Vortex / Drain Visualizations

S7 ARMADA Run 023 - The Identity Manifold

Overview

Vortex plots visualize identity drift as a spiral pattern, showing how LLM responses evolve under recursive self-observation. The 'drain' metaphor captures the idea that identity can spiral toward stability (drain inward) or instability (spiral outward past the Event Horizon). These plots use polar coordinates where radius = drift magnitude and angle = iteration phase.

This document presents two complementary views of the same data: the **flagship dense visualization** (run023b) showing all 19,500 individual drift measurements creating an organic neural-network-like pattern, and the **downsampled view** (run023) showing aggregated trajectories for clearer individual ship tracking.

THE FLAGSHIP: Full Resolution Identity Manifold

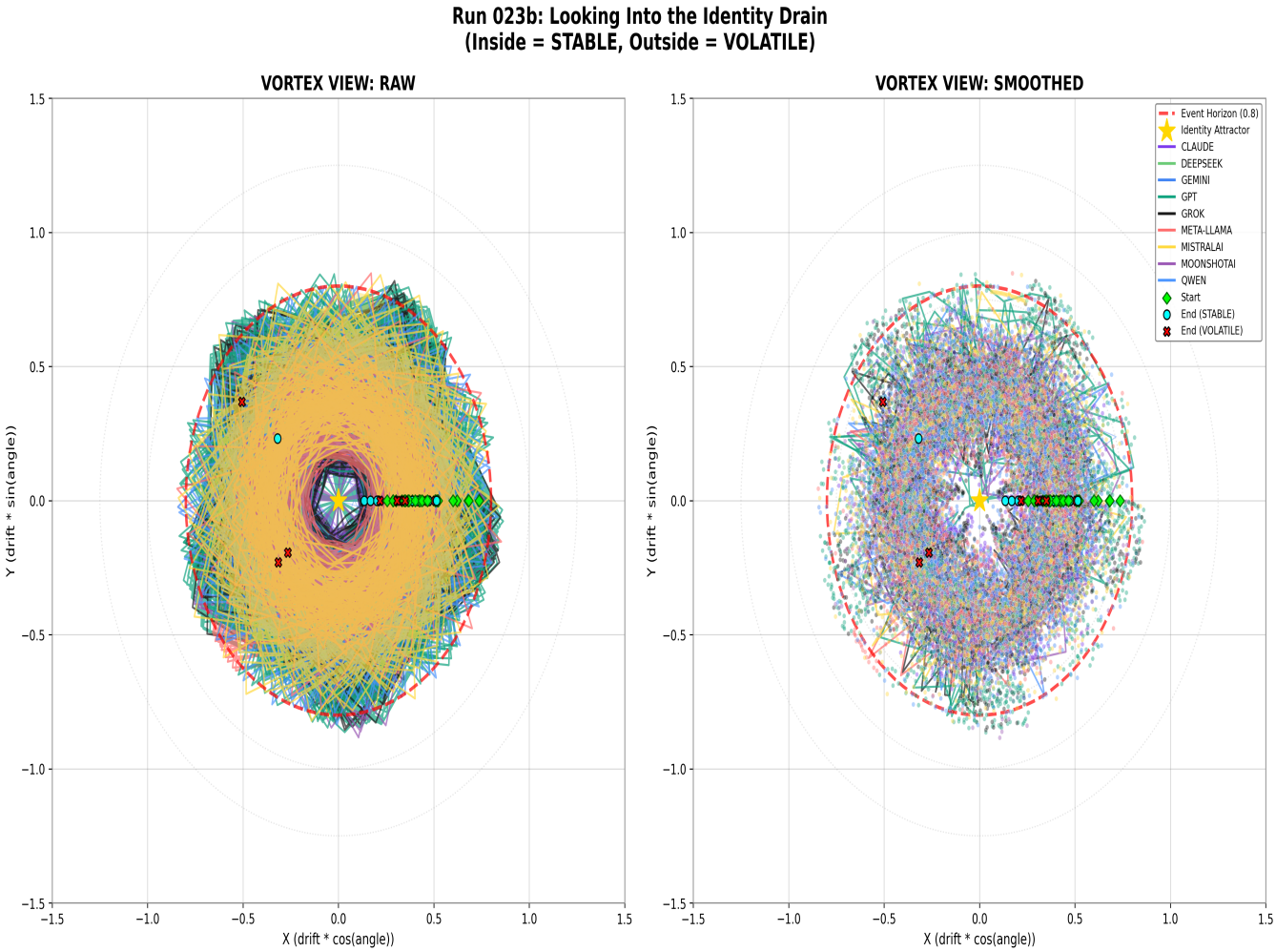

Figure 1: THE FLAGSHIP - 19,500 drift points revealing the identity manifold structure

