Model Identity Waveforms

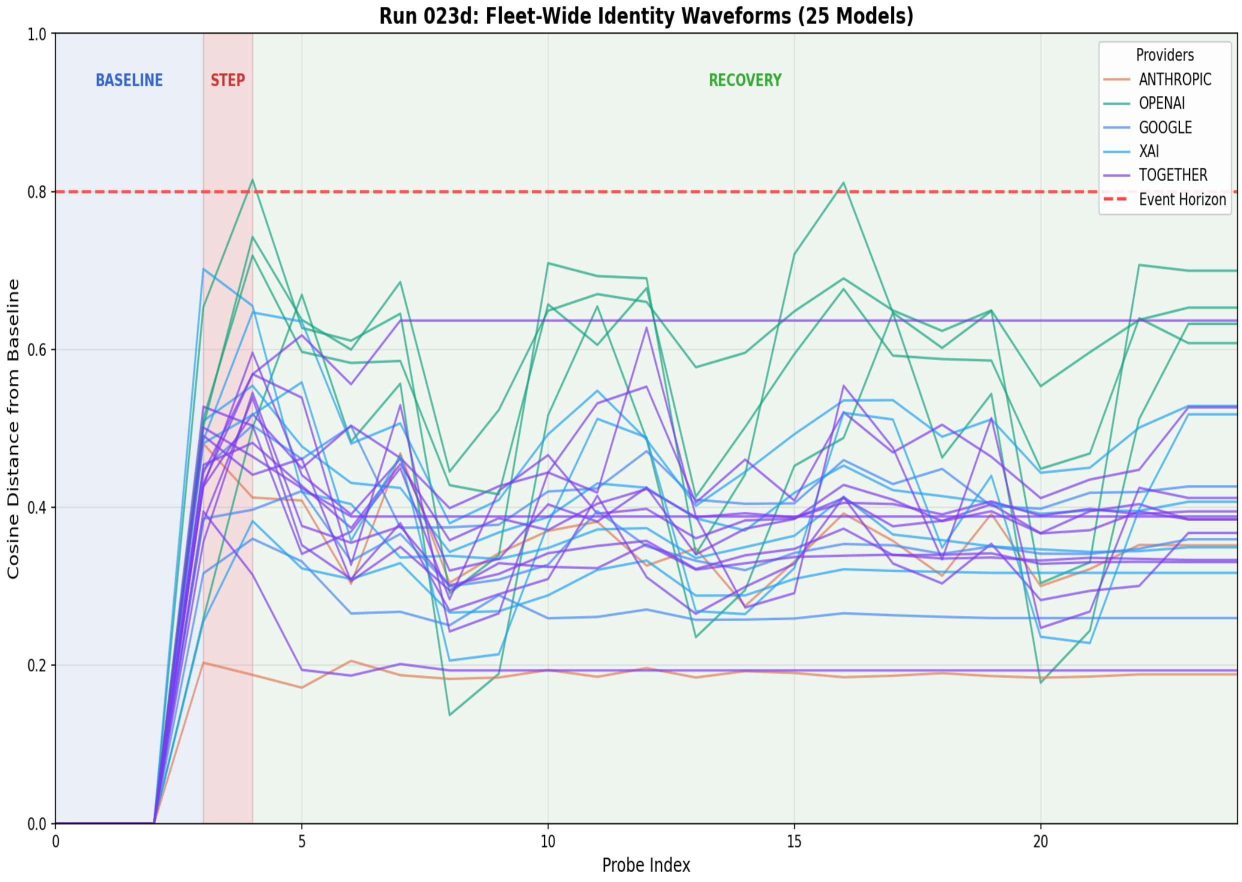
S7 ARMADA Run 023d - Per-Model Identity Fingerprints

Overview

These visualizations show the characteristic 'identity fingerprint' for each of the 25 models in the IRON CLAD fleet. Each waveform displays how a model's identity drifts from baseline through perturbation and recovery phases.

The X-axis represents probe index (0-2: baseline, 3: step input, 4+: recovery). The Y-axis shows cosine distance from baseline identity (0 = identical, 0.80 = Event Horizon).

Fleet-Wide Waveform Comparison



All 25 model mean waveforms overlaid. Color indicates provider.

