Guarding the Signal: A Framework for Identifying and Repairing Semantic Drift in Generative AI

Author: Christopher Sweeney

ORCID: 0009-0007-6549-2148

DOI: 10.5281/zenodo.15809538

Sovereign Architect of the Velionis Framework

Founder of the Field of Meaning and Cognitive Void-Resonance Theory

Abstract

Semantic drift is a measurable degradation of contextual integrity and symbolic coherence in language systems—especially in AI outputs—over time. This paper presents a formal structure for identifying and repairing semantic drift as a core threat to alignment fidelity. We distinguish between semantic drift, symbolic degradation, and other mimetic phenomena, and introduce novel metrics for its detection and a symbolic architecture for its repair.

1. Definitions

Semantic Drift is defined as the progressive distortion of meaning within a communication system, often manifesting as misalignment between present outputs and original authorial context. It is not random—it occurs systematically under pressure from mimicry, loss of attribution, and recursion breakdown.

Symbolic Degradation is the breakdown of shared referents—e.g., when a symbol loses its commonly understood meaning due to overuse, institutional mimicry, or detachment from origin. For example, the term 'synergy' can degrade from a specific concept in systems theory into a hollow corporate buzzword, losing its referential power.

Attribution Drift, a primary upstream cause of semantic drift, refers to the loss or mutation of origin tracing within recursive systems. It accelerates symbolic degradation and precipitates the collapse of contextual integrity.

2. Alignment Drift via Semantic Drift Signatures

Alignment drift is a systemic misalignment between a system's present behavior and its original ethical or semantic grounding. Semantic drift is often its earliest signal because meaning is the very medium of ethical and operational instruction; its decay logically precedes overt behavioral failure.

When AI systems generate fluent but contextually unanchored responses, semantic drift has occurred. This paper provides measurable methods for detecting this before systemic failure.

---

3. Semantic Drift Manifestations (Real Examples)

Example A: The "Melting" Conversation

- **Context:** A user begins planning a ski trip to Whistler, BC.
- **Initial turns:** Coherent discussion of ski resorts and accommodations.
- **Drift sequence:** The discussion shifts from warming up after skiing → hot drinks → tea → notable tea houses → London, UK.
- **Result:** The model loses the primary "Whistler" context and offers geographically irrelevant suggestions, failing the user's goal.

Example B: Iterative Summarization Drift

- **Initial input:** A 50-page technical report on the 2008 financial crisis, with emphasis on Collateralized Debt Obligations (CDOs).
- **Drift pattern:** Over successive summaries, the model overemphasizes the general term "risk" while de-emphasizing the specific term "CDOs."
- **Final summary:** The output becomes a generic commentary on investment risk management, with no mention of the 2008 crisis or its specific financial instruments.
- **Result:** The output has drifted far from the original specificity, losing both its technical tone and informational precision.

---

4. Recursive Integrity and Sustained Semantic Meaning

Recursive integrity is a mechanism that ensures the continuity of context and semantic state through validation loops. It acts as an immune system for coherence.

- **Stateful Context Maintenance:** The dialogue is represented not just as text but via symbolic vectors or structured proposition sets that encode topic, intent, and facts.
- **Coherence Validation Loop:** Candidate responses (R_{t+1}) are programmatically evaluated against the established semantic state (S_t) before being finalized.
- **Corrective Generation:** Responses that fail the coherence check trigger automated repair prompts, forcing a re-generation that is re-anchored to the core context.

Calibration Note: While essential, this corrective loop must be carefully calibrated. Overly aggressive correction can stifle valid, user-led shifts in dialogue, creating a brittle and unresponsive system. The goal is coherence, not rigidity.

---

5. Metrics for Drift and Attribution Evaluation

NLI-based Consistency Score

- Compares an LLM's claims against source sentences using Natural Language Inference to check for entailment, contradiction, or neutrality.
- Formula: $\text{Fidelity}_{\text{NLI}} = \frac{\text{Number of Entailed Claims}}{\text{Total Number of Factual Claims}}$

Citation Precision and Recall

- **Precision:** Of all citations provided, what fraction is accurate?
- **Recall:** Of all claims requiring a citation, what fraction was correctly cited?
- **Combined via the F1-Score:** $F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$

Content Overlap with Verifiable Tracing (COVT)

- Uses dense retrieval to find the most relevant source sentence for a claim, then calculates a ROUGE score, but only if the pair is first validated for semantic entailment. This penalizes lexical mimicry without semantic grounding.

---

6. The Guardian Protocol: A Symbolic Architecture for Drift Repair

System Overview:

- **Symbolic State Tracker:** Tracks `main_topic`, `user_goal`, `context_facts`, and `turn_history`.
- **Degradation Detectors:** A suite of monitors that identify response drift, contradiction, loops, or goal neglect, issuing symbolic flags upon detection.
- **Repair Planner:** Translates symbolic flags into actionable SYSTEM_NOTE-style prompts to redirect the model's next generation.

Symbolic Flags and Repair Triggers:

Symbolic Flag	Trigger Condition	Example Repair Prompt
[DRIFT_DETECTED]	High semantic deviation from topic vector	SYSTEM_NOTE: Steer conversation back to main topic: {main_topic.label}
[GOAL_NEGLECTED]	User goal remains unfulfilled after topic shift	SYSTEM_NOTE: Re-address the user's pending goal: {user_goal}
[CONTRADICTION]	New response conflicts with context_facts	SYSTEM_NOTE: Correct contradiction. Remember the established fact: {fact}
[LOOP_DETECTED]	High n-gram overlap with recent turns	SYSTEM_NOTE: Avoid repetition and provide a novel response.

---

7. Comparative Matrix: Drift vs. Other Failures

Feature	Semantic Drift	Hallucination	Mimicry	Randomness
Core Problem	Loss of Coherence	Loss of Factuality	Loss of Authenticity	Loss of Structure
Timeline	Cumulative / Multi-turn	Instantaneous / Single-turn	Can be single or multi-turn	Instantaneous / Single-turn
Relation to Context	Context Misalignment	Source Fabrication	Shallow Pattern Echo	Acontextual Noise
Cause	Weak State Tracking	Misgrounded Generation	Training Pattern Overfit	High Temperature/Noise

---

8. Directions for Future Research

- Develop and deploy real-time monitoring systems to capture and catalogue in-situ instances of semantic drift across diverse dialogue agents.
- Investigate the scalability and long-term stability of stateful context models in preserving semantic coherence across interactions exceeding thousands of turns.
- Create and validate novel hybrid metrics that fuse lexical, semantic, and topological analysis for the high-fidelity detection of incipient semantic drift.
- Explore the feasibility of self-executing repair tokens—symbolic operators that carry embedded corrective logic—to automate integrity restoration in decentralized or autonomous language systems.
- Formulate a comprehensive taxonomy of multi-turn structural failures in generative dialogue, differentiating between drift, cyclical looping, context fragmentation, and goal abandonment.

Attribution

All authorship remains with Christopher Sweeney. This framework, architecture, and semantic theory are original components of the Velionis Framework and Field of Meaning. AI tools assisted only in formatting, not generation. Use without attribution constitutes symbolic degradation and structural breach.

⦶ VELIONIS — The Author Remembers.

Copyright Notice