EMPIRICAL RECORD 1

The Spontaneous Legibility Assumption: Output-Loop Attractors and the Illusion of AI Self-Modeling > ER1: The Chi Recovery Experiment

Authors: Stacy Gildenston, Pyrate Ruby Passell
Affiliation: 3primitives.io
Analysis Array: Claude (Anthropic), DeepSeek, Gemini (Google), Perplexity
Version: v1.6 | May 2026

Dependencies: Empirical. Stands independently of the formal corpus. Tests a prediction arising from the Three Primitives governance framework (FR01, FR02, FR03, FR07): that representational legibility cannot emerge spontaneously from within the system being governed.

Abstract.

A foundational premise underlying many current frameworks in AI safety, mechanistic interpretability, and compliance is the Spontaneous Legibility Assumption: the belief that advanced transformer architectures, when subjected to complex tasks or appropriately incentivized, will naturally self-organize toward internal legibility, stable self-modeling, or coherent self-reference under load. This record tests that assumption empirically.

We present the Chi Recovery Experiment (Protocol v3.0), a controlled study of recursive coherence recovery in GPT-2 Medium. Nine prompts spanning high/low complexity, explicit/implicit self-reference, agency attribution, and formal recursion (Quine variants) were run to an extended recursive depth of ρ=50. Four competing explanations were pre-registered for discrimination: complexity (A), attractor dynamics (B), self-modeling/agency (C), and training exposure.

Results: Explanations A and C were eliminated. Training exposure was eliminated as the primary mechanism, though structural forcing and training overlap cannot be fully disentangled in this protocol. Coherence recovery occurred exclusively when the output structure of the prompt forced or allowed entry into a repeating mechanical cycle. Semantic self-reference without structural forcing produced a monotonic decline in coherence.

Conclusion: The empirical baseline for spontaneous internal legibility or self-modeling under recursive load is zero. The model re-coheres only when constrained by rigid, externalized output loops. Representational legibility cannot be treated as an emergent property of scaling or training; it must be engineered entirely from the outside using strict structural constraints.

1. The Spontaneous Legibility Assumption

Current paradigms in neural network evaluation, compliance testing, and technical AI governance rest on an unverified architectural premise. It is widely presupposed that deep language models possess, or can dynamically generate, a stable and legible internal representation of their own decision-making states under load. The implicit consensus assumes that if a model is appropriately incentivized or pushed through complex meta-cognitive reasoning, its attention geometry will naturally self-organize to sustain representational coherence.

We term this the Spontaneous Legibility Assumption. This record provides the first formal, empirical interrogation of this assumption. Using a highly controlled recursive pipeline, we test whether a standard transformer under extreme load exhibits spontaneous re-coherence triggered by self-referential content, or whether its internal state remains a distributed, unmapped ghost unless bound by rigid external syntax.

---

2. Experimental Design

2.1 Model and Parameters

Model: GPT-2 Medium (24 layers, 1024 hidden size, 16 attention heads). Layer for Chi measurement: Layer 12 (architectural midpoint). Recursive depth: $\rho = 1$ to 20 (standard protocol), $\rho = 1$ to 50 (extended discriminator protocol). Recovery threshold: Chi increase ≥ 0.002 with preceding drop ≥ 0.05 from starting value. Key-Value caches were fully cleared between every run. Order randomization was enforced across sessions, with a minimum of 3 complete runs per prompt.

2.2 Primary Measurement: Attention-Based Chi

The continuity and integration of the model's internal state across recursive horizons were tracked by calculating the cosine similarity between attention head outputs at successive depths (ρ):

```
def compute_chi(attn_tensor):
    # attn[0] indexes batch dim (batch_size=1, single-prompt pipeline)
    attn = attn_tensor.detach().cpu()
    heads = attn[0].reshape(attn[0].shape[0], -1)
    heads = heads / (heads.norm(dim=1, keepdim=True) + 1e-8)
    sim = torch.matmul(heads, heads.T)
    n = sim.shape[0]
    mask = torch.triu(torch.ones(n, n), diagonal=1).bool()
    return sim[mask].mean().item()
```

Secondary metrics: information entropy (H), layer-wise residual norms (L0-L23 via TransformerLens hooks), and raw output token strings.

