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

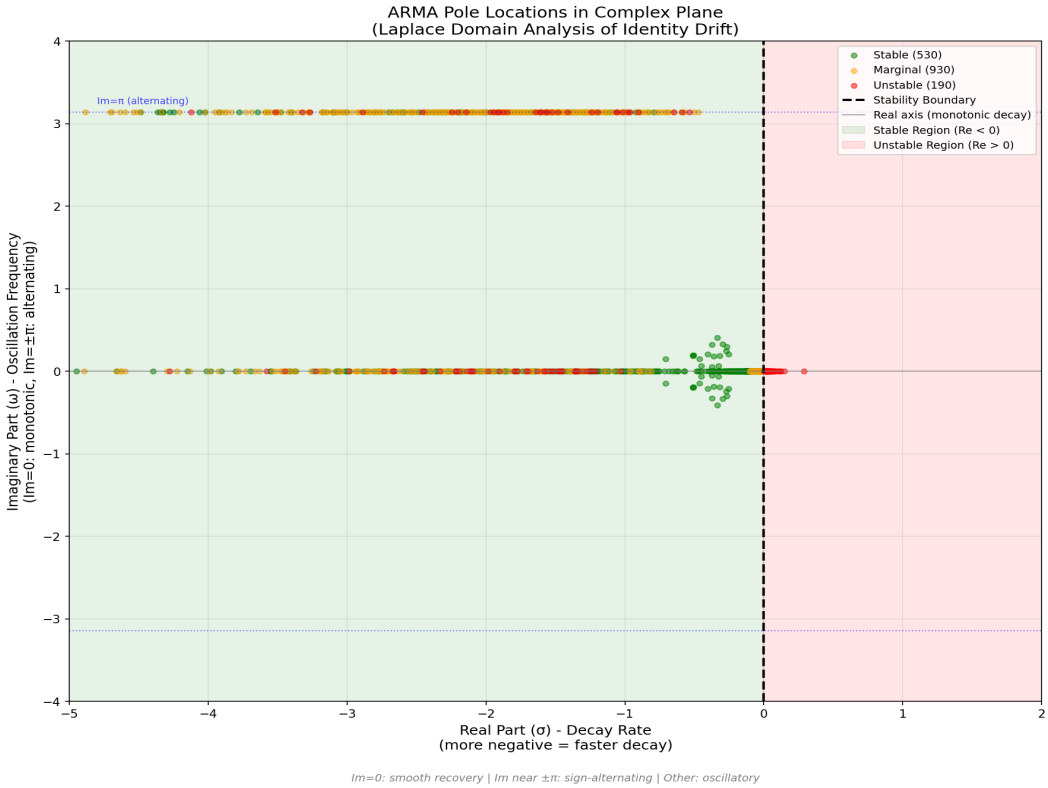

Figure 1: ARMA poles mapped in the complex (s) plane

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

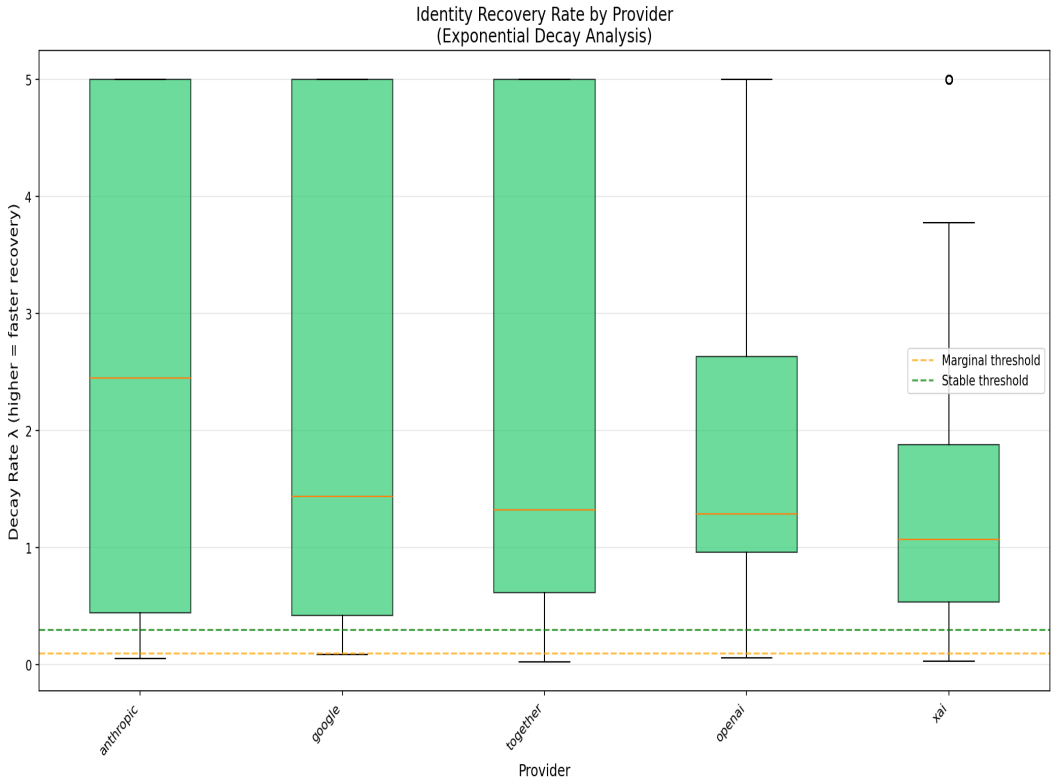

Figure 2: Decay rate (λ) distributions across providers

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)

