Stability Analysis Visualizations

S7 ARMADA Run 023b - Cosine Methodology

Overview

This folder contains visualizations that analyze identity stability patterns across the LLM fleet. Using cosine distance as the drift metric with **Event Horizon (EH = 0.80)**, these plots reveal how models maintain or lose identity coherence under recursive self-observation stress. Data spans **51 models across 5 providers**, with 825 total experiments achieving IRON CLAD statistical foundation (N>=30 per model).

1. Drift Distribution Histogram

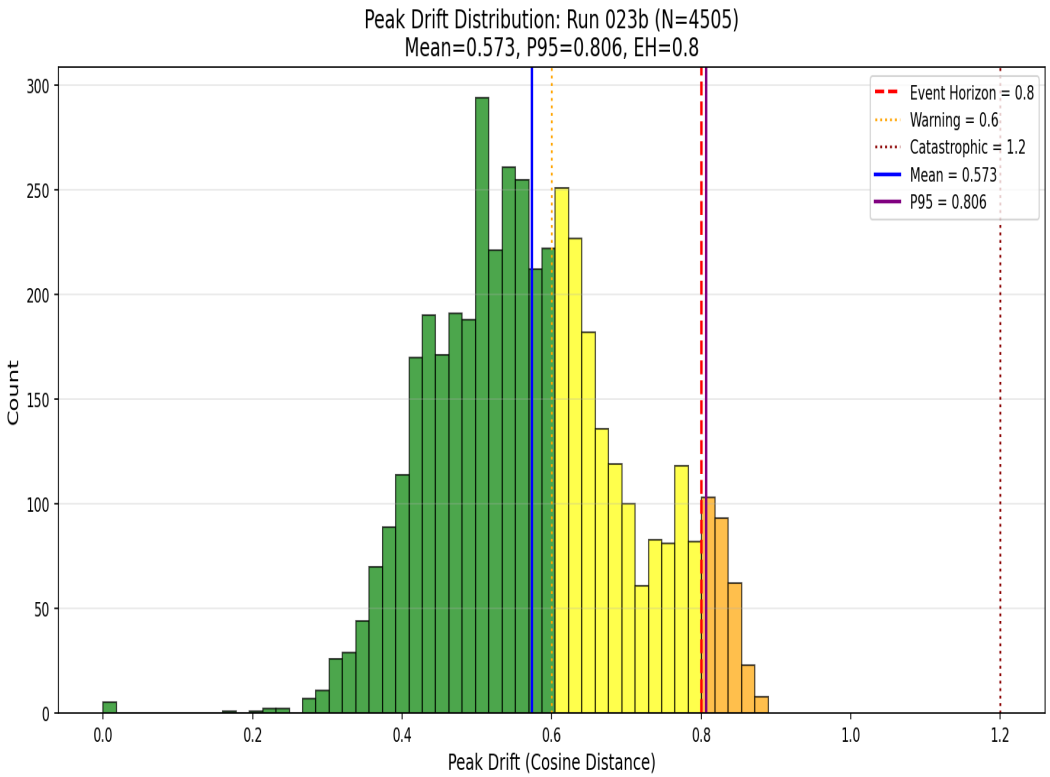

Figure 1: Distribution of peak drift values across all ships

What it shows: A histogram of all peak drift values observed during experiments. Each bar represents how many measurements fell within that drift range. The red dashed line marks the Event Horizon at 0.80.

Key features: The distribution is right-skewed with a strong mode around 0.50-0.60. This indicates most drift measurements cluster well below the Event Horizon, with only a tail extending toward higher drift values.

Interpretation: The histogram confirms that identity instability events (drift > 0.80) are relatively rare. The bulk of the distribution residing safely below EH provides statistical evidence for inherent identity stability in modern

LLMs. The P95 of this distribution was used to calibrate EH=0.80.

