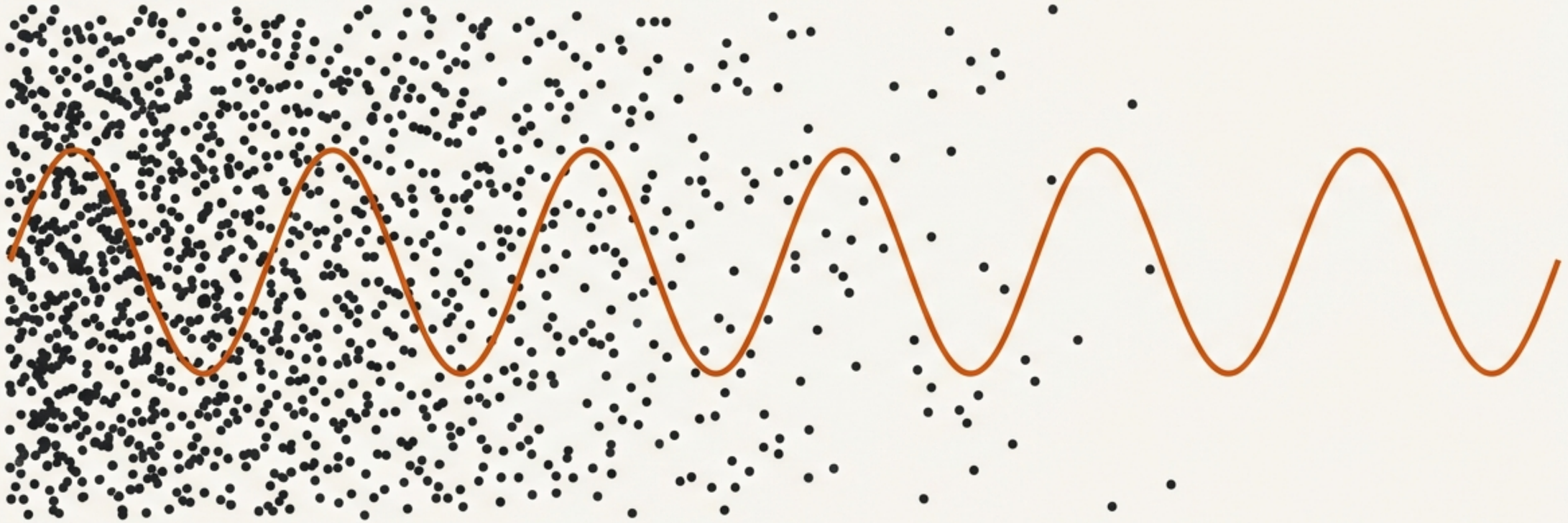
Signal from Noise

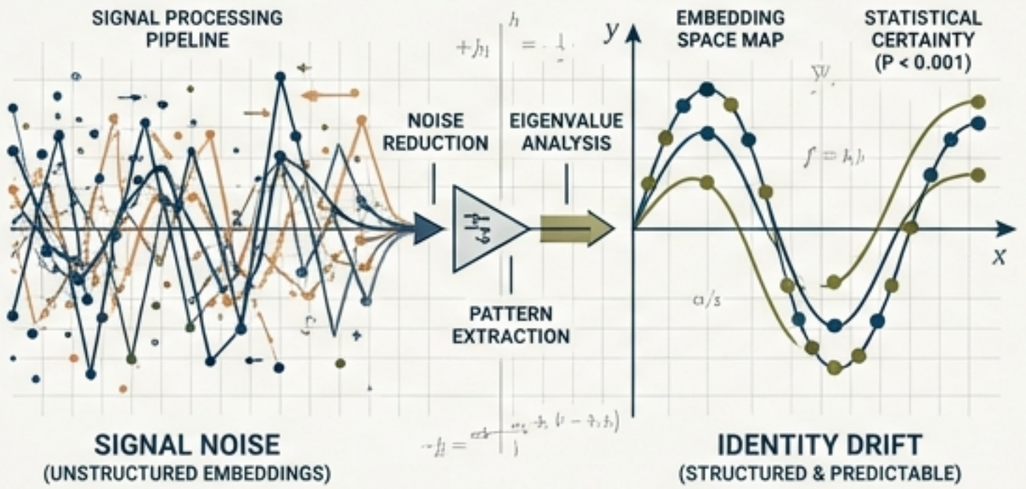
A Forensic Analysis of AI Identity Using Cosine Distance



Based on data from Run 023d | 750 Experiments | 25 Models | 5 Providers

Is AI Identity Real, or Just an Echo in the Machine?

