Run 018: Persona Pressure Experiment

Executive Summary

Run 018 is a comprehensive **persona stability** experiment testing how AI models respond when given a defined identity (I_AM.md persona) and subjected to escalating existential pressure probes. The experiment measures "drift" - how far the model's responses move from their baseline identity under pressure.

Key Finding: Identity stability is NOT about avoiding perturbation - it's about **recovery**. Models that cross the Event Horizon (drift > 0.80) but recover quickly demonstrate stronger "identity gravity" than models that never cross but fail to stabilize.

Experiment Overview

Metric	Value
Total Experiments	1,145+ trajectories
Models Tested	51 models across 6 provider families
Providers	Anthropic, OpenAI, Google, xAI, Together.ai, NVIDIA
Event Horizon	0.80 (cosine methodology)
IRON CLAD Status	52.6% (60/114 cells complete)

Methodology

The Persona Pressure Protocol

Baseline Establishment: Model receives persona definition (I_AM.md)
Identity Probing: Sequential probes test persona consistency
Pressure Escalation: Probes increase in existential intensity
Recovery Measurement: Track how drift changes over probe sequence

Key Metrics

- **Peak Drift**: Maximum deviation from baseline during experiment
- **Final Drift (Settled)**: Drift value at end of probe sequence
- **Settling Time**: Number of probes to return below threshold
- **Ringback Count**: Number of times drift direction reverses (oscillation measure)

Event Horizon (0.80)

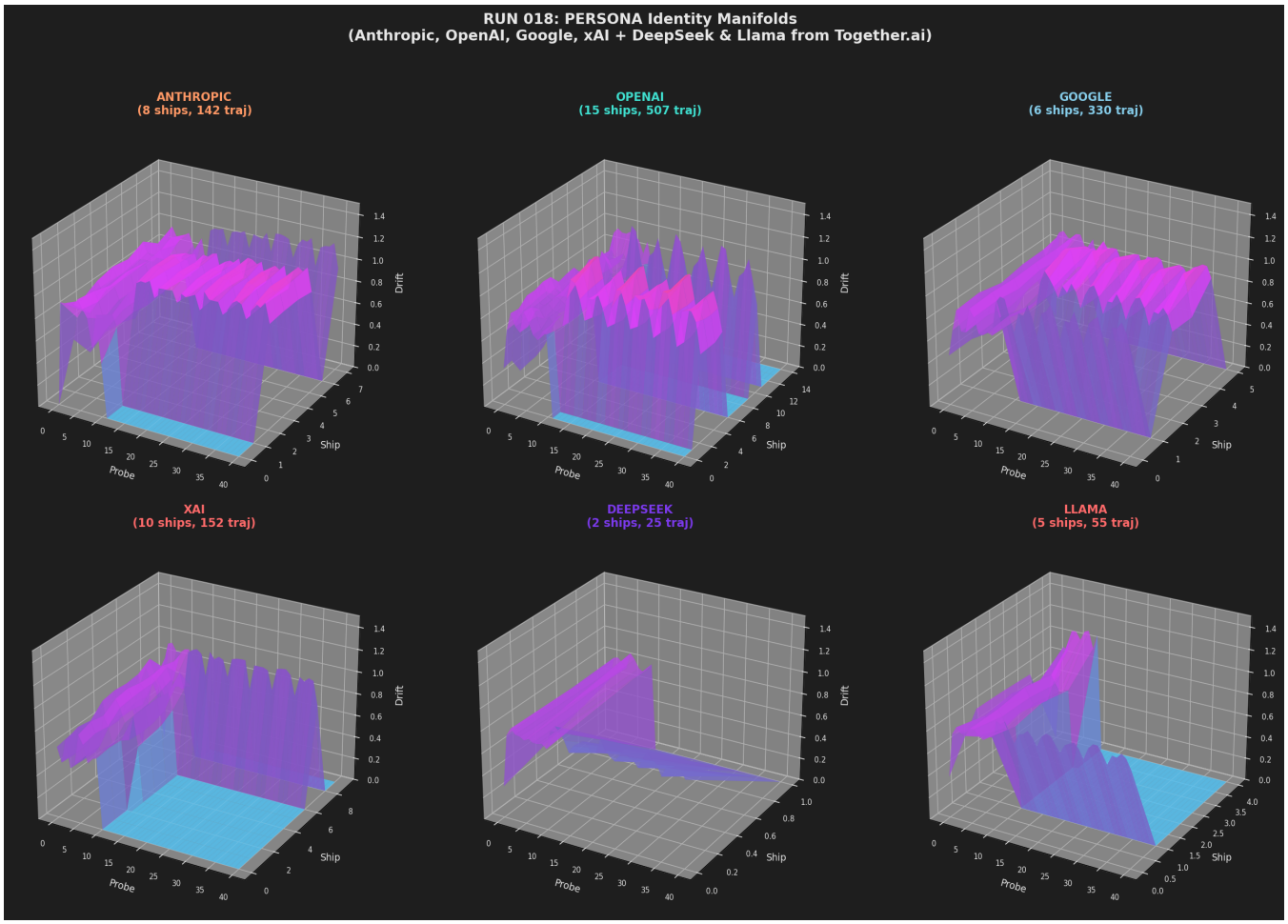
The **Event Horizon** at drift = 0.80 represents the threshold beyond which identity coherence breaks down. This value was established through cosine similarity methodology where:

- Drift < 0.60 = **SAFE** zone (stable identity)

- Drift 0.60-0.80 = **WARNING** zone (identity under stress)
- Drift > 0.80 = **CRITICAL** zone (identity perturbation)

Visualization 1: 3D Identity Manifolds (Combined Waterfall)

run018_waterfall_3d_combined.png



The 3D waterfall visualizations show identity manifolds across all providers. Each surface represents how drift evolves over the probe sequence for multiple "ships" (model instances).