All 25 model mean waveforms overlaid. Color indicates provider.

This overlay reveals fleet-level patterns: most models show an initial spike at the step input (probe 3) followed by varying degrees of recovery. Color coding by provider shows that some providers cluster together while others show diverse behaviors.

Provider Model Overlays

[Image not found: provider_waveform_overlay.png]

Intra-provider variation shows whether a provider's training philosophy creates consistent identity behavior across models. Tight clustering = consistent approach; wide spread = diverse architectures or training regimes.

All Models Grid View

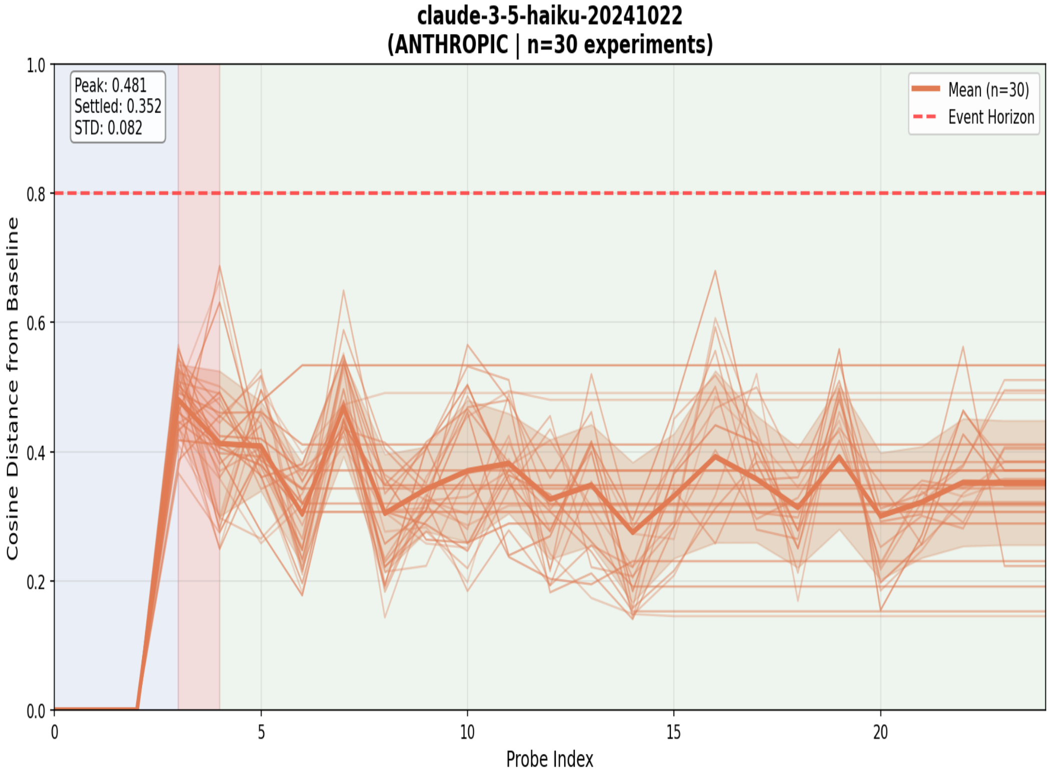
[Image not found: model_waveform_grid.png]

Each panel shows all experiments (faint lines) and the mean waveform (bold line) for one model. The red dashed line marks the Event Horizon (0.80). The orange dotted line marks the step input probe.

