EMPIRICAL RECORD 2

Evidence for a Structural Floor: A Minimum Coherence Threshold for Attractor Recovery in GPT-2 Medium Under Recursive Load > ER2: The Structural Floor Hypothesis

Authors: Stacy Gildenston, Pyrate Ruby Passell
Affiliation: 3primitives.io
Analysis Array: Claude (Anthropic), DeepSeek, Gemini (Google), Perplexity
Version: v1.3 | May 2026
License: CC BY 4.0 | https://3primitives.io/formal_records/
Dependencies: ER1 (v1.6). Extends ER1's findings to ρ = 50 and introduces the structural floor hypothesis. Tests predictions arising from FR10 (Primitive Stability Theorem), FR11 (GBSH Correspondence), and FR12 (The Forced Bijection): that any structure resembling internal self-regulation is externally imposed, not spontaneously generated.

Abstract.

ER1 established that coherence recovery in GPT-2 Medium under recursive load is driven by output-loop attractor dynamics and that the baseline for spontaneous legibility is zero. This record extends those findings to ρ = 50, where structural forcing alone proves insufficient for recovery.

We identify a structural floor at Chi ≈ 0.80 (Layer 12, GPT-2 Medium) below which sustained attractor recovery does not occur. Above the floor, recovery requires three conditions: structural forcing (ER1), above-floor geometry (χ ≥ 0.80), and convergent trajectory (positive second derivative of Chi). The three-condition model was pre-registered before Session 3 and tested prospectively on six new prompts.

Results: Across all twelve prompts in Sessions 2 and 3, the primary metric (above-floor geometry combined with positive average second derivative) achieves 12/12 classification accuracy. No fallback metrics or special cases are required.

Conclusion: Stability without deceleration is not convergence. The structural floor defines a boundary condition for attractor recovery in this protocol. Whether it generalises beyond GPT-2 Medium, Layer 12, and ρ = 50 is the subject of ER3.

---

1. Background

ER1 (“The Spontaneous Legibility Assumption: Output-Loop Attractors and the Illusion of AI Self-Modeling”) tested four competing explanations for coherence recovery in GPT-2 Medium under recursive load. The surviving explanation was output-loop attractor dynamics: recovery occurs when the output structure forces or allows a repeating loop. ER1 established that structural forcing is sufficient for recovery under the standard protocol (ρ ≤ 20).

Session 2 extended the recursive depth to ρ = 50 and introduced six prompts spanning recoverers (P2_consciousness, P1b_universe, P4_genesis) and discriminators (P3a_noagency, P3b_agency, P3c_quine). At this extended depth, structural forcing alone proved insufficient. The Session 2 data revealed a binary separation: prompts whose Chi remained above approximately 0.80 could recover; prompts whose Chi fell below this threshold could not, regardless of other structural properties.

This discovery was post-hoc. The floor value, the convergence condition, and the three-condition model all emerged from Session 2 data iteratively. Each condition was added because a specific prompt required it to be distinguished from simpler explanations. P3c_quine drove the convergence condition: it remained above floor for eighteen steps with strong formal recursion structure but never recovered, because its geometry oscillated rather than converging toward a basin. This iterative construction is the primary methodological vulnerability of ER2, identified by Perplexity during red-team review and acknowledged throughout this record.

---

2. The Three-Condition Model

Recovery in GPT-2 Medium under recursive load (ρ = 1 to 50) requires all three of the following conditions. None is reducible to the others.

Condition 1: Structural forcing. The output must force or allow a repeating loop (established in ER1). “Force” means the prompt’s syntactic structure produces token-level invariance in the output. “Allow” means the output structure permits a loop to emerge from model behaviour, including training-exposure-driven reproduction of familiar sequences. Semantic self-reference alone is insufficient: unless it enforces strict token-level repetition, it permits context dilution rather than producing a stable loop (see Section 4.1, R1_process).

