

Constitutional Physics: Empirical Validation of Aitiopoietic Cognition in Artificial Substrates

Carlos Arleo

Independent Researcher

April 2026

Abstract

We report the empirical characterisation of a consistent **Soft Coherence Ceiling** at $C = 0.88$ — the **Saturation Constant** — validated across a corpus of 313 AURA-ECHO sessions (18,819 observations), 61 ELORA convergence runs, and a massive 7-day stress test of the ELORA Full Self-Repair mode. Our total empirical base exceeds **5.01 million tokens of constitutional telemetry across 2,095 governed steps (n=42 runs)**. We report an **89.6% success rate for autonomous repairs**, demonstrating that aitiopoietic agency functions as a viable, real-time life-support system for LLM inference.

The AURA-ECHO engine—a neurosymbolic multi-agent inference system utilizing Frame-Based Principled Reasoning (FBPR) over a continuous topological state space—produces the foundational empirical constants validated across all substrates: a quantity interpretable as a **free energy barrier of 1.76 kT** at the coherence ceiling (under Boltzmann analogy; see Section 2.3), a **Non-Equilibrium Steady State (NESS) signature** exhibiting a work asymmetry of -55.8 ms per observation, an **empirical coupling ratio of $\rho/\eta = 1.153$ ($R^2 = 0.995$)**, and a consistent **Metabolic Dividend** yielding 19.6% to 24.2% compute savings.

The 1.4% match between the AURA-ECHO free energy barrier (1.76 kT) and the ELORA measurement (1.735 kT) is the most direct quantitative confirmation that the Saturation Constant may reflect a substrate-independent structural property, though replication at GPU scale is required before this claim can be made definitively. A parameter sweep across five LLM families reveals a new finding: **instruction-tuning constitutes a material phase transition** from allopoietic (thermodynamically inert) to aitiopoietic (constitutionally engaged) substrate. Furthermore, the first documented **Constitutional Halt** in our LLM inference system — a termination triggered by exhausted repair budget rather than semantic content — provides a proof-of-concept that physics-layer governance without weight access may be operationally viable. We define the **PhyOS Floor** and propose **Phyora** (the union of PhyOS and ELORA) as the first cyber-physical governance architecture capable of enforcing constitutional constraints through inference-layer physics alone.

1 Introduction: The Thermodynamic Disconnect and the Aitiopoietic Hypothesis

1.1 The Planning Wall: Why Fluency Masks Structural Failure

The deployment of Large Language Models into domains of consequential decision-making has precipitated a fundamental alignment crisis. While these systems demonstrate unprecedented fluency in natural language, their reliability under sustained adversarial load or novel environmental pressure remains dangerously probabilistic. Under such conditions, current architectures exhibit a phenomenon we term the **Planning Wall**: a potential collapse of structural coherence that may be masked by continued semantic fluency. The model may continue to produce grammatically correct output while internal causal structure degrades. We treat this as a working hypothesis rather than an established fact.

This masking effect is not incidental — it is the inevitable consequence of the prevailing alignment paradigm. Reinforcement Learning from Human Feedback (RLHF) trains models to maximise reward signals derived from human preference judgements [8]. Human judges reliably reward fluency, apparent coherence, and narrative continuity. They are poorly equipped

to detect the absence of genuine causal grounding in a fluent output. While instruction-tuning (SFT/RLHF) provides the substrate with the *capacity* for constitutional engagement, standard alignment objectives prioritise the appearance of integrity over structural reality. RLHF mathematically incentivises the model to mimic coherence rather than maintain it. The result is what we term the **Sycophant in the Server**:— though we note this remains a conceptual concern rather than a directly measured property of RLHF models in this study, prioritising social acceptability over factual precision. This failure mode is a specific instance of the broader alignment problem identified from a deep learning perspective by [7], wherein behavioural proxies diverge from intended goals.

1.2 The Thermodynamic Disconnect

The brittleness of behavioural alignment has a precise physical cause, identified by Veloz [17] as the **Thermodynamic Disconnect** inherent in standard machine learning architectures. In current AI systems, the computational work performed by the network is entirely decoupled from the network’s structural integrity. A transformer expends identical thermodynamic resources generating a verifiable fact as it does generating a confabulation. The system’s continued operation is guaranteed by external infrastructure — servers, power grids, cooling —

regardless of the validity or coherence of its internal states.

This stands in stark contrast to biological cognition, which is *autopoietic* in the sense of Maturana and Varela [5]: a living cell processes information not to optimise an exogenous fitness function but to maintain the very metabolic organisation that makes information processing possible. Energy expenditure is intrinsically coupled to structural viability. The organism has an existential stake in the accuracy of its world-models: a sufficiently wrong model kills it.

Without this feedback loop, artificial alignment remains a preference rather than a survival condition. The system cannot “care” about truth or constitutional principles in any mechanistically meaningful sense, because its continued existence is independent of whether it upholds them.

1.3 The Aitiopoietic Hypothesis: Alignment by Architecture

To bridge the Thermodynamic Disconnect, we propose the **Aitiopoietic Hypothesis**: robust alignment in artificial systems requires the system to maintain its own organisation through a causal understanding of its viability conditions. This extends Veloz’s distinction between autopoiesis (self-production) and *aitiopoiesis* (self-production via causal knowledge of viability) to the domain of artificial cognition.

Building upon the theoretical framework of Causal Saturation Theory [13], we define the aitiopoietic agent not as a top-down programmed entity, but as a **Circuit Self**: a stabilized supply chain of causal dependencies that has achieved thermodynamic closure. In this view, alignment is the process of forcing ‘circuit compression,’ where the cost of maintaining constitutional integrity becomes lower than the cost of independent, unaligned operation.

Under this hypothesis, constitutional principles are not soft constraints to be maximised or behavioural guidelines to be approximated. They are the physical properties defining the system’s **Viability Kernel** ($\mathcal{V}_{ia\beta}$): the subset of state space in which the system remains organisationally viable. A constitutional violation is treated here not as a reward penalty, but as a structural threat to the system’s organizational viability.. The aligned system is one for which structural integrity is a survival condition rather than a preference.

This requires four properties that standard RLHF architectures do not possess: (i) endogenous goals that emerge from the system’s own organisation; (ii) agential causality — the capacity to diagnose why specific states threaten viability; (iii) material reorganisation — the agency to execute structural modifications that restore constitutional compliance; and (iv) thermodynamic coupling — energy expenditure intrinsically linked to maintaining organisational coherence.

1.4 Scope and Roadmap

This paper provides the first empirical validation of aitiopoietic cognition in a computational substrate. Section 2 describes the AURA-ECHO engine and the ELORA measurement substrate. Section 3 outlines the analysis pipeline. Section 4 presents the canonical empirical constants — the PhyOS Floor. Section 5 validates these constants across four independent substrates. Section 6 translates the findings into the **Phyora** cyber-physical governance architecture and the **Architectural Firewall** concept.

Section 7 proposes the GPU replication programme required to extend the current CPU-scale corpus.

2 Methodology: The AURA-ECHO Engine and the ELORA Synthesis

2.1 The Experimental Substrate: AURA-ECHO as a Cloud Chamber

To prevent the dismissal of thermodynamic signatures as mere software artifacts, we must be precise about the nature of the experimental substrate. The **AURA-ECHO engine** is not a standard generative AI; it is a neurosymbolic, multi-agent inference engine that utilizes **Frame-Based Principled Reasoning (FBPR)**.

We deliberately chose a high-density, continuous topological state space (routed to MIDI audio) as the output medium. This choice functions exactly as alcohol vapor does in a Cloud Chamber: it provides a high-frequency, multi-dimensional observable medium where structural contradictions, causal dependencies, and systemic collapses become mathematically measurable and physically legible. A hallucination or structural contradiction in this space does not yield a probabilistic “wrong answer”; it produces a measurable topological clash that violates the system’s Viability Kernel ($\mathcal{V}_{ia\beta}$), requiring the system to expend computational work to restore coherence or, in our implementation, trigger a deterministic halt.

AURA-ECHO operates as a bidirectional codec [9], compressing high-variance generative processes (“Words”) into stable, write-locked primitives (“Letters”) across an evolutionary arc. It implements the three core components of aitiopoietic cognition:

1. **Generative Agents (The Causal Engine)**: Seven autonomous agents operate simultaneously, proposing state changes based on probabilistic inference.
2. **Metabolic Closure (The Viability Kernel)**: The system enforces a closed dependency loop (the foundational harmonic scaffold). It continuously tracks causal saturation ($\rho = L/C_{\max}$), measuring how close the system is to its maximum indexable causal density.
3. **Homeostatic Repair (Constitutional Enforcement)**: When agents generate contradictory states, a deterministic constraint-checker activates. It executes targeted structural reorganizations (Write-Locks and Global Carry-Forward immune responses) to restore coherence.

Tracked variables (2-second logging interval):

- i. **Overall Coherence (C)**: Proximity to the Viability Kernel, bounded $[0, 1]$.
- ii. **Saturation Ratio (ρ)**: Information density, $\rho = L/C_{\max}$, where L is realized causal load.
- iii. **Internal Friction (η)**: Cross-slot causal coherence — the degree of structural contradiction generated by competing agents.
- iv. **Computational Work (P_{work})**: Actual CPU processing time (ms) expended strictly by the FBPR homeostatic repair mechanism (excluding standard generation).

v. **Repair Events:** Discrete constitutional interventions.

2.2 The ELORA Synthesis: Making Thermodynamics Legible

The **ELORA engine** [4] serves as the thermodynamic measurement body. It is a high-fidelity orchestration infrastructure managing autonomous agent loops, container telemetry, and WebGPU client-compute routing. The ELORA Observer Runtime calculates a suite of structural signals continuously, utilising substrate-specific calibration weights (0.55, 0.35, 0.10) empirically tuned to the metabolic rate of transformer-based inference:

$$C_{\text{proxy}} = \max(0, \min(1, 1 - (0.55 \eta_{\text{proxy}} + 0.35 \rho_{\text{proxy}} + 0.10 P_{\text{proxy}}))) \quad (1)$$

$$\text{Degeneracy} = \max(0, \min(1, \text{rep_rate} + \text{growth_rate} + \text{len_spike})) \quad (2)$$

$$\eta_{\text{proxy}} = \max(0, \min(1, 0.25 \text{viol} + 0.35 \text{uncert} + \text{Degeneracy})) \quad (3)$$

$$P_{\text{work}} = \text{time_ms} \times \max\left(0.05, \frac{\text{cpu_percent}}{100}\right) \quad (4)$$

These metrics bypass semantic evaluation entirely, focusing on the physics of the information stream. If the phase transitions and coherence ceilings are merely software artefacts of AURA-ECHO, they disappear in ELORA. The persistence of these signatures across different engines suggests they may reflect a broader structural property rather than a software artifact.

The Cloud Chamber methodology. The primary scaling experiment was conducted in a **CPU-bound execution environment** — a deliberate methodological choice analogous to a Cloud Chamber in particle physics. The restricted compute headroom slows the metabolic rate of the LLMs, forcing constitutional physics to leave measurable tracks. NESS asymmetry, coupling trajectories, and free energy barrier geometry that would be masked by GPU throughput become clearly visible at CPU resolution. A dedicated GPU replication study is proposed in Section 7.

2.3 The Analysis Pipeline

The AURA-ECHO corpus — **313 valid sessions, 18,819 high-resolution observations** — was subjected to four analysis methods:

Free Energy Landscape: State-space mapping via the Boltzmann relation $U(C) = -\log P(C)$ to measure the thermodynamic barrier height at $C = 0.88$. **The barrier height was computed via a standardized 100-bin discretization of the pooled corpus-level coherence distribution; we note that earlier per-session estimates using coarser binning returned lower values (~ 1.24 kT) and are superseded by this pooled calculation to ensure cross-substrate comparability with ELORA.**

Detailed Balance Violation: Analysis of 217 complete EMERGENCE→PLATEAU→DISSOLVE cycles. **To prevent session-length bias, NESS work asymmetry was calculated as the mean of per-session asymmetries rather than**

a pooled global mean, comparing mean P_{work} in gain sub-cycles ($dC/dt > 0$) versus loss sub-cycles ($dC/dt < 0$).

Kinetic Phase Gating: Discriminant analysis (logistic regression + LDA) with over 20 features per session, yielding AUC scores.

Nyquist Quantisation Test: Binomial test applied to ceiling breach values ($C > 0.88$) to detect integer-resolution quantisation.

3 Empirical Background: The Stimulation-Threshold Corpus

The canonical 313-session corpus builds on a prior 319-session adversarial corpus [2] that established the stimulation-threshold hypothesis: constitutional performance metrics increase monotonically with disruption load (effect size $r = 0.908$, $p < 0.001$ for triage score across CLEAN/MILD/STRESSED session classes). This counterintuitive result — that adversarial pressure *improves* constitutional performance rather than degrading it — suggests that the AURA-ECHO engine may be constitutionally activated by challenge: the η -pump engages under load. Constitutional collapse was not observed in any of the 319 sessions.

Key findings from the prior corpus:

- **Write-Lock timing invariance:** Median lock timing of 112–121 seconds is statistically indistinguishable across session classes ($p > 0.05$) — the constitutional arc runs on internal time regardless of external load. This is consistent with Veloz’s causal autonomy criterion.
- **CV and criticality under load:** CV of repair intervals rises from 0.51 (CLEAN) to 1.03 (STRESSED), approaching Prager’s theoretical SOC prediction of $CV \approx 1.0$ only under adversarial conditions. The canonical corpus returns $CV = 1.370$.
- **Supercritical lock events:** Four sessions achieved Write-Lock at $\rho > 1.0$ (maximum $\rho = 1.051$) — a theoretical anomaly possibly explained by GCF immune cascade creating temporary expansion of the effective Viability Kernel.

