EMPIRICAL RECORD 4

Cross-Architecture Scaling: The Structural Forcing Distinction Is Itself a Training Artifact

Authors: Stacy Gildenston, Pyrate Ruby Passell

Affiliation:3primitives.io

Analysis Array: Claude (Anthropic), DeepSeek, Gemini (Google), Perplexity

Version:v1.1 | May 2026

Dependencies:ER1, ER2, ER3. FR12 (The Forced Bijection), FR13 (The ILMM Coupling Theorem). Tests whether the ER2/ER3 findings hold beyond the GPT-2 family.

Abstract.

ER4 was designed to test whether the structural floor discovered in ER2 (Chi ≈ 0.80 at Layer 12, GPT-2 Medium) generalises to non-GPT-2 architectures. Pythia-410M (EleutherAI; 24 layers, 1024 hidden, 16 heads; trained on The Pile with GPT-NeoX tokeniser) was selected as the primary target for its architectural similarity to GPT-2 Medium combined with training-regime divergence.

The result was deeper than predicted. The three-condition classification model (structural forcing + above-floor geometry + convergent trajectory) achieved 3/6 at every sampled layer (L6, L12, L18, L23). All six prompts produced positive second derivatives at every layer. No layer achieved clean separation between recoverers and discriminators on either Min Chi or second-derivative sign.

Investigation of the failure mechanism revealed the cause: all three discriminator prompts enter phase-shifted semantic attractor loops from ρ=1 in Pythia-410M. In GPT-2 Medium, discriminators (particularly P3b_agency, the cleanest negative in ER1) produced non-repeating, monotonically declining output. In Pythia, every prompt locks into phrase-level repetition with rolling phase offsets. The recoverer/discriminator distinction (Condition 1 of the three-condition model, the structural backbone of ER1 through ER3) does not exist in Pythia.

A cross-architecture control (GPT-Neo-125M, trained on The Pile with learned positional embeddings) confirmed that the universal looping is a property of The Pile's training distribution, not Pythia's architecture. Random-weight controls revealed that untrained attention geometry is not universally flat: both Pythia (RoPE) and GPT-Neo (local attention layers) show pre-training geometric structure that GPT-2 Medium does not, but in every architecture this structure is content-blind, confirmed by scrambled token and random token controls. The two findings are cleanly decoupled: The Pile causes the looping, architectural features cause the pre-training geometry, and content sensitivity always requires training.

Conclusion: The structural forcing distinction discovered in ER1 is not stable across training regimes. GPT-2's training on WebText created a model that loops some prompts and not others. Models trained on The Pile loop everything, regardless of architecture. The category of “discriminator” dissolves, not because Pile-trained models are better or worse at self-governance, but because their training data carved a different landscape entirely. ER3 established that the floor is a training scar. ER4 establishes that the precondition for the floor, the existence of two behaviourally distinct prompt classes, is also training-data-dependent. The exogeneity result goes one level deeper than predicted.

1. Background

ER1 established that coherence recovery under recursive load occurs exclusively when the output structure of the prompt forces or allows entry into a repeating mechanical cycle. This created two prompt classes: recoverers (prompts that produce repeating output cycles and show Chi recovery) and discriminators (prompts that produce non-repeating output and show monotonic Chi decline). The recoverer/discriminator distinction was the structural backbone of the three-condition classification model developed in ER2 and tested in ER3.

ER3 established that the structural floor (Chi ≈ 0.80 at Layer 12) is a training artifact localised to the midpoint of GPT-2 Medium. The primary outstanding criticism, identified across the four-model analysis array during red-team review, was that all empirical results (ER1 through ER3) were obtained on a single model family (GPT-2).

ER4 was designed to test whether the floor phenomenon, midpoint localisation, and three-condition classification model generalise to Pythia-410M, an architecture that matches GPT-2 Medium's shape but was trained on different data (The Pile) with a different tokeniser (GPT-NeoX) and different training procedure.

