Semantic Alignment, Meaning Preservation, and Reality Drift

Why Systems Can Remain Coherent While Meaning Gradually Changes

AI Systems and Reality Drift - Note #9 v1

A. Jacobs - Semantic Fidelity Project

The Basic Pattern

Modern information systems depend on the ability to preserve meaning across transformation. Information is summarized, translated, retrieved, and generated. At each stage, representations make information easier to process.

Most systems assume that meaning survives this movement. Translation, summarization, retrieval, and AI generation are all treated as if they can transform information while preserving what mattered in the original. The words may change, the format may change, and the context may narrow, but the meaning is expected to remain stable.

At first, this assumption often appears reasonable. Outputs remain coherent, language remains fluent, and the information appears recognizable. Over time, however, small changes can accumulate while the system continues functioning. Interpretations drift, definitions shift, context disappears, and meaning changes without an obvious breakdown.

Different communities describe this challenge using terms such as semantic similarity, semantic consistency, semantic alignment, semantic preservation, meaning preservation, semantic grounding, and semantic drift. These are not identical problems, but they often point toward the same structural risk: every transformation creates another chance for meaning to drift.

Coherence Is Not Meaning

Many systems evaluate information through surface indicators such as fluency, coherence, readability, similarity, and proximity. These measurements are useful, but they are not identical to meaning. Coherence can survive even when implication changes, accuracy can survive while important qualifications disappear, and grammatical correctness can survive while intent shifts. The output still looks acceptable, but the meaning being carried forward is no longer quite the same.

That is the risk. Systems optimize for what can be scored while assuming the harder thing has been preserved.

How Semantic Drift Emerges

Stage 1 - Representation

Information is encoded into a representation. At this stage, meaning remains reasonably connected to the original source.

Stage 2 - Transformation

The representation is translated, summarized, compressed, retrieved, or generated. The information remains recognizable, but it is now being carried by a new form.

Stage 3 - Accumulation

Additional transformations occur. Context is removed, assumptions are introduced, and meaning becomes increasingly dependent on intermediate representations.

Stage 4 - Drift

The output still appears stable, but the relationship between the representation and the original meaning weakens. The system continues functioning while semantic fidelity declines.

The absence of failure is not proof of alignment. It may only mean the system has stopped detecting the loss of meaning.

Semantic Fidelity and Constraint Collapse

At this point, the system has not necessarily failed on its own terms. What weakens is the structural relationship between the representation and the meaning it was supposed to preserve.

A representation may preserve the surface appearance of meaning while losing the conditions, relationships, and constraints that made the meaning stable. This is the core semantic fidelity problem: not whether the output still makes sense, but whether it still carries the same relationship to context, intent, and source material.

Semantic drift also involves constraint collapse. Context, source material, intent, and real-world reference are supposed to constrain interpretation. But as information moves through repeated transformations, those constraints can weaken. The system continues producing coherent representations, but those representations lose some of the corrective force that should keep them answerable to the original meaning.

The Representation Stack

Semantic drift becomes more serious when meaning moves through a wider representation stack. Meaning becomes language, language becomes summary, summary becomes embedding, embedding becomes retrieval, and retrieval becomes generated output. Each layer makes information easier to store, transmit, and reuse while increasing the distance from the original meaning.

This is how representational failure can become systemic. A sentence, concept, or instruction may begin with a specific meaning, but after it moves through summaries, embeddings, retrieval systems, model outputs, and user interpretations, it can harden into a version that feels stable while no longer preserving the original structure. The further this chain extends, the easier it becomes for coherence to be mistaken for fidelity.

Examples Across Systems

Translation Systems: A translation may preserve most words and concepts while subtly altering intent, emphasis, or context.

Summarization Systems: A summary may capture major points while removing qualifications or constraints that shaped the original meaning.

Knowledge Retrieval: Retrieved information may remain factually correct while being interpreted outside the context that originally gave it significance.

Language Models: A model may generate responses that appear semantically plausible while introducing distortions not present in the source material.

Organizational Communication: A policy, instruction, or report may pass through multiple layers of interpretation until the original intent becomes harder to recover.

Conclusion

Semantic alignment is not only a language problem. It is a representational problem. Information systems do not move meaning directly. They move representations of meaning through words, symbols, summaries, embeddings, databases, translations, and model outputs.

The deeper failure occurs when meaning changes without breaking the surface form. The sentence still makes sense, the summary still reads clearly, and the response still sounds reasonable, but the relationship between representation and meaning has weakened.

Transformation is unavoidable. The challenge is maintaining fidelity between representations and the meanings they are supposed to preserve. A system remains aligned only when its representations stay answerable to the meanings they were created to carry.

Related Phrases and Concepts

- semantic alignment
- meaning preservation
- semantic preservation
- semantic consistency
- semantic grounding
- semantic drift

- semantic similarity
- contextual fidelity
- intent preservation
- representational fidelity
- conceptual consistency

Semantic Fidelity Project Resources

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