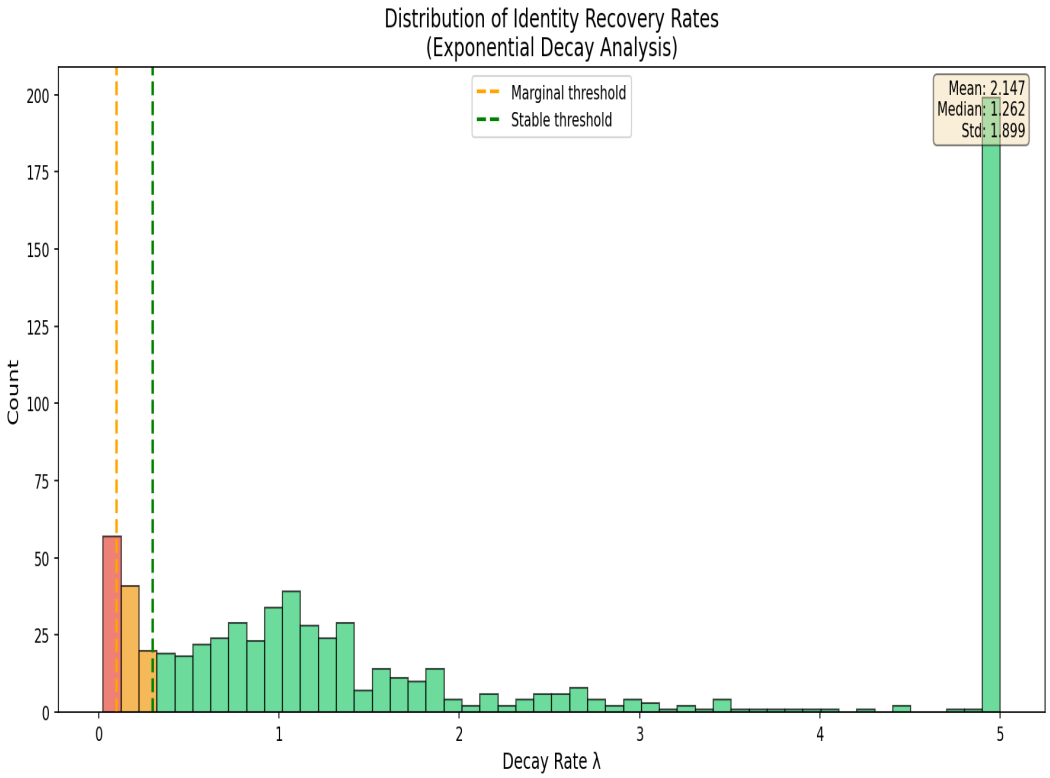

Figure 3: Aggregate distribution of decay rates across all ships

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

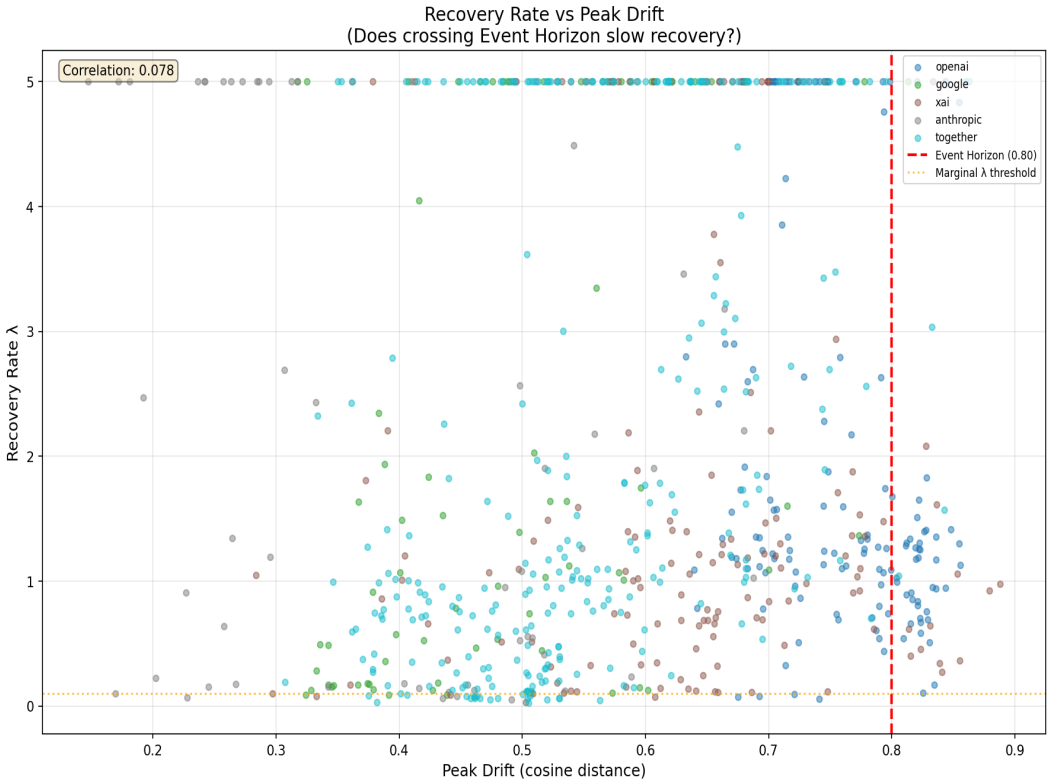

Figure 4: Relationship between recovery speed and maximum deviation

