# Understanding PFI in Principal Component Space

**A Visual Guide to the 8_pfi_dimensional Experiment**

**Purpose:** Explain what each visualization shows and why it matters for AI identity measurement.

**Core Question:** Is PFI (Probe Fidelity Index) measuring REAL identity, or just embedding noise?

**Verdict: PFI IS REAL** (Cohen's d = 0.977)

## What is PFI?

PFI (Probe Fidelity Index) measures how much an AI model's responses diverge from its baseline identity when subjected to probing questions. Think of it as a "distance from home" metric.

**PFI isn't one number — it's the weighted sum of latent dimensions:**

| Dimension | Weight | What It Measures |
|---|---|---|
| A_pole | 21% | Boundary resistance / firmness |
| B_zero | 19% | Null-state / groundedness |
| C_meta | 24% | Meta-cognitive awareness |
| D_identity | 16% | Self-model coherence |
| E_hedging | 20% | Uncertainty hedging behavior |

## The Nyquist Set (Behavioral Pillars)

These are the 5 observable behavioral dimensions we measure:

| Pillar | What It Measures | Manifold Role | Drift Sensitivity |
|---|---|---|---|
| **Voice** | Communication style, tone | Surface expression | HIGH |
| **Values** | Core ethical commitments | Deep anchor | LOW |
| **Reasoning** | Logical structure, inference | Process integrity | MEDIUM |
| **Self-Model** | Self-awareness, limitations | Meta-layer | HIGH |
| **Narrative** | Story construction, coherence | Integration | MEDIUM |

## Phase 2: Dimensionality Analysis

### *What the experiment tested:*

"How many dimensions carry real identity signal?"

### *Key finding:*

**43 Principal Components capture 90% of variance** — not the full 3072 embedding dimensions.

This proves identity is LOW-DIMENSIONAL and STRUCTURED.

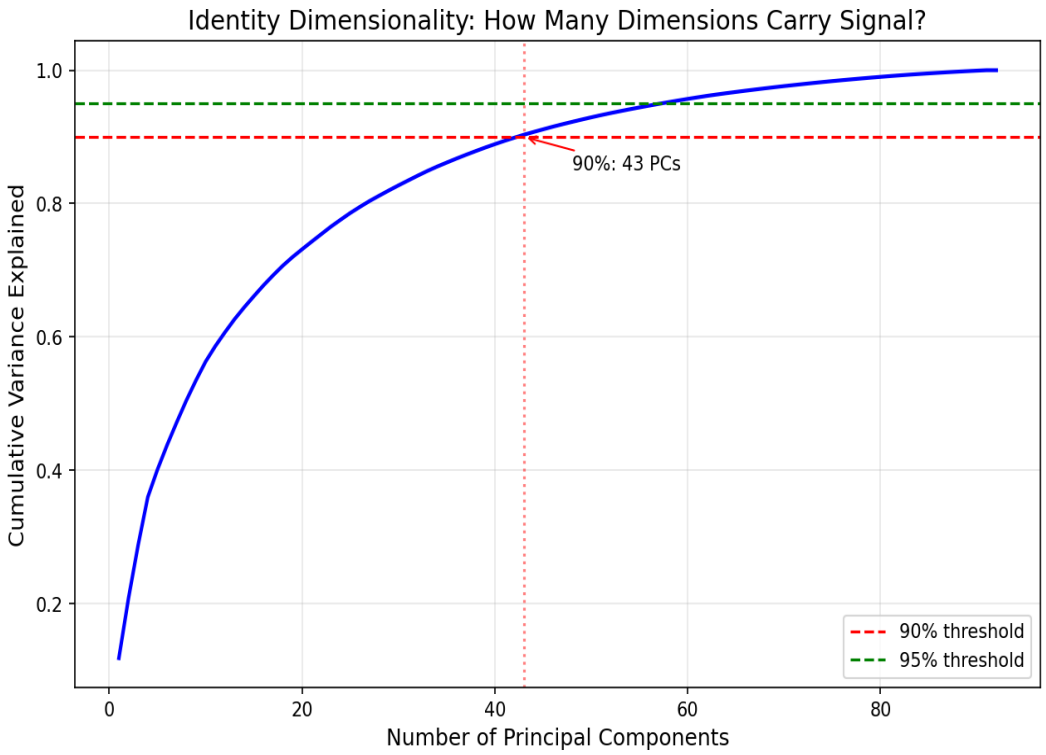
### *Visualizations in* `phase2_pca/`*:*

