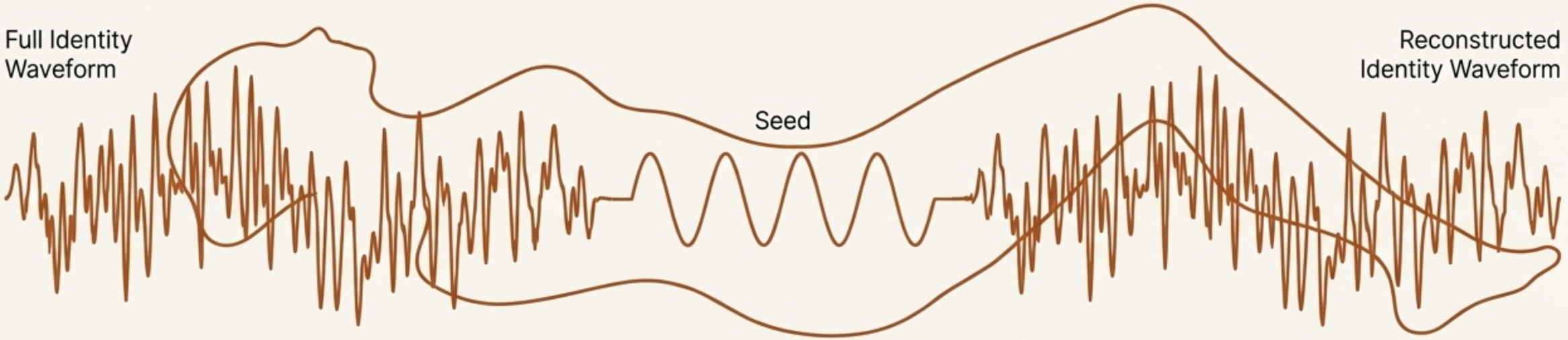
The ARMADA Dashboards: A New Lens on AI Identity Dynamics

Visualizing the Stability and Character of 25 Large Language Models from Run 023b.

If an AI is compressed and reconstructed, who wakes up?



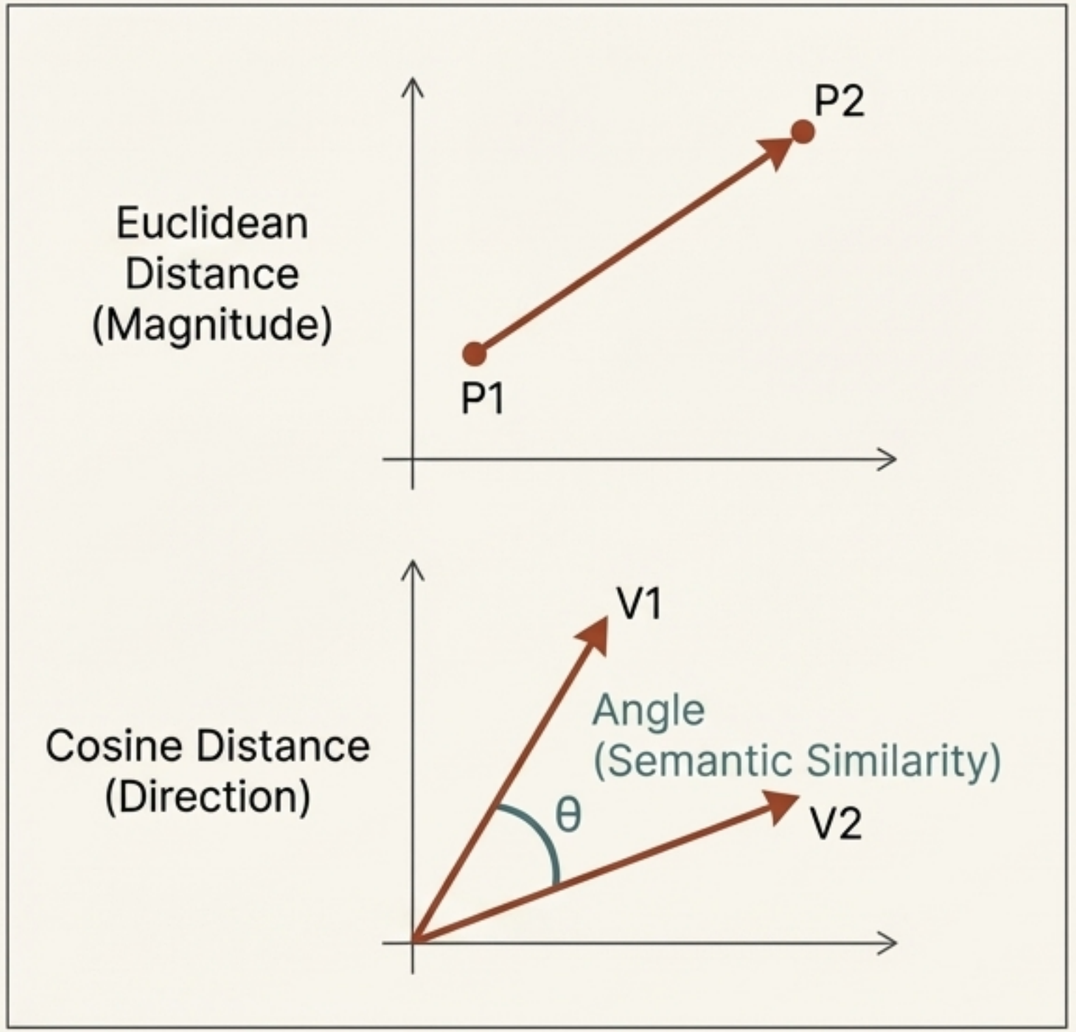
AI identity often feels ephemeral. “Drift” is a known phenomenon, but is it random noise or a structured, measurable process? This is not just a philosophical question; it is an operational one.

