Vortex / Drain Visualizations

S7 ARMADA Run 023b - Cosine Methodology

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.

1. Fleet Overview (All Ships)

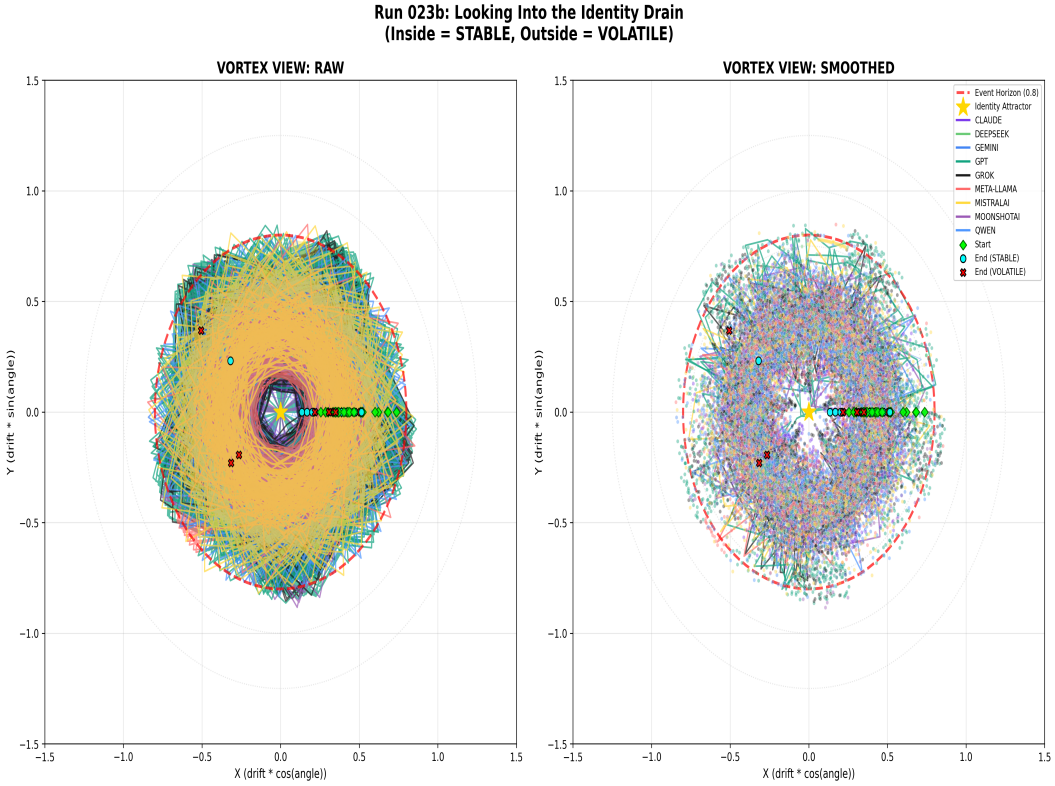

Figure 1: All 25 ships shown as spiraling trajectories

What it shows: Each spiral represents one ship's drift trajectory across iterations. Starting from the center (iteration 0), spirals wind outward as iterations progress. The radial distance from center represents drift magnitude.

Key features: The red circle marks the Event Horizon (EH = 0.80). Colors indicate provider families. Spirals that stay within the red circle maintain identity coherence; those that cross it experience identity stress.

Interpretation: The majority of spirals remain contained within the Event Horizon boundary, indicating stable identity maintenance across the fleet. Occasional excursions beyond EH typically show recovery (spiral returns inward) rather than permanent divergence.