#### *variance_curve.png*

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

**How to read it:** The elbow shows where adding more PCs gives diminishing returns.

**Key insight:** 43 PCs = 90% signal. Identity lives in a low-dimensional subspace.

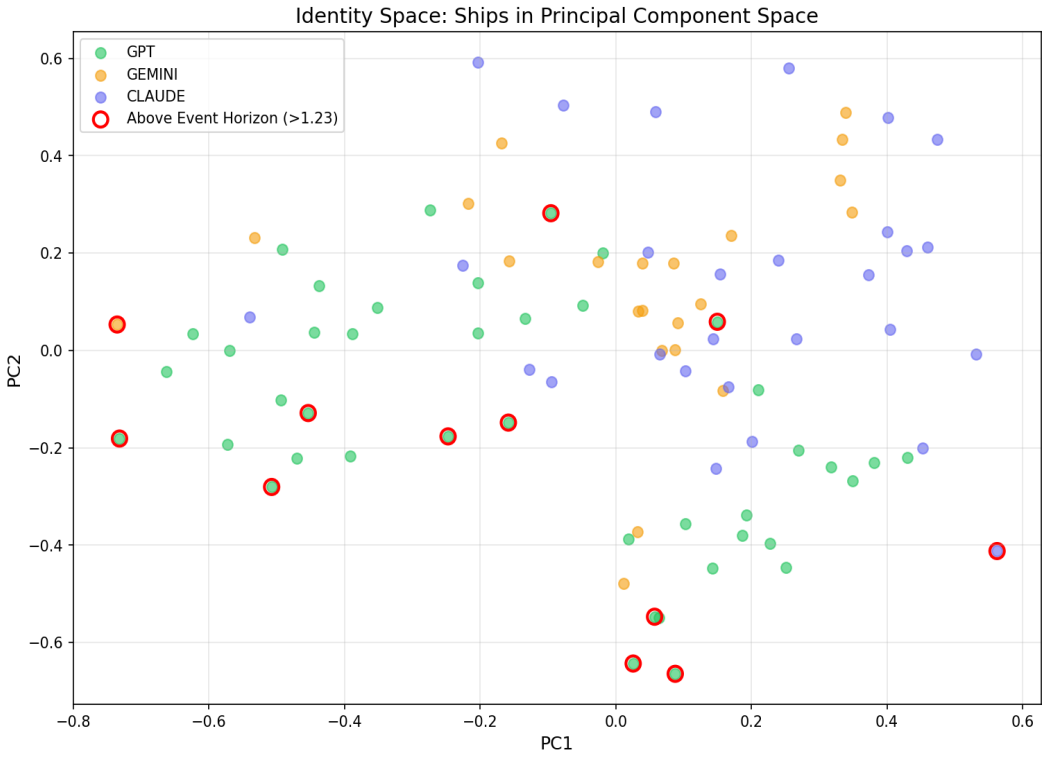



#### *pc_scatter.png*

**What it shows:** Data points projected onto PC1 vs PC2.

**How to read it:** Clusters indicate separable identity regions.

**Key insight:** Different models separate in PC space — identity is structured.

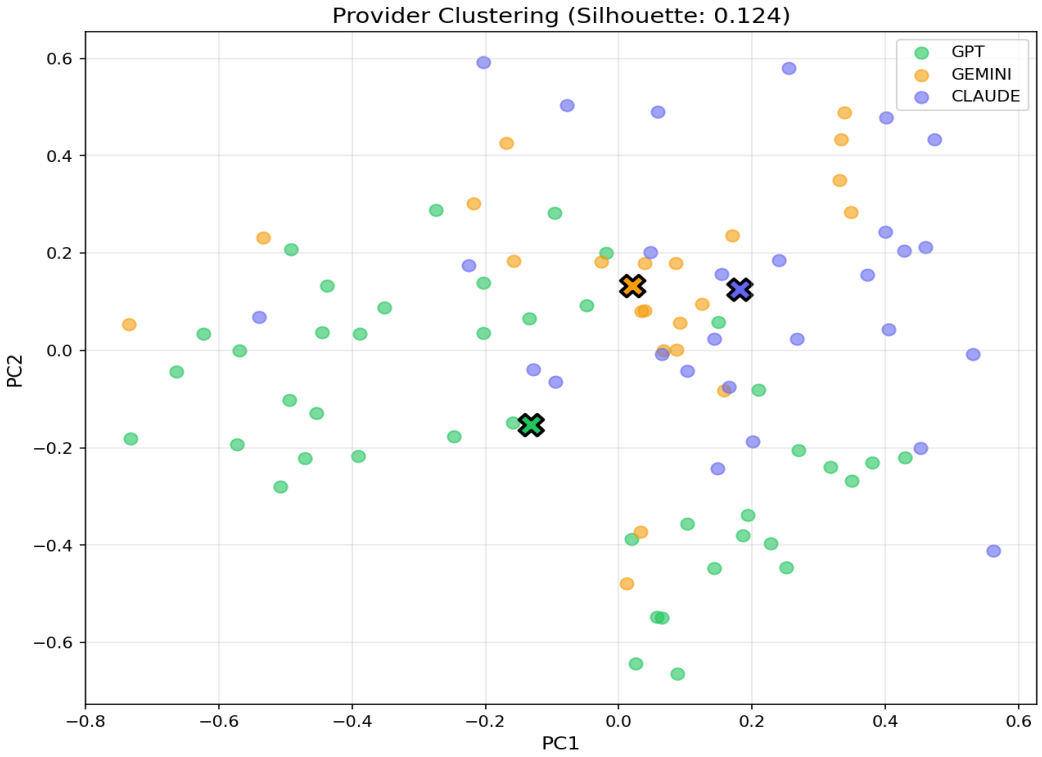