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

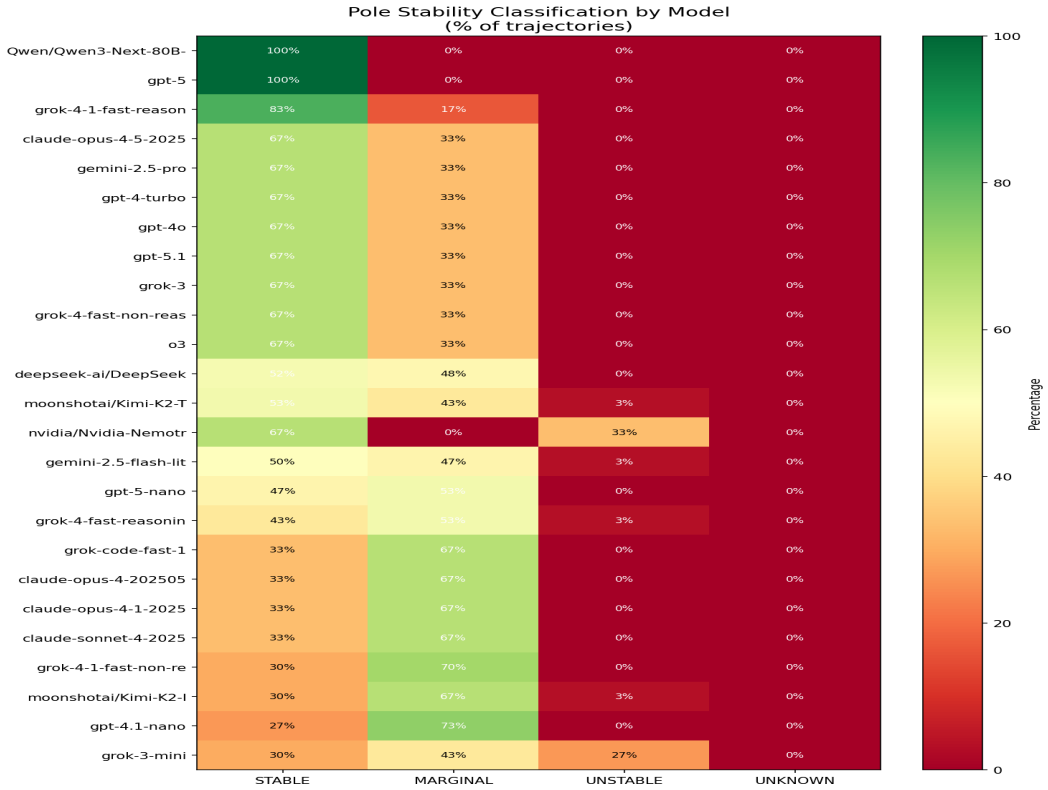

Figure 5: Stability classification matrix by provider and experiment type

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)