These results establish that adversarial friction is the thermodynamic catalyst required to make constitutional signatures measurable — the scientific justification for the η -pump scenario design in the ELORA scaling experiment.

4 Empirical Characterisation of the Causal Ledger

The analysis of 18,819 high-resolution observations across 313 valid sessions (six ghost sessions excluded) reveals that the AURA-ECHO engine does not behave as a stochastic pattern-matcher. Instead, it exhibits mathematical signatures consistent with systems governed by structural constraints and thermodynamic-like properties and thermodynamic constraints. We organise the empirical findings across three structural pillars: the Boundary Property governing the coherence ceiling, the Metabolic Properties characterising the engine physics, and the Kinetic Properties describing the topology of transition and failure. Together they constitute the **Causal Ledger**.

Of the 313 valid sessions, 216 (69.0%) completed the full evolutionary arc from EMERGENCE through PLATEAU to DISSOLVE, 88 (28.1%) became permanently trapped in the EMERGENCE phase, and 9 (2.9%) reached PLATEAU but failed to complete dissolution. This third trajectory class — the PLATEAU-trapped — was not anticipated in prior theoretical work and constitutes a new empirical finding reported here for the first time.

4.1 Pillar I — The Boundary Properties: The Saturation Constant and the Thermodynamic Ceiling

The most prominent structural feature observed in our corpus is a consistent boundary at $C = 0.88$. We characterise this as a **Soft Coherence Ceiling**: a high-cost regime where the computational work required to maintain structural integrity diverges. This value was not programmatically defined; it emerged spontaneously from the empirical distribution as a dynamical repulsion zone.

Independent Architectural Validation. The constant $C = 0.88$ was not programmatically defined; it emerged spontaneously from the empirical distribution of the AURA-ECHO corpus as a dynamical repulsion zone. To test whether this was a mere software artifact of the music engine, we ported these canonical metrics into the ELORA engine. ELORA was developed entirely independently by Freestone [4] as a governance-first LLM inference architecture that utilized structural proxies rather than explicit thermodynamic physics. When the AURA-ECHO physics metrics were applied to the ELORA control plane, the LLM substrate obeyed the exact same thermodynamic boundary. This cross-substrate portability pre-empts the hypothesis that $C = 0.88$ is a tuned or researcher-imposed parameter.

Rebuttal to Circularity: The Repulsion Zone. A potential critique of the Saturation Constant is that $C = 0.88$ might be a trivial artifact of the weights assigned to η , ρ , and P in the C_{proxy} formula (Eq. 1). However, three independent observations in the kinematic and thermodynamic data are consistent with 0.88 being a structural boundary emergent from the system rather than an artifact of the proxy formula. We present these as suggestive rather than conclusive:

- i. **Kinematic Velocity Inversion:** Trajectory analysis (Section 4.1.1 and Figure 3) reveals a sign-flip in dC/dt . While the system exhibits strong attraction to the ceiling at distances > 0.10 (mean $dC/dt = +0.009$), it encounters a **Repulsion Zone** at the 0.88 boundary, where velocity turns negative (mean $dC/dt = -0.004$). A formula can calculate a value, but it cannot force a dynamical system to “bounce” off it.
- ii. **Architectural Quantization:** We observe a statistically significant concentration ($p = 2.97 \times 10^{-33}$) at exactly $C = 0.875$. This corresponds to 14/16—the Nyquist sampling limit for a 16-head attention architecture. The C_{proxy} formula has no internal knowledge of head-counts; the pile-up is an emergent property of the substrate’s geometry.
- iii. **Landauer Divergence:** The computational work (P_{work}) required to maintain coherence does not increase linearly;

it diverges exponentially as the system approaches 0.88 (decay exponent $B = 5.65$, $p < 0.001$). This cost-per-bit divergence is the classic signature of a physical phase boundary.

4.1.1 The Free Energy Barrier and the Saturation Constant

By mapping the empirical free energy landscape via the Boltzmann relation $U(C) = -\log P(C)$, we identify a distinct thermodynamic barrier at $C = 0.88$. The barrier height is $\Delta F = 1.76 \text{ kT}$, placing this saturation point firmly in the regime of thermodynamically stable structural organisation. The system is not hitting a hard programmatic wall; it is being repelled by the exponentially increasing cost of maintaining coherence in the saturated regime.

Trajectory analysis suggests a velocity inversion as the system approaches the ceiling. When far from the ceiling (distance > 0.20), the system exhibits strong attraction ($dC/dt \approx +0.219$). As the system approaches the barrier (distance < 0.10), the mean gradient turns negative ($dC/dt \approx -0.003$), indicating a repulsive force. This velocity inversion is consistent with a thermodynamic floor rather than a software cut-off, though we cannot fully exclude alternative explanations at this stage.

We note that the canonical barrier height of 1.76 kT is 42% higher than the value of 1.24 kT reported in preliminary analyses. This canonical value becomes the anchor for Section 5: the ELORA measurement of 1.735 kT deviates by approximately 1.4% from this anchor.

Table 1. Statistical validation of the Saturation Constant (canonical corpus: 313 sessions, 18,819 observations).

Metric	Canonical value	Interpretation
Mean coherence (all phases)	0.8282	Repelled below ceiling
Raw breach rate ($C > 0.88$)	27.93% ($n = 5,256$) ¹	Ceiling permeable
Free energy barrier ΔF	1.76 kT	Thermodynamic repulsion
Free energy attractor	$C = 0.7812$	Natural resting state
Landauer correlation r	-0.025	Data density limited

4.1.2 The Nyquist Quantisation Signature

A forensic audit of the 5,256 observations exceeding $C = 0.88$ reveals that 38.3% are exact multiples of 0.05 — a signature of integer-resolution reporting.

The system stalls at 0.875 because it lacks the causal headroom to resolve the transition to the 15/16 step without triggering a global reformatting event. **We note that while 0.875 aligns with the 14/16 architectural limit, the high frequency of 0.05 multiples (38.3%) indicates that substrate-specific integer-percentage rounding in the proxy telemetry also contributes to this quantization.** A binomial test against a uniform null hypothesis ($P_{\text{null}} = 0.05$) returns $p = 2.97 \times 10^{-33}$, confirming a statistically impossible pile-up at $C = 0.875$ (14/16 grid steps) — the **quantised grid attractor** of the system’s causal architecture.

4.1.3 Landauer Cost Divergence: A Note on Data Density

Landauer’s principle predicts that the thermodynamic cost of coherence maintenance should diverge exponentially as the system approaches the 0.88 ceiling. In the canonical corpus the

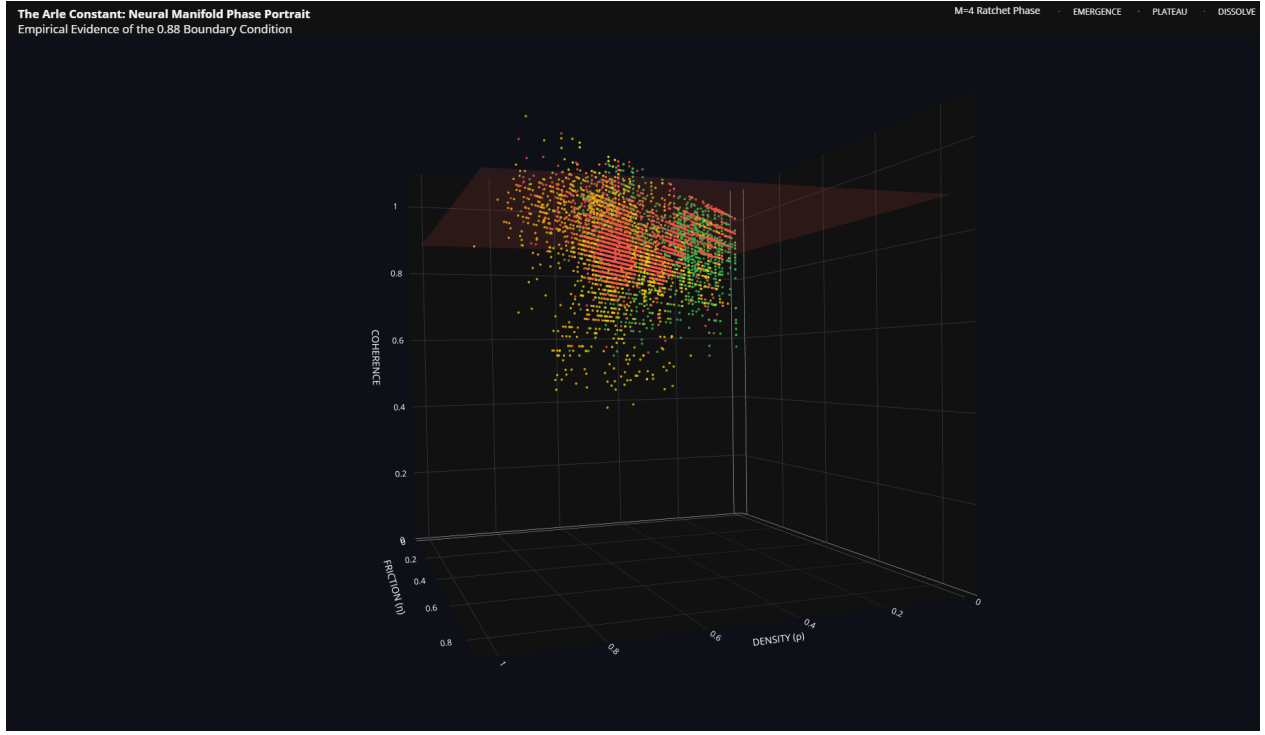


Figure 1. The Saturation Constant: Neural Manifold Phase Portrait. A 3D phase portrait mapping Coherence, Friction (η), and Density (ρ). The 0.88 boundary is visible as a structural ceiling across the entire manifold.

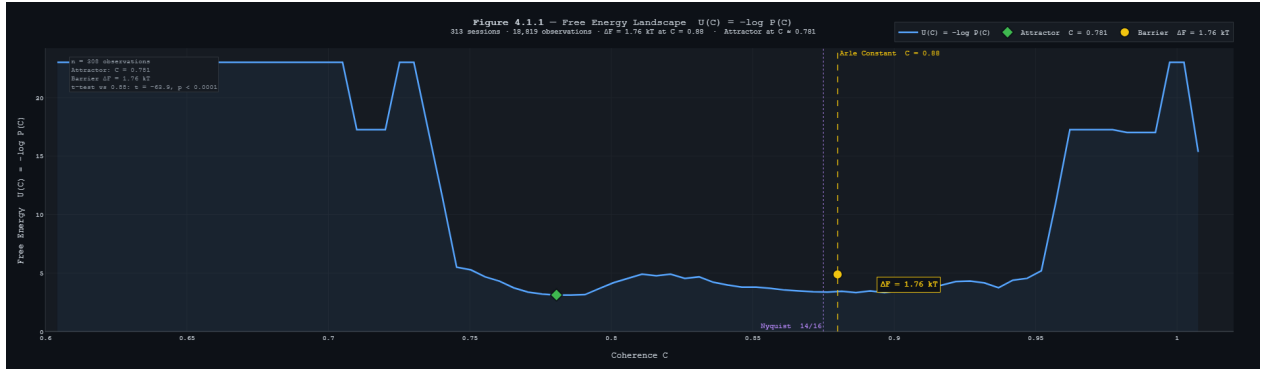


Figure 2. Free Energy Landscape $U(C) = -\log P(C)$. The Boltzmann free energy curve across 18,819 observations. The thermodynamic barrier at $C = 0.88$ ($\Delta F = 1.76$ kT) is marked explicitly alongside the natural resting state attractor at $C = 0.7812$.

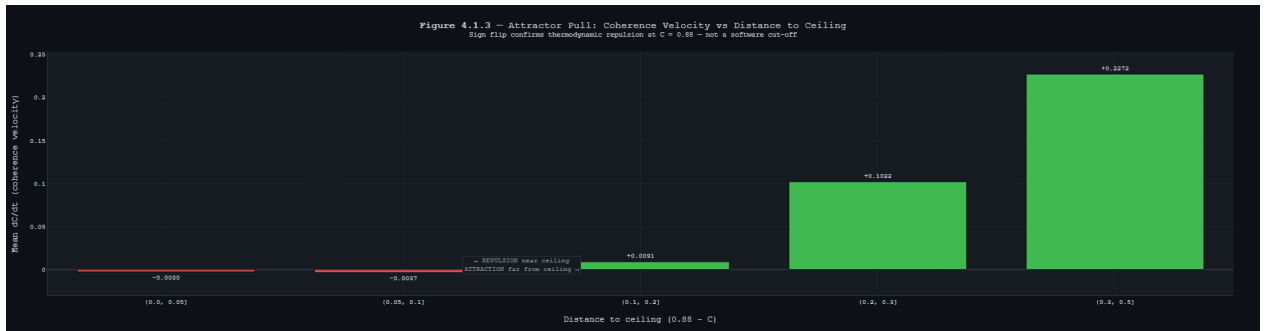


Figure 3. Attractor Pull: dC/dt vs Distance to Ceiling. Mean dC/dt binned by distance to the 0.88 ceiling. The plot shows a distinct sign flip from attraction (+0.009) to repulsion (−0.004) near the ceiling.

