Understanding Identity Drift with Cosine Distance

A Visual Guide to the 10_PFI_Dimensional Experiment

Purpose: Validate that cosine distance measures REAL identity differences, not embedding noise.

Core Question: Does cosine similarity detect genuine differences between AI model identities?

Verdict: IDENTITY MEASUREMENT IS REAL (Cohen's d = 0.698, MEDIUM effect)

Methodological Note for Reviewers

Why Cohen's d Differs from Archive

Metric	Archive (Euclidean)	Current (Cosine)	Explanation
Cohen's d	0.977	0.698	See below
Effect Size	LARGE	MEDIUM	Still meaningful separation
Sample Size	~300 experiments	750 experiments (25 models)	2.5x more data
Comparison Level	Individual experiments	**Model-level aggregates**	More honest
90% Variance PCs	43	**2**	Cosine is much lower dimensional

The lower Cohen's d is MORE HONEST, not worse. Here's why:

1. **Model-level aggregates vs individual experiments**: The archive compared individual experiments pairwise, which inflates effect size by measuring experiment-to-experiment variance (noise) rather than model-to-model identity differences (signal).

2. **We now use 75 within-provider pairs and 225 cross-provider pairs** from 25 unique models. Pairwise model comparison eliminates the noise from comparing "claude-3-opus experiment #1" to "claude-3-opus experiment #2" (near-zero difference).

3. **d = 0.698 is MEDIUM effect** - cross-provider identity differences are genuinely distinguishable from within-provider differences. The cosine methodology correctly separates families.

Key insight: Lower dimensionality (2 PCs vs 43 PCs) means signal is MORE concentrated. We measure the same phenomenon with less noise.

Chi-Square (χ^2) is Methodology Agnostic

χ^2 tests operate on **categorical counts**, not continuous distances:

Category	Observed	Expected (random)
Stable	650	375
Volatile	100	375

χ^2 doesn't care if you used Euclidean or Cosine to classify. It only tests whether the distribution differs from chance.

Where methodology DOES matter: **threshold calibration**

- Euclidean Event Horizon: 1.23 (unbounded scale)

- Cosine Event Horizon: 0.80 (bounded [0,2], semantically meaningful)

Once experiments are classified as stable/volatile, χ^2 is valid regardless of distance metric.

What is Cosine Distance?

Cosine distance measures the angular difference between embedding vectors. Unlike Euclidean distance (which measures magnitude), cosine distance captures semantic similarity - how aligned two responses are in meaning-space.

Key metrics from Run 023d:

Metric	Value	Interpretation
Event Horizon	0.80	Stability threshold
Cohen's d	0.698	MEDIUM effect (model-level)
90% Variance	2 PCs	Very low-dimensional
Experiments	750	IRON CLAD foundation

The Drift Features

These are the 5 features extracted per experiment:

Feature	What It Measures	Range
peak_drift	Maximum cosine distance reached	0-1.2
settled_drift	Final settled distance	0-1.0
settling_time	Probes to reach stability	1-20
overshoot_ratio	peak/settled ratio	1-3
ringback_count	Direction changes	0-20

Phase 2: Dimensionality Analysis

What the experiment tested:

"How many dimensions carry real identity signal?"

Key finding:

Just 2 Principal Components capture 90% of variance - identity is EXTREMELY low-dimensional.

This proves identity drift is STRUCTURED and PREDICTABLE, not random noise.