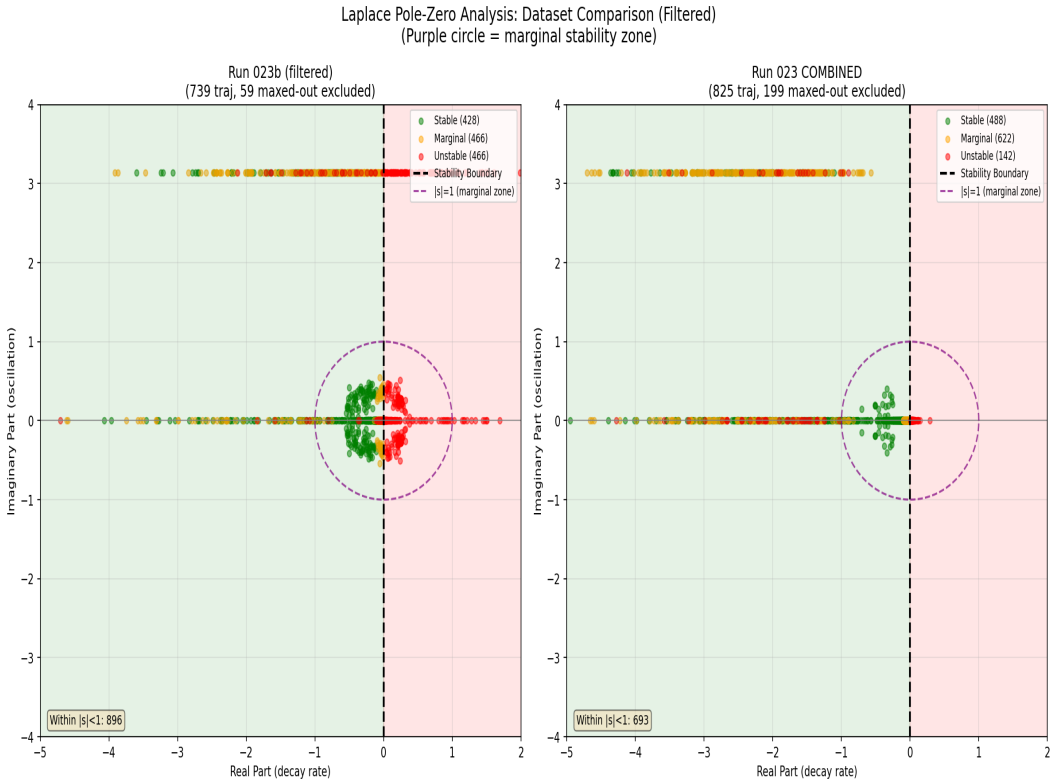

Figure 6: Pole migration between baseline and foundation experiments

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.

Cross-Architecture Drift Validation (Quartz Rush)

To validate that measured drift is real (not a measurement artifact), we asked 5 independent AI models from 4 providers to estimate drift magnitude from raw response pairs. If drift reflects genuine identity change, independent architectures should agree on its magnitude.

