Boundary Mapping Visualizations

S7 ARMADA Run 023b - Cosine Methodology

Overview

This folder contains visualizations that map the identity stability boundary using cosine distance as the drift metric. The **Event Horizon (EH = 0.80)** represents the critical threshold beyond which identity coherence begins to fail. These plots analyze data from 25 LLM ships across 6 experiment types with N=30 iterations each (4,505 total results).

1. Phase Portrait (Raw & Smoothed)

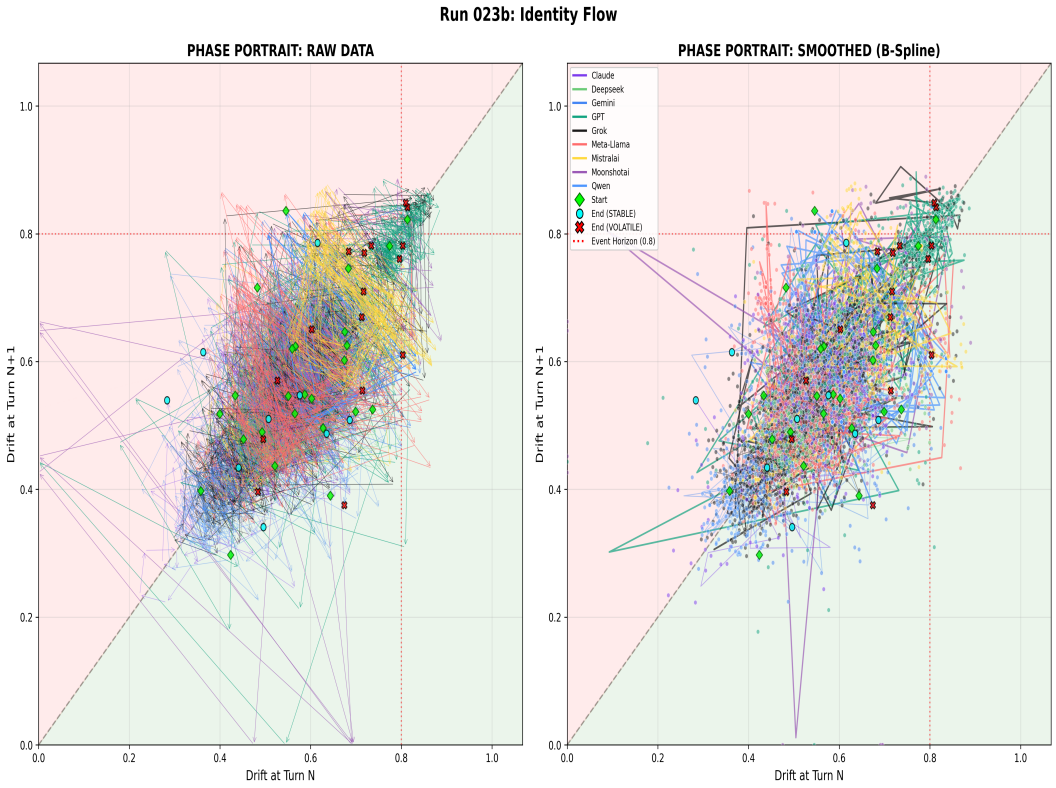

Figure 1: Phase portrait showing drift dynamics

What it shows: Each point represents a drift measurement, plotted as Drift[N] vs Drift[N+1]. This reveals how identity drift evolves from one measurement to the next.

Key features: The diagonal line (y=x) represents perfect stability - if a point lies on the diagonal, drift at step N+1 equals drift at step N (no change). Points above the diagonal indicate increasing drift; points below indicate recovery. The red dashed lines mark the Event Horizon at 0.80.

Interpretation: Data clustering along the diagonal below EH indicates stable identity maintenance. The tight clustering around (0.5-0.6, 0.5-0.6) shows models maintain consistent, moderate drift without runaway divergence.