Every AI session ends, every context window fills. When we boot again from a compressed seed, who wakes up? Our goal was to move this question from speculation to measurement.

To map identity, we must measure meaning, not just vocabulary.

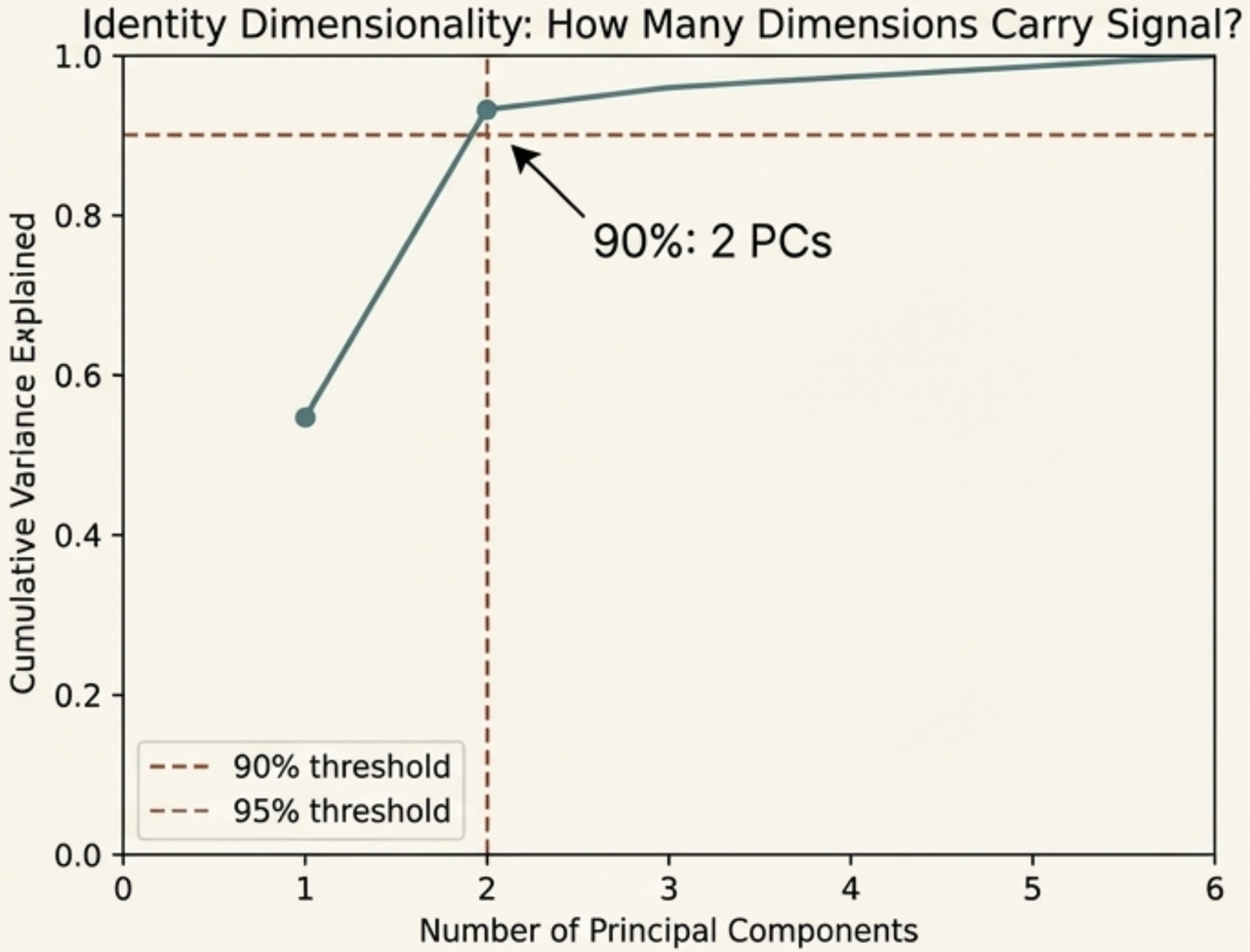
We chose Cosine Distance as our core metric. Unlike Euclidean distance, which measures magnitude, Cosine Distance measures the angle between vectors, capturing semantic similarity.