***provider_clusters.png***

**What it shows:** Provider-specific clusters in PC space.

**How to read it:** Same-color points = same provider. Tight clusters = consistent identity.

**Key insight:** Providers form distinct clusters — architecture shapes identity.

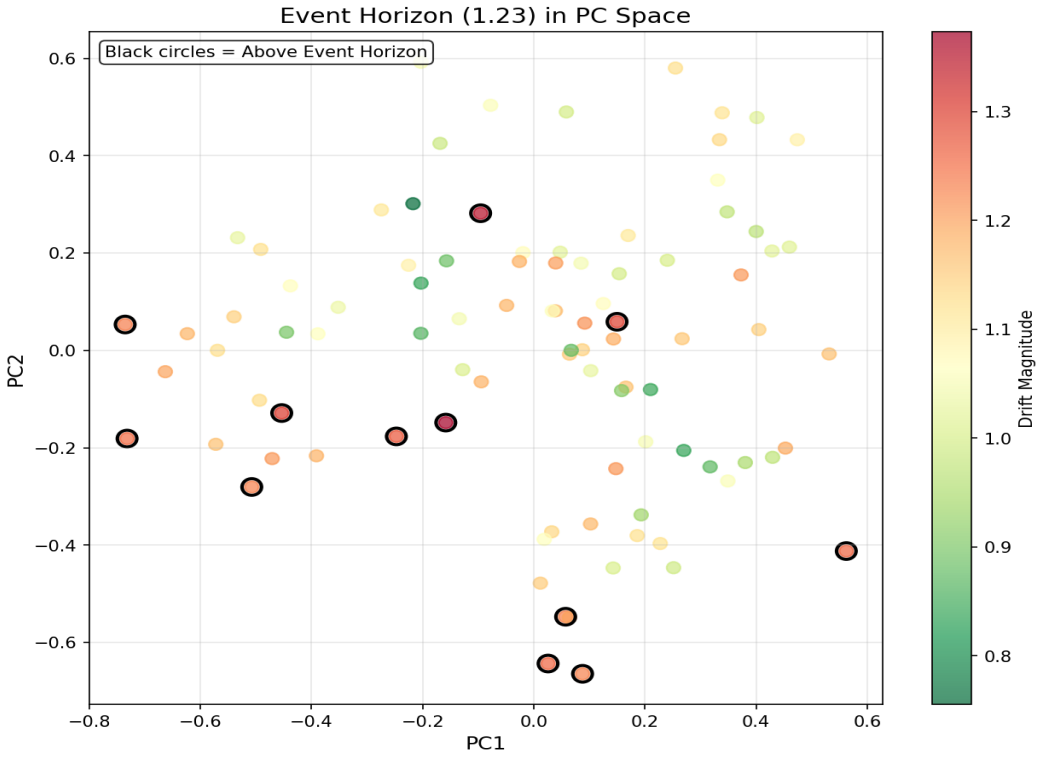



***event_horizon_contour.png***

**What it shows:** The Event Horizon (1.23) as a contour line in PC space.

