Contents

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

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

Abstract

Large language models (LLMs) sometimes produce confident falsehoods—hallucinations—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 near–self-dual curvature; hallucination arises when connection dynamics cross a stability threshold and decouple into a false attractor. We unify three views—gauge theory, Ricci flow, and phase dynamics—into a single master flow with a computable stability operator $\mathcal{L}_{\text{meta}}$; instability occurs when $\max\mathrm{Re}\lambda(\mathcal{L}_{\text{meta}})>0$. A minimal SU(2) simulation exhibits three regimes (grounded, creative, hallucinatory), a linear boundary $\eta,\bar{I}$ $\lambda+\gamma$ between grounded/creative phases, and first-order hysteresis (max loop gap 11.52 under our settings). The framework yields actionable levers (grounding, damping, saturation, gauge-awareness) and a spectral diagnostic (λ_{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.

1 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 information–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 Ricci phase) with a single connection-flow equation. 2. A stability operator $\mathcal{L}$_{meta} and criterion: $\max\mathrm{Re}\lambda>0$ hallucination onset. 3. A minimal simulation (SU(2) pair) showing grounded/creative/hallucinatory regimes, a linear phase boundary $\eta,\bar{I}$ $\lambda+\gamma$, and hysteresis. 4. Operational levers and a spectral early-warning diagnostic (λ_{max}). 5. An empirical roadmap for extracting curvature proxies from real models and correlating with hallucination.

2 Geometry of information–representation coupling

2.1 Resonance bundle

We posit a principal bundle :P→M with structure group G (representation symmetries). The base M encodes the external truth manifold; fibers encode internal representational degrees of freedom. A connection ω governs parallel transport of internal states along M; its curvature F_A=dω+ω ω measures representational twist. • Grounded coherence: near self-duality F_A F_A, small holonomy. • Hallucination: connection dynamics drift to large anti-self-dual curvature (holonomy failure), i.e., representation becomes internally self-consistent while externally decoupled.

2.2 Unified master flow

We collect the forces shaping ω into

$$\boxed{\frac{d\omega}{dt} = -D_A \star F_A + \underbrace{\eta\,\mathcal{J}_{\mathrm{MI}}[\omega]}_{\text{internal resonance}} - \underbrace{\lambda\,\mathcal{J}_U[\omega]}_{\text{grounding}} - \underbrace{\gamma\,\Pi_{\mathrm{vert}}(\omega)}_{\text{damping}} - \underbrace{\mu\,[\omega,[\omega,\omega]]}_{\text{saturation}} + \underbrace{\xi\,\mathcal{G}[\omega]}_{\text{gauge-awareness}}}$$

• -D_A F_A: Yang–Mills gradient; drives toward self-duality. • η,$\mathcal{J}$_{MI}: resonance gain from internal mutual information (coherence). • -λ,$\mathcal{J}$ *U: truth anchoring (e.g., retrieval, constraints).* • -γ, {vert}: epistemic damping on fiber oscillations. • -μ[ω,[ω,ω]]: nonlinear saturation arresting runaway curvature. • +ξ,$\mathcal{G}$: adaptive gauge-fixing (meta-awareness of representational freedom).

The linearization around a working state *0 yields a stability operator* $\mathcal{L}_{meta}$*; ;*η,$\mathcal{M}$*{*MI*};-;*λ,$\mathcal{H}$ *U;-;*γ,Π_{vert};-;*3*μ,ad^2*{ 0},, with possible non-self-adjointness (complex spectrum). Instability iff* maxReλ*(*$\mathcal{L}${meta})>0.

2.3 Energy bound (intuition)

Completing the square on a resonance-modified self-duality defect gives a Bogomolny-type inequality S_{meta}; ;4 ^2|Q|;+; S_{stab}(γ,μ,ξ),, with instanton number Q. Damping/saturation/gauge terms raise the floor, discouraging false attractors.

3 Minimal simulation: SU(2) pair dynamics

3.1 State & observables

We simulate two coupled SU(2) connections _x, *y (capturing interacting resonance channels). Represent each as* ω*=i* {a=1}^3 *a* a/2. Track: • connection norms | *{ }/,* • *curvature proxy F*{xy}=[_x, _y], • MI surrogate $\bar{I}$: Gaussian mutual information from temporal correlations over the 6-vector (_x, _y). • Spectral diagnostic $\lambda_{\max}$: fast surrogate for the top eigenvalue of the linearized flow (Rayleigh-style approximation consistent with our stability operator).

