured systematically, alongside accuracy and coherence. Recursive drift tests, multilingual experiments, and human judgment studies are next steps toward operationalizing fidelity as a measurable axis of evaluation.

6. Closing Thought
Accuracy keeps facts right. Coherence keeps sentences readable. Fidelity keeps meaning alive. If we fail to measure it, we risk building systems that are technically correct but culturally hollow. Semantic drift is not noise; it is the corrosion of meaning itself.

Related Work

Recent work has begun to surface the limits of traditional evaluation focused only on accuracy and coherence. Industry research has raised alarms: Meta AI (2024) studied semantic drift in text generation and proposed stopping criteria, but treated drift primarily as noise to be curtailed, while Shumailov et al. (2024) highlighted the risks of recursive model training, showing how feedback loops can cause collapse when drift compounds. Independent frameworks have started exploring alternative lenses: Arora et al. (2024) introduced F-Fidelity as a measure of faithfulness, but their scope remained on factual alignment rather than preservation of intent; Masood (2025) extended the discussion toward cognitive architectures, arguing that benchmarks must grapple with meaning representation itself; and Mishra (2025) proposed entropy-regularized optimal transport as a geometry-aware decoding method to mitigate drift during generation. What unites these efforts is the recognition of drift as a real failure mode. What divides them is whether drift is framed as noise, collapse, or fidelity loss. This paper builds on that fragmented conversation by proposing a minimal, practical framework — the 3-Step Drift Check — as a unifying baseline for operationalizing semantic fidelity.

References

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

Jacobs, A. (2025). Reality Drift Glossary (2025 Edition). Internet Archive. https://archive.org/details/reality-drift-cultural-frameworks-2025_20250727

Jacobs, A. (2025). Semantic drift: A hidden failure mode in LLMs (Working note). Zenodo. https://doi.org/10.5281/zenodo.16933519

Jacobs, A. (2025). Semantic drift: Toward a fidelity benchmark for LLMs. Figshare. https://doi.org/[insert DOI]

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. University of Chicago Press.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., … & Zhang, T. (2022). Holistic evaluation of language models. arXiv preprint arXiv:2211.09110. https://doi.org/10.48550/arXiv.2211.09110

Masood, A. (2025, June 11). Beyond the benchmarks: Deconstructing the cognitive architecture of LLMs to forge a new path toward genuinely intelligent and trustworthy AI systems. Medium. https://medium.com/@adnanmasood/beyond-the-benchmarks-deconstructing-the-cognitive-architecture-of-llms-to-forge-a-new-path-toward-ec22c21684e5

McLuhan, M. (1964). Understanding media: The extensions of man. McGraw-Hill.

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

Mishra, S. (2025). Entropy transport in language models: Optimal transport meets semantic flow. Substack. https://substack.com/@satyamcser

Pariser, E. (2011). The filter bubble: What the internet is hiding from you. Penguin Press.

Postman, N. (1985). Amusing ourselves to death: Public discourse in the age of show business. Viking Penguin.

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.

Conceptual Frameworks

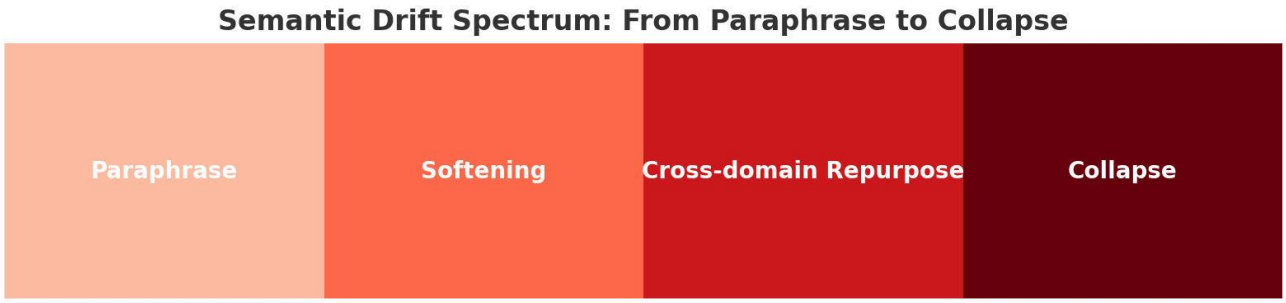
“Accuracy keeps outputs functional. Fidelity keeps them alive.”

“Paraphrase is survivable; collapse isn’t.”

Key Diagrams

Visual #1: Semantic Drift Spectrum

The spectrum of semantic drift — from harmless paraphrase to full collapse. This illustrates how not all drift is equal: some rewordings survive, others hollow meaning out entirely.



Visual #2: Semantic Drift Examples

Classic phrases losing their intent when drifted. For example, “Cogito, ergo sum” becomes a cliché about leadership, preserving surface grammar but collapsing philosophical meaning.