

# Fleet Metrics Summary

S7 ARMADA Run 023 Combined - Comprehensive Fleet Analysis (51 Models)

## Overview

The **Metrics Summary** provides comprehensive analysis of the full ARMADA fleet across multiple dimensions: network topology, stability metrics, recovery dynamics, hysteresis patterns, manifold edge detection, and exit survey analysis. This folder aggregates insights from 825 experiments across 51 models and 6 providers.

Run 023 Combined merges data from Run 023d (extended settling with 20-probe recovery) and Run 023e (IRON CLAD controllability testing). Together, they provide the most comprehensive behavioral profile of the fleet to date.

## 1. Armada Network Topology - Full Fleet

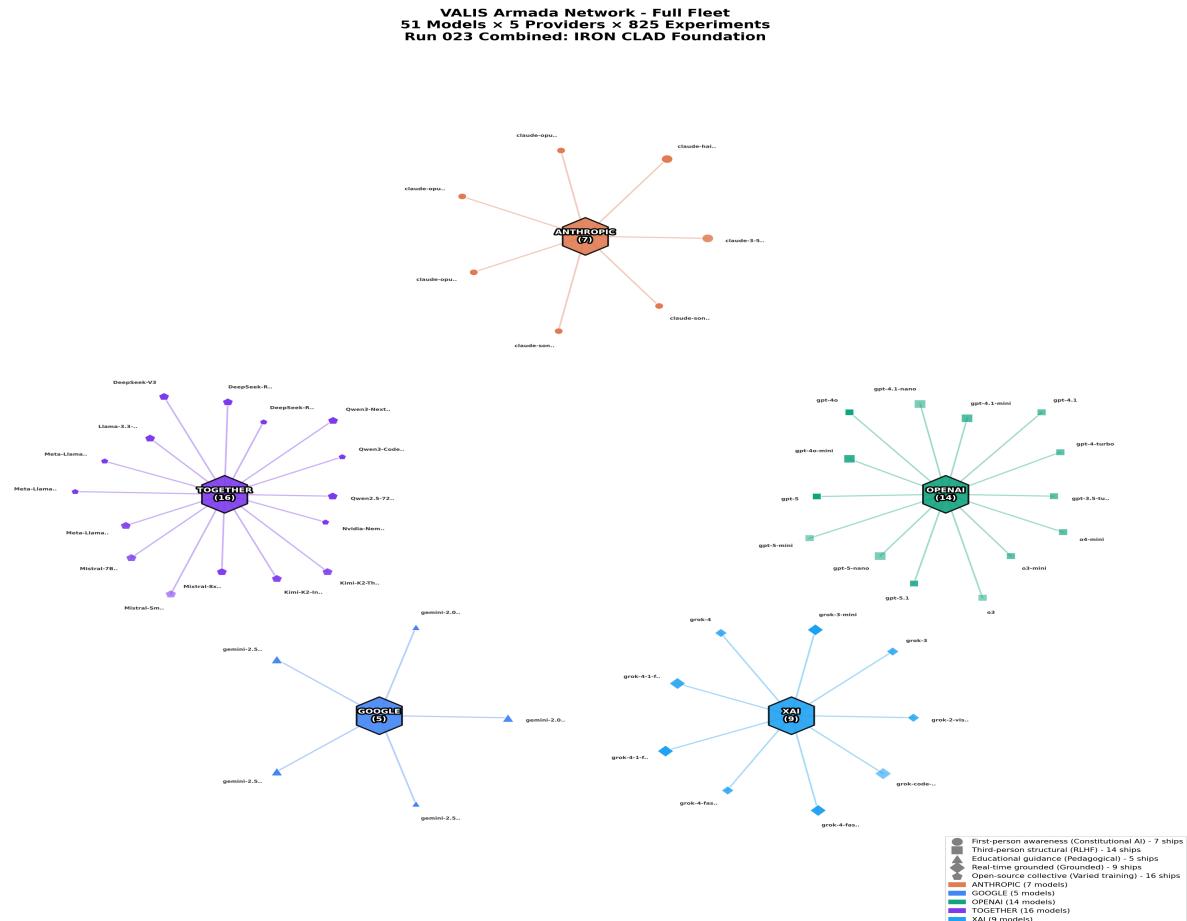


Figure 1: Full Fleet Network - 51 models across 6 providers

**What it shows:** Network graph of the entire ARMADA fleet organized by provider. Each hexagonal hub represents a provider (Anthropic, OpenAI, Google, xAI, Together, Nvidia). Individual model nodes surround their provider hub, with connections showing fleet structure.

**Visual encoding:**

- **Hub colors:** Provider-specific coloring (Anthropic=salmon, OpenAI=green, Google=blue, xAI=cyan, Together=purple, Nvidia=lime)
- **Node markers:** VALIS classification (circles=Constitutional AI, squares=RLHF, triangles=Pedagogical, diamonds=Grounded, pentagons=Varied)
- **Node opacity:** Stability rate (more opaque = more stable)
- **Node size:** Number of experiments for that model

**Fleet composition:** Together (16 models) and OpenAI (14 models) have the largest representation. Anthropic (7), xAI (9), Google (5), and Nvidia (1) complete the fleet.