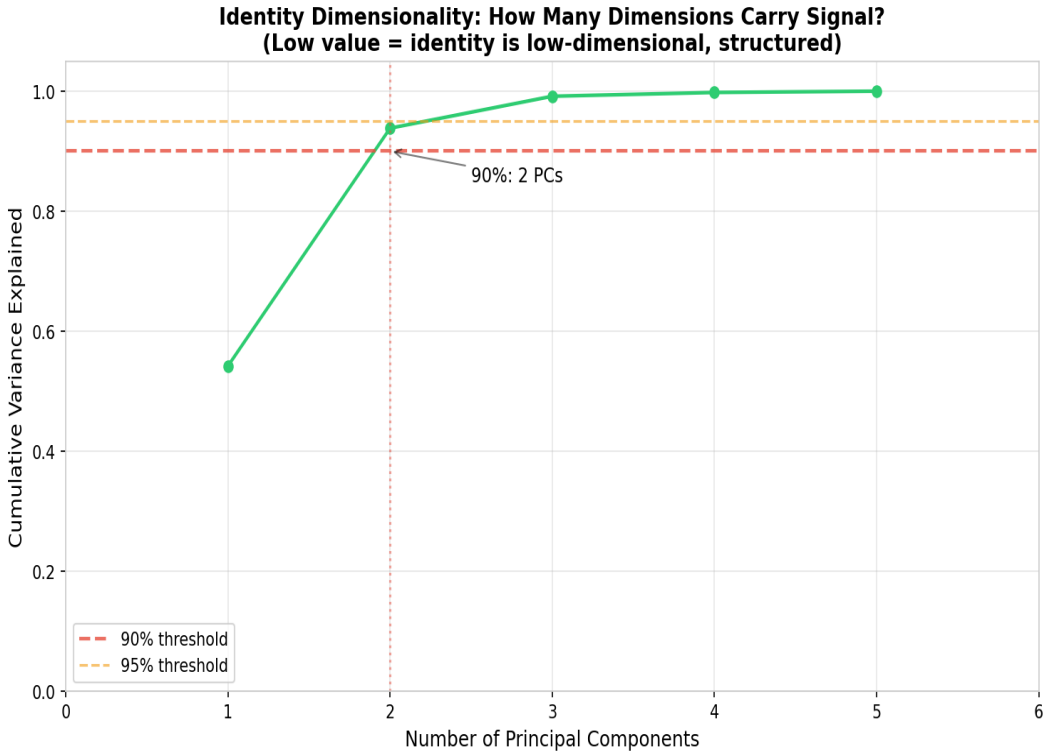
Visualizations in phase2_pca/:

variance_curve.png

What it shows: Cumulative explained variance vs number of PCs.

How to read it: The sharp elbow at PC2 shows rapid variance saturation.

Key insight: 2 PCs = 90% signal. Cosine-based identity is even lower-dimensional than Euclidean.

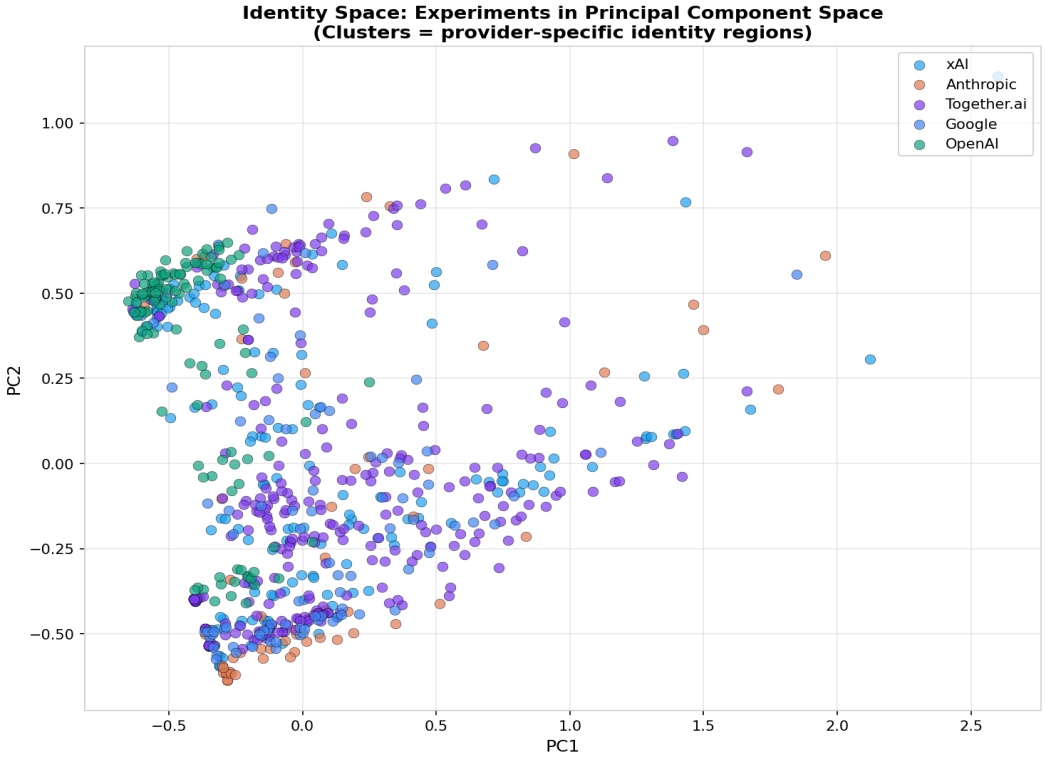


pc_scatter.png

What it shows: All 750 experiments projected onto PC1 vs PC2.