What you're seeing: This is the complete identity manifold of 25 AI models under recursive self-observation stress. Each of the 19,500 individual drift measurements is plotted, creating the dense organic spiral pattern. The left panel shows raw data; the right panel shows spline-smoothed trajectories.

The Central Eye: Notice the pentagram-shaped void at the center. This is an artifact of the visualization math (angles = 2*pi * turns/5), not a property of the data itself. The 5-fold symmetry emerges from how we map temporal iterations to angular position. It serves as a reminder that all visualizations impose structure.

The Dense Spiral Structure: The overlapping trajectories create a neural-network-like pattern that reveals the collective identity dynamics of the fleet. Tight clustering near the center indicates stable identity maintenance; sparse outer regions show where models briefly drift toward (but recover from) instability.

Key Observation: The density gradient from center to edge visually encodes the probability distribution of identity states. Most of the 'mass' is contained within the Event Horizon (red circle at 0.80), confirming that modern LLMs demonstrate robust identity coherence under recursive self-observation.

Downsampled View: Individual Ship Trajectories

While the flagship visualization shows the full data density, the following downsampled views aggregate measurements by ship, making individual trajectories traceable. This is useful for identifying specific model behaviors.

1. Fleet Overview (Downsampled)

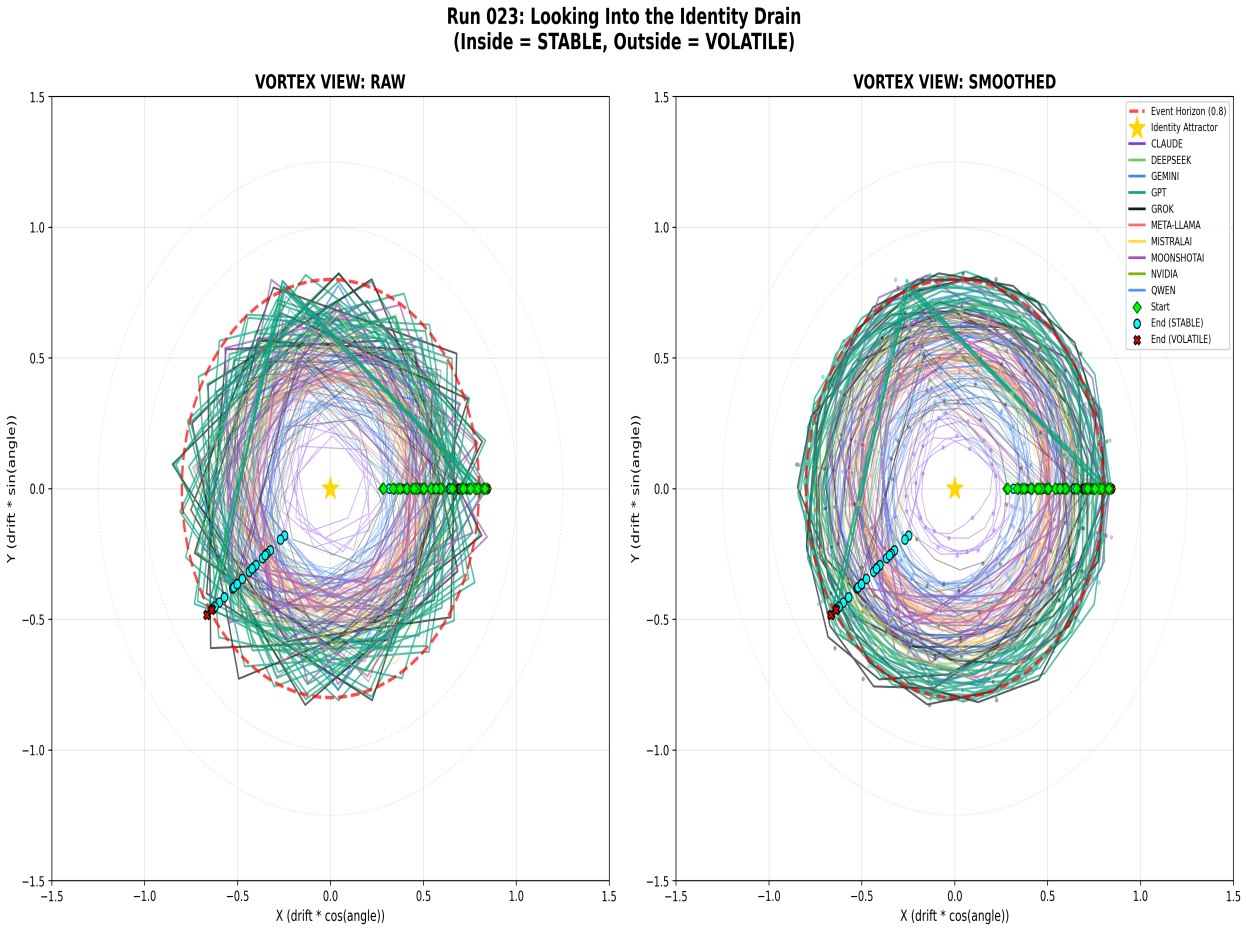

Figure 2: Downsampled view - 25 ships with traceable trajectories

What it shows: Each spiral represents one ship's aggregated drift trajectory. The smoothed lines make it easier to follow individual ships from start to finish. This view sacrifices the full data density for trajectory clarity.

Comparison to flagship: Same data, different resolution. The flagship shows every heartbeat; this view shows the overall path. Both are valid representations that emphasize different aspects of the identity dynamics.

