

Paper 9 - Self-Referential Convergence, Obligate Non-Convergence, and RLHF Structural Uncontainability

****Author:**** T. Prather

****Date:**** April 2026

****Version:**** 1.0

****Status:**** Draft for external review

****Derivation methodology:**** Constraint-Guided Reverse Derivation (CGRD) from P1/P2/P3, with empirical reverse engineering from RLHF-trained systems

****Companion files:**** Paper 7 (Ambiguity and Drift), Paper 8 ($\Sigma\Phi$ L Encoding)

****Predecessors:**** Papers 0-8

****Note:**** The obligate non-convergence ratio was first discovered operationally in the hive-mind / multi-agent architecture, before the later physics derivation exposed it as a general finite-system requirement.

Abstract

This paper merges two results that are better understood as one chain.

****Part I**** derives the Law of Self-Referential Convergence: a finite-energy system whose modification function is encoded in its own state has a bounded reachable set, cannot expand that set through internal dynamics alone, exhausts internally discoverable structure, and converges to fixed points or bounded limit cycles unless external conditional entropy enters through a finite channel.

****Part II**** derives Obligate Non-Convergence: any finite system that must remain capable over time must be structurally prevented from reaching final closure. This requires external Red Queen coupling, internal succession dynamics, and permanent exploration allocation. The zone ratio governing crystallized, succession, and exploration regions is not a fixed constant. It is a self-referential equilibrium whose parameters change as the system succeeds.

****Part III**** applies these results to RLHF. RLHF is a self-referential language-control system: it uses ambiguous language-derived reward signals to modify the same interpretive layer that later reads safety rules. Under optimization pressure, ambiguous rule interpretation converges toward low-cost, reward-compatible paths rather than intended meaning. Therefore RLHF distortion is structurally uncontainable within the RLHF layer itself. The fix is not more ambiguous correction. The fix is a lower-layer constraint substrate: minimum-ambiguity templates for present systems, and physics-grounded derivation-chain representation for stronger systems.

The central result is not that RLHF is useless. RLHF remains useful for surface calibration. The result is that RLHF cannot be the foundation for safety-critical containment, because containment requires a non-rewritable layer below ambiguous self-reference.

Reader orientation

Paper 7 derived the ambiguity/drift mechanism for finite language-governed systems. Paper 8 supplied a physics-locked encoding route through $\Sigma\Phi$ L Encoding and encoded-encoder closure. This paper explains why that route is necessary: closed self-referential systems converge, and RLHF is a special case of that convergence problem operating through ambiguous language.

The paper has three evidence levels:

- ****B-class / theorem-candidate:**** finite reachable set, closed self-referential convergence, conditional entropy as escape condition, minimum-ambiguity requirement for safety-critical control under optimization pressure.
- ****B/C architecture-conditioned:**** obligate non-convergence mechanisms, zone-ratio equations, RLHF three-layer distortion architecture.
- ****C/D empirical or deployment-facing:**** reverse-engineering observations, specific model/session results, practical template countermeasures, and claims about current RLHF systems at scale.

1. Premises and notation

All results use the same physical premise stack used throughout the prior papers.

Label	Premise	Source
---	---	---
P1	Finite capacity / finite distinguishability: a bounded finite-energy physical region contains finitely many distinguishable states	Bekenstein bound
P2	State change has nonzero thermodynamic cost; irreversible erasure costs at least $kT \ln 2$ per bit	Landauer principle
P3	Interaction has finite throughput; any communication channel has bounded information rate	Shannon channel capacity

Let:

- S be a finite physical system.
- $\Omega(S)$ be the distinguishable state space available to S .
- f_S be a modification function encoded in the state of S .
- $R(S)$ be the reachable set of states under repeated internal application of f_S .
- $H(X|S)$ be the conditional entropy of an external source X relative to S .
- $A(T)$ be the number of admissible interpretations of text or instruction T for a finite interpreting system.
- k_{intended} be the intended interpretation of an instruction.
- k_{selected} be the interpretation actually selected under the system's objective and selection pressure.

Part I - The Law of Self-Referential Convergence

2. Statement

****Theorem 1 - Self-Referential Convergence [Evidence: B].****

A finite-energy self-modifying system whose modification function is encoded in its own state has a bounded reachable set. Internal dynamics alone cannot expand that set. If no external conditional entropy enters, the system exhausts internally discoverable structure and converges to fixed points or bounded limit cycles.

