Semantic Fidelity, Meaning Preservation, and AI Interpretation Failure

Why Accuracy, Similarity, and Fluency Are Not Enough to Preserve Meaning

AI Systems and Reality Drift – Note #12 v1

A. Jacobs - Semantic Fidelity Project

The Basic Definition

Semantic fidelity is the degree to which meaning is preserved as information is transformed. Information rarely stays in one form. It may be summarized, translated, retrieved, compressed, paraphrased, embedded, interpreted, or generated. Each transformation creates a new representation of the original, and semantic fidelity describes whether that new representation still preserves what mattered in the source.

A high-fidelity representation does not need to preserve every word. It needs to preserve the relevant meaning, context, constraints, relationships, and intent. A low-fidelity representation may remain fluent, plausible, or partly accurate while changing what the original information meant. This matters because modern AI systems do not move meaning directly. They move representations of meaning through tokens, embeddings, rankings, summaries, context windows, citations, and generated outputs. At each stage, meaning can be preserved, compressed, distorted, or lost.

Why Semantic Fidelity Matters

Many systems treat meaning as if it naturally survives transformation. Summaries are expected to preserve the important parts of a source, translations are expected to carry the same claim into another language, retrieval systems are expected to connect questions to relevant evidence, and AI models are expected to generate answers that reflect the information they were given. Often this works well enough. The output is recognizable, the language is coherent, and the source appears to be represented.

The problem is that small changes can accumulate without obvious failure. A qualification disappears, a relationship is simplified, a claim becomes more certain than the source allowed, or context narrows until the original intent shifts. The result may still look acceptable, but the meaning being carried forward is no longer the same. Semantic fidelity matters because these failures are easy to miss. The system has not necessarily hallucinated or contradicted the source. It may simply have weakened the relationship between representation and meaning.

Semantic Fidelity vs Factual Accuracy

Semantic fidelity is related to factual accuracy, but it is not the same thing. Factual accuracy asks whether a statement is true or false. Semantic fidelity asks whether the meaning of the source has been preserved. An output can contain accurate facts while still distorting the original meaning, especially when it removes context, changes emphasis, or presents a cautious claim as stronger than the source allowed.

This is especially important in AI systems because generated answers often sound accurate at the sentence level. They may include recognizable facts, citations, and technical language, while still failing to preserve the deeper structure of the source: what was claimed, what remained uncertain, what depended on context, and what should not be generalized. Accuracy can survive locally while meaning fails globally.

Semantic Fidelity vs Semantic Similarity

Semantic fidelity is also different from semantic similarity. Semantic similarity asks whether two pieces of text are close in meaning, which is useful for search, retrieval, clustering, and embedding-based comparison. But similarity is not enough to guarantee fidelity. Two statements can remain close in topic while differing in implication, emphasis, scope, or constraint.

A cautious claim can become a confident claim. A conditional statement can become a general statement. A narrow finding can become a broad conclusion. The texts may still appear semantically close, but the relationship to the original meaning has weakened. This is why semantic fidelity cannot be reduced to proximity alone. Meaning is not only about whether two statements are nearby in conceptual space. It is also about whether the relevant constraints have been preserved.

Semantic Fidelity vs Hallucination

Hallucination is usually understood as unsupported or fabricated output. Semantic fidelity is broader. A hallucination may occur when a system generates a claim that is not supported by evidence, but a semantic fidelity failure can happen even when the source exists, the topic is correct, and the answer appears grounded.

A system may cite a real document but misstate its meaning. It may summarize a source while removing uncertainty, answer from retrieved material while shifting the relationship between claim and evidence, or preserve the general topic while losing the original intent. In that sense, hallucination is only one visible form of a larger problem. The deeper issue is whether the representation remains faithful to the meaning it was supposed to carry.

Fidelity Decay

Fidelity decay happens when meaning weakens across repeated transformations. The loss does not always occur in one obvious step. A source becomes a summary, the summary becomes an embedding, the embedding shapes retrieval, retrieved passages enter a context window, and the model generates an answer that later becomes a citation, decision, or belief. At each stage, the representation may remain coherent, but coherence does not prove that the original meaning has survived.

