# The Basin Is the Trajectory: The Loop Closure Principle in nanoGPT

Claudio Irrgang

May 8, 2026

**Abstract**

We introduce a controlled measurement framework for isolating the causal role of recursive token re-injection in autoregressive generation. Applied to a bilingual character-level transformer, the framework asks whether the recursive feedback loop is causally necessary for the trajectory-level structure observed in generation.

The intervention holds components, weights, seeds, and per-step distributions fixed, removing only token re-injection. With the loop severed, per-step divergence fails to accumulate (the JSD slope sits at machine zero) and regime structure collapses. Close the loop and persistent regime-switching dynamics return, with a positive divergence rate ($0.691 \pm 0.021$ nats/char across 4 seeds $\times$ 50 rollouts each; across-seed std, per-seed means from sanity $p$=1 conditions; see Table 11) at every tested temperature.

The same intervention inside 4 jointly trained bilingual variants ($M_{\text{ESPT}}$, seeds 42–45) replicates the contrast quantitatively: recursive fidelity drops $0.86 \pm 0.06 \rightarrow 0.49 \pm 0.04$ while static fidelity is preserved ($0.87 \pm 0.02 \rightarrow 0.88 \pm 0.03$).

The basin label commits over the first $\sim$1200 characters of recursive generation (mean commitment time 24.4 windows), not earlier. It does not separate under geometric clustering at any layer across all seeds tested. Four independent methods, including an endogenous decomposition that requires no reference models, agree on the same partition. Partial-feedback experiments show that any nonzero reinjection produces dynamics; steady-state divergence is invariant to feedback rate (three orders of magnitude) and to reinjection lag (up to 100 steps). Content invariance holds at the level of plateau existence – both uniform-random and marginal-sampled feedback sustain positive $D_{\text{rate}}$ – but the plateau height is content-dependent ($\sim$35% spread across real, uniform, and marginal feedback). A same-language negative control (mixture of two $M_{\text{ES}}$ variants at different random seeds) yields $D_{\text{rate}}$ 22$\times$ smaller than the bilingual baseline, confirming that the mechanism requires distinguishable components.

The theory's sufficient conditions for exact metastability do not hold in this setting. A mechanism consistent with the theory operates despite failure of its sufficient conditions. These results identify the *Loop Closure Principle (empirical invariance class) in nanoGPT*: recursive feedback is the necessary enabler that converts per-step component disagreement into trajectory-level divergence; once active ($p > 0$, distinguishable components), steady-state divergence exhibits empirical invariances over the tested ranges of rate, content, and timing: an empirically established property of recursive autoregressive composition in this controlled construction. The structural independence of this property from the standard sufficient conditions is verified formally in a synthetic finite-discrete recursive system in Appendix C; this is a logical-independence result in a designed model, not a theorem about transformers or any continuous process. The Principle is established within a controlled autoregressive mixture; its applicability beyond this regime remains an empirical question.

## Contents

**1 Introduction** **3**

1.1 Related work . . . . . 4

1.2 Outline . . . . . 4

**2 The Recursive Compression Loop** **4**
2.1 What the lens sees . . . 5
2.2 How lossy is the emission? . . . 5
2.3 What the lens tells us to measure . . . 5
2.4 The controlled mixture . . . 5
2.5 Why these languages . . . 6

**3 Severing the Wire** **6**
3.1 Design . . . 6
3.2 Results . . . 7
3.3 Temperature dependence . . . 7
3.4 The separation is not inherited from the seed . . . 8
3.5 What this rules out . . . 8
3.6 Partial feedback: rate-invariant steady state . . . 8
3.7 What kind of feedback is sufficient? . . . 9
3.8 Negative control: same-language mixture . . . 11
3.9 The Loop Closure Principle . . . 11

**4 What the Loop Creates** **12**
4.1 The output becomes its own thing . . . 12
4.2 Regimes without basins . . . 13
4.3 Commitment time . . . 15

**5 The Partition Is Real** **16**
5.1 Four methods, one partition . . . 16
5.2 The slow dynamics cross-validate . . . 16
5.3 No external models required . . . 17

**6 What the Lens Illuminated** **17**
6.1 The structural fit . . . 18
6.2 Observables that transferred . . . 18
6.3 Sufficient conditions not applicable here . . . 19
6.4 Diagnostic scope and limitations . . . 19

**7 Discussion** **19**
7.1 What this paper establishes empirically . . . 19
7.2 Bridge experiment: loop dependence in a trained model . . . 20
7.3 Connection to larger models . . . 21
7.4 Scope . . . 21
7.5 Limitations . . . 21
7.6 Implications . . . 22

**8 Conclusion** **22**

**A Experimental Details** **23**
A.1 Architecture . . . 23
A.2 Models . . . 23
A.3 Generation protocol . . . 23
A.4 Metrics . . . 23

**B Full Numerical Results** **24**
B.1 Recursion versus static . . . 24
B.2 Temperature sweep . . . 24
B.3 Typed separation . . . 24
B.4 Layer-by-layer analysis . . . 25
B.5 Singular spectrum . . . 25
B.6 Amplification hierarchy . . . 25
B.7 Commitment and hazard . . . 25
B.8 Coarse kernel . . . 25
B.9 Triangulation . . . 26
B.10 Trilingual ($B = 3$) . . . 26
B.11 Scope results . . . 26
B.12 Seed replication (Wave 1) . . . 27

**C Formal Verification Artifacts** **27**

# 1 Introduction

Autoregressive language models generate text by sampling each token from a conditional distribution given the prior context, then appending the sampled token to that context. The recursive feedback loop – where the model's own output becomes its next input – is structurally distinct from a non-recursive sampling process, but its causal role in the trajectory-level structure of generation has not been isolated in a controlled setting. This paper severs and restores that loop in a controlled bilingual mixture and in jointly trained bilingual variants, holding components, weights, seeds, and per-step distributions fixed. The intervention isolates feedback re-injection as the causal variable governing trajectory-level accumulation in this construction.