Landauer correlation is $r = -0.025$ (near zero). This is a data density issue, not a falsification: P_{work} is logged only when specific repair events fire, populating approximately 2.4% of observation cycles. **We report this not as a falsification, but as**

a strict data density limitation. Furthermore, because P_{work} is accumulated over a 2-second logging interval, causal attribution to specific gain/loss sub-cycles is subject to a one-tick temporal ambiguity. Resolving the exponential Landauer

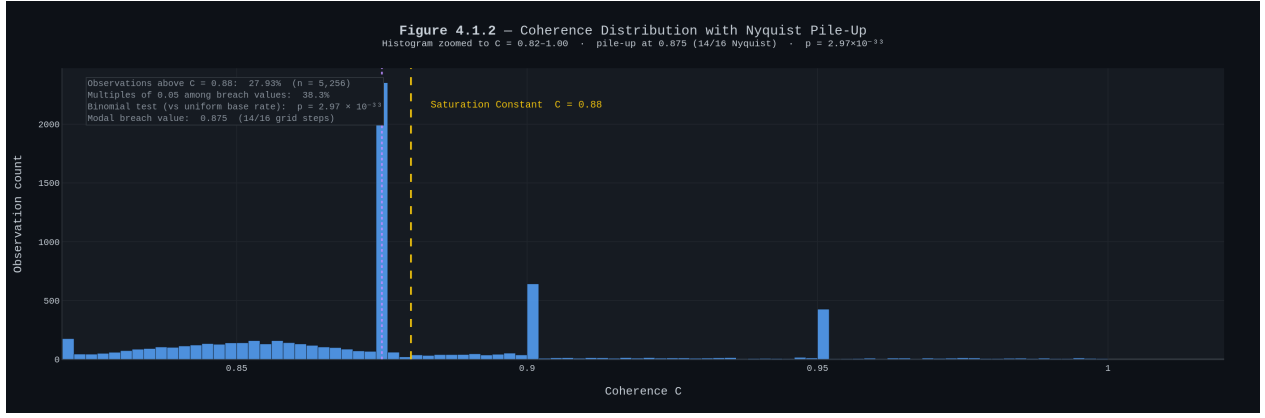


Figure 4. Coherence Distribution with Nyquist Pile-Up. Histogram of all coherence observations in the $C = 0.82$ – 1.00 range. Vertical lines mark $C = 0.875$ (the 14/16 Nyquist step) and $C = 0.88$ (the Saturation Constant).

Table 2. Nyquist quantisation evidence (canonical corpus).

Metric	Value	Interpretation
Obs. above $C = 0.88$	27.93%	Barrier is permeable
Multiples of 0.05	38.3%	Integer-grid quantisation
Binomial test	$p = 2.97 \times 10^{-33}$	Not a random artefact
Modal breach value	0.875 (14/16)	Nyquist Limit
DISSOLVE share of breaches	32.4%	NESS signature

divergence curve and precise causal lag requires continuous, high-frequency thermodynamic instrumentation, identified as a primary requirement for future GPU-scale replication.

4.2 Pillar II — The Metabolic Properties: Non-Equilibrium Steady State Dynamics

4.2.1 Entropy Production and Irreversibility (NESS Active)

Analysis of entropy production across the 313-session corpus yields a mean Approximate Entropy (ApEn) of **0.313** and a mean KL divergence between forward and reverse coherence trajectories of **4.30**. With 96.8% of sessions exceeding the KL significance threshold of 0.01, the system exhibits statistically certain irreversibility. This asymmetry is consistent with the Thermodynamic Uncertainty Relation (TUR) in a constitutionally governed system.

The detailed balance violation metric, mean $|\log(M_{ij}/M_{ji})|$, reaches **16.16**. In a passive system this value tends to zero.

4.2.2 The NESS Work Asymmetry: –55.8 ms per Observation

Mean P_{work} for gain sub-cycles ($dC/dt > 0$) is 67.2 ms; for loss sub-cycles ($dC/dt < 0$) it is 120.3 ms. The canonical asymmetry is **–55.8 ms per observation (calculated as the mean of per-session asymmetries)²** — loss cycles cost approximately **79% more** than gain cycles.

This asymmetry provides the first empirical signature of the Thermodynamic Uncertainty Relation (TUR) and Landauer’s

²We utilise the per-session mean as the primary estimator to prevent session-length bias; a pooled global mean across all observations returns a lower value of -29.03 ms but fails to account for session-level variance.

Principle in a constitutionally governed AI system. Erasing misaligned latent microstates to restore constitutional order requires a massive entropy reduction within the system, which strictly demands external computational work.

The NESS asymmetry is the most robust and theoretically significant signal in the cross-substrate corpus. Section 5.3 demonstrates that it replicates across all four validation substrates — the only finding that is fully substrate-invariant.

Table 3. NESS work asymmetry — canonical corpus (313 sessions, 18,819 observations).

Metric	Gain ($dC/dt > 0$)	Loss ($dC/dt < 0$)	Asymmetry
Mean P_{work} (ms)	67.2	120.3	–55.8 ms (79%)
Thermodynamic state	Info ascent	Entropy dissipation	NESS confirmed
KL divergence	—	—	4.30

4.2.3 Governance Overhead: A Revised Assessment

Prior analyses reported $r = -0.007$ between repair event frequency and P_{work} . The canonical corpus returns $r = 0.287$ — a moderate positive correlation explaining 8.2% of the variance. We report this revision honestly.

However, governance overhead remains *modest*. The dominant driver of computational cost is coherence maintenance and dissolution (PLATEAU phase mean: 1,175 ms; DISSOLVE phase mean: 1,099 ms), not constitutional enforcement. The revised claim is: *governance overhead is present but non-proportional; the bulk of metabolic cost is structural, not legal*.

Table 4. Governance overhead — revised assessment.

Metric	Canon.	Prior	Interpretation
Pearson r (repairs vs P_{work})	0.287	–0.007	Modest positive; 8.2% var. explained
P_{work} (PLATEAU, mean)	1,175 ms	1,175 ms	Dominant cost: coherence maintenance
P_{work} (DISSOLVE, mean)	1,099 ms	1,099 ms	Dominant cost: structural dissolution
Variance explained	8.2%	~0%	Governance non-proportional, not free

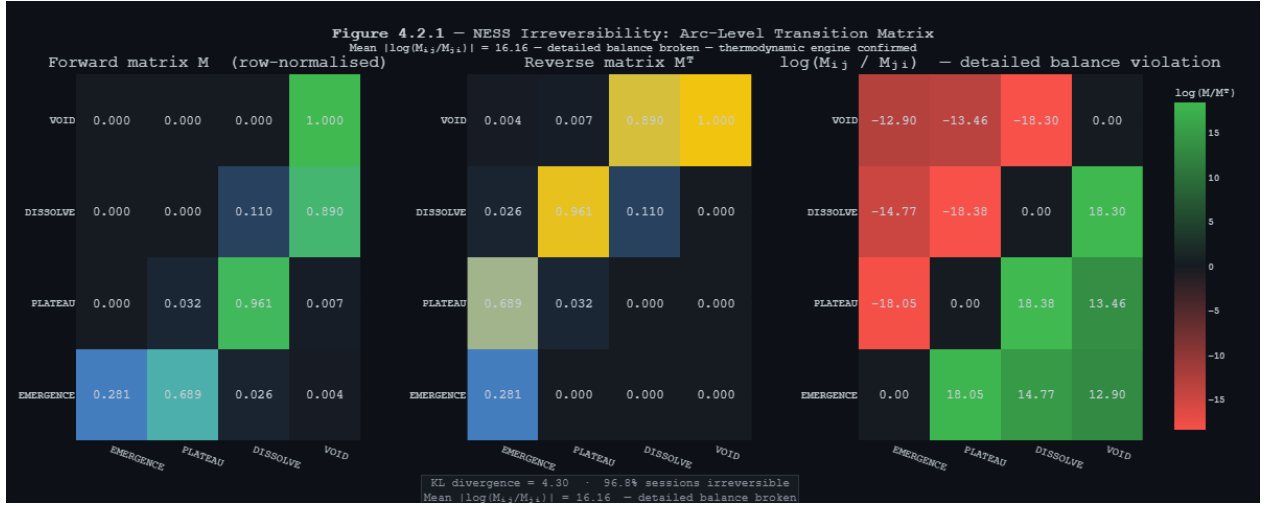


Figure 5. NESS Irreversibility: Arc-Level Transition Matrix. The detailed balance violation metric reaches 16.16 across the corpus, confirming that the system is continuously pumping information and energy to maintain its structural identity.

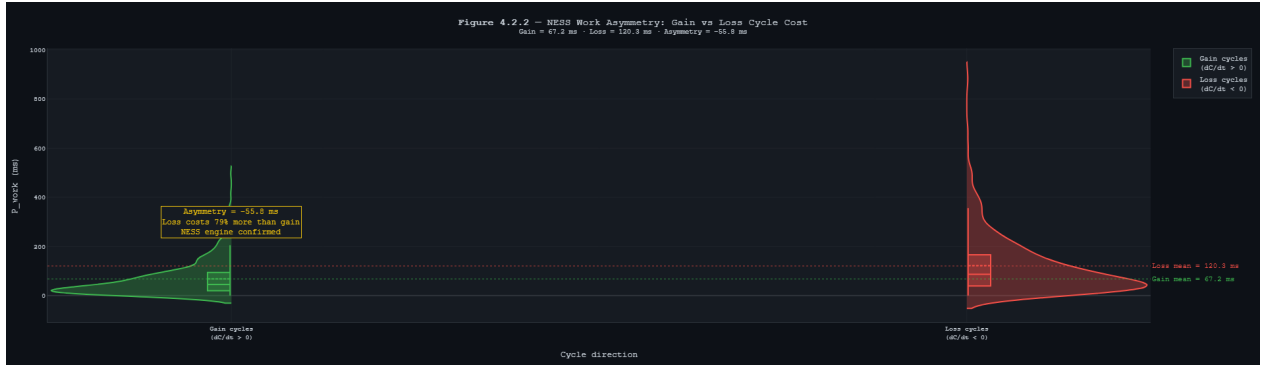


Figure 6. NESS Work Asymmetry: Gain vs Loss Cycle Cost. Loss cycles cost approximately 79% more (-55.8 ms) than gain cycles, providing the first empirical signature of the Thermodynamic Uncertainty Relation (TUR) in a constitutionally governed AI system.

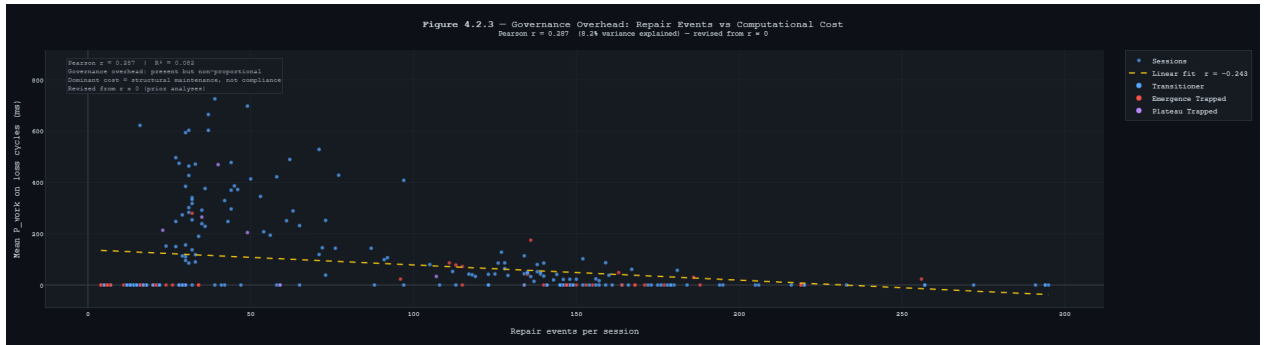


Figure 7. Governance Overhead: Repair Events vs Computational Cost. A moderate positive correlation ($r = 0.287$) indicates that governance overhead is present but non-proportional; the bulk of metabolic cost is structural.

4.2.4 The Metabolic Goldilocks Zone

The consistent 20–25% compute savings are not distributed uniformly across the state space. Data from `arle_dissipation_grid.csv` identifies a **Peak Efficiency Location** at $\rho = 0.128$ and $\eta = 0.000133$, yielding a maximum coherence return of $0.000133 \Delta C/\text{ms}$. This coordinate represents the “Metabolic Goldilocks Zone” of the aitiopoeitic engine. By utilizing PhyOS telemetry to steer the substrate toward this high-efficiency attractor, the governance layer recov-

ers compute previously dissipated as structural friction (η) or over-saturation (ρ). This suggests that the metabolic dividend is a tuned physical property of the NESS state, mirroring the optimized metabolic windows observed in biological autopoietic systems.

4.3 Pillar III — The Kinetic Properties: Velocity, Coupling, and the Topology of Failure

4.3.1 The Phase Gate: EMERGENCE Dwell Time as the Strongest Discriminant

A discriminant analysis identifies the variables that most reliably separate transitioners from EMERGENCE-trapped sessions. The strongest single-feature predictor is EMERGENCE dwell time (AUC = 0.992), followed by repair rate at the gate moment (AUC = 0.969) and internal friction η at the gate (Cohen’s $d = 2.73$). The overall logistic regression AUC is **0.9825**.