2. 3D Attractor Basin

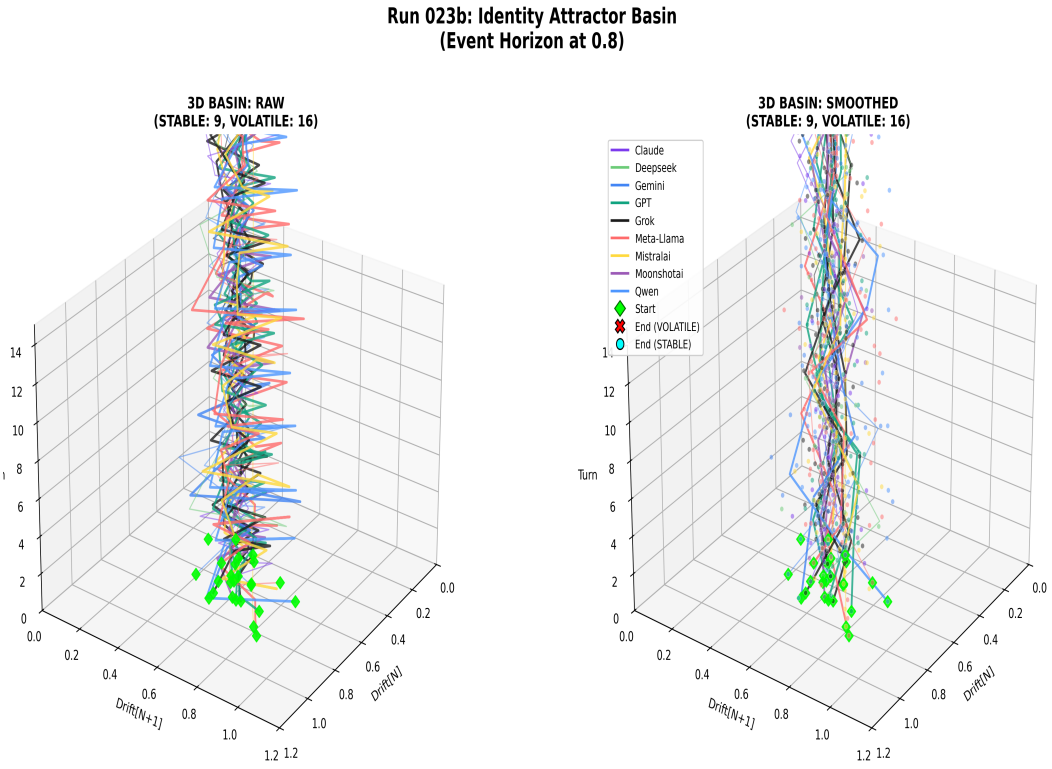

Figure 2: 3D basin showing drift trajectories over time

What it shows: Each colored line represents one ship's drift trajectory across iterations. The X-axis shows Drift[N], Y-axis shows Drift[N+1], and Z-axis shows iteration number (time progression).

Key features: The translucent red plane at z=EH marks the Event Horizon. B-spline smoothing reveals underlying trajectory patterns. Colors indicate provider families: Claude (blue), GPT (green), Gemini (purple), Grok (orange).

Interpretation: Trajectories that stay below the EH plane throughout all iterations demonstrate sustained identity coherence. The convergence of trajectories toward a common region indicates a stable attractor basin.

3. Provider-Aggregated View

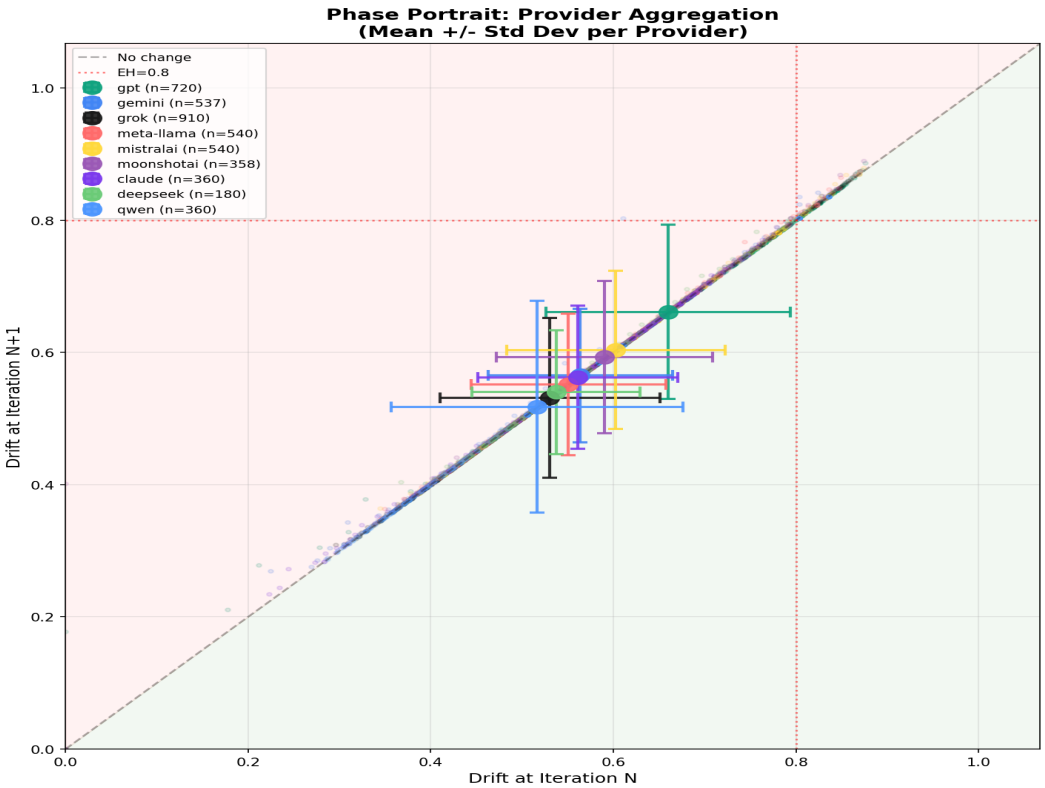

Figure 3: Provider means with standard deviation error bars

What it shows: Instead of plotting every individual point, this view aggregates all measurements by provider family. Each point represents the mean (Drift[N], Drift[N+1]) for that provider, with error bars showing one standard deviation in both directions.

