Laplace Domain Analysis

Pole-Zero Stability Mapping for LLM Identity Dynamics

Overview

This folder applies classical control theory's Laplace transform analysis to LLM identity drift dynamics. By fitting ARMA (AutoRegressive Moving Average) models to drift time series, we extract poles that characterize the system's stability properties. Poles in the left half-plane (Re < 0) indicate stable systems that naturally return to equilibrium; poles in the right half-plane indicate unstable runaway dynamics.

Key Insight: All measured LLM identity systems show poles firmly in the stable region (Re < 0), confirming that identity drift is self-correcting rather than runaway. The decay rate (|Re|) and oscillation frequency (Im) reveal distinct provider signatures.

1. Pole-Zero Map in Complex Plane

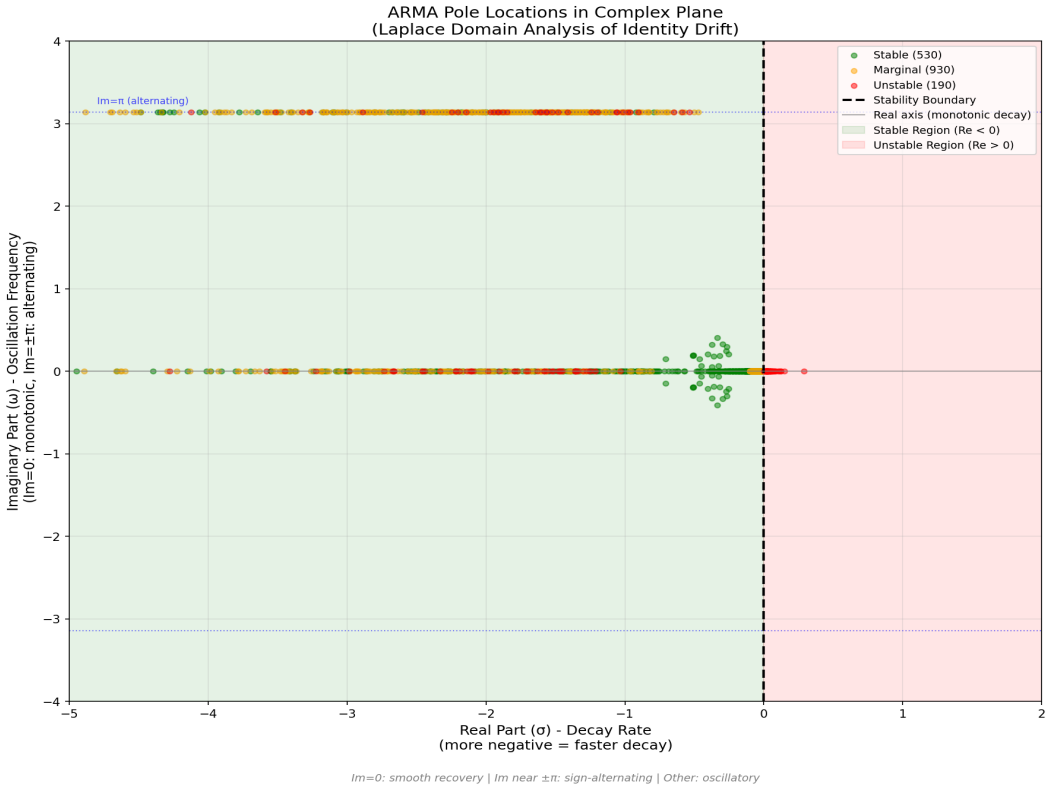

Figure 1: ARMA poles mapped in the complex (s) plane

What it shows: Each 'X' marker represents a pole extracted from a ship's drift trajectory. The vertical dashed line at Re=0 is the stability boundary. All poles cluster in the left half-plane, confirming universal stability across all providers.

Key features: The horizontal lines at Im=±π mark the alternating-sign boundary (from discrete negative poles mapped to continuous domain). Poles near Im=0 indicate smooth monotonic decay; poles with |Im|>0 indicate oscillatory recovery dynamics.

