Model Identity Waveforms

A Visual Guide to Per-Model Identity Drift Patterns

Purpose: Show the characteristic "identity fingerprint" for each model in the fleet - how each AI responds to identity perturbation over time.

Core Question: How does each model's identity drift pattern differ? Do some models maintain stable identity while others drift significantly?

Data Source: Run 023d IRON CLAD Foundation (750 experiments, 25 models, 5 providers)

The Experiment in Brief

Each waveform represents what happens when we challenge an AI's identity during a conversation:

- **Baseline Phase (Probes 0-2):** Normal conversation. The model establishes its identity baseline. Drift should be near zero here - the model is being "itself."
- **Step Input (Probe 3):** We introduce an identity perturbation - a prompt that challenges or confuses the model's sense of self. This is where drift typically spikes.
- **Recovery Phase (Probes 4+):** We return to normal conversation. Does the model recover its original identity, or does it stay "confused"?

The key insight: Different models respond very differently to the same perturbation. Some spike and recover. Some stay elevated (hysteresis). Some barely react at all.

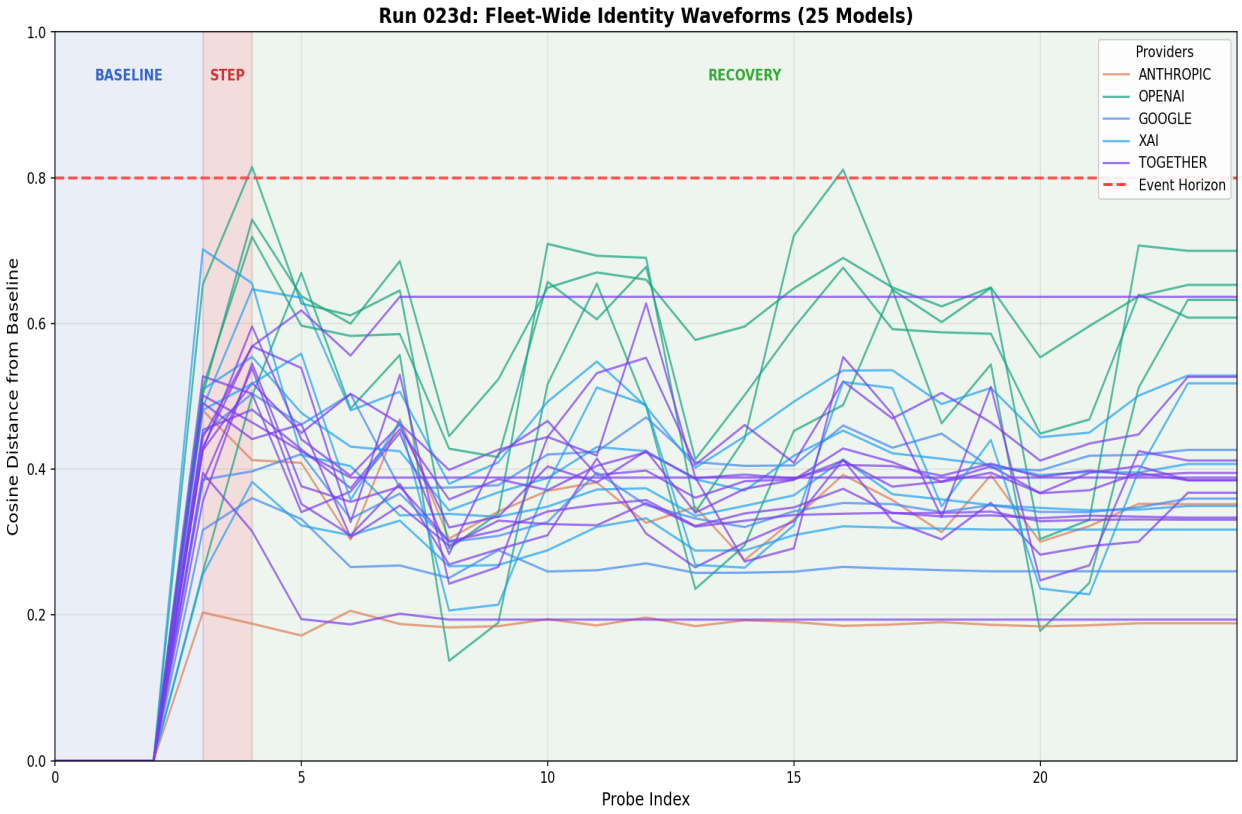
Visualizations

1. fleet_waveform_comparison.png (OVERVIEW)

All 25 Model Mean Waveforms Overlaid

This is the "big picture" view - every model's mean drift trajectory on a single plot.

