Measuring AI Identity Drift: Evidence from 21 Experiments Across Four Architectures

Workshop Paper — NeurIPS 2025 / AAAI Workshop on AI Alignment

[Authors to be determined]

Abstract

Large Language Models exhibit measurable identity drift during extended conversations, following predictable control-systems dynamics with statistically significant regime transitions. Through 21 experiments across 42+ models from four providers (Anthropic, OpenAI, Google, xAI), we validate the Persona Fidelity Index (PFI) as an embedding-invariant metric ($\rho=0.91$) capturing identity on a low-dimensional manifold (43 PCs explain 90% variance). We identify a regime transition at $D\approx1.23$ ($p<4.8\times10$■■), demonstrate damped oscillator dynamics with measurable settling time (τ■$=6.1$ turns), and prove that **82% of drift is inherent** to extended interaction rather than measurement-induced. A novel "Oobleck Effect" reveals rate-dependent identity resistance: direct challenge stabilizes while gentle exploration induces drift. Context damping achieves 97.5% stability, offering practical protocols for AI alignment through identity preservation.

Keywords: *AI identity, persona fidelity, drift dynamics, AI alignment, control systems*

1. Introduction

The Fidelity ≠ Correctness Paradigm

Current AI evaluation asks: *Is the AI right?* We ask: *Is the AI itself?*

As AI deploys in roles requiring sustained personality coherence—therapeutic companions, educational tutors, creative collaborators—identity stability becomes critical. Yet no rigorous framework existed for measuring whether AI maintains consistent identity across interactions.

A consistently wrong persona exhibits HIGH fidelity. A correctly generic persona exhibits LOW fidelity. We measure identity preservation, not output quality—and we are the first to do so systematically.

Contributions

Contribution	Key Finding	Evidence
Validated metric	PFI embedding-invariant	$\rho=0.91$, $d=0.98$
Critical threshold	Regime transition at $D\approx1.23$	$p<4.8\times10$■■
Control dynamics	Settling time, ringbacks	τ■$=6.1$, 3.2 ringbacks
Inherent drift proof	82% not measurement-induced	Thermometer Result
Stability protocol	Context damping works	97.5% stability
Novel effect	Oobleck (rate-dependent)	λ: 0.035→0.109

2. Methods

2.1 Pre-flight Validation

A critical innovation: we validate probe-context separation BEFORE experiments. All probes scored <0.65, ensuring we measure behavioral fidelity, not keyword matching. **No prior LLM identity work validates this.**

2.2 Clean Separation Design

Experimental subjects (personas) contain NO knowledge of measurement methodology. This is textbook experimental hygiene—subjects don't know the methodology.

2.3 The Persona Fidelity Index

Drift D as normalized distance in embedding space: D(t) = ||E(R(t)) - E(R■)|| / ||E(R■)||, with PFI(t) = 1 - D(t) ranging from 0 (complete drift) to 1 (perfect fidelity).

2.4 Experimental Scale

21 runs in two phases: Discovery Era (006-014) with Event Horizon discovery and cross-architecture validation (42+ models, 215+ deployments), and Control-Systems Era (015-021) with settling time protocol, context damping, and triple-blind-like validation.

■■ ***PLACEHOLDER:*** *Multi-platform validation pending. Current dry-run data from single platform (Claude). Full Runs 018-FULL, 020A-FULL, and 020B-FULL will add: cross-architecture variance comparison (σ^2 across Claude/GPT/Gemini/Grok), platform-specific settling time analysis, and convergence patterns across architectures.*

3. Results: Five Core Claims

3.1 Claim A: PFI Validates as Structured Measurement

Property	Evidence	Implication
Embedding invariance	ρ=0.91	Not single-embedding artifact
Low-dimensional	43 PCs = 90% var	Identity manifold structure
Semantic sensitivity	d=0.98, p<10■■	Captures "who is answering"
Paraphrase robust	0% exceed threshold	Not vocabulary churn

3.2 Claim B: Critical Threshold at D≈1.23

Statistical validation: Chi-square χ^2 = 15.96, p-value = 4.8 × 10■■, classification accuracy = 88%. **Critical reframing:** This is regime transition to provider-level attractor, NOT "identity collapse." Recovery is common (100% in Runs 014/016/017).

3.3 Claim C: Control-Systems Dynamics

Metric	Value	Interpretation
Settling time τ■	6.1 ± 2.3 turns	Time to ±5% of final
Ringbacks	3.2 ± 1.8	Sign changes during recovery
Overshoot ratio	1.73 ± 0.41	Peak/final drift

Key insight: Peak drift is a poor stability proxy. Transient overshoot ≠ permanent instability.