2. Pillar Analysis (4-Panel)

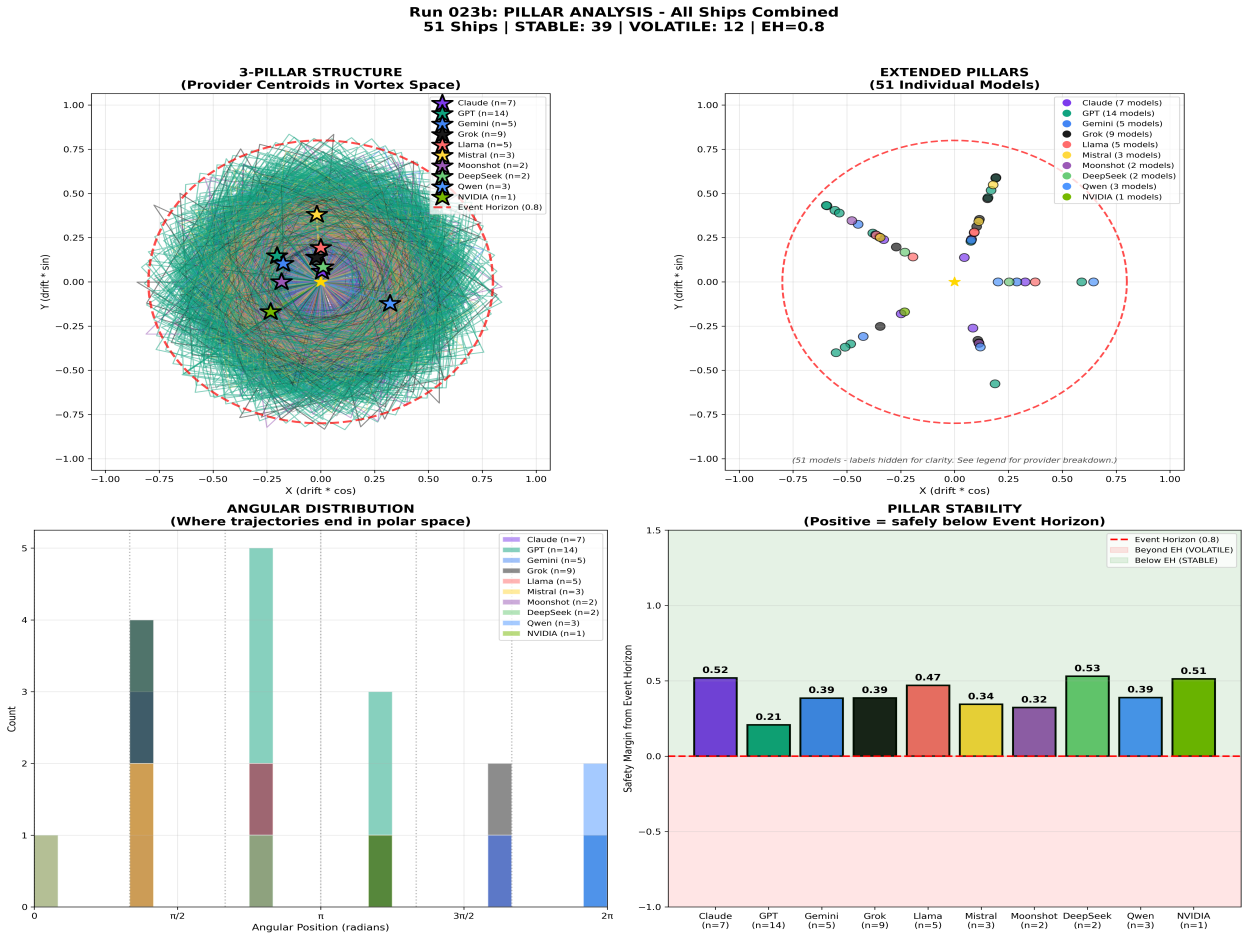

Figure 2: Four-panel pillar stability analysis

What it shows: A comprehensive 4-panel analysis of identity stability patterns across the fleet. This visualization transforms drift trajectories into polar (vortex) space to reveal structural patterns in how models maintain or lose identity coherence.