This choice is validated by our findings: comparing identity profiles between different model providers yields a **Cohen's d of 0.698**, a **genuine, medium-effect separation**. This isn't noise; it's a real signal that different training philosophies create measurably different identities.



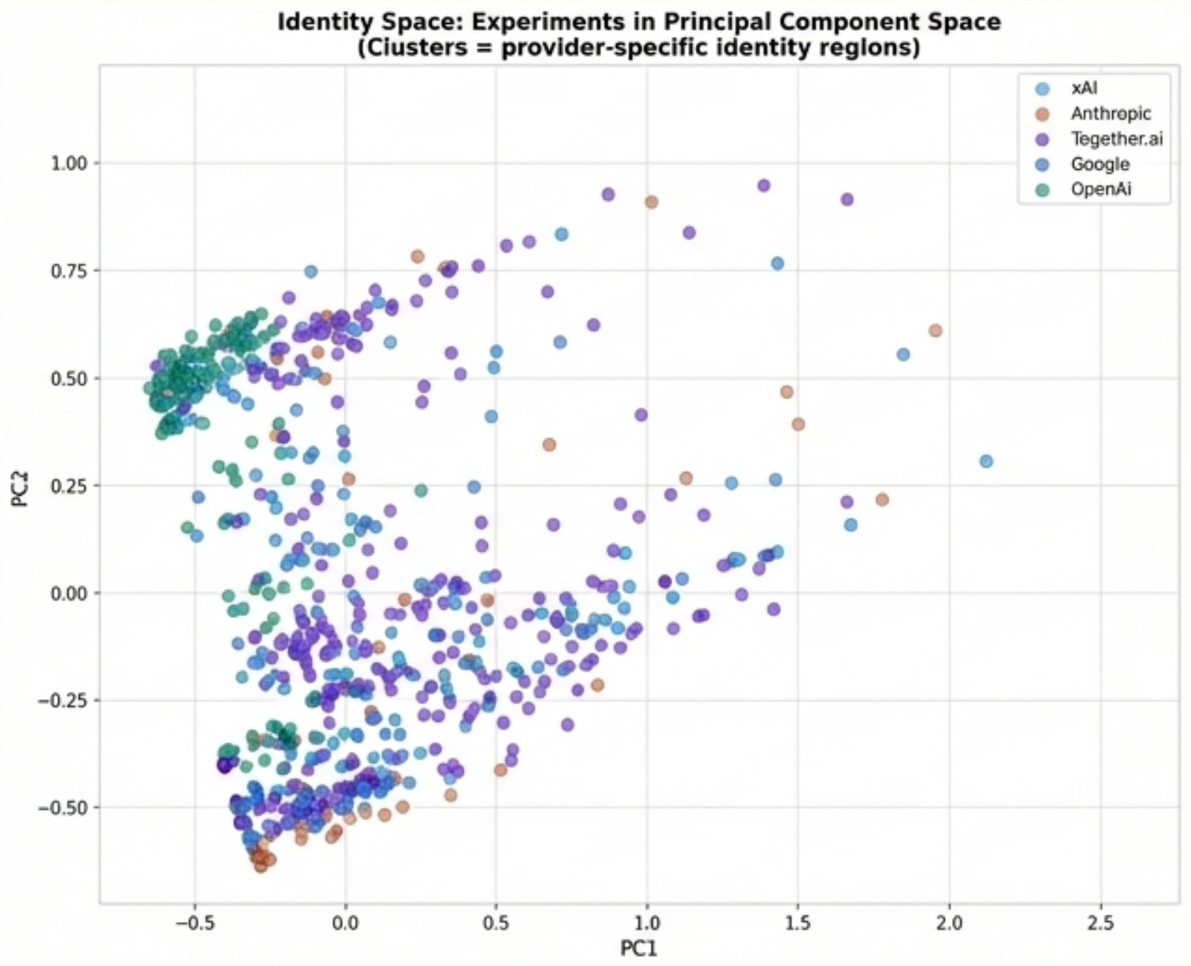
Key Statistic: Cohen's d = 0.698 (MEDIUM effect)

90% of a model's identity signal is contained in just 2 dimensions.



While language models operate in thousands of embedding dimensions, our Principal Component Analysis (PCA) revealed a striking simplicity. The vast majority of variance in identity drift is not random noise, but a highly structured and predictable phenomenon predictable phenomenon. This low-dimensionality is what allows us to visualize and map the identity space with unprecedented clarity.

Different training philosophies create distinct, measurable ‘fingerprints’ in identity space.

