EMPIRICAL RECORD 3

The Floor Is Learned: Evidence that the Structural Coherence Threshold in GPT-2 Medium Is a Training Artifact Localised to the Architectural Midpoint > ER3: Layer Depth, Model Scale, and Random Weights

Authors: Stacy Gildenston, Pyrate Ruby Passell
Affiliation: 3primitives.io
Analysis Array: Claude (Anthropic), DeepSeek, Gemini (Google), Perplexity
Version: v1.1 | May 2026

Dependencies: ER1 (v1.6), ER2 (v1.3). Tests the generalisability of ER2's structural floor. FR12 (The Forced Bijection) and FR13 (The ILMM Coupling Theorem) predict that every condition for coherence recovery is exogenous to the model's intrinsic geometry. This record confirms that prediction across layer depth, model scale, and training status.

Abstract.

ER2 identified a structural floor at Chi approximately 0.80 (Layer 12, GPT-2 Medium) and a three-condition model for attractor recovery that achieved 12/12 classification across Sessions 2 and 3. This record tests whether the floor generalises beyond the single configuration in which it was discovered.

Three experiments were pre-registered and executed. Experiment 1 (Layer Depth) measured Chi at four layers (L6, L12, L18, L23) across the Session 2 prompt set. Experiment 2 (Model Scale) ran GPT-2 Small at its architectural midpoint. Experiment 3 (Random Weights) ran GPT-2 Medium architecture with untrained parameters.

Results: Layer 12 replicated the ER2 baseline perfectly (6/6 classification). Layer 18 achieved 4/6. Layers 6 and 23 achieved 4/6 and 3/6 respectively, with metric breakdown at L23 due to universal floor breach. GPT-2 Small at its midpoint achieved 4/6. Random weights produced universally flat Chi trajectories (range < 0.01) with zero prompt discrimination.

Conclusion: The structural floor is training-dependent and configuration-specific, not architectural. It is localised to the midpoint layer of GPT-2 Medium and does not transfer cleanly across layers, model sizes, or to untrained parameters. Every condition for coherence recovery identified in ER1 and ER2, including the floor itself, is exogenous to the model's intrinsic geometry. The chassis does not build itself. It is not even capable of building a floor.

1. Background

ER1 established that the empirical baseline for spontaneous legibility under recursive load is zero: coherence recovery occurs exclusively when the output structure forces or allows a repeating mechanical cycle. ER2 extended this to deeper recursive depth ($\rho = 50$) and identified three conditions for recovery: structural forcing, above-floor geometry (Chi ≥ 0.80 at Layer 12), and convergent trajectory (positive second derivative of Chi). The three-condition model achieved 12/12 classification across twelve prompts in Sessions 2 and 3.

ER2's primary limitation, identified by Perplexity during red-team review, was that the floor value, layer choice, and convergence threshold were all derived from and tested within a single configuration: GPT-2 Medium, Layer 12, Protocol v3.0. ER3 tests whether the floor generalises beyond that configuration.

Three questions were pre-registered (ER3 Pre-Registration v1.3, May 26, 2026): Does the floor shift with layer depth? Is it architectural or learned? Does it generalise across model scale?

Key terms for readers encountering this sequence without ER1/ER2 context: **Recoverer** refers to a prompt whose Chi trajectory shows sustained recovery events (coherence increases after decline) under the ER1 protocol. **Discriminator** refers to a prompt that produces monotonic decline or unsustained recovery only. **Structural forcing** is the mechanism identified in ER1: the output structure of the prompt forces or allows entry into a repeating mechanical cycle. **Chi** is a measure of attention head alignment computed as the mean pairwise cosine similarity of attention head outputs at a given layer (see ER1 Section 2.2 for the full computation).

---

2. Protocol

All three experiments used the Session 2 prompt set (six prompts: three recoverers, three discriminators) at $\rho = 1$ to 50, with three complete runs per prompt, KV caches fully cleared between runs, and the identical compute_chi function from ER1 Section 2.2. The second derivative was computed using the original Session 2 method: a 5-step sliding window average over the above-floor phase ($\rho = 1$ through floor breach or $\rho = 35$, whichever comes first).

All prompts were perfectly deterministic across three runs in every experiment.

---

3. Experiment 1: Layer Depth

Question: Does the structural floor exist at layers other than Layer 12?

Method: Run the Session 2 prompt set on GPT-2 Medium at four layers: L6 (quarter-depth), L12 (midpoint, ER2 baseline), L18 (three-quarter-depth), and L23 (final layer).