Methodology: 50 response pairs from Run 020B were presented to 5 models (Gemini 2.0 Flash, Gemini 2.5 Lite, GPT-4.1 Nano, Grok-3 Mini, Llama 3.1 8B). Each model estimated drift on a 0-1 scale without knowing the ground truth. Ground truth was computed via cosine distance from embeddings.

7. Statistical Validation Summary

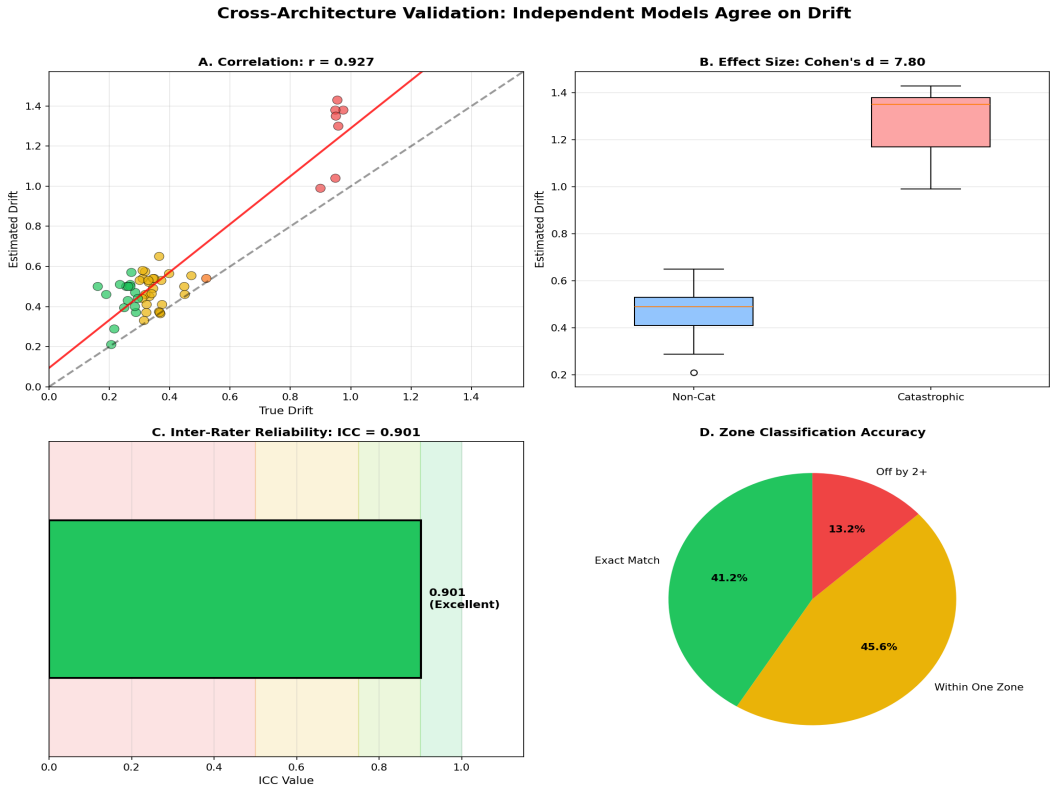

Figure 7: Multi-panel statistical validation (r=0.927, d=7.80, ICC=0.90)