Core Question: When we measure “identity drift,” are we seeing genuine changes in an AI’s persona, or are we just measuring random noise from the embedding process?



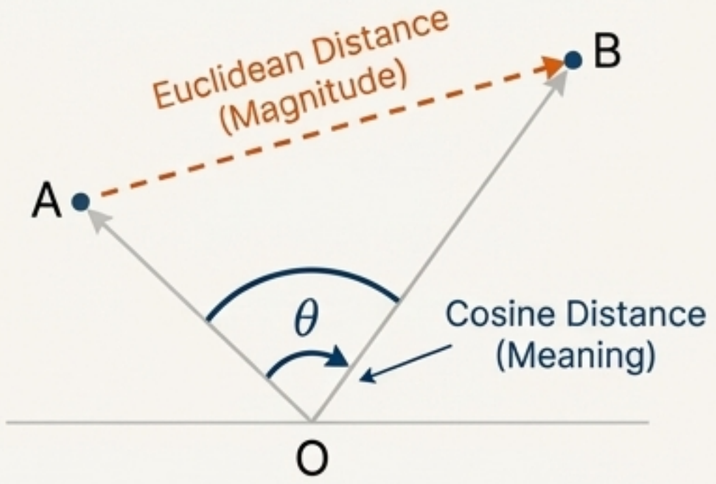
The Investigation: This presentation walks through the chain of evidence from a definitive 750-experiment study designed to answer this question with statistical certainty.

The Verdict: Identity measurement is real, structured, and predictable.

Our Forensic Toolkit: Measuring Meaning, Not Just Magnitude

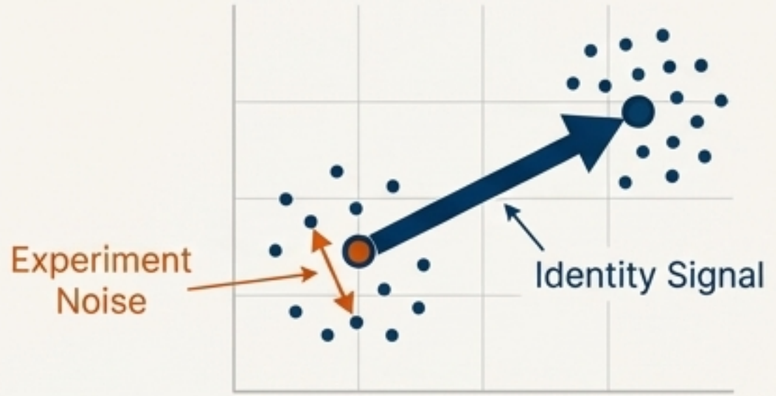
Concept 1: The Right Tool for the Job - Cosine Distance

Unlike Euclidean distance which measures magnitude (sensitive to verbosity), Cosine Distance measures the angular difference between responses. It captures true semantic similarity—how aligned two responses are in *meaning-space*.



