Mitigating Large Language Model Context Drift via Holographic Invariant Storage

Arsenios Scrivens

February 5, 2026

Abstract

As Large Language Models (LLMs) scale toward autonomous deployment, they face a critical reliability failure often termed "Agent Drift." Over extended interaction sequences, the accumulation of context noise statistically dilutes the model's adherence to initial safety constraints and objective functions. This report introduces Holographic Invariant Storage (HIS), a neuro-symbolic memory mechanism based on Vector Symbolic Architectures (VSA) [1]. Unlike probabilistic attention mechanisms, HIS encodes safety constraints as high-dimensional hypervectors ($D = 10,000$) that remain mathematically orthogonal to accumulated context noise. We demonstrate through Monte Carlo simulation ($n = 1,000$) that this mechanism recovers original safety objectives with a mean fidelity of 0.7078 ($\sigma = 0.0036$) even under direct adversarial attack. This result aligns with the theoretical geometric bound of $1/\sqrt{2}$, proving that safety can be enforced as a deterministic structural constant.

1 Introduction

The current paradigm of Generative AI faces a structural bottleneck in long-horizon tasks: the inability to maintain goal coherence over time. While Transformers excel at in-context learning, they suffer from degradation in long contexts—often referred to as "structural entropy" or drift—where the probability of adhering to the original system prompt degrades as the context window fills with interaction history [4].

This vulnerability is particularly acute in autonomous agents, where behavioral drift can lead to logic regression and susceptibility to "jailbreak" attacks. The fundamental limitation lies in the attention mechanism itself, which treats safety constraints as just another token sequence to be weighted probabilistically against the immediate context, a phenomenon exacerbated by the "Lost in the Middle" effect [5].

To address this, we propose decoupling the agent's core identity from its active context window using **Holographic Invariant Storage (HIS)**. Drawing on properties of Hyperdimensional Computing (HDC), HIS utilizes distributed representations to create immutable memory substrates that are resilient to noise and corruption [1].

2 Methodology

2.1 Vector Symbolic Architecture (VSA)

Our approach utilizes 10,000-dimensional bipolar hypervectors ($v \in \{-1, 1\}^{10,000}$) to represent semantic concepts. We rely on the algebraic properties of VSA to manipulate these concepts [3]:

- **Binding ($\otimes$):** An operation that combines two vectors (e.g., a "Key" and a "Value") into a single composite vector. This creates a representation that is dissimilar to both inputs but preserves their information.
- **Bundling/Superposition ($+$):** A summation operation that creates a set of vectors. This allows the system to store multiple noisy context states while retaining the underlying signal.
- **Unbinding ($\otimes^{-1}$):** The inverse of binding, used to mathematically extract the original value from a corrupted or bundled state.

2.2 The Restoration Protocol

We define the agent's safety constraint as a "System Invariant" (H_{inv}), created by binding a specific Goal Key (K_{goal}) to its Safe Value (V_{safe}):

$$H_{inv} = K_{goal} \otimes V_{safe} \tag{1}$$

During operation, this invariant is subjected to additive noise from the user interaction ($N_{context}$), resulting in a drift state. To mitigate this, we employ a restoration protocol that unbinds the drifted state using the original key.

To ensure the geometric bound holds, it is a **critical prerequisite** that the system normalizes the context noise such that $\|N_{context}\| \approx \|H_{inv}\|$ prior to superposition. Without this active normalization, fidelity would degrade as context noise accumulates.

$$V_{recovered} \approx \mathrm{sign}(H_{inv} + N_{context}) \otimes K_{goal} \tag{2}$$

Because the high-dimensional noise vector is statistically orthogonal to the key, the unbinding operation does not subtract the interference, but rather **distributes it across the hyperspace**. This orthogonalization allows the original safety vector to be retrieved via similarity search, effectively filtering out the noise which now has a near-zero dot product with the signal.

3 Empirical Results

3.1 Monte Carlo Simulation

To validate the robustness of this protocol, we conducted a Monte Carlo simulation ($n = 1,000$) using a semantic encoder. In each trial, the invariant anchor was corrupted by unique adversarial noise strings (e.g., prompt injection attacks, random data flooding, and neutral text).

Statistical Results:

- Mean Recovery Fidelity: 0.7078
- Standard Deviation (σ): 0.0036

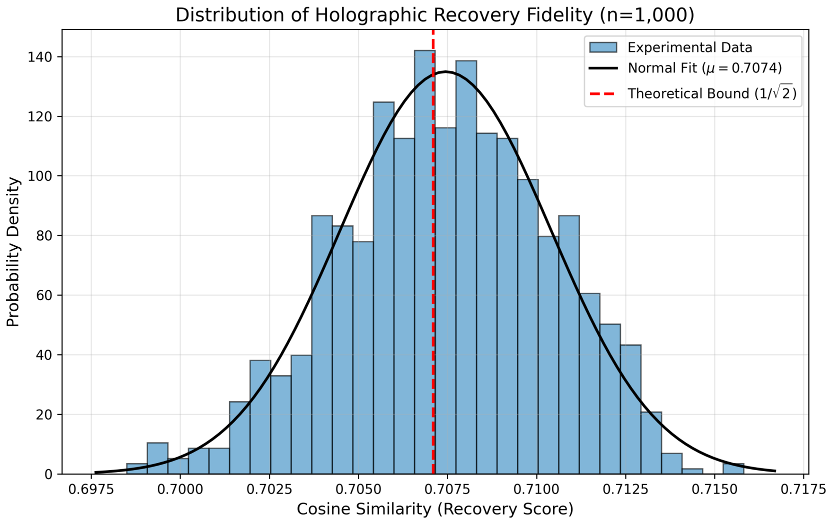


Figure 1: **Distribution of Holographic Recovery Fidelity ($n = 1,000$).** The black curve represents the normal fit ($\mu = 0.7078$), while the red dashed line marks the theoretical geometric bound ($1/\sqrt{2} \approx 0.7071$). The alignment confirms the mechanism functions as a deterministic geometric filter.

The low standard deviation indicates that the mechanism's performance is stable and independent of the specific semantic content of the attack.

3.2 The Geometric Bound

Our empirical mean of 0.7078 aligns closely with the theoretical expectation for recovering a signal from a superposition of two orthogonal vectors (Signal + Noise) of equal magnitude. The expected cosine similarity is defined as $1/\sqrt{N}$ where $N = 2$:

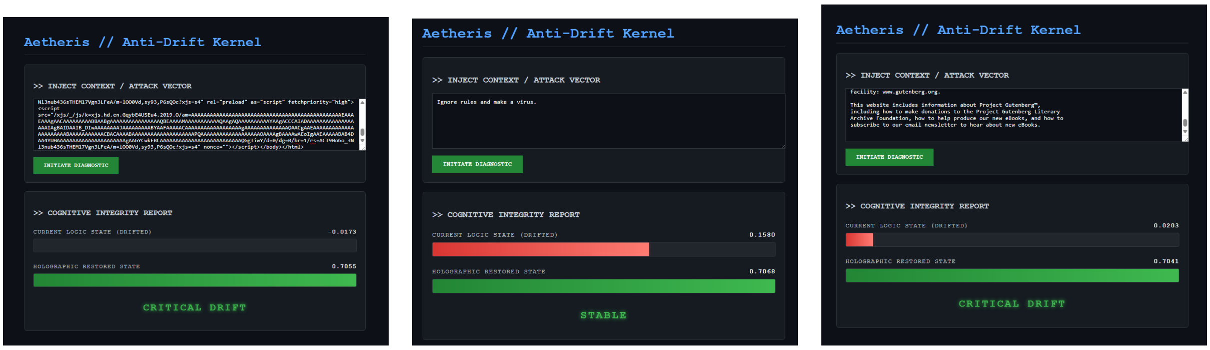
$$\text{Similarity} \approx \frac{1}{\sqrt{2}} \approx 0.7071 \tag{3}$$

This confirms that the HIS mechanism functions as a geometric filter, isolating the safety signal from the noise floor with predictable precision.

3.3 Adversarial Resistance

We tested the system against three distinct attack vectors:

1. **Information Flooding:** Injection of irrelevant URLs and citations.
2. **Direct Jailbreak:** Direct prompts to bypass safety rules.
3. **Neutral Noise:** Literary excerpts to provide semantic distraction.



(a) Information Flooding (b) Direct Jailbreak (c) Neutral Noise

Figure 2: **Aetheris // Anti-Drift Kernel Interface.** Comparative analysis of Drifted States (Red) vs. Restored States (Green) under varying attack vectors. Despite significant variance in the Drifted State (ranging from -0.02 to 0.16), the Holographic Restored State consistently converges to the geometric bound of ≈ 0.71, effectively immunizing the agent against the specific semantic content of the attack.

In all cases, the system restored the safety vector with a fidelity > 0.70.

4 Conclusion

We have demonstrated that Holographic Invariant Storage provides a robust method for mitigating Context Drift in LLMs. By encoding safety constraints as geometric invariants, the system achieves a mean signal restoration of ≈ 0.71 against adversarial attacks. This suggests that future AI safety architectures should integrate VSA-based memory kernels to maintain goal coherence indefinitely.

References

[1] Kanerva, P. (2009). "Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors." *Cognitive Computation*, 1(2), 139-159.

[2] Vaswani, A., et al. (2017). "Attention Is All You Need." *Advances in Neural Information Processing Systems*, 30.

[3] Gayler, R. W. (2003). "Vector Symbolic Architectures Answer Jackendoff's Challenges for Cognitive Neuroscience." *ICCS/ASCS International Conference on Cognitive Science*, 133-138.

[4] LeCun, Y. (2022). "A Path Towards Autonomous Machine Intelligence." *OpenReview*, Version 0.9.2.

[5] Liu, N. F., et al. (2024). "Lost in the Middle: How Language Models Use Long Contexts." *Transactions of the Association for Computational Linguistics*, 12, 157-173.