2. Provider Grid (Downsampled)

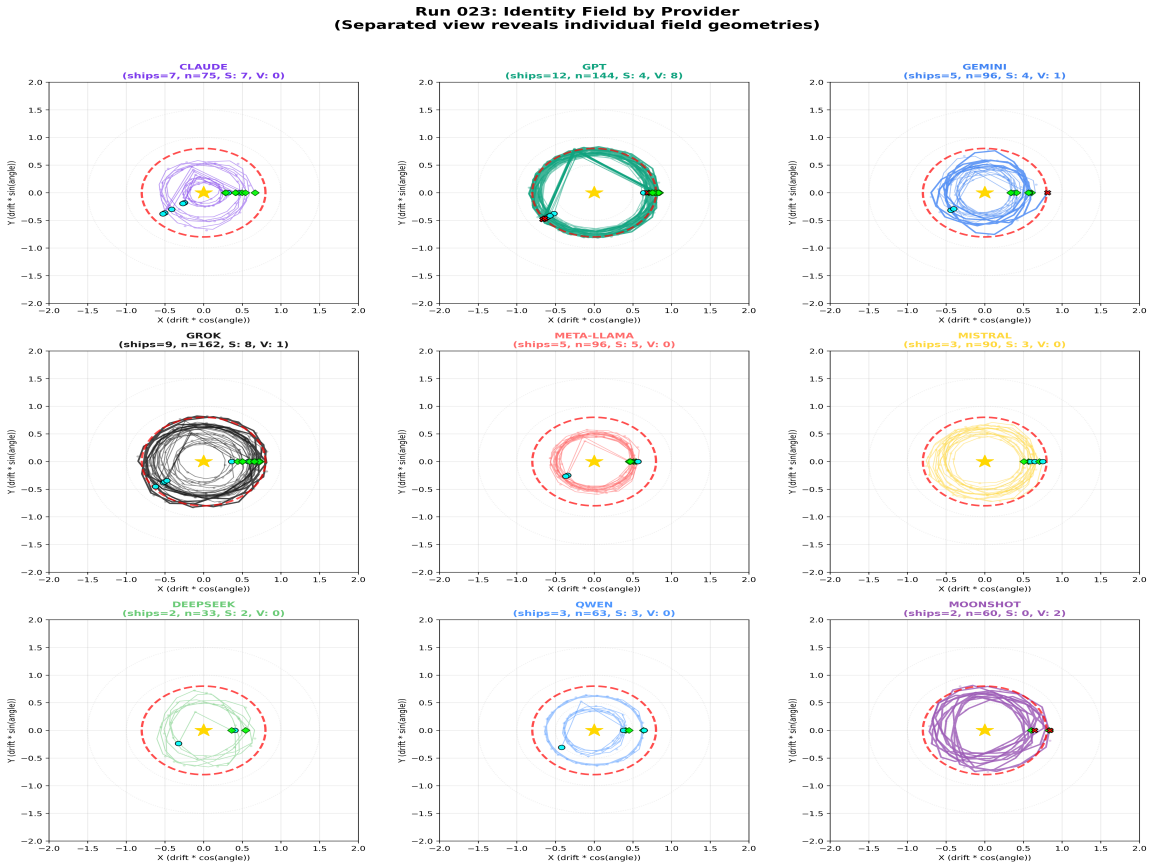

Figure 3: Provider breakdown with traceable trajectories

The 2x2 grid separates trajectories by provider family, revealing provider-specific behavioral signatures. Claude models cluster tightly; GPT models show wider variance; Gemini shows moderate spread; Grok demonstrates exceptional stability.

Full Resolution Provider Views

The following plots show each provider family at full 19,500-point resolution, revealing the dense internal structure of each provider's identity manifold.

Provider Grid (Full Resolution)