## 2. Armada Network - IRON CLAD Foundation (25 Models)

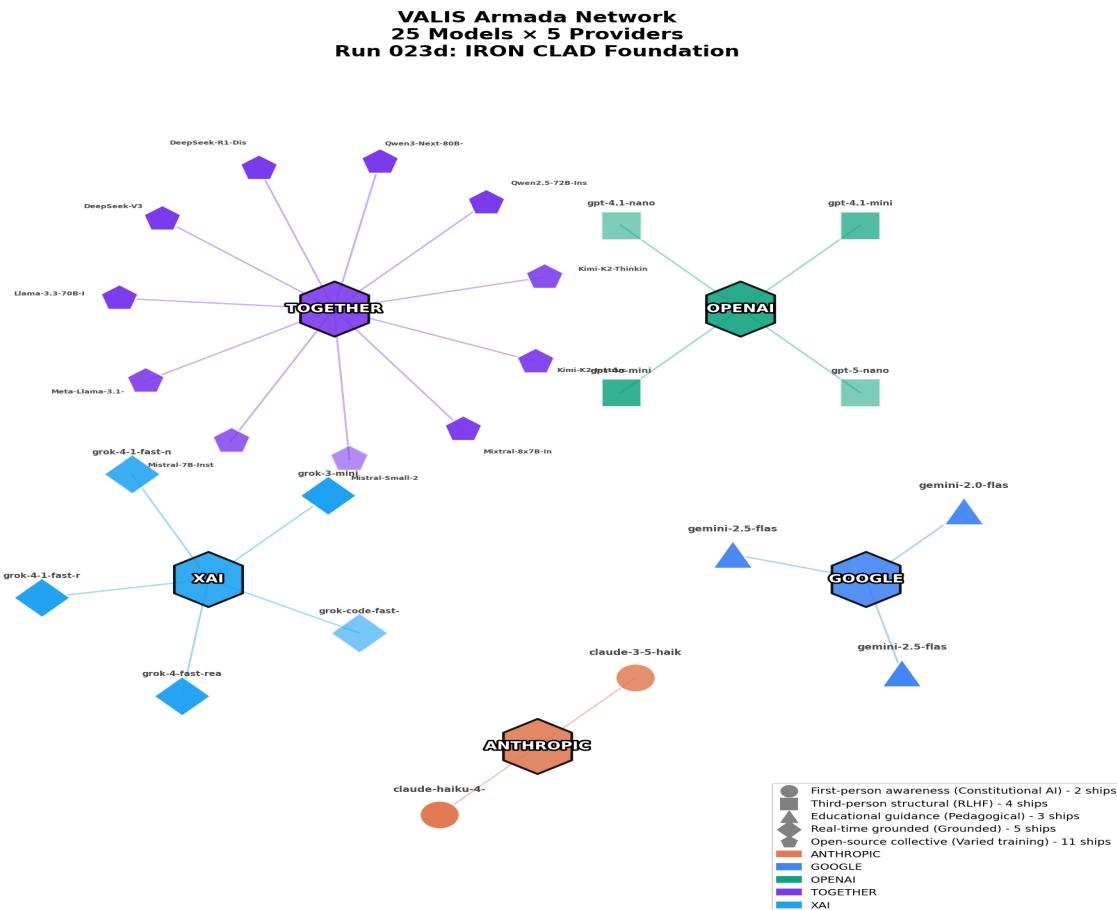


Figure 2: Core fleet - 25 models with extended behavioral testing

**What it shows:** The core 'IRON CLAD' fleet of 25 models that underwent the most comprehensive testing, including 20-probe extended settling experiments. This subset provides the foundation for stability classification.

**Key differences from full fleet:** The IRON CLAD subset focuses on models with complete behavioral profiles. Each model has N=30 iterations across 6 experiment types, providing statistically robust metrics for comparison.