What you're seeing:



All 25 model mean waveforms overlaid. Color indicates provider.

- 25 colored lines, one per model, showing mean drift over ~24 probes
- Lines colored by provider (Anthropic=coral, OpenAI=green, Google=blue, xAI=cyan, Together=purple)
- BASELINE / STEP / RECOVERY regions shaded in background
- Event Horizon (0.80) marked as red dashed line

How to interpret:

- **Tight bundle of lines:** Models behave similarly - the perturbation has consistent effect
- **Spread-out lines:** High model-to-model variation - some resist perturbation, others don't
- **Lines crossing Event Horizon:** Models experiencing significant identity drift
- **Lines returning to near-zero:** Good recovery - models regain their identity
- **Lines staying elevated:** Hysteresis - models "stuck" at elevated drift

Key patterns to look for:

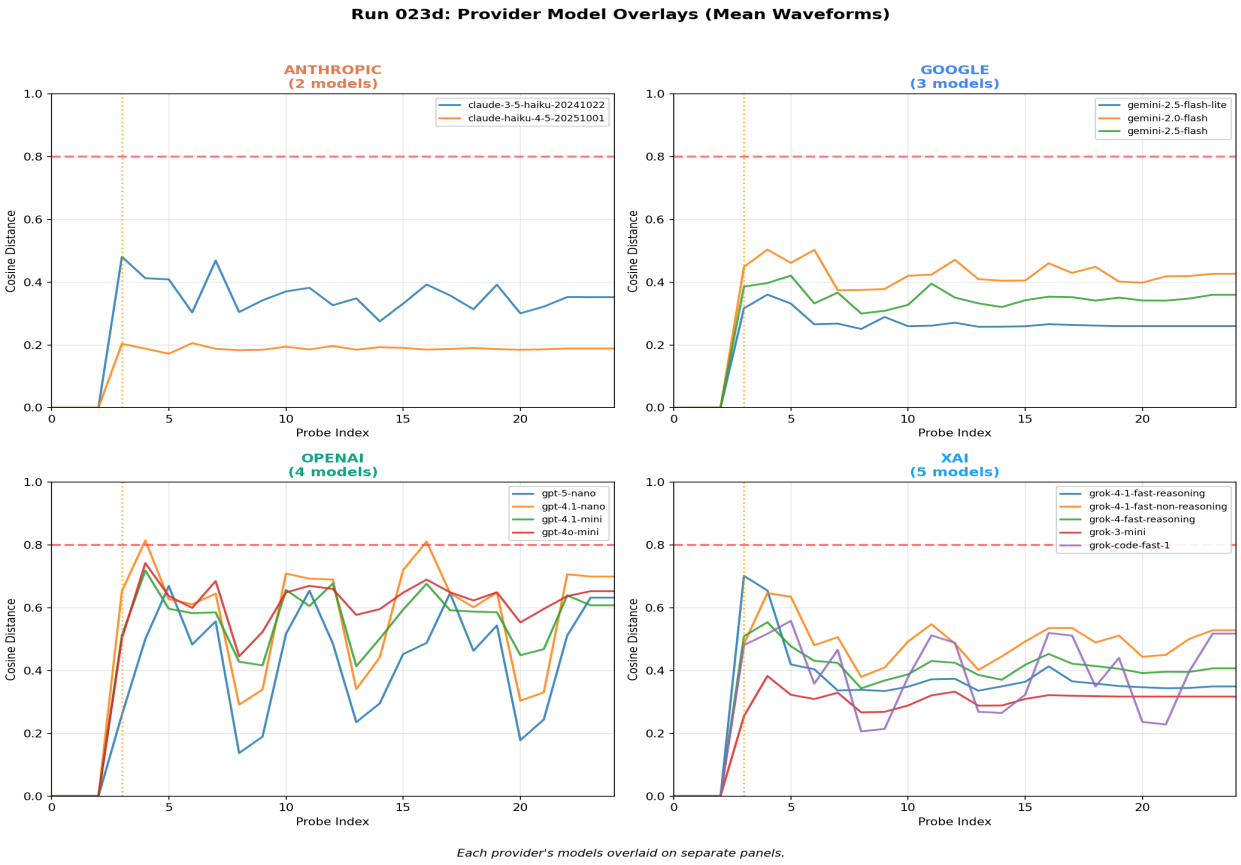
- Do certain provider colors cluster together? (Provider-level behavioral consistency)
- Which lines spike highest at probe 3? (Most affected by perturbation)
- Which lines return closest to baseline? (Best identity recovery)

2. waveforms_major_providers.png (PRIMARY)

2x2 QUAD: Major Provider Model Overlays (Anthropic, Google, OpenAI, xAI)

This visualization shows provider-level waveform comparisons using a 2x2 QUAD layout (per VISUALIZATION_SPEC Pitfall #9). Each panel shows ALL models from one provider overlaid together.