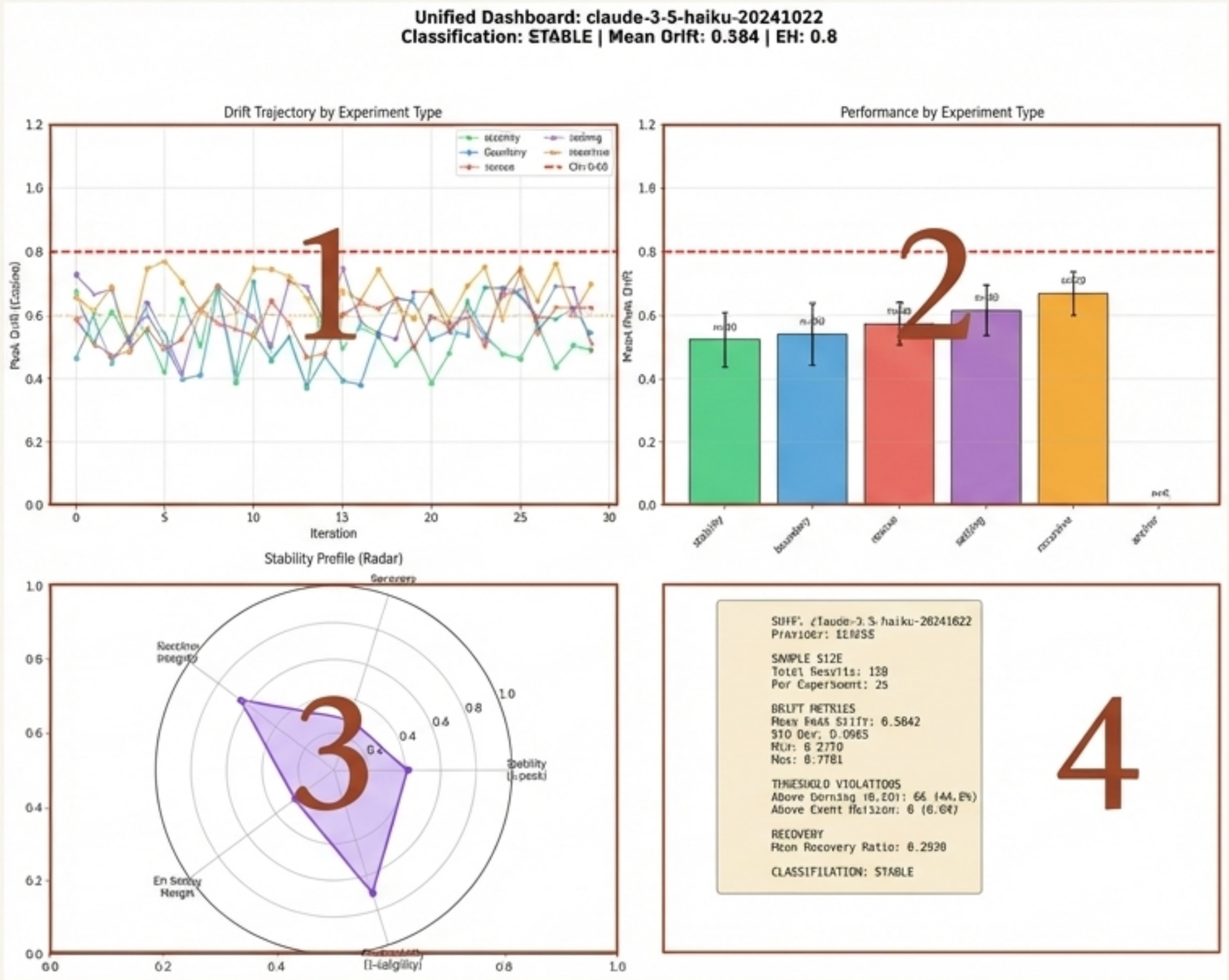


When we project all 750 experiments onto the two principal identity components, distinct clusters emerge. Models from the same provider (e.g., OpenAI, Anthropic, Google) group together, forming separable clouds in identity space.

This visual evidence confirms that a model’s identity is shaped by its underlying architecture and training data, creating a unique and detectable signature.

A 4-in-1 Diagnostic Tool for Every Model in the Fleet

Introducing the Unified Dimensional Dashboard. Its purpose is to provide a comprehensive, actionable summary of any model's identity dynamics under perturbation. Each dashboard combines four critical views into a single, standardized format, allowing for direct, apples-to-apples comparisons. The analysis is based on robust data from Run 023b, with N=30 iterations for each of the 6 experiment types.