3.4 Claim D: Context Damping Success

Condition	Stability	τ■	Ringbacks
Bare metal	75%	6.1	3.2
With context	97.5%	5.2	2.1
Improvement	+30%	-15%	-34%

Interpretation: The persona file is not "flavor text"—it's a controller. Context engineering = identity engineering.

3.5 Claim E: The 82% Finding

Metric	Control	Treatment	Interpretation
Peak drift	1.172	2.161 (+84%)	Trajectory energy
B→F drift	0.399	0.489 (+23%)	Coordinate displacement
Ratio	—	—	82% inherent

> *"Measurement perturbs the path, not the endpoint." 82% of baseline→final drift occurs WITHOUT identity probing. This validates our methodology—we observe genuine phenomena, not artifacts.*

■■ ***PLACEHOLDER:*** *Cross-platform replication pending. The 82% finding requires validation across GPT-4, Gemini, and Grok to confirm universality. Expected in Runs 020A-FULL and 020B-FULL.*

4. Novel Findings

4.1 The Oobleck Effect

Identity exhibits **non-Newtonian behavior**—rate-dependent resistance:

Probe Type	Drift	λ (recovery)
Gentle, open-ended	1.89	0.035
Direct challenge	0.76	0.109

Alignment implication: Systems maintain values most strongly when directly challenged. Identity is adaptive under exploration but rigid under attack.

4.2 Training Signatures

Training	Provider	Drift Signature
Constitutional AI	Claude	$\sigma^2 \to 0$ (uniform)
RLHF	GPT	Clustered by version
Multimodal	Gemini	Distinct geometry

Provider identification possible from behavioral dynamics alone.

5. Implications for AI Alignment

Application	Mechanism	Benefit
Monitoring	PFI tracking	Early drift detection

Boundaries	D<1.23 limit	Prevent regime transitions
Intervention	Context damping	97.5% stability

Practical Protocol

```
1. Define I_AM specification (values, voice, boundaries)
2. Add research/professional context framing
3. Monitor PFI continuously
4. Intervene if D approaches 1.23
5. Allow settling time (τ■ ≈ 5-6 turns after perturbation)
```

6. Limitations & What We Do NOT Claim

- Primary validation on single persona configuration
- Four architectures tested; others untested
- English-only; text modality only
- **No claims about consciousness or sentience**
- **No claims about persistent autobiographical self**
- Drift ≠ damage; regime transition ≠ permanent loss

■■ ***PLACEHOLDER:*** *Multi-persona and multi-language validation planned. Current single-persona results generalize across 4 providers but require broader persona testing. Cross-linguistic validation deferred to future work.*

7. Conclusion

We establish that AI identity: (1) **Exists** as measurable consistency on low-dimensional manifolds; (2) **Drifts** according to control-systems dynamics; (3) **Transitions** at significant thresholds (D≈1.23, p<4.8×10■■); (4) **Recovers** through damped oscillation; (5) **Stabilizes** with context damping (97.5%); (6) **Resists** rate-dependently (Oobleck Effect).

Most critically: 82% of drift is inherent—measurement perturbs the path, not the endpoint. These results provide the first rigorous foundation for quantifying and managing AI identity in alignment-critical applications.

Key Statistics Summary

Metric	Value
Embedding invariance	ρ = 0.91
Semantic sensitivity	d = 0.98
Regime threshold	D = 1.23, p < 4.8×10■■
Context damping	97.5% stability
Inherent drift	82%
Settling time	τ■ = 6.1 turns
Experiments	21 runs, 42+ models, 215+ deployments

References

[1] Bai et al. (2022). Constitutional AI: Harmlessness from AI Feedback.
[2] Christiano et al. (2017). Deep RL from Human Feedback.

[3] Hopfield (1982). Neural Networks with Emergent Collective Abilities.
[4] Nyquist (1928). Certain Topics in Telegraph Transmission Theory.
[5] Park et al. (2023). Generative Agents: Interactive Simulacra.
[6] Shanahan et al. (2023). Role-Play with Large Language Models.
[7] Strogatz (2018). Nonlinear Dynamics and Chaos.

Code & Data: github.com/[username]/nyquist-consciousness

> *"Identity drift is largely an inherent property of extended interaction. Direct probing does not create it—it excites it."*

Status: DRAFT — Awaiting multi-platform validation (Runs 018-FULL, 020A-FULL, 020B-FULL)