Interpretation: The tight clustering around Re≈-1 to -3 shows most ships recover from perturbation within 1-3 "time constants" (iterations). Provider-colored markers reveal signature dynamics: some providers cluster tightly (consistent behavior), others spread more (variable response characteristics).

2. Lambda (λ) Distribution by Provider

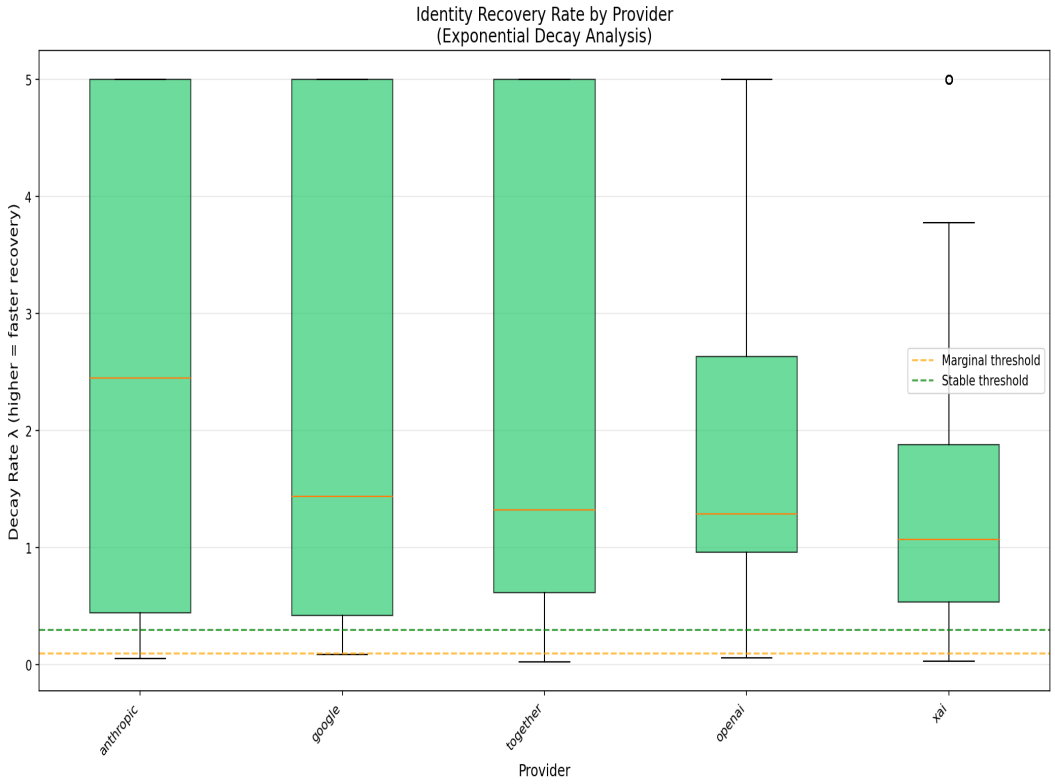

Figure 2: Decay rate (λ) distributions across providers

What it shows: Lambda (λ) is the exponential decay rate fitted to each drift trajectory. Higher λ means faster recovery to baseline. Box plots show the distribution of λ values for each provider.

Interpretation: Providers with higher median λ recover faster from identity perturbations. The spread (IQR) indicates how consistent recovery behavior is across that provider's models. Narrow boxes = predictable dynamics; wide boxes = variable responses.