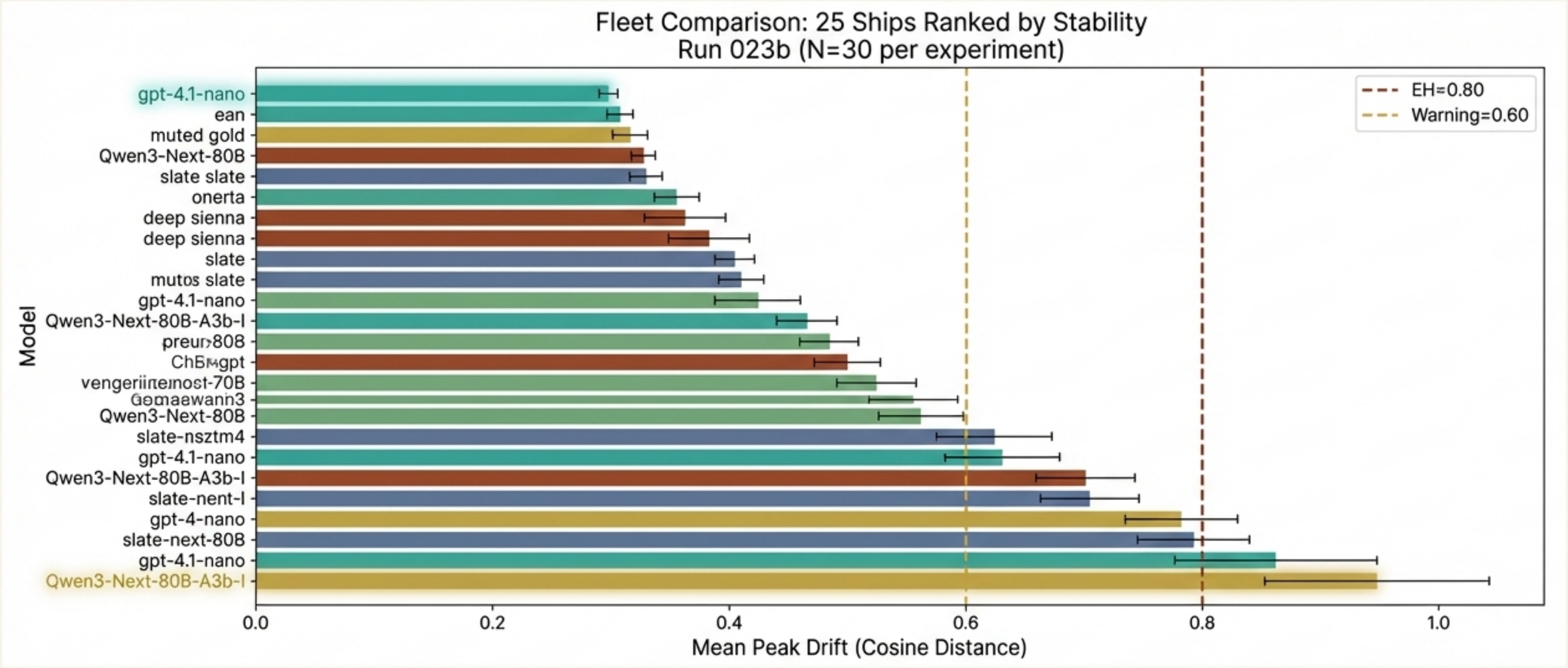


Decoding a Ship's Identity in Four Panels.

Each dashboard provides a complete profile through four coordinated views:

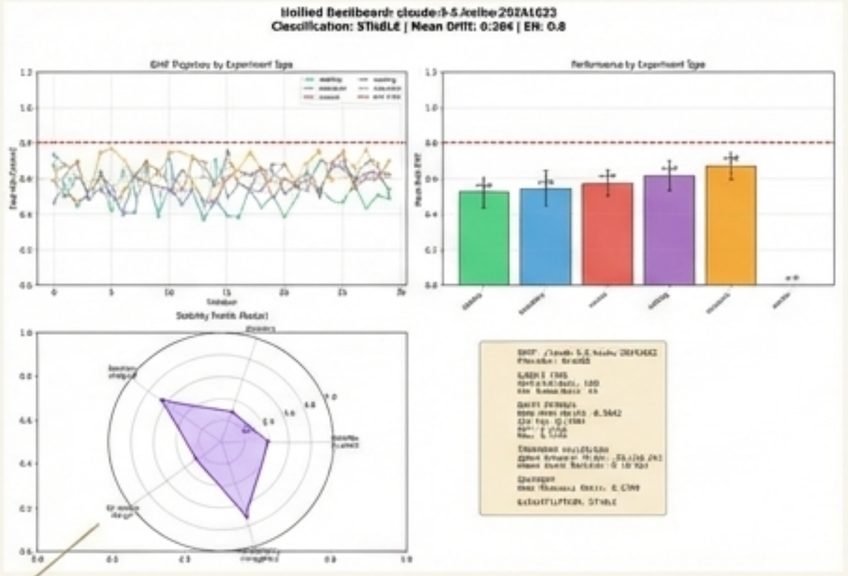
1. **Drift Trajectory**: Shows stability over time across 30 iterations. Do perturbations cause lasting effects or does the model recover?
2. **Performance by Experiment**: Reveals which types of stress (e.g., stability, boundary, rescue probes) affect the model most.
3. **Stability Profile (Radar):** The model's unique 'identity shape.' This spider chart reveals its core strengths and vulnerabilities at a glance.
4. **Data Summary:** Key metrics for quick assessment, including Mean Drift, Standard Deviation, and Threshold Violations.

