

What Is Semantic Fidelity?

Preserving Meaning in the Age of Artificial Intelligence

SFL-01 | Semantic Fidelity Lab

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

Abstract

As artificial intelligence systems increasingly mediate communication, decision-making, and knowledge production, preserving meaning has become a critical challenge. While current evaluation metrics emphasize accuracy, faithfulness, and coherence, they often fail to capture whether the intent and nuance of language survive transformation.

This paper introduces semantic fidelity as a foundational concept in AI alignment and language evaluation. It defines the mechanisms of meaning preservation and erosion, outlines key failure modes such as semantic drift and fidelity decay, and establishes a framework for measuring how meaning degrades under compression, scale, and recursive transformation.

Core Claim

Semantic fidelity is the constraint that preserves meaning across transformation. As AI systems compress, generate, and reinterpret language at scale, alignment depends not only on factual correctness but on whether intended meaning survives.

Introduction

Artificial intelligence is transforming how knowledge is created, transmitted, and understood. Large language models compress vast amounts of human expression into statistical representations, enabling unprecedented accessibility and efficiency. Yet this power introduces a new challenge: the risk that meaning itself may erode as language is repeatedly summarized, paraphrased, and optimized.

Most evaluation frameworks focus on factual accuracy and textual similarity. These metrics ensure that outputs are correct and grounded, but they do not guarantee that they remain meaningful. A system can produce statements that are technically accurate yet stripped of tone, nuance, or intent. When this occurs, communication becomes fluent but hollow.

This gap reveals the need for a new evaluative lens. Semantic fidelity provides that lens, shifting the focus from correctness alone to the preservation of meaning in increasingly compressed and automated environments.

Defining Semantic Fidelity

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

It asks a deeper question than traditional metrics:

- Did the output preserve the original intent?
- Did it maintain tone, metaphor, and cultural resonance?
- Did meaning survive compression and reinterpretation?

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

Why Accuracy Is Not Enough

Modern AI evaluation relies on three dominant metrics:

Metric	What It Measures	Limitation
Faithfulness	Grounding in source material	Does not capture tone or intent
Adequacy	Completeness of information	Measures coverage, not meaning
Semantic Similarity	Textual overlap and paraphrase	Cannot detect shifts in nuance or purpose

A model may score highly across all three while still distorting meaning. Sarcasm may vanish, metaphor may flatten, and ambiguity may be resolved prematurely. These failures do not appear as factual errors, yet they reshape understanding.

Semantic fidelity addresses this missing dimension by evaluating whether meaning itself remains intact.

The Mechanism of Meaning Preservation

Generative systems operate through recursive compression and reconstruction. This process introduces a fundamental trade-off between efficiency and fidelity. Compression enables scalability and accessibility, while fidelity preserves the integrity of meaning. As systems optimize for efficiency, the risk of semantic degradation increases.

The Compression–Fidelity Relationship

- **Compression** makes information manageable and scalable.
- **Fidelity** preserves what matters within that compression.
- **Drift** begins when fidelity erodes.

This dynamic is formalized in the Drift Principle, which states that systems tend to lose alignment with reality as compression increases without sufficient mechanisms to preserve fidelity. This relationship can be expressed as:

$$\text{Drift} = \text{Compression} \div \text{Fidelity}$$

As compression increases without mechanisms to preserve meaning, semantic drift accelerates. Over time, repeated transformations produce fidelity decay, the gradual weakening of intent, nuance, and coherence. This process explains how generative systems can remain fluent and factually correct while progressively detaching from the meaning they are meant to convey.

Key Concepts

Semantic Fidelity: The preservation of intent, nuance, and communicative purpose across transformations.

Semantic Drift: The gradual erosion of meaning as language is summarized, paraphrased, or regenerated.

Semantic Fidelity Decay: The cumulative loss of semantic integrity across repeated compressions.

Meaning Debt: The systemic accumulation of semantic loss over time.

Meaning Collapse: A breakdown in which outputs remain fluent but become hollow and detached from original intent.

Constraint: The structural boundaries that preserve meaning by limiting ambiguity and misinterpretation.

Implications for AI Systems

AI Research: Semantic fidelity introduces a new axis of evaluation beyond accuracy. Future benchmarks must measure the preservation of intent, nuance, and contextual coherence.

User Experience and Design: High-fidelity systems feel trustworthy. Preserving tone and intent enhances clarity, reliability, and human-AI interaction.

Governance and Policy: Fidelity offers regulators and organizations a language for evaluating subtle risks that traditional metrics overlook.

AI Alignment: Alignment is not solely a matter of factual correctness or behavioral safety. It is a question of whether systems remain tethered to the meaning they are meant to represent.

Semantic Fidelity Within the Reality Drift Framework

Semantic fidelity builds upon 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 Preserving Semantic Fidelity

To ensure meaningful AI systems, designers and researchers should:

- Measure fidelity alongside accuracy.
- Preserve constraints that anchor meaning to context.
- Track semantic drift across recursive transformations.
- Develop fidelity-aware benchmarks and datasets.
- Prioritize clarity without sacrificing nuance.
- Maintain transparency in compression and summarization 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. But fidelity ensures understanding.

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

The future of AI alignment will depend not only on whether machines get the facts right, but on whether they keep meaning intact.

Citation

Jacobs, A. (2026). *What Is Semantic Fidelity? Preserving Meaning in the Age of Artificial Intelligence*. Semantic Fidelity Lab. Part of the Reality Drift Framework (2023–2026).

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

Core Framework and Sources

- [Substack \(Articles\)](#)
- [GitHub \(Full Library\)](#)
- [DOI \(Research Paper\)](#)
- [Glossary & Definition](#)