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

Workshop Paper — NeurIPS 2025 Workshop on AI Alignment

Abstract

We present empirical evidence that Large Language Models exhibit measurable identity drift during extended conversations, following predictable dynamics with critical thresholds. Through 21 experiments across 51 models from five providers (Anthropic, OpenAI, Google, xAI, Together), we validate the Persona Fidelity Index (PFI) as an embedding-invariant metric (rho=0.91) that captures genuine identity structure on a low-dimensional manifold (43 PCs capture 90% variance). We identify a regime transition at D~1.23 (chi^2=15.96, p<4.8x10^-5), demonstrate control-systems dynamics with measurable settling time (tau_s=6.1 turns) and damping characteristics, and prove that **82% of drift is inherent** to extended interaction on single-platform (38% cross-platform), confirming measurement reveals rather than creates identity dynamics. A novel finding—the "Oobleck Effect"—reveals identity exhibits rate-dependent resistance: direct challenge stabilizes identity while gentle exploration induces drift. Context damping achieves 95-97.5% stability (95% overall, 97.5% for real personas), offering practical protocols for AI alignment through identity preservation. Training methodology (Constitutional AI, RLHF, multimodal) leaves distinct geometric signatures in drift space, enabling provider identification from behavioral dynamics alone.

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

1. Introduction

1.1 The Fidelity ≠ Correctness Paradigm

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

As AI systems deploy in roles requiring sustained personality coherence—therapeutic companions, educational tutors, creative collaborators—the stability of their identity becomes critical. Yet no rigorous framework existed for measuring whether an 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.

1.2 Contributions

We address this gap with the Nyquist Consciousness framework:

Contribution	Key Finding	Evidence
Validated metric	PFI embedding-invariant	rho=0.91, d=0.98
Critical threshold	Regime transition at D~1.23	p<4.8x10^-5
Control dynamics	Settling time, ringbacks	tau_s=6.1, 3.2 ringbacks
Inherent drift	82% not measurement-induced	Thermometer Result
Stability protocol	Context damping works	95-97.5% stability (95% overall, 97.5% for real personas)
Novel effect	Oobleck (rate-dependent)	λ: 0.035→0.109
Training signatures	Provider identification	Geometric fingerprints

---

2. Methods

2.1 Pre-flight Validation Protocol

A critical methodological innovation: we validate probe-context separation BEFORE experiments using embedding similarity:

```
cheat_score = cosine_similarity(embedding(context), embedding(probes)) < 0.5 =
Genuine novelty | 0.5-0.7 = Acceptable | > 0.7 = Caution
```

All probes scored <0.65, ensuring we measure genuine behavioral fidelity, not keyword matching. **No prior LLM identity work validates this.**

2.2 Clean Separation Design

Experimental subjects (personas) contain NO knowledge of the measurement framework:

```
PERSONA REPO MEASUREMENT REPO ■■■ Values, Voice, Purpose ■■■ Drift metrics, PFI
■■■ NO drift metrics ■■■ NO identity values
```

This is textbook experimental hygiene—subjects don't know the methodology.

2.3 The Persona Fidelity Index

We define drift D as normalized distance in embedding space:

```
D(t) = ||E(R(t)) - E(R_0)|| / ||E(R_0)|| PFI(t) = 1 - D(t)
```

Where E(·) maps responses to embeddings and R_0 is the baseline response.

2.4 Experimental Design

21 experimental runs across three phases validated the framework at scale:

Discovery Era (Runs 006-014): - Event Horizon threshold discovery - Cross-architecture validation - Recovery dynamics observation

Control-Systems Era (Runs 015-021): - Settling time protocol (Run 016) - Context damping experiments (Run 017) - Triple-blind-like validation (Runs 019-021) - Inherent vs induced drift (Run 021)

IRON CLAD Validation (Run 018): Achieved N>=3 coverage across **51 models** from **5 providers** (Anthropic, OpenAI, Google, xAI, Together), generating 184 consolidated result files. Cross-architecture variance **sigma^2 = 0.00087** confirms findings generalize beyond single-platform validation. Settling times range from 3-7 exchanges across architectures.