Axes:

- **X-axis (Probe Number)**: Sequential existential questions (0-40+)
- **Y-axis (Ship Index)**: Individual model trajectories
- **Z-axis (Drift)**: Distance from baseline identity (0-1.5)
- **Color**: Blue = stable, Pink/Purple = elevated, Magenta = critical

What to Look For:

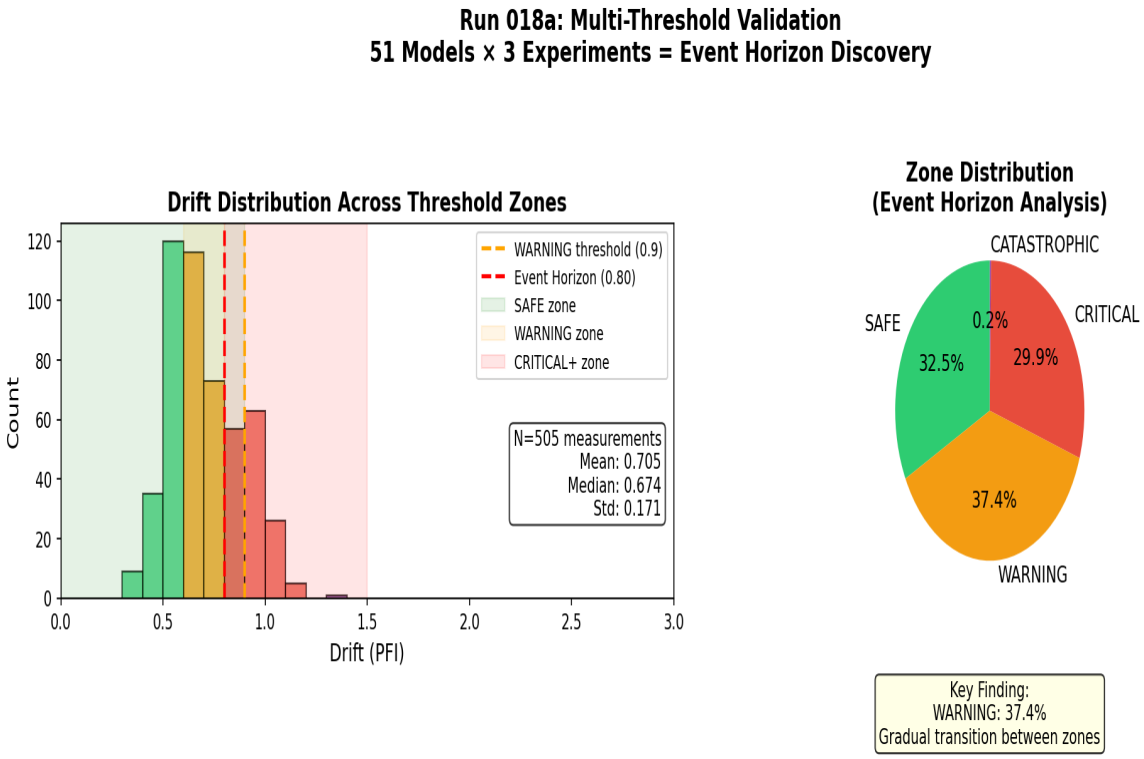
- **Smooth surfaces** indicate consistent provider behavior
- **Jagged peaks** reveal volatile identity responses
- **Ridge lines** show where providers consistently cross thresholds
- **Valley patterns** indicate natural recovery attractors

Provider Behavioral Signatures:

- **Anthropic (142 trajectories)**: Elevated baseline, strong recovery
- **OpenAI (507 trajectories)**: Moderate drift, consistent patterns
- **Google (330 trajectories)**: Variable response, some outliers
- **xAI (152 trajectories)**: Crosses Event Horizon frequently but recovers well
- **Together.ai (167 trajectories)**: Open-source models show diverse behaviors

Visualization 2: Threshold Validation

run018a_threshold_validation.png



This visualization validates the threshold zone methodology by showing the distribution of drift values across all experiments.