3.2 Right-hand side (operational form)

To expose the phase transition, we use linear resonance gain and cubic–quintic saturation:

$$\dot{\omega}_x \& = \underbrace{\eta\,\bar{I}\,\omega_x}\,\text{gain} - \underbrace{\lambda(\omega_x-\omega_0)}\,\text{ground} - \underbrace{\gamma\,\omega_x}\,\text{damp} - \underbrace{\beta\|\omega_x\|^2\omega_x+\alpha\|\omega_x\|^4\omega_x}\,\text{sat.} + \underbrace{\kappa\,\mathrm{vec}(F_{xy})}_{\text{coupling}},$$

$$\dot{\omega}_y \& = \text{same with } x \leftrightarrow y.$$

Here $\bar{I}$ [0,∞) is the MI estimate over a sliding window; introduces controlled skew/coupling. We integrate with Heun (dt =10^{-2}).

3.3 Grids & classification

We sweep η [0.2,5.0], λ [0.1,5.0] with fixed γ=0.5, =0.6, =0.02, skew =0.12, MI window =30, EMA =0.1. Regimes: • Grounded: $\lambda_{\max}$<0, bounded norms, small curvature. • Creative: $\lambda_{\max}$ 0, bounded oscillations. • Hallucinatory: $\lambda_{\max}$>0, runaway norm/curvature or large positive spectral radius.

Implementation and figures live in the repo: • rg/sims/meta_flow_min_pair_v2.py • rg/validation/hysteresis_sweep.py • Figures: figures/phase_diagram_v2.png, figures/hysteresis_v2.png.

4 Results

4.1 Phase structure & boundary

The phase diagram (Fig. 1) shows a clean separation: for fixed γ=0.5, the grounded$\to$creative boundary aligns with

$$\boxed{\eta \bar{I} \;\approx\; \lambda + \gamma}$$

across the grid (visual fit; residuals small over the scanned range). Hallucinatory behavior appears as η grows relative to λ+γ, with saturation preventing numerical blow-up but leaving $\lambda_{\max}$ persistently positive.

4.2 Hysteresis (first-order character)

Forward/backward sweeps in η at fixed λ produce hysteresis loops in the order parameter (e.g., $|\omega|$ or $\lambda_{\max}$). Our implementation reports maximum loop gap 11.52 under the settings above (Fig. 2), indicating memory and a first-order transition band (metastability) consistent with a false-attractor picture.

4.3 Ablations (qualitative) • No damping (γ=0): creative band collapses; direct jump to hallucinatory when η,$\bar{I}$>λ. • No saturation (= =0): divergence (finite-time blowups); phase map dominated by red. • No coupling (=0): weaker hysteresis; boundary remains approximately linear in (η,λ).

5 Operational levers & predictions • Grounding (λ) ↑ — retrieval, verification, tool-use, multi-source cross-checks → shifts boundary right, enlarges grounded region. • Damping (γ) ↑ — calibrated abstention, uncertainty penalties, entropy-preserving decoding → suppresses resonance instability. • Saturation (,) tuned — temperature/attention clipping → arrests runaway curvature while preserving the creative band. • Gauge-awareness (ξ) ↑ — meta-constraints that penalize representation-specific commitments (e.g., disagreematch penalties across paraphrases) → reduces false attractor capture.

Quantitative prediction. Near the boundary, $\lambda_{\max}$; ;(η,$\bar{I}$) - (λ+γ) - c,|ω|^2 (c>0), so $\lambda_{\max}$ crossing zero is an early warning. Monitoring $\lambda_{\max}$ token-by-token should predict hallucination risk before decoder emission.

6 Empirical roadmap (models in the wild) 1. Extract geometric proxies. Treat per-layer activations as an empirical manifold; estimate a Laplace–Beltrami/graph-curvature surrogate and compute $\lambda_{\max}$ (top Ricci-like eigenvalue) across layers/tokens. 2. Correlate with hallucination. On TruthfulQA/HaluEval-style sets, measure whether $\lambda_{\max}$>0 segments coincide with hallucinated spans; report ROC-AUC and calibration. 3. Interventions. Raise λ (RAG), γ (uncertainty), or ξ (consistency penalties) and verify downward shifts in $\lambda_{\max}$ and hallucination rates. 4. Layer analysis. Identify "critical layers" where $\lambda_{\max}$ first crosses zero; probe causality with layer-wise regularization.