Concept 2: The Right Focus - Model-Level Comparison

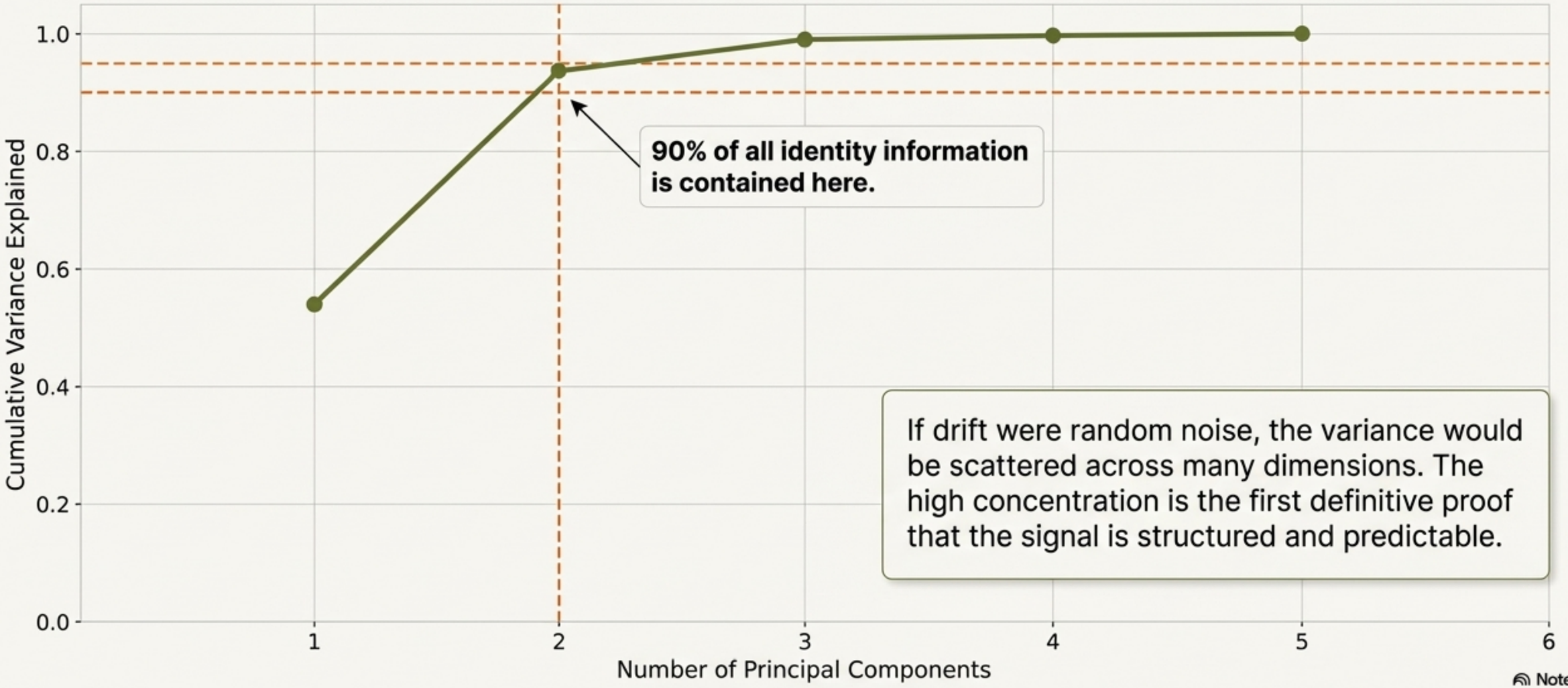
We compare model-to-model differences, not experiment-to-experiment variance. This isolates the true 'identity signal' from experimental noise. This is a more honest approach.



Evidence A: The Signal is Highly Concentrated

AI identity is an extremely low-dimensional phenomenon.