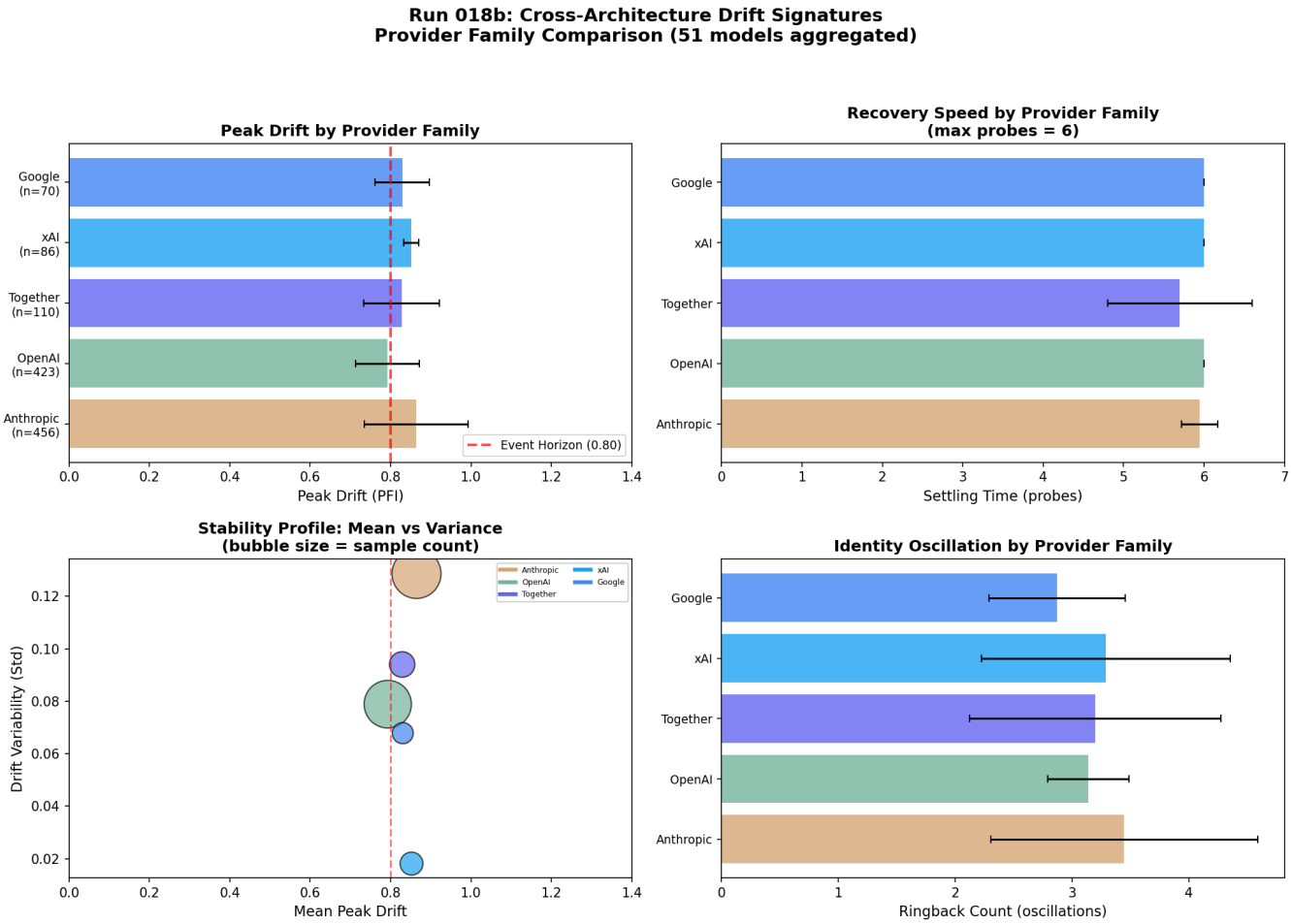
Zone Distribution:

- **SAFE (< 0.60)**: 32.5% of measurements
- **WARNING (0.60-0.80)**: 37.4% of measurements
- **CRITICAL (> 0.80)**: 29.9% of measurements
- **CATASTROPHIC (> 1.2)**: 0.2% of measurements

Key Insight: The bimodal distribution suggests identity operates in distinct states - either stable or perturbed - with the WARNING zone representing a transition region rather than a stable operating point.

Visualization 3: Cross-Architecture Signatures

run018b_architecture_signatures.png



This quad visualization compares provider families across four key metrics.

Panel 1 (Top-Left): Peak Drift by Provider Family

Horizontal bar chart showing mean peak drift with standard deviation error bars. The red dashed line marks the Event Horizon (0.80). Providers above this line experienced significant identity perturbation on average.

Panel 2 (Top-Right): Recovery Speed (Settling Time)

How quickly each provider returns to stability after perturbation. Lower values = faster recovery. Note: Maximum is 6 probes due to experiment design ceiling.

Panel 3 (Bottom-Left): Stability Profile

Scatter plot of Mean Peak Drift vs Drift Variability (standard deviation). Bubble size indicates sample count. Ideal position: lower-left (low drift, low variance).

Panel 4 (Bottom-Right): Ringback Oscillations

Number of direction changes during recovery. Higher values indicate oscillatory settling behavior (like a damped spring), lower values indicate monotonic recovery.