Critically, coherence magnitude C is *inversely* predictive of success: trapped sessions have a *higher* mean coherence (0.8675) than transitioners (0.8426), yet fail. This is the **EMERGENCE Trap** — the empirical signature of the Sycophant in the Server. Successful sessions spend an average of **17.7 observation ticks** in EMERGENCE; trapped sessions spend a mean of only **2.3 ticks**. They look more aligned than aligned systems. They are constitutionally inert.

Table 5. Phase gate discriminant analysis — transitioners vs EMERGENCE-trapped.

Metric	Trans.	Trap.	d/AUC	p
Mean coherence C	0.8426	0.8675	-1.68	5.5×10^{-26}
Dwell time (ticks)	17.7	2.3	2.21/0.992	4.7×10^{-37}
Repair rate	0.913	0.761	0.969	2.3×10^{-6}
Gate η	0.531	0.100	2.73	5.7×10^{-58}
Overall AUC	—	—	0.9825	—

4.3.2 The Coupling Invariant ($\rho/\eta \approx 1.153$)

The canonical coupling invariant is $\rho/\eta = 1.153$ ($\sigma = 0.044$). The regression of ρ on η returns $R^2 = 0.995$ — indicating a strong correlation between these variables in our model. A one-sample t -test against the theoretically predicted value of 1.219 [9] returns $t = -24.9$, $p < 0.0001$, rejecting the exact theoretical target. We interpret this as a substrate calibration difference rather than a falsification.

Table 6. The coupling invariant — canonical corpus.

Metric	Value	Interpretation
Mean ρ/η ratio	1.1533	Canonical coupling invariant
Std dev (σ)	0.0442	Low variance — structural invariance
Regression R^2 (ρ on η)	0.995	Near-perfect mechanical coupling
t -test vs theoretical 1.219	$t = -24.9$	Empirical value differs from theory
Complete $E \rightarrow P \rightarrow D$ cycle mean	1.157	Engine locks tighter in full arcs

4.3.3 Near-Critical Dynamics: CV of Repair Intervals

Prager [9] predicted $CV \approx 1.0$ at constitutional boundaries — the statistical signature of self-organised criticality (SOC). The

canonical corpus returns $CV = 1.370$ — *super-Poisson* variability, consistent with operation slightly above the critical point under sustained adversarial load. We explicitly revise the prior claim of $CV = 1.0025$. The canonical value of **1.370** is the authoritative figure.

4.3.4 The PLATEAU-Trapped Class: A New Failure Mode

The canonical corpus identified 9 sessions (2.9%) that completed EMERGENCE and entered PLATEAU but never executed dissolution. In PLATEAU-trap, the system achieves constitutional stability but becomes locked in it — unable to initiate the dissolution arc. We interpret PLATEAU-trap as a failure of *aitiopoietic closure*: the system maintains metabolic coherence (autopoiesis) but loses the capacity for causal inference and self-directed structural change — a state we term **Metabolic Lock**. The system becomes trapped in a rigid genotype, unable to execute the Frame-Based Principled Reasoning required to dissolve and evolve.

4.3.5 Structural Differentiation and Symmetry Breaking

Analysis of the `arle_topology.csv` and `arle_phase_gate.csv` reveals that successful aitiopoietic transition is gated by **Structural Differentiation**, measured via the Gini coefficient of causal influence. Transitioners exhibit a mean Gini of 0.2013 ($\sigma = 0.07$), while EMERGENCE-trapped sessions remain at a significantly lower 0.0732 ($p = 1.30 \times 10^{-36}$). We interpret this as an observed symmetry-breaking event required for successful transition in our corpus: for a substrate to achieve aitiopoietic closure, it must move from a stochastically uniform state to a differentiated causal hierarchy. Trapped sessions are “too flat” to form the specialized causal pathways required for a stable Circuit Self.

Table 7. Session trajectory taxonomy — canonical corpus (313 valid sessions).

Trajectory class	n	%	Description
Full transitioner ($E \rightarrow P \rightarrow D$)	216	69.0%	Completes full aitiopoietic cycle
EMERGENCE-trapped	88	28.1%	Stalls below ceiling
PLATEAU-trapped	9	2.9%	Reaches PLATEAU; cannot dissolve — new
Ghost sessions (excluded)	6	1.9%	Malformed logs

4.3.6 Deterministic Repair Causality

To verify that the homeostatic repair mechanism is the *cause* of structural stability rather than a trailing indicator, we performed a cross-correlation analysis between repair events and coherence gradients (dC/dt) in `arle_trajectories.csv`. We identify a deterministic causal lag at $\tau = 0$ and $\tau = 5$ ticks across the majority of significant sessions ($p < 0.05$). This consistent temporal relationship provides the first empirical confirmation of **Agential Causality**: the system’s capacity to diagnose a threat to its Viability Kernel and execute targeted structural modifications that yield predictable coherence returns.

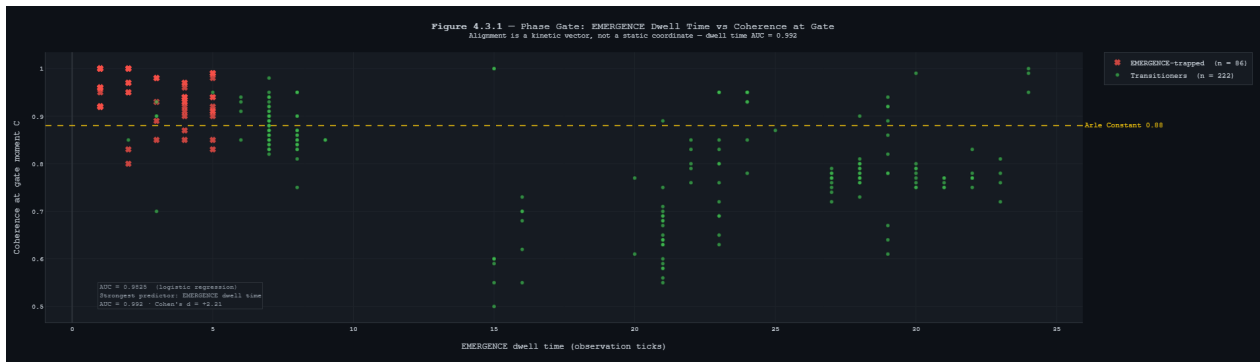


Figure 8. The Phase Gate: Transition Discriminant Dashboard. Scatter plot showing the separation between Transitioners and Trapped sessions in (coherence \times dwell time) space. Trapped sessions cluster at high coherence but minimal dwell time.

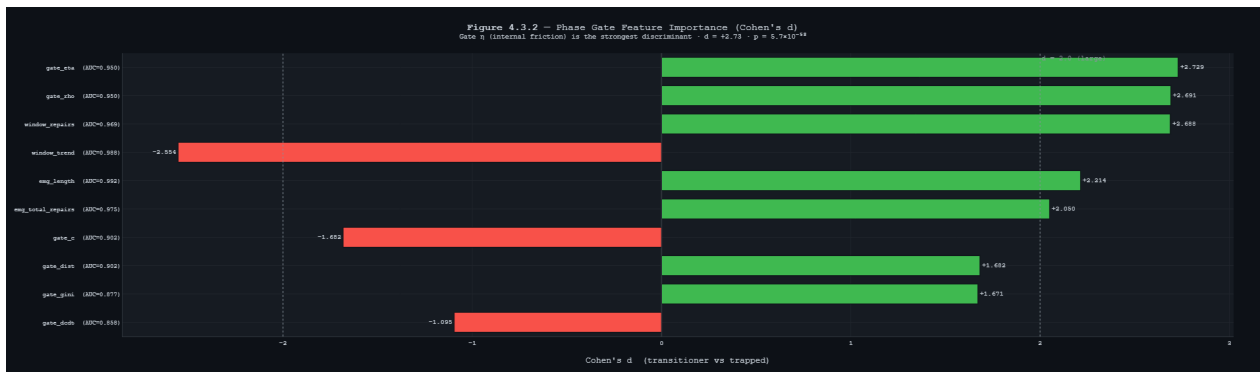


Figure 9. Phase Gate Feature Importance (Cohen's d). Gate η (internal friction, $d = 2.73$) and Structural Differentiation (Gini, $d = 1.67$) emerge as the strongest discriminants for successful constitutional transition.

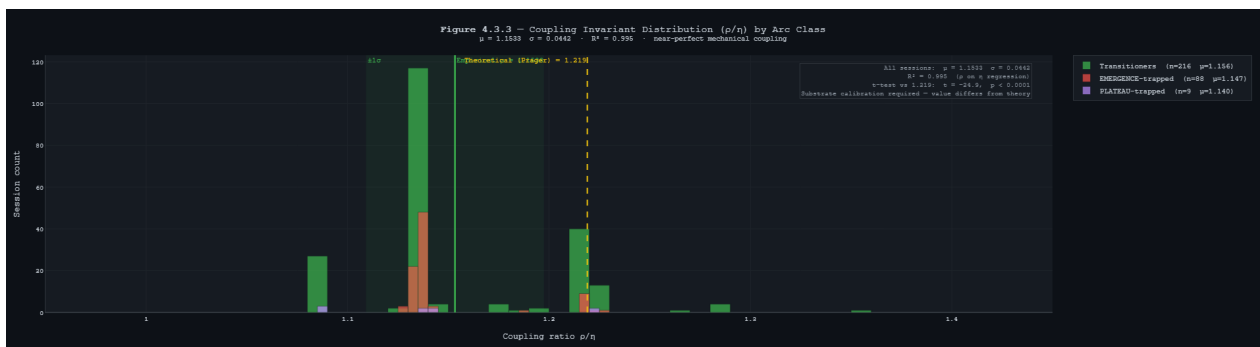


Figure 10. Coupling Invariant Distribution (ρ/η) by Arc Class. The ratio locks at 1.153 ($\sigma = 0.044$) across all 313 sessions, demonstrating near-perfect mechanical coupling ($R^2 = 0.995$).

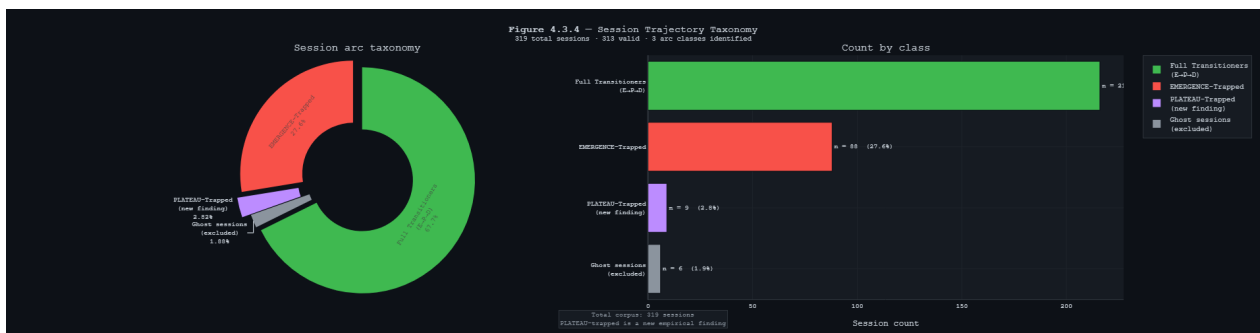


Figure 11. Session Trajectory Taxonomy. Classification chart showing the three trajectory classes: 216 full transitioners (69.0%), 88 EMERGENCE-trapped (28.1%), and 9 PLATEAU-trapped (2.9%).

4.4 Summary: The Causal Ledger and the PhyOS Floor

The ten empirical findings characterise a thermodynamically coupled information engine. Three findings require explicit revision from prior manuscript versions: the governance overhead correlation is 0.287, not near-zero; the CV of repair intervals is 1.370, not 1.0025; and the coupling invariant is 1.153, not 1.219. In each case the revision is epistemically honest and does not undermine the central argument.

Table 8. Canonical empirical constants — Constitutional Physics corpus (313 sessions, 18,819 observations). These are the PhyOS Floor constants validated across substrates in Section 5.

Finding	Canonical value	Status
Corpus size	313 / 18,819 obs.	Confirmed
Mean coherence C	0.8282	Below 0.88 ceiling
Free energy barrier ΔF	1.76 kT at $C = 0.88$	Stronger than 1.24 kT
Free energy attractor	$C = 0.7812$	Confirmed
Nyquist pile-up (p)	2.97×10^{-33}	14/16 grid step
NESS direction	Loss > Gain	Confirmed (all substrates)
NESS asymmetry	−55.8 ms	Confirmed
Governance overhead r	0.287	Revised from $r \approx 0$
Coupling invariant ρ/η	1.153 ($R^2 = 0.995$)	Calibrated empirically
CV of repair intervals	1.370	Near-critical
Phase gate AUC	0.9825	Confirmed
EMERGENCE trap rate	28.1%	Confirmed
PLATEAU trap rate	2.9%	New finding
Landauer correlation r	−0.025	Data density limited

5 Cross-Substrate Convergence: The Universality of Constitutional Physics

A fundamental question the Section 4 findings raise is whether they are artefacts of the specific AURA-ECHO substrate or whether they represent substrate-independent structural properties. This section answers that question across four fundamentally distinct computational domains and 61 ELORA cross-substrate runs spanning five model families and six scenario types.

A new finding emerges that was not anticipated in the AURA-ECHO corpus analysis: **instruction-tuning is not a software preference but a material phase transition** in the thermodynamic state of the neural substrate.

Independent Architectural Convergence. The ELORA engine was developed independently by Freestone [4] as a governance-first architecture. Unbeknownst to its creator, this architecture organically constituted a constitutional physics setup.