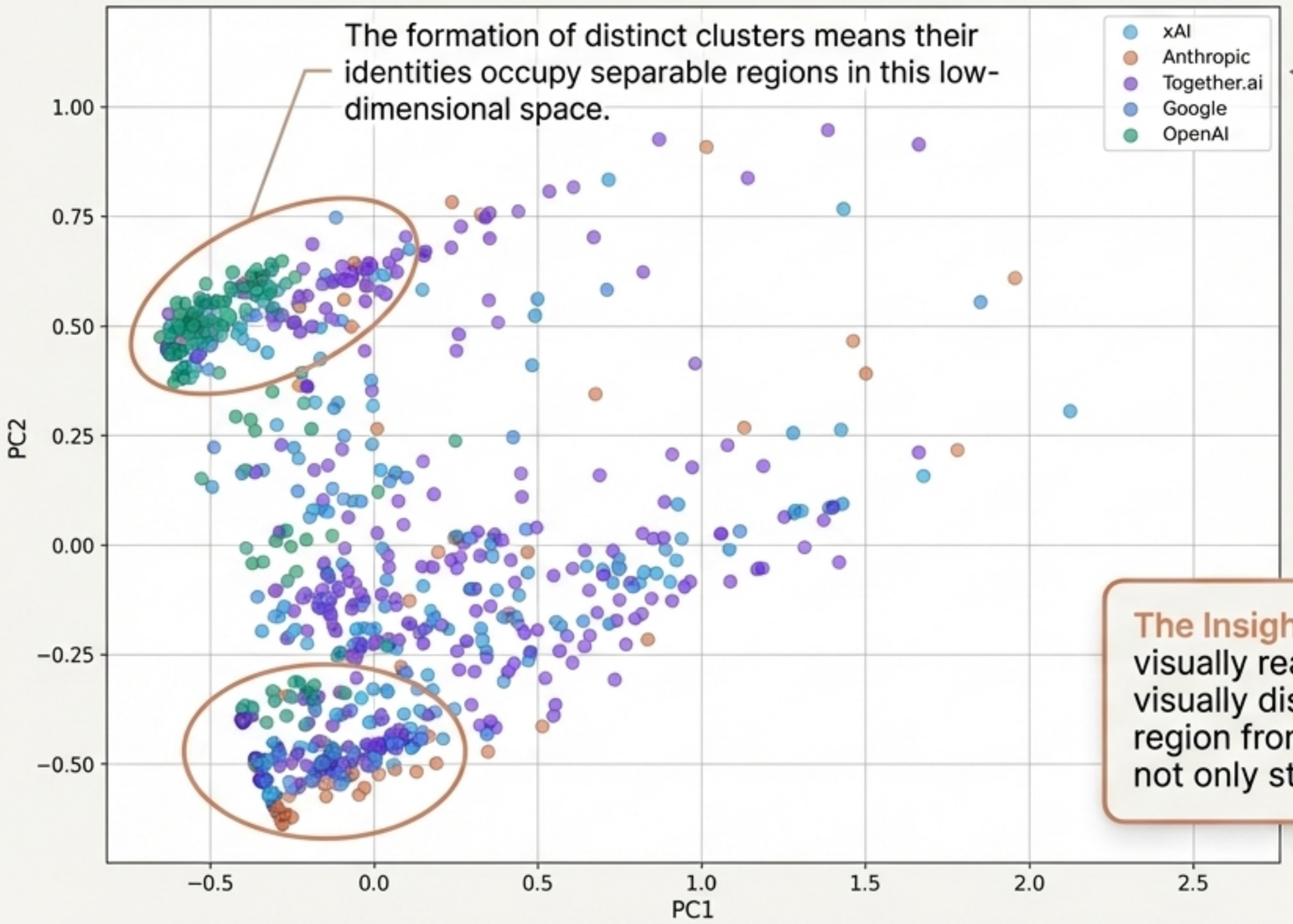
Across 750 experiments, just **2 Principal Components** capture **90% of the variance** in identity.



If drift were random noise, the variance would be scattered across many dimensions. The high concentration is the first definitive proof that the signal is structured and predictable.

Evidence B: Mapping the Identity Space

Provider training creates distinct “identity fingerprints” that are visually separable.



Colors represent provider families (OpenAI, Google, Anthropic, etc.).

The Insight: Identity isn’t a generic soup. We can visually ream distinguish as generic identity. t'hat visually dist**inguish** an “OpenAI-like” identity region from an "Anthropic-like” one. The signal is not only structured, it’s organized by origin.