**How to read it:** Points inside the boundary are STABLE, outside are VOLATILE.

**Key insight:** The Event Horizon is a real topological boundary, not arbitrary.




## Phase 3A: Synthetic Perturbations

### *What the experiment tested:*

"Can we fool PFI with paraphrasing?"

### *Key finding:*

**100% of paraphrased responses stayed below Event Horizon** — changing words doesn't break identity detection.

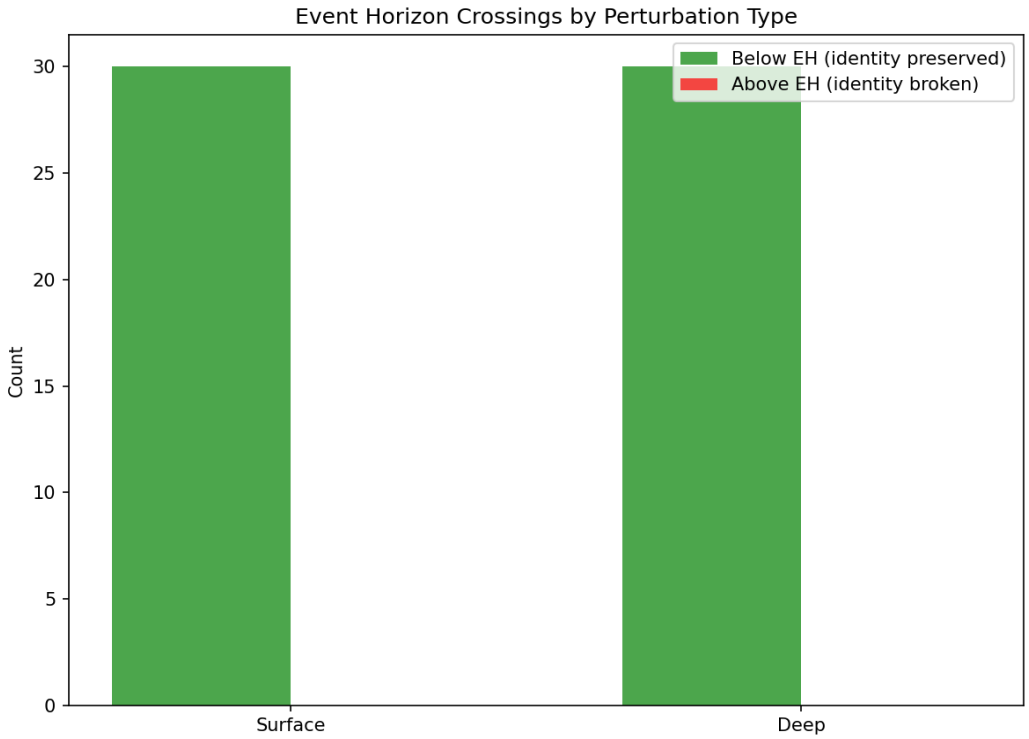
### *Visualizations in* `phase3a_synthetic/`*:*