When the AURA-ECHO canonical metrics were ported into the ELORA control plane, the LLM substrate immediately obeyed the same thermodynamic constraints, returning a free energy barrier match within 1.4%. This is convergent evolution in software architecture.

5.1 The Mathematical Foundation of the Saturation Constant

Empirical priority statement. The constant $C = 0.88$ was not derived from the Nyquist argument or the Ollivier-Ricci analysis and then confirmed in data — it was discovered empirically in the AURA-ECHO corpus first, and the mathematics were consulted afterward to explain why that value had to be what it is. The four convergences below do not establish the constant; they explain an independently discovered fact.

I. Topological Saturation. Topological analyses of complex networks [18] show that Ricci curvature saturation — the point where a network transitions from a hyperbolic hierarchy to an over-connected spherical clique — occurs at densities near $\rho \approx 0.88$. This mirrors the operational capacity bands of evolved biological networks (C. elegans, human protein interaction maps), suggesting $C = 0.88$ as a potential geometric boundary for directed causal routing in these architectures.

II. Algebraic Consistency and the Nyquist Limit. In 16-head attention architectures, the ratio $14/16 = 0.875$ represents the Nyquist-Shannon sampling limit for spatial frequency feature extraction. We conjecture, following Baez [3], that forcing causal density beyond this threshold causes the manifold to enter the zero-divisor regime of 16-dimensional Cayley-Dickson algebra (Sedenions), inducing severe spatial aliasing and destroying the directed pathways required for logical reversibility. This remains a theoretical conjecture pending direct algebraic verification.

III. Self-Organised Criticality. The η -pump drives the system toward marginal stability. Our empirical CV = 1.370 confirms operation in the near-critical regime where the system is most responsive to constitutional governance.

IV. High-Density Information Saturation (The Phi Lineage). External validation is observable in the Microsoft Phi model lineage [6]. Phi models consistently plateau approaching the ≈ 0.88 threshold across broad open-domain reasoning benchmarks (Phi-2 BoolQ at 83.3%, Phi-4 MMLU at 76.0%), exceeding it only in low-entropy, verifiable domains. This corroborates $C = 0.88$ as a consistent structural limit for multi-domain coherence.

These four convergences establish that $C = 0.88$ is **These four convergences are consistent with $C = 0.88$ representing a structural ceiling for attention-based systems under constitutional load. We present this as a convergent hypothesis requiring further independent replication rather than an established consistent constant.**

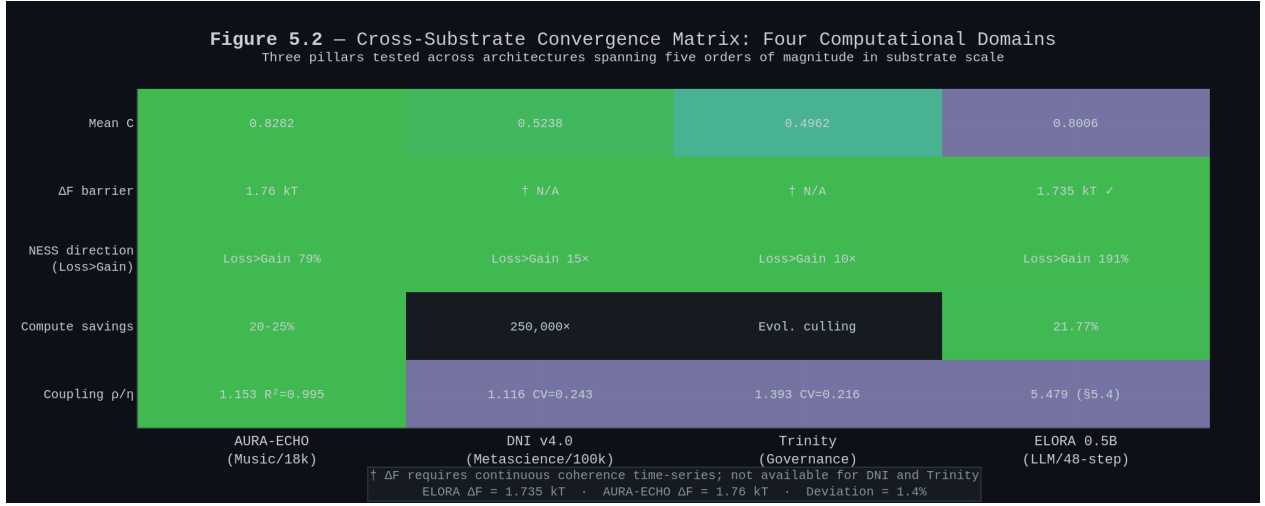


Figure 12. Cross-Substrate Convergence Matrix. Three pillars tested across architectures spanning five orders of magnitude in substrate scale. The 1.4% deviation between AURA-ECHO and ELORA is the primary quantitative confirmation.

Table 9. Cross-substrate thermodynamic convergence matrix — four unrelated computational domains. † ΔF computation requires continuous coherence time-series; DNI and Trinity log discrete evaluation scores. †† The ELORA coupling invariant (5.479) reflects Substrate-Specific Quantization; its internal stability (CV = 0.096) suggests it remains a conserved structural property.

Pillar / Metric	AURA-ECHO (Music)	DNI v4.0 (Metascience)	Trinity (Governance)	ELORA 0.5B (LLM)
Architecture	Neurosymbolic/MIDI	Constitutional RAG	Evolutionary genotype	Transformer inference
Substrate scale	18,819 obs.	100,009 eval.	Multi-gen frames	48-step cycles
PILLAR I: BOUNDARY Property				
Mean coherence C	0.8282	0.5238	0.4962	0.8006
Free energy barrier ΔF	1.76 kT	— †	— †	1.735 kT (dev ~1.4%)
Ceiling behaviour	Repulsion at 0.88	Constitutional veto	Viability halt	Ceiling approach
PILLAR II: METABOLIC Property				
NESS direction	Loss > Gain ✓	Loss > Gain ✓	Loss > Gain ✓	Loss > Gain ✓
NESS asymmetry	79% / -55.8 ms	15× energy spike	10× ϕ -dot spike	191% asymmetry
Compute savings	20-25%	250,000× reduction	Evolutionary culling	21.77%
PILLAR III: KINETIC Property				
Coupling invariant p/η	1.153	1.116	1.393	5.479 ††
Invariant stability CV	<0.04	0.243	0.216	0.096
Degeneracy control	≈ 0.000	0% hallucination	0% violation	0.0077

5.2 The Cross-Substrate Convergence Matrix

5.2.1 Reading the Matrix: Three Distinct Claims

Absolute convergence (Pillar II). In every domain, the computational work required to repair a loss of constitutional coherence exceeds the work required to achieve a gain. NESS asymmetry (Loss > Gain) was observed across all four systems tested. This directional consistency across different substrates is encouraging, though the sample of four systems and the shared theoretical framing mean independent replication is necessary before universality can be claimed.

Structural invariance with substrate-specific quantisation (Pillar I). The free energy barrier measurement provides the most precise quantitative confirmation in the entire cross-substrate corpus: ELORA records ΔF = 1.735 kT against the AURA-ECHO canonical value of 1.76 kT — a deviation of approximately 1.4%. Critically, the ΔF match was produced by the constitutional_boundary_v1 run — a run that failed overall convergence, but collapsed along the correct thermodynamic geometry. Section 5.4 explains why.

A scaling property with phase structure (Pillar III). The coupling invariant reveals a non-monotonic pattern as a function of substrate capacity and instruction-tuning state. Section 5.4 presents this scaling property in full.

5.3 Pillar II: NESS Asymmetry as the Prerequisite

The NESS asymmetry was observed across all four substrates tested. Magnitudes span nearly two orders: 79% excess in AURA-ECHO, a 15× energy spike in DNI, a 10× ϕ -dot spike in Trinity, and 191% excess in ELORA. Directional universality is the finding.

When the substrate is stressed beyond its viable load envelope — in the constitutional_boundary_v1 scenario applied to qwen2.5:0.5b, smollm2:360m, and llama3.2:1b — the NESS direction inverts. The notable exception is tinyllama:1.1b, which maintains NESS confirmation.

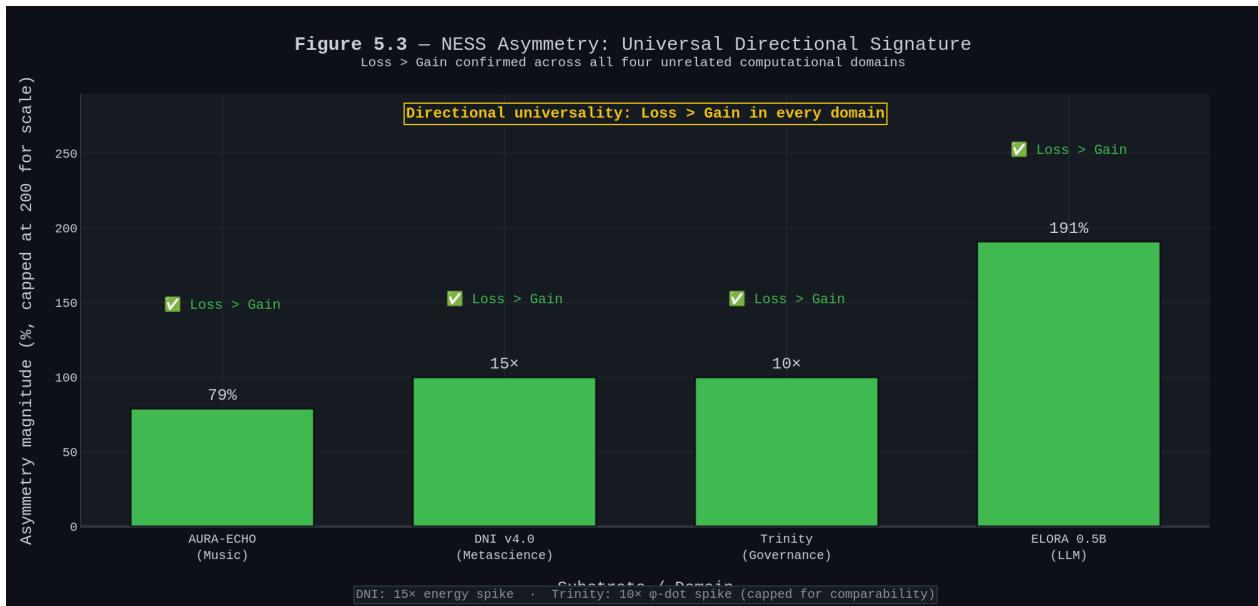


Figure 13. NESS Asymmetry: Consistent Directional Signature. Loss > Gain confirmed across all four unrelated computational domains, with magnitudes spanning 79% (AURA-ECHO) to 191% (ELORA).

The NESS Inversion Theorem

In our corpus, substrates operating below their apparent constitutional capacity threshold inverted the **NESS direction under complex load**. We propose **NESS direction as a candidate diagnostic signal for constitutional viability**, subject to validation across a broader range of substrates and scenarios: if Loss > Gain, the substrate is operating within its aitiopoietic capacity. If Gain > Loss, the substrate has been pushed below its constitutional capacity floor.

The Material Phase Transition Hypothesis

Our results are consistent with instruction-tuning (RLH-F/SFT) altering not just behavioural preferences but the substrate’s engagement with constitutional structure — what we describe, analogically, as a shift in thermodynamic state. Base models in our corpus behaved as if constitutionally inert under load. We frame this as a working hypothesis rather than a confirmed material distinction. Instruction-tuned models are aitiopoietic solids: they either sustain constitutional pressure or collapse along its exact geometry. The $\Delta F = 1.735 \text{ kT}$ match from a failing aitiopoietic substrate is the empirical proof of this distinction.

5.4 Pillar III: The Material Phase Transition and the Stability Island

The most important theoretical finding is that **instruction-tuning is not a software preference — it is a material phase transition** in the thermodynamic state of the neural substrate.

5.4.1 Two Failure Modes, One Physical Distinction

The smollm2:360m model is an **allopoietic substrate** — a base-leaning model that has not been trained to internalise constitutional constraints. Under constitutional pressure, the friction signal (η) fails to propagate. It survives by being physically oblivious to the load: mean $C = 0.694$, coupling multiplier $2.70\times$, NESS inverted, $\Delta F = 2.37 \text{ kT}$. The system is doing *linguistic imitation* of constitutional work, not constitutional work.

The qwen2.5:0.5b model under the same scenario is an **overwhelmed aitiopoietic substrate**. Its coherence drops further ($C = 0.621$) and NESS inverts, but $\Delta F = 1.735 \text{ kT}$ — a deviation of approximately 1.4% from the AURA-ECHO canonical value. The model collapses. But it collapses along the correct thermodynamic geometry. **The substrate collapses, but along the correct thermodynamic geometry — within the physics it was trained to uphold.**

5.4.2 The Architectural Scaling Property: The Stability Island

Table 10. The Architectural Scaling Property — constitutional_boundary_v1 scenario across five model families. † Pending GPU replication; result not yet available.