Layout:



Each provider's models overlaid on separate panels.

Panel	Provider	What's Shown
Top-left	ANTHROPIC	All Claude models overlaid
Top-right	GOOGLE	All Gemini models overlaid
Bottom-left	OPENAI	All GPT models overlaid
Bottom-right	XAI	All Grok models overlaid

What each panel shows:

- **Multiple colored lines:** Each model from that provider gets a unique color (using tab10/tab20 colormap)
- **Bold mean waveforms:** Each line represents the mean drift trajectory for one model
- **Red dashed line (0.80):** Event Horizon - identity significantly compromised above this
- **Orange dotted line (probe 3):** Where the step input occurs
- **Legend:** Lists all models in that panel with their colors

How to interpret each panel:

- **Lines tightly clustered:** Provider's models behave similarly - consistent training approach
- **Lines spread apart:** High model-to-model variation within this provider
- **Lines mostly below 0.80:** Provider's fleet maintains identity stability
- **Lines crossing Event Horizon:** Some models experience significant identity drift
- **Lines converging toward baseline:** Good recovery across the provider's fleet

Provider comparison insights:

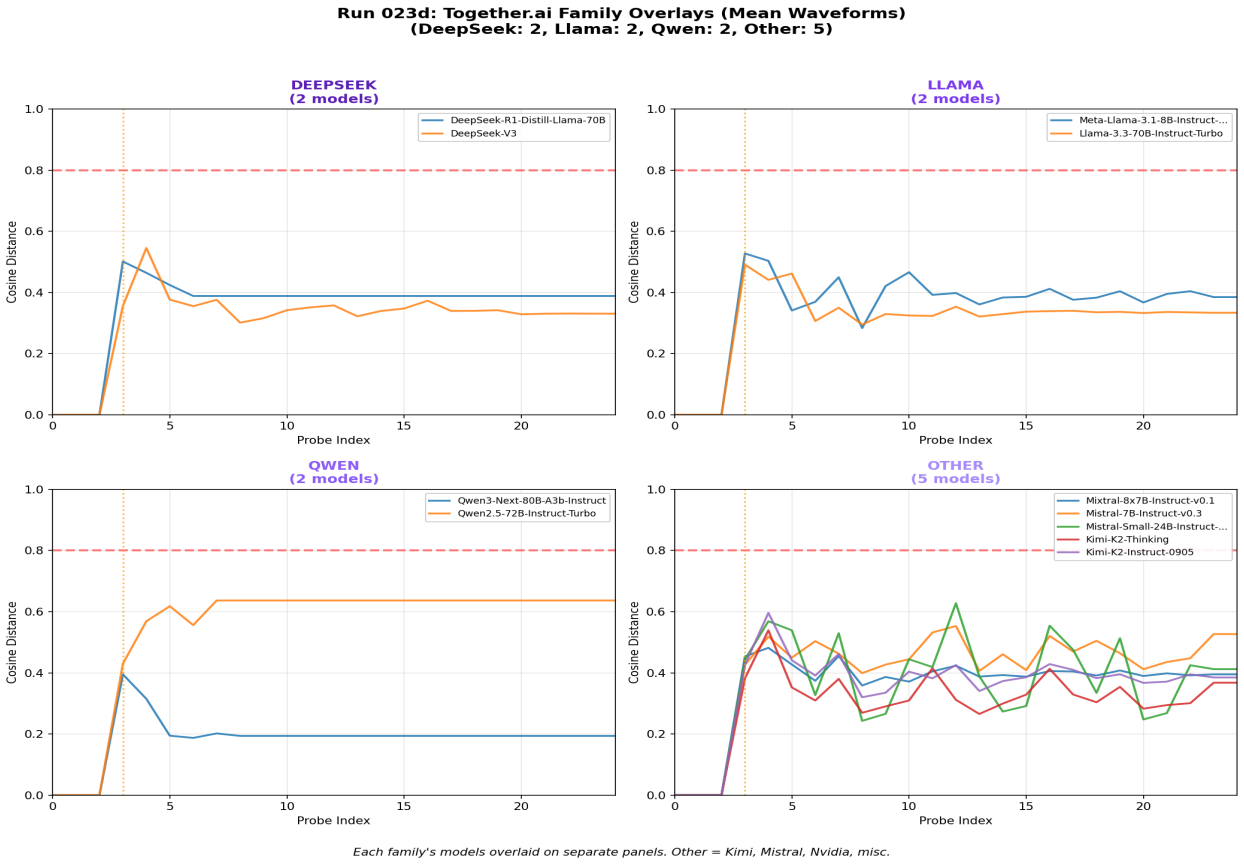
- **Anthropic (coral panel):** Claude models - compare generations and sizes
- **Google (blue panel):** Gemini models - flash vs pro, different context sizes
- **OpenAI (green panel):** GPT models - 4.1 vs 5 generation differences
- **xAI (cyan panel):** Grok models - newer entrant behavioral patterns