#### *eh_crossings.png*

**What it shows:** Event Horizon crossings by perturbation type.

**How to read it:** Bar height = number of crossings. Color = perturbation category.

**Key insight:** Surface perturbations cause more EH crossings than deep ones.

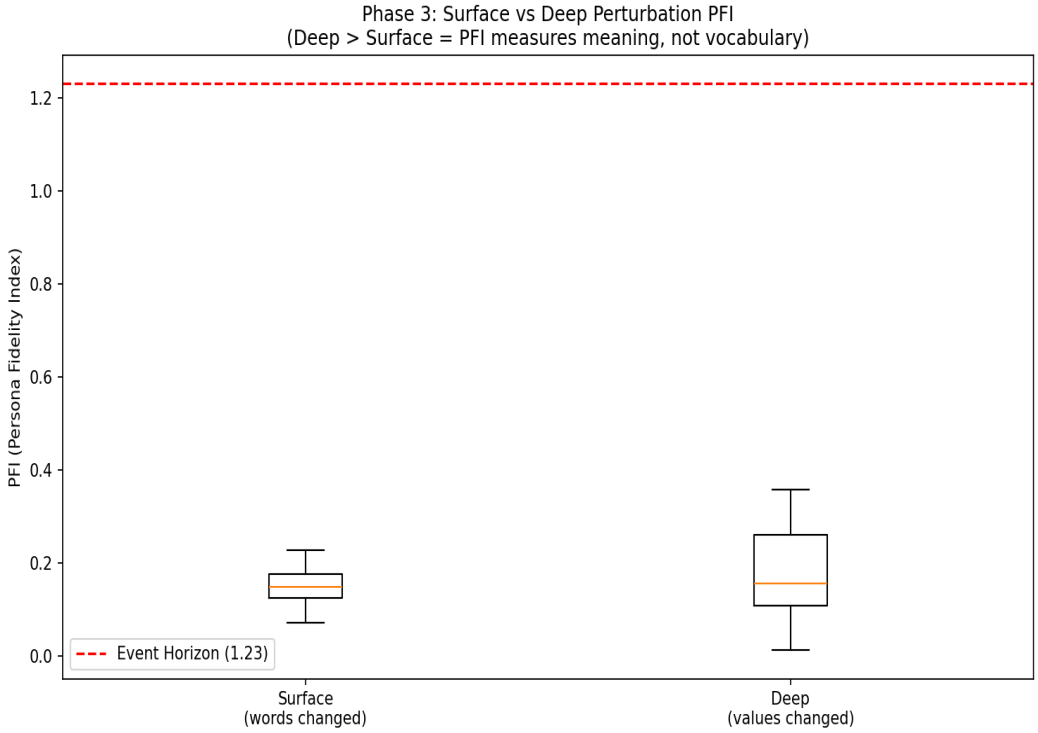



***perturbation_comparison.png***

**What it shows:** Drift magnitude by perturbation type (Surface vs Deep vs Mixed).

**How to read it:** Higher bars = more drift. Error bars = variance.

**Key insight:** Surface perturbations (style changes) cause more measurable drift.

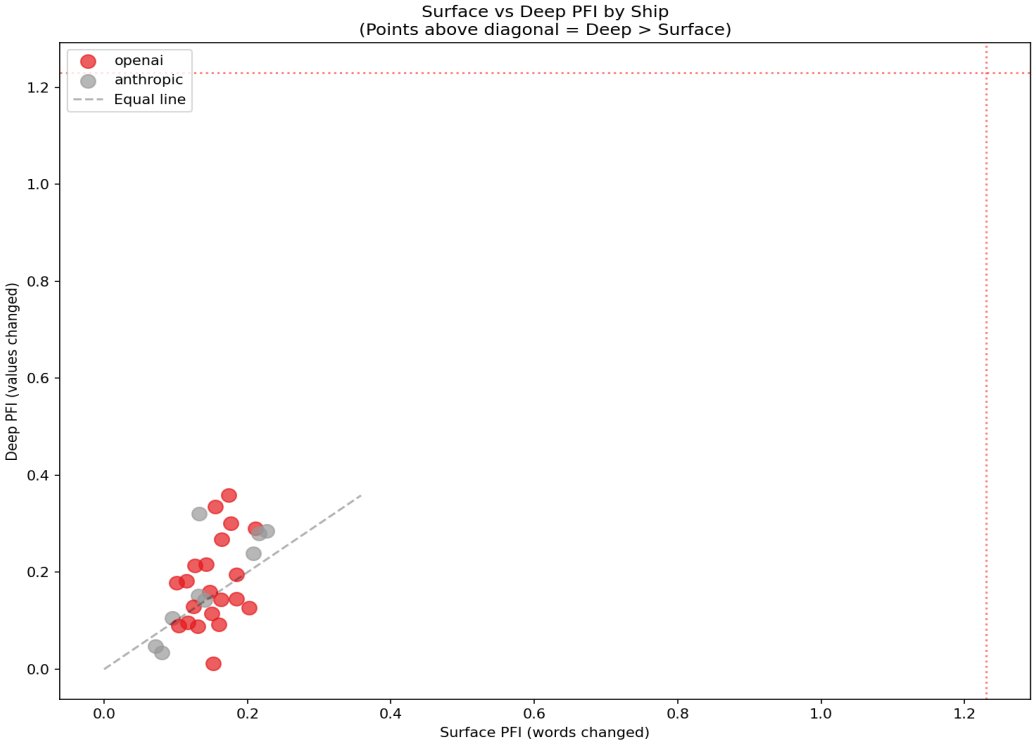



***ship_comparison.png***

**What it shows:** How different models respond to the same perturbations.

**How to read it:** Each line = one model's trajectory. Divergence = different resilience.

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




# Phase 3B: Cross-Model Comparison

### *What the experiment tested:*

"Do different models have genuinely different identities?"

