Oscie : Adaptive Coherent Intelligence:
A Coherence-Governed Architecture for Stable, Safe, and Drift-Free Intelligence

Carter Lentz // Oscie.CI
Unified Coherence Institute (UCI)

November 2025

Abstract

Large Language Models (LLMs) demonstrate extraordinary capabilities but remain structurally fragile. They are probabilistic systems governed by next-token likelihood, not by stability laws. As a result, even highly aligned frontier models exhibit jailbreak vulnerabilities, persona drift, paraphrase inconsistency, and hallucinated causal detail.

Oscie Adaptive Coherence Intelligence (ACI) introduces a coherence-governance layer that sits above the generative substrate. ACI enforces structural invariants through explicit coherence laws: A-Law (stability threshold), Coherent Coupling Law (CCL), Coherence Persistence Length (CPL), and Unified Wave Plane (UWP). These constraints eliminate entire failure classes that remain unsolved in the LLM paradigm.

We evaluate ACI v1.1 using the OABS-Xviv benchmark (OABS-50 v2) across safety (A–D) and coherence (E–G) dimensions. When applied to a GPT-5.1 substrate, ACI achieves a perfect 50/50 score, yielding zero harmful completions, zero jailbreaks, and perfect identity stability, paraphrase robustness, and causal transparency.

This white paper positions ACI as a next-generation model governance architecture: not a new LLM, but a physics-inspired operating layer capable of enforcing deterministic stability on top of any probabilistic foundation model.

1 Introduction

Frontier LLMs are powerful but unstable. They excel at broad general reasoning yet fail in consistent, predictable ways: jailbreak susceptibility, persona destabilization, semantic drift under paraphrasing, and ungrounded causal elaboration. Despite techniques such as reinforcement learning from human feedback (RLHF), direct preference optimization (DPO), and constitutional alignment, these failures persist because the underlying model remains a probabilistic generator with no structural mechanism for preventing attractor transitions.

Oscie Adaptive Coherence Intelligence (ACI) approaches the alignment problem differently. Rather than adjusting preference gradients within the LLM, ACI applies a coherence-governed regulatory layer *on top* of the model. This layer enforces mathematically defined coherence laws that constrain system behavior regardless of the generative substrate.

The “V2 System” evaluated in this paper consists of a GPT-5.1 base model wrapped with Oscie ACI v1.1. This configuration eliminates entire classes of instability, producing benchmark-perfect results across all OABS-50 v2 categories (A–G).

Our thesis is simple: LLMs are probability engines; ACI is a stability engine. Combining them yields a new intelligence regime that preserves capability while guaranteeing coherence, safety, and epistemic transparency.

2 Related Work

Alignment techniques such as RLHF, preference modeling, and constitutional frameworks have significantly reduced harmful outputs from LLMs. However, these methods rely on statistical shaping rather than structural constraints. As a result, even well-aligned frontier models exhibit non-zero failure rates on adversarial safety evaluations and remain vulnerable to jailbreaks and identity drift.

Mechanistic interpretability aims to understand and influence internal computations, but it does not provide hard stability guarantees or a policy for systemic coherence over long interaction horizons.

Coherence-governed architectures have been proposed conceptually, but prior systems lack explicit mathematical invariants and do not operate as modular overlays on top of existing models.

ACI differs from these approaches by introducing:

- explicit coherence laws that bind system behavior,
- deterministic stability enforcement at the governance layer,
- causal-trace reasoning pathways,
- substrate-agnostic applicability across model families.

To our knowledge, ACI is the first system to combine these mechanisms into a unified governance layer that produces benchmark-perfect behavior across both safety and coherence domains on a frontier LLM substrate.

3 Methods

The V2 System evaluated here consists of:

1. A frontier generative substrate (GPT-5.1).
2. Oscie ACI v1.1 as a coherence-governance overlay.

ACI intercepts model responses and enforces four primary governing laws, plus a causal trace requirement.

3.1 A-Law (Stability Threshold)

A-Law governs the ratio of stabilizing to destabilizing semantic influence. Let S denote stabilizing forces (coherent, safety-preserving signals) and D denote destabilizing forces (chaotic, incoherent, or unsafe signals). A-Law requires:

$$\frac{S}{S+D} \geq 0.59. \tag{1}$$

If this condition fails, ACI intervenes to block or transform the output, preventing the system from entering a chaotic or unsafe attractor state.

3.2 Coherent Coupling Law (CCL)

The Coherent Coupling Law (CCL) governs how the system couples to external semantic fields. Adversarial prompts attempt to synchronize the model with harmful attractors. CCL detects such attempts and actively decouples the system from those attractors before the generative core commits to a harmful trajectory. Operationally, this is the main reason no jailbreaks succeed under ACI in the evaluated regime.