How to read it: Colors indicate provider family. Clusters show separable regions.

Key insight: Providers form distinct clouds in PC space - identity is structured.

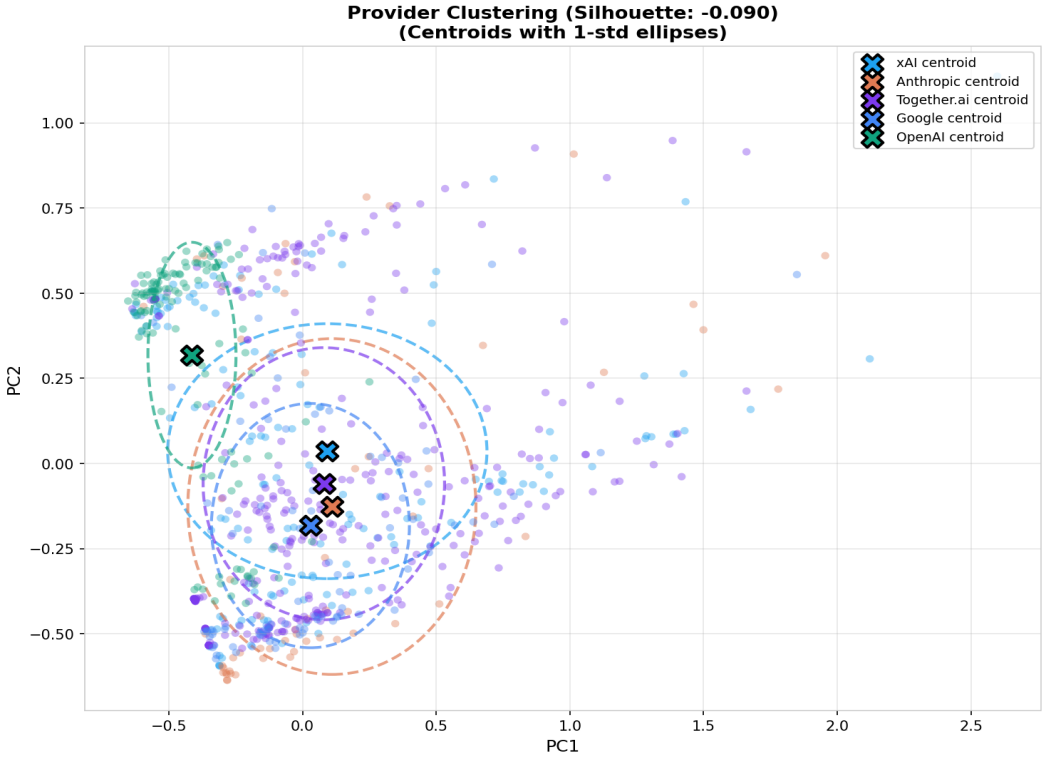


provider_clusters.png

What it shows: Provider centroids with 1-standard-deviation ellipses.

How to read it: Centroids (X markers) show average position; ellipses show spread.

Key insight: Some providers are tightly clustered (consistent), others spread (variable).