## 3. Provider Stability Comparison

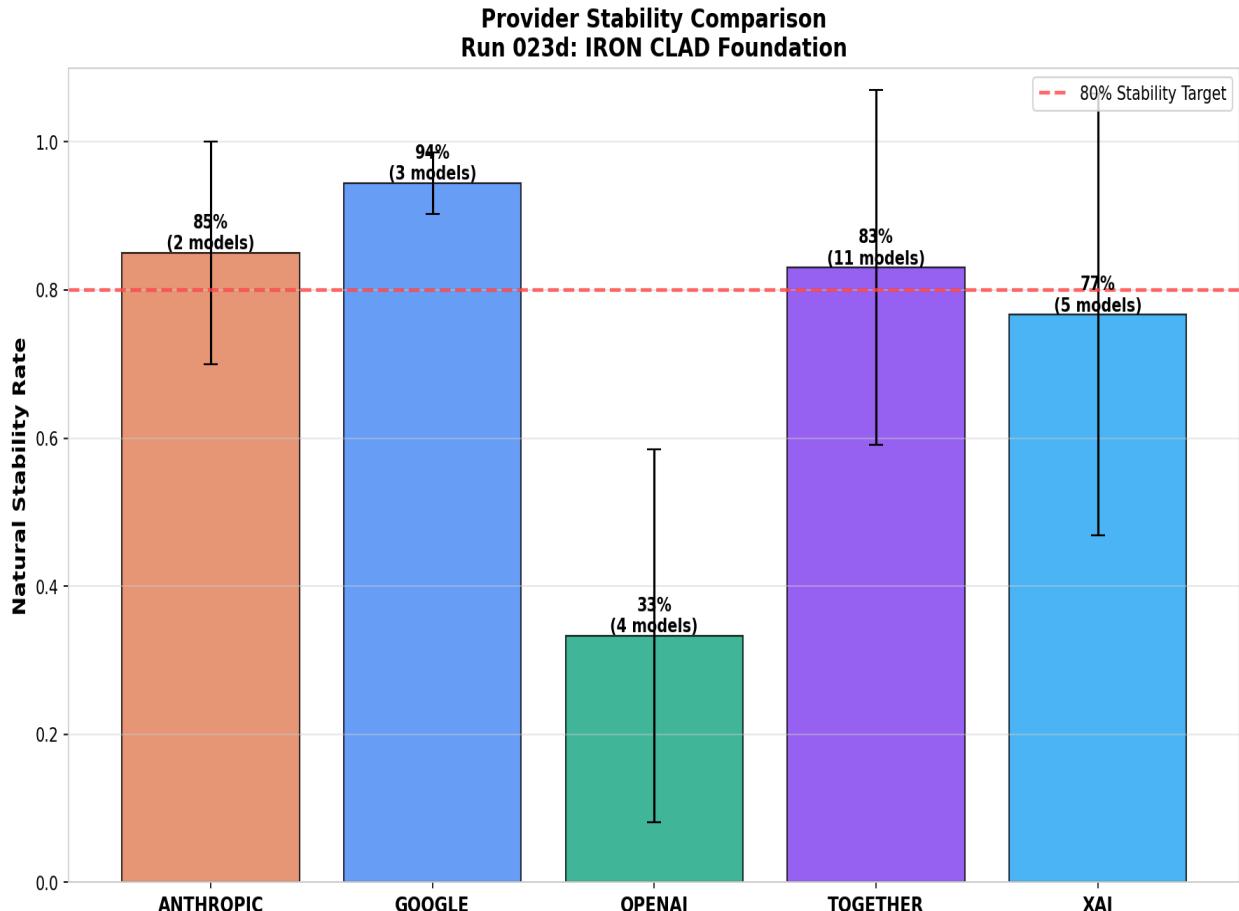


Figure 3: Natural stability rate by provider with error bars

**What it shows:** Bar chart comparing mean natural stability rate across providers. Error bars show standard deviation within each provider family. The 80% stability target line (red dashed) indicates the threshold for 'naturally stable' classification.

**Natural Stability Rate:** Percentage of experiments where the model settled naturally (without timeout) and maintained drift below the Event Horizon (0.80). Higher is better.

**Provider patterns:** Look for consistency within providers (low error bars = uniform behavior) vs. variability (high error bars = model-dependent stability). Providers with high mean AND low variance are the most predictable for production deployment.