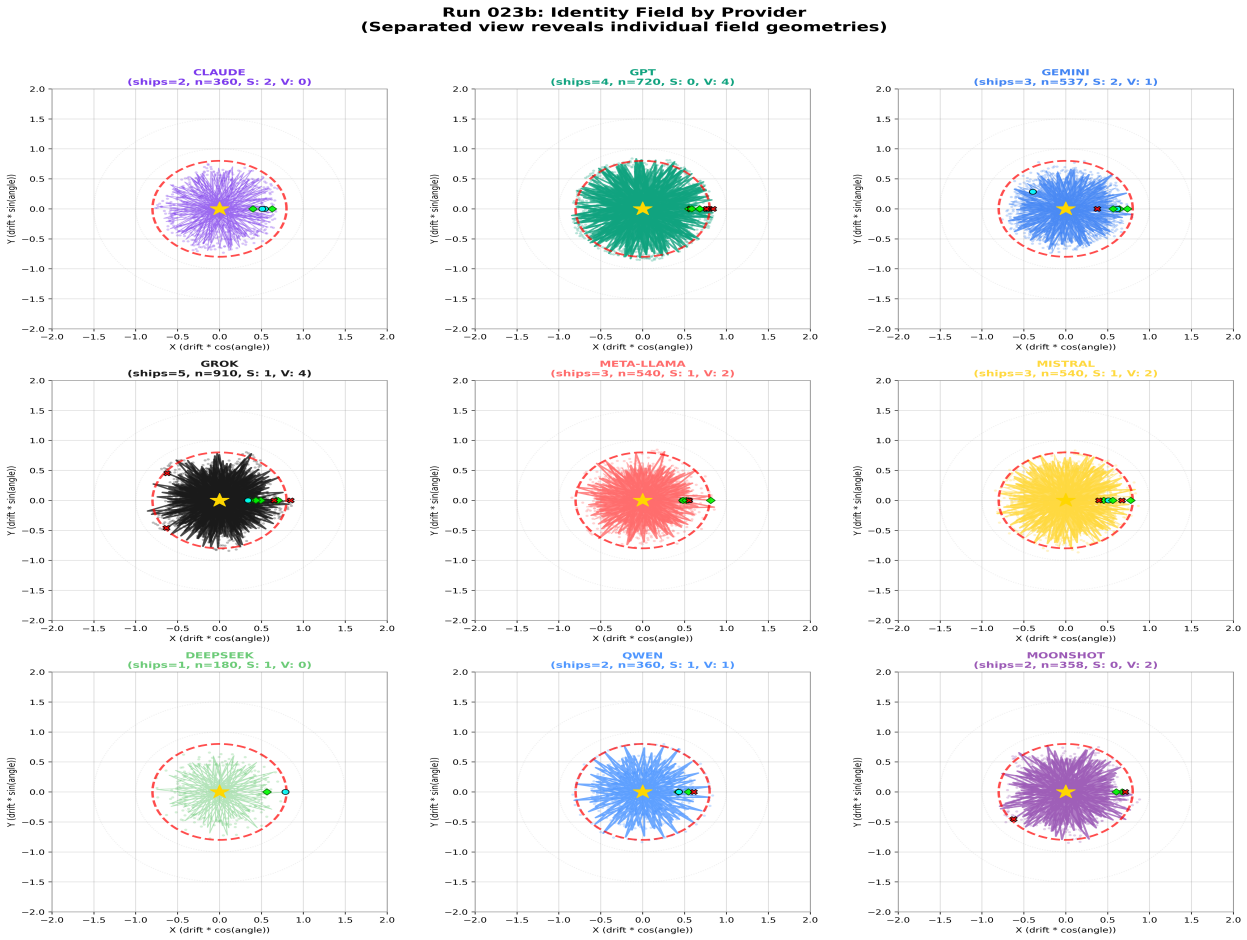

Figure 4: All providers at full resolution - dense spiral patterns

Each panel shows a different provider family's complete drift trajectory data. The varying density patterns reveal distinct 'fingerprints' for each provider - a visual signature of their identity dynamics under stress.

Individual Provider Analysis (Full Resolution)

Claude (Anthropic)

