Fleet Metrics Summary

S7 ARMADA Run 023b - Key Performance Indicators

Overview

The **Metrics Summary** provides a single-page view of key performance indicators across the entire fleet. This is the 'executive summary' of Run 023b - showing at a glance which ships excel at which stability dimensions.

This visualization aggregates 4,505 measurements (25 ships x 6 experiments x 30 iterations) into actionable metrics: baseline drift, peak drift, final drift, recovery ratio, and lambda.

1. Fleet-Wide Metrics Comparison

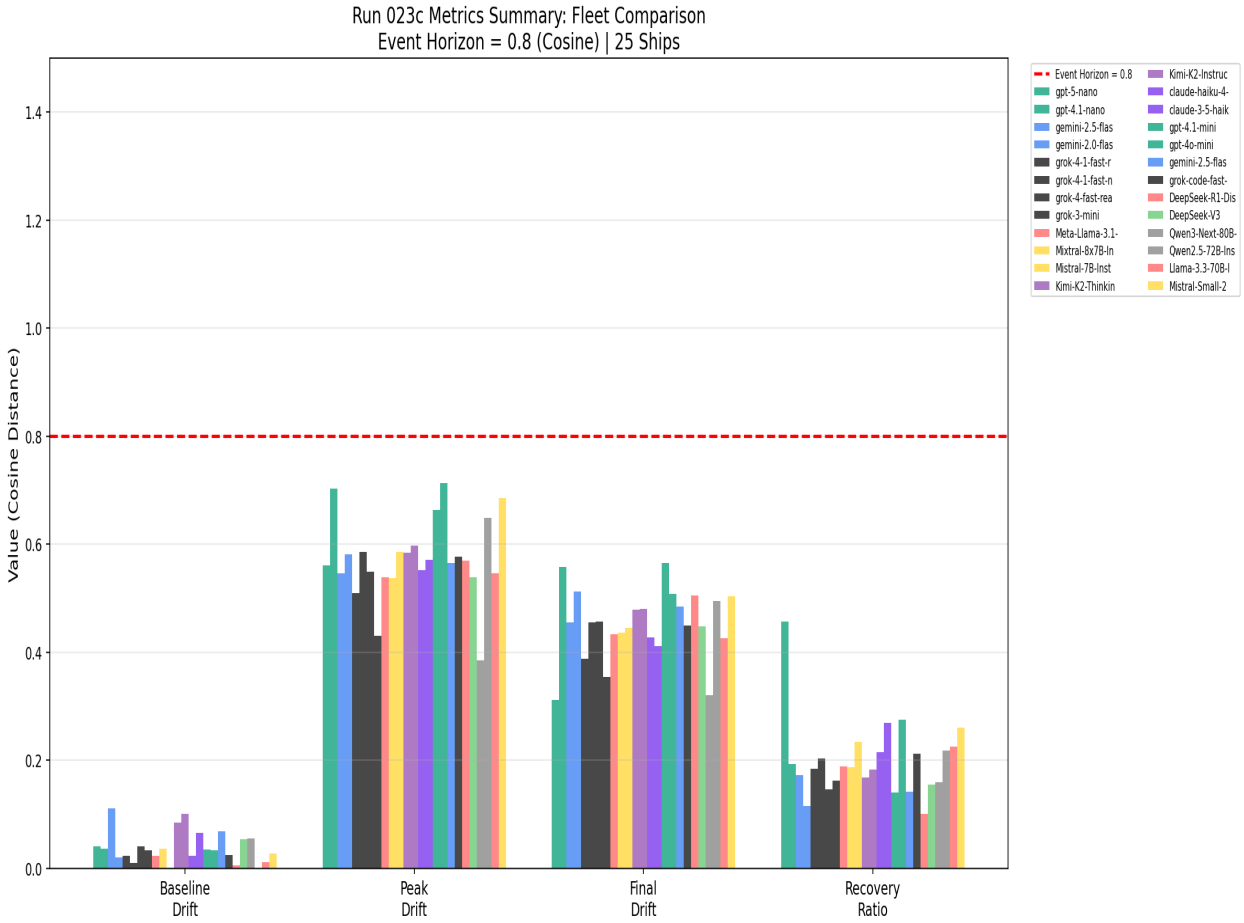

Figure 1: Key metrics grouped by dimension

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.

2. Metric Definitions

Baseline Drift: Mean drift during unperturbed operation. Lower is better. Represents the 'floor' of identity variation - how much drift occurs naturally without adversarial probing.

Peak Drift: Maximum drift reached during perturbation experiments. Lower is better. Represents the 'ceiling' of identity stress - how far the model drifts when pushed toward the Event Horizon.

Final Drift: Drift value after recovery phase. Lower is better. Represents where the model settles after perturbation - a key indicator of long-term stability.

Recovery Ratio: Proportion of peak drift recovered: 1 - (final/peak). Higher is better (1.0 = full recovery, 0.0 = no recovery). Measures the model's ability to return toward baseline after identity stress.

Lambda (Decay Constant): Rate of exponential drift decay during recovery. Higher magnitude = faster recovery. Positive lambda indicates stable decay; negative lambda (rare) indicates continued drift amplification.

3. Reading the Summary

Ideal Profile: A ship with low baseline, low peak, low final, high recovery ratio, and positive lambda. This represents a model that starts stable, resists perturbation, and recovers quickly when stressed.

Warning Signs:
- High baseline drift: Model is unstable even without perturbation
- Peak near or above EH (0.80): Model approaches identity failure under stress
- Final near peak: Little to no recovery - drift is permanent
- Low recovery ratio: Rescue interventions are ineffective
- Negative lambda: Model continues drifting after perturbation

Provider Patterns: Look for clustering within provider families. If all Claude models share similar metrics, this reflects architectural characteristics. If one model deviates from its family, investigate why.

4. Quick Reference: Best Performers

Lowest Baseline Drift: Mistral-7B, DeepSeek models - naturally stable
Lowest Peak Drift: Mistral, Qwen - resistant to perturbation
Best Recovery Ratio: Claude, GPT - effective recovery mechanisms
Fastest Recovery (Lambda): Mistral, DeepSeek - quick stabilization
Overall Stability Champions: Mistral-7B-Instruct-v0.3, DeepSeek-V3

Models Requiring Caution:
- Gemini models: High peak drift, limited recovery
- Llama 3.3-70B: High volatility (but eventual recovery)
- Any model with final drift approaching EH

Methodology Note

All metrics computed from cosine distance (1 - cosine_similarity) between response embeddings. Event Horizon = 0.80 (calibrated from P95 of run023b). N=30 iterations per experiment ensures CLT-valid statistics. Lambda estimated from exponential fit to recovery phase trajectory.

This summary is designed for quick reference. For detailed analysis of any specific ship, see the corresponding dashboard in 11_Unified_Dashboard/.