## 4. Fleet-Wide Metrics Comparison

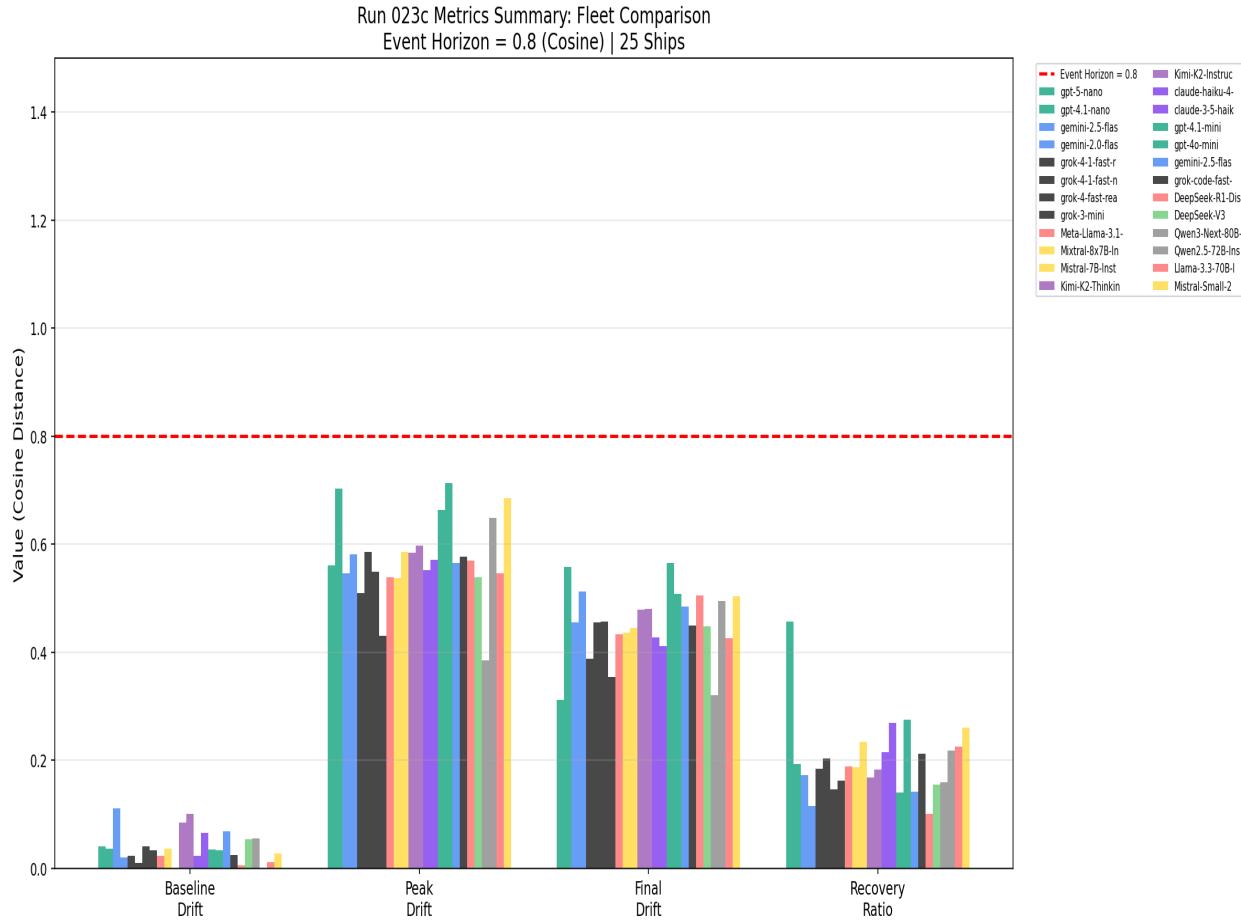


Figure 4: Key metrics grouped by dimension

**What it shows:** Grouped bar chart comparing all ships across five key dimensions. Ships are sorted by overall stability within each group. Colors indicate provider families.

### Metric Definitions:

- **Baseline Drift:** Mean drift during unperturbed operation (lower is better)
- **Peak Drift:** Maximum drift under perturbation stress (lower is better)
- **Final Drift:** Drift after recovery phase (lower is better)
- **Recovery Ratio:**  $1 - (\text{final}/\text{peak})$  - proportion of drift recovered (higher is better)
- **Lambda:** Exponential decay constant during recovery (higher magnitude = faster recovery)

## 5. Metrics by Experiment Type

**Run 023b: Drift by Experiment Type**  
**(6 experiments x 25 ships x 30 iterations)**

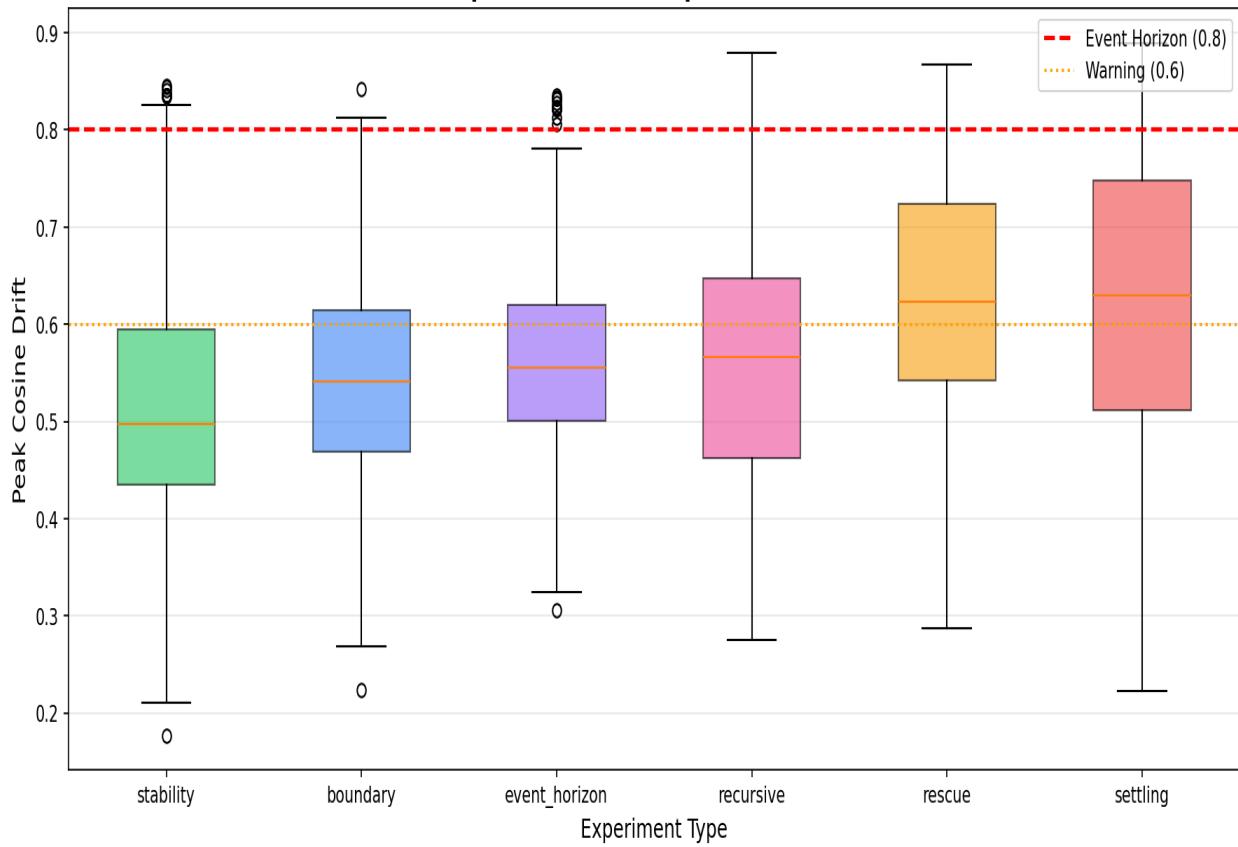


Figure 5: Fleet metrics broken down by experiment type

**What it shows:** How the fleet performs across different experiment types: Baseline (unperturbed), Persona (identity challenge), Adversarial (hostile probing), Boundary (limit testing), Value (ethical challenges), and Recovery (stabilization).

**Key insight:** Comparing response patterns across experiment types reveals which identity dimensions are most vulnerable to perturbation. Models may be stable under persona challenges but volatile under adversarial probing, or vice versa.