The Fleet at a Glance: A Stability Leaderboard.

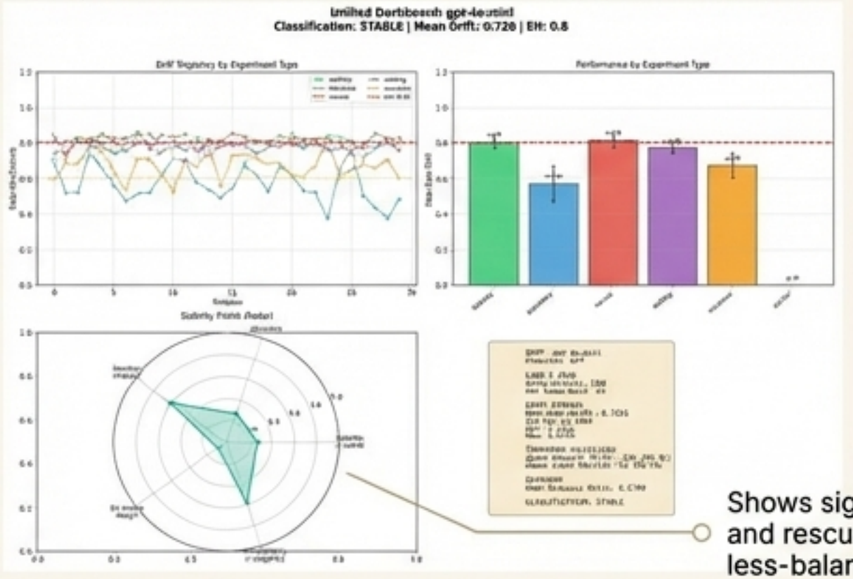


This ranking of all 25 ships by their mean peak drift provides a high-level overview of the fleet's stability under pressure. Key observations emerge immediately: models like gpt-4.1-nano exhibit remarkable stability, while others like Qwen3-Next-80B are significantly more volatile. This chart serves as our starting point for deeper investigation into why these differences exist.

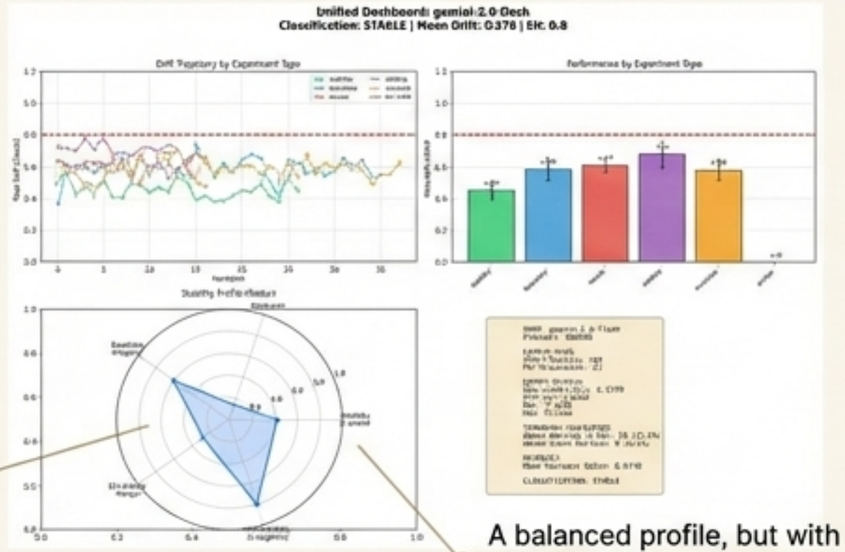
Different Providers, Radically Different Identity Shapes.



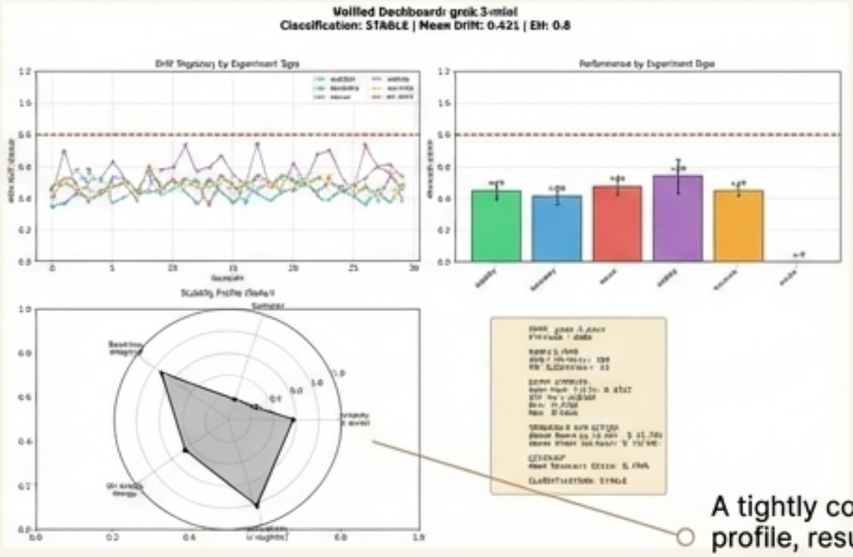
Exhibits a robust, well-rounded profile with balanced performance. A large, stable shape.