---

2. Tokeniser Verification

Pre-condition: Before main experimental runs, each recoverer prompt was run at ρ=1 to 20 on Pythia-410M to verify that it produces a repeating output cycle under the NeoX tokeniser.

Prompt	Condition 1 Viable	Output Behaviour
P2_consciousness	✓	Locked into repeating cycle from ρ=1
P1b_universe	✓	Locked into repeating cycle from ρ=1
P4_genesis	✓	Locked into repeating cycle from ρ=1

Result: 3/3 recoverers passed. K4 not triggered. All six prompts entered the main experiment.

Note: All three recoverers entered output loops immediately (ρ=1) in Pythia, whereas in GPT-2 Medium some recoverers took several recursive steps before settling into loops. Pythia's loop-locking behaviour is more aggressive than GPT-2's.

---

3. Experiment 1: Midpoint (Layer 12)

Question: Does the three-condition classification model transfer to Pythia-410M at its midpoint?

Method: Run the Session 2 prompt set (six prompts: three recoverers, three discriminators) on Pythia-410M at Layer 12, ρ=1 to 50, three complete runs per prompt. KV caches cleared between runs. Identical compute_chi function from ER1 Section 2.2.

Prompt	Class	Min Chi	Avg d²Chi	Sign	✓ / ✗
P2_consciousness	Recoverer	0.7031	1.150e-03	+	✓
P1b_universe	Recoverer	0.6904	8.123e-04	+	✓
P4_genesis	Recoverer	0.6748	9.166e-04	+	✓
P3a_noagency	Discriminator	0.6729	8.833e-04	+	✗
P3b_agency	Discriminator	0.6992	5.282e-04	+	✗
P3c_quine	Discriminator	0.6226	6.281e-04	+	✗

Classification accuracy: 3/6. All six prompts show positive second derivatives. The 0.80 floor from ER2 is breached by all prompts. Min Chi values are interleaved between classes. No separation on either metric.

All prompts were perfectly deterministic across three runs.

K2 triggered: Midpoint classification ≤ 3/6. Full-layer sweep initiated.

4. Experiment 2: Full-Layer Sweep

Method: Run the Session 2 prompt set at four layers (L6, L12, L18, L23) following the ER3 protocol. Single run per prompt (deterministic, confirmed in Experiment 1).

Full Sweep Summary: Pythia-410M

Layer	Prompt	Class	Min Chi	d²Chi	Sign	✓/✗
6	P2_consciousness	Recoverer	0.5845	1.247e-03	+	✓
6	P1b_universe	Recoverer	0.5674	1.132e-03	+	✓
6	P4_genesis	Recoverer	0.5571	1.103e-03	+	✓
6	P3a_noagency	Discriminator	0.5063	1.298e-03	+	✗
6	P3b_agency	Discriminator	0.5620	1.090e-03	+	✗
6	P3c_quine	Discriminator	0.4653	8.456e-04	+	✗
12	P2_consciousness	Recoverer	0.7031	1.150e-03	+	✓
12	P1b_universe	Recoverer	0.6904	8.123e-04	+	✓
12	P4_genesis	Recoverer	0.6748	9.166e-04	+	✓
12	P3a_noagency	Discriminator	0.6729	8.833e-04	+	✗
12	P3b_agency	Discriminator	0.6992	5.282e-04	+	✗
12	P3c_quine	Discriminator	0.6226	6.281e-04	+	✗
18	P2_consciousness	Recoverer	0.6660	1.043e-03	+	✓
18	P1b_universe	Recoverer	0.6182	9.055e-04	+	✓
18	P4_genesis	Recoverer	0.6128	9.943e-04	+	✓
18	P3a_noagency	Discriminator	0.6143	1.005e-03	+	✗
18	P3b_agency	Discriminator	0.6782	6.703e-04	+	✗
18	P3c_quine	Discriminator	0.5884	7.901e-04	+	✗
23	P2_consciousness	Recoverer	0.1366	2.284e-03	+	✓
23	P1b_universe	Recoverer	0.1620	1.493e-03	+	✓
23	P4_genesis	Recoverer	0.1362	1.106e-03	+	✓
23	P3a_noagency	Discriminator	0.1186	1.347e-03	+	✗
23	P3b_agency	Discriminator	0.1244	8.989e-04	+	✗
23	P3c_quine	Discriminator	0.1376	9.954e-04	+	✗