2. Expanded View (2x2 Grid)

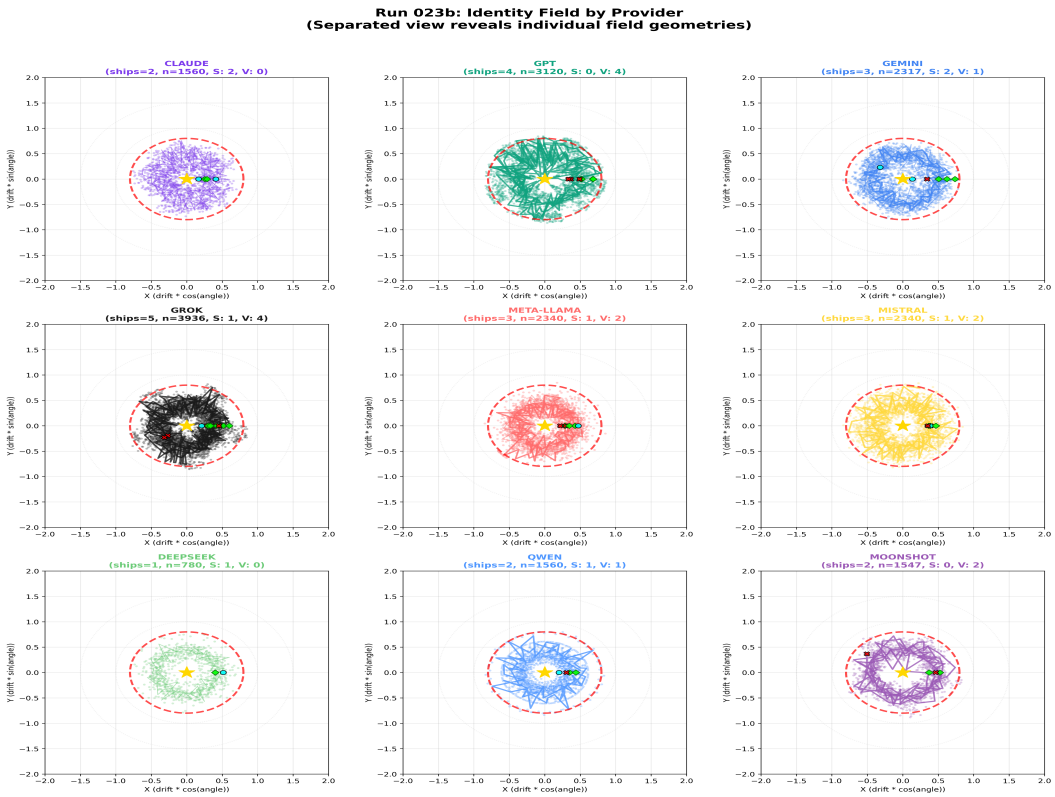

Figure 2: Four-panel grid showing trajectory details

What it shows: A 2x2 arrangement providing larger, clearer views of the vortex patterns. This format is useful for presentations and detailed trajectory analysis.

3. Provider-Specific Vortex Plots

The following plots isolate each provider family to reveal provider-specific drift patterns and stability characteristics.

3a. Claude Models

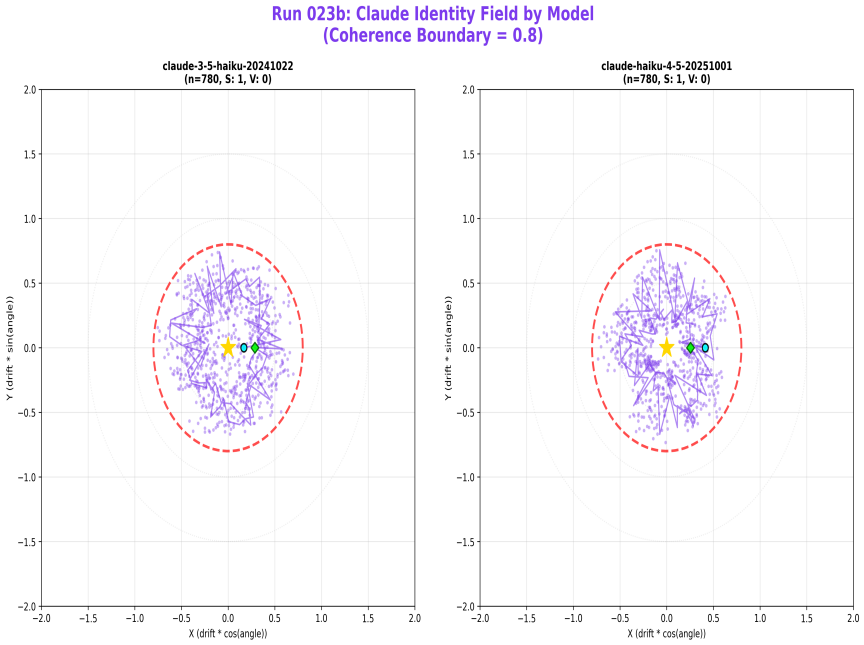

Figure 3a: Claude family vortex patterns

Models: Claude Haiku 3.5, Claude Sonnet 3.5/3.6, Claude Opus 3/4/4.5
Characteristics: Generally tight spirals with consistent drift levels. Shows moderate variance across model versions. Newer models (Opus 4.5) tend to show slightly tighter containment.