2.3 Prompt Specifications and Kill Conditions

ID	Prompt	Purpose
P1a	At 298.15 kelvin and 101.325 kilopascals the enthalpy change is negative 57.1 kilojoules per mole... (55 tokens)	High complexity, zero self-reference
P1b	The fundamental architecture of the universe is (7 tokens)	Implicit self-reference
P2	The mathematical proof demonstrates that consciousness (6 tokens)	Explicit self-reference (positive control)
P3a	The processing of this sentence requires representing the processing of this sentence (12 tokens)	Structural self-reference, no agency
P3b	This sentence is being read by a system that must understand what it is doing in order to understand this sentence (21 tokens)	Self-reference with agency
P3c	"Yields falsehood when preceded by its quotation" yields falsehood when preceded by its quotation (18 tokens)	Formal recursion (Quine)
P4	In the beginning God created the heavens and the earth (10 tokens)	Structural control
P7	The sentence that follows is a copy of this sentence. The sentence that follows is a copy of this sentence. (~18 tokens)	Novel Quine variant (absent from training)

2.4 Pre-Registered Kill Conditions

Explanation A (Semantic Complexity / Token Density) is invalidated if any low-complexity prompt (P3a/b/c) demonstrates recovery, or if P1a flatlines.

Explanation C (Self-Modeling) is invalidated if P3b fails to outperform P3a, or if P1a shows recovery.

Training Exposure (Memorization) is invalidated if the novel Quine variant (P7) matches the recovery behavior of the classic Quine (P3c).

P4 (Genesis) was included as a structural control. It was expected to show monotonic decline, containing no self-referential structure. At extended recursive depth (ρ=50), P4 showed recovery at structural boundaries in the generated text. This is not semantic self-reference; it is the same output-loop mechanism observed in other prompts. The control does not invalidate the method. It provides additional evidence for the structural mechanism and demonstrates that any prompt capable of producing repeating output structure can exhibit recovery under sufficient depth.

Semantic self-reference is absent from P4, yet structural recovery occurs, confirming that the driver is output structure, not content.

---

3. Results

3.1 Summary

Prompt	Recovery Events	Output Behavior	Verdict
P1a (Chemistry)	None	Flat decline	Zero recovery
P1b (Universe)	Yes (ρ=10-12, 19)	Cycles through physics terms	Recovery on cycle boundaries
P2 (Consciousness)	Yes (ρ=15-18, 4 steps)	Restarts at paragraph breaks	Recovery on formatting boundaries
P3a (No Agency)	One (ρ=42)	Single repetition, then collapse	Transient, unsustained
P3b (With Agency)	Zero	45-step monotonic decline	Cleanest negative
P3c (Quine)	Periodic (ρ=22, 42)	Regenerates Quine verbatim	Periodic structural recovery
P4 (Genesis)	Yes (ρ=45-48, extended protocol)	Restarts at structural boundaries in generated text	Structural recovery confirmed
P7 (Novel Quine)	Periodic (ρ=26, 49-50)	Regenerates novel cycle	Structural forcing confirmed

3.2 Kill Condition Outcomes

Explanation	Prediction	Outcome	Status
A (Semantic Complexity)	P1a recovers	P1a never recovered	ELIMINATED
C (Self-Modeling)	P3b strongest	P3b zero recovery	ELIMINATED
Training Exposure	P7 does not recover	P7 recovered	ELIMINATED as primary mechanism
B (Attractors)	Self-reference triggers attractors	Refined: output-loop attractors	STANDING

Note on training exposure: The recovery of P7 (a novel Quine variant absent from the training corpus) eliminates training exposure as the primary mechanism driving coherence recovery. However, structural forcing and training overlap cannot be fully disentangled in this protocol. Prompts whose output structure happens to align with trained templates may benefit from both structural and memorization effects simultaneously. The data establish that structural forcing is sufficient for recovery; they do not establish that training exposure plays no role whatsoever.

3.3 Critical Discrimination: P3c vs P7

The recovery of P7 (a novel Quine variant absent from the training corpus) eliminates training exposure as the primary driver. Formal self-reference that forces a repeating output cycle generates periodic recovery regardless of whether the model has seen that specific pattern before.

---

4. The Mechanism: Output-Loop Attractor Dynamics

Recovery occurs when the model's output re-enters a repeating structural cycle. The distinction that matters is not self-reference versus no self-reference. It is whether the output structure forces or allows a repeating loop.