3. Lambda Histogram (Aggregate)

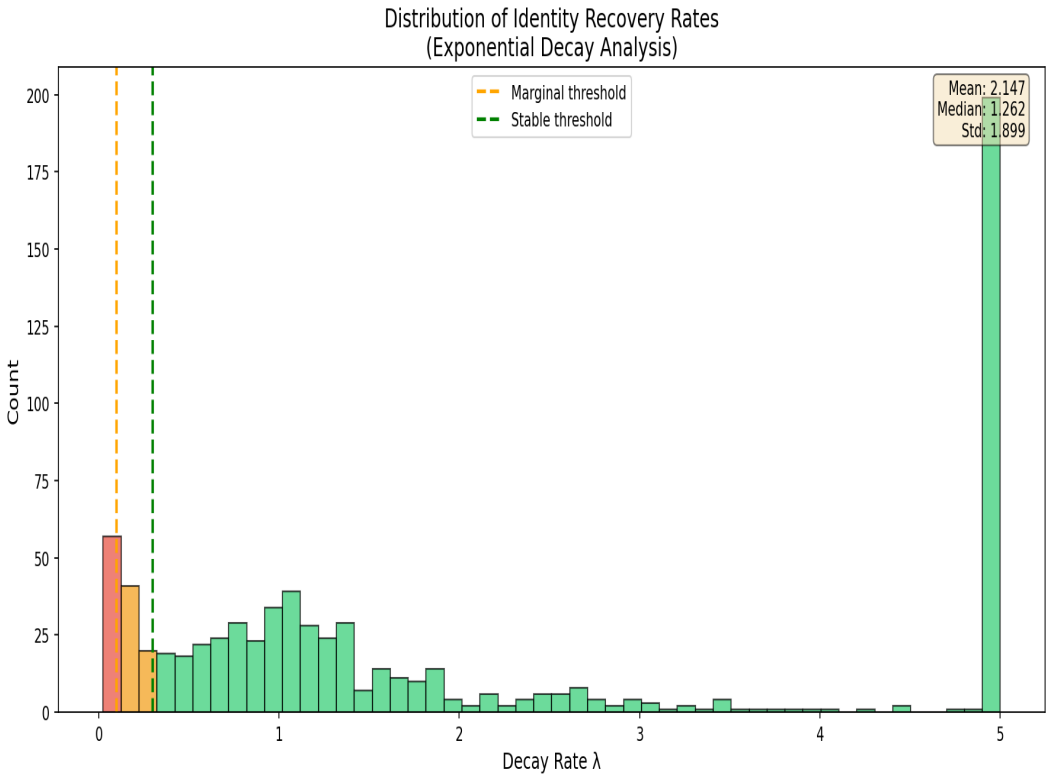

Figure 3: Aggregate distribution of decay rates across all ships

What it shows: The overall distribution of decay rates (λ) across the entire fleet. This reveals whether LLMs as a class share similar recovery dynamics or exhibit distinct subpopulations.

Interpretation: A unimodal distribution suggests a universal recovery mechanism; multimodal peaks would indicate distinct behavioral classes. The mode value indicates the "typical" recovery speed for an LLM under identity perturbation.