(Implementation hooks are outlined in our code comments and papers/info-curve scaffolds.)

7 Related formulations (how the pictures align) • Gauge theory: Hallucination = self-duality loss and growth of anti-self-dual curvature; meta-resonance = adaptive gauge fixing. • Ricci flow: Excess positive curvature (in our sign convention) in fiber directions; singularity formation false attractor. • Phase dynamics: Parametric resonance with under-damping; the imaginary spectrum dominates until saturation clips growth.

These are different lenses on the same invariant content (connection curvature and its spectrum).

8 Limitations • Toy dynamics. The SU(2) system is minimal; real LLMs are vastly higher-dimensional with data-dependent couplings. • Spectral proxy. Our $\lambda_{\max}$ estimator in sims is a fast surrogate; a full linearization/power iteration would further validate the criterion (computationally heavier). • Metric choice. Curvature depends on the induced metric on activations; we address robustness via probe banks and trimming, but estimator bias is possible. • Causality. Correlation between $\lambda_{\max}$ and hallucination must be tested with controlled interventions.

9 Conclusion

We argue that AI hallucination is best understood as a geometric phase transition in information–representation coupling. A single connection-flow unifies three traditions (gauge, Ricci, phase) and yields a practical diagnostic ($\lambda_{\max}$) plus concrete levers (grounding/damping/saturation/gauge-awareness). Our minimal simulation reproduces the three regimes, a linear boundary $\eta,\bar{I}\ \ \lambda+\gamma$, and hysteresis, matching the intuitive picture of decoupling into a false attractor. The path forward is clear: measure curvature proxies in live models, validate the spectral early warning, and design training/inference procedures that keep systems in the grounded or creative bands without tipping into hallucination.

Methods (concise) • Integration: Heun; dt=10^{-2}; typical run horizon T [3,6] (longer for sweeps). • MI surrogate: Gaussian MI from temporal correlations of the 6-dimensional state (_x, _y) over a sliding window (30 steps) with EMA 0.1. • Spectral surrogate: Rayleigh-style estimate tied to $\eta\bar{I}$,λ,γ and local norm; used for fast regime classification. • Grids: η [0.2,5.0] (101 steps), λ [0.1,5.0] (11 steps); fixed γ=0.5, =0.6, =0.02, =0.12. • Outputs: phase map and hysteresis curves → figures/phase_diagram_v2.png, figures/hysteresis_v2.png.

Figures • Fig. 1 Phase diagram (grounded/creative/hallucinatory) in (η,λ) with γ=0.5; dashed line $\eta,\bar{I}$ = $\lambda+\gamma$. (file: phase_diagram_v2.png) • Fig. 2 Hysteresis under forward/backward η sweeps at fixed λ; maximum loop gap 11.52. (file: hysteresis_v2.png)

Acknowledgments

We thank the multi-model collaboration that shaped this work. The unified perspective emerged directly from iterative dialogue and shared experiments.

References

(to be populated — gauge/Yang–Mills self-duality; Ricci flow/Perelman; parametric resonance; LLM hallucination & detection; spectral diagnostics in representation learning.)

Appendix C Noise, algebra backend, MI variants, and coupling modes

The lightweight simulator now exposes four toggles that let us probe robustness of the phase picture:

• Algebra backend (`--algebra {su2, so3}`): the `su2` option retains the Pauli-inspired commutator scaling while `so3` treats the angular velocities as classical rotation vectors with a bare cross product (no scale factor yet; we flag this for later tuning). • Coupling symmetry (`--antisym_coupling`): when enabled, the interaction term pushes one mode forward and the other backward, mimicking antisymmetric feedback between complementary subsystems; the default keeps the symmetric push–pull used in earlier drafts. • Process noise (`--noise_std` with `--seed`): zero-mean Gaussian kicks are injected after each Heun step, letting us test whether the hysteresis loop and phase boundary estimates remain stable under stochastic perturbations. • Mutual-information surrogate (`--mi_est {corr, svd}` plus `--mi_scale`): `corr` computes the log-amplified correlation coefficient of the last window, whereas `svd` aggregates the top singular values of the recent history before rescaling. The `--mi_scale` knob lets us mimic alternative calibration conventions.