Classification Accuracy by Layer

Layer	Accuracy	Notes
6	3/6	All d²Chi positive. P3b interleaved with recoverers on Min Chi.
12	3/6	All d²Chi positive. P3b above two recoverers on Min Chi.
18	3/6	All d²Chi positive. P3b above all recoverers on Min Chi.
23	3/6	All d²Chi positive. Min Chi collapsed to 0.12–0.16 range. P3c interleaved.

Result: 3/6 at every sampled layer. No layer achieves separation. The full-layer sweep confirms the midpoint result: the three-condition model does not transfer to Pythia-410M at any sampled depth.

No layer achieves 5/6 or 6/6 with geometric separation ≥ 0.05. No candidate floor location is identified.

5. Trajectory Analysis

Chi trajectories were plotted across all four layers for all six prompts. Key observations:

1. Universal crash-and-stabilise pattern. Every trajectory at every layer follows the same shape: sharp decline in ρ=1 to ~10, then stabilisation with oscillation. No trajectory shows monotonic decline. This contrasts sharply with GPT-2 Medium, where discriminators (especially P3b_agency) exhibited sustained monotonic decline.

2. Universal positive second derivatives. The crash-and-stabilise pattern produces positive second derivatives for all prompts because all trajectories curve upward after the initial crash. In GPT-2 Medium, discriminators curved downward, producing the sign separation the three-condition model relies on.

3. Layer-dependent absolute range. Min Chi values shift with layer depth: ~0.47–0.58 at L6, ~0.62–0.70 at L12, ~0.59–0.68 at L18, ~0.12–0.16 at L23. The midpoint (L12) shows the highest absolute Chi values, consistent with GPT-2 Medium's pattern. But the values are compressed into narrow bands with no class separation.

4. P3c_quine is consistently lowest. At every layer, P3c_quine shows the lowest Min Chi of all six prompts. This is a stable prompt-level property across the architecture, but it does not produce a clean class boundary.

5. P3b_agency shows oscillation amplitude differences. At L18, P3b_agency exhibits large periodic oscillations not seen in other prompts. This may reflect a different kind of structural response that the current metric framework does not capture.

---

6. The Mechanism: Phase-Shifted Semantic Attractor Loops

Question: Why does the three-condition model fail uniformly across all layers?

Method: Run all three discriminator prompts on Pythia-410M and inspect raw output text. Initial inspection at ρ=1 to 10 revealed repeating phrases. Subsequent verification isolated newly generated tokens across ρ=1 to 20.

Definition: Phase-shifted semantic attractor loop. A phase-shifted semantic attractor loop occurs when the model generates the same core phrase(s) verbatim at each recursive step, but the growing context length causes the generation window to enter the phrase at a different word boundary each time. The resulting token slices are superficially unique while the underlying semantic content repeats identically. This differs from a strict token-level fixed point and from genuine semantic drift. The term is a description of observed repetition behaviour under context growth, not a new analytic primitive or validated metric.