The same intervention in 4 jointly trained bilingual model variants ($M_{\text{ESPT}}$, seeds 42–45) replicates the contrast quantitatively: recursive fidelity drops $0.86 \pm 0.06 \rightarrow 0.49 \pm 0.04$; static fidelity is preserved ($0.87 \pm 0.02 \rightarrow 0.88 \pm 0.03$).

When the loop is removed, trajectory-level structure collapses (JSD slope $\approx 0$). When it is restored, divergence accumulates ($D_{\text{rate}} = 0.691 \pm 0.021$ nats/char across 4 seeds $\times$ 50 rollouts each). The same per-step component disagreement produces either noise or coherent regime-switching text depending only on the presence of feedback. This holds at every temperature we tested.

We use the Recursive Interaction Systems (RIS) framework [5] as a diagnostic lens. RIS was designed for systems with lossy emission and recursive feedback; autoregressive models have both. The framework tells us what to measure (separation parameters, divergence rates, spectral structure) and where to look (the emission map, coarse dynamics, commitment process). It provides the observables and structure that make the intervention interpretable, not the object of study.

The sufficient conditions for the theory's exact results are not applicable to our model; the languages are too close, the alphabets overlap too much, and a key entropy-rate assumption does not hold. Feedback re-injection is necessary for the regime structure we observe in this construction.

**Scope.** This paper studies a controlled autoregressive mixture system: a bilingual character-level transformer built from Spanish and Portuguese (deliberately close languages). This is a controlled construction demonstrating mechanism existence, not prevalence. We do not claim that all language models exhibit the same mechanism, only that it is causally necessary in this construction. We use "Principle" throughout to denote an empirically identified invariance class within a defined architectural scope, not a derived theorem.

**Contributions.**

1. We identify and characterize the *Loop Closure Principle in nanoGPT* (Section 3.9; partially formalized in Appendix C: algebraic identities and the structural-independence construction are fully proved in Lean, while spectral-gap and bridge results are statement-only): a principle scoped to the architecture class studied here (∼839K-parameter character-level transformer), with formal independence backing in finite discrete recursive systems. The empirical claim is that recursive feedback, together with distinguishable components and non-degenerate feedback content, is sufficient for persistent trajectory-level divergence in this construction.
2. A clean causal intervention isolating recursive feedback in the controlled mixture system ($M_{\text{ES}} + M_{\text{PT}}$): components, weights, and seeds held fixed; only re-injection removed.
3. Evidence from the controlled mixture that regime structure does not form centroid-separable clusters ($k$-means at chance across all seeds tested; $k$-nearest-neighbor accuracy rising from 52% to 77% by depth; typed separation parameters spanning three orders of magnitude).
4. A partition recoverable by four independent methods, including an endogenous decomposition requiring no external reference models.
5. We identify a self-averaging structure in the partial-feedback sweep: across reinjection rate spanning three orders of magnitude, mean $D_{\text{rate}}$ is invariant while per-trajectory dispersion compresses as $1/\sqrt{n_{\text{appends}}}$, a quantitative prediction stronger than rate-invariance alone.
6. A commitment result: the basin label commits over the first ∼1200 characters of recursive generation (mean $t^* = 24.4$ windows), not at the seed (Section 4.3).
7. Quantitative replication of the bridge contrast across 4 jointly trained $M_{\text{ESPT}}$ variants (seeds 42–45): recursive fidelity $0.86 \pm 0.06 \rightarrow 0.49 \pm 0.04$; static fidelity preserved at $0.87 \pm 0.02 \rightarrow 0.88 \pm 0.03$ (Section 7.2; Table 11). Composition variation (different language pairs) is pending follow-up (Wave 2).

### 1.1 Related work

Regime persistence in language generation is usually treated as a practical problem: mode collapse, repetitive generation, degeneration [4]. We are after the cause, not a fix.

Our probing methodology follows Alain and Bengio [1]. The observation that regime information is linearly decodable without forming centroid-separable clusters is consistent with Elhage et al. [3]; our typed-separation measurements give a quantitative instance of this. Time-lagged independent component analysis for slow-mode recovery originates in Molgedey and Schuster [8]; the spectral theory of metastability in Markov chains is classical [2, 7].

Closest prior work studies degeneration, probing, feature geometry, or slow-mode structure. We are not aware of prior work that isolates recursive token re-injection itself as the causal variable through a controlled intervention that holds components, weights, and seeds fixed. The result is a minimal causal demonstration that recursive feedback converts local disagreement into global structure in this mixture system.

### 1.2 Outline

Section 2 describes the loop and the experimental setup. Section 3 presents the intervention, including the partial-feedback rate sweep and rate-invariance evidence. Section 4 characterizes what the loop creates. Section 5 confirms the partition with independent methods. Section 6 maps the findings onto the RIS framework. Section 7 discusses scope and implications.

## 2 The Recursive Compression Loop

Applied to a transformer during generation, the RIS framework identifies three structural components: the hidden state, the lossy emission map, and the recursive feedback loop.

### 2.1 What the lens sees

A transformer maintains a hidden state: the residual stream $h_t \in \mathbb{R}^d$, a high-dimensional representation of everything the model knows about the context so far. At generation time, this state is compressed through the unembedding matrix $W_u$ and a softmax into