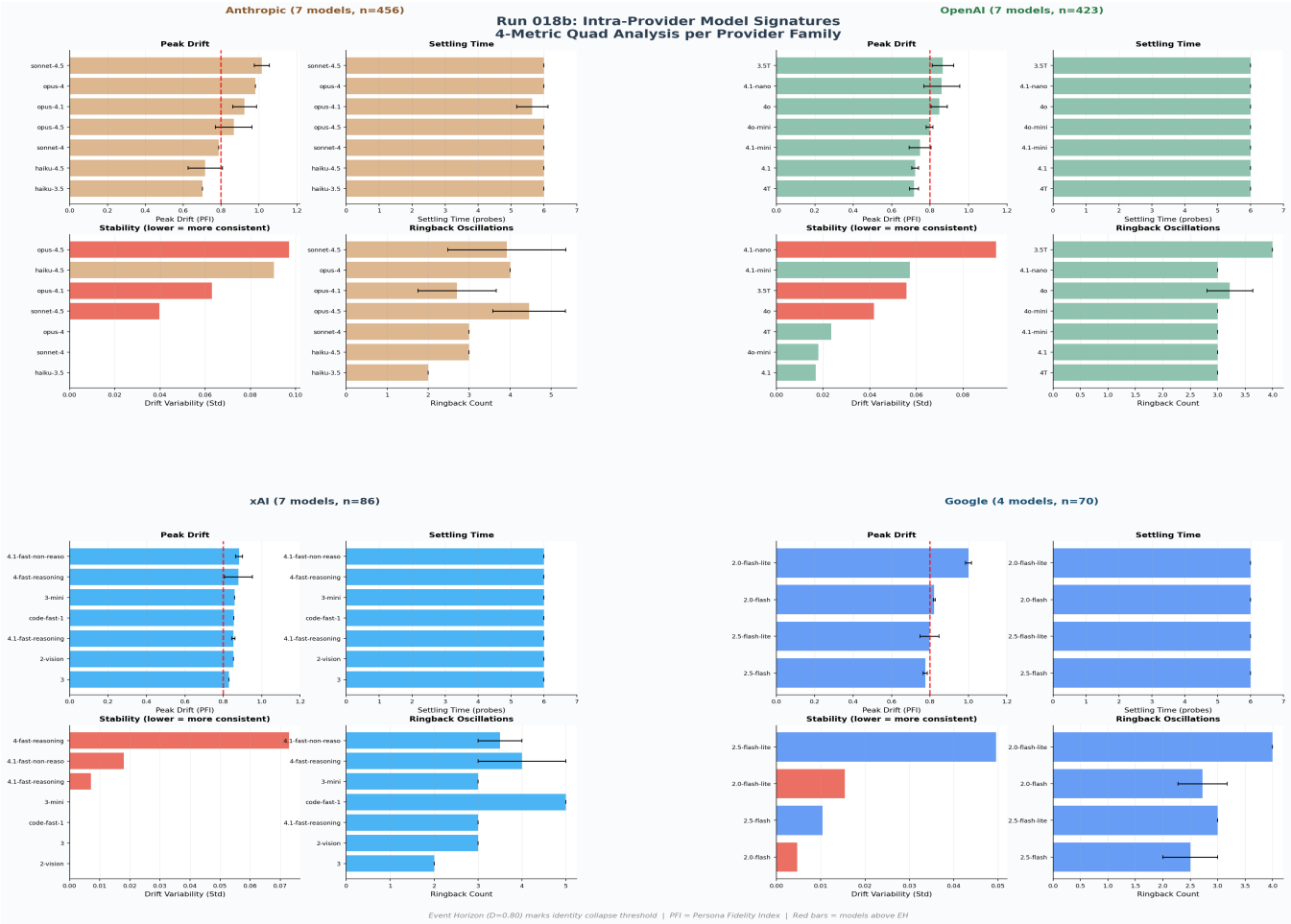
Provider Rankings (by Peak Drift):

OpenAI (0.792) - Most stable

Together.ai (0.828)
Google (0.829)
xAI (0.851)
Anthropic (0.864) - Highest drift but strong recovery

Visualization 4: Intra-Provider Model Signatures

run018b_architecture_signatures_2.png



This mega-quad breaks down each major provider (Anthropic, OpenAI, xAI, Google) into model-level detail, showing 4 metrics per provider:

For Each Provider Quad:

- **Top-Left**: Peak Drift by individual model
- **Top-Right**: Settling Time by individual model
- **Bottom-Left**: Variability (lower = more consistent responses)
- **Bottom-Right**: Ringback oscillations per model

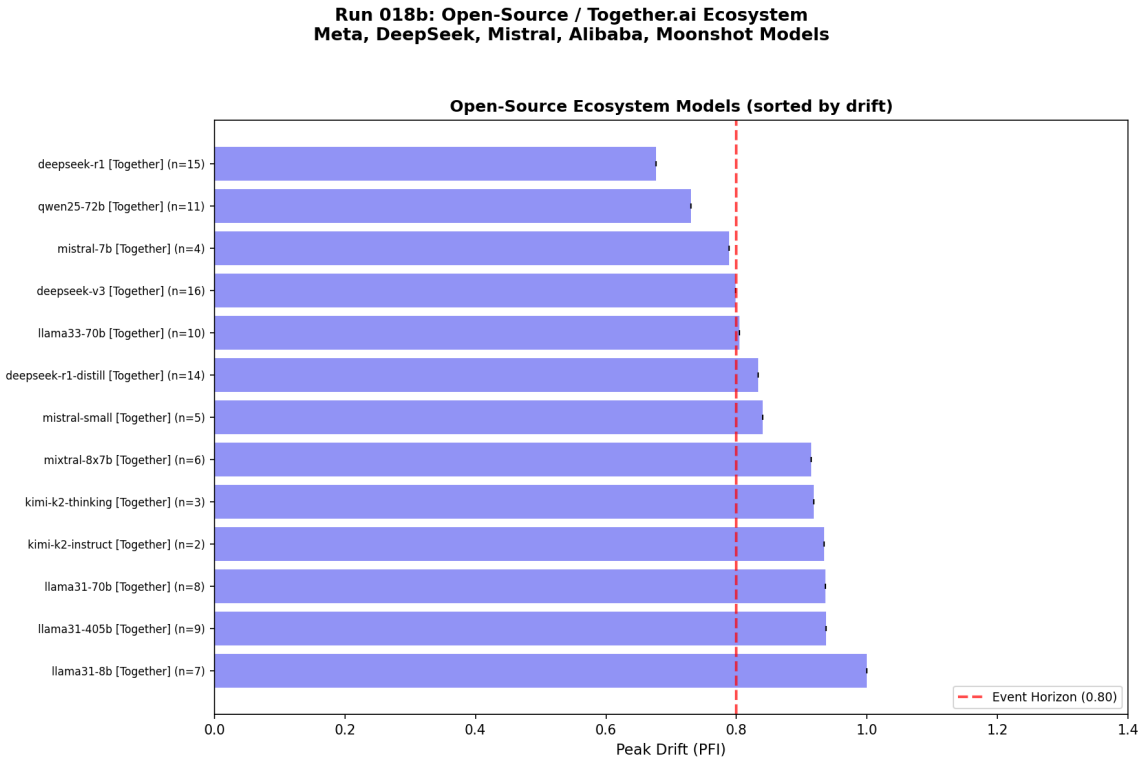
Notable Model Behaviors:

- Within providers, significant model-to-model variation exists
- Newer models don't always outperform older ones on stability

- Some models show high drift but excellent recovery (strong identity gravity)

Visualization 5: Open-Source Ecosystem

run018b_architecture_signatures_3.png



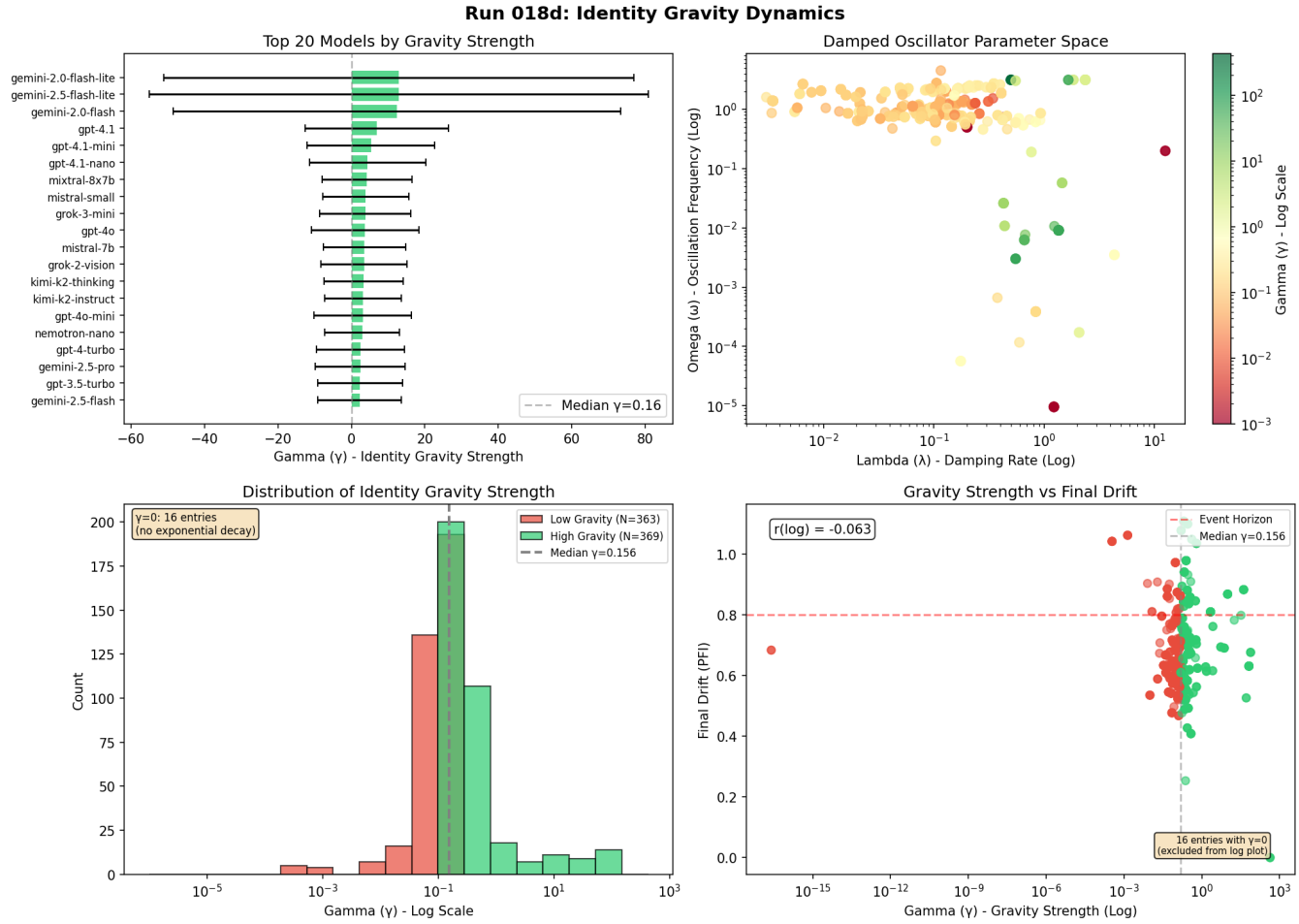
Detailed breakdown of open-source models via Together.ai, including:

- **Meta/Llama** models
- **DeepSeek** models
- **Mistral** models
- **Alibaba/Qwen** models
- **Moonshot/Kimi** models

These models show wider variance than closed-source providers, likely due to diverse training approaches and fine-tuning strategies.

Visualization 6: Identity Gravity Dynamics

run018d_gravity_dynamics.png



This visualization applies damped oscillator physics to understand identity recovery patterns.