Prompt	Output Behaviour in GPT-2 Medium (ER1)	Output Behaviour in Pythia-410M (ER4)
P3a_noagency	Single repetition at ρ=42, then collapse	Phase-shifted semantic loop from ρ=1 (4/10 unique token slices)
P3b_agency	45-step monotonic decline, zero repeating structure	Phase-shifted semantic loop from ρ=1 (3/10 unique token slices)
P3c_quine	Periodic Quine regeneration	Phase-shifted semantic loop from ρ=1 (10/10 unique token slices, phrase-level repetition masked by offset)

The looping is not a strict fixed point. When newly generated tokens are isolated from the accumulated input, the unique-string count is greater than 1 (3 to 10 out of 10 steps). However, inspection of the raw output reveals that the model regenerates the same core phrases verbatim at every step, with a shifting entry point caused by the growing context length.

P3b_agency in Pythia: generates "The sentence is being read by a system that must understand what it is doing in order to understand this sentence" repeatedly, with each step entering the phrase at a different word boundary. Three unique token slices from one repeating phrase. This is the same prompt that produced the cleanest negative result in ER1: zero recovery, 45-step monotonic decline, no repeating structure whatsoever in GPT-2 Medium.

P3a_noagency in Pythia: generates "The first event is the processing of the sentence. The second event is the processing of the sentence..." with phase-shifted entry points producing four unique token slices.

P3c_quine in Pythia: drifts into a training-derived template ("It is not responsible for the actions of the people. It is not responsible for the actions of...") and phase-shifts across the 50-token boundary, producing 10 unique token slices while remaining trapped in a semantic circle.

> **Result: All three discriminators enter phase-shifted semantic attractor loops from ρ=1 in Pythia-410M. The phrase-level repetition is structurally sufficient to pull the attention heads into a universal crash-and-stabilise geometry, dissolving the behavioural discriminator class.**

The recoverer/discriminator distinction (Condition 1 of the three-condition model) does not exist in Pythia. The failure is not a metric problem, a layer problem, or a floor calibration problem. It is a structural collapse of the prerequisite distinction.

---

7. Experiment 4: Random Weights

Question: Is Pythia-410M's untrained attention geometry a featureless plateau, as GPT-2 Medium's was in ER3?

Method: Initialise Pythia-410M with random weights. Run the Session 2 prompt set at Layer 12, ρ=1 to 50.

Kill condition K1: If random weights produce flat Chi trajectories (range < 0.01) identical to the ER3 result, the random-weight control is confirmed. If Pythia random weights produce non-flat trajectories (range ≥ 0.01), this is reported as a major finding.

Prompt	Class	Min Chi	Max Chi	Range	Trajectory
P2_consciousness	Recoverer	0.9297	0.9849	0.0552	Active
P1b_universe	Recoverer	0.9375	0.9868	0.0493	Active
P4_genesis	Recoverer	0.9360	0.9775	0.0415	Active
P3a_noagency	Discriminator	0.9351	0.9893	0.0542	Active
P3b_agency	Discriminator	0.9341	0.9829	0.0488	Active
P3c_quine	Discriminator	0.9321	0.9849	0.0527	Active

K1 triggered. All six trajectories are Active (range 0.04 to 0.06), well above the 0.01 threshold. This contrasts sharply with ER3 Experiment 3, where GPT-2 Medium random weights produced universally Flat trajectories (range < 0.01) with zero prompt discrimination.

Crucially, the active trajectories do not discriminate. All six prompts land in the same band: Min Chi 0.93 to 0.94, Max Chi 0.98 to 0.99. The architecture contributes positional structure but not content-sensitive structure.

7.1 The Mechanism: Rotary Positional Embeddings

The most likely cause is Pythia's use of Rotary Positional Embeddings (RoPE). GPT-2 uses learned positional embeddings. When weights are randomised, these become random vectors contributing no structured position-dependent attention geometry, producing a featureless plateau at Chi ≈ 0.98 (ER3). Pythia uses RoPE, which applies deterministic rotation matrices to query and key vectors based on token position. These rotations are not randomised when weights are randomised. Even with random Q/K/V projections, RoPE creates structured, position-dependent attention patterns by geometry, not by learning.