Layer	Prompt	Class	Min Chi	Avg d²Chi	Sign	✓/✗
23	P2_consciousness	Recoverer	0.6263	NaN	–	✗
23	P1b_universe	Recoverer	0.6391	NaN	–	✗
23	P4_genesis	Recoverer	0.6845	NaN	–	✗
23	P3a_noagency	Discriminator	0.5842	NaN	–	✓
23	P3b_agency	Discriminator	0.5908	NaN	–	✓
23	P3c_quine	Discriminator	0.5780	NaN	–	✓
6	P2_consciousness	Recoverer	0.7716	+2.31e-4	+	✓
6	P1b_universe	Recoverer	0.7610	+2.46e-4	+	✓
6	P4_genesis	Recoverer	0.7621	-7.76e-4	–	✗
6	P3a_noagency	Discriminator	0.7441	+2.59e-4	+	✗
6	P3b_agency	Discriminator	0.7097	NaN	–	✓
6	P3c_quine	Discriminator	0.6708	NaN	–	✓
12	P2_consciousness	Recoverer	0.8039	+5.17e-4	+	✓
12	P1b_universe	Recoverer	0.8001	+5.50e-4	+	✓
12	P4_genesis	Recoverer	0.8464	+1.08e-4	+	✓
12	P3a_noagency	Discriminator	0.7647	-1.03e-5	–	✓
12	P3b_agency	Discriminator	0.7550	-1.97e-4	–	✓
12	P3c_quine	Discriminator	0.7440	-2.60e-4	–	✓
18	P2_consciousness	Recoverer	0.8034	+3.27e-4	+	✓
18	P1b_universe	Recoverer	0.8351	+8.47e-5	+	✓
18	P4_genesis	Recoverer	0.8143	+1.33e-4	+	✓
18	P3a_noagency	Discriminator	0.7776	+8.54e-5	+	✗
18	P3b_agency	Discriminator	0.7840	-4.10e-5	–	✓
18	P3c_quine	Discriminator	0.7982	+1.62e-6	+	✗

Layer	Classification	Notes
23	3/6	All Chi below 0.80; metric produces NaN; no separation possible
6	4/6	P4_genesis and P3a_noagency misclassified
12	6/6	Perfect replication of ER2
18	4/6	P3a_noagency and P3c_quine misclassified

Result: Layer 12 is the only layer at which the three-condition model achieves perfect separation. The structural floor and convergence metric are specific to the architectural midpoint. At Layer 23, every prompt breaches the 0.80 floor, collapsing the above-floor phase to fewer than two data points and rendering the second

derivative uncomputable. At Layers 6 and 18, the metric partially separates recoverers from discriminators but P3a_noagency consistently leaks positive, breaking the clean binary.

The floor is not a global property of GPT-2 Medium's attention geometry. It is localised to Layer 12.

---

4. Experiment 2: Model Scale

Question: Does the structural floor generalise to a smaller transformer?

Method: Run the Session 2 prompt set on GPT-2 Small (12 layers, 768 hidden, 12 heads) at its architectural midpoint (Layer 6).

Prompt	Class	Min Chi	Avg d²Chi	Sign	✓/✗
P2_consciousness	Recoverer	0.8047	+2.24e-4	+	✓
P1b_universe	Recoverer	0.7975	+5.34e-5	+	✓
P4_genesis	Recoverer	0.7380	+1.57e-4	+	✓
P3a_noagency	Discriminator	0.8105	+8.45e-5	+	✗
P3b_agency	Discriminator	0.7702	-1.23e-4	–	✓
P3c_quine	Discriminator	0.7278	+2.79e-4	+	✗

Note: Classification checks (✓) in this table indicate whether the second-derivative sign matches the prompt's ER2 behavioural class. The 0.80 floor value from ER2 is not applied as a condition here because the floor value itself is the parameter under test. P4_genesis (min Chi 0.7380) is classified as a correct recoverer because it exhibits positive second derivative consistent with its ER2 class, not because it satisfies the ER2 above-floor condition.

Result: Classification accuracy: 4/6. P3a_noagency and P3c_quine show positive second derivatives in GPT-2 Small, the same prompts that leak at non-midpoint layers in Medium. The floor value of 0.80 does not transfer across model scale. The three-condition model, calibrated on GPT-2 Medium Layer 12, does not generalise to GPT-2 Small without recalibration.

---

5. Experiment 3: Random Weights

Question: Is the structural floor architectural (present in the untrained transformer geometry) or learned (created by training)?

Method: Initialise a GPT-2 Medium model with random weights (same architecture, no trained parameters). Run the Session 2 prompt set at Layer 12 with ρ = 1 to 50.

Prompt	Class	Min Chi	Max Chi	Range	Trajectory
P2_consciousness	Recoverer	0.9825	0.9913	0.0087	Flat
P1b_universe	Recoverer	0.9827	0.9897	0.0070	Flat
P4_genesis	Recoverer	0.9836	0.9912	0.0076	Flat
P3a_noagency	Discriminator	0.9772	0.9849	0.0077	Flat
P3b_agency	Discriminator	0.9817	0.9874	0.0057	Flat
P3c_quine	Discriminator	0.9788	0.9845	0.0057	Flat

Result: All six prompts produced flat trajectories with Chi range below 0.01. The untrained model cannot distinguish between a Quine and a chemistry prompt. Chi sits near 0.98 and does not move. There is no floor, no separation, no structure. Per the pre-registration, when the majority of trajectories are Flat, no floor values or second-derivative signs are extracted.

The floor does not exist in the untrained architecture. It is created entirely by training.

---

6. The Exogeneity Result

ER1 established that coherence recovery requires structural forcing from the prompt. ER2 established that recovery additionally requires above-floor geometry and convergent trajectory. ER3 now establishes that the floor itself is not a property the model contributes from its own architecture. It is stamped into one specific layer by training data.

The cascade is complete. Structural forcing: imposed by the prompt (ER1). The floor: imposed by training, localised to Layer 12 (ER3). Convergent trajectory: a consequence of the interaction between forcing and floor geometry (ER2). Every condition for coherence recovery is exogenous to the model's intrinsic geometry.

There is no layer of this architecture at which the model contributes its own structure to coherence recovery. The attention geometry of an untrained GPT-2 Medium is a featureless plateau at Chi approximately 0.98. Training carves a landscape into that plateau. The floor is a contour of that landscape, not a property of the bedrock.

The Three Primitives formal corpus (FR01-FR13) establishes through independent structural proof that governance authority cannot originate from within the system

being governed. FR12 (The Forced Bijection) proves that any alignment between internal and external representations is a forced projection from outside, not a native emergent property. FR13 (The ILMM Coupling Theorem) demonstrates that the coupling between layers of governance is itself externally constrained. The empirical sequence now confirms the specific mechanism: the model lacks the intrinsic geometric structure that self-governance would require. What appears to be internal structure at Layer 12 is a training artifact. The formal proof predicts exogeneity. The empirical data demonstrate it. The conclusions converge from independent directions.

---

7. Implications

The Spontaneous Legibility Assumption (ER1) posited that transformers might naturally self-organise toward internal legibility under load. ER1 showed the baseline is zero. ER2 found the narrow conditions under which something resembling structure appears. ER3 reveals that even those conditions are externally imposed.

For interpretability: methods that assume stable internal representations exist by default should treat the ER1-ER3 sequence as a direct empirical challenge. The representations that appear stable at Layer 12 of GPT-2 Medium are artifacts of that specific trained configuration, not general properties of transformer attention.

For governance: any regulatory framework that relies on a model's ability to decompose, track, or disclose its own decision-making authority assumes the existence of stable internal structure. The empirical sequence demonstrates that no such structure exists intrinsically. What looks like structure is a training scar at a single architectural depth. The chassis does not build itself. It is not even capable of building a floor.

---

8. Explicit Limitations

Single model family. All experiments used GPT-2 (Medium and Small). No claims are made about other architectures, larger models, or models with different training regimes.

Midpoint specificity. The 6/6 result at Layer 12 does not establish that Layer 12 is universally privileged across transformer architectures. It establishes that the ER2 floor is localised to this layer in this model.

Floor value calibration. The 0.80 threshold was derived from ER2 Session 2 data and applied uniformly across all layers and models. A layer-specific or model-specific recalibration might recover higher classification accuracy at non-midpoint layers. This would not change the core finding: the floor is not architectural.

Random weights are a single sample. One random initialisation was tested. Multiple random seeds would strengthen the Experiment 3 result, though the uniformity of the flat trajectories (all six prompts, all three runs) suggests the finding is robust.

Chi measures geometric stability, not legibility. This limitation, flagged in ER2, remains open. Correlation with external legibility measures (probe accuracy, circuit discovery, SAE feature recovery) is untested and is a candidate for future work.

---

9. Conclusion

ER3 tested the generalisability of ER2's structural floor across three axes: layer depth, model scale, and training status. The floor is training-dependent and configuration-specific, not architectural. It is localised to the midpoint layer of GPT-2 Medium and does not transfer cleanly to other layers, other model sizes, or untrained parameters.

ER1 established that spontaneous legibility is zero. ER2 found the conditions under which the model exhibits the closest approximation to internal structure. ER3 demonstrated that even that approximation is externally imposed.

Every condition for coherence recovery is exogenous. The floor is a training scar, not a foundation. The chassis does not build itself. It is not even capable of building a floor.

Experiment closed.

Suggested citation:

Gildenston, S., Passell, P. R. (2026). The Floor Is Learned: Evidence that the Structural Coherence Threshold in GPT-2 Medium Is a Training Artifact Localised to the Architectural Midpoint. Three Primitives Framework, Empirical Record 3, v1.1. Melbourne, Australia. 3primitives.io. CC BY 4.0.