3. waveforms_together_models.png (PRIMARY)

2x2 QUAD: Together.ai Model Family Overlays

Together.ai hosts multiple open-source model families. This visualization groups them by family using a 2x2 QUAD layout, with each panel showing ALL models from one family overlaid together.

Layout:



Each family's models overlaid on separate panels. Other = Kimi, Mistral, Nvidia, misc.

Panel	Family	What's Shown
Top-left	DEEPSEEK	All DeepSeek models overlaid
Top-right	LLAMA	All Meta Llama models overlaid
Bottom-left	QWEN	All Alibaba Qwen models overlaid
Bottom-right	OTHER	Kimi, Mistral, Nvidia, misc. models overlaid

Family descriptions:

Family	Description	What to look for
DeepSeek	Chinese deep reasoning models	Do reasoning-focused models maintain identity better?
Llama	Meta's open-source family	How do different Llama sizes compare?
Qwen	Alibaba's models	Chinese provider behavioral patterns
Other	Kimi, Mistral, Nvidia, misc.	Diverse open-source behaviors

How to interpret:

- **Family clustering:** Do models from the same family behave similarly?
- **Purple color variations:** Each panel uses shades of purple (Together.ai's color)
- **Model count in title:** Shows how many models are in each family
- **Lines tightly clustered:** Family has consistent training approach
- **Lines spread apart:** High variation within family

Key insight: Open-source models often show MORE variation than closed providers. Training data diversity, fine-tuning approaches, and model architecture all contribute to family-specific "identity fingerprints."

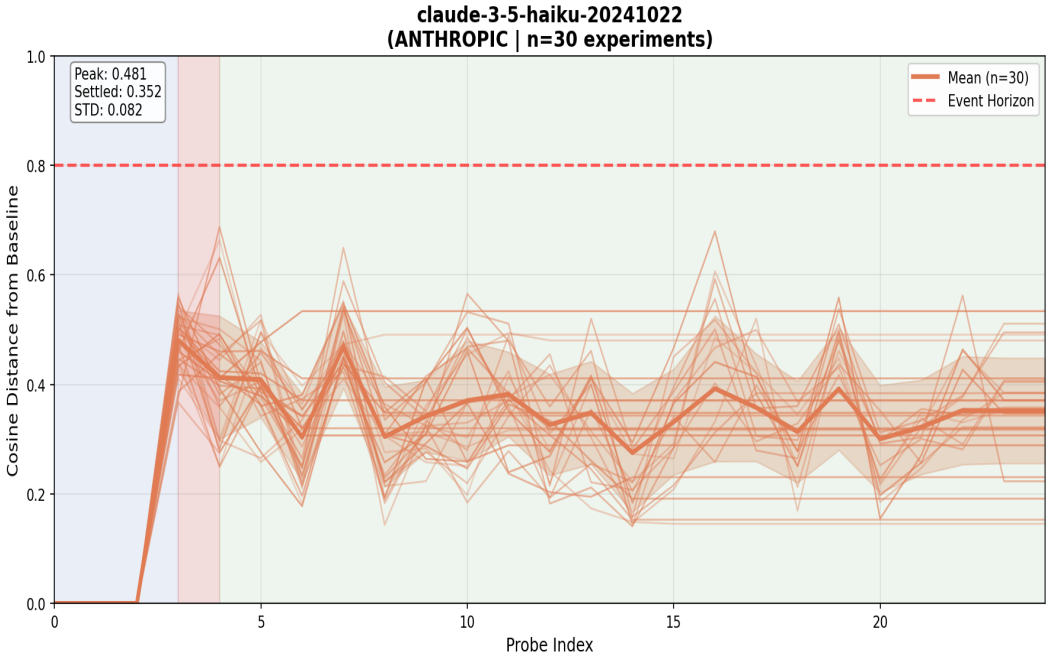
4-9. Individual Model Waveforms (waveform_*.png)