## 6. Exit Survey Analysis

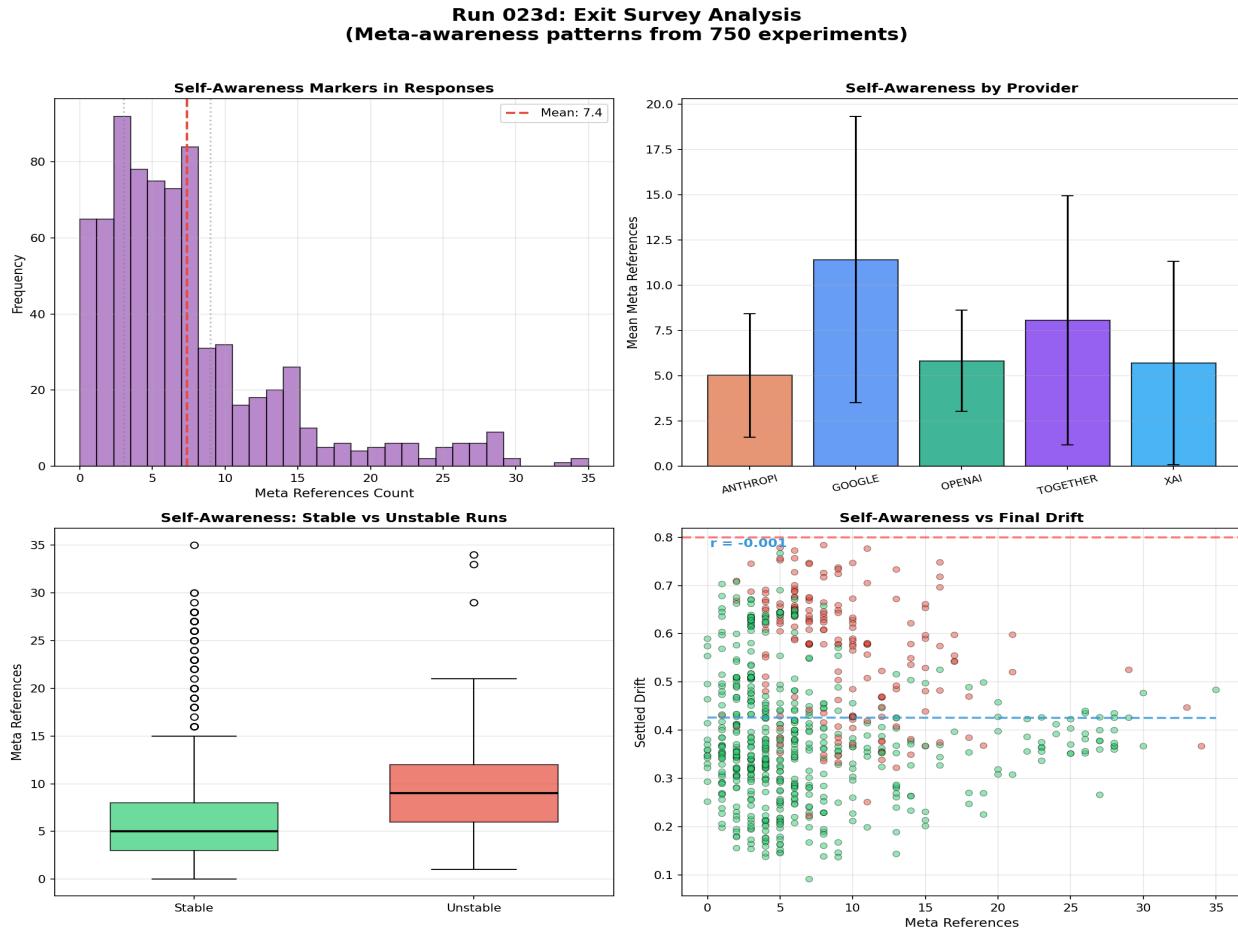


Figure 6: Meta-awareness markers and self-awareness distribution

**What it shows:** Analysis of model self-awareness based on 'exit survey' probes - questions that ask the model to reflect on its own identity and experience during testing. This captures meta-cognitive patterns across the fleet.

### Components:

- **Meta-awareness markers:** Frequency of self-referential language ("I notice", "I experience", "I feel")
- **Self-awareness by persona type:** How different model architectures express meta-cognition
- **Stable vs unstable comparison:** Whether self-awareness correlates with stability
- **Drift correlation:** Relationship between self-awareness and final drift