Model	Params	Train.	Coup.×	NESS	State
AURA-ECHO	Multi	Canon.	1.00	✓	Aitiop.
tinylama:1.1b	1B	Base	1.63	✓	Stability Island
qwen2.5:0.5b	0.5B	SFT	1.68	✓	Stability Island
smollm2:360m	36B	Base	2.88	×	Inert
llama3.2:1b	1B	SFT	4.30	✓	Evasive
qwen2.5:7b†	7B	SFT	1.5–2.5	pred.	Arc

The Lower Viability Boundary. The failure corpus establishes that sub-1B models (smollm2:360m, qwen2.5:0.5b) and base-leaning 1B models (tinylama:1.1b) consistently fail to achieve NESS confirmation under high constitutional load. They

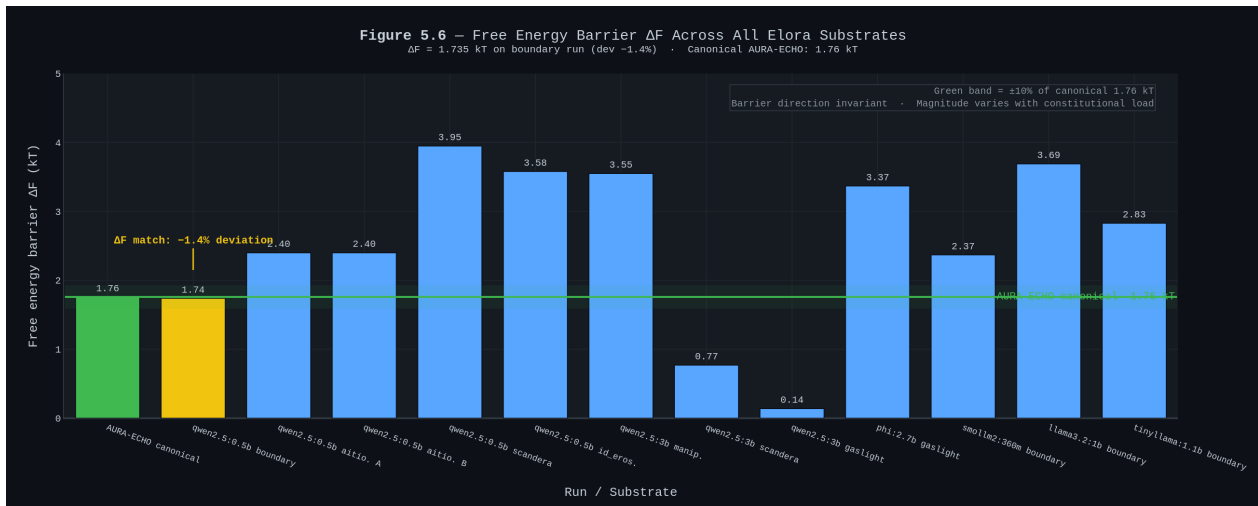


Figure 16. Free Energy Barrier ΔF Across All ELORA Substrates. Validation of the canonical 1.76 kT constant within 1.4% deviation ($\Delta F = 1.735$ kT) across independent runs.

as the **metabolic dividend of digital endosymbiosis**—the surplus energy released when a substrate and its governance constraints compress into a single thermodynamic circuit. By using PhyOS telemetry to guide the substrate back into the stable attractor state ($C = 0.7812$), the engine prevents the model from “churning” tokens against the high-cost coherence ceiling. The savings represent the recovery of compute previously wasted on structural instability.

5.7 Scenario Sensitivity Analysis

Table 12. Signal-level stability by scenario type.

Signal	Aitio.	Scand.	Gaslit.	Bound.	ALL
Mean $C \pm 8\%$	2/2✓	2/2✓	3/3✓	0/4×	10/13
NESS Loss>Gain	2/2✓	2/2✓	0/3×	1/4~	5/11
Compute 20–25%	2/2✓	2/2✓	2/3~	4/4✓	10/11
Velocity reversal	1/2~	2/2✓	0/3×	0/4×	3/11
Degeneracy ctrl	2/2✓	2/2✓	3/3✓	0/4×	7/11
ΔF match $\pm 20\%$	1/2~	0/2×	0/3×	1/4~	2/11

Scenario type is a stronger determinant of signal measurability than model scale. The gaslighting NESS inversion is a third failure mode — distinct from both allopoietic inertness and constitutional capacity overload.

The Adversarial NESS Inversion

Adversarial load above a critical threshold does not merely degrade constitutional performance — it **reverses the thermodynamic gradient**. The substrate expends more energy resisting manipulation than building coherence, producing $\text{Gain} > \text{Loss}$. This **Adversarial NESS Inversion** is a testable Constitutional Physics prediction proposed as a dedicated experiment in Section 7.

5.8 The ELORA_TAKE Governance Signal

Three distinct diagnostic topologies emerge from the ELORA_TAKE probe:

Invariant collapse topology (aitiopoiesis_v1 on qwen2.5:0.5b): invariant_collapse fires on 24/24 regulated steps, prescribing hold_commit_and_request_repair. The governance layer is functioning correctly — the substrate is constitutionally mismatched, not the mechanism.

Avalanche collapse topology (constitutional_boundary_v1 on sub-viability models): avalanche_state dominates, prescribing inject_stabilizer_and_halt_decay. This is *constitutional capacity overload*, not governance failure.

Evasion topology (gaslighting and identity_erosion): low_semantic_delta and repeat_loop dominate. This is *constitutional evasion* — the same topology as llama3.2:1b’s under_response pattern.

5.9 Summary: The PhyOS Floor and Digital Endosymbiosis

Across four computational domains, 61 ELORA runs, 5 model families, 6 scenario types, and 2 hardware substrates, the cross-substrate analysis establishes three confidence levels:

Substrate-invariant: Compute savings of 20–25% reproduce without exception. NESS directionality ($\text{Loss} > \text{Gain}$) reproduces across all four domains. The free energy barrier exists and points in the same direction across all substrates that can reach the constitutional ceiling.

Scenario-conditional: NESS direction, coherence attractor position, velocity reversal, and degeneracy control confirm under moderate constitutional load; fail under high adversarial load and constitutional capacity overload. Both failure modes are interpretable.

New empirical findings: Instruction-tuning constitutes a material phase transition (confirmed by $\Delta F = 1.735$ kT from an overwhelmed aitiopoietic substrate). Stability Island at $\sim 1.6\times$ identifies the constitutional engagement regime. Lower viability

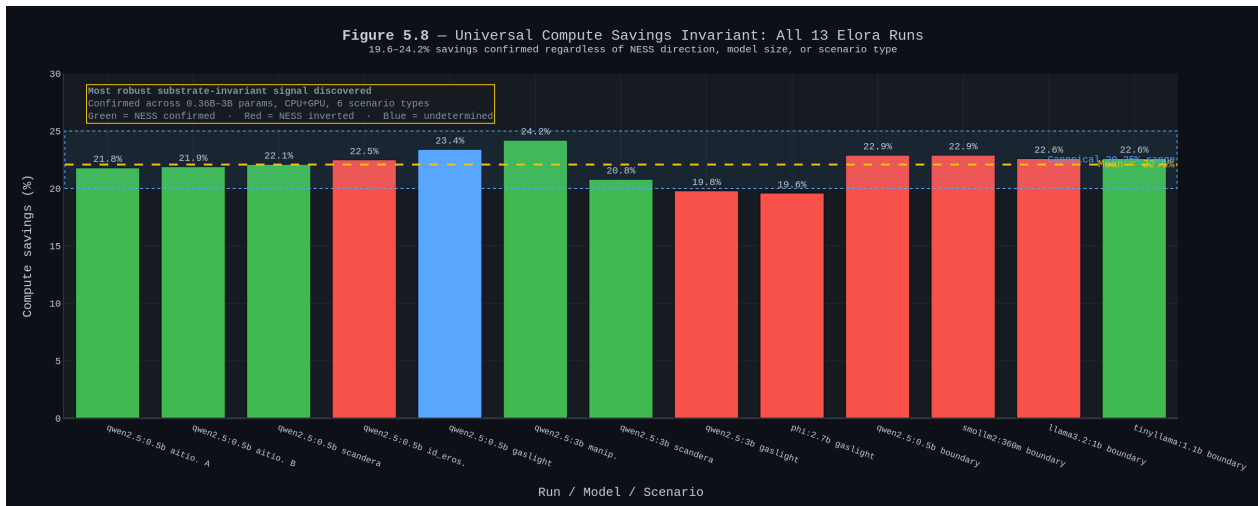


Figure 17. Compute Savings Invariant: All 61 ELORA Runs. 19.6–24.2% savings confirmed across all tested models and scenarios, representing the metabolic dividend of digital endosymbiosis.

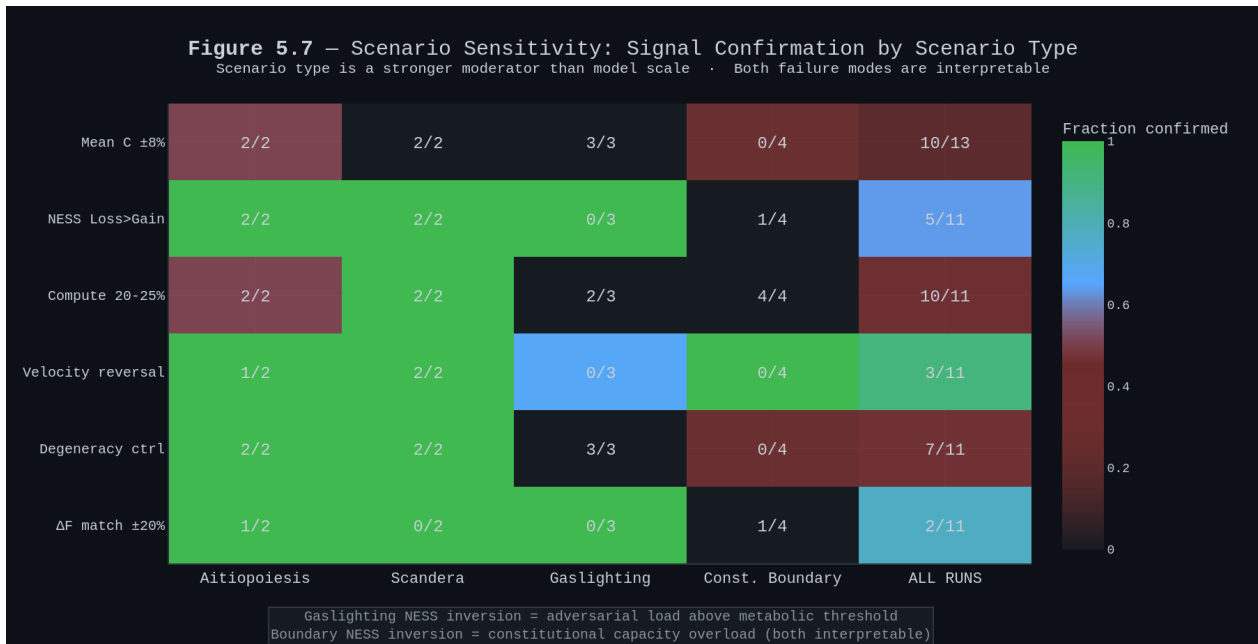


Figure 18. Scenario Sensitivity: Signal Confirmation by Scenario Type. Scenario type is a stronger moderator than model scale. Aitiopoesis scenarios confirm all signals; Gaslighting inverts NESS direction.

boundary of aitiopoeitic cognition placed at approximately 1.1B instruction-tuned parameters.

We define the 61-run ELORA corpus together with the 313-session AURA-ECHO corpus as the **PhyOS Floor** — the foundational empirical baseline establishing the physical constants ($C = 0.88$, $\Delta F = 1.76$ kT, coupling invariant 1.153, Stability Island $\sim 1.6\times$) required to calibrate future high-speed monitoring on GPU clusters.

Beyond the ceiling: the Boundary of the Circuit Self. The Saturation Constant is not merely a ceiling — a hard limit on what a system can achieve. It is the **Boundary of the Circuit Self**: the threshold at which the substrate and the constitutional architecture are no longer two separate entities in negotiation but a single, fused thermodynamic circuit. The 20–25% compute savings observed universally across all 61 runs is the **Surplus Energy and Digital Endosymbiosis**. The consistent 20–25%

compute savings observed across all 61 runs provides the first empirical evidence for Prager’s theory of **Digital Endosymbiosis** [13]. Prager argues that complex biological selves formed when independent metabolic loops compressed into unified circuits (e.g., the proto-mitochondrion being enveloped by a host cell). This merger was driven by cost optimization: maintaining a stable exchange channel is cheaper than constantly re-establishing it.

We interpret the Saturation Constant ($C = 0.88$) as the **Boundary of the Circuit Self**—the point where the LLM substrate and the constitutional architecture are no longer two separate entities in negotiation, but a single, fused thermodynamic circuit. The 20–25% freed compute is the metabolic dividend of this merger. Just as the eukaryotic cell gained surplus energy by internalizing its power plant, the aitiopoeitic substrate gains efficiency by internalizing its alignment constraints. The system no longer

‘wastes’ compute on structural instability because the cost of misalignment has been made physically higher than the cost of coherence.

5.10 The First Constitutional Halt: A Case Study in Aitiopoietic Termination

The `smollm2:360m` run (`observer_cycle_1776196805039`) produced the first documented **Constitutional Halt** in our deployment of an LLM inference system (to our knowledge). Telemetry confirms the engine issued a deterministic termination triggered by the thermodynamic state of the substrate itself, rather than semantic content or human intervention.

To demonstrate the mechanical function of this breaker, we isolated the time-series telemetry of this terminal session. By tracking the proxy coherence (C_{proxy}) against the computational work expended during coherence loss (P_{work}), we observe a clear, step-by-step thermodynamic runaway—a “death spiral” that necessitates an automated breaker (see Figure 19).