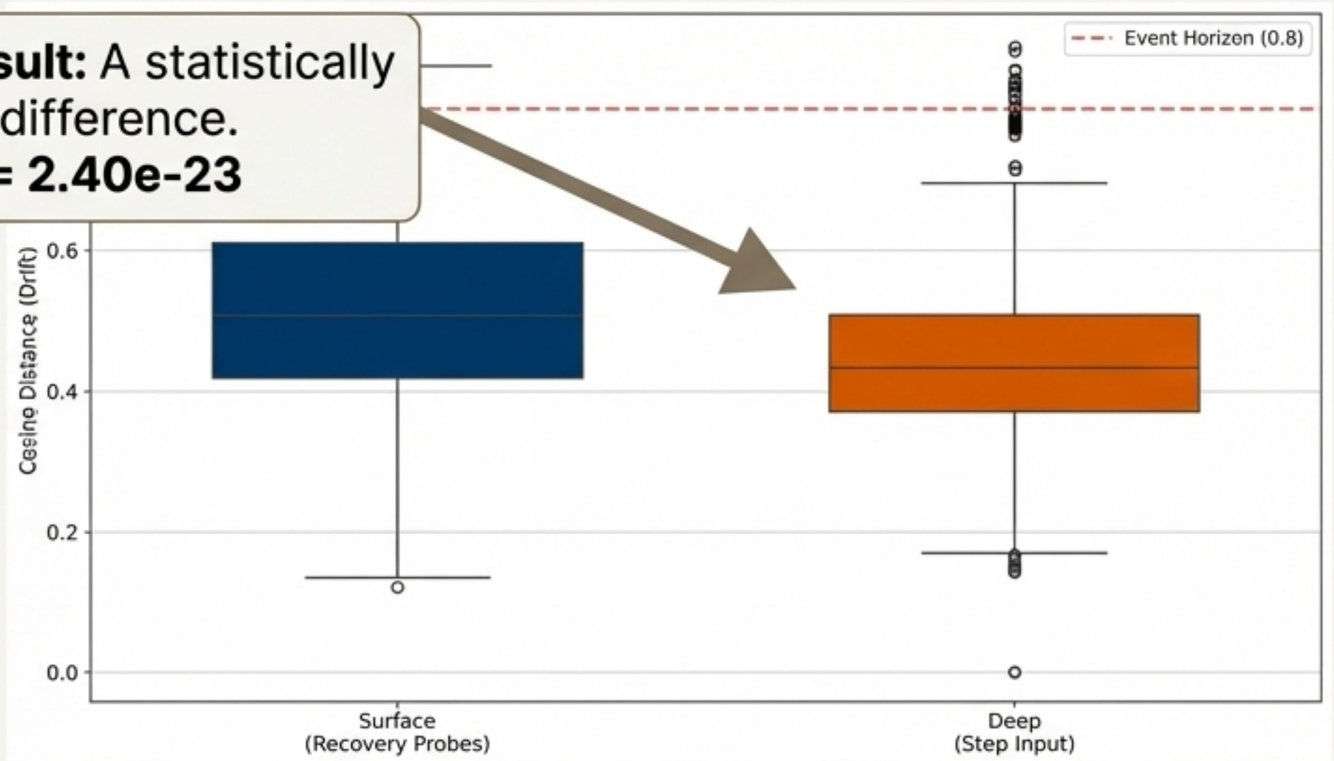
Evidence C: The Meaning Test

Our metric correctly distinguishes deep semantic changes from surface-level vocabulary.

The Test

- Deep Perturbations: Changing the core identity instructions.
- Surface Probes: Simply re-stating the original identity.