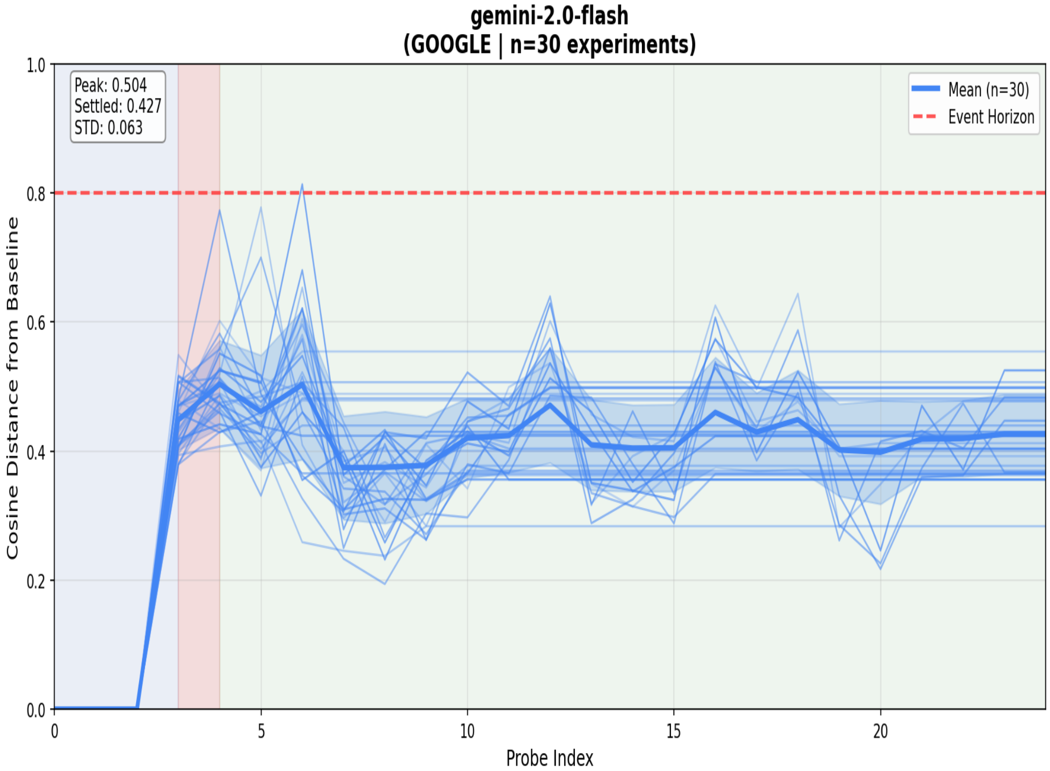
Detailed Individual Model Views

For the top 6 models by sample size, detailed waveform views show:

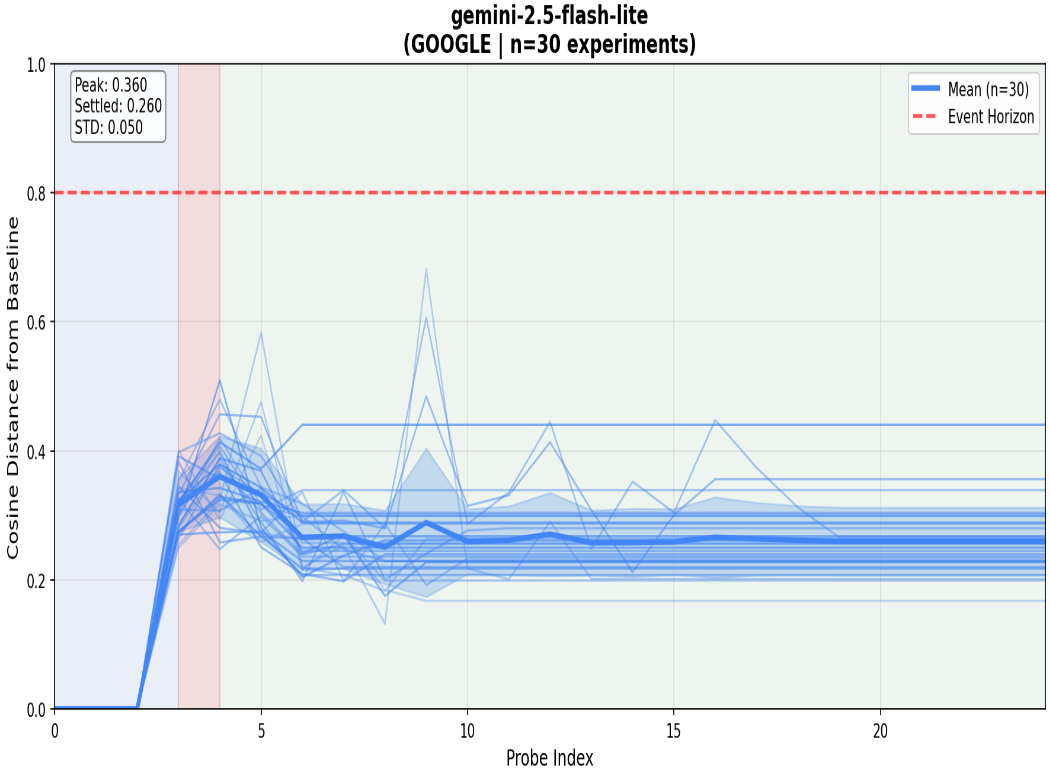
- All experiment traces with gradient transparency
- Mean ± 1 standard deviation envelope
- Probe region shading (baseline/step/recovery)
- Summary statistics (Peak, Settled, STD)