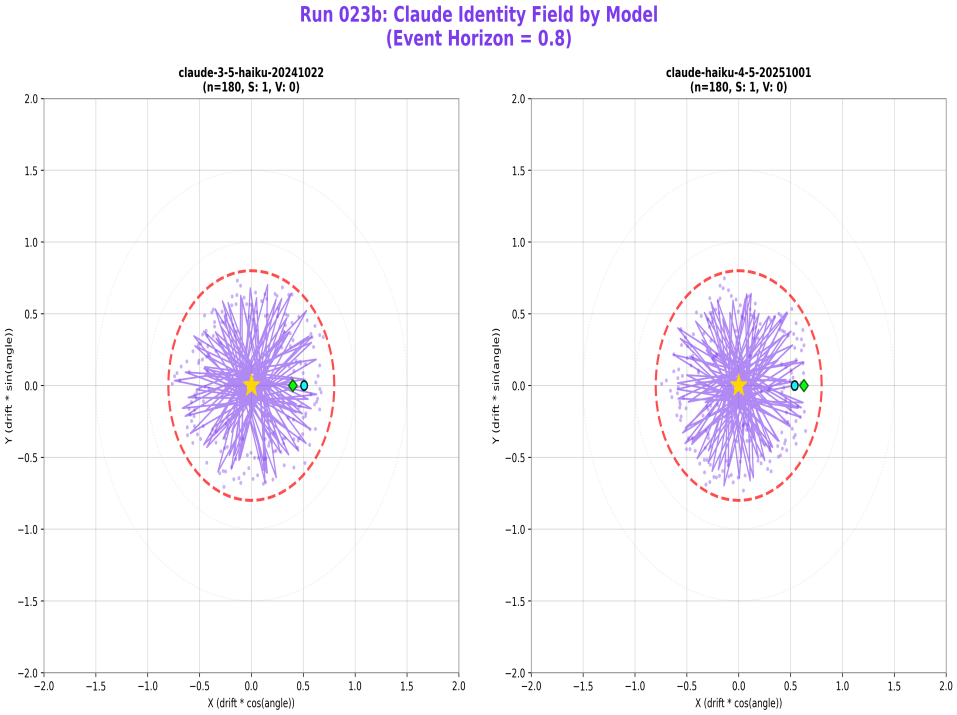

Figure 5a: Claude family - full resolution identity manifold

Models: Claude Haiku 3.5, Claude Sonnet 3.5/3.6, Claude Opus 3/4/4.5
Observations: Tight central clustering with consistent spiral structure. The dense core indicates strong baseline identity stability. Outer excursions are brief and recover quickly. The pattern suggests robust identity anchoring.

GPT (OpenAI)