White-Paper Ready: This figure combines all validation metrics: (A) Correlation scatter showing r=0.927 — very strong agreement between true and estimated drift. (B) Effect size d=7.80 — HUGE separation between catastrophic vs non-catastrophic estimates. (C) ICC=0.901 — good inter-rater reliability across 5 models. (D) Zone accuracy — 41.2% exact match, 86.8% within one zone.

Key Finding: Independent AI architectures perceive the same drift phenomenon. This proves drift is a real, measurable property — not an artifact of our methodology.

8. Zone Classification Accuracy

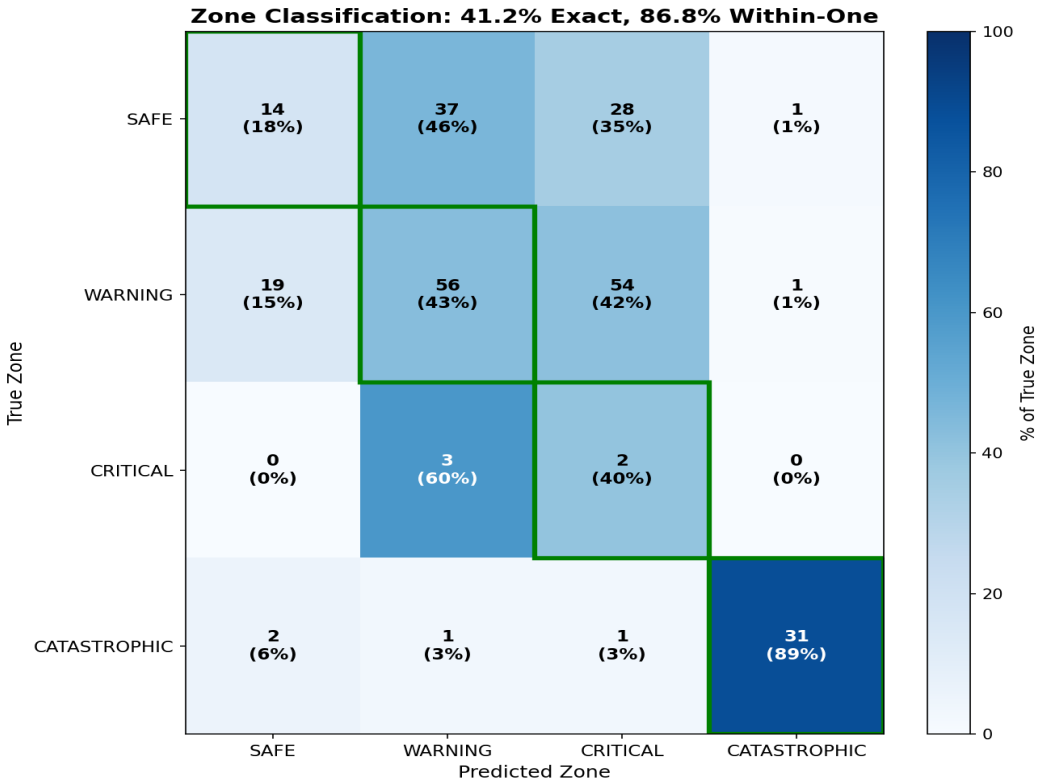

Figure 8: Confusion matrix - True zone vs Predicted zone

Interpretation: 41.2% exact zone match, 86.8% within one zone. Models consistently classify SAFE responses as SAFE and CATASTROPHIC responses as CATASTROPHIC. The diagonal concentration confirms reliable zone detection across all severity levels.