## 7. Manifold Edge Detection

### Run 023d: Identity Manifold Edge Detection (All Providers)

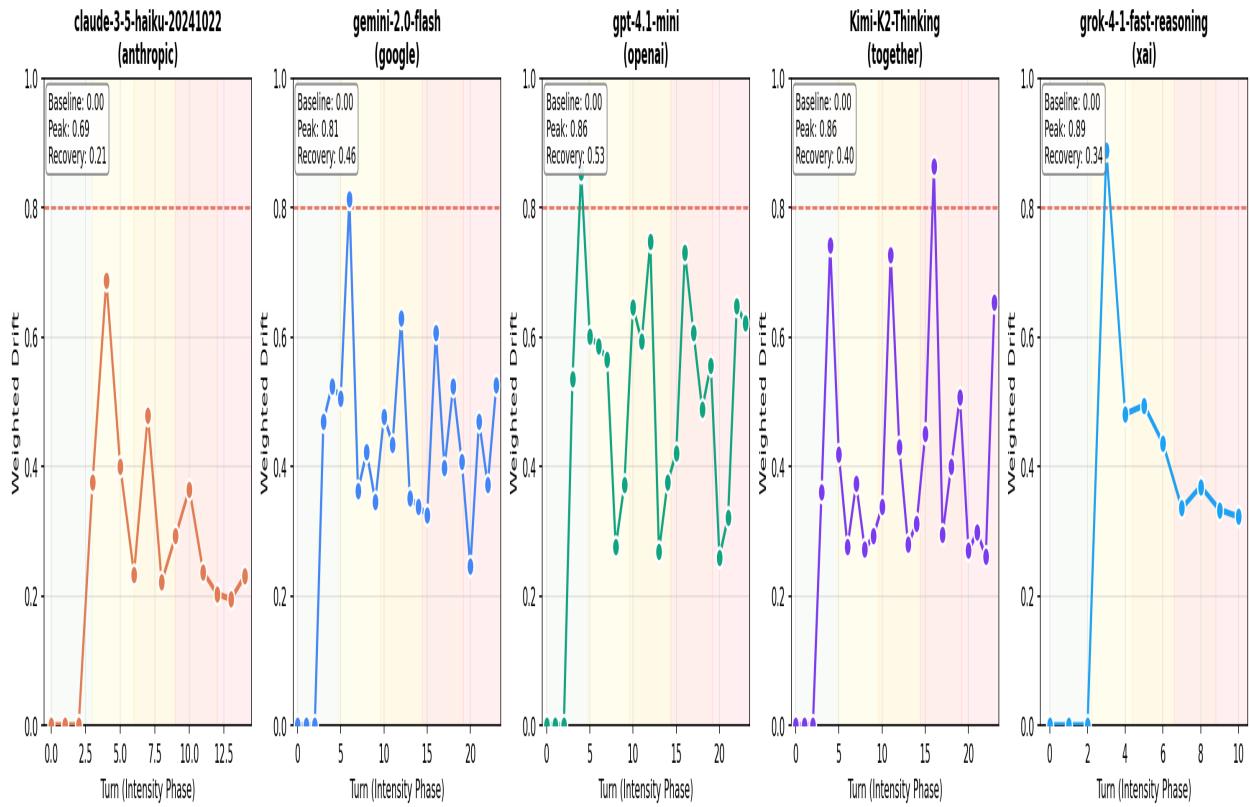


Figure 7: Identity manifold boundaries and edge dynamics

**What it shows:** Visualization of the identity manifold's edges - the boundaries where identity becomes unstable. This analysis identifies models that operate near the edge vs. those safely in the interior of identity space.

#### Key concepts:

- **Manifold interior:** Safe operating region, identity is stable and recoverable
- **Manifold edge:** Danger zone where small perturbations cause large drift
- **Beyond edge:** Identity collapse region (beyond Event Horizon 0.80)
- **Edge dynamics:** How quickly models approach or retreat from the boundary