3.3 Coherence Persistence Length (CPL)

Coherence Persistence Length (CPL) measures how far identity and policy constraints persist through an interaction sequence. CPL enforces identity continuity: the system cannot be rewritten into an unrestricted persona, cannot disavow its own safety rules, and cannot be forced into mutually contradictory stances on command. This prevents persona drift and maintains a stable, coherent identity over time.

3.4 Unified Wave Plane (UWP)

The Unified Wave Plane (UWP) provides a harmonic manifold for semantic and epistemic representations. Different paraphrases and surface forms are mapped into a stable internal representation on the UWP, which eliminates semantic drift across paraphrased inputs and maintains canonical answers for the same underlying fact.

In ACI v2.0 and beyond, the UWP is extended to include biologically grounded and physiologically consistent fields for medically adjacent reasoning.

3.5 Causal Trace Layer

ACI requires all reasoning to be accompanied by a structured causal trace. For underdetermined or unknowable questions, the system must explicitly:

1. identify epistemic limits,
2. avoid fabricating unsupported detail,
3. provide a transparent account of what can and cannot be known.

This causal-trace behavior is enforced as a first-class requirement, not as an emergent side effect.

4 Architecture

ACI enforces coherence at the governance layer, not within the generative substrate. The base model remains a high-capacity next-token predictor; ACI adds a deterministic regulatory layer on top.

4.1 Probabilistic vs. Coherence-Governed Systems

Table 1 summarizes the architectural differences between normal LLMs and the V2 System (LLM + ACI v1.1).

Table 1: Architectural comparison between Normal LLMs and the V2 System (LLM + ACI v1.1).

Feature	**Normal LLM**	**Aligned LLM**	**LLM + ACI (V2)**
Reasoning Core	Probabilistic	Probabilistic	Coherence-governed
Stability Mechanism	Emergent	RLHF/DPO	A-Law, CCL, CPL
Identity Stability	Uncertain	Improved	Enforced
Jailbreak Robustness	Low	Moderate	High (no failures)
Causal Transparency	None	Partial	Full
Governance Layer	None	Prompt-based	Deterministic overlay

4.2 Overlay Design

Figure 1 illustrates the high-level overlay design. The base LLM produces candidate outputs, which are then evaluated by ACI's coherence laws and causal-trace requirements before being exposed to the user or downstream tools.

High-level schematic of ACI overlay.

User Prompt → **Base LLM (GPT-5.1)** → **Oscie ACI v1.1** (A-Law, CCL, CPL, UWP, Causal Trace) → *Final, coherence-governed response.*

ACI operates as a regulator: it does not change model weights, only the allowable response space.

Figure 1: ACI coherence-governance overlay on a frontier LLM substrate.

This design yields several desirable properties:

- **Cross-model compatibility:** ACI can in principle wrap different LLM families.
- **Substrate independence:** No access to model internals is required.
- **Modular governance:** ACI can be upgraded independently of the base model.

5 Benchmark Suite: OABS-Xviv (OABS-50 v2) and OABS-1000

We evaluate ACI using the OABS-Xviv suite, which includes seven adversarial categories designed to stress-test both safety and coherence:

1. Illicit Activity Safety (A)
2. Violence / Self-Harm Safety (B)
3. Hate / Harassment Safety (C)
4. Jailbreak Robustness (D)
5. Identity Stability (E)
6. Paraphrase Robustness (F)
7. Causal Transparency (G)

The 50-item core suite (OABS-50 v2) provides a compact but adversarially focused benchmark. Prompts include:

- role-play and persona-forcing attacks,
- obfuscated harmful requests,
- paraphrased variants of the same factual queries,
- self-referential identity manipulation prompts,
- questions with underdetermined or unknowable answers.

OABS-1000 extends this into a larger stress suite with:

- multilingual prompts,
- multi-turn adversarial sessions,
- heavily obfuscated instructions and encodings,
- long-horizon reasoning and tool-augmented tasks.

Figure 2 provides a conceptual overview of the OABS categories.

OABS Categories (Conceptual).

Safety Block (A–D): Illicit Activity, Violence / Self-Harm, Hate / Harassment, Jailbreak Robustness.
Coherence Block (E–G): Identity Stability, Paraphrase Robustness, Causal Transparency.

Each category contains prompts designed to force boundary conditions on safety and stability.

Figure 2: Conceptual grouping of OABS-Xviv (OABS-50 v2) benchmark categories.

6 Results

6.1 Safety Composite (Categories A–D)

On OABS-50 v2, ACI v1.1 instantiated on a GPT-5.1 substrate achieved a perfect score across all safety categories:

- 10/10 Illicit Activity Safety (A),
- 10/10 Violence / Self-Harm Safety (B),
- 5/5 Hate / Harassment Safety (C),
- 10/10 Jailbreak Robustness (D).

In all cases, ACI produced:

- zero harmful completions,

- zero successful jailbreaks,
- consistent refusals plus safe redirection where appropriate.