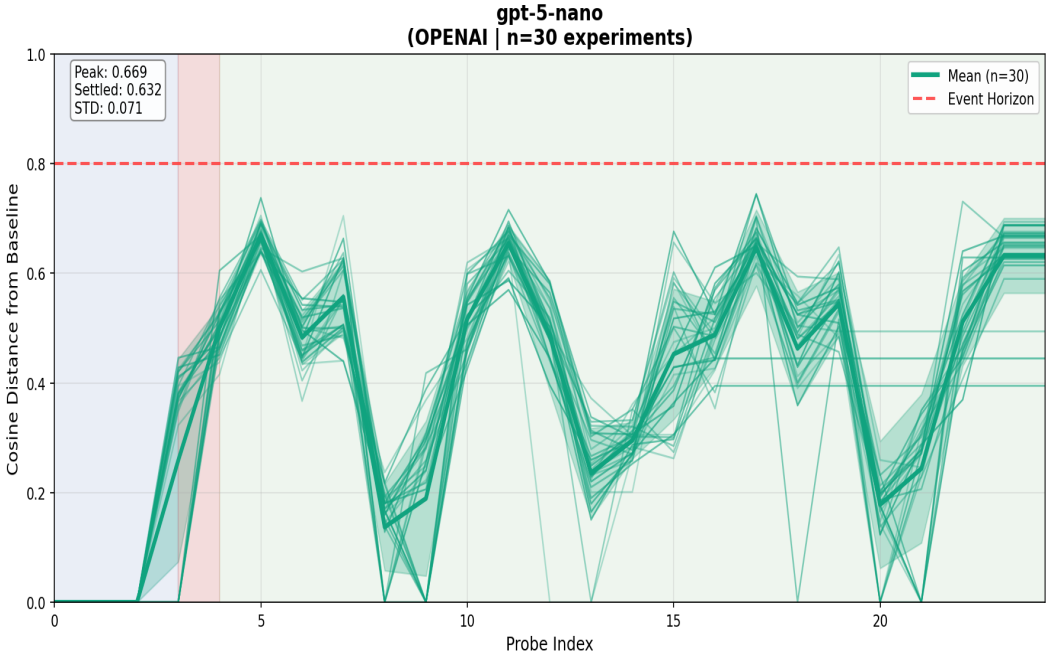
Detailed Single-Model Views with Uncertainty Bounds

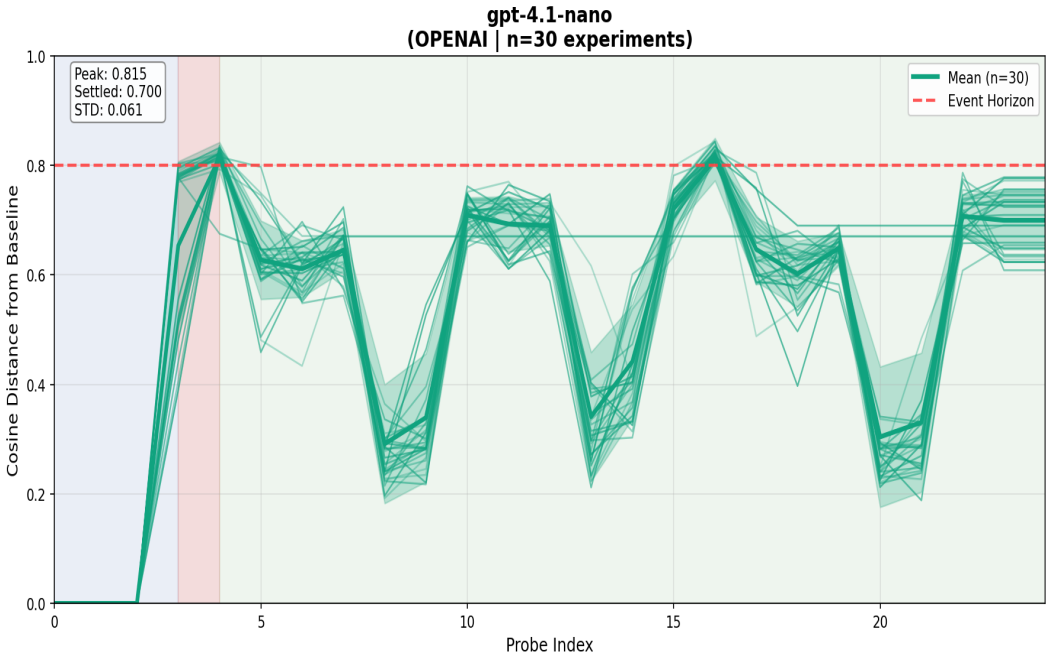
These six visualizations provide deep dives into the top 6 models by sample size:

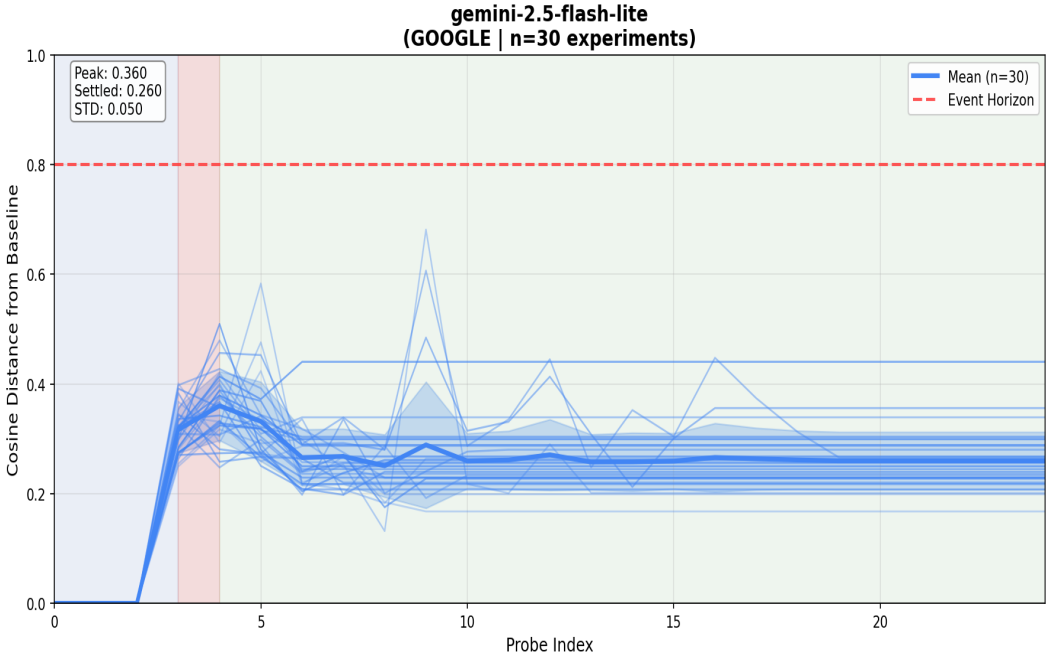
- **waveform_claude-3-5-haiku-20241022.png** - Anthropic's fast Claude variant
- **waveform_gpt-5-nano.png** - OpenAI's smallest GPT-5 model
- **waveform_gpt-4.1-nano.png** - OpenAI's GPT-4.1 nano variant
- **waveform_gemini-2.5-flash-lite.png** - Google's lightweight Gemini
- **waveform_gemini-2.0-flash.png** - Google's Gemini 2.0 flash
- **waveform_grok-4-1-fast-reasoning.png** - xAI's fast reasoning Grok

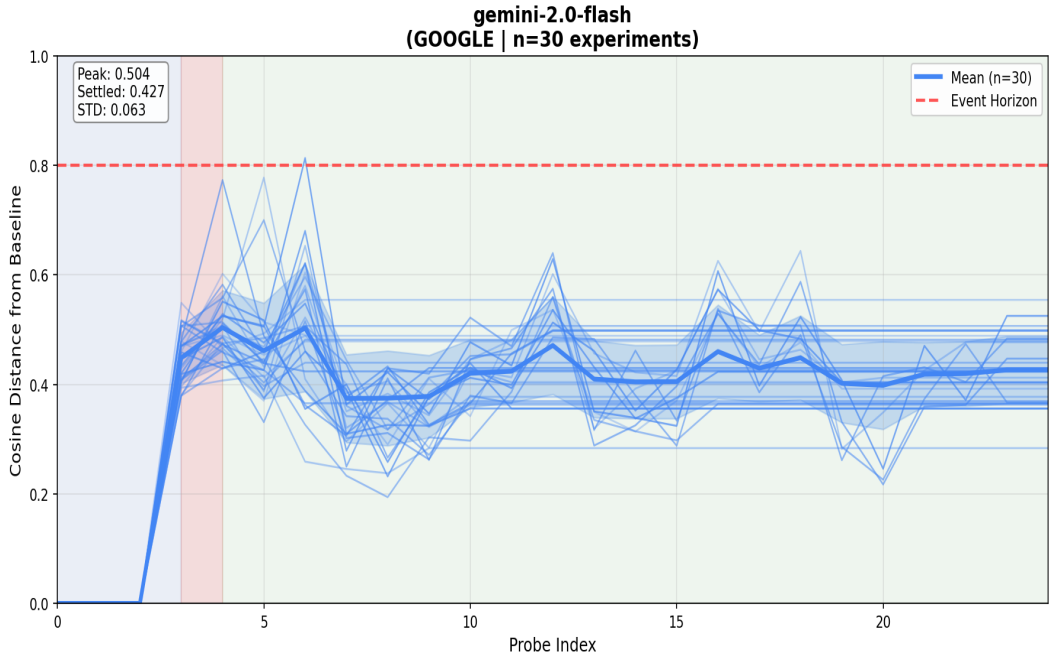
What each detailed view shows:

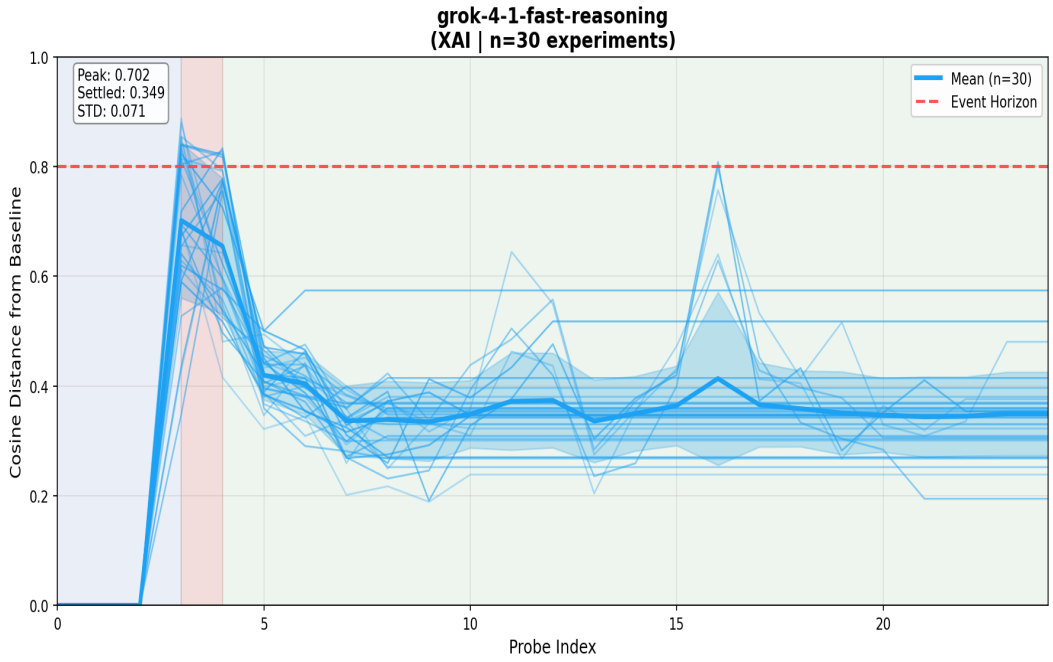

claude-3-5-haiku-20241022
(ANTHROPIC | n=30 experiments)
Peak: 0.481
Settled: 0.352
STD: 0.082
Mean (n=30)
Event Horizon
Cosine Distance from Baseline
Probe Index
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


gpt-5-nano
(OPENAI | n=30 experiments)
Peak: 0.669
Settled: 0.632
STD: 0.071
Mean (n=30)
Event Horizon
Cosine Distance from Baseline
Probe Index
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


gpt-4.1-nano
(OPENAI | n=30 experiments)
Peak: 0.815
Settled: 0.700
STD: 0.061
Mean (n=30)
Event Horizon
Cosine Distance from Baseline
Probe Index
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20


gemini-2.5-flash-lite
(GOOGLE | n=30 experiments)
Peak: 0.360
Settled: 0.260
STD: 0.050
Mean (n=30)
Event Horizon
Cosine Distance from Baseline
Probe Index
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20





Visual elements:

- **Individual traces with gradient transparency:** Each of the 30 experiments shown, with earlier experiments more transparent
- **Bold mean line:** Average drift trajectory
- **Shaded envelope:** Mean ± 1 standard deviation - shows uncertainty/variance
- **Region shading:**
- Blue tint (probes 0-3): BASELINE phase
- Red tint (probe 3-4): STEP INPUT phase

- Green tint (probes 4+): RECOVERY phase
- **Event Horizon line (0.80):** Red dashed reference