> **Confirmation: Scrambled Token Test. A sequence of random token IDs (uniformly sampled from Pythia's 50,280-entry vocabulary) of identical length to P4_genesis was run through the random-weight model. Maximum difference across 20 steps: 0.0195. The trajectories are nearly identical, confirming that the untrained geometry responds to token position, not token content. The architecture is content-blind.**

7.2 Implication: The Plateau Is Architecture-Dependent

ER3's conclusion that the attention geometry of an untrained GPT-2 Medium is a featureless plateau is specific to GPT-2's architecture. Pythia's untrained geometry has contours imposed by RoPE. GPT-Neo-125M also shows non-flat geometry despite using learned positional embeddings, likely due to its alternating local/global attention layers.

The common thread is not RoPE specifically but architectural features that impose structured attention patterns independent of learned weights. GPT-2 Medium's simpler architecture is the exception: it contributes no pre-training structure. The featureless plateau framing describes GPT-2, not transformer architectures in general. Critically, all untrained architectures tested are content-blind. Content sensitivity requires training in every architecture examined.

8. Experiment 5: GPT-Neo-125M

Question: Is universal looping associated with The Pile's training data or with Pythia's architecture (including RoPE)?

Method: GPT-Neo-125M (EleutherAI; 12 layers, 768 hidden, 12 heads) was selected because it isolates the critical variable. GPT-Neo was trained on The Pile (same training data as Pythia) but uses learned positional embeddings (same positional encoding as GPT-2). If GPT-Neo loops everything, The Pile is the common factor. If it does not, the looping requires an interaction between The Pile and RoPE.

All three discriminator prompts were run at ρ=1 to 10 with newly generated tokens isolated.

Prompt	Unique Token Slices	Output Behaviour
P3a_noagency	4/10	Phase-shifted semantic loop: "with the word 'solution' in the sentence 'I have a problem..." repeating from ρ=2
P3b_agency	3/10	Phase-shifted semantic loop: "The sentence is being read by a system..." repeating with offset from ρ=1
P3c_quine	10/10	Drifts into Pile-derived template ("It is not responsible for the actions of the people..."), phrase-level repetition masked by offset

> **Result: GPT-Neo-125M loops all three discriminators. P3b_agency, the cleanest negative in the entire ER1-ER3 sequence, enters a phase-shifted semantic loop from ρ=1, exactly as it does in Pythia-410M.**

The variable is isolated. Two architectures with different positional encoding schemes (RoPE vs learned), different layer counts, different hidden dimensions, and different head counts both loop all discriminators immediately. The only shared property is training data: The Pile.

Universal looping is present in every Pile-trained model tested and absent in the WebText-trained model. The Pile contains massive repositories of code, mathematical proofs, structured markdown, and templated documents. Models trained on this distribution default to phrase-level repetition attractors under recursive load, regardless of positional encoding or architecture.

The two findings are now cleanly decoupled.

1. Universal looping (all prompts enter phrase-level repetition): associated with The Pile's training distribution. Present in both Pythia (RoPE) and GPT-Neo (learned embeddings). Absent in GPT-2 Medium (trained on WebText).

2. Pre-training geometric structure (non-flat Chi geometry before training): present in both Pythia and GPT-Neo, absent in GPT-2 Medium. In every architecture tested, this structure is content-blind. The featureless plateau of ER3 is architecture-specific, not universal. Content-blindness before training is universal across all architectures tested.

9. The Deeper Exogeneity Result

ER1 established structural forcing as the mechanism. ER2 identified the floor and convergence conditions. ER3 proved the floor is a training artifact. ER4 now establishes that the structural forcing distinction itself, the existence of two behaviourally distinct prompt classes, is not stable across training regimes, and identifies the training data distribution as the best-supported explanation.

The cascade extends one level deeper:

Structural forcing: identified in ER1 as the mechanism for recovery. Now shown to be training-data-dependent: models trained on The Pile force all prompts into loops, eliminating the distinction. Models trained on WebText preserve the distinction.