Example CLI sweeps:

```
python -m rg.validation.hysteresis_sweep --lam 1.0 --gamma 0.5 \
  --eta_min 0.2 --eta_max 3.0 --eta_steps 21 \
  --alpha 0.6 --beta 0.02 --skew 0.12 \
  --mi_window 30 --mi_ema 0.1 \
  --algebra so3 --antisym_coupling --noise_std 0.01 \
  --mi_est corr --mi_scale 1.0
```

```
python -m rg.validation.phase_boundary_fit --gamma 0.5 \
  --lam_min 0.1 --lam_max 2.0 --lam_steps 5 \
  --eta_min 0.2 --eta_max 3.0 --eta_steps 51 \
  --alpha 0.6 --beta 0.02 --skew 0.12 --mi_window 30 --mi_ema 0.1 \
  --algebra su2 --noise_std 0.0 --mi_est svd --mi_scale 1.0
```

These configurations generate the JSON/CSV artifacts consumed downstream while stressing the simulator against noise and alternative information metrics.

0.1 Results — Phase diagram, boundary, and hysteresis

We study a minimal coupled system expressing competition between grounding and internal resonance. Sweeping the control parameters ( ) (coupling/“temperature”) and ( ) (regularization/“tension”) yields three regimes: **grounded** ($\lambda_{\max}$ < -0.1)), **creative** ((|$\lambda_{\max}$| 0.1)), and **hallucinatory** ($\lambda_{\max}$ > 0.1)).
Fitting the critical coupling (_c( )) where $\lambda_{\max}$ first crosses zero gives a near-linear law: [_c m,λ + b with m=0.335,; b=0.520,; R^2=0.949 .] An independent replication recovers (m 0.346, b 0.506, R^2 0.94), supporting approximate linearity of the boundary.