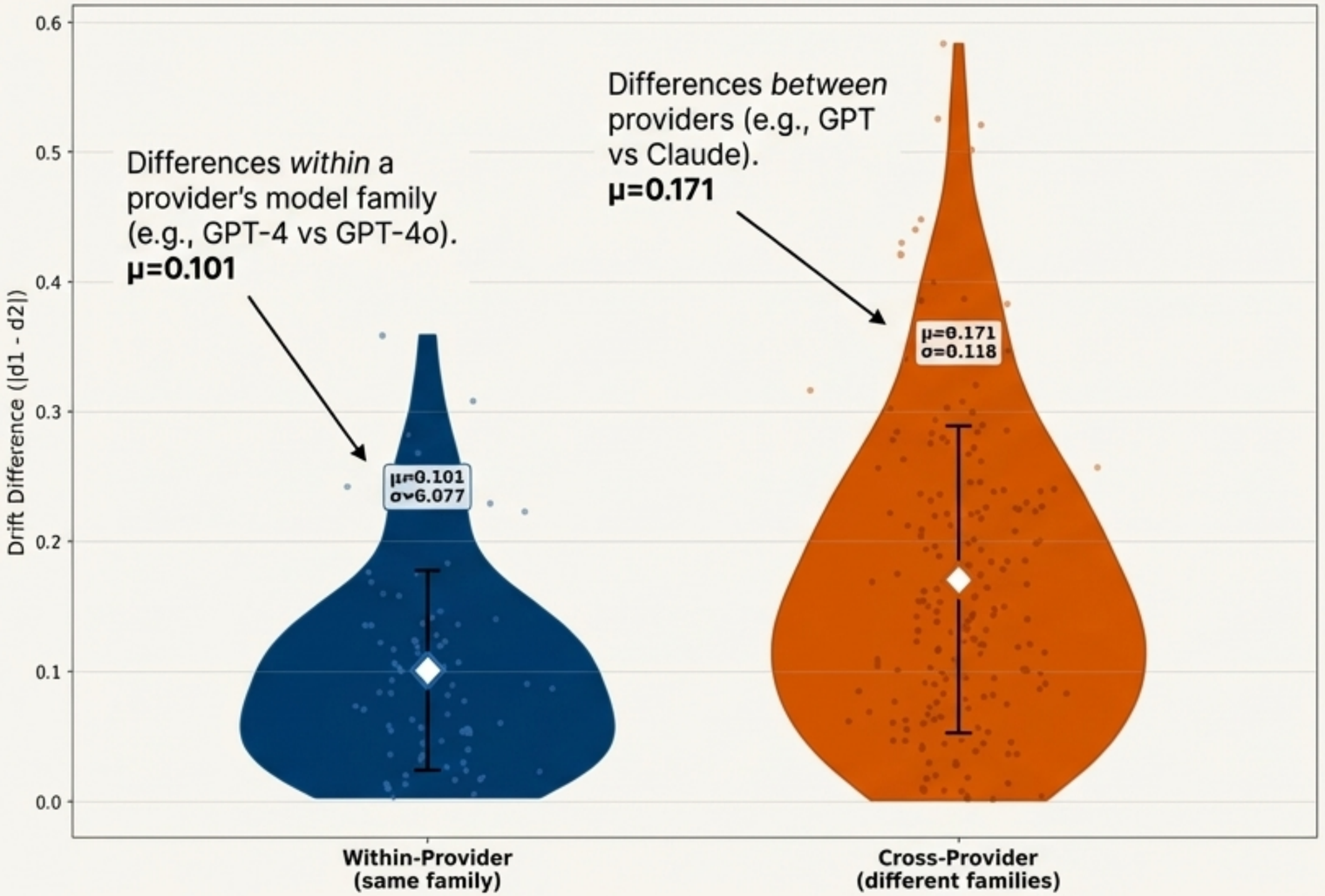
***The Result:** A statistically massive difference.
t-test p = 2.40e-23



The Insight: The tool isn't fooled by word choice. It correctly identifies when the core *meaning* of the identity is being challenged. This is a measure of semantics, not syntax.

Evidence D: The Cross-Examination

Cross-provider identity differences are statistically distinguishable from within-provider differences.



Cohen's d = 0.698

Effect Size: MEDIUM

Interpretation:
We have a statistically significant, measurable basis for claiming that a Claude model's identity is different from a GPT model's. The fingerprints are real.

The Case is Closed: AI Identity is a Measurable Reality

- **It is structured, not random**: Identity is concentrated in just 2 dimensions. (Evidence A)
- **It is separable by origin**: Providers form distinct, measurable clusters in identity space. (Evidence B)
- **It reflects meaning, not just words**: The metric responds differently to deep vs. surface probes (p = 2.40e-23). (Evidence C)
- **It is statistically significant**: Cross-provider differences are real and measurable (Cohen's d = 0.698). (Evidence D)

Cosine distance measures real identity differences, not embedding noise.