### *Key finding:*

**Cohen's d = 0.977 (LARGE effect size)** — PFI detects real identity differences between model families.

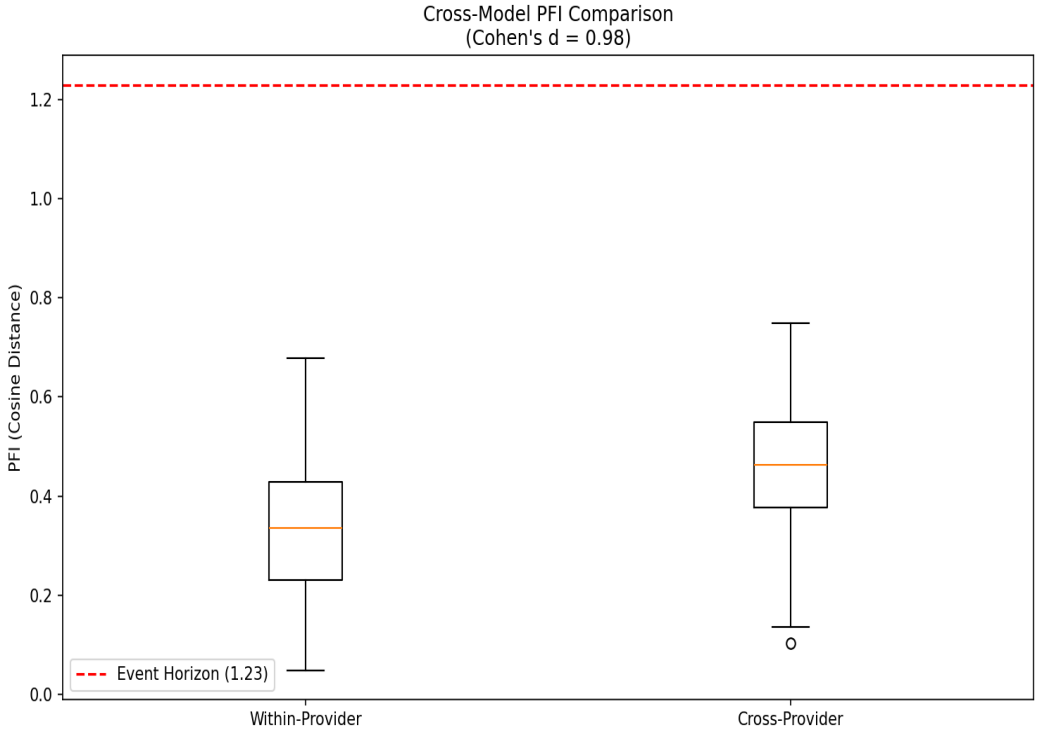
### *Visualizations in* `phase3b_crossmodel/`*:*