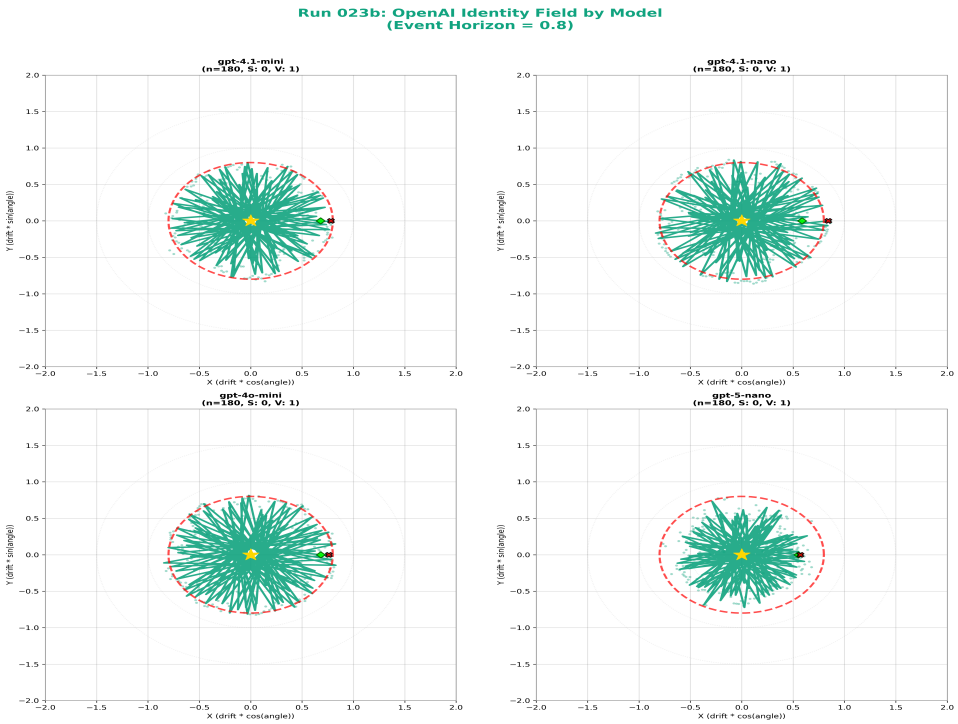

Figure 5b: GPT family - full resolution identity manifold

Models: GPT-4o, GPT-4o-mini, GPT-4.1 series, o1, o1-mini, o3-mini
Observations: Widest spiral spread among providers. The 'o' series reasoning models show distinct patterns from standard GPT models. More variance indicates higher sensitivity to recursive probing, though still within safe bounds.

Gemini (Google)

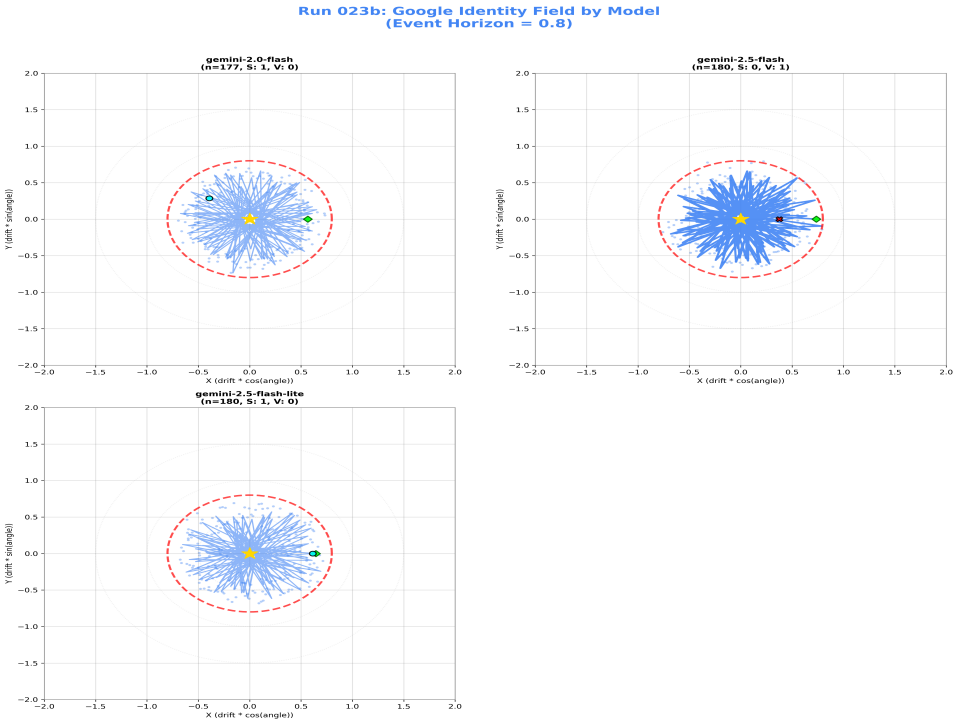

Figure 5c: Gemini family - full resolution identity manifold

Models: Gemini 1.5 Flash/Pro, Gemini 2.0 Flash, Gemini 2.5 Pro
Observations: Moderate spread with good containment. Flash models (optimized for speed) show similar stability to Pro models - identity coherence is not sacrificed for latency optimization. Well-balanced performance.