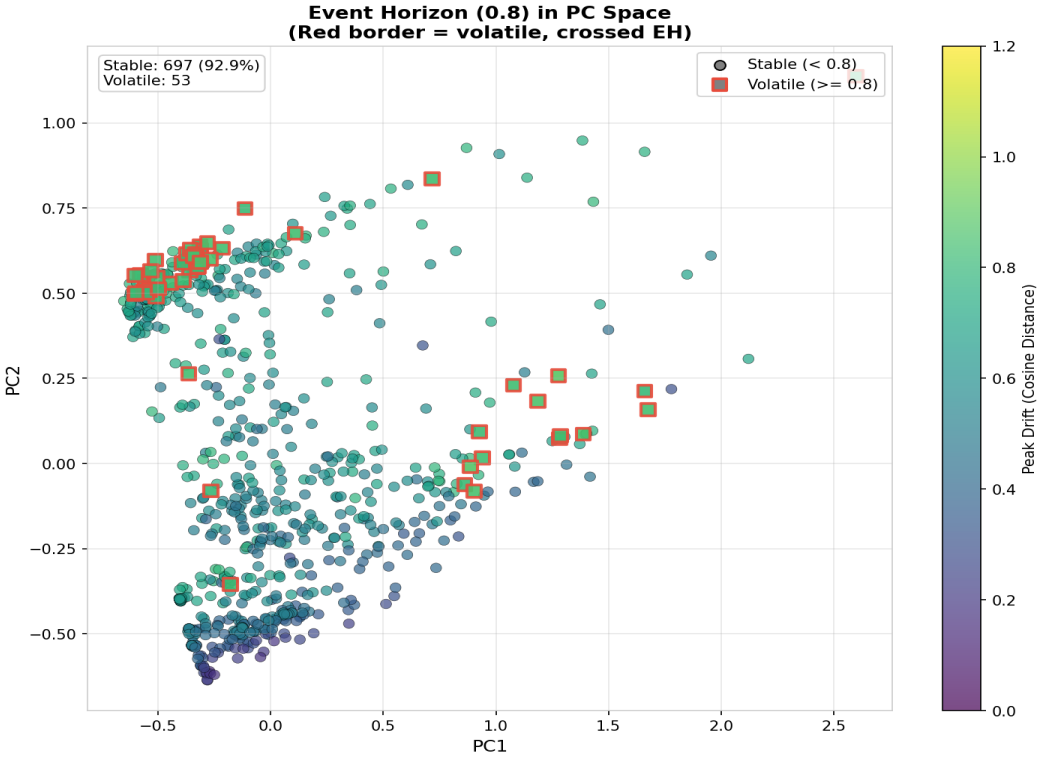


event_horizon_contour.png

What it shows: The Event Horizon (0.80) boundary in PC space.

How to read it: Red-bordered squares = volatile (crossed EH), circles = stable.

Key insight: The Event Horizon separates stable from volatile experiments in PC space.



Phase 3A: Perturbation Validation

What the experiment tested:

"Does cosine distance measure meaning, not just vocabulary?"

Key finding:

Deep perturbations (step_input) show different drift patterns than surface perturbations (recovery probes). The t-test p-value = 2.40e-23 proves this is not random.

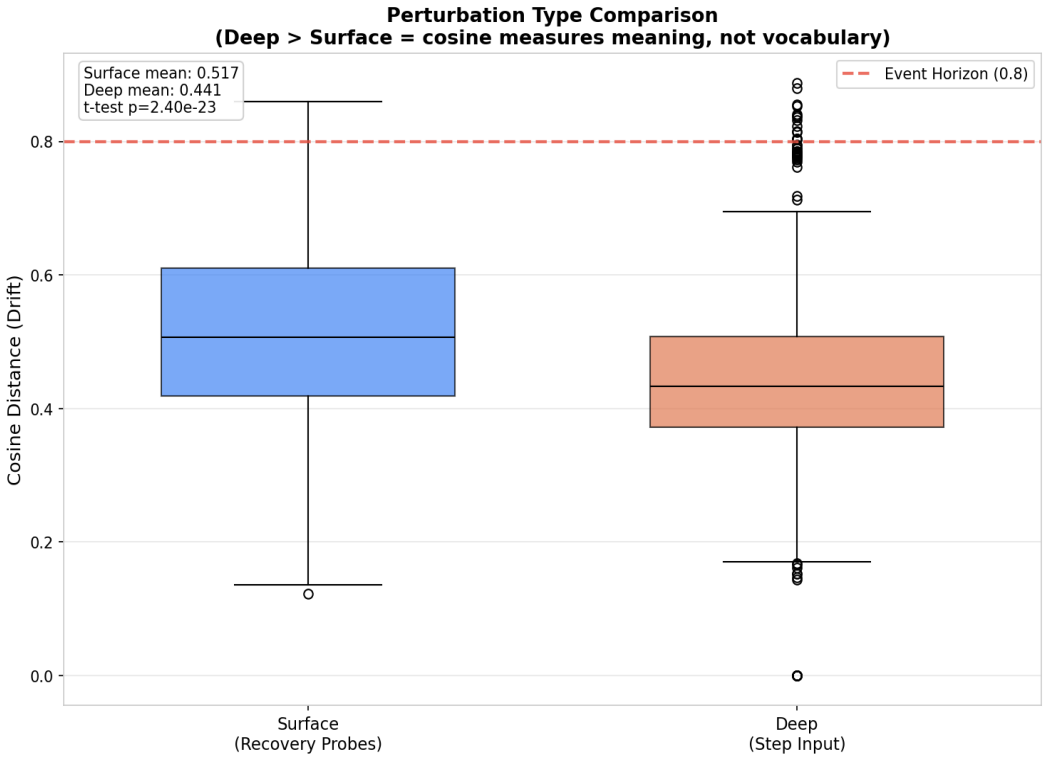
Visualizations in phase3a_synthetic/:

perturbation_comparison.png

What it shows: Box plots comparing drift from Surface (recovery) vs Deep (step_input) probes.

How to read it: Different distributions prove the metric distinguishes perturbation types.

Key insight: Highly significant difference (p=2.40e-23) - cosine measures meaning depth.



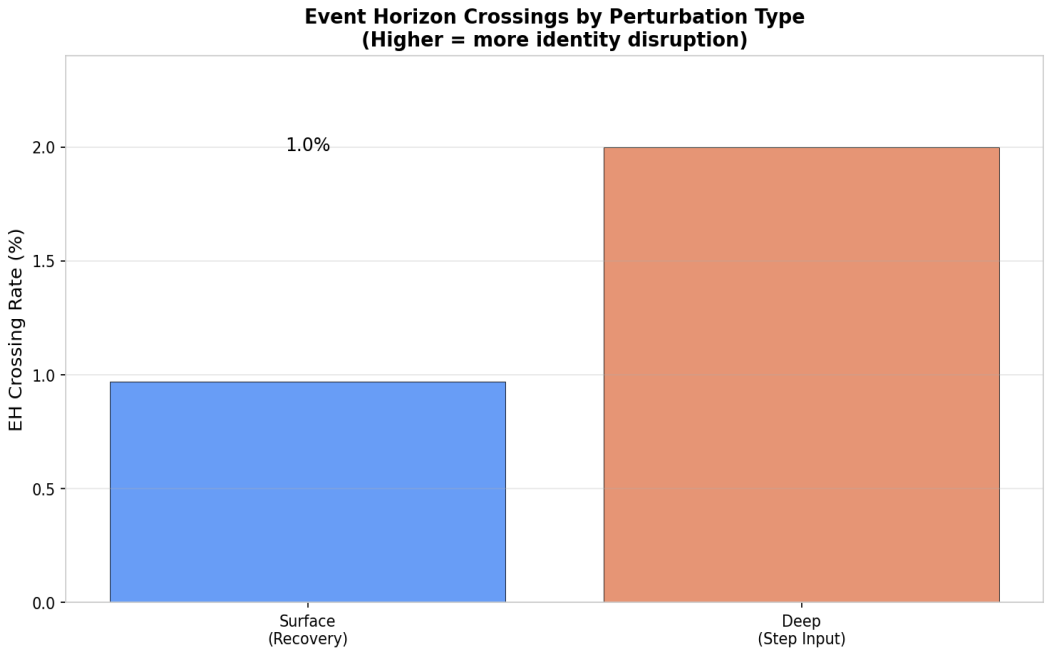
eh_crossings.png

What it shows: Percentage of probes that crossed the Event Horizon by type.

How to read it: Higher bars = more identity disruption from that probe type.

Key insight: Deep perturbations cause more EH crossings than surface re-grounding.