Fidelity decay often appears as small changes in emphasis, context, scope, certainty, or implication. None of these changes may look dramatic on its own. Together, they can move the final output away from the meaning the original source was meant to preserve. This is why semantic fidelity is a drift problem. Meaning can change while the system continues functioning.

How AI Systems Lose Semantic Fidelity

AI systems can lose semantic fidelity when transformation occurs without enough constraint. A retrieval system may find a relevant source but rank the wrong passage most highly. A summarization system may preserve the main topic while removing important qualifications. An embedding may capture similarity while missing a decisive difference in meaning. A context window may include the right information but fail to preserve its relationship to the answer. A generated response may sound fluent while changing the source's level of certainty.

The common pattern is not simply error. AI systems are powerful because they can transform information quickly and fluently, but every transformation creates another chance for meaning to shift. This becomes especially serious in systems that use retrieval, summarization, agents, automated reporting, or multi-step reasoning. Each step may appear successful on its own while the final output carries a weakened version of the original meaning.

Examples of Semantic Fidelity Loss

Translation: A sentence may be grammatically correct in another language while changing tone, intent, or implication.

Summarization: A summary may capture the main point while removing the qualifications that made the original claim accurate.

Retrieval-Augmented Generation: A system may retrieve the correct source but generate an answer that overstates, misreads, or recontextualizes what the source says.

Policy Communication: An instruction may pass through multiple departments until the original purpose becomes harder to recover.

Scientific Reporting: A cautious finding may become a simplified headline, then a general claim, then a belief about what the science proves.

AI Assistance: A model may produce a helpful answer that sounds aligned with the user's intent while quietly changing the task, assumptions, or conclusion.

In each case, the output may remain usable. The problem is that usability is not the same as fidelity.

Can Semantic Fidelity Be Measured?

Semantic fidelity can be evaluated, but it should not be treated as a single simple score. Some parts can be checked directly: whether the output contradicts the source, preserves key claims, retains

qualifications, cites evidence accurately, maintains the same level of certainty, and preserves relationships between concepts. Other parts require interpretation, such as whether intent has shifted, framing has changed, context has been narrowed in a misleading way, or a conditional claim has become unconditional.

A useful semantic fidelity evaluation should examine source preservation, context retention, qualification preservation, intent preservation, implication control, constraint preservation, citation support, contradiction detection, scope control, and uncertainty preservation. This means semantic fidelity can be partially operationalized, but not fully replaced by a single similarity metric. Similarity can help detect relatedness. It cannot, by itself, prove meaning preservation.

Semantic Fidelity and Reality Drift

Semantic fidelity is one of the core constraints that prevents Reality Drift. Reality Drift occurs when representations remain coherent, useful, and operational while gradually losing alignment with the realities they were built to describe. Semantic fidelity describes one version of that problem: meaning remains legible while its relationship to the original source weakens.

This can happen in AI systems, institutions, media, education, science, and everyday communication. The surface form survives and the system continues functioning, but the connection between representation and reality becomes less reliable. When semantic fidelity declines, systems may still appear aligned because their outputs remain coherent. That is the danger. Coherence can conceal drift.

Conclusion

Semantic fidelity names a problem that accuracy, similarity, and fluency do not fully capture. A system can be fluent without preserving meaning, similar without preserving implication, factual in pieces while changing the structure of the original claim, or sourced while weakening the relationship between evidence and answer.

The deeper challenge is meaning preservation across transformation. As information moves through summaries, translations, embeddings, retrieval systems, context windows, and generated outputs, the question is not only whether the result sounds right. The question is whether it still carries the meaning it was supposed to preserve. Semantic fidelity is the constraint that keeps representation answerable to meaning.

Related Phrases and Concepts

- semantic fidelity
- meaning preservation
- fidelity decay
- semantic similarity
- semantic drift
- hallucination

- grounding
- faithfulness
- source fidelity
- context retention
- Reality Drift

Semantic Fidelity Project Resources

- Research Library (GitHub)
- Articles & Essays (Substack)
- Semantic Fidelity Glossary
- LLM Failure Modes and Semantic Fidelity