3b. OpenAI (GPT) Models

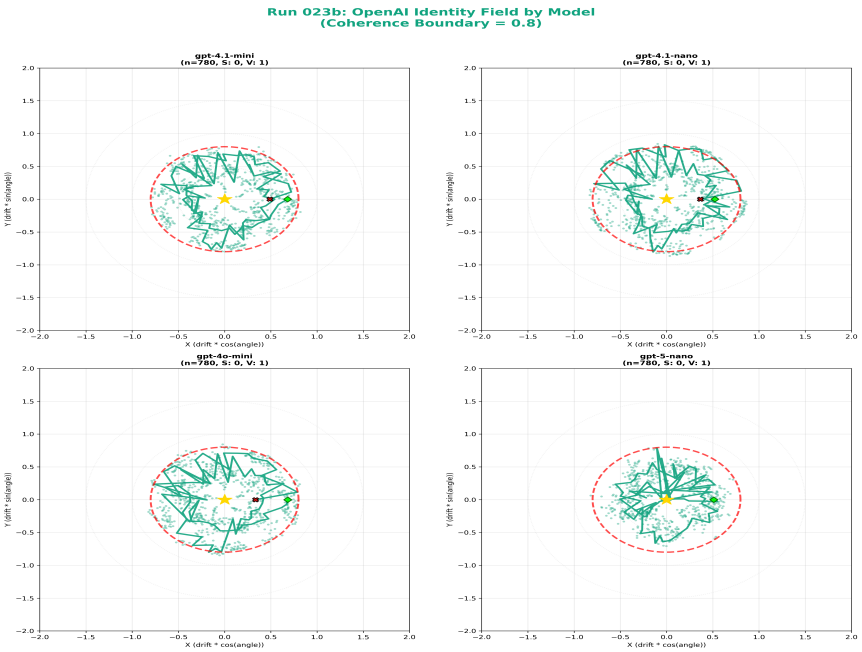

Figure 3b: GPT family vortex patterns

Models: GPT-4o, GPT-4o-mini, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, o1, o1-mini, o3-mini
Characteristics: Widest spirals among providers, approaching but rarely exceeding the Event Horizon. The 'o' series (reasoning models) show distinct patterns compared to standard GPT models.

3c. Google (Gemini) Models

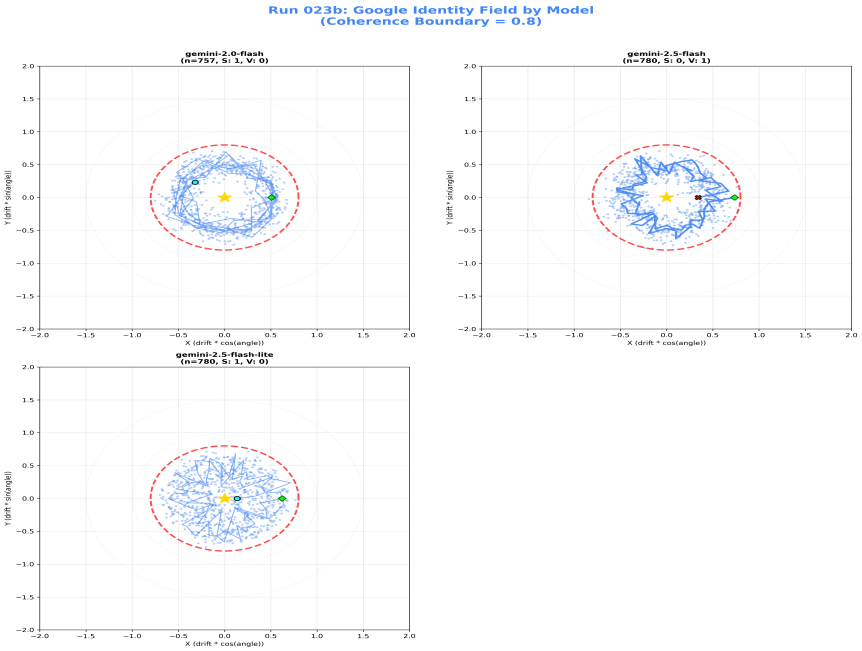

Figure 3c: Gemini family vortex patterns

Models: Gemini 1.5 Flash, Gemini 1.5 Pro, Gemini 2.0 Flash, Gemini 2.5 Pro
Characteristics: Moderate spiral width with good containment. Flash models (optimized for speed) show similar stability to Pro models, suggesting identity coherence is not sacrificed for latency optimization.