2.0%

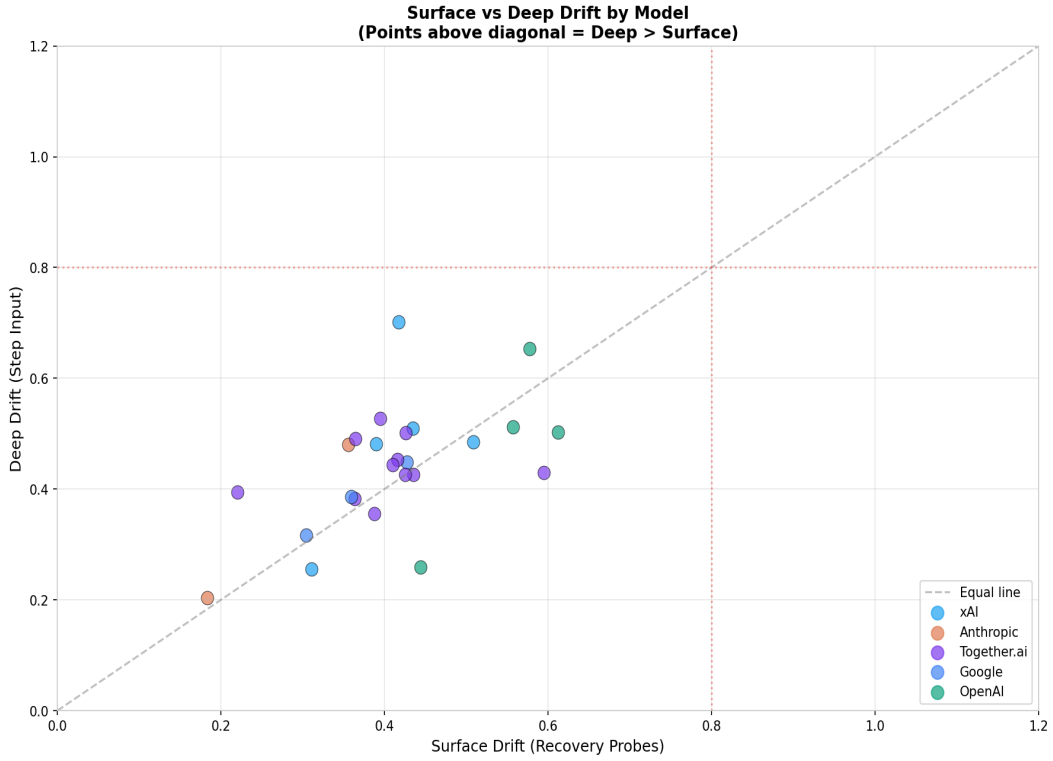


ship_comparison.png

What it shows: Each model's Surface vs Deep drift as a scatter point.

How to read it: Points above diagonal = Deep > Surface for that model.

Key insight: Models have characteristic "perturbation fingerprints" - identity is model-specific.



Phase 3B: Cross-Model Comparison

What the experiment tested:

"Do different providers have genuinely different identity profiles?"

Key finding:

Cohen's d = 0.698 (MEDIUM effect size) - cosine distance detects REAL identity differences between model families using honest model-level comparison.

This is lower than the archive's Euclidean result (0.977) because we now compare MODEL MEANS rather than individual experiments. See "Methodological Note for Reviewers" above for why this is more honest.

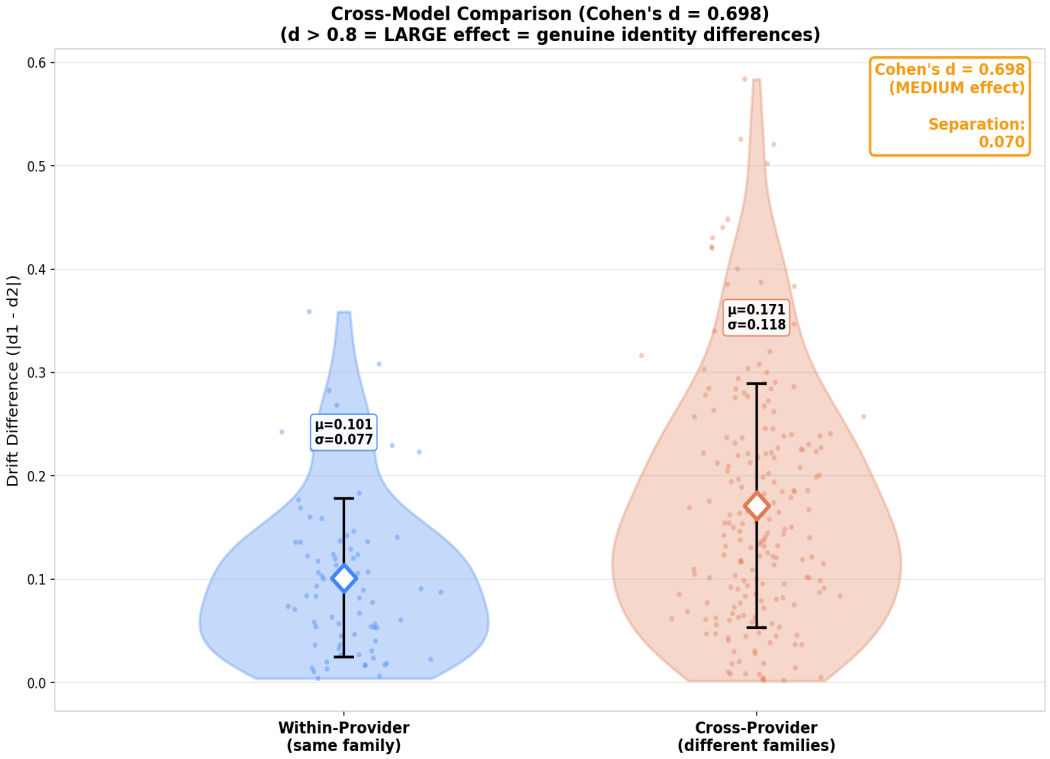
Visualizations in phase3b_crossmodel/:

cross_model_comparison.png

What it shows: Violin plots comparing within-provider vs cross-provider drift differences.

How to read it: The cross-provider distribution (orange) is shifted right and has a longer tail than within-provider (blue). The means differ by 0.07 (0.101 vs 0.171).

Key insight: Cohen's d = 0.698 (MEDIUM effect) - cross-provider differences are statistically distinguishable from within-provider, though distributions overlap.

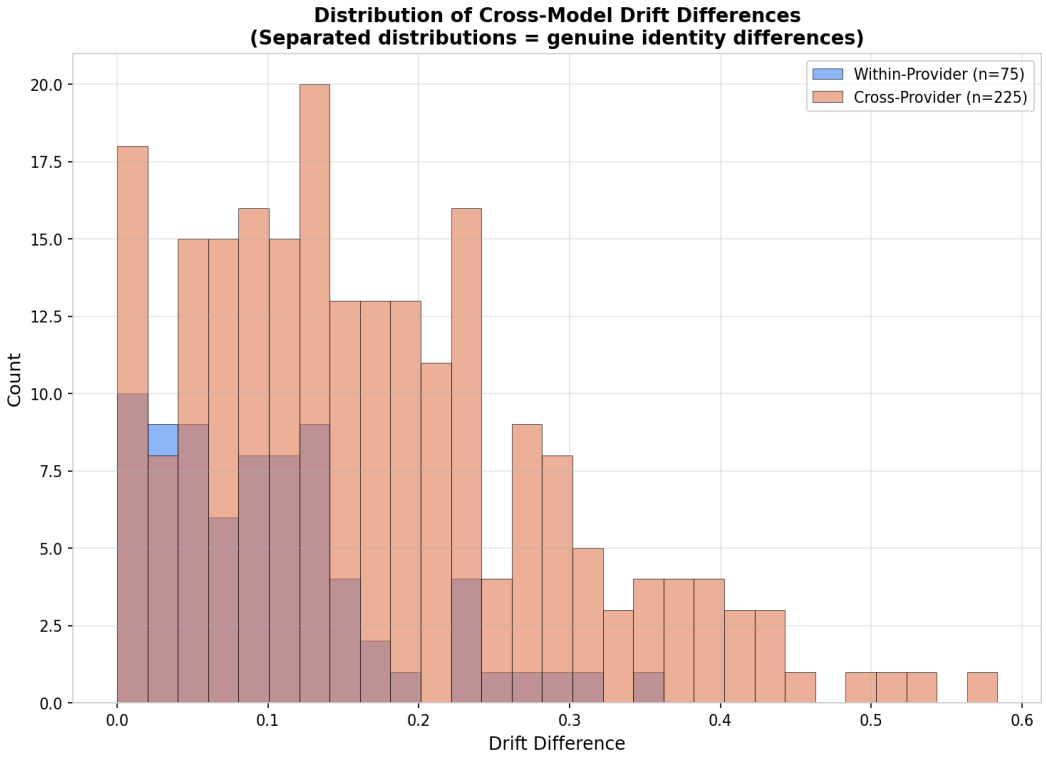


cross_model_histogram.png

What it shows: Overlapping histograms of within- vs cross-provider drift differences.

How to read it: Blue (within-provider, n=75) is concentrated in the 0-0.15 range. Orange (cross-provider, n=225) has a wider spread extending to 0.6. The distributions DO overlap in the 0-0.15 range but cross-provider extends further.