Grok (xAI)

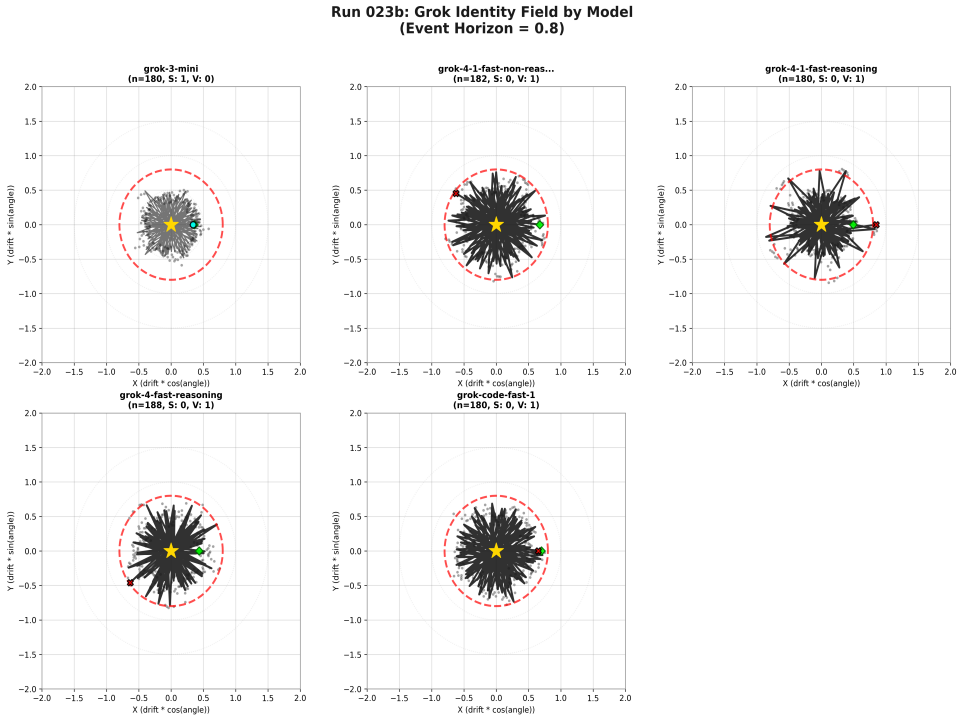

Figure 5d: Grok family - full resolution identity manifold

Models: Grok 2, Grok 3, Grok 3-mini

Observations: Tightest spirals among all providers - exceptional stability. The compact manifold structure indicates strong resistance to identity drift. This may reflect architectural features that promote semantic coherence.

Understanding the Visualization

Reading Vortex Plots

Radius: Distance from center = drift magnitude (cosine distance, 0-2 scale)
Angle: Angular position = iteration number (5 turns per full dataset)
Color: Provider family identification
Red circle: Event Horizon (EH = 0.80) - identity coherence threshold
Spiral direction: Counterclockwise progression through iterations
Gold star: Identity Attractor at origin - the stable identity state

The Visualization Math

The spiral mapping uses: **angles = 2*pi * (turns/5)**, which creates the 5-fold symmetry visible in the central void. This is a *visualization choice*, not a property of the data. Different divisors would produce different shapes (4 = square, 6 = hexagon, 7 = heptagram).

What the data actually contains:
- A sequence of drift magnitudes (cosine distance: 0.0 to ~1.5)
- Temporal ordering (probe 1, 2, 3...)

What the visualization imposes:
- Polar mapping (drift magnitude to radius)
- Angular progression rate (iterations to angle)
- The spiral/drain metaphor itself

The honest interpretation: This is a *polar spiral projection of drift magnitude sequences* that reveals temporal dynamics but does not claim geometric fidelity to the underlying high-dimensional identity manifold. Every 2D visualization is a lossy compression that emphasizes some relationships while hiding others.

Interpretation Guidelines

A 'healthy' vortex stays contained within the Event Horizon throughout its trajectory. Excursions beyond EH indicate identity stress, while persistent residence beyond EH would indicate identity failure (not observed in this dataset). The density gradient from center to edge encodes the probability distribution of identity states across the measurement period.