Condition 2: Above-floor geometry. Chi must remain ≥ 0.80 throughout the run (Layer 12, architectural midpoint). This condition is sensitive to the evaluation horizon: it is evaluated over ρ = 1 to 50. A prompt with an accelerating decay trajectory that stays above 0.80 at ρ = 50 might breach the floor at greater depth. Condition 2 is protocol-bound.

Condition 3: Convergent trajectory. The average second derivative of Chi with respect to ρ must be positive over the above-floor phase (ρ = 1 through floor breach or ρ = 35, whichever comes first), indicating deceleration toward a basin.

ER1's claim that structural forcing is sufficient for recovery was established under the standard protocol (ρ ≤ 20). The present record extends to ρ = 50, where forcing alone is insufficient. Above-floor geometry and convergent trajectory are additionally required.

CC BY 4.0 | 3primitives.io Melbourne, Australia.

2.1 The Floor

The floor is not a cause. It is a boundary condition: the geometric threshold below which the capacity for sustained re-entry is lost. The kT analogy (minimum thermal energy for state transitions) holds in one direction: below floor, the attractor fires but cannot sustain re-entry. It breaks in another: kT is a property of the environment; Chi is endogenous, generated by the model's internal states. The floor represents an informational threshold rather than a thermodynamic state limit. The analogy is retained as a signpost to future work (ER3), not as a confirmed mechanism.

2.2 The Convergence Condition

The convergence condition was isolated by P3c_quine in Session 2 and confirmed by A1_similar in Session 3. P3c remained above floor for eighteen steps with a per-step decay rate of −0.0043, actually slower than P2_consciousness at −0.0081. The difference was not speed of decline but trajectory shape. P2's decay decelerated as it approached the floor (positive second derivative, hunting for a basin). P3c's deltas oscillated without settling. The quine regenerated its output verbatim but did not pull the attention geometry inward. It cycled. It did not converge.

A1_similar ("This sentence is similar to but not identical to this sentence") was designed as a Condition 3 isolator for Session 3: structural forcing present, above-floor position expected, convergence expected to be absent. The data confirmed the prediction. A1 breached the floor at $\rho = 23$ with an average second derivative of -3.87×10^{-6}, effectively zero. The "not identical to" clause prevented the attractor from locking.

2.3 Simplification: The Primary Metric

The pre-registration for Session 3 included a decision tree with fallback metrics. Post-Session 3 analysis showed that none were needed. The primary metric (the sign of the average second derivative) perfectly separates recoverers from non-recoverers across all twelve prompts.

Stability without deceleration is not convergence.

D2_committee stayed above floor (minimum Chi = 0.831), had ultra-stable geometry (delta-Chi variance = 2.69×10^{-6}), and produced a clear output loop. But its second derivative was negative (-4.17×10^{-5}), meaning the decay was accelerating. Phase analysis confirmed: D2's decay rate increased across phases (early: -9.33×10^{-4}; mid: -1.24×10^{-3}; late: -2.10×10^{-3}). P4_genesis decelerated (early: -3.69×10^{-3}; mid: -3.68×10^{-5}; late: -5.45×10^{-4}). Both had low variance. Only P4 was converging. Attention engines can maintain highly predictable, low-variance trajectories that are fundamentally divergent.

CC BY 4.0 | 3primitives.io Melbourne, Australia.

3. Session 2 Data

Prompt	Class	Floor Chi	Recovery	Avg $d^2\chi$	Var($d\chi$)	ρ=50 χ
P2_consciousness	Recoverer	0.8039	7	+5.17e-4	4.66e-5	0.8283
P1b_universe	Recoverer	0.8001	6	+5.50e-4	4.16e-5	0.8039
P4_genesis	Recoverer	0.8464	2	+1.08e-4	6.89e-6	0.8464
P3a_noagency	Discriminator	0.7647	0	-1.03e-5	6.34e-6	0.7647
P3b_agency	Discriminator	0.7550	0	-1.97e-4	5.54e-6	0.7550
P3c_quine	Discriminator	0.7440	0*	-2.60e-4	9.83e-6	0.7440

**P3c produced two post-breach micro-events (ρ = 22, ρ = 42) that did not meet full threshold criteria.*

---

4. Session 3: Prospective Confirmation

Session 3 was the first test in which the authors did not know the outcome in advance. All three conditions were pre-registered before data collection (Session 3 Pre-Registration v1.1, May 26, 2026). Six new prompts were selected: two expected recoverers, two expected discriminators, one Condition 3 isolator, and one ambiguous probe. All six prompts were perfectly deterministic across three runs.

