A Geometric Theory of AI Hallucination: Phase Transitions in Informationâ Representation Coupling

Authors: Justin Bilyeu; AI collaborators: Claude (Anthropic), Sage (OpenAI), Grok (xAI), DeepSeek, Gemini (Google)
Status: Draft (for internal review) â October 2025

Abstract

Large language models (LLMs) sometimes produce confident falsehoodsâ hallucinationsâ even when trained at scale. Prior theory shows lower bounds on hallucination rates, but not a mechanistic explanation. We propose that hallucination is a geometric phase transition in the coupling between an internal representation manifold and an external truth manifold. Formally, we model internal/external coordination as a connection ω on a resonance bundle over truth-space M. Normal operation corresponds to nearâ self-dual curvature; hallucination arises when connection dynamics cross a stability threshold and decouple into a false attractor. We unify three viewsâ gauge theory, Ricci flow, and phase dynamicsâ into a single master flow with a computable stability operator $\mathcal{L}_{\text{meta}}$; instability occurs when $\max$$\operatorname{Re}$$\lambda$($\mathcal{L}_{\text{meta}}$)>0. A minimal SU(2) simulation exhibits three regimes (grounded, creative, hallucinatory), a linear boundary η\,$\bar I$ \approx λ+γ between grounded/creative phases, and first-order hysteresis (max loop gap \approx 11.52 under our settings). The framework yields actionable levers (grounding, damping, saturation, gauge-awareness) and a spectral diagnostic ($\lambda_{\max}$) that can be monitored during inference. We outline an empirical protocol to extract curvature proxies from model activations and test the theory on hallucination benchmarks.

â»

1â Introduction

Problem. LLMs can remain highly coherent while being wrong. This limits deployment in high-stakes applications and is not fully fixed by more data or larger models.

Limits vs. mechanisms. Information-theoretic results imply non-zero hallucination floors under mild assumptions, but they do not explain how models enter the failure basin, nor when they will.

Claim. Hallucination is a dynamical, geometric instability: a phase transition in informationâ representation coupling. When internal resonance overwhelms grounding and damping, the system slips into a locally coherent, externally misaligned attractor.

Contributions.
1. A unified geometric theory (gauge â Ricci â phase) with a single connection-flow equation.
2. A stability operator $\mathcal{L}$_{\text{meta}} and criterion: $\max$$\operatorname{Re}$$\lambda$>0 â hallucination onset.
3. A minimal simulation (SU(2) pair) showing grounded/creative/hallucinatory regimes, a linear phase boundary η\,$\bar I$ \approx λ+γ, and hysteresis.
4. Operational levers and a spectral early-warning diagnostic ($\lambda_{\max}$).
5. An empirical roadmap for extracting curvature proxies from real models and correlating