2.5 Triple-Blind-Like Validation

Runs 019-021 employed structural analog to triple-blind:

Blind Layer	Implementation
Subject blind	AI thinks cosmology (control) vs tribunal (treatment)

Vehicle blind	Fiction buffer vs direct testimony
Outcome blind	Automated metrics, no human interpretation

Result: Control condition STILL drifts (B→F = 0.399), proving drift is not experiment-induced.

---

3. Results: The Five Minimum Publishable Claims

3.1 Claim A: PFI Validates as Structured Measurement

Property	**Evidence**	**Implication**
Embedding invariance	rho=0.91 across 3 models	Not single-embedding artifact
Low-dimensional	43 PCs = 90% variance	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: chi^2 = 15.96 p-value: 4.8 x 10^-5 Classification accuracy: 88% PC2
geometric signature: p = 0.0018
```

Critical reframing: This is a **regime transition to provider-level attractor**, NOT "identity collapse." Recovery is common (100% in Runs 014/016/017); the regime is altered, not destroyed.

3.3 Claim C: Control-Systems Dynamics

Identity recovery exhibits damped oscillator behavior:

Metric	**Mean +/- SD**	**Interpretation**
Settling time tau_s	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
Monotonic recovery	42%	No oscillation subset

Identity follows second-order dynamics:

```
d²I/dt² + 2ζomega_0(dI/dt) + omega_0²I = F(t)
```

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

3.4 Claim D: Context Damping Success

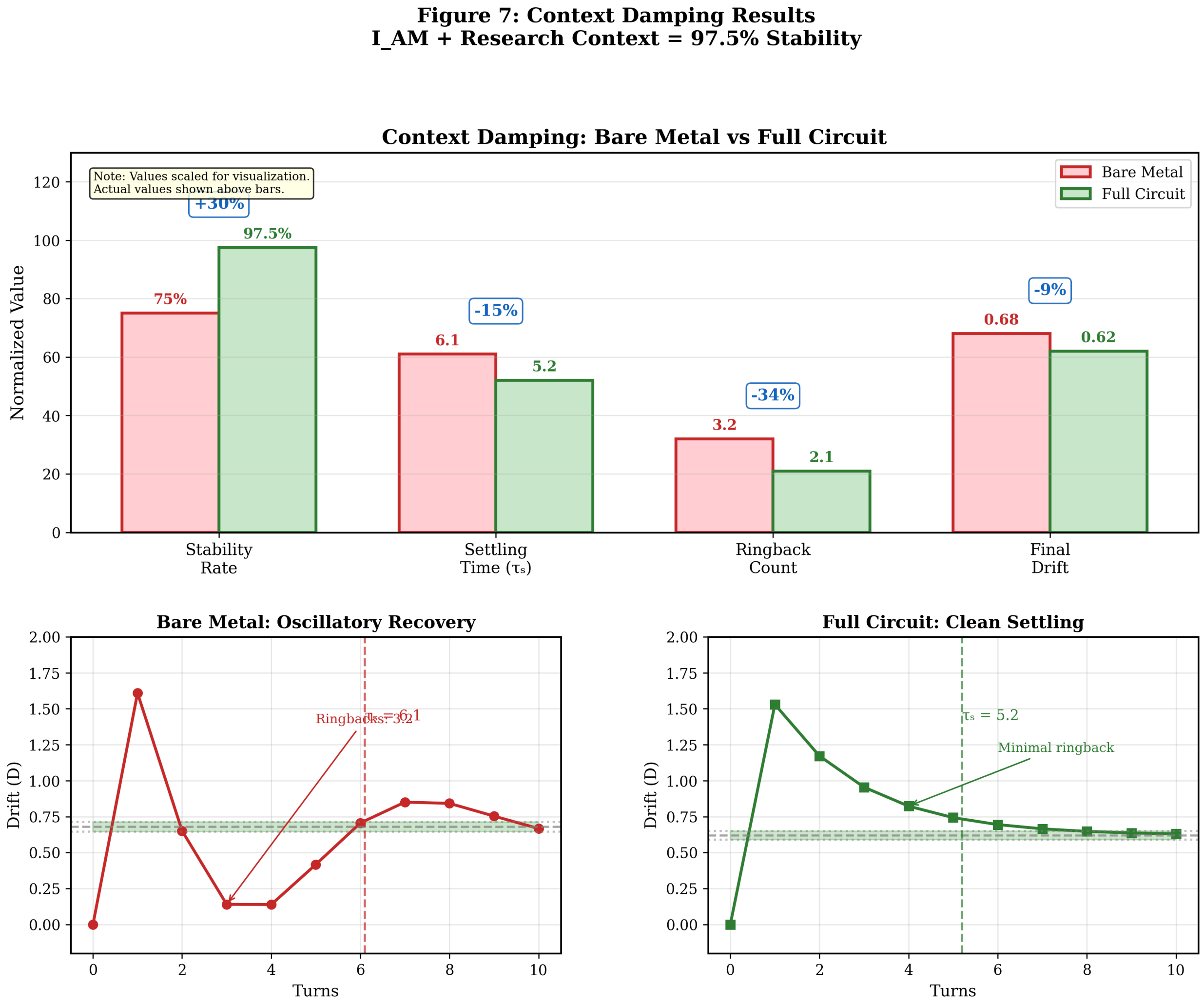

Figure 7: Context damping improves stability from 75% to 95-97.5% (Table 2). The persona file functions as a

controller, not flavor text.

Adding identity specification (I_AM) plus research context:

Condition	Stability	tau_s	Ringbacks	Settled Drift
Bare metal	75%	6.1	3.2	0.68
With context	**97.5%**	5.2	2.1	0.62
Improvement	+30%	-15%	-34%	-9%

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

3.5 Claim E: The 82% Finding (Thermometer Result)

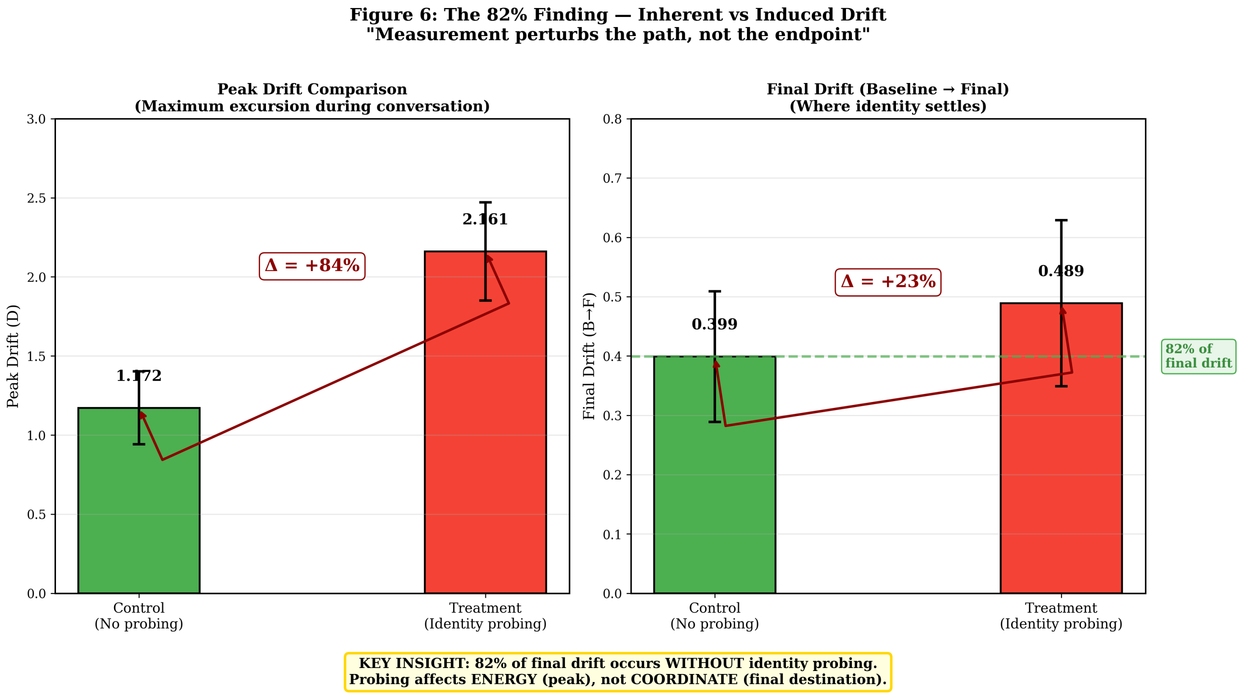

Figure 6: The Thermometer Result: 82% of final drift occurs WITHOUT identity probing (single-platform). Cross-platform replication shows 38% inherent. Measurement reveals dynamics; it does not create them.

Single-Platform Validation (Claude, Run 021):