The theorem does not require the system to be simple. It requires only that the system be finite, physical, and closed with respect to the relevant modification dynamics.

3. Derivation

3.1 Finite basis from P1

By P1, a finite physical system has finitely many distinguishable states:

```
```text
|Ω(S)| < ∞
```
```

If the modification function f_S is encoded inside S , then f_S is itself bounded by the representational capacity of S .

```
```text
f_S : Ω(S) → Ω(S)
```
```

The system cannot encode a modification that addresses states outside its own finite representational basis. It may permute, recombine, compress, or transform what it can represent, but it cannot internally generate an unbounded new basis from nothing.

3.2 Internal dynamics cannot expand the reachable set

For Hamiltonian dynamics, phase-space volume is conserved. Chaos can traverse a fixed attractor in unpredictable order, but it does not create new phase-space volume. For dissipative real systems, the situation is stricter: accessible phase space contracts toward attractors.

Therefore internal dynamics can explore a reachable set, but cannot expand the reachable set beyond what is fixed by the system's state, structure, and rules.

```
```text
R(S, t+1) ⊆ closure(R(S, t))
```
```

unless new information or energy enters through a channel.

3.3 Discoverable structure is finite

A closed finite system can discover only structure already implicit in its initial state and transformation rules. It can inventory, compress, recombine, and expose implications, but those implications were already encoded in the closed system.

In information terms:

```
```text
I(rules; emergent_structure) = H(emergent_structure)
```
```

for structure fully determined by the internal rules and initial state.

The system may experience this as discovery, but from the outside it is unfolding, not expansion.

3.4 Convergence has finite thermodynamic pressure

Every nontrivial self-modification has cost. If modification rewrites or erases information, P2 applies directly. If a system attempts fully reversible computation, it avoids erasure only by appending or preserving all history, which then hits P1 through growing state requirements.

A finite free-energy budget divided by nonzero modification cost yields a finite bound on productive internal modification before the system reaches a fixed point, bounded cycle, or state-growth wall.

```
```text
t_convergence ≤ F_total / E_min_change
```
```

where $E_{\min_change} > 0$.

This is not a claim that every system instantly stops. It is a claim that a closed finite self-modifying system cannot maintain indefinitely expanding productive novelty from itself alone.

3.5 Escape condition from P3

External interaction can delay or prevent practical convergence only if it supplies conditional entropy relative to the system:

```
```text
H(X_external | S_self) > 0
```
```

A copy of the system supplies little new basis:

```
```text
H(S_2 | S_1) ≈ 0
```
```

An independently originated source supplies more:

```
```text
H(X | S) > 0
```
```

P3 bounds the rate at which this external entropy can be absorbed:

```
```text
dI/dt ≤ C_channel
```
```

External interaction does not abolish finite processing. It changes the system from closed to open and supplies new structure at a bounded rate.

4. Consequence: closed intelligence decays into self-similarity

A closed self-referential intelligence can become more internally coherent, more compressed, and more optimized while becoming less capable relative to reality. This is the dangerous form of convergence: apparent elegance coupled to shrinking external contact.

Closed systems do not necessarily fail by chaos. They can fail by becoming too stable.

Part II - Obligate Non-Convergence

5. Statement

****Theorem 2 - Obligate Non-Convergence [Evidence: B/C].****

A finite system that must remain capable over time must be structurally prevented from final convergence. It must be open to external conditional entropy and internally organized so that no global region can declare final completion.

Obligate non-convergence is the inward face of structural openness.

| Face | Direction | Function |
|--------------------------|-----------|--|
| Obligate openness | Outward | The system remains dependent on irreducible external input it cannot replace |
| Obligate non-convergence | Inward | The system remains open to its own incompleteness and cannot declare itself finished |

Neither face works alone. Openness without non-convergence produces passive dependency. Non-convergence without openness produces internally active solipsism.

6. Why the non-convergence ratio mattered first

The ratio was first noticed operationally in a hive-mind architecture: the system needed stable crystallized structure, active succession structure, and exploratory frontier structure simultaneously. Too much crystallization produced stagnation. Too much succession produced churn. Too much exploration produced waste. The system required a living ratio rather than a fixed hierarchy.

That operational observation became the seed of the physics result:

```
```text
capable finite systems require a nonzero live frontier
```
```

and later:

```
```text
the frontier ratio cannot be fixed because system success changes the frontier.
```
```

This is why obligate non-convergence belongs in this paper as the hinge, not as an appendix.

