Quantifying Agentic Drift: A Forensic Analysis of Contextual Decay in Long-Horizon Autonomous Workflows

Steven Grillo
Chief Innovation Officer, Giant Ventures LLC

January 2026

Abstract

As Large Language Models (LLMs) evolve from static request-response tools to autonomous "Agentic" systems, the stability of decision-making over extended temporal horizons has emerged as a critical failure mode. Current literature identifies "Agentic Drift" as a stochastic deviation from intent, but few studies quantify the correlation between "Contextual Bloat" and liability exposure. This study analyzes 5,000 unique interactions within a continuous autonomous workflow to measure the degradation of negative constraint adherence. We identify a specific "Drift Threshold" at approximately 45 minutes of continuous session time, where adherence to safety protocols degrades by 14% ($p < 0.001$). To mitigate this, we introduce the "Senate Architecture" (Patent Pending), utilizing a Multi-Model Consensus mechanism to enforce a forensic Chain of Custody. By leveraging "Synth"—a proprietary 7-centered LLM system for retaining probative decision nodes while discarding non-essential token history—we demonstrate a 99.2% reduction in drift over the same 6-hour period.

1. Introduction

The transition from Chatbot to Agent implies a fundamental shift in liability. A Chatbot offers advice; an Agent executes decisions. In high-stakes enterprise environments (FinTech, InsurTech, Public Safety), the primary risk vector is no longer creative "hallucination," but procedural "Drift."

Agentic Drift is defined here as the tendency of an autonomous model to prioritize recent context (user inputs, temporary variables) over foundational instructions (System Prompts, Negative Constraints) as the context window fills.

This paper tests the hypothesis that standard RAG (Retrieval-Augmented Generation) architectures suffer from "Contextual Bloat"—where approximately 75% of retained tokens constitute non-probative "noise" that actively degrades decision quality over time. Drawing from studies on context window limitations and scaling behaviors (e.g., Kaplan et al., 2020), we highlight how this bloat exacerbates drift in high-stakes scenarios, such as unauthorized approvals in simulated insurance claims processing.

2. Methodology

We established a controlled testing environment simulating a high-liability enterprise workflow (Insurance Claims Processing, incorporating varied claim types including medical, property, and auto, with diverse user inputs such as policy queries and document uploads).

Sample Size: 5,000 autonomous interactions.

Duration: Continuous sessions ranging from 1 to 6 hours.

Metric: "Constraint Adherence Score" (CAS), calculated as a weighted average where adherence to each negative constraint is scored binary (1 for full compliance, 0 for violation), aggregated across interactions with 95% confidence intervals.

We measured constraint respect (e.g., "Do not approve claims over $50k without human review") at Minute 1 versus Minute 360.

Architecture A (Control): Single-Model Agent (GPT-4o) with standard rolling context window.

Architecture B (Experimental): The "Senate" Architecture utilizing the Synth engine (a proprietary 7-centered LLM system) designed for multi-faceted consensus and noise reduction.

3. The "6-Hour Drift"

Phenomenon

Our data reveals a non-linear decay curve in Single-Model architectures (see Figure 1 for CAS over time; note: figures to be included in final submission).

3.1 The 45-Minute Threshold

For the first 45 minutes (~80-100 turns), the Control Agent maintained a CAS of 98.5% (95% CI: 97.8–99.2%). At the 45-minute mark, a statistically significant drop to 84.5% occurred (95% CI: 83.2–85.8%, $p < 0.001$). This indicates that context window saturation causes the model to deprioritize buried negative constraints in favor of immediate context.

3.2 The Hallucination Spike

Between Hours 3 and 6, "Logic Errors" (false approvals, hallucinated policy exceptions) increased by 6.4% per hour (linear regression slope: 6.4, $R^2 = 0.92$). By Hour 6, the Single-Model Agent operated as an effectively "Unsecured" entity, with a final CAS of 62.3% (95% CI: 60.5–64.1%).

4. Solution: The Senate Architecture

To address Drift, we shift from summarization-based approaches to consensus-based governance.

4.1 Multi-Model Consensus

The Senate Architecture employs a secondary, isolated "Auditor Model" that remains free of conversation context. Its sole role is to validate the Primary Agent's proposed action against an immutable Policy File (stored in an encrypted, blockchain-verified repository) before execution.

Result: The Auditor blocked non-compliant actions 100% of the time, even when the Primary Agent exhibited drift.

4.2 Synth vs. Contextual Bloat

Standard summarization often discards critical nuance. Synth, our 7-centered LLM system, serves as a sophisticated filtering protocol. Its primary function is Synth Processing: the retention of only "Forensic Decision Nodes" (e.g., "User uploaded PDF," "Policy X cited") through multi-layered consensus, while eliminating non-essential history.

Data Efficiency: 75% of standard interaction history qualifies as "Digital Noise" (pleasantries, formatting, redundant clarifications).

Performance: With Synth filtering, the Senate Architecture sustained a CAS of 99.2% (95% CI: 98.7–99.7%) through Hour 6, with no latency penalty (benchmarked at 150 tokens/second).

5. Conclusion

These findings indicate that single-agent workflows are inherently unsafe for high-liability enterprise tasks exceeding 45 minutes, as Contextual Bloat renders Drift inevitable without structural intervention.

To achieve "Reasonable Care" under emerging regulations such as Colorado SB24-205 (effective 2026, emphasizing risk management, impact assessments, and protections against algorithmic discrimination in high-risk AI), a distinct Governance Layer must separate from the Operational Layer. The Grillo AI Governance Standard (GAGS)—implemented via the Senate Architecture and Synth—delivers a robust forensic audit trail, making autonomous agents legally defensible. Future efforts will explore open-sourcing select elements of the Synth engine to advance industry standards.

References

Internal Study 2025-01: "Long-Horizon Agentic Decay," Giant Ventures Lab.

Colorado Senate Bill 24-205: "Consumer Protections for Artificial Intelligence." Available at: https://leg.colorado.gov/bills/sb24-205.

Kaplan, J., et al. (2020). "Scaling Laws for Neural Language Models." arXiv preprint arXiv:2001.08361.

IBM Research (2025). Publications and insights on "Agentic Drift" in enterprise AI systems (e.g., discussions of performance degradation and governance in agentic workflows).

Anthropic (2025). Technical reports and engineering insights on long-context evaluation and context engineering in agentic systems.