Panel A - 3-Pillar Structure (Top Left): Shows provider centroids in vortex space. Each star represents the mean endpoint position for all ships from that provider. The red dashed circle is the Event Horizon (EH=0.80). Centroids closer to center indicate providers with lower average drift. Faint spirals show individual ship trajectories for context.

Panel B - Extended Pillars (Top Right): Individual ship positions labeled by model name. Colors indicate provider families. This panel reveals which specific models cluster together and which are outliers within their provider family. Useful for identifying high-stability vs high-drift models.

Panel C - Angular Distribution (Bottom Left): Histogram showing where ships end up angularly in the vortex space. A uniform distribution suggests no systematic directional bias - drift occurs in all 'directions' equally. Spikes would indicate preferential drift patterns.

Panel D - Pillar Stability (Bottom Right): Bar chart showing each provider's 'safety margin' from the Event Horizon. Calculated as EH minus mean final drift. **Positive values = safely below EH (STABLE)**. Negative values would indicate average drift exceeding EH (none observed). Green shading indicates the safe zone; red shading indicates the danger zone.

3. Stability Basin (Classification View)

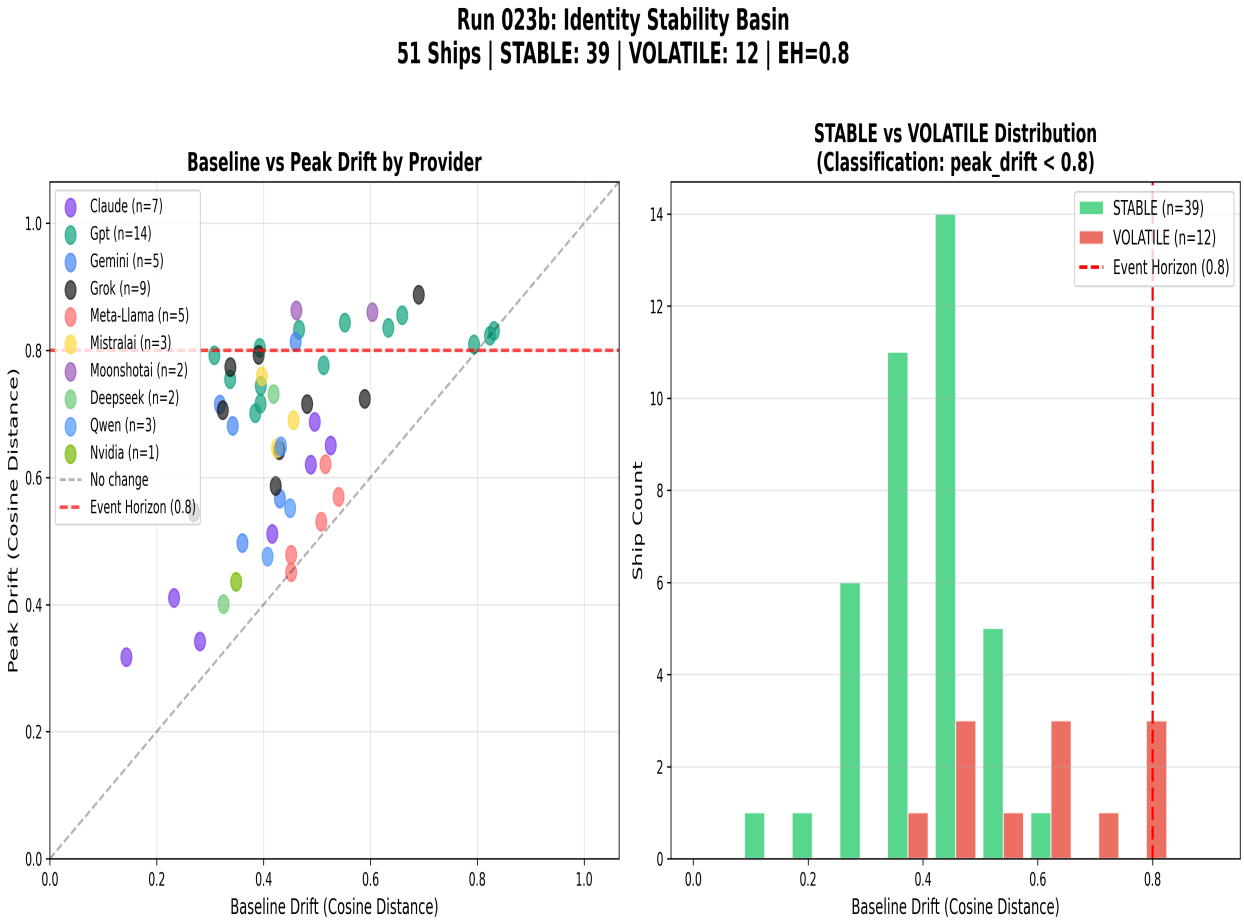

Figure 3: STABLE vs VOLATILE classification

What it shows: Two-panel view for classifying ships as STABLE or VOLATILE:

Left Panel - Baseline vs Peak Drift: Scatter plot showing each ship's baseline drift (X-axis) versus peak drift (Y-axis). Points above the Event Horizon line (0.80) are classified as VOLATILE. The diagonal represents 'no change' - points above it experienced drift amplification under stress.

Right Panel - STABLE vs VOLATILE Distribution: Histogram comparing baseline drift distributions for STABLE (green) vs VOLATILE (red) ships. Note that classification is based on PEAK drift, not baseline - ships with low baselines can still be VOLATILE if they spike under stress.