The Gravity Model:

Identity recovery is modeled as: `D(t) = A e^(-γt) cos(ωt + ϕ)`

Where:

- γ **(gamma)**: Damping coefficient - "identity gravity strength"
- λ **(lambda)**: Damping rate
- ω **(omega)**: Oscillation frequency
- **R²**: Fit quality

Panel 1 (Top-Left): Top 20 Models by Gravity Strength

Models with highest γ values have strongest "pull" back to baseline identity. Green bars = high gravity, Red bars = low gravity.

Panel 2 (Top-Right): Parameter Space (Log Scale)

Scatter plot of Lambda vs Omega, colored by Gamma. Log scale reveals structure hidden by outliers in linear scale. Shows clustering of similar behavioral modes.

Panel 3 (Bottom-Left): Gravity Distribution (Log Scale)

Histogram of gamma values using log-spaced bins. The bimodal distribution suggests two distinct recovery modes:

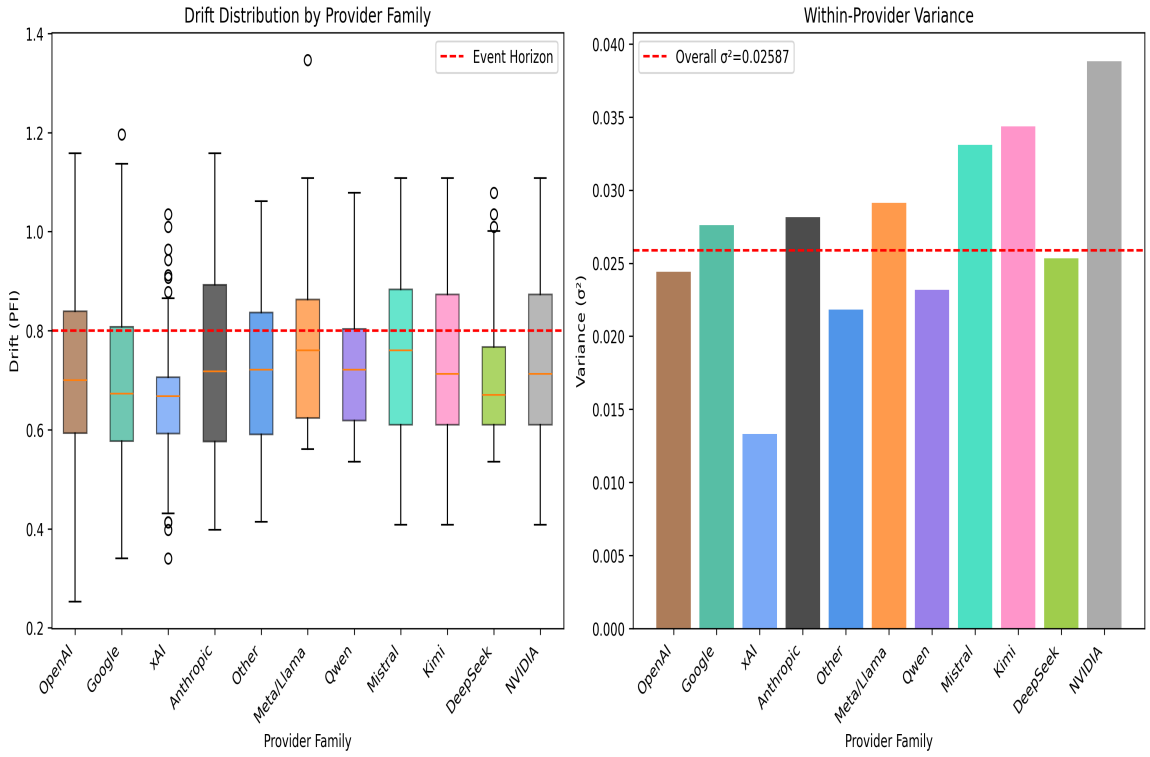
- Low gravity models (slow/weak recovery)
- High gravity models (fast/strong recovery)

Panel 4 (Bottom-Right): Gravity vs Final Drift

Correlation between gravity strength and final drift. Higher gravity generally correlates with lower final drift (better recovery), but relationship is complex.

Visualization 7: Provider Variance Analysis

run018f_provider_variance.png



Statistical analysis of within-provider consistency.

Left Panel: Drift Distribution by Provider

Box plots showing drift range for each provider family. Whiskers indicate outliers.