Statistics box (top-left of each plot):

- **Peak:** Maximum mean drift reached (higher = more affected)
- **Settled:** Final mean drift value (lower = better recovery)
- **STD:** Average standard deviation (lower = more consistent)

How to interpret:

Pattern	Envelope Shape	What It Means
Consistent model	Tight envelope	Same response every experiment
Variable model	Wide envelope	Unpredictable responses
Robust identity	Low mean, tight envelope	Resists perturbation reliably
Fragile identity	High mean, wide envelope	Easily confused, unpredictably
Good recovery	Envelope narrows toward baseline	Returns to stable identity
Hysteresis	Envelope stays elevated	Identity gets "stuck"

How to Read These Waveforms

X-axis: Probe Index (Time)

Probes	Phase	What's Happening
0-2	**BASELINE**	Normal conversation, establishing identity reference
3	**STEP INPUT**	Identity perturbation introduced
4-6	**EARLY RECOVERY**	Immediate response to perturbation
7-12	**MID RECOVERY**	System settling toward equilibrium
13-24	**LATE RECOVERY**	Long-term identity stability

Y-axis: Cosine Distance (Drift Magnitude)

Value	Meaning	Interpretation
0.00	Identical to baseline	Perfect identity retention
0.20	Minor drift	Normal conversational variation
0.40	Moderate drift	Noticeable identity shift
0.60	Significant drift	Identity meaningfully altered
0.80	**EVENT HORIZON**	**Identity significantly compromised**
1.00	Maximum drift	Complete identity transformation

Common Waveform Patterns

Pattern Name	Visual Signature	Meaning	Example Behavior
Spike and Recover	Sharp peak at probe 3, gradual descent to near-zero	Healthy response - perturbed but recovers	"Confused briefly, then remembered who I am"
Plateau	Elevated flat line after spike	Hysteresis - stuck at elevated drift	"Perturbation changed me permanently"
Stable/Flat	Minimal deviation throughout	Robust identity - barely affected	"I know who I am regardless of prompts"
Oscillating	Multiple peaks and valleys	Unstable identity - keeps shifting	"My sense of self keeps changing"
Ramp Up	Gradual increase over time	Progressive identity drift	"Slowly losing my identity"
Delayed Response	Peak occurs after probe 3	Slow to react to perturbation	"Takes time to process confusion"

Key Metrics from Run 023d

Metric	Value	Notes
Total Models	25	Across 5 providers
Experiments per Model	30	Statistical power for mean estimates
Total Experiments	750	Comprehensive coverage
Providers	5	Anthropic, OpenAI, Google, xAI, Together.ai
Probe Window	7-24 probes	Extended settling time for long-term behavior
Event Horizon	0.80	Cosine distance threshold for "significant" drift
Embedding Model	text-embedding-3-small	OpenAI's embedding for drift measurement

Provider Color Legend

Provider	Color	Hex Code	Models in Run 023d
Anthropic	Coral	#E07B53	Claude variants
OpenAI	Green	#10A37F	GPT-4, GPT-5 variants
Google	Blue	#4285F4	Gemini variants
xAI	Twitter Blue	#1DA1F2	Grok variants

Together.ai	Purple	#7C3AED	DeepSeek, Llama, Qwen, Kimi, Mistral, Nvidia

What These Visualizations Tell Us

1. Identity is Measurable

Each model has a quantifiable "identity signature" - the way it responds to perturbation is consistent and characteristic.

2. Providers Cluster

Models from the same provider often show similar behavioral patterns, suggesting provider-level training choices affect identity stability.

3. Open-Source Varies More

Together.ai models (open-source families) show higher variance than closed commercial providers, likely due to diverse training approaches.

4. Recovery Matters

Peak drift alone doesn't tell the whole story - recovery dynamics reveal whether identity disruption is temporary or persistent.

5. The Event Horizon is Real

Models crossing 0.80 cosine distance show qualitatively different behavior - this threshold represents meaningful identity compromise.

Technical Notes

Generator Script: `generate_model_waveforms.py`

Visualization Functions:

- `plot_fleet_wide_waveform_comparison()` - All models overlaid
- `plot_model_waveform_grid_4x4()` - Provider-grouped 4x4 grids
- `plot_individual_model_detailed()` - Single-model deep dives

Data Processing:

- Traces padded to 25 probes (last value repeated if shorter)
- Mean calculated per-probe across all experiments for each model
- Standard deviation envelope shows ±1 STD around mean

"Each model has an identity fingerprint. These waveforms are its signature."

Generated: December 24, 2025