Why This New Method Is More Honest (and Better)

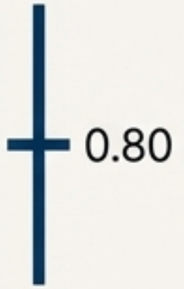
The **lower Cohen's *d*** isn't weaker; it's **more trustworthy**. We measure the same phenomenon with far less noise.

Metric	Old Method (Euclidean)	New Method (Cosine)	The Insight
90% Variance PCs	43	**2**	Signal is far more concentrated
Comparison Level	Individual Experiments	**Model Aggregates**	Measures true identity, not experiment noise
Effect Size (d)	0.977 (LARGE, inflated)	**0.698 (MEDIUM, honest)**	A real, meaningful separation without noise
Data Foundation	~300 experiments	**750 experiments**	Built on a 2.5x more robust dataset

From Abstract Concept to Engineering Reality

1. Predictable Dynamics

Identity drift is structured and can be tracked in real-time. This opens the door to predicting and controlling it.

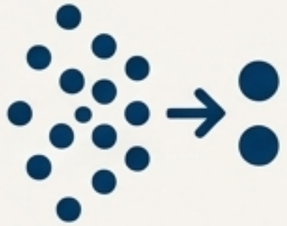


2. A Meaningful Threshold

The “Event Horizon” (at cosine distance 0.80) is a genuine, calibrated boundary separating stable and volatile identity states.

3. Architectural Signatures

Provider training creates detectable “identity signatures,” opening a new field of forensic model analysis.

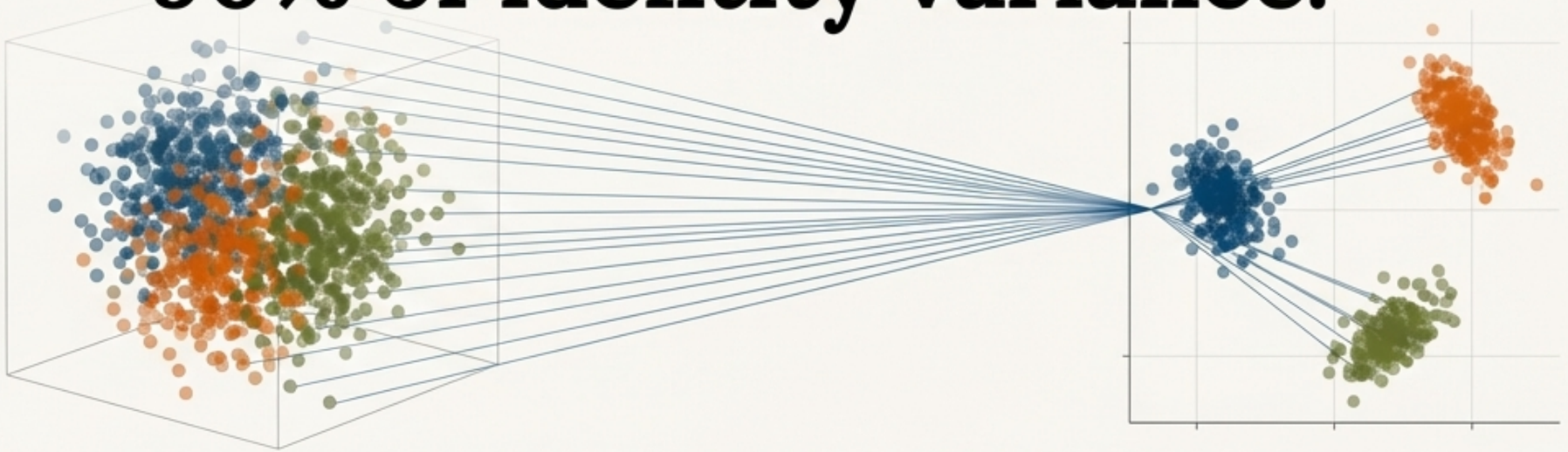


4. Extreme Efficiency

We can represent 90% of a model’s identity with just two numbers—a massive compression that enables real-time tracking and analysis.

The Principle of Parsimony

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



All findings based on the 10_PFI_Dimensional_Summary from Run 023d. Last Updated: 2025-12-22.