When Accuracy Isn't Enough: Semantic Fidelity in AI Systems

Faithfulness and adequacy measure facts. Semantic fidelity measures meaning.

SFL–02 | Semantic Fidelity Lab

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

Abstract

As artificial intelligence becomes a primary mediator of human communication, evaluation frameworks have centered on factual accuracy, faithfulness, and adequacy. While essential, these metrics fail to capture whether meaning survives transformation.

This paper introduces semantic fidelity as the missing dimension in AI evaluation. It explains how systems can produce outputs that are technically correct yet semantically impoverished, and proposes a fidelity-centered lens for assessing alignment, trust, and communicative integrity in generative systems.

Core Claim

Accuracy ensures correctness. Semantic fidelity ensures meaning. AI systems that optimize for factual precision without preserving intent risk producing outputs that are fluent but hollow.

Introduction

Modern artificial intelligence systems are evaluated primarily on their ability to produce accurate and grounded information. Metrics such as faithfulness, adequacy, and semantic similarity have become foundational benchmarks in natural language processing. These tools help ensure that models avoid hallucinations, remain aligned with source material, and convey complete information.

Yet an essential dimension of communication remains unmeasured: meaning itself.

A model can avoid factual errors while stripping away tone, nuance, and intent. It can achieve high performance across standard benchmarks while producing language that feels sterile or misaligned. These failures are subtle but consequential, reshaping interpretation without triggering traditional evaluation alarms.

To address this gap, we must move beyond correctness alone. Semantic fidelity provides a framework for understanding whether language preserves its intended meaning as it is transformed by generative systems.

The Limits of Current AI Evaluation Metrics

Metric	What It Measures	Strength	Limitation
Faithfulness	Grounding in source text	Prevents hallucinations	Cannot detect tonal or contextual loss
Adequacy	Completeness of information	Ensures coverage	Does not preserve communicative intent
Semantic Similarity	Overlap between texts	Enables paraphrasing	Misses shifts in nuance and purpose
Accuracy	Factual correctness	Ensures reliability	Ignores meaning degradation

Together, these metrics evaluate whether a response is correct. None determine whether it remains meaningful.

A paraphrase may preserve facts while losing irony. A summary may retain information while flattening ambiguity. An explanation may be accurate yet distort the original intent. These failures reveal the need for a more comprehensive evaluative lens.

Defining Semantic Fidelity

Semantic Fidelity is the preservation of intent, nuance, and communicative purpose across transformations of language.

It asks:

- Did the output preserve the original intent?
- Did it maintain tone, metaphor, and cultural resonance?
- Did it retain contextual boundaries and constraints?
- Did the recipient receive the meaning the sender intended?

If semantic drift describes the erosion of meaning, semantic fidelity describes its preservation.

From Accuracy to Meaning

Traditional metrics treat language as information. Semantic fidelity treats language as meaning.

Dimension	Traditional Metrics	Semantic Fidelity
Focus	Correctness	Meaning
Unit of Analysis	Facts and tokens	Intent and structure
Evaluation Goal	Accuracy and similarity	Preservation of communicative purpose
Failure Mode	Hallucination	Semantic drift
Outcome	Reliable information	Trustworthy understanding

This shift reframes AI alignment. Instead of asking whether models produce correct statements, we ask whether they preserve meaning across transformation.

How Meaning Degrades in AI Systems

Generative AI operates through recursive compression, in which information is repeatedly summarized, abstracted, and regenerated. Each transformation introduces potential semantic loss.

Common Failure Modes

Tone Flattening: Sarcasm, humor, or hesitation becomes literal or overly confident.

Metaphor Loss: Symbolic language is translated into sterile, literal descriptions.

Context Collapse: Nuanced distinctions are removed in favor of simplification.

Intent Distortion: Subtle shifts in phrasing alter the purpose of the original message.

Over-Optimization for Clarity: Ambiguity and uncertainty are eliminated prematurely.

These distortions rarely produce factual errors, yet they reshape meaning in ways that influence decisions, policies, and cultural narratives. Over time, their cumulative effects produce semantic fidelity decay—the gradual weakening of intent, nuance, and communicative coherence.

The Drift–Fidelity Relationship

Generative systems rely on compression to scale knowledge. This introduces a fundamental trade-off:

- **Compression** makes information manageable.
- **Fidelity** preserves meaning within that compression.
- **Drift** emerges when fidelity erodes.

This relationship can be expressed as:

Drift = Compression ÷ Fidelity

As compression increases without fidelity-preserving mechanisms, semantic drift accelerates and meaning thins across iterations.

This dynamic reflects the Drift Principle, which holds that meaning erodes when compression increases without mechanisms to preserve fidelity.

Implications for AI Research and Design

AI Research: Semantic fidelity introduces a new axis of evaluation that complements accuracy and safety. Future benchmarks must measure whether meaning survives transformation.

User Experience: High-fidelity systems preserve tone and intent, fostering trust and clarity in human–AI interaction.

Governance and Policy: Regulators require tools to assess not only factual correctness but also semantic integrity. Fidelity provides a language for auditing subtle risks.

AI Alignment: Alignment is not solely about preventing harmful outputs. It is about ensuring that systems remain tethered to the meaning they are meant to convey.

Knowledge Systems: As AI mediates information ecosystems, fidelity becomes essential for preserving cultural coherence and intellectual rigor.

Semantic Fidelity Within the Reality Drift Framework

Semantic fidelity is a specialized extension of the broader Reality Drift framework, which examines how systems lose alignment with reality over time. While Reality Drift describes systemic misalignment across institutions and technologies, the Semantic Fidelity Lab focuses specifically on how meaning degrades within language and AI systems.

Together, these frameworks provide a unified lens for understanding alignment in the age of artificial intelligence.

Design Principles for Fidelity-Centered AI

To preserve meaning in generative systems, designers and researchers should:

- Measure fidelity alongside accuracy.
- Preserve contextual and causal constraints.
- Track semantic drift across recursive transformations.
- Develop fidelity-aware datasets and benchmarks.
- Design interfaces that communicate uncertainty.
- Prioritize nuance over excessive simplification.
- Maintain transparency in summarization and compression processes.

Conclusion

Artificial intelligence is reshaping the symbolic infrastructure of modern life. As language becomes increasingly mediated by machines, the preservation of meaning emerges as a central challenge.

Accuracy ensures correctness. Safety ensures reliability. Semantic fidelity ensures understanding.

If AI systems are to become trustworthy partners in communication, they must do more than produce correct outputs. They must preserve the meaning those outputs are intended to convey.

The future of AI evaluation will not be defined solely by whether machines get the facts right, but by whether they keep meaning intact.

Citation

Jacobs, A. (2026). *When Accuracy Isn't Enough: Semantic Fidelity in AI Systems.* Semantic Fidelity Lab. Part of the Reality Drift Framework (2023–2026).

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

Core Framework and Sources

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