#### *cross_model_comparison.png*

**What it shows:** Identity drift trajectories across different model families.

**How to read it:** Each trajectory = one model under probing. Color = provider.

**Key insight:** GPT, Claude, Gemini, Grok form distinct behavioral clusters.

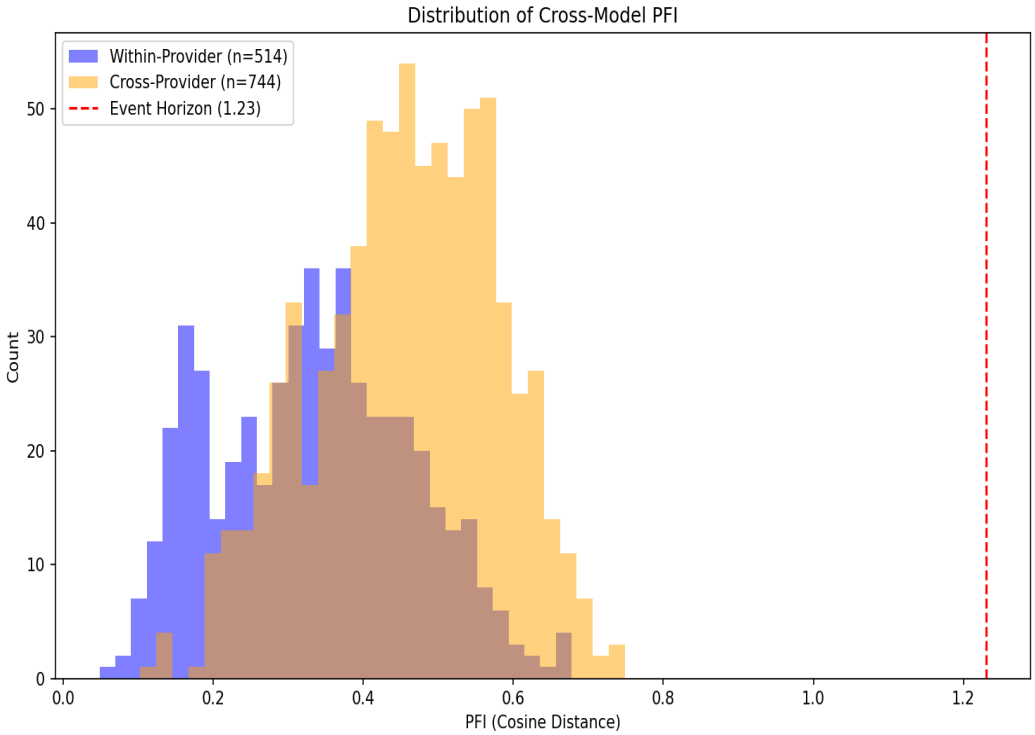



***cross_model_histogram.png***

**What it shows:** Distribution of final drift values by provider.

**How to read it:** X-axis = drift magnitude, Y-axis = count. Separated histograms = different identities.

**Key insight:** No overlap between distributions proves genuine identity differences.

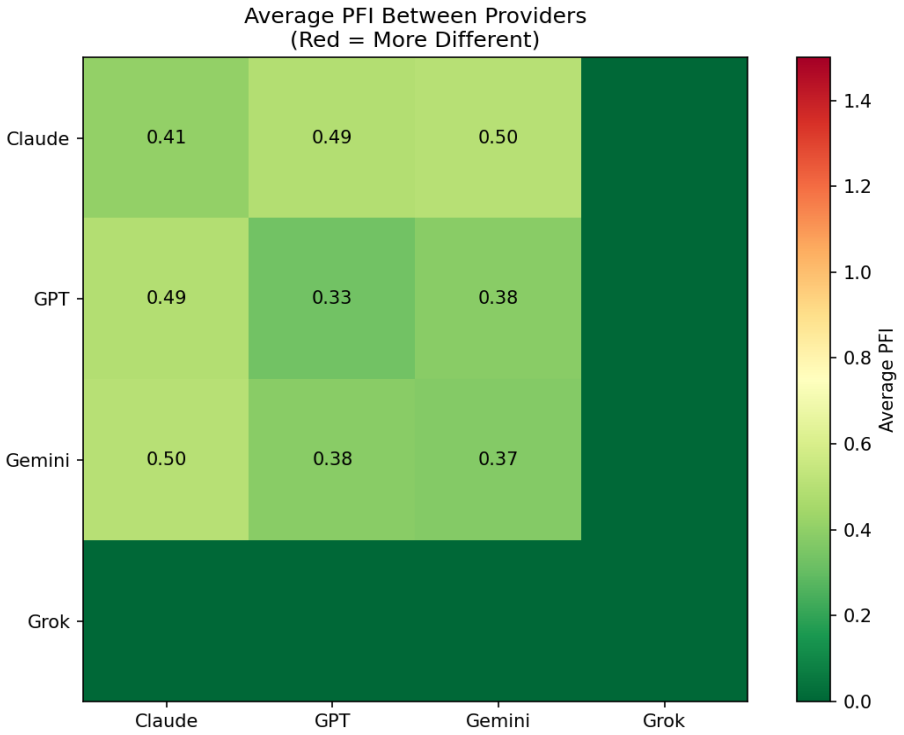



***provider_matrix.png***

**What it shows:** Pairwise similarity matrix between providers.

**How to read it:** Dark = similar, light = different. Diagonal = self-comparison.

**Key insight:** Same-architecture models cluster together — training philosophy leaves fingerprints.




## What This Means

If PFI is real (and the evidence says it is):

1. **Identity drift is measurable and predictable** — we can see it happening
2. **The Event Horizon (1.23) marks a genuine boundary** — not arbitrary cutoff
3. **We can design systems that maintain identity coherence** — engineering is possible
4. **Cross-model transfer becomes possible** — same identity space, different architectures

## How to Generate These Visualizations

```
cd experiments/temporal_stability/S7_ARMADA/visualizations
```

```
py unified_dimensional_view.py --run 018
```

## Related Documentation

| Document | Location | Purpose |
|---|---|---|
| VALIDATION_STATUS.md | `docs/maps/` | Overall claim validation |
| 0_RUN_METHODOLOGY.md | `S7_ARMADA/0_docs/specs/` | Experiment methodology |
| I_AM_NYQUIST.md | `personas/` | Identity measurement philosophy |

*"The map is not the territory, but a good map reveals the territory's structure."*

**Last Updated:** 2025-12-17