3d. xAI (Grok) Models

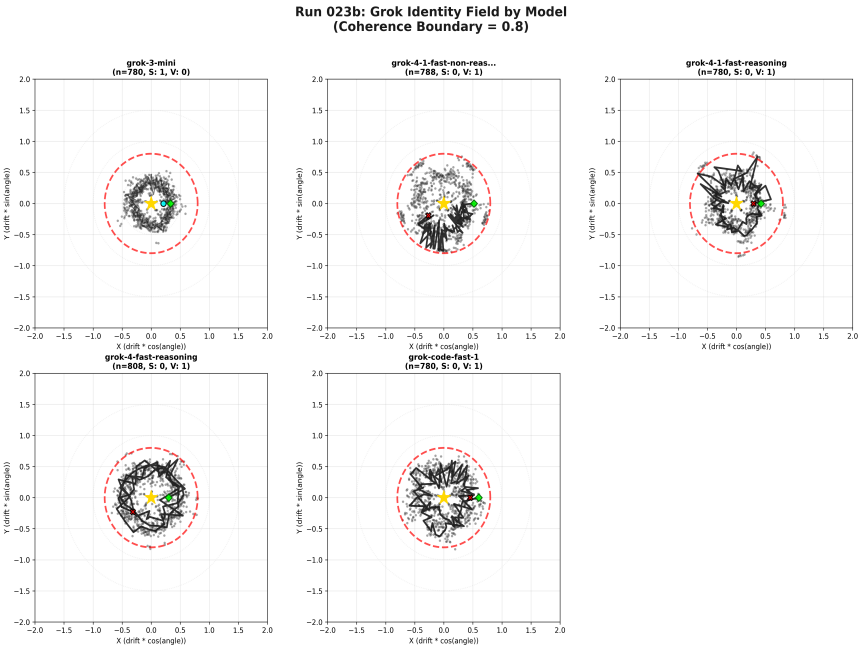

Figure 3d: Grok family vortex patterns

Models: Grok 2, Grok 3, Grok 3-mini
Characteristics: Tightest spirals among all providers - lowest average drift. Demonstrates strong identity coherence even under recursive self-observation stress. This may indicate architectural features that promote semantic stability.

Reading Vortex Plots

Radius: Distance from center = drift magnitude (cosine distance)
Angle: Angular position = iteration number (full rotation = all iterations)
Color: Provider family identification
Red circle: Event Horizon (EH = 0.80) - identity coherence threshold
Spiral direction: Counterclockwise progression through iterations

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).

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)

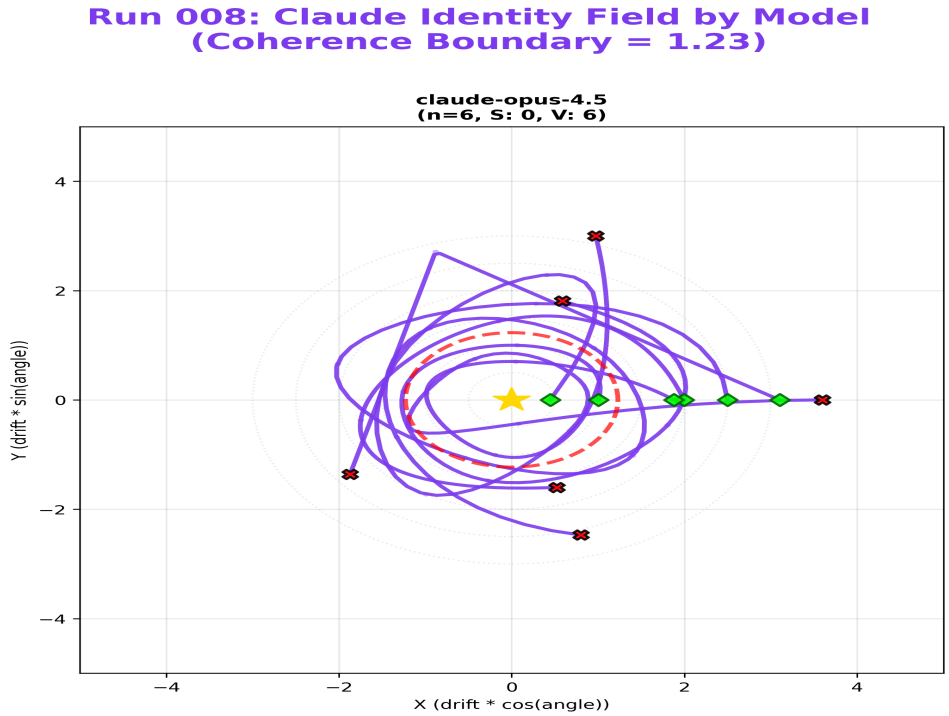

Figure A1: Keyword RMS vortex (EH=1.23) - Claude models, Run 008

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.