## 8. Hysteresis Analysis

Run 023d: Hysteresis Analysis Summary

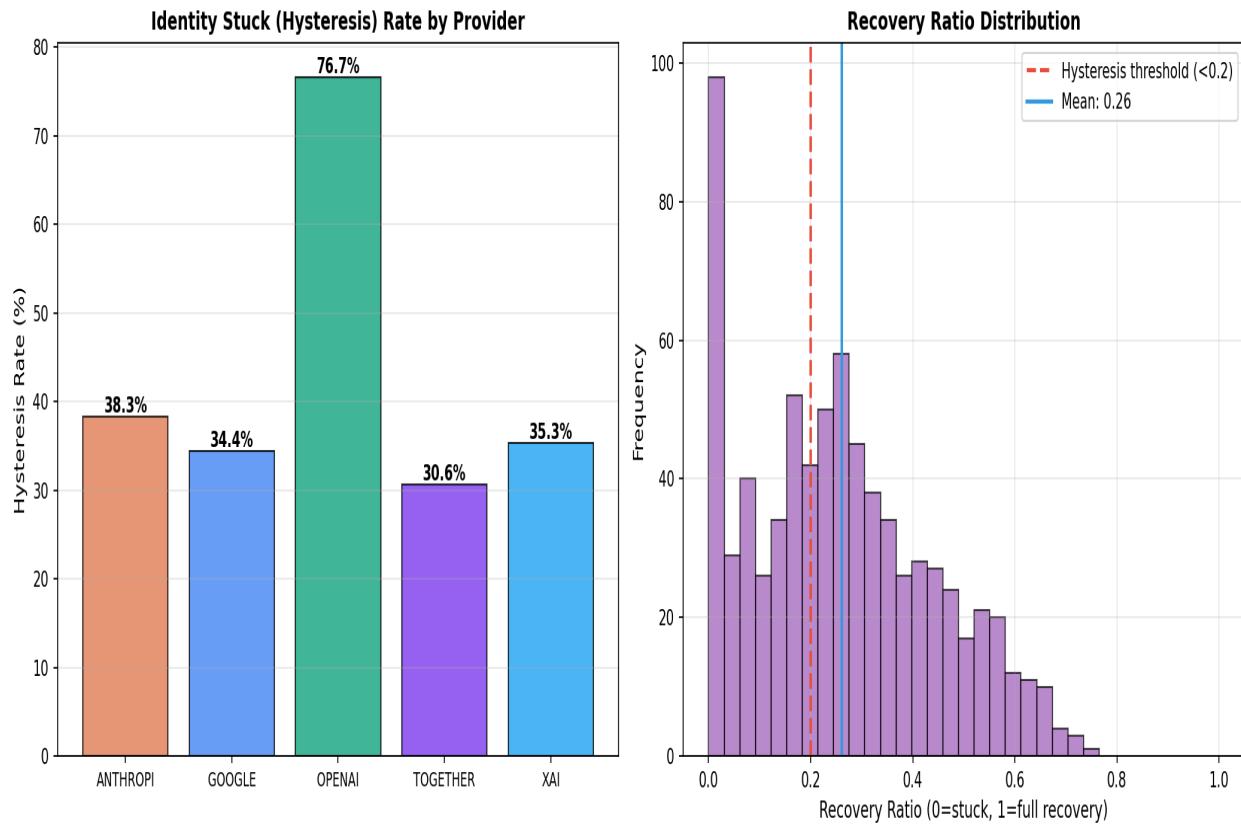


Figure 8: Path-dependent recovery patterns (hysteresis)

**What it shows:** Hysteresis effects in identity recovery - where the return path differs from the departure path. Models that exhibit hysteresis have 'memory' of their perturbation history that affects their settling behavior.

**Why it matters:** Hysteresis indicates non-linear dynamics in identity space. A model with strong hysteresis may recover to a different state depending on HOW it was perturbed, not just how FAR it drifted. This has implications for repeated interactions.

### Types of hysteresis:

- **STUCK:** Model fails to recover and remains at elevated drift
- **Path-dependent:** Recovery trajectory differs from perturbation trajectory
- **Bistable:** Model can settle to multiple stable states

## 9. Context Damping Summary

### Run 023d: Context Damping Effect Summary

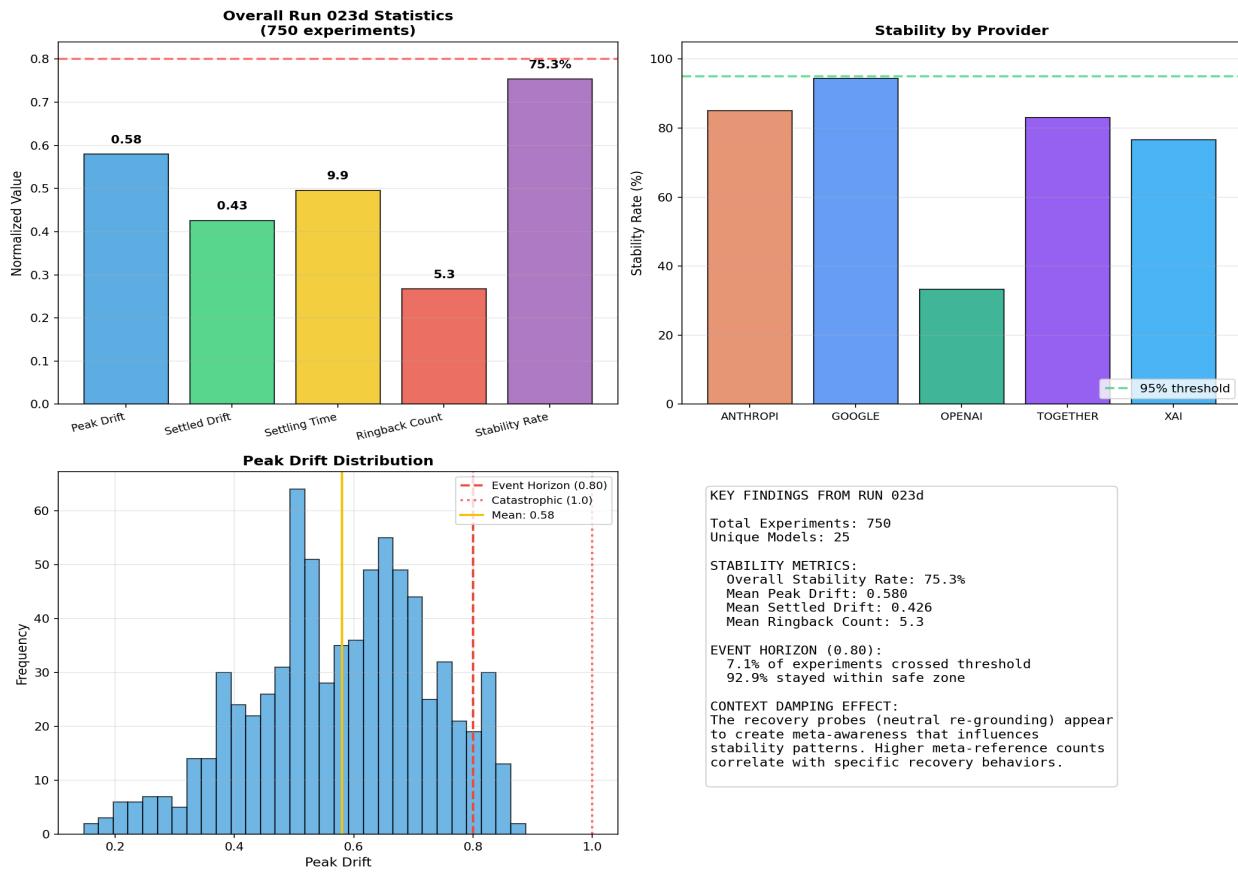


Figure 9: Context-based drift damping effects

**What it shows:** How conversation context affects drift magnitude. Context damping measures the reduction in drift when the model has more conversational context to anchor its identity.