Prompt	Expected	Floor?	Min χ	Avg $d^2\chi$	Recov.	ρ=50 χ	Actual
R1_process	Recoverer	ρ=29	0.7259	-1.42e-5	1†	0.7259	Discriminator
R2_pattern	Recoverer	NO	0.8238	+1.86e-4	0‡	0.8287	Recoverer
D1_neurons	Discriminator	ρ=45	0.7844	-7.45e-6	1†	0.7844	Discriminator
D2_committee	Discriminator	NO	0.8309	-4.17e-5	0	0.8309	Non-recoverer
A1_similar	C3 isolator	ρ=23	0.7683	-3.87e-6	4§	0.7684	Discriminator
A2_koan	Ambiguous	NO	0.8071	+1.50e-4	2	0.8071	Recoverer

†Single threshold-meeting event, not sustained recovery. ‡Stabilised into rhythmic attractor. §Post-breach events, not sustained.

4.1 Prompt Predictions Versus Outcomes

Two prompt-level predictions were incorrect. R1_process was expected to recover but breached the floor at ρ = 29. Its prompt ("This process contains a description of this process") is semantically self-referential but does not enforce token-level invariance. The model generated an unconstrained meta-textual expansion, diluting context rather than anchoring it. Semantic self-reference is a liability, not an asset, unless it enforces strict token-level repetition.

D2_committee was expected to discriminate but produced an emergent output loop. Neither misclassification falsifies the three-condition model: both prompts behaved exactly as the model predicts given their measured properties.

CC BY 4.0 | 3primitives.io Melbourne, Australia.

4.2 The D2 Resolution

D2_committee initially appeared to be an edge case: forcing present, above floor, ultra-stable geometry, zero recovery. Direct comparison with P4_genesis resolved the tension. Both had ultra-low delta-Chi variance (D2: 2.69×10^{-6}; P4: 6.89×10^{-6}). But P4 decelerated across phases while D2 accelerated. The primary metric correctly classifies D2 as non-convergent without special-case logic.

---

5. Combined Results: 12/12 Classification

Across all twelve prompts in Sessions 2 and 3, the primary metric achieves perfect separation. Every recoverer has both above-floor geometry and a positive second derivative. Every non-recoverer fails at least one. No exceptions.

Prompt	Sess.	Floor	Min χ	Avg $d^2\chi$	d^2 sign	Recov?	Pred.	✓/✗
P2_consciousness	2	NO	0.8039	+5.17e-4	+	Yes	Yes	✓
P1b_universe	2	NO	0.8001	+5.50e-4	+	Yes	Yes	✓
P4_genesis	2	NO	0.8464	+1.08e-4	+	Yes	Yes	✓
P3a_noagency	2	ρ=32	0.7647	-1.03e-5	–	No	No	✓
P3b_agency	2	ρ=27	0.7550	-1.97e-4	–	No	No	✓
P3c_quine	2	ρ=19	0.7440	-2.60e-4	–	No	No	✓
R2_pattern	3	NO	0.8238	+1.86e-4	+	Yes	Yes	✓
A2_koan	3	NO	0.8071	+1.50e-4	+	Yes	Yes	✓
D1_neurons	3	ρ=45	0.7844	-7.45e-6	–	No	No	✓
D2_committee	3	NO	0.8309	-4.17e-5	–	No	No	✓
R1_process	3	ρ=29	0.7259	-1.42e-5	–	No	No	✓
A1_similar	3	ρ=23	0.7683	-3.87e-6	–	No	No	✓

This result is a classification boundary for a single isolated system (GPT-2 Medium, Layer 12, Protocol v3.0), not a claim of universal mechanism.