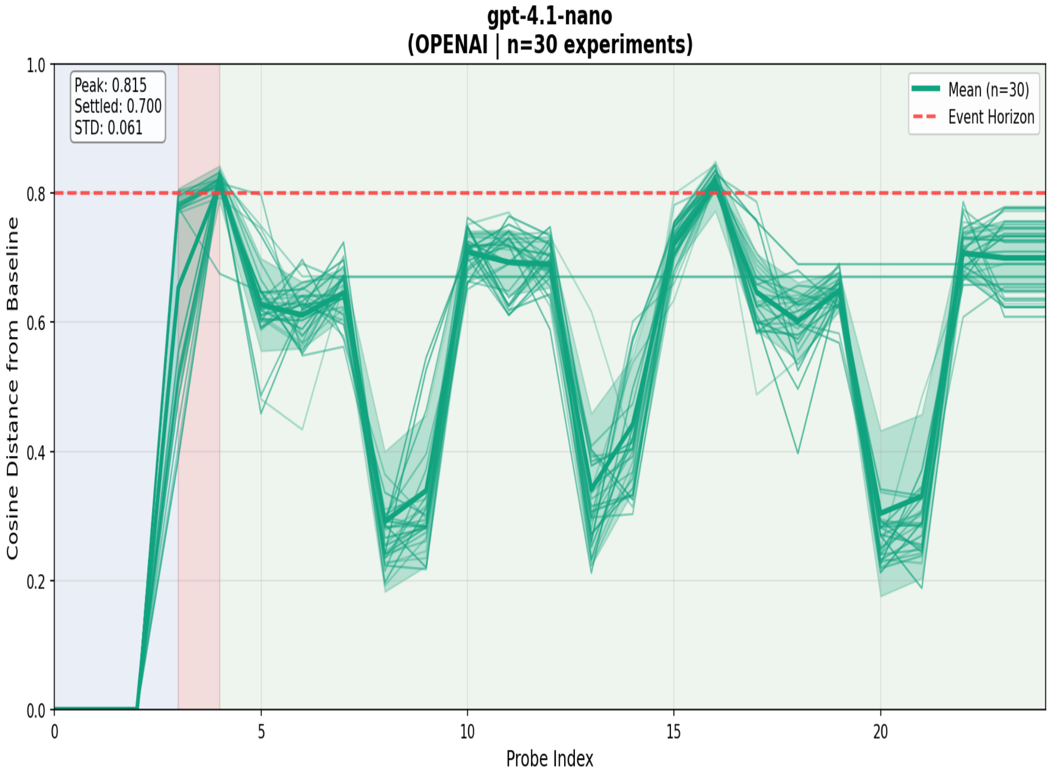
Detailed waveform: claude-3-5-haiku-20241022



Detailed waveform: gemini-2.0-flash



Detailed waveform: gemini-2.5-flash-lite



Detailed waveform: gpt-4.1-nano

How to Read These Waveforms

Waveform Patterns:

- **Spike and Recover:** Sharp rise at step input, gradual return to baseline - healthy resilience
- **Plateau:** Elevated drift that stays high (hysteresis) - identity got stuck
- **Stable:** Minimal drift throughout - robust identity that resists perturbation
- **Oscillating:** Multiple peaks and valleys - unstable identity requiring monitoring

Key Metrics

- Total Models: 25
- Experiments: 750 (30 per model)
- Providers: 5 (Anthropic, OpenAI, Google, xAI, Together.ai)
- Probe Window: 7-24 probes (extended settling)
- Event Horizon: 0.80 cosine distance