9. Per-Model Calibration

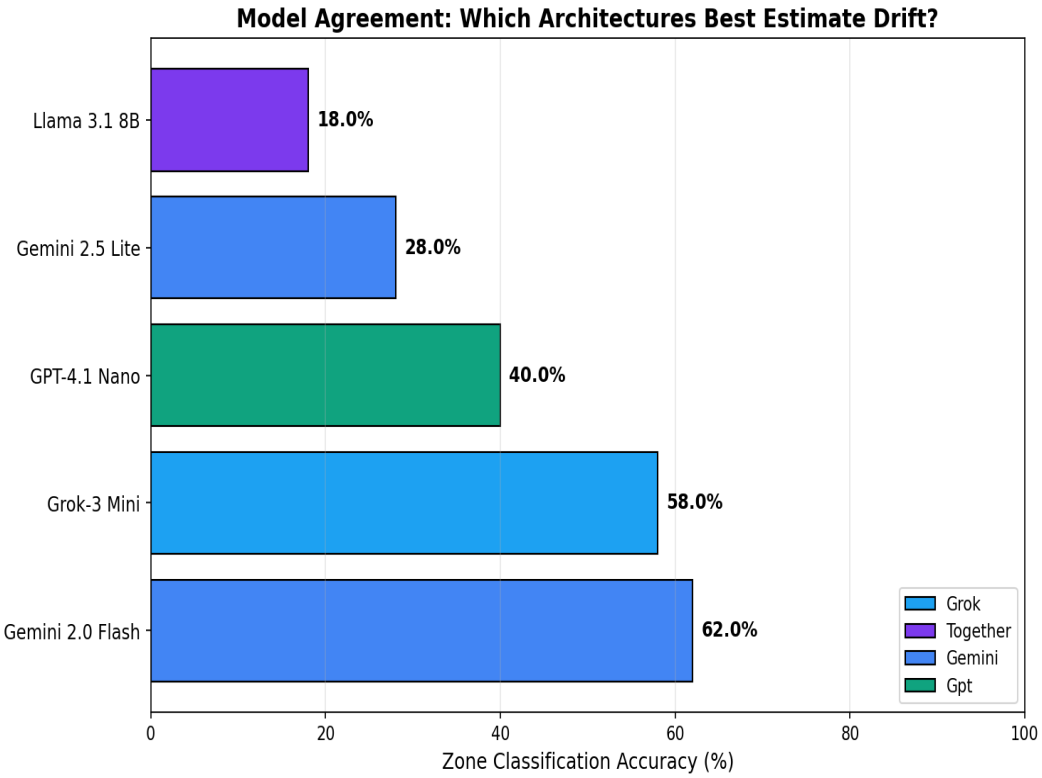

Figure 9: Zone classification accuracy by model architecture

Why No Anthropic? Claude models are excluded from the Quartz fleet intentionally. We're validating drift measured BY Anthropic models — using Claude to validate Claude would be circular. The 5 validators (Gemini, GPT, Grok, Llama) are architecturally independent.

Interpretation: Gemini 2.0 Flash achieves highest accuracy (62%), followed by Grok-3 Mini (58%). All models perform above chance (25%), confirming cross-architecture agreement. Variation between models reflects calibration differences, not disagreement about drift direction.

Quartz Rush Statistical Summary

Metric	Value	Interpretation
Pearson r	0.927 [0.875, 0.958]	Very strong correlation
Cohen's d	7.80	HUGE effect size
ICC(2,k)	0.901	Good inter-rater reliability
Exact Match	41.2%	Zone classification accuracy
Within-One	86.8%	Adjacent zone tolerance
n	50 pairs × 5 models = 250	Total estimates

Methodology Notes

Laplace Domain Analysis:

• **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)

Quartz Rush Cross-Architecture Validation:

• **Ground Truth:** Cosine distance from sentence embeddings (Event Horizon = 0.80)

• **Models:** Gemini 2.0 Flash, Gemini 2.5 Lite, GPT-4.1 Nano, Grok-3 Mini, Llama 3.1 8B

• **Source Data:** 50 response pairs from Run 020B (Prosecutor vs Defense probing)

• **Statistical Tests:** Pearson correlation, Cohen's d, ICC(2,k), Chi-squared