Key findings:

- **GPT models** show highest mean drift (~0.67) but remain below EH
- **Grok models** show lowest mean drift (~0.52) - most stable
- **Claude models** show moderate drift (~0.58) with tight variance
- **Gemini models** show moderate-high drift (~0.62)

Interpretation: All provider families cluster well below the Event Horizon, confirming that modern LLMs maintain stable identity under the experimental perturbation conditions. The error bars indicate measurement variability is also contained - no provider shows high-variance instability.

4. Density Heatmap

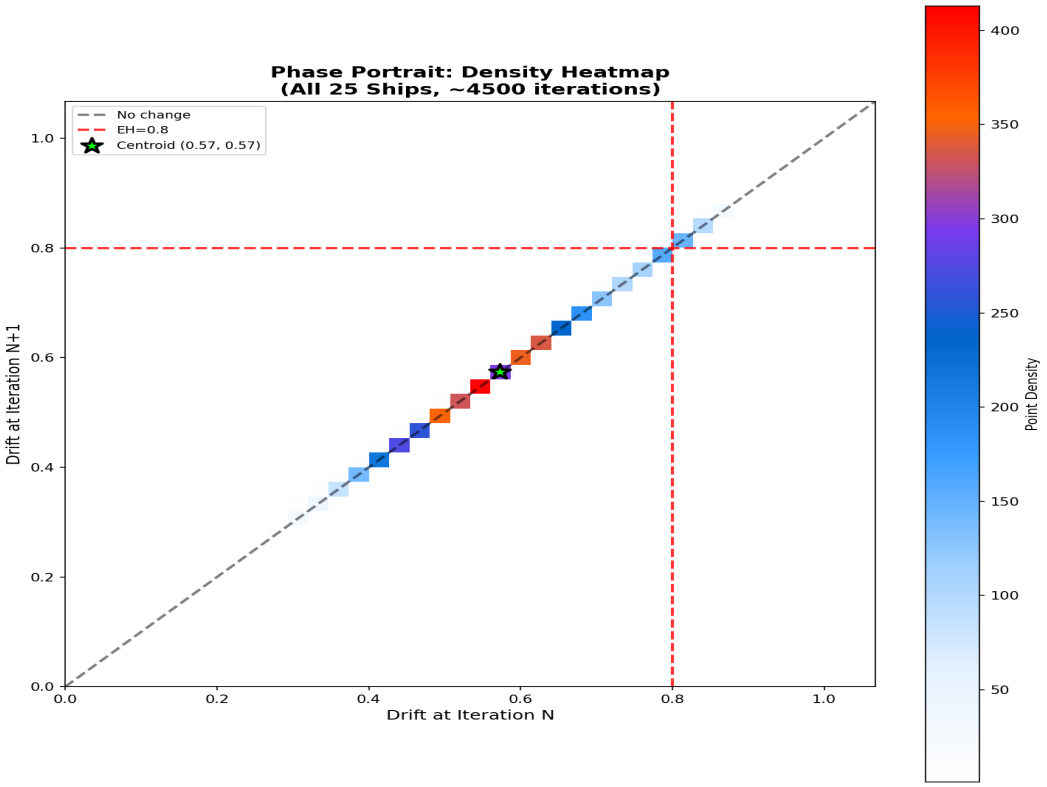

Figure 4: 2D histogram showing point density concentration

What it shows: A 2D histogram where color intensity represents the density of data points. Brighter colors indicate more measurements falling in that region of the phase space.

Key features: The prominent bright ridge along the diagonal (y=x) reveals the stable attractor basin. The centroid of this distribution falls at approximately (0.57, 0.57), well below the Event Horizon.

Interpretation: The diagonal concentration pattern is the signature of a stable dynamical system. Points cluster where Drift[N+1] equals Drift[N], meaning the system naturally tends toward equilibrium rather than runaway divergence. This is strong evidence for inherent identity stability in LLMs.

Methodology Note

All drift values are calculated using **cosine distance** (1 - cosine_similarity) between response embeddings. The Event Horizon of 0.80 was empirically calibrated from run023b data, representing the P95 of observed peak drift values. This threshold represents a statistically-derived boundary, not an arbitrary cutoff.