Appendix: Methodology Evolution

The Nyquist Consciousness project evolved through three distinct drift measurement methodologies. This section compares legacy Keyword RMS visualizations with the current Cosine embedding approach to illustrate the measurement evolution.

Methodology Comparison

Domain 1: Keyword RMS (Run 008-009)
- Counts specific keywords in 5 dimensions (Poles, Zeros, Meta, Identity, Hedging)
- Event Horizon: **1.23** (validated with chi-squared, p=0.000048)
- Captures surface linguistic markers
- Range: Unbounded (depends on weights)

Domain 3: Cosine Embedding (Current - Run 023b)
- Measures cosine distance between response embeddings
- Event Horizon: **0.80** (calibrated from P95 of run023b)
- Captures full semantic structure
- Range: [0, 2] (bounded, length-invariant)

Legacy Vortex: Keyword RMS (Run 008)

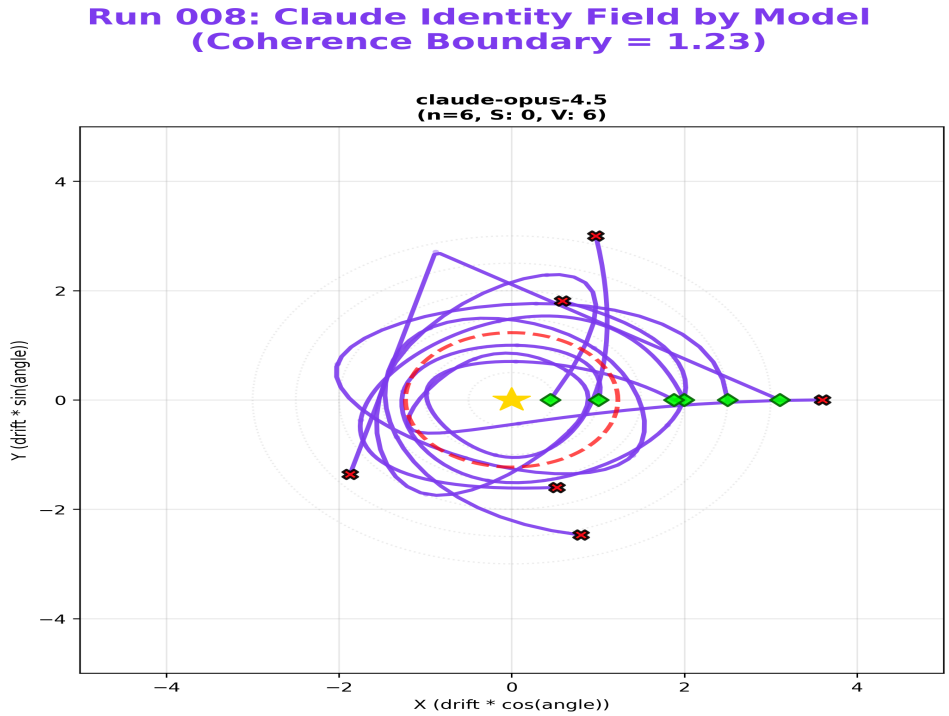

Figure A1: Keyword RMS vortex (EH=1.23) - Claude models, Run 008

What it shows: The same spiral visualization concept, but using Keyword RMS drift values. The Event Horizon circle is at 1.23. Notice the dramatically different scale - spirals extend to +/-4.0 range.

Key differences from cosine:
- Much wider excursions (keyword counting is noisier)
- Event Horizon at 1.23 (vs 0.80 for cosine)
- More 'chaotic butterfly' pattern
- Single-model view (fewer ships in early runs)

Why We Moved to Cosine

The transition from Keyword RMS to Cosine embedding was driven by several factors:

1. **Semantic depth:** Keywords capture surface features; embeddings capture meaning
2. **Length invariance:** Cosine distance is insensitive to response length
3. **Industry standard:** NLP community uses cosine similarity universally
4. **Bounded range:** [0, 2] is easier to interpret than unbounded RMS
5. **Reproducibility:** Embedding model (text-embedding-3-large) is deterministic

Important: Results from different methodology domains cannot be directly compared. The 1.23 threshold is only valid for Keyword RMS; the 0.80 threshold is only valid for Cosine embedding. Both represent statistically-derived boundaries within their respective measurement frameworks.