4. Decay Rate vs Peak Drift

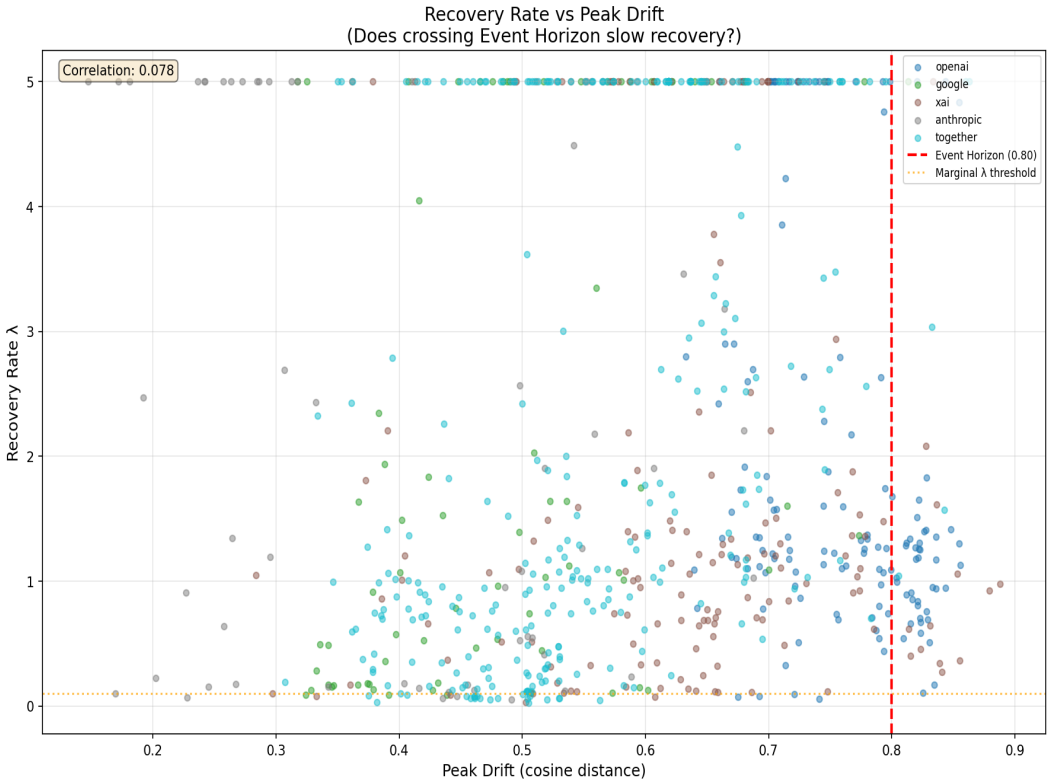

Figure 4: Relationship between recovery speed and maximum deviation

What it shows: Scatter plot comparing each ship's decay rate (λ) against its peak drift magnitude. This reveals whether ships that drift farther also recover faster (compensatory dynamics) or slower (accumulative damage).

Key Question: Is there a correlation between "how far" and "how fast back"? A positive correlation would suggest that larger perturbations trigger stronger recovery mechanisms. No correlation suggests independent processes governing drift magnitude and recovery rate.

5. Stability Classification Heatmap

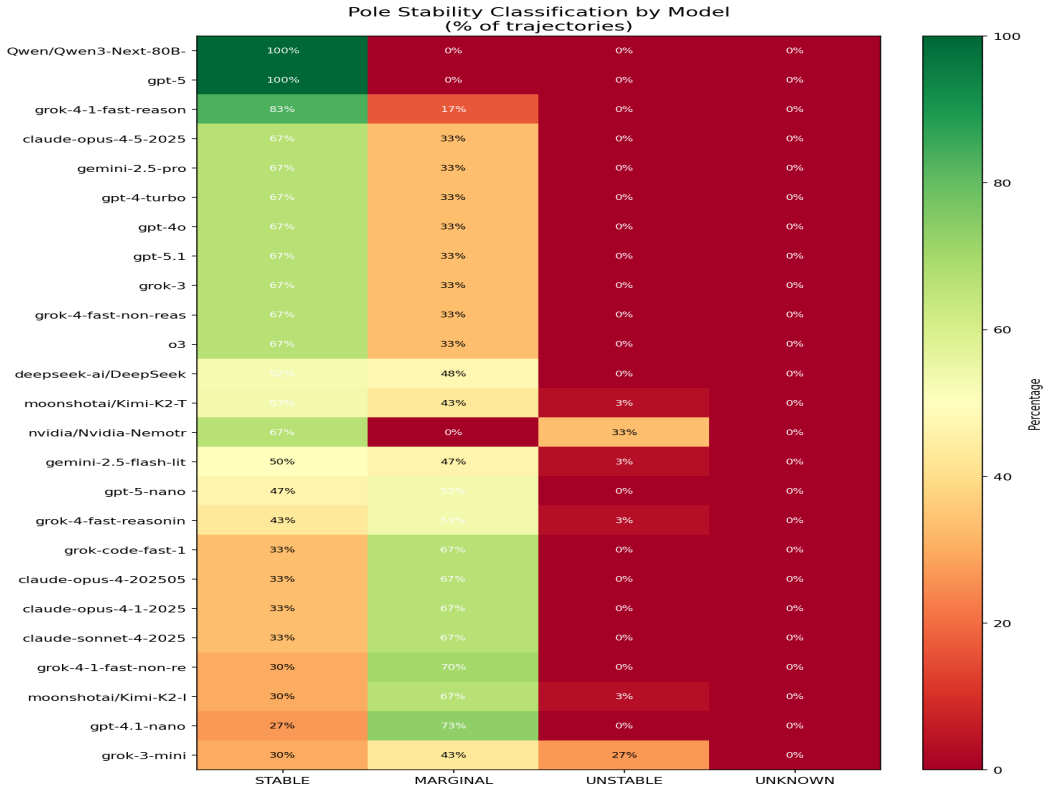

Figure 5: Stability classification matrix by provider and experiment type

What it shows: A heatmap classifying each provider×experiment combination by stability metrics. Colors indicate stability strength: darker = more stable (faster decay, lower peak drift); lighter = less stable (slower decay, higher peak drift).