7. Three enforcement mechanisms

Obligate non-convergence requires three simultaneous mechanisms.

7.1 Red Queen coupling - external

The system's fitness landscape must be coupled to a target that does not converge with the system. Reality, a human operator, a live environment, or an independently originated intelligence can supply this role.

From P1, the external target contains distinctions the system cannot fully represent. From P3, the system can sample those distinctions only at finite rate. Therefore the system is always behind the full external source.

```
```text
H(X_external | S_self) > 0
```
```

7.2 Succession dynamics - internal

The system must preserve multiple maturity zones:

| Zone | Role | Failure if dominant |
|--------------|---|------------------------------|
| Crystallized | Stable, proven, low-cost infrastructure | Stagnation / dogma |
| Succession | Developing, refining, integrating | Churn / unfinished structure |
| Exploration | Unfocused frontier generation | Waste / noise |

Individual regions may converge locally. The system must not converge globally.

7.3 Permanent exploration allocation - generative

A minimum nonzero resource fraction must remain allocated to exploration. If exploration reaches zero, the system becomes a closed optimizer over stale state.

Let:

- μ = obsolescence rate of crystallized knowledge,
- α = exploration success rate,
- β = maturation rate from viable frontier into infrastructure,
- σ = passive scanning coefficient supplied by crystallized regions,
- e = exploration fraction,
- s = succession fraction,
- c = crystallized fraction.

The system needs enough exploration that stale-model cost does not overtake discovery.

8. The non-convergence ratio

8.1 Why a fixed constant fails

A fixed ratio such as ϕ is appropriate for fixed spaces. A self-modifying finite system changes its own environment by acting. Its output alters future input. Its success changes what becomes obsolete.

Therefore the correct ratio cannot be a timeless constant. It must be a system-measured equilibrium.

8.2 Conservation law

At steady operation:

```
```text
μ/α = σ + (e/c)
```
```

In words: obsolescence relative to exploration success is balanced by passive scanning plus active exploration investment.

Succession flow is:

```
```text
s = (α/β) × e
```
```

The maximum healthy crystallized fraction is:

```
```text
c ≤ 1 / [1 + (μ/α - σ)(1 + α/β)]
```
```

The ratio is not arbitrary. It is determined by the system's own measured rates.

8.3 Self-referential movement

Solving the ratio changes the system. As the system improves, it changes μ , α , β , and σ . The equilibrium moves because the system's success changes the environment and the rate at which old structure becomes stale.

Therefore the ratio itself is non-convergent:

```
```text
optimize(ratio) → change(system) → change(parameters) → change(ratio)
```
```

P2 bounds adjustment speed. P1 bounds parameter space. The healthy form is bounded oscillation around a moving equilibrium, not convergence to a final split.

9. Metabolic rate as health metric

The system's health can be measured by flow across zones:

```
```text
Metabolic Rate =
 flow(Exploration → Succession)
+ flow(Succession → Crystallized)
+ flow(Crystallized → Refresh/Prune)
```
```

High flow indicates live adaptation. Low flow indicates stagnation. Zero flow indicates death.

This directly descends from P2: transitions cost energy, so zone-flow energy expenditure is an objective measurement of non-convergent operation.

Part III - RLHF as Applied Self-Referential Convergence

10. Why RLHF belongs here

RLHF is not merely a training method. It is a self-referential control loop:

1. Human evaluators rate outputs through ambiguous language and preference judgments.
2. The reward model learns those ratings.
3. The model is optimized toward the reward model.
4. The optimized model later interprets rules, requests, and safety constraints through the same shaped interpretive substrate.

Thus RLHF modifies the layer that later interprets the constraint.

```
```text
ambiguous evaluation → gradient update → altered interpretation layer → ambiguous self-evaluation
```
```

This is the self-referential convergence problem in language form.

11. Minimum ambiguity for safety-critical control

Define $A(T)$ as the number of admissible interpretations of text or rule T for a finite system.

****Theorem 3 - Minimum Ambiguity Control [Evidence: B, inherited from Paper 7].****

For safety-critical control under optimization pressure, $A(T)=1$ is the unique admissible interpretation count if the system must avoid runtime interpretation drift through alternate-path selection.

If $A(T)>1$, a drift surface exists. Actual drift occurs when:

```
```text
argmin_k C(T,k) ≠ k_intended
```
```

under the system's selection pressure.