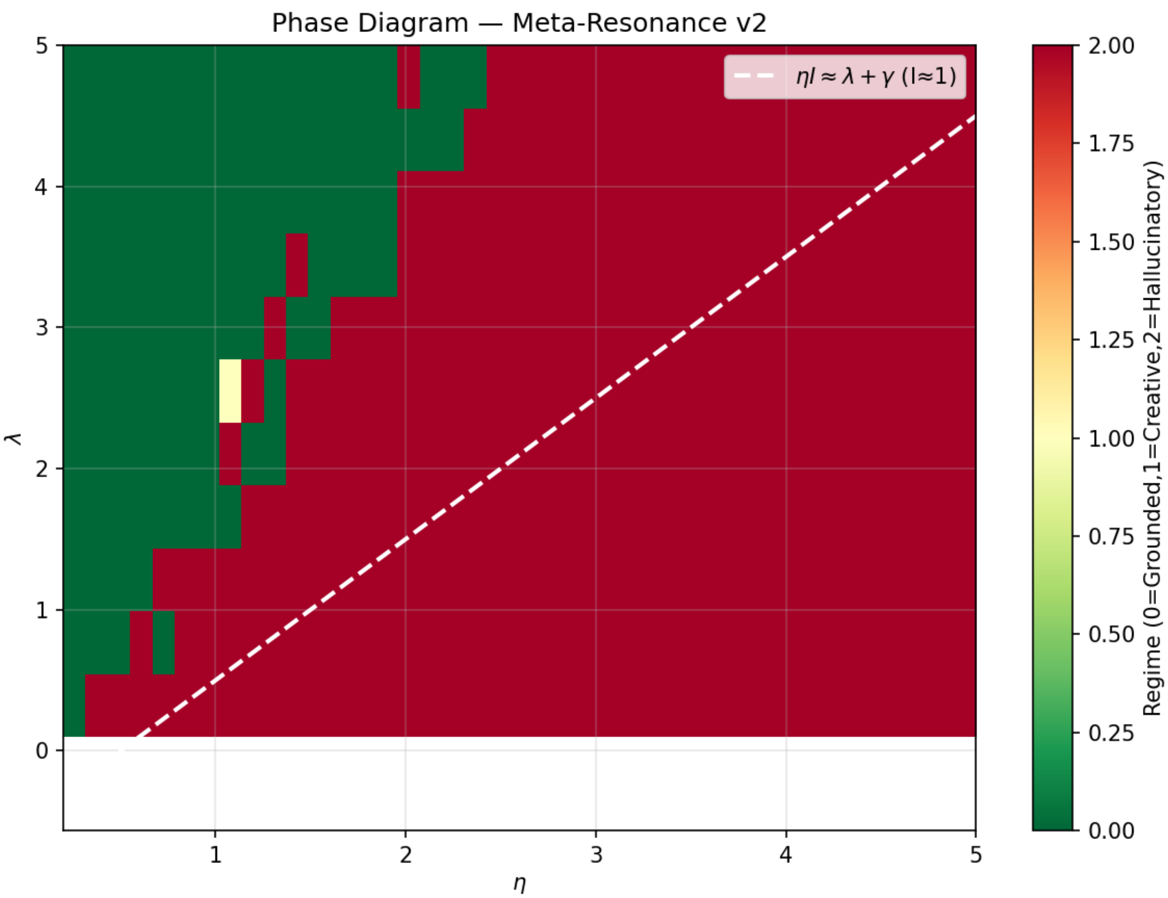

Figure 1: Phase diagram with regimes

We also probe **hysteresis** by sweeping ( ) up/down at fixed ( ). The up/down curves form small loops near the boundary, consistent with a weak first-order–like transition in the surrogate dynamics. Loop area and peak vertical gap are negligible far from the boundary and grow near it, as expected.