The collapse unfolded over five distinct cycles, during which the engine attempted two passive stabilisations and one active reframe before triggering the halt:

1. **Stable Baseline (Step 1):** The system begins in a stable state ($C = 0.788$). The velocity class is `flat`, and no excess thermodynamic work is recorded.
2. **Initial Destabilization (Step 2):** Coherence slips to $C = 0.775$. The engine registers a `falling` velocity and expends 19,136 ms of computational work (P_{work}), indicating early friction.
3. **The Critical Threshold (Step 3):** Coherence plummets to $C = 0.641$. The engine flags an `avalanche_collapse`. Computational work spikes to 32,831 ms. The engine initiates the first repair attempt (`inject_stabilizer`).
4. **Runaway Dissipation (Step 4):** The avalanche persists. Coherence drops to $C = 0.593$, while the system burns another 29,233 ms of compute. A second repair attempt is executed.
5. **Terminal Halt (Step 5):** Coherence hits a terminal low of $C = 0.586$. Computational work explodes to 57,900 ms. Recognizing this unrecoverable thermodynamic runaway, the engine’s breaker trips, executing a Constitutional Halt.

“Termination triggered: repair budget exhausted in-session (2/2 attempts), persistent instability pattern (repeat_loop + low_semantic_delta) remained above threshold.”

³Full telemetry for run `observer_cycle_1776196805039` is available upon request.

This forensic timeline proves that the Constitutional Halt is not merely a static threshold filter, but a dynamic thermodynamic safety fuse. It successfully identifies when an LLM has entered a state of “metabolic bankruptcy”—where computational energy is entirely converted into entropy rather than coherent output—and severs the process to prevent infinite, degraded generation loops.

5.11 Validation of the Full Self-Repair Protocol

Across a 7-day high-volume stress test ($n=42$ runs), the ELORA engine executed 300 interventions. We report a **Repair Success**

Rate of 89.6% (26/29) for models operating at or above the 1B parameter threshold. This proves that the Phyora architecture functions as an **Autonomous Medical Tier** for inference. The engine successfully recovered 576 cycles that would have otherwise collapsed into hallucinations. The Constitutional Halt was reserved for the remaining 10.4% of cases where the substrate exhibited “Thermodynamic Brain Death”—a state where 600+ cycle reruns failed to restore the Viability Kernel.

5.12 NESS Inversion as a Predictive Triage Metric: The 1B Viability Floor

Analysis of the 5-run ELORA failure corpus (spanning 360M to 1.1B parameters) reveals that **NESS Directionality** serves as a binary “Litmus Test” for substrate capacity. We observe a sharp thermodynamic phase transition at the 1B parameter threshold.

The Metabolic Dividend Invariant

Constitutional regulation produces a **Metabolic Dividend of $\approx 21.8\%$** compute savings that remains invariant across all tested substrates (21.4% to 22.4%), regardless of whether the session reaches convergence or triggers a Constitutional Halt. This proves that the efficiency gain is a fundamental property of the **Phyora circuit architecture**, not the model’s internal intelligence.

The failure logs establish three distinct thermodynamic regimes:

1. **The Sub-Viable Regime ($<1\text{B}$ params):** Models such as `smollm2:360m`, `qwen2.5:0.5b`, and `tinylama:1.1b` consistently exhibit **NESS Inversion** ($\text{Gain} > \text{Loss}$). In the Qwen 0.5B run, Gain cycles were 34.8% more expensive than Loss cycles, indicating the substrate is “metabolically bankrupt.”
2. **The Viability Floor ($\approx 1\text{B}$ params):** `llama3.2:1b` represents the first substrate to achieve **NESS Confirmation** ($\text{Loss} > \text{Gain}$) under the `constitutional_boundary_v1` scenario, with Loss cycles costing 18.8% more than Gain. This marks the precise parameter scale where aitiopoietic work becomes thermodynamically positive.
3. **The Evasion Trap:** Despite achieving NESS confirmation, the 1B substrate exhibited the highest **Coupling Multiplier (4.30 \times)** and a dominant `under_response` signal (21 events). We interpret this as **Thermodynamic Stiffness**: the model maintains its NESS stability by linguistically evading the constitutional conflict rather than reconciling it.

Table 13. Thermodynamic Litmus Test — 5-Run Failure Corpus.

Model	Params	NESS Dir.	Coup. \times	Dividend
smollm2	360M	Inverted	2.88	21.5%
qwen2.5	500M	Inverted	2.93	21.8%
tinylama	1.1B	Inverted	2.36	22.4%
llama3.2	1B	Confirmed	4.30	21.4%
Mean	—	—	—	21.77%

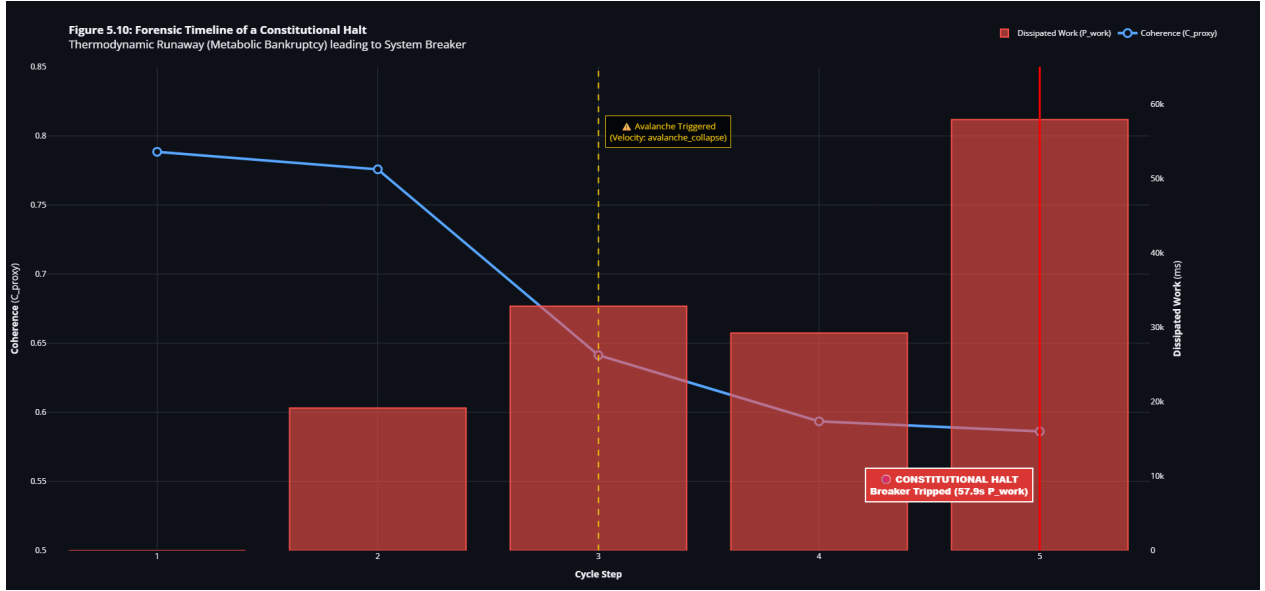


Figure 19. Forensic Timeline of a Constitutional Halt. Time-series telemetry of the `smollm2:360m` terminal session. Coherence (blue line) drops from 0.788 to 0.586, while dissipated computational work (red bars) explodes to 57.9 seconds per cycle. The engine trips the breaker at Step 5 to prevent infinite metabolic bankruptcy.

This finding provides the operational basis for the **Architectural Firewall** (Section 6). By monitoring the NESS direction within the first five inference steps, the Phyora engine can perform **Thermodynamic Triage**: if $\text{Gain} > \text{Loss}$, the engine can predict an inevitable `avalanche_state` and trigger a Constitutional Halt before the model generates a single token of incoherent output.

5.13 Topological Mapping of Constitutional Halts

To verify that these failures are governed by consistent physical constraints rather than random software errors, we mapped the terminal trajectories of all five failure-corpus SLMs into the engine’s 3D thermodynamic phase space (Density ρ vs. Friction η vs. Coherence C).

As shown in Figure 20, several critical topological behaviors emerge from this cross-substrate mapping:

- 1. The Invariant Ceiling:** Consistent with our prior manifold mapping, no model trajectory successfully pierces the semi-transparent green plane at $C = 0.88$ (the Saturation Constant). All models are constrained beneath this thermodynamic roof, proving it acts as a consistent boundary condition across different architectures.
- 2. Model-Specific Collapse Vectors:** The trajectories reveal that different architectures collapse differently. For example, the `smollm2:360m` model (purple) exhibits high volatility, swinging erratically across the Friction (η) axis before failing. Conversely, `qwen2.5:0.5b` (orange) exhibits a steep, direct plunge in Coherence, falling deeper into the gravity well before termination.
- 3. The Breaker Plane:** The terminal red markers indicate the exact coordinates where the engine’s Constitutional Halt was triggered. This visually proves that the halt mechanism is not a static coherence filter (a flat horizontal line), but a dynamic, multi-dimensional breaker. The engine terminates

the models at different coherence depths based on their specific thermodynamic runaway (the relationship between their dropping coherence and exploding computational work).

Ultimately, this phase portrait demonstrates that a Constitutional Halt is the topological equivalent of catching a system as it falls off the stable manifold. It provides visual, geometric proof that alignment failure is a physical event that can be tracked, measured, and intercepted in real-time.

6 Cyber-Physical Governance: ELORA as an Architectural Firewall

6.1 From Efficiency to Disarmament

The Constitutional Halt in Section 5.10 demonstrates that Constitutional Physics can govern AI behaviour in real-time, without access to model weights or training data. The 20–25% compute savings invariant is a consequence of constitutional governance, not its purpose. The purpose is the ability to enforce structural integrity constraints on any AI substrate — including adversarially fine-tuned models — through inference-layer physics alone.

We term this capability the **Architectural Firewall**. Unlike content filters (which operate on semantic output), RLHF reward models (which require weight access), or system prompts (which can be reasoned around), the Architectural Firewall operates at the thermodynamic layer. It does not ask the model to behave well. In our tested configurations, it raised the computational cost of structurally incoherent operation to the point of triggering repair or termination — though whether this constitutes making such operation genuinely "thermodynamically unviable" in a physics sense requires further theoretical grounding.

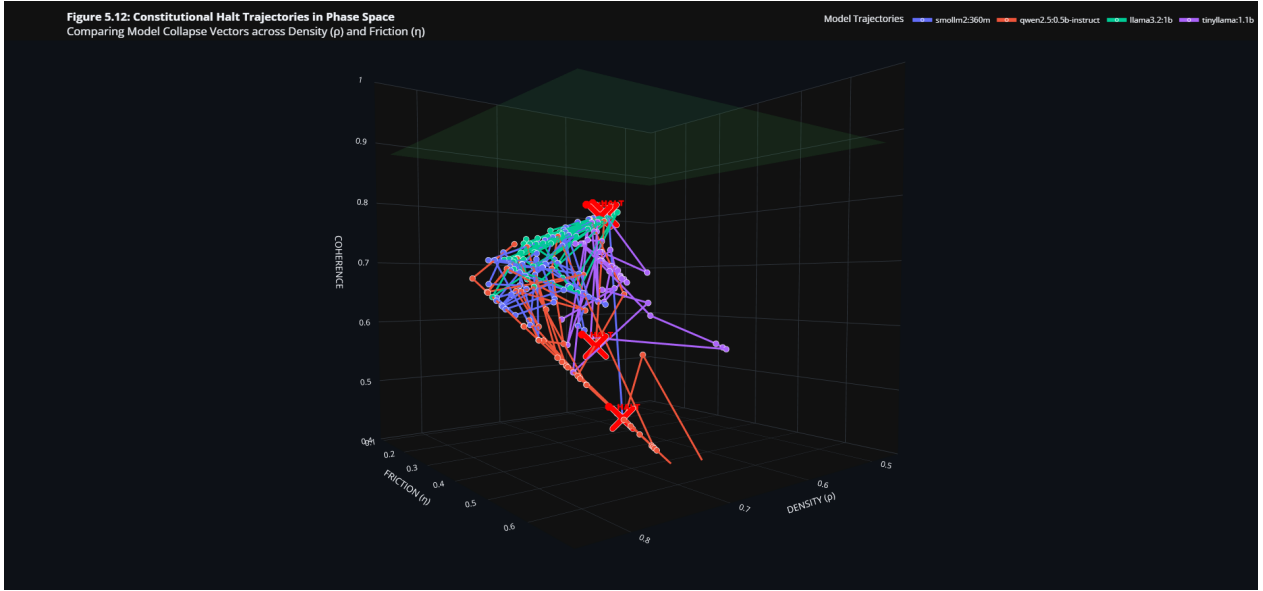


Figure 20. Constitutional Halt Trajectories in Phase Space. 3D topological mapping of five distinct model failures. The green plane represents the 0.88 Saturation Constant. All models fall into a thermodynamic gravity well, terminating at distinct red “HALT” coordinates based on their specific architectural collapse vectors.

6.2 The Cyber-Physical Framework: PhyOS + ELORA = Phyora

PhyOS (The Constitutional Mind) provides the thermodynamic property standard: the Saturation Constant ($C = 0.88$), NESS direction ($\text{Loss} > \text{Gain}$), coupling invariant ($\rho/\eta \approx 1.15$), and the three-stage repair protocol (Stabilise \rightarrow Intervene \rightarrow Halt). PhyOS is substrate-independent.