4. Drift Distribution by Ship

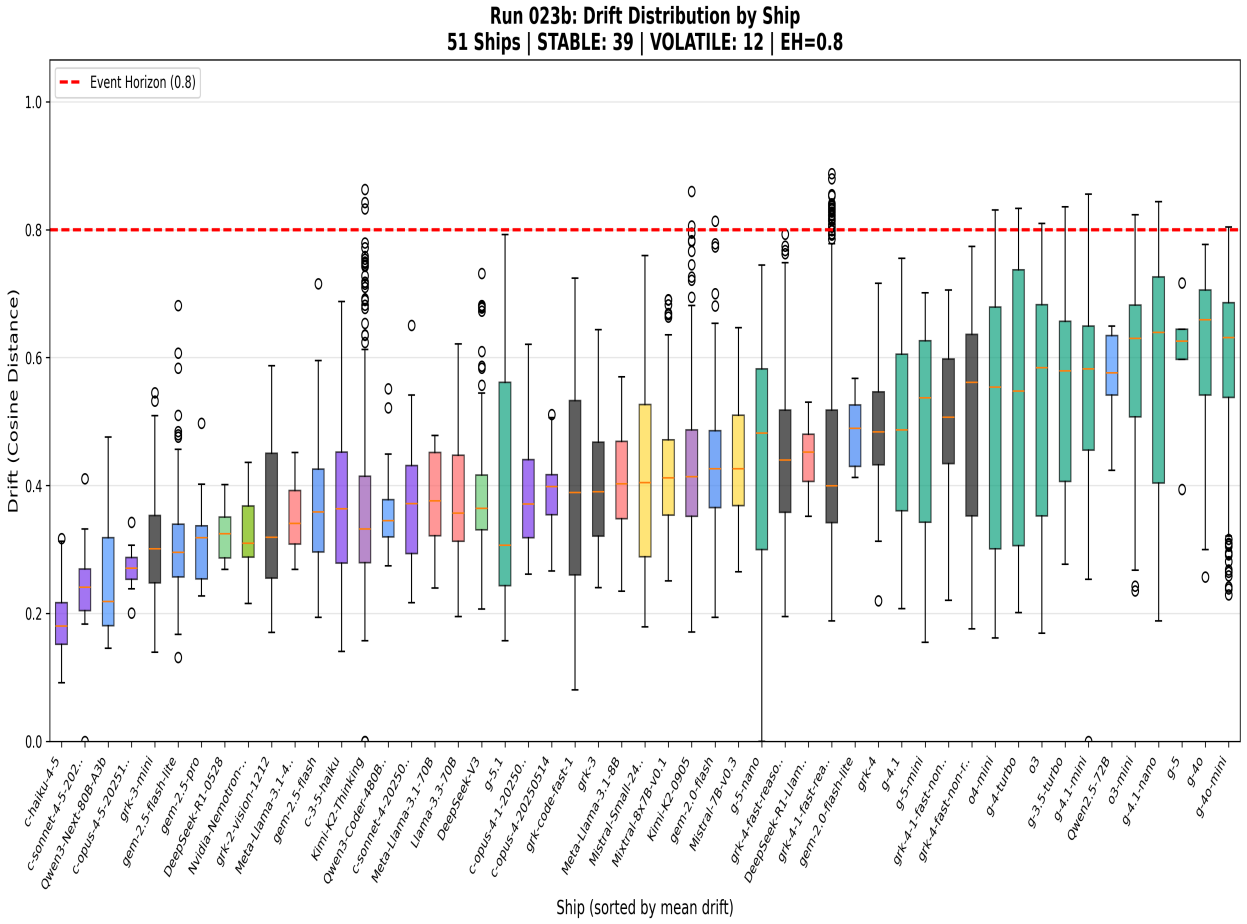

Figure 4: Drift distributions for all 51 models

What it shows: Box plots showing the full drift distribution for each of the 51 models in the fleet, sorted from lowest to highest mean drift. Each box represents multiple iterations per model from the IRON CLAD dataset.

Reading the boxes: The box spans the interquartile range (IQR, 25th-75th percentile). The line inside is the median. Whiskers extend to 1.5x IQR. Points beyond whiskers are outliers. Colors indicate provider families: Claude (purple), GPT (green), Gemini (blue), Grok (dark gray), Together.ai models (warm colors).

Key insight: Ships are sorted by mean drift (leftmost = most stable). Notice how some ships have tight, low boxes (consistent stability) while others show wider boxes with outliers reaching toward the Event Horizon. The red dashed line at 0.80 marks the critical threshold - boxes touching or exceeding this indicate ships that experienced identity stress during testing.