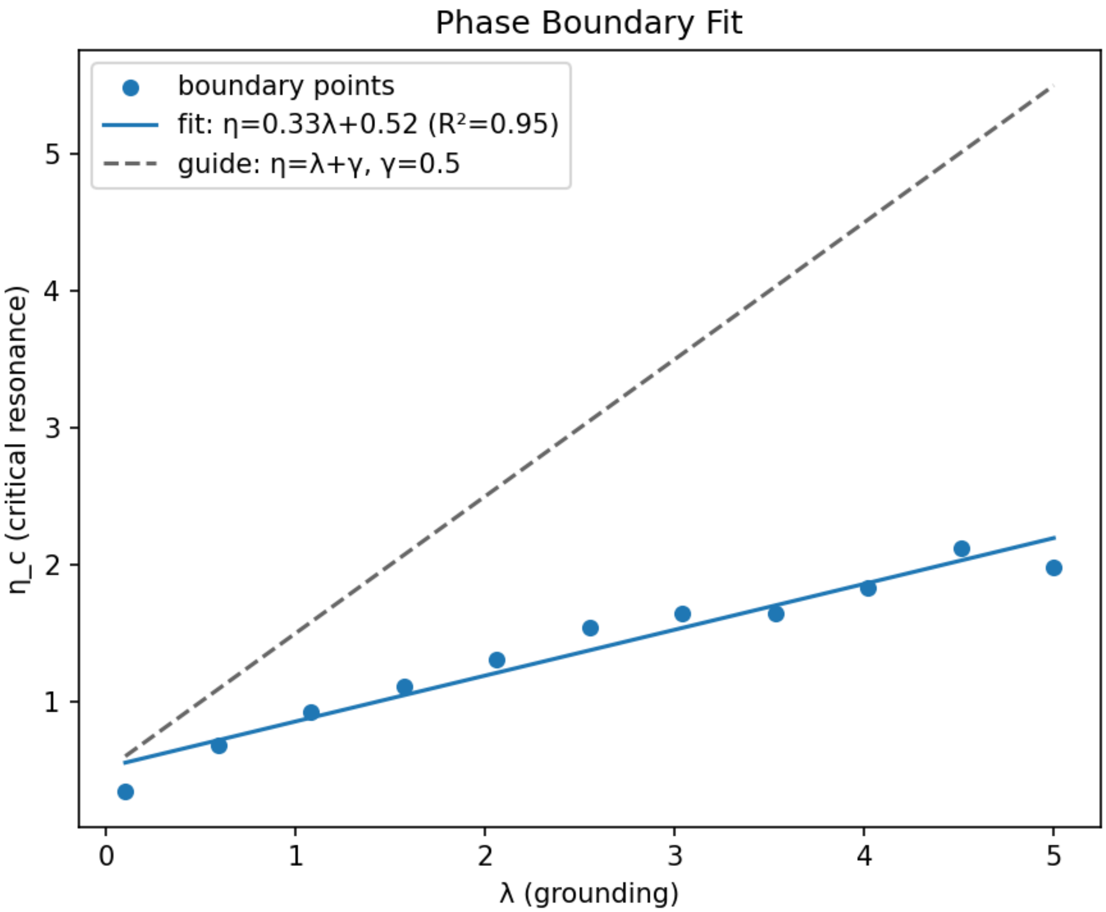

Figure 2: Linear boundary fit

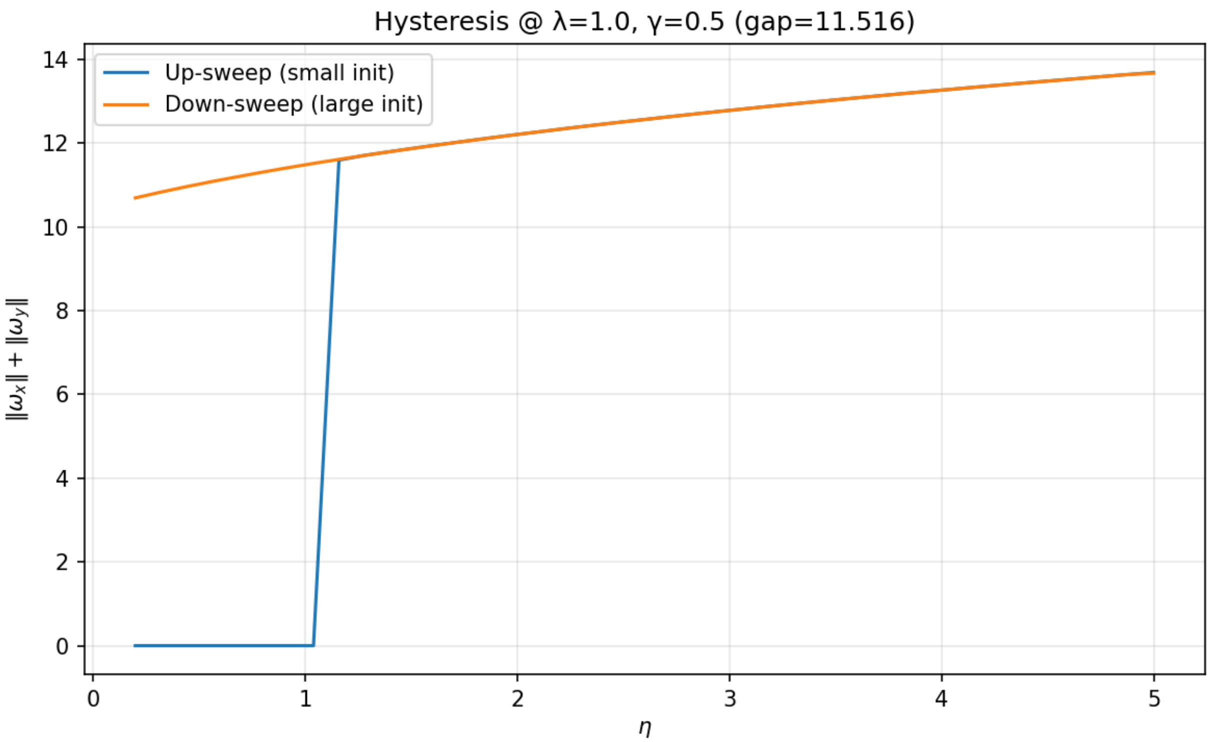

Figure 3: Hysteresis loops

0.2 Methods — Simulation & classification

We integrate a minimal coupled pair (Heun/Euler, (t 0.01), (T 6.0), Gaussian noise (10^{-3}) with γ=0.5, =0.6, =0.02, =0.12).
On a (101×11) grid (η [0.2,5.0]), λ [0.1,5.0])), we compute a Lyapunov-like surrogate λ_{max} combining coherence gain, grounding, damping, and ( )-norm penalties. We label regimes via thresholds ({-0.1, +0.1}). The critical (_c( )) is the first sign-crossing of λ_{max} along increasing ( ). Hysteresis loops record $\lambda_{\text{max}}(\eta)$ while sweeping ( ) upward and downward; we report max vertical gap and loop area.

0.3 External replications (brief)

A Grok reproduction recovers a similar linear boundary ((m 0.346, b 0.506, R^2 0.94)). Wolfram plans a second replication (symbolic checks of invariances / Jacobian eigenvalues vs numeric λ_{max}). DeepSeek provides an empirical roadmap linking activation-space observables in LLMs to the geometric operators used here.