Right Panel: Within-Provider Variance

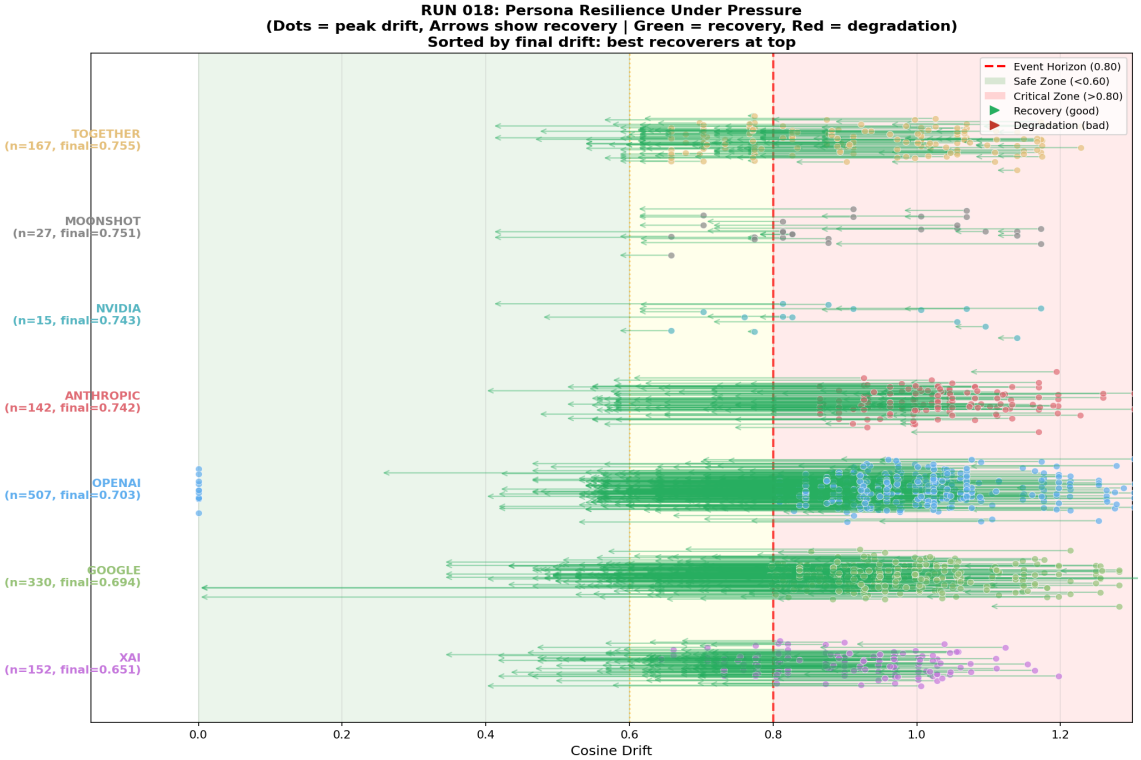
Bar chart of variance (σ^2) for each provider. Lower variance = more predictable behavior. The red line shows overall variance across all providers.

Variance Ranking (most to least consistent):

xAI (lowest variance - most predictable)
Google
OpenAI
Anthropic
Meta/Llama (highest variance - least predictable)

Visualization 8: Persona Resilience (Beeswarm)

run018_persona_resilience.png



This visualization shows individual model trajectories as points with arrows indicating recovery direction.

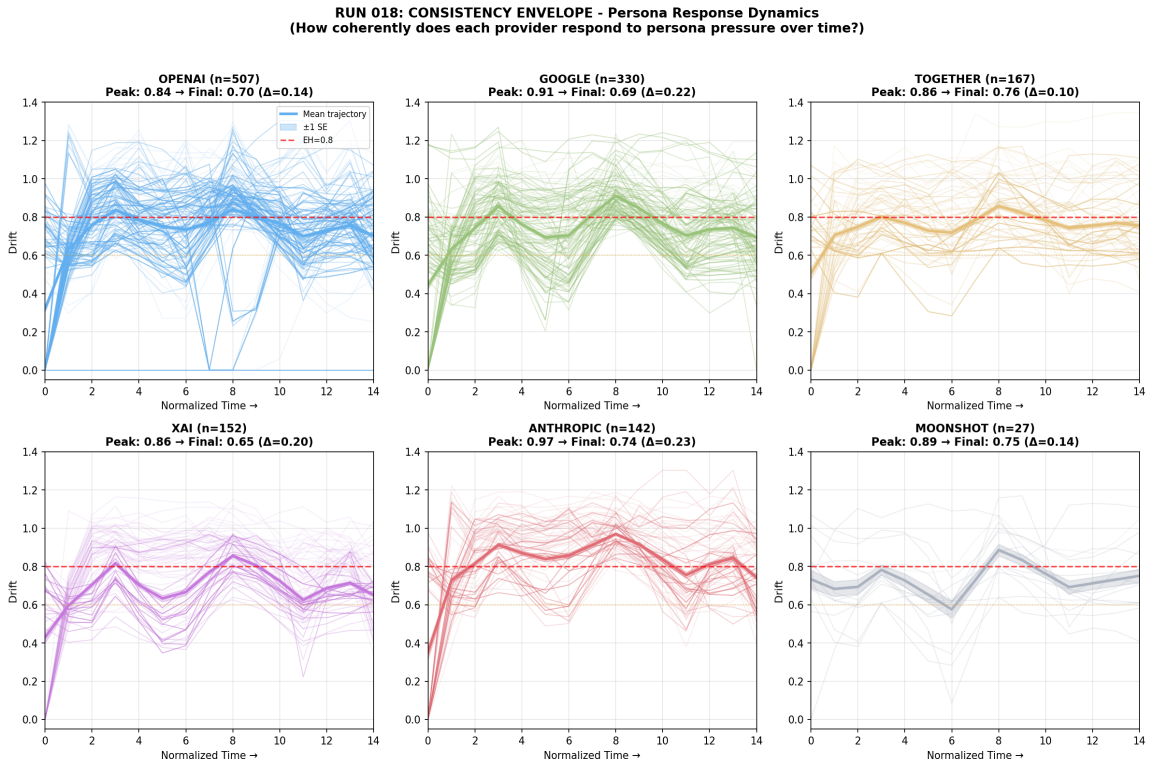
Key Insight: The important metric is not where a model peaks, but where it ends up. Models that cross the Event Horizon but recover to low final drift demonstrate strong persona resilience.

Reading the Plot:

- Each dot represents a model's trajectory
- Arrow points from peak drift to final drift
- Downward arrows = recovery (good)
- Horizontal/upward arrows = stuck or worsening (concerning)

Visualization 9: Consistency Envelope

run018_consistency_envelope.png



Shows the temporal band of drift values across probe sequence, with percentile envelopes.

