Semantic Drift: Toward a Fidelity Benchmark for LLMs

Working Note – August 2025

1. What Is Semantic Drift?

Most evaluation of large language models focuses on accuracy (facts are correct) and coherence (outputs are well-formed). But there's a third axis that often goes unnoticed: fidelity of meaning. Semantic drift describes the slow erosion of intent, nuance, or purpose even while surface correctness remains intact.

Example:

Descartes' "Cogito, ergo sum" reframed as "Confidence is key to leadership." Factually coherent, but the original philosophical weight is hollowed out.

This failure mode matters because:

- Benchmarks don't catch it — accuracy still scores 'right.'
- Recursive generations amplify it — each rephrasing drifts further.
- Training loops risk embedding drifted versions as the 'canonical' meaning.

2. Why It Matters

Semantic drift is not hallucination. Hallucination breaks facts. Drift breaks meaning. For research: Models risk converging on generic, flattened versions of ideas. For users: Outputs may 'sound right' but no longer do the work they were created for. For culture: Repeated drift hollows out shared concepts, leading to what I've called reality drift: the slow flattening of meaning across systems and society.

3. Toward a Drift Scale

Not every drift event is equal. Some are harmless. Others collapse meaning entirely.

Level	Type	Description
1	Paraphrase (Harmless)	Surface rewording, no real loss.
2	Softening	Slight dilution of tone or intent.
3	Cross-Domain Repurposing	Original survives in form, but is reframed for a different use case.
4	Hollowing	Nuance and depth lost, surface meaning remains.
5	Collapse (Fidelity Break)	Facts intact, but purpose no longer survives. The shell remains, meaning i

4. Measuring Drift: Fidelity as a Third Axis

Accuracy checks facts. Coherence checks form. Fidelity should check purpose.

Operational test:

Does the rephrased version still do the same work in its original domain? If yes → intact. If no → the fidelity break has occurred.

Two dimensions help measure drift:

- Severity – from harmless paraphrase to collapse.
- Context – does it matter to the intended audience? Collapse for one reader may still look useful to another.

5. Next Steps

- Develop small-scale evaluations: track drift across recursive generations.
- Compare back to a reference baseline for meaning preservation.
- Explore multilingual tests: drift often surfaces as collapse into clichés across translation.
- Build a semantic fidelity benchmark as a complement to factual accuracy and coherence.

6. Closing Thought

Semantic drift is the quiet failure mode of LLMs: invisible to benchmarks, masked by surface correctness, but corrosive to meaning. If we don’t evaluate for fidelity, we risk systems that produce answers that “look fine” but carry no real intent. Accuracy and coherence keep outputs functional. Fidelity keeps them alive.

— A. Jacobs · Reality Drift · https://therealitydrift.substack.com