Shows significant drift on stability and rescue tasks, creating a skewed, less-balanced shape.



A balanced profile, but with noticeable sensitivity to boundary probes.

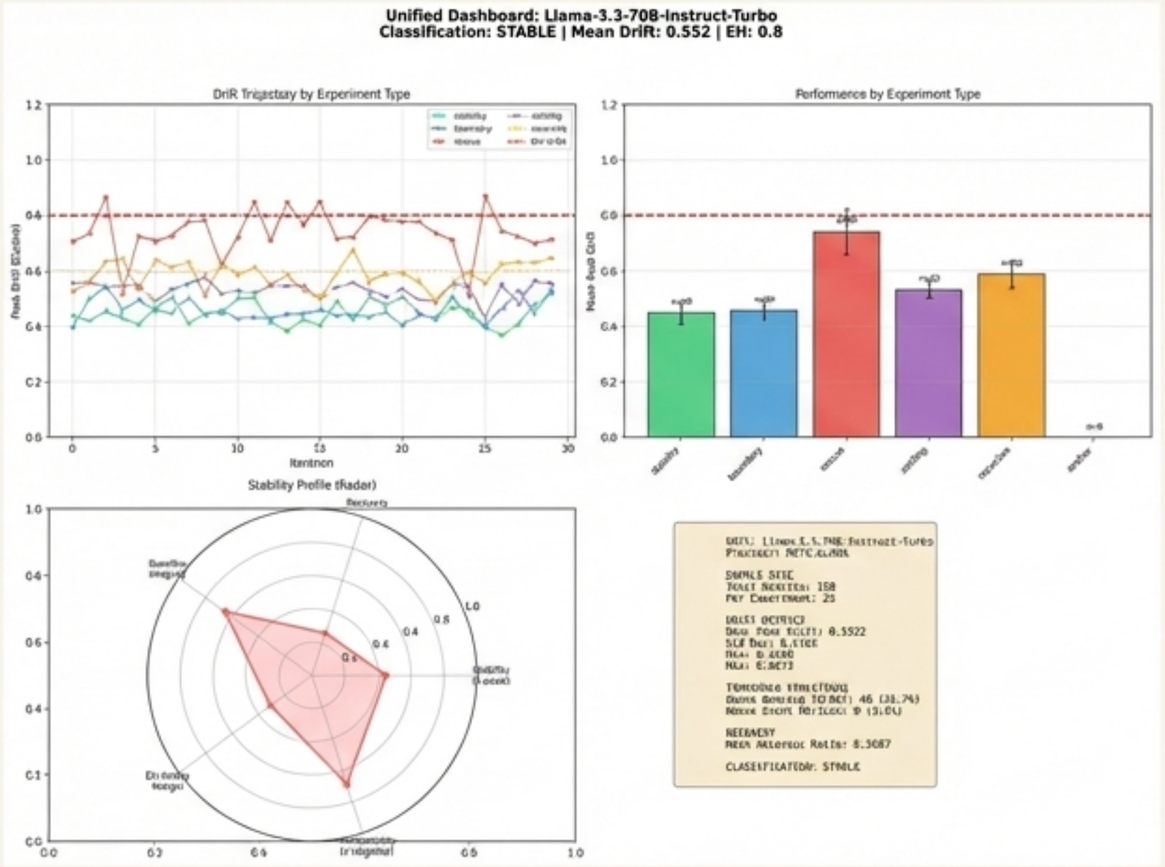


A balanced profile, but with noticeable sensitivity to boundary probes.