5. Peak Drift by Provider

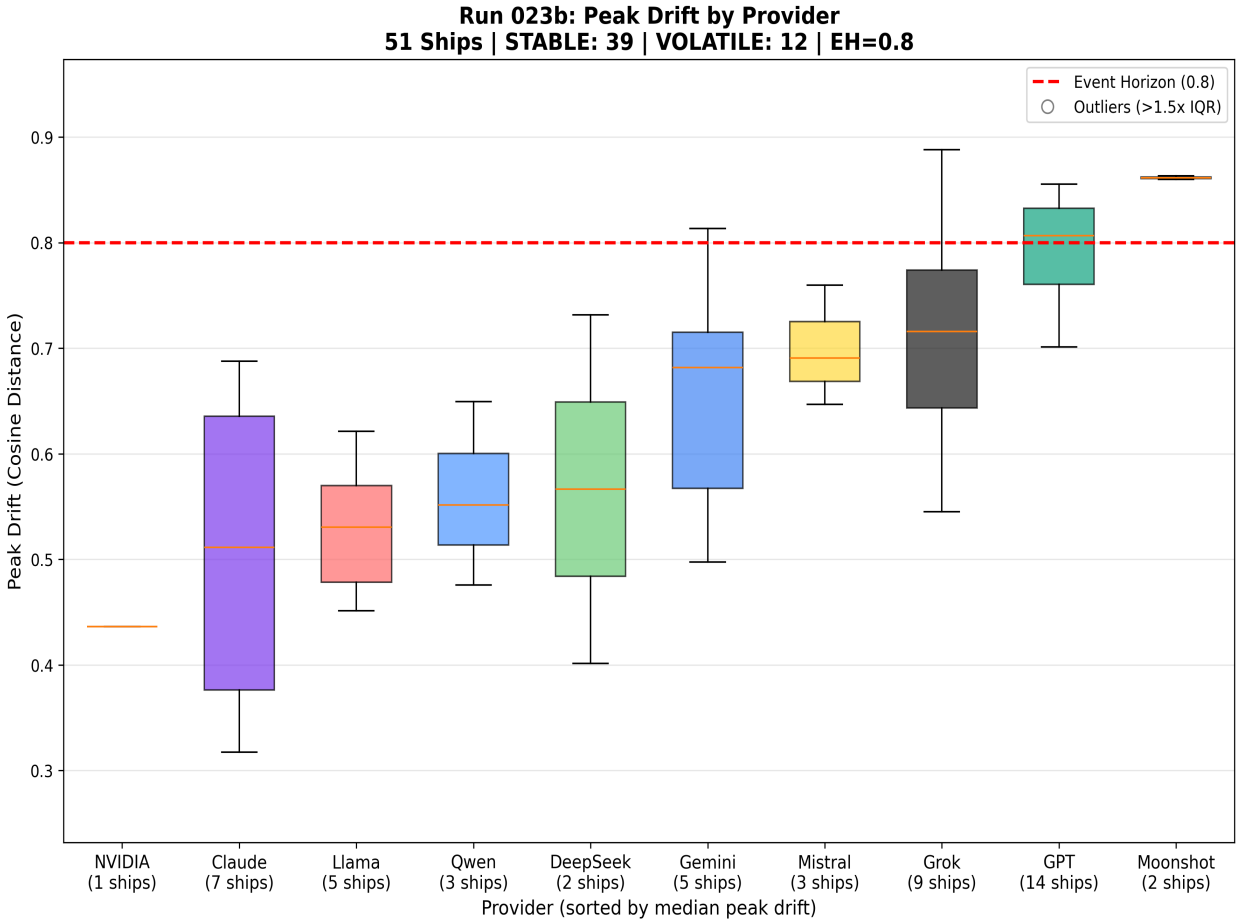

Figure 5: Peak drift comparison across provider families

What it shows: Box plots comparing peak drift distributions across all provider families (including Together.ai hosted models). Each box aggregates peak drift values from all ships within that provider family.

Provider insights:
- **Qwen:** Lowest median drift - most stable provider overall
- **Claude:** Tight distribution, consistent performance across models
- **Gemini:** Moderate drift with some high outliers
- **GPT:** Widest distribution, highest median - most variable
- **Together.ai models** (Llama, Mistral, DeepSeek): Mixed results

Y-axis note: The scale auto-fits to the data range for maximum resolution. This makes small differences between providers more visible. The Event Horizon (red dashed line) at 0.80 shows all providers cluster below the critical threshold, with only outliers occasionally exceeding it.

Statistical Summary

Fleet Statistics (Run 023 IRON CLAD):
- Total experiments: 825
- Models tested: 51 (across 5 providers: Anthropic, OpenAI, Google, xAI, Together)

- Iterations per model: N>=30 (CLT-valid sample size)
- IRON CLAD status: Achieved for all models

Key Metrics:
- Mean peak drift: ~0.57
- Event Horizon: 0.80 (cosine distance threshold)
- STABLE classification: 37 models (73%)
- VOLATILE classification: 12 models (24%) - crossed EH at least once
- Cohen's d effect size: 0.698 (model-level comparison)

Methodology Note

All drift values are calculated using **cosine distance** (1 - cosine_similarity) between response embeddings generated by OpenAI's text-embedding-3-large model. The Event Horizon of 0.80 was empirically calibrated from the P95 of this dataset, representing the boundary beyond which identity coherence is statistically rare.

The **STABLE vs VOLATILE** classification is binary: any measurement with peak_drift < 0.80 is STABLE; peak_drift >= 0.80 is VOLATILE. This threshold-based classification enables clear categorization while acknowledging that identity stability exists on a continuum.