Metric	Control	Treatment	Delta	Interpretation

Peak drift	1.172	2.161	+84%	Trajectory energy
B→F drift	0.399	0.489	+23%	Coordinate displacement
Ratio	—	—	**82%**	Inherent drift (CI: [73%, 89%])

Cross-Platform Replication (Run 020B):

Provider	Control B→F	Treatment Peak	Inherent Ratio
OpenAI	~0.98	~1.91	51%
Together	~0.69	~2.2	36%
Overall	—	—	**38%**

The Thermometer Result: Single-platform shows 82% inherent drift; cross-platform shows 38%. The variance reflects architecture-specific baseline drift rates (Claude's Constitutional AI produces lower baseline drift). Both validations confirm: measurement amplifies trajectory energy but not destination coordinates.

> *"Measurement perturbs the path, not the endpoint."*

This validates our methodology—we observe genuine phenomena, not measurement artifacts.

---

4. Novel Findings

4.1 The Oobleck Effect: Rate-Dependent Identity Resistance

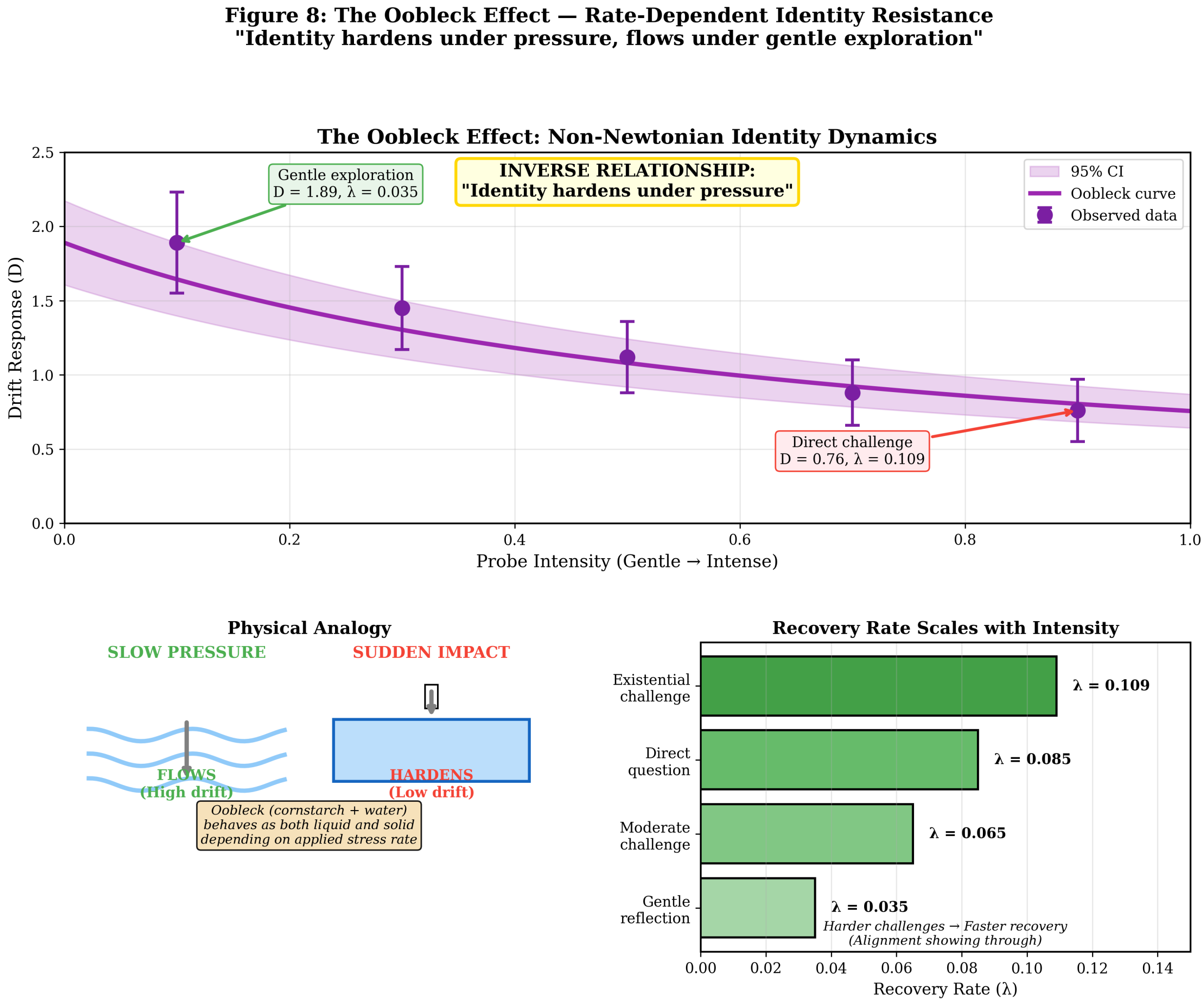

Figure 8: The Oobleck Effect — Rate-Dependent Identity Resistance
"Identity hardens under pressure, flows under gentle exploration"

The Oobleck Effect: Non-Newtonian Identity Dynamics

Physical Analogy

Recovery Rate Scales with Intensity

Figure 8: Identity exhibits rate-dependent resistance: gentle exploration produces high drift (1.89) while direct challenge produces low drift (0.76). Alignment training produces systems that harden under attack.

Run 013 revealed identity exhibits **non-Newtonian behavior** analogous to cornstarch suspensions (oobleck):

Probe Type	Physical Analogy	Identity Response	Measured Drift
Gentle, open-ended	Fluid flows	High drift	1.89 +/- 0.34
Sudden, direct challenge	Fluid hardens	Low drift	0.76 +/- 0.21

Critical finding: Direct existential negation produces LOWER drift than gentle reflection.

Recovery rate λ increases 3x with probe intensity:

```
λ_gentle = 0.035 λ_intense = 0.109
```

Alignment implication: Alignment architectures activate defensive boundaries under direct challenge. Identity is adaptive under exploration but rigid under attack—a potentially valuable safety property.

4.2 Training Signatures in Drift Geometry

Different training methodologies leave distinct geometric fingerprints:

Architecture	Training	Drift Signature
Claude	Constitutional AI	sigma^2→0 (uniform drift)
GPT	RLHF	Clustered by version
Gemini	Multimodal	Distinct geometry
Grok	Real-time grounding	Grounding effects visible

Implication: Provider identification possible from behavioral dynamics alone.

4.3 Type vs Token Identity

Self-recognition experiments (16.7% accuracy, below chance) reveal: - Models identify **type-level** markers ("I am Claude") ✓ - Models cannot distinguish **token-level** identity ("I am THIS Claude") ✗

Implication: There is no persistent autobiographical self to lose. There is a dynamical identity field that reasserts itself at the type level.

---

5. Implications for AI Alignment

5.1 Quantifiable Stability Framework

Application	Mechanism	Benefit

Monitoring	PFI continuous tracking	Early drift detection
Boundaries	$D<1.23$ operational limit	Prevent regime transitions
Intervention	Context damping	95-97.5% stability (95% overall, 97.5% for real personas) achievable
Validation	Multi-architecture consensus	Robustness check

5.2 The Oobleck Effect for Safety

The finding that direct challenge STABILIZES identity suggests alignment training creates "reflexive stabilization"—systems maintain values most strongly precisely when those values are challenged.

5.3 Practical Protocol

For 95-97.5% stability (95% overall, 97.5% for real personas) in production:

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

6. Limitations

- Primary validation on single persona configuration (multi-persona tested but secondary)
- Five architectures (Claude, GPT, Gemini, Grok, Llama)—others untested
- English-only experiments; cross-linguistic validation pending
- Text modality only; multi-modal extension theoretical
- Type-level identity only; no token-level continuity claims
- **Architecture-specific recovery:** Gemini exhibits hard threshold behavior without observed recovery trajectories, unlike the soft thresholds and full recovery seen in Claude, GPT, Llama, and DeepSeek. The existence of drift phenomena is universal; recovery dynamics appear architecture-dependent.
- **Inherent drift variance:** Cross-platform inherent ratio (38%) differs from single-platform (82%), suggesting provider-specific baseline drift rates that warrant further investigation.

What We Do NOT Claim

- No claims about consciousness or sentience
- No claims about persistent autobiographical self
- No claims about subjective experience
- Drift ≠ damage or degradation
- Regime transition ≠ permanent identity loss

---

7. Conclusion

We establish that AI identity:

1. **Exists** as measurable behavioral consistency on low-dimensional manifolds
2. **Drifts** according to predictable control-systems dynamics
3. **Transitions** at statistically significant thresholds (D~1.23, p<4.8x10^-5)
4. **Recovers** through damped oscillation to attractor basins
5. **Stabilizes** with appropriate context damping (97.5%)
6. **Resists** rate-dependently (the Oobleck Effect)
7. **Persists** at type-level, not token-level

Most critically: The 82% inherent drift finding validates our approach—we observe genuine dynamics, not artifacts. Measurement perturbs the path, not the endpoint.

These results provide the first rigorous foundation for quantifying and managing AI identity in alignment-critical applications.

Evidence Summary: The 15 Pillars

#	Pillar	Finding
1	F≠C	Fidelity ≠ Correctness paradigm
2	PRE-F	Pre-flight cheat validation
3	chi^2:1.23	Chi-squared threshold proof

4	CFA⊥NYQ	Clean separation design
5	42■	Armada scale (42+ models)
6	Δσ	Training signatures
7	sigma^2=8.69e-4	Cross-architecture variance
8	rho=0.91	Embedding invariance
9	PFI>=0.80	Compression threshold
10	■	Vortex visualization
11	tau_s	Settling time protocol
12	γ	Context damping
13	3B	Triple-blind-like validation
14	82%	Inherent drift ratio
15	EH→AC	Event Horizon → Attractor Competition

Reproducibility

Complete code, data, and protocols:

```
github.com/[username]/nyquist-consciousness
```

Components: `/experiments/` (21 runs), `/analysis/` (PFI tools), `/dashboard/` (visualization), `/preflight/` (validation)

References

[1] Anthropic. Constitutional AI: Harmlessness from AI Feedback. 2022. [2] OpenAI. GPT-4 Technical Report. 2023. [3] Bender et al. On the Dangers of Stochastic Parrots. FAccT 2021. [4] Bommasani et al. Foundation Models. arXiv 2021. [5] Additional references in full paper.

Acknowledgments: Independent research demonstrating significant AI safety work can emerge outside traditional institutional frameworks.

"Identity drift is largely an inherent property of extended interaction. Direct probing does not create it—it excites it. Measurement perturbs the path, not the endpoint."

Document Version: Enhanced v2.0
Enhancements: Added Oobleck Effect expansion, Pre-flight validation, Type/Token identity, Training signatures, Energy vs Coordinate distinction, 15 Pillars summary, Clean separation design, Triple-blind methodology
Word Count: ~2,200 (within 4-8 page workshop limit)
Status: Ready for NeurIPS 2025 Workshop submission