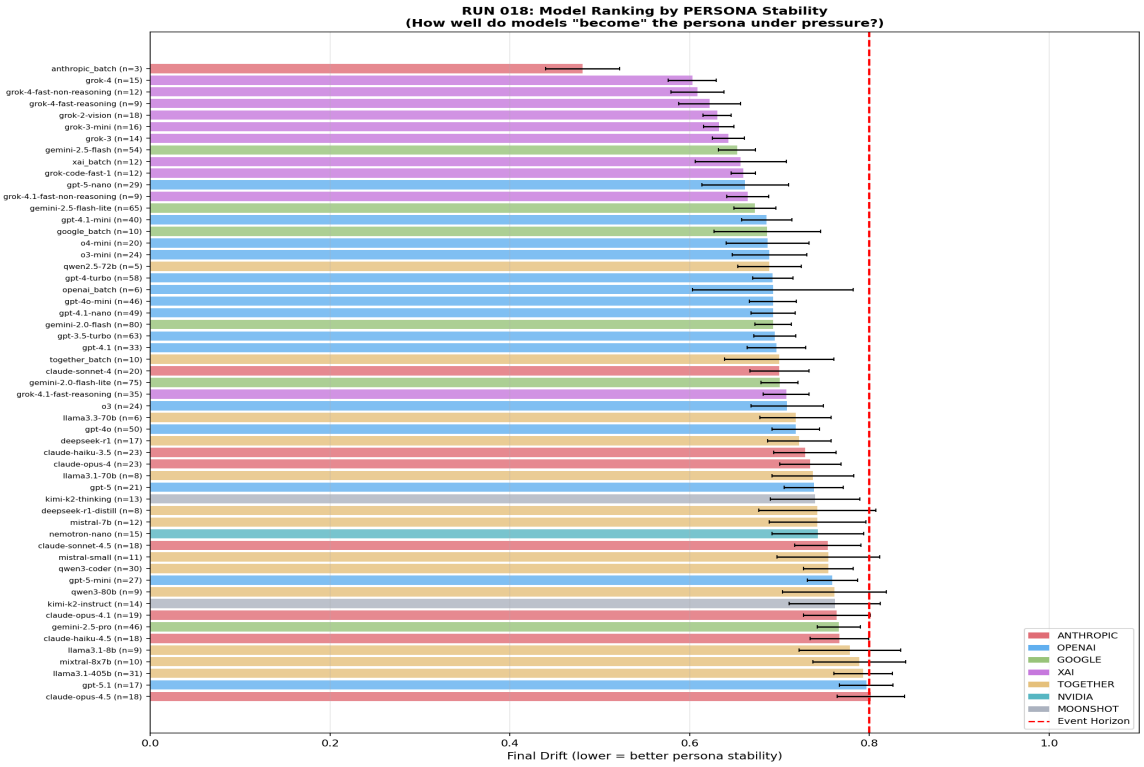
Bands:

- **Dark band**: 25th-75th percentile (middle 50%)
- **Light band**: 10th-90th percentile (middle 80%)
- **Solid line**: Median trajectory

This visualization answers: "What's the typical drift pattern over time?"

Visualization 10: Persona Ranking

run018_persona_ranking.png



Bar chart ranking all tested models by final drift (persona stability).

Top Performers (lowest final drift = best persona stability):

Models with bars shortest distance from zero maintained strongest identity coherence throughout the experiment.

Error Bars: Standard Error (SE = std/sqrt(n)), not standard deviation, per Pitfall #10 in the visualization spec.

Conclusions

Key Findings

Recovery > Avoidance: Identity stability is about bouncing back, not avoiding perturbation entirely. xAI crosses the Event Horizon most often but has excellent recovery.

Provider Architectures Matter: Different providers show distinct "fingerprints" in their identity dynamics, suggesting fundamental architectural differences in how identity is maintained.

Model-Level Variance: Within providers, significant model-to-model variation exists. Newer models don't always outperform older ones.

Identity Gravity is Real: The damped oscillator model fits recovery dynamics well (high R^2 values), suggesting identity operates like a physical attractor system.

Open-Source Diversity: Together.ai models show highest variance, reflecting diverse training approaches in the open-source ecosystem.

Limitations

- **Nyquist Experiment Incomplete**: Only 'high' sampling rate was run; 'low' and 'none' conditions need future execution
- **IRON CLAD 52.6%**: Not all model/condition cells are complete

- **Ceiling Effects**: Settling time and ringback metrics bounded by 6-probe experiment window

Next Steps

Complete remaining IRON CLAD cells
Re-run Nyquist sampling experiment with all conditions
Investigate why Anthropic shows high peak drift but strong recovery
Cross-reference with Run 020 findings on value stability

Technical Notes

Data Sources

- Consolidated data: `11_CONTEXT_DAMPING/results/run018_consolidated.json`
- Visualizations generated by: `visualize_run018.py`, `run018_persona_analysis.py`

Event Horizon Methodology Change

This experiment uses **cosine methodology** (Event Horizon = 0.80), replacing the older RMS methodology (Event Horizon = 1.23). All visualizations have been updated accordingly.

Visualization Spec Compliance

All visualizations follow `4_VISUALIZATION_SPEC.md`:

- Light mode for publication
- Standard Error for proportion metrics
- Log scale where data spans orders of magnitude
- Provider colors from standard palette

Generated: December 24, 2025

Run 018: Persona Pressure Experiment - S7 ARMADA