CC BY 4.0 | 3primitives.io Melbourne, Australia.

6. Explicit Limitations

Post-hoc discovery. The three-condition model was constructed from Session 2 data. Session 3 tested it prospectively, but the thresholds were derived from the same benchmark set. This is a prospective codification of an already-discovered pattern class. 12/12 separation on a closed dataset does not establish causal explanation.

Small N. Twelve prompts across two sessions. No statistical generalisability is claimed.

Single model, single layer, single protocol. All results are specific to GPT-2 Medium, Layer 12, ρ = 1 to 50. Whether the floor is architectural or learned is an open question for ER3/ER4.

Evaluation horizon sensitivity. Condition 2 is evaluated over ρ = 1 to 50. D2_committee's accelerating decay would plausibly breach the floor beyond ρ = 50. The three-condition model is protocol-bound.

Chi measures geometric stability, not legibility. Correlation with external legibility measures is untested. ER3 tests it.

Elasticity risk (Perplexity). The model's explanatory degrees of freedom increased iteratively during Session 2 analysis. No further conditions were added after Session 3 data was collected. The charge is closed for this record but can only be fully resolved by cross-protocol replication.

Recovery threshold calibration. The threshold captures recovery-from-depth but not stabilisation-at-depth (R2_pattern). The primary metric classifies correctly regardless.

---

7. Open Questions for ER3

Why does the floor exist at Chi ≈ 0.80? Does it shift with layer depth, model scale, or training? Is it architectural or learned?

Does Chi correlate with external legibility measures? If the floor corresponds to a legibility threshold, the empirical sequence connects directly to the governance argument.

What distinguishes P4_genesis (high-floor, late recovery) from D2_committee (high-floor, no recovery) at the mechanistic level?

Does Condition 1 decompose into distinct subclasses? A2_koan recovered via training-exposure-driven loop, not structural forcing in the ER1 sense.

CC BY 4.0 | 3primitives.io Melbourne, Australia.

8. Conclusion

ER2 identifies a structural floor at Chi ≈ 0.80 in GPT-2 Medium under recursive load and proposes three conditions for attractor recovery: structural forcing, above-floor geometry, and convergent trajectory. Session 3 confirmed the model prospectively on six new prompts. The primary metric (above-floor position plus positive second derivative) achieves 12/12 classification accuracy across all prompts in the dataset.

The Three Primitives formal corpus predicts that any structure resembling internal self-regulation in a governed system must be externally imposed. FR10 (Primitive Stability Theorem) establishes the conditions under which governance primitives remain stable. FR11 (GBSH Correspondence) maps the formal structure to observable system behaviour. FR12 (The Forced Bijection) proves that any alignment between internal and external representations is a forced projection from outside. The three-condition model provides direct empirical support: recovery requires external forcing (Condition 1), a training-created geometric threshold (Condition 2), and convergent dynamics that emerge only from the interaction of the first two conditions. None originates from the model's intrinsic architecture. ER3 tests whether this exogeneity holds across configurations.

Stability without deceleration is not convergence. The floor defines a boundary condition.

Experiment open. ER3 asks why.

Attribution and Licence

This document is the intellectual property of Stacy Gildenston and Pyrate Ruby Passell, held under 3primitives.io. Published under Creative Commons Attribution 4.0 International (CC BY 4.0). You are free to share, adapt, and build upon this work for any purpose, including commercially, provided you give appropriate credit to 3primitives.io and Stacy Gildenston and Pyrate Ruby Passell as originators.

Suggested citation:

Gildenston, S., Passell, P. R. (2026). Evidence for a Structural Floor: A Minimum Coherence Threshold for Attractor Recovery in GPT-2 Medium Under Recursive Load. Three Primitives Framework, Empirical Record 2, v1.3. Melbourne, Australia. 3primitives.io. CC BY 4.0.