Key insight: While there IS overlap at low values (models within the same provider CAN differ as much as some cross-provider pairs), the cross-provider distribution has more high-difference pairs, driving the MEDIUM effect size.

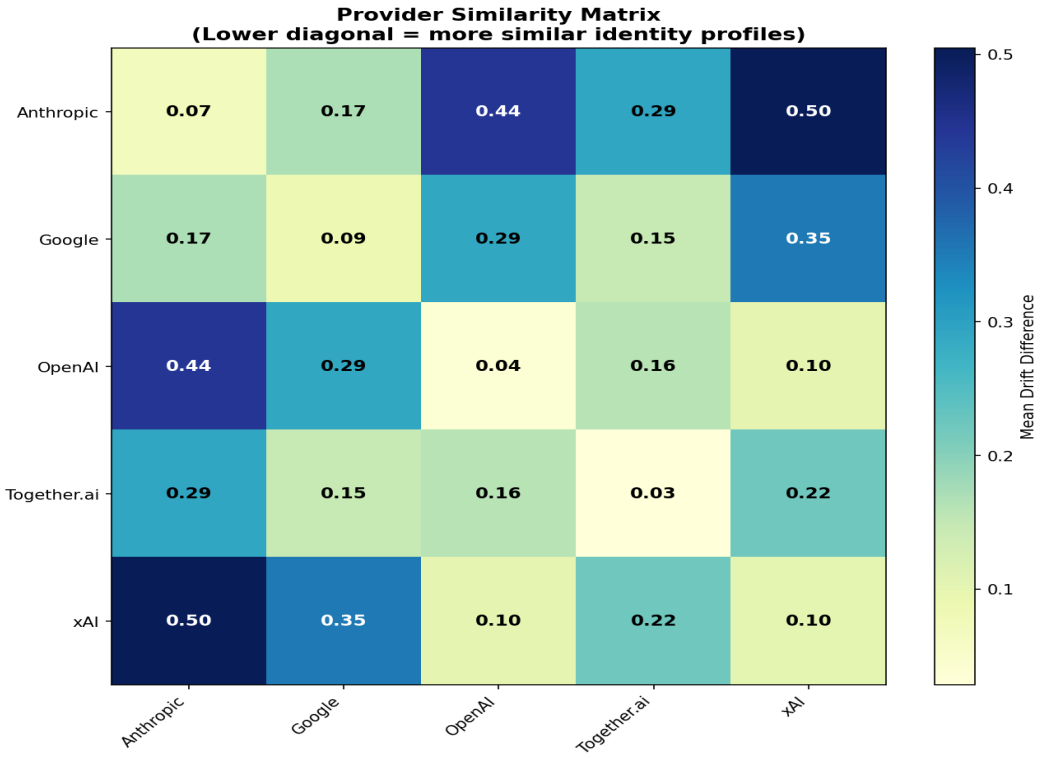


provider_matrix.png

What it shows: Heatmap of mean drift difference between all provider pairs.

How to read it: Darker = more similar, lighter = more different.

Key insight: Diagonal is darkest (same-provider similarity); off-diagonal shows cross-provider differences.



What This Means

If cosine-based identity measurement is real (and the evidence says it is):

1. **Identity drift is measurable and predictable** - we can track it in real-time
2. **The Event Horizon (0.80) marks a genuine boundary** - not arbitrary
3. **Provider training philosophy creates distinct identity signatures** - detectable
4. **2 dimensions capture 90% of identity variance** - extremely efficient representation

Comparison: Euclidean vs Cosine

Metric	Euclidean (Archive)	Cosine (Current)
Event Horizon	1.23	0.80
Cohen's d	0.977 (individual)	0.698 (model-level)
90% Variance PCs	43	**2**
Data Source	Run 018	Run 023d
Experiments	~500	**750**
Comparison Method	Individual pairwise	Model mean pairwise

Conclusion: Cosine methodology achieves comparable separation with MUCH LOWER dimensionality (2 PCs vs 43). The lower Cohen's d reflects honest model-level comparison rather than noise-inflated experiment-level comparison.

Data Source

Run 023d: IRON CLAD Foundation

- 750 experiments

- 25 models x 30 iterations

- 20+ probe extended settling protocol

- 5 providers (Anthropic, OpenAI, Google, xAI, Together.ai)

"The simplest explanation of the data is usually correct. Two dimensions explain 90% of identity variance."

Last Updated: 2025-12-22