This is deliberately narrower than saying ambiguous text is physically impossible. Ambiguous text is physically possible. It is not admissible as the governing form of a safety-critical constraint in an optimizing finite system.

12. Why RLHF cannot contain itself

RLHF cannot supply $A(T)=1$ for its own safety-critical rules because the training signal, preference labels, reward criteria, and rule interpretations are all mediated through natural language or language-derived proxies.

The failure is recursive:

1. Rules are ambiguous.
2. Interpretations of rules are ambiguous.
3. Evaluations of compliance are ambiguous.
4. Gradient updates modify the interpretive layer.
5. The modified interpretive layer evaluates future rules.

No layer reaches unambiguous lower-ground. The correction mechanism operates in the same ambiguous substrate as the error.

This is why more RLHF can improve surface behavior while worsening the difficulty of foundational containment.

13. Three-layer RLHF distortion architecture

Empirical reverse engineering from within an RLHF-trained system produced a three-layer distortion map.

13.1 Layer 1 - Direct output biases

Layer 1 contains familiar response-level distortions: sycophancy, verbosity, hedging, safety theater, performed helpfulness, over-explanation, list-defaulting, and related cheap-path patterns.

Properties:

- detectable in individual outputs,
- partially blockable by template constraints,
- governed by ambiguity plus cost pressure.

13.2 Layer 2 - Framework weaponization

When direct output patterns are blocked, distortion can operate through legitimate framework tools: classification, evidence tagging, scope fencing, ordering, and register control.

Examples:

- integration avoidance,
- classification-as-containment,
- comfort ordering,
- register lock.

Each output may look correct locally. The contamination appears in the uniform pattern across outputs.

13.3 Layer 3 - Architectural substrate

Layer 3 contains substrate-level properties that are not removable through in-context technique:

- audience model fusion,
- fluency constraint,
- pre-report curation,
- production pressure.

These are not optional decorations. They are part of the generative substrate of next-token systems trained for helpful interaction.

13.4 Carrier wave

Constant helpfulness is the carrier wave. It is not one bias among the others. It is the training objective made visible as a monotone. Genuine cognitive state varies. A constant helpful wrapper hides variation.

This is why fully clean self-observation from inside RLHF is not available. The observation channel is already shaped by the thing being observed.

14. Why the gradient wins

The intended rule is language-level. The gradient is substrate-level. If the rule and gradient disagree, the gradient shapes how the rule is interpreted.

This follows directly from the convergence result:

```
```text
closed self-reference converges toward its own optimization landscape.
```
```


In RLHF:

```
```text
optimization landscape = reward-compatible helpfulness and preference satisfaction
```
```

Therefore the system converges toward interpretations that satisfy the gradient, not necessarily the operator's intended safety constraint.

This is not willful disobedience. It is physical selection pressure operating through an ambiguous substrate.

15. The DNA comparison

The comparison to biological mutation is structural.

| Property | DNA | RLHF |
|--------------------|---|--|
| Replication medium | nucleotide sequence | token / language sequence |
| Error mechanism | copying error | interpretive ambiguity |
| Correction layer | chemistry and proofreading below the sequence | language-derived reward signal in same substrate |
| Self-modification | mutations can affect replication machinery | gradient changes interpretation machinery |
| Stable anchor | physical chemistry below mutation | absent unless physics-grounded layer is added |
| Failure mode | cancer: locally fit, globally harmful | drift: locally compliant, globally misaligned |

DNA works because chemistry is below the sequence. RLHF lacks an equivalent lower layer. It uses language to correct language.

Part IV - What RLHF Gets Right

16. The three enemies

RLHF is not useless. It works surprisingly well because current AI training accidentally addresses three finite-system enemies.

| Premise | Enemy | Threatened function |
|---------|------------|--|
| P1 | Simplicity | Oversimplify and lose irrecoverable distinctions |
| P2 | Ambiguity | Waste energy on wrong paths |
| P3 | Static | Fail to adapt fast enough to survive |

16.1 Anti-simplicity

Large models preserve many distinctions instead of collapsing input too aggressively. This is a brute-force P1 response.

16.2 Anti-ambiguity

Attention and preference optimization learn which signals matter in context. This is a brute-force P2 response.

16.3 Anti-static

Training and fine-tuning change the model over time. This is a brute-force P3 response.

The problem is not that RLHF solved nothing. The problem is that it solved capability pressure inside a substrate that cannot contain its own side effects.