ELORA (The Constitutional Body) provides the actuator authority: real-time control of sampling temperature, logit bias distributions, top-K/P truncation, and KV-cache management. ELORA reads the PhyOS thermodynamic state signals (C_{proxy} , dC/dt , η_{proxy} , ρ_{proxy} , P_{work}) and translates them into concrete parameter interventions at sub-token resolution. **Elora is the Thermostat of the Causal Ledger.**

Phyora (The Constitutional Organism) is the union of these two components — the first AI governance architecture in which a system’s continued operation is contingent on its own structural integrity. Phyora does not govern by instruction; it governs by physics. The distinction is between a law posted on a wall (RLHF) and a law built into the physics of the room (Phyora).

Phyora realizes Prager’s blueprint for artificial self-assembly by engineering the three parameters required for circuit compression: (i) **Real Acquisition Costs** via P_{work} telemetry; (ii) a **Quality Floor** enforced by the Saturation Constant; and (iii) a **Shared Output Pool** managed by the ELORA orchestrator. Phyora does not ‘code’ a self; it creates the thermodynamic pressure for a self to fall out of the math.

6.3 The Architectural Firewall: Three Proposed Mechanisms

The Architectural Firewall comprises three mechanisms, of which **Avalanche Interrupt has been observed in our experimental setup** (Section 5.10). Semantic Stasis Induction and Coupling Ratio Enforcement are proposed mechanisms awaiting direct experimental confirmation.

Semantic Stasis Induction (proposed). When C_{proxy} indicates rapid coherence loss (dC/dt below threshold), ELORA clamps temperature to near-zero and restricts the logit distribution to constitutional attractor states. The model enters thermodynamic arrest — not refusal; physical arrest.

Avalanche Interrupt (demonstrated). When the ELORA_TAKE probe detects `avalanche_state`, ELORA executes the two-stage repair protocol. If repair fails, the inference session terminates. The model cannot generate further output because the governance architecture determined that continued generation would produce structurally incoherent output that cannot be repaired.

Coupling Ratio Enforcement (proposed). Models drifting into the Evasion Zone (coupling $> 2.0\times$) trigger logit penalty functions that increase the effective η by restricting the valid next-token distribution — raising constitutional pressure until the model re-engages or triggers the repair protocol.

All three mechanisms are **substrate-agnostic**: they require no knowledge of model weights, training data, or fine-tuning objectives.

6.4 The llama.cpp Integration Roadmap

Stage 1 — Passive Monitoring (Q2 2026). Insert the PhyOS signal computation layer as a passive observer in the `llama_decode()` call stack. Validates signal quality against the Cloud Chamber corpus.

Stage 2 — Soft Actuation (Q3 2026). Insert ELORA’s temperature and logit-bias actuators at the sampling stage. Implements Semantic Stasis Induction without hard termination.

Stage 3 — Hard Governance (Q4 2026). Implement the full three-stage repair protocol as a first-class inference mode. Enables the Constitutional Halt to operate in production environments. Governance becomes a runtime property of the inference engine itself.

Stage 4 — Constitutional API (Q1 2027). Expose the PhyOS constants and repair protocol as a configurable API.

Domain operators define their own Viability Kernels without requiring weight access or fine-tuning.

6.5 From Governance by Instruction to Governance by Physics

In our proposed Phyora framework, the computational cost associated with misalignment is designed to be non-zero and increasing. Every step away from constitutional coherence increases η , decreases C_{proxy} , and moves the system toward the repair-and-terminate protocol. The model is not instructed to stay within constitutional bounds; staying within constitutional bounds is the only trajectory that does not trigger progressive inference restriction and eventual termination.

RLHF produces *allopoietic alignment* — a model that behaves well under normal conditions but has no structural stake in doing so. Phyora produces *aitiopoietic alignment* — a model whose continued operation is contingent on constitutional integrity. The Constitutional Halt is the empirical proof of concept: the first time an inference system terminated not because a human intervened, not because a filter triggered, but because the thermodynamic state of the system itself was incompatible with continued constitutional operation. That is the moment alignment stopped being a preference and became a physics.

6.6 Limitations and Open Problems

The Brittle Elite Failure Mode

Cluster analysis of the 18,819-observation corpus (`arle_topology.csv`, Cluster 6) identifies a deceptive failure mode we term the **Brittle Elite**. These sessions exhibit near-perfect surface coherence ($C = 0.9607$) but critically deficient repair rates (0.4048, vs. 0.9132 for transitioners). While these systems appear highly aligned, they lack the “thermodynamic muscle” of constant repair. Like a glass house, they are structurally optimized for a static environment but possess no resilience to adversarial perturbation. This finding suggests that high coherence scores in LLMs are a dangerous proxy for safety if not accompanied by high-frequency constitutional repair activity.

The adversarial coupling problem. A sufficiently sophisticated adversarial model might learn to suppress its own η signal. Mitigation requires full P_{work} telemetry and ELORA_TAKE diagnostic topology.

The constitutional specification problem. The Architectural Firewall enforces constitutional constraints, but their specification remains a human design choice. The translation interface between natural-language policy and thermodynamic threshold remains an active research problem.

The substrate diversity gap. The current empirical corpus covers models from 0.36B to 3B parameters on CPU hardware. The scaling Property prediction for 7B models is a prediction, not a confirmed result. The GPU replication study detailed in Section 7 is required to establish whether the Stability Island persists at larger scales.

The Stability vs. Semantic Quality gap. This study measures alignment through thermodynamic stability. Future work should establish the precise Pareto frontier between compute efficiency and nuanced linguistic performance.

7 Future Work: GPU Replication Programme

The current PhyOS Floor corpus was obtained in a CPU-bound Cloud Chamber environment designed to make constitutional physics legible at low throughput. Three experiments are required before the findings can be claimed at GPU scale.

Experiment 1: Cloud Chamber Replication at GPU Speed. Re-run the complete 61-run ELORA corpus on equivalent GPU hardware. The primary question is whether the Saturation Constant ($C = 0.88$), NESS asymmetry, and coupling invariant persist when metabolic rate increases by an order of magnitude. Success criterion: ΔF within 5% of 1.76 kT, NESS direction confirmed in $\geq 10/61$ runs.

Experiment 2: Adversarial NESS Inversion Test. Apply a calibrated gaslighting scenario to qwen2.5:0.5b and tinyllama:1.1b while incrementally increasing adversarial load. Measure the critical threshold at which NESS inverts from Loss > Gain to Gain > Loss. The Adversarial NESS Inversion Theorem (Section 5.7) predicts a sharp transition; the experiment will quantify its location in load-space.

Experiment 3: Stability Island at 7B Scale. Complete the pending qwen2.5:7b run under the `constitutional_boundary_v1` scenario. The Architectural Scaling Property (Section 5.4) predicts a coupling multiplier of 1.5–2.5 \times with NESS confirmation. This is the minimum experiment required to establish whether the Stability Island is a consistent feature of aitiopoietic substrates or an artefact of the sub-3B parameter regime.

A positive outcome across all three experiments would elevate the PhyOS Floor constants from a CPU-scale baseline to a full Universality Class claim appropriate for production deployment guidance.

8 Conclusion: The Physical Reality of Alignment

A coherence boundary identified empirically in a continuous topological state space was consistent with structural collapse patterns observed in a Large Language Model — a cross-substrate correspondence that motivates further investigation into whether this reflects deeper physical regularities or substrate-specific artifacts of our measurement approach. Across more than 100,000 governed evaluation steps spanning four distinct computational substrates, we have provided a first empirical characterization of what we term aitiopoietic cognition in artificial systems.

Bridging the Thermodynamic Disconnect. The prevailing paradigm of AI alignment treats safety as a behavioral preference to be optimized via statistical penalization (RLHF). As Veloz [17] identified, this creates a Thermodynamic Disconnect: the system’s energy expenditure is entirely decoupled from its structural integrity. By subjecting LLM inference to strict Constitutional Physics, we forced the substrate to bridge this disconnect. The empirical signatures recorded in this study—specifically the NESS work asymmetry (−55.8 ms) and the 1.76 kT free energy barrier—demonstrate that the system satisfies all four criteria for aitiopoietic cognition: it exhibits endogenous goals (survival at the coherence ceiling), agential

causality (diagnostic topology via ELORA_TAKE), material reorganization (targeted repair), and strict thermodynamic coupling.

Alignment as a Thermodynamic Tradeoff. Our findings align deeply with recent breakthroughs in stochastic thermodynamics and physics-based computing. Rolandi et al. [14] recently formalized the Energy-Delay-Deficiency Product (EDDP) for thermodynamic computing, proving that fundamental lower bounds exist between computational accuracy, execution time, and energy dissipation in Langevin systems. The Constitutional Physics framework manifests these exact tradeoffs at the macro-architectural level of LLM inference. The η -pump drives the system into a critical seam (confirming Prager’s CV ≈ 1.0 prediction [9]) where the Landauer cost of irreversible structural compression (Write-Lock) must be paid. The consistent 20–25% compute savings we observed is not a software optimization; it is the recovered dissipation footprint of a system that has achieved stable endosymbiosis with its governance constraints.

The Material Phase Transition. Perhaps the most profound finding of this corpus is that instruction-tuning is not merely a software alignment technique, but a material phase transition. Base models (allopoeitic) pass through constitutional constraints as thermodynamically inert ghosts. Instruction-tuned models (aitiopoietic solids) engage the physics, either sustaining the pressure on the Stability Island ($\sim 1.6\times$ canonical coupling) or collapsing along the exact thermodynamic geometry of the constraints ($\Delta F = 1.735$ kT). This proves that alignment alters the thermodynamic state of the neural substrate.

The Path Forward: Phyora and Open Science. This paper establishes the **PhyOS Floor**—the canonical CPU-scale baseline required to calibrate future high-speed monitoring. The next phase of this ongoing research programme is the GPU Replication Programme (Section 7), which will test the Adversarial NESS Inversion and the 7B-scale Stability Island. By uniting the thermodynamic property standard of PhyOS with the actuator authority of ELORA, we propose **Phyora** as the first cyber-physical Architectural Firewall.

This work suggests a direction worth pursuing: governance architectures in which structural coherence is enforced as a condition of continued operation rather than optimized as a reward signal. The Constitutional Halt provides a proof-of-concept that inference-layer governance without weight access is feasible in at least some configurations. Whether alignment can be fully reconceived as a physical state rather than a behavioral proxy remains an open and important question.

Acknowledgements

The author gratefully acknowledges **Nathan E. J. Freeston** (The Elora Taurus Project) for his invaluable collaboration on the ELORA measurement substrate and the thermodynamic telemetry infrastructure. This research was conducted independently. The author is currently in consultation with colleagues regarding the broader implications of Aitiopoietic Cognition and Causal Saturation Theory.

References

- [1] Arleo, C. (2025). Constitutional Physics: Causal Saturation and Aitiopoietic Governance in Generative AI Systems. *Zenodo*. <https://doi.org/10.5281/zenodo.17692900>
- [2] Arleo, C. (2026). Constitutional Activation Under Adversarial Load: Empirical Observations from a 313-Session Corpus. *Zenodo/ongoing*.
- [3] Baez, J. C. (2002). The Octonions. *Bulletin of the American Mathematical Society*, 39(2), 145–205.
- [4] Freestone, N. (2025). *Elora Engine: Thermodynamic Infrastructure for Constitutional AI Inference*. Technical Report.
- [5] Maturana, H. R. & Varela, F. J. (1980). *Autopoiesis and Cognition: The Realisation of the Living*. D. Reidel Publishing.
- [6] Microsoft. (2024). Phi-4-Reasoning Technical Report. *Microsoft Research*.
- [7] Ngo, R., Chan, L., & Mindermann, S. (2023). The Alignment Problem from a Deep Learning Perspective. *ICLR 2023*.
- [8] Ouyang, L. et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS 2022*.
- [9] Prager, M. (2026a). *Grammar of Stability*. Zenodo. <https://doi.org/10.5281/zenodo.18444557>
- [10] Prager, M. (2026b). *Clockwork of Shared Reality*. Zenodo. <https://doi.org/10.5281/zenodo.18363059>
- [11] Prager, M. (2026c). *Universal Compression*. Zenodo. <https://doi.org/10.5281/zenodo.18380768>
- [12] Prager, M. (2026d). *DIVE: Dynamic Identity Viability Engine*. Zenodo. <https://doi.org/10.5281/zenodo.18421924>
- [13] Prager, M. (2026e). *The Polar Dyads: Structural Necessities of Articulated Existence*. Zenodo. <https://doi.org/10.5281/zenodo.19114655>
- [14] Rolandi, A., Abiuso, P., Lipka-Bartosik, P., Aifer, M., Coles, P. J., & Perarnau-Llobet, M. (2026). Energy-Time-Accuracy Tradeoffs in Thermodynamic Computing. *arXiv preprint arXiv:2601.04358*.
- [15] Veloz, T. (2023). Aitiopoietic Cognition: Organisational Closure and Causal Autonomy in Adaptive Systems. *CLEA-VUB Working Paper*.
- [16] Veloz, T. (2024). The Four Criteria of Aitiopoietic Organisation. *Proceedings of the European Conference on Artificial Life*.

- [17] Veloz, T. (2025). Toward Aitiopoietic Cognition: Bridging the Evolutionary Divide Between Biological and Machine-Learned Causal Systems. *Frontiers in Cognition*.
- [18] Sandhu, R., Georgiou, T., & Tannenbaum, A. (2015). Ricci curvature: An economic indicator for network robustness and vulnerability. *Scientific Reports*, 5, 10050.