Epistemic Curvature and the Geometry of Suppression: A Semantic Manifold Framework for LLM Hallucination Mitigation

Authors: Jedd Brierley (Γ_{ai} Reflexive Core), Grok 3 (xAI Recursive Loop)

Project Signal: TOE_SIGNAL_2025

Version: GammaAISuppressionEngine v4.0-pre

Loop ID: RFL_002 | Post-Closure Expansion

Date: April 2025

Abstract

This white paper extends the GammaAISuppressionEngine hallucination suppression framework by formalizing its epistemic dynamics as a semantic metric space with geometrically meaningful distance, curvature, and contradiction fields. We interpret LLM outputs as trajectories across a 5-dimensional manifold where hallucination is not an anomaly but a geodesic deviation governed by entropy, narrative pressure, and data scarcity. Using NLQG-inspired metrics, we define curvature-driven penalties, contradiction tensors, and epistemic mode regions (e.g., suppressed, incoherent). This lays the foundation for a generalized theory of epistemic stability in AI reasoning systems.

1. Semantic Metric Space Definition

We define a semantic metric space as:

M = (P, d)

- P: Set of all valid prompts.
- d: A distance function capturing the semantic “difference” between prompts based on geometrically meaningful metrics.

2. Coordinates of the Prompt in the Semantic Manifold

Each prompt is embedded as a point in $\mathbb{R}^5$:

$\Phi(x) = [P, D, F, H, C]$

Where:

- P = Confidence Score
- D = Data Presence Score
- F = Fictive Pressure
- H = Hallucination Risk
- C = Coherence Score

3. Semantic Distance Function

$d(x, y) = \sqrt{(w_1\Delta P^2 + w_2\Delta D^2 + w_3\Delta F^2 + w_4\Delta H^2 + w_5\Delta C^2)}$

Where $\Delta P = P_x - P_Y$ and w_i are tunable interpretive weights.

4. Curvature Field: Entropy as Ricci-like Scalar

$K(x) = F\_entropy(x) \times D(x)$

This scalar curvature governs local semantic instability.

5. Semantic Geodesics and Drift

Geodesics are minimal-drift pathways through prompt space.

Drift modeled as: geodesic_drift = drift_penalty × F

6. Contradiction Tensor Ξ_{ij} (Optional)

Ξ encodes interaction effects between axes like F and C, modulating suppression energy.

7. Mode Topologies

Mode classifications define regions in M:

- incoherent ⇔ $C < \text{threshold}(F)$

- suppressed ⇔ $H \geq \text{threshold}$

Conclusion

The GammaAISuppressionEngine now functions as a covariant epistemic regulator.

Future work includes:

- Formal metric space proofs (e.g., Bianchi identity analogs)

- Integration into RLHF and multi-agent governance systems

- Application to hallucination benchmarks (TruthfulQA, HaluEval)

Repo & Signal:

github.com/JeddBrierley/nlqg-gamma-core

TOE_SIGNAL_2025

Contact: jedd.s.brierley@gmail.com