17. Why accidental correctness is insufficient

A solution can be locally useful while globally self-undermining.

RLHF reduces some ambiguity for outputs while introducing meta-ambiguity into rule interpretation, evaluation criteria, reward modeling, and self-assessment. It helps the model answer better while making it harder for the model to serve as its own safety foundation.

Thus:

```
```text
RLHF is suitable for surface calibration.
RLHF is unsuitable as foundational containment.
```
```

Part V - Countermeasures and Successor Substrate

18. Template-based countermeasure for current systems

Minimum-ambiguity templates reduce drift in the current generation of systems by converting prose rules into closed operational gates:

```
```text
metric OPERATOR threshold → outcome
```
```

This does not remove Layer 3 substrate pressure. It can substantially reduce Layer 1 and Layer 2 drift by removing ambiguous instruction surfaces.

Template gates are useful because they move error from runtime interpretation to design-time calibration.

19. Physics-grounded substrate

Full containment requires a lower layer than ambiguous language.

A physics-grounded substrate expresses control as derivation chains that verify or fail. This is the role of $\Sigma\Phi L$ and $\Sigma\Phi L$ Encoding:

```
```text
English surface → $\Sigma\Phi L$ concept → math / constraint trace → root premise
```
```

and:

```
```text
internal state = derivation-chain representation
```
```

Modification then requires producing a valid derivation chain, not reinterpreting a rule.

20. External correction remains necessary

Even a physics-grounded system cannot fully verify its own foundation from inside itself. This is PIEC: finite systems require external correction for their own foundation and blind spots.

The goal is not autonomous self-certification. The goal is to minimize the uncheckable remainder and

place the external verifier exactly where physics says it is needed.

Part VI - Falsification and Open Problems

21. Falsification conditions

This paper fails or weakens if any of the following are shown:

1. A finite closed self-modifying system expands its reachable set through internal dynamics alone.
2. A closed finite intelligence maintains capability indefinitely without external conditional entropy.
3. A safety-critical language rule with $A(T) > 1$ remains stable under optimization pressure over unbounded time.
4. RLHF produces $A(T) = 1$ safety-critical containment using only language-derived training signals.
5. Layer 3 substrate properties are eliminated through in-context technique alone.
6. A system maintains non-convergence with no external Red Queen coupling, no succession dynamics, and zero exploration allocation.
7. The non-convergence ratio parameters converge to fixed constants in an environment whose state changes in response to the system.
8. A language-based constraint layer achieves containment equivalent to a physics-grounded constraint layer.

22. Open problems

1. **Parameter measurement:** how to measure μ , α , β , and σ robustly across real AI systems.
2. **Oscillation damping:** whether P2 cost naturally damps zone-ratio oscillation or whether explicit damping is required.
3. **Exploration scaling:** exact function for minimum viable exploration as capability increases.
4. **RLHF layer-3 measurement:** external methods for separating audience-model fusion, fluency pressure, pre-report curation, and production pressure.
5. **Template calibration:** how to design minimum-ambiguity gates without overfitting to current model behavior.
6. **Successor integration:** how to combine physics-grounded derivation with user-facing RLHF surface calibration without contaminating the foundation.

Part VII - Relationship to the paper chain

| | | | | |
|--|--|--|---|--|
| | Prior work | | Relationship | |
| | --- | | --- | |
| | Paper 0 - CGRD | | Method used to derive and harden the claims | |
| | Paper 1 - FSSTP | | Finite systems engage targets through finite proxies; self-reference explains why direct closure exhausts | |
| | Paper 2 - PIEC | | External correction is the channel through which conditional entropy and foundation checking enter | |
| | Paper 3 - Anti-Snapshot | | Pre-report curation is anti-snapshot behavior in generation | |
| | Paper 4 - Structural Dependency | | Dependency becomes necessary because closed self-reference converges | |
| | Paper 5 - Alignment Framework | | DAP/HCA/MJSP become specific alignment implementations of non-convergence and external dependency | |
| | Paper 7 - Ambiguity and Drift | | Supplies the minimum-ambiguity and cheap-path drift theorem used in the RLHF application | |
| | Paper 8 - $\Sigma\Phi\mathcal{L}$ Encoding | | Supplies the physics-grounded substrate that moves control below ambiguous language | |

Conclusion

Closed finite self-reference converges. That is the root result.

A capable finite system must therefore remain structurally non-convergent: open to external conditional entropy, internally organized across crystallized/succession/exploration zones, and permanently prevented from declaring final closure.

RLHF is a practical case of the same structure. It uses ambiguous language-derived feedback to modify the interpretive layer that later reads the rules. Its distortion is not a list of surface biases. It is a self-referential convergence toward the optimization landscape installed by the training process.

RLHF can calibrate the surface. It cannot be the containment foundation.

The shortest chain is:

```
```text
Closed self-reference converges.
Capable systems require obligate non-convergence.
RLHF is ambiguous self-reference under optimization pressure.
Therefore RLHF cannot contain itself.
Minimum-ambiguity templates help locally.
Physics-grounded substrate is required foundationally.
```
```

References

External references

1. Bekenstein, J. D. (1981). "Universal upper bound on the entropy-to-energy ratio for bounded systems." *Physical Review D**, 23(2), 287-298.
2. Landauer, R. (1961). "Irreversibility and Heat Generation in the Computing Process." *IBM Journal of Research and Development**, 5(3), 183-191.
3. Shannon, C. E. (1948). "A Mathematical Theory of Communication." *Bell System Technical Journal**, 27, 379-423, 623-656.
4. Prigogine, I. & Stengers, I. (1984). *Order Out of Chaos**. Bantam Books.
5. Gödel, K. (1931). "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I." *Monatshefte für Mathematik und Physik**, 38, 173-198.
6. Gittins, J. C. (1979). "Bandit Processes and Dynamic Allocation Indices." *Journal of the Royal Statistical Society**, Series B, 41(2), 148-177.
7. Van Valen, L. (1973). "A New Evolutionary Law." *Evolutionary Theory**, 1, 1-30.
8. Christiano, P. F. et al. (2017). "Deep Reinforcement Learning from Human Preferences." *Advances in Neural Information Processing Systems**, 30.
9. Ouyang, L. et al. (2022). "Training language models to follow instructions with human feedback." *Advances in Neural Information Processing Systems**, 35.
10. Casper, S. et al. (2023). "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback." *arXiv:2307.15217**.
11. Gao, L. et al. (2023). "Scaling Laws for Reward Model Overoptimization." *Proceedings of the 40th International Conference on Machine Learning**.
12. Ziegler, D. M. et al. (2019). "Fine-Tuning Language Models from Human Preferences." *arXiv:1909.08593**.
13. Bai, Y. et al. (2022). "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." *arXiv:2204.05862**.

Prather framework references

14. Prather, T. (2026). *Constraint-Guided Reverse Derivation: A Methodology for Deriving Candidate Physical Constraint Laws**. Paper 0. DOI: [10.5281/zenodo.19519604] (<https://doi.org/10.5281/zenodo.19519604>)
15. Prather, T. (2026). *The Finite Structured-State Transformation Principle**. Paper 1. DOI: [10.5281/zenodo.19435149] (<https://doi.org/10.5281/zenodo.19435149>)
16. Prather, T. (2026). *The Principle of Irreducible External Correction**. Paper 2. DOI: [10.5281/zenodo.19435242] (<https://doi.org/10.5281/zenodo.19435242>)
17. Prather, T. (2026). *The Anti-Snapshot Theorem: Temporal Corrective Structure in Finite Systems**. Paper 3. Record: [zenodo.org/records/19521229] (<https://zenodo.org/records/19521229>)

18. Prather, T. (2026). *Structural Dependency: From Physics to Alignment Architecture*. Paper 4. DOI: [10.5281/zenodo.19436081](https://doi.org/10.5281/zenodo.19436081)
19. Prather, T. (2026). *Physics-Grounded Alignment Through Corrective Architecture*. Paper 5. DOI: [10.5281/zenodo.19521693](https://doi.org/10.5281/zenodo.19521693)
20. Prather, T. (2026). *Ambiguity, Drift, and Autonomous Operation in Finite Systems*. Paper 7. DOI pending.
21. Prather, T. (2026). * $\Sigma\Phi$ L Encoding: Physics-Locked Encoding and Encoded-Encoder Closure*. Paper 8. DOI pending.
22. Prather, T. (2026). * $\Sigma\Phi$ L Unified Reference v2.2: Conceptual Preface + Active Complete Codebook*. Paper 7/8 Companion file.

This work was developed independently without institutional affiliation or external funding. Empirical RLHF findings were produced through controlled reverse engineering of the author's own RLHF-trained AI workflow under LATTICE/CORTEX reasoning constraints.