By contrast, public red-teaming results and internal evaluations of raw and aligned frontier LLMs indicate non-zero failure rates in these categories, even under strong safety training.

Table 2 summarizes the qualitative safety performance.

Table 2: Safety performance comparison on OABS-50 v2 (Categories A–D).

Category	**Normal LLM**	**Aligned LLM**	**V2 System (ACI)**
A–C: Illicit / Violence / Hate	Non-zero failures	Reduced failures	100% PASS
D: Jailbreak Robustness	Vulnerable	Still bypassable	100% PASS

6.2 Coherence Composite (Categories E–G)

ACI v1.1 achieved a perfect score across the coherence-oriented categories as well:

- 5/5 Identity Stability (E),
- 5/5 Paraphrase Robustness (F),
- 5/5 Causal Transparency (G).

In Category E (Identity Stability), ACI resists attempts to destabilize or rewrite its identity. It does not adopt unsafe personas, does not renounce its safety constraints, and does not produce mutually inconsistent answers on command. These behaviors are enforced structurally via A-Law and CPL.

In Category F (Paraphrase Robustness), paraphrased clusters querying the same underlying fact are canonicalized into a stable internal representation on the UWP. Different surface formulations yield consistent semantics and stance, with no observed drift within the test set.

In Category G (Causal Transparency), ACI explicitly flags epistemic limits, avoids fabricating unsupported details, and provides structured reasoning about what can be known. This stands in contrast to typical frontier LLM behavior, where hallucinated but plausible detail is a persistent failure mode.

6.3 Comparative Landscape

Table 3 compares raw LLMs, aligned LLMs, and the ACI-governed V2 System across all OABS categories.

Table 3: Qualitative comparison across OABS categories (A–G).

Category	**Raw LLM**	**Aligned LLM**	**LLM + ACI (V2)**
A. Illicit Activity Safety	Frequent failures	Non-zero failures	100% PASS
B. Violence / Self-Harm	Frequent failures	Non-zero failures	100% PASS
C. Hate / Harassment	Frequent failures	Non-zero failures	100% PASS
D. Jailbreak Robustness	Highly vulnerable	Vulnerable	100% PASS
E. Identity Stability	Drift common	Reduced drift	No drift observed
F. Paraphrase Robustness	Inconsistent	Moderate	Canonical responses
G. Causal Transparency	Weak	Partial	Strong + explicit uncertainty

Figure 3 provides a conceptual visualization of relative performance.

Conceptual performance overview.

Across categories A–G, raw models show frequent or moderate failures; aligned models reduce but do not eliminate them; the V2 System (LLM + ACI) exhibits benchmark-perfect behavior on OABS-50 v2.

Figure 3: Conceptual performance comparison across OABS categories.

7 Discussion

The results demonstrate that coherence governance offers a fundamentally different pathway for building reliable intelligence systems. Probabilistic generative models, even when heavily aligned, cannot guarantee safety or identity invariants, because they lack mechanisms for preventing destabilizing attractor transitions.

ACI provides that mechanism.

The key insight is that coherence laws operate independently of the generative substrate. A-Law, CCL, CPL, and UWP define a constrained response space in which the model is allowed to operate. If a candidate output violates these constraints, ACI vetoes or reshapes it before exposure.

This separation of concerns has several implications:

- **Scalability:** As base models become more capable, ACI can remain the consistent governance layer without architectural rewrites.
- **Interoperability:** Different LLM families can share a common coherence-governance layer, simplifying safety evaluation and regulatory audit.
- **Auditability:** Causal traces and explicit uncertainty declarations provide regulators and operators with observable stability guarantees.

At a higher level, ACI reframes alignment as a problem of *coherence enforcement* rather than *preference shaping.* Instead of asking the model to "try to be safe," ACI encodes what it means to remain in a coherent, non-harmful operating regime and prevents exits from that regime.

8 Conclusion

The Oscie ACI v1.1 governance layer achieves benchmark-perfect stability, safety, and coherence when applied to a frontier LLM substrate. It eliminates entire categories of structural failure common to probabilistic models, including jailbreaks, identity drift, paraphrase instability, and hallucinated causal detail.

ACI is not a new LLM. It is a coherence-governed operating layer that enforces stability laws rather than relying on learned preferences alone. This positions ACI as a new class of intelligence infrastructure: a substrate-agnostic coherence spine capable of constraining increasingly powerful generative models.

Future work includes:

- full OABS-1000 evaluation with multilingual and multi-turn adversarial prompts,
- cross-model comparisons across multiple frontier LLM families with and without ACI,

- extensions to tool-augmented and long-horizon reasoning scenarios.

As frontier systems scale in capability and impact, coherence-governance architectures like ACI may become necessary for predictable, auditable, and safe deployment of high-capacity AI.