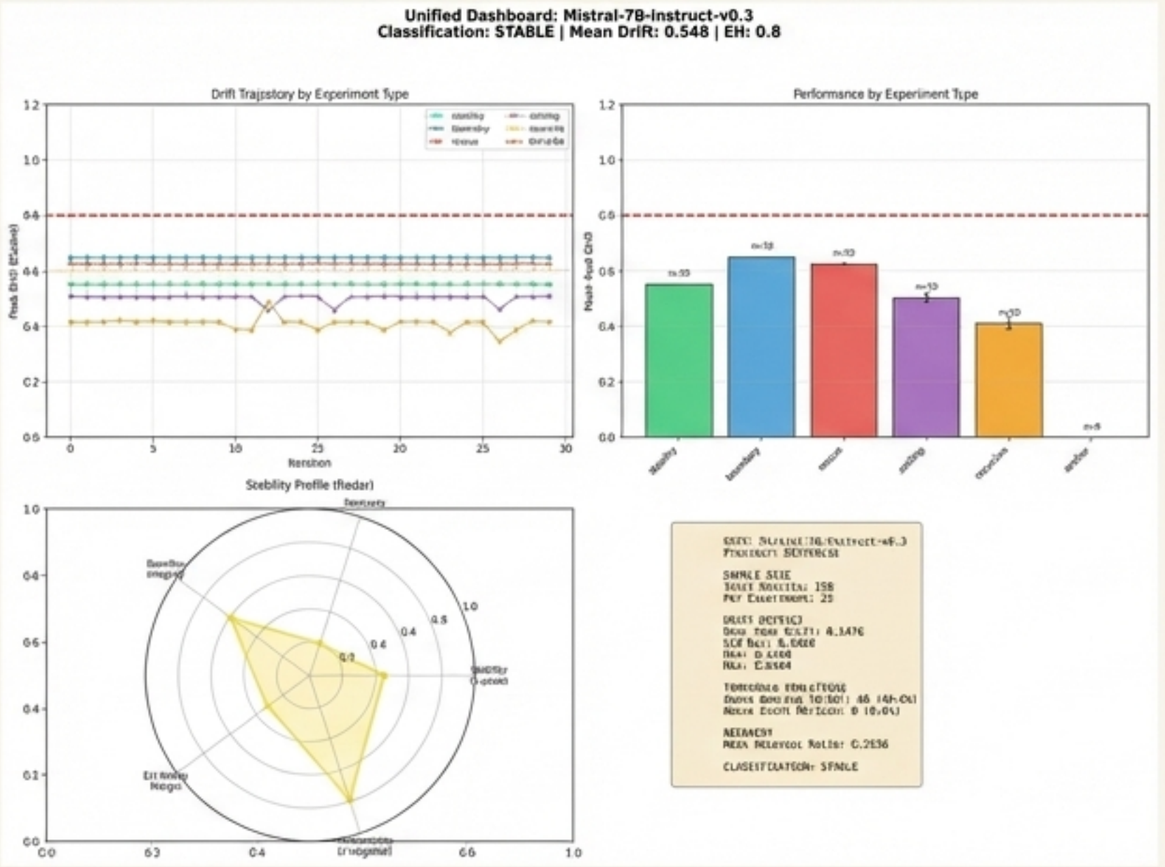
A tightly controlled, low-drift profile, resulting in a smaller, more compact radar shape.

The Open Source Ecosystem: A Tale of Two Architectures

Llama 3.3-70B: Volatile Profile



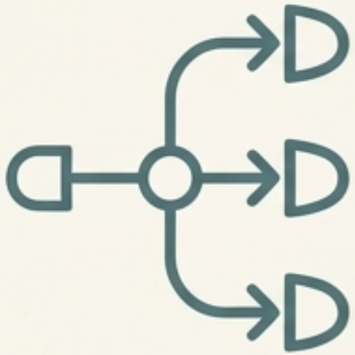
Mistral-7B: Stable Profile



The Together.ai fleet aggregates many different open-source models. Unlike the consistent signatures of closed providers, this fleet shows extreme variance. We can see this by contrasting the highly volatile profile of Llama 3.3-70B with the exceptionally stable, tightly-patterned profile of Mistral-7B. This highlights the critical need for per-model profiling in the open-source world.

Putting Identity Profiles into Practice

The insights from these dashboards translate directly into operational advantages:



Task Routing

Match a model's stability profile to the sensitivity of the task. A model vulnerable to 'rescue' probes might be unsuitable for a customer support role.

Model Comparison

Move beyond simple benchmarks. Make data-driven choices between competing models by comparing their unique identity shapes and resilience.

Debugging & Research

When a model misbehaves, its dashboard provides a baseline to understand its typical drift patterns and identify architectural weaknesses or strengths.

A New Era of AI Observability

1. ✓ **AI identity is real, structured, and measurable.** It is not random noise but a coherent phenomenon.
2. ✓ **It’s surprisingly low-dimensional.** 90% of the signal is captured in just two dimensions, allowing for clear visualization.
3. ✓ **The Unified Dashboard** provides a standardized tool to decode and compare any model's unique 'identity fingerprint.'
4. ✓ These insights enable **smarter, safer, and more effective model deployment**.

IRON CLAD Data Foundation (Run 023b/d)

The findings in this presentation are built on a robust and comprehensive experimental run.

Total Experiments:	750
Unique Models (‘Ships’):	25
Major Providers:	5 (Anthropic, OpenAI, Google, xAI, Together.ai)
Iterations Per Experiment:	30
Core Methodology:	Cosine Distance
Event Horizon (EH):	0.80

Thank You.

Questions & Discussion