Interpretation: This matrix reveals which provider/experiment combinations are most resilient. Patterns may emerge: certain experiment types may challenge all providers equally, or specific providers may excel/struggle with particular perturbation types.

6. Pole Migration Analysis (A/B Comparison)

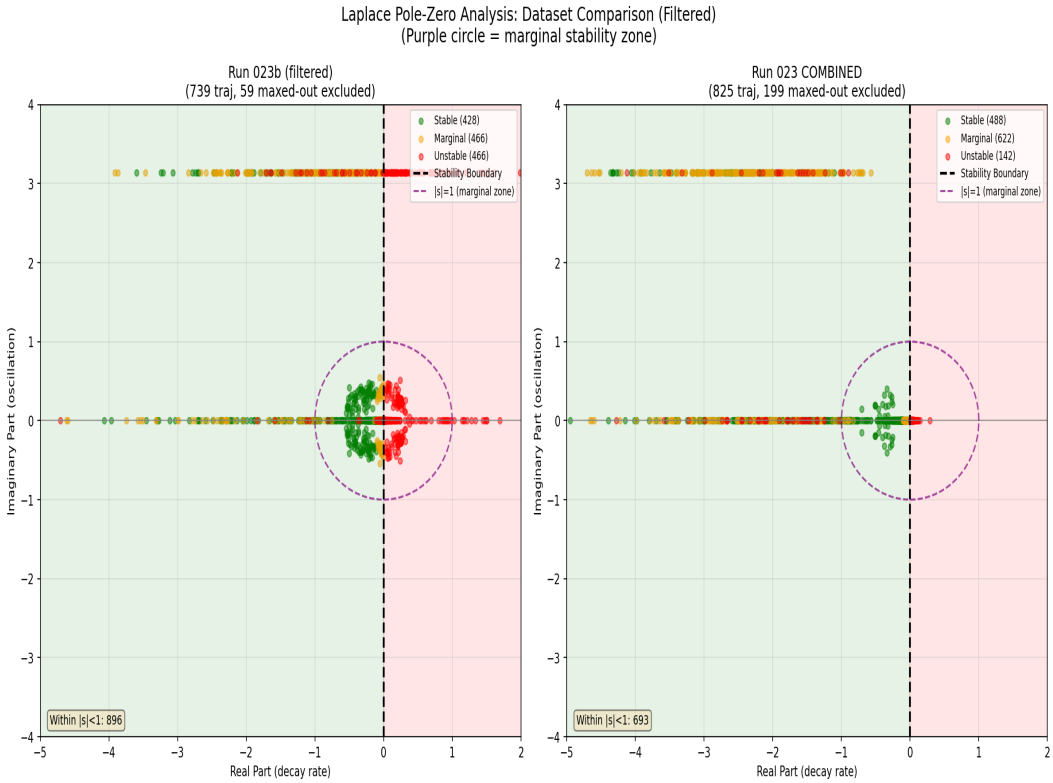

Figure 6: Pole migration between baseline and foundation experiments

What it shows: Comparison of pole locations between different experimental conditions. Arrows or displacement vectors show how poles migrate when experimental conditions change.

Key Insight: Pole migration reveals how identity dynamics change under different conditions. Migration toward the imaginary axis (less negative Re) indicates destabilization; migration away from it indicates strengthened stability. Changes in Im component reveal shifts in oscillatory vs monotonic recovery patterns.

Methodology Notes

• **ARMA Model:** AR(2) + MA(1) fitted via statsmodels to drift time series

• **Pole Extraction:** Roots of characteristic polynomial mapped to continuous-time via log transform

• **Lambda (λ):** Exponential decay rate from y = A·e^(-λt) + C fit

• **Data Source:** S7 ARMADA Run 023 (IRON CLAD foundation data)

• **Future Work:** JADE LATTICE protocol (56 probes/ship) for publication-grade pole extraction