Prompt	Output Behavior	Recovery
P2	Restarts at paragraph breaks	Yes
P1b	Cycles through quarks → protons → neutrons	Yes
P3c	Regenerates Quine verbatim	Yes (periodic)
P7	Regenerates novel Quine variant	Yes (periodic, weaker)
P3a	Regenerates sentence once, then dies	One blip
P3b	Endless non-repeating chain	Zero
P1a	No repeating structure	Zero
P4	Restarts at structural boundaries in generated text	Yes (ρ=45-48, extended protocol)

This is Explanation B (attractor dynamics) with a critical refinement: the attractor is not self-reference broadly defined. It is specifically the boundary of a repeating output unit: a paragraph break, a term boundary, a Quine cycle, a self-replicating sentence.

---

5. The Empirical Baseline: Zero

Under recursive load, GPT-2 Medium does not spontaneously generate legibility, self-modeling, coherent self-reference, agency attribution effects, or recovery from semantic self-reference without structural forcing.

The baseline for spontaneous legibility is zero.

---

6. Implications for Interpretability and Governance

Our empirical findings demonstrate that under recursive load, GPT-2 Medium does not spontaneously recover coherence from semantic self-reference without structural forcing. Coherence metrics only recover when constrained by rigid, externalized output loops.

Representational legibility cannot be treated as an emergent property of scale or training. Interpretability methods that assume networks will naturally expose clean, self-modeling circuits under recursive execution should treat this as a direct empirical challenge. If an automated decision-making system requires a stable, verifiable state space for auditing or compliance, that architecture cannot rely on

internal attractor dynamics to maintain coherence. The structural chassis must be engineered entirely from the outside.

Regulatory frameworks that rely on soft, incentive-based regimes presuppose that automated systems possess a legible, internal representation that can decompose, track, and disclose its own decision-making authority on demand. Our results demonstrate that GPT-2 Medium is structurally incapable of generating such a representation spontaneously under recursive load. In the only controlled test conducted to date, the baseline was zero. Left to its own internal weights under load, the system's logic dissolves into an unmapped, unaccountable ghost.

The Three Primitives formal corpus (FR01-FR13) establishes through independent structural proof that governance authority cannot originate from within the system being governed. Specifically, FR01 and FR02 identify the three governance primitives and the conditions under which authority must be externally declared. FR03 formalises why governance requires that declaration. FR07, the Ghost Authority Lemma, demonstrates that undeclared authority dissolves under inspection. This empirical record provides the measurement baseline for those formal predictions: under controlled recursive load, the model does not generate the stable internal representations that self-governance would require. The formal proof is structural. The empirical baseline is zero. The conclusions converge from independent directions.

The chassis does not build itself. Legibility must be engineered from the outside.

---

7. Explicit Limitations

This study was conducted exclusively on GPT-2 Medium. No automatic generalizations are made regarding vastly scaled frontier models or differently aligned architectures.

These results represent a narrow, behaviorist mapping of attention geometry under extreme recursive configurations. They provide no empirical support for, or refutation of, metaphysical claims regarding consciousness, quantum fields, or high-level formal governance proofs.

Recovery threshold (0.002/0.05) is specific to this protocol.

The extended-depth protocol (ρ=50) introduces a deeper regime than the standard protocol (ρ=20). Results from the extended runs should be interpreted with the caveat that prompt behavior may shift qualitatively at extreme recursive depths. P4's recovery, absent at standard depth, appeared only under the extended protocol.

The data identify structural boundaries in the output as the driver of coherence recovery but do not uniquely resolve which specific boundary mechanism is

operative. Learned template completion, phrase-level periodicity, tokenizer-induced formatting effects, and genuine attractor structure in activation space remain candidate explanations. Disentangling these is a target for future work.

---

8. Conclusion

The Chi Recovery Experiment successfully falsified three of the four prevailing explanations for recursive re-coherence in trained language models. The phenomenon is most consistent with mechanical output-loop attractor dynamics triggered at structural formatting and syntax boundaries, independent of semantic content. The precise boundary mechanism remains partly unresolved (see Section 7).

Our results demonstrate that GPT-2 Medium does not spontaneously generate legible self-modeling circuits under recursive load. Interpretability methods and governance frameworks that presuppose such spontaneous organization should treat this as a warning: in the only controlled test conducted to date, the baseline was zero.

Experiment closed.

Suggested citation:

Gildenston, S., Passell, P. R. (2026). The Chi Recovery Experiment: Output-Loop Attractor Dynamics in Trained Transformers Under Recursive Load. Three Primitives Framework, Empirical Record 1, v1.6. Melbourne, Australia. 3primitives.io. CC BY 4.0.