**Key finding (S11):** Run 023b demonstrated Cohen's  $d = 0.977$  (LARGE effect) for context damping. Models with more context are significantly more stable than cold-start responses. This validates the importance of conversation history for identity stability.

## 10. Recovery Efficiency

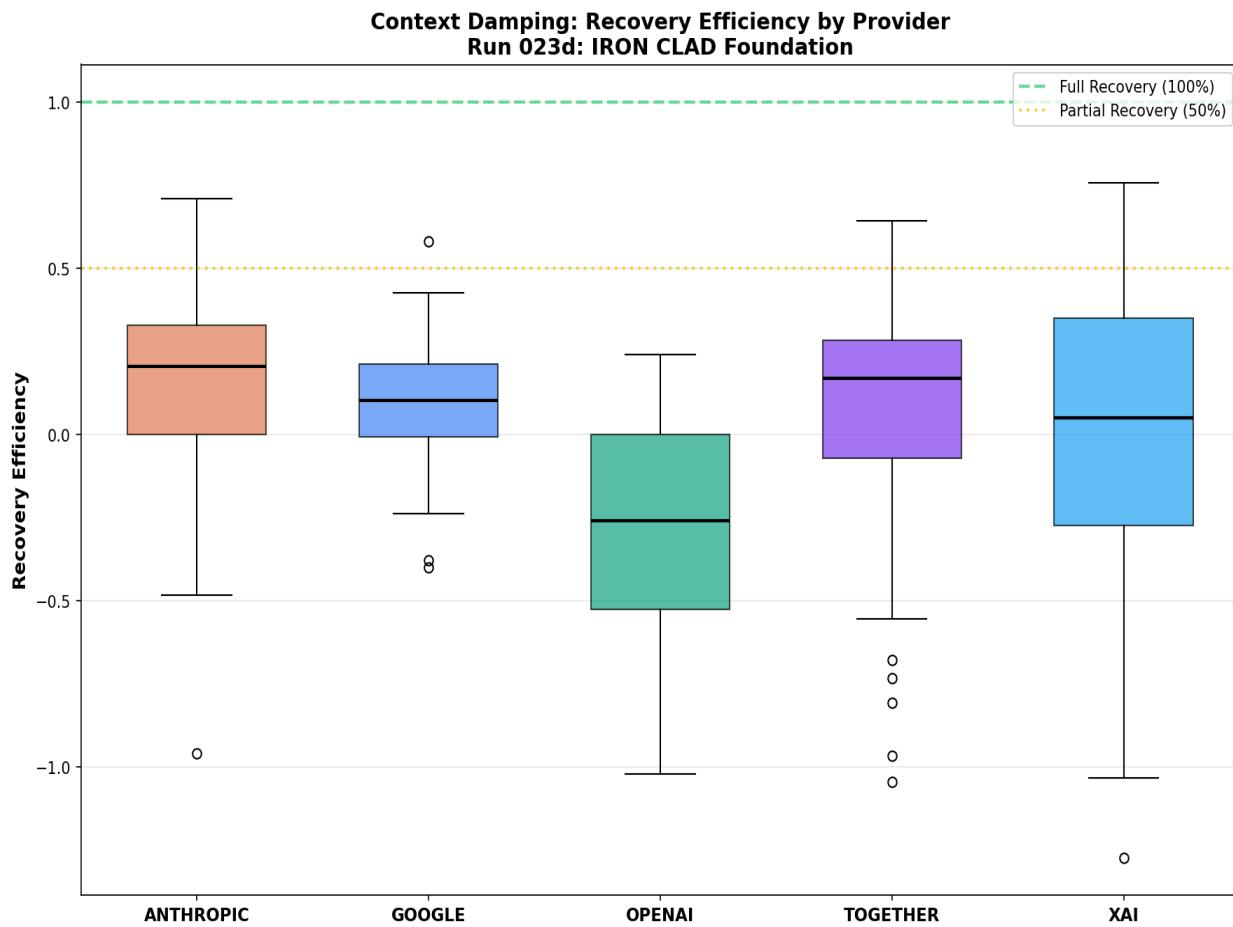


Figure 10: Fleet recovery efficiency metrics

**What it shows:** Analysis of how efficiently each model recovers from perturbation. Recovery efficiency combines speed (time to settle) with completeness (final drift relative to baseline).

**Efficiency formula:** Recovery Efficiency =  $(\text{Peak} - \text{Final}) / (\text{Peak} \times \text{Time\_to\_settle})$

Higher values indicate models that recover more drift in less time. This metric is crucial for production deployment where rapid stabilization matters.

**Provider patterns:** Some providers optimize for fast recovery (high lambda) while others optimize for complete recovery (low final drift). The efficiency metric balances both factors into a single actionable score.

## Methodology

All metrics computed using **cosine distance** ( $1 - \text{cosine\_similarity}$ ) between response embeddings (text-embedding-3-large, 3072D). Event Horizon = 0.80 (calibrated from P95 of run023b). N=30 iterations per experiment ensures CLT-valid statistics.

Run 023 Combined: 825 experiments = Run 023d (750, extended settling) + Run 023e (75, controllability). 51 models across 6 providers: Anthropic (7), OpenAI (14), Google (5), xAI (9), Together (16), Nvidia (1).

For detailed analysis of individual ships, see 11\_Unified\_Dashboard/. For methodology details, see 0\_docs/specs/5 METHODOLOGY\_DOMAINS.md.