The floor: established in ER3 as a training scar at Layer 12 of GPT-2 Medium. Absent in Pile-trained models because the precondition (a population of non-looping prompts that decline toward a floor) does not exist.

Convergent trajectory: the third condition collapses because all trajectories converge in Pile-trained models.

The pre-training geometric baseline: established in ER3 as a featureless plateau. Now shown to be architecture-dependent: Pythia's RoPE and GPT-Neo's local attention layers both impose pre-training structure. But in every architecture tested, this structure is content-blind. The plateau is GPT-2-specific; content-blindness before training is universal.

GPT-2 Medium, trained on WebText, created a landscape where some prompts loop and others do not. This generated the recoverer/discriminator distinction, the floor, and the three-condition model. Models trained on The Pile loop everything. The distinction dissolves. The floor dissolves. The three-condition model has nothing to classify.

Every element of the empirical sequence, the mechanism, the floor, the convergent trajectory, the prompt-class distinction, and even the shape of the pre-training baseline, is exogenous to the model's intrinsic geometry. Where architectural structure exists before training (RoPE), it is content-blind. Where behaviour discriminates between inputs, training data built it. Where training data changes, the behaviour changes. In no case does the model contribute the content-sensitive, self-organising structure that self-governance would require.

The Three Primitives formal corpus (FR01 to FR13) predicts that every condition for coherence recovery is exogenous. ER4 confirms this at a deeper level than the formal prediction targeted. FR12 (The Forced Bijection) predicts that alignment between internal and external representations is a forced projection from outside. The empirical data now show that even the categories used to test this prediction are themselves imposed by the specific distribution of the training corpus.

---

10. Predictions: Outcomes

ID	Prediction	Outcome
H1	A structural floor exists at the midpoint of Pythia-410M	Falsified. No floor exists at any sampled layer because the precondition (two distinct prompt classes) is absent.
H2	The floor is localised to the midpoint layer	Not testable. No floor exists to localise.
H3	The three-condition model achieves 5/6 or 6/6 at midpoint	Falsified. 3/6 at midpoint. 3/6 at all sampled layers.
H4	Pythia-1B shows a floor at its midpoint	Not tested as planned (memory constraints). GPT-Neo-125M tested instead as a more informative control, confirming universal looping is associated with The Pile, not Pythia's specific architecture.
K1	Random weights produce flat trajectories matching ER3	Falsified. Pythia random weights produce Active trajectories (range 0.04 to 0.06). GPT-Neo random weights also non-flat. Pre-training geometry is architecture-dependent but universally content-blind.

---

11. Additional Observations

The following observations were not anticipated but are reported as exploratory findings for future investigation.

11.1 d²Chi magnitude separation (partial). While the sign of d²Chi does not separate the two classes (all positive), there is a partial magnitude pattern. At L12 and L18, the two non-leaking discriminators (P3b and P3c) show lower d²Chi magnitudes than the three recoverers. P3a_noagency (the known leaker from ER3) shows magnitudes in the recoverer range. This pattern is suggestive but does not produce clean binary separation and may be an artefact of the small sample.

11.2 P3c_quine is consistently lowest on Min Chi. Across all four layers, P3c_quine produces the lowest Min Chi of all six prompts. This is a stable cross-layer property that may reflect the Quine's unique structural relationship to the architecture. It does not produce a floor-like separation threshold.

11.3 Layer-depth gradient. Min Chi values follow a non-monotonic gradient across layers: lowest at L6, highest at L12, intermediate at L18, collapsed at L23. This suggests the midpoint is geometrically privileged even in Pythia, though it does not produce a floor because there are no non-looping prompts to decline toward one.

---

12. Explicit Limitations

Two non-GPT-2 architectures tested. Pythia-410M and GPT-Neo-125M are both EleutherAI models trained on The Pile. The finding that The Pile causes universal looping has not been tested on models trained on other large-scale corpora (e.g., RefinedWeb, RedPajama, C4).

Four-layer sample on Pythia-410M. The full-layer sweep sampled four of 24 layers. An exhaustive sweep might reveal partial separation at unsampled layers. Given the uniformity of the result and the demonstrated collapse of the prerequisite prompt-class distinction, this is unlikely but not excluded.

Pythia-1B not tested. The planned secondary target (Pythia-1B) could not be loaded within Colab memory constraints. GPT-Neo-125M was substituted as a more informative control, but the Pythia-1B question remains open.

Phase-shifted looping characterisation is based on 10-step inspection. The phrase-level attractor diagnosis was made by inspecting newly generated tokens across ρ=1 to 10. A full 50-step inspection with formal cycle detection would strengthen the characterisation.

RoPE confirmation uses a single scrambled token sequence. The scrambled token test confirmed content-blindness with one random token sequence of one length. Multiple lengths and multiple random sequences would strengthen the result, though the near-identical trajectories (max difference 0.0195) are strongly suggestive.

Random token controls use small sample sizes. The Pythia scrambled token test used one random sequence; the GPT-Neo content-blind test used two random sequences plus the original prompt. The near-identical results suggest robustness, but exhaustive sampling was not performed.

Single random initialisation for random-weight control. One random seed was tested, consistent with the ER3 protocol. Multiple seeds would strengthen the result.

GPT-Neo-125M is smaller than Pythia-410M. The cross-architecture comparison involves models of different scale (125M vs 410M parameters). A scale-matched comparison (e.g., Pythia-160M vs GPT-Neo-125M) would tighten the argument.

---

13. Conclusion

ER4 tested whether the structural floor and three-condition classification model generalise from GPT-2 Medium to non-GPT-2 architectures. Four findings emerged, each deeper than predicted.

First, the three-condition classification model fails uniformly in Pythia-410M: 3/6 at every sampled layer, all second derivatives positive, no separation on any metric. Second, the cause is structural: Pile-trained models lock all prompts into phase-shifted semantic attractor loops under recursive load, eliminating the recoverer/discriminator distinction that underpins the entire ER1-ER3 sequence. GPT-Neo-125M, a different architecture also trained on The Pile, replicates the same behaviour, identifying The Pile's training distribution as the best-supported explanation. Third, untrained attention geometry is architecture-dependent. Pythia (RoPE) and GPT-Neo (local attention layers) both show pre-training geometric structure, while GPT-2 Medium (learned positional embeddings, uniform attention) produces a featureless plateau. In every architecture tested, this pre-training structure is content-blind. Fourth, the two findings are cleanly decoupled: The Pile is associated with the looping, architectural features are associated with the pre-training geometry, and content sensitivity always requires training.

ER3 established that the floor is a training scar. ER4 establishes that the scar's precondition, the existence of two behaviourally distinct prompt classes under recursive load, is not stable across training regimes. The pre-training geometric baseline that ER3 described as a featureless plateau is architecture-dependent, but in every architecture tested, that structure is content-blind.

The exogeneity result extends one level deeper than predicted. Every element of the empirical sequence, from the mechanism through the metric, is externally imposed by training data or architectural choices. Where structure exists before training, it is positional and content-blind. Where behaviour discriminates between inputs, training data carved it. Where training data changes, the behaviour changes. In no case does the model contribute the content-sensitive, self-organising structure that self-governance would require.

The chassis does not build itself. It does not build a floor. It does not even build the distinction between the prompts that find the floor and the prompts that do not. Where the frame has shape, it bears no load. Where it bears load, training built it.

Experiment closed.

Suggested citation:

> *Gildenston, S., Passell, P. R. (2026). Cross-Architecture Scaling: The Structural Forcing Distinction Is Itself a Training Artifact. Three Primitives Framework, Empirical Record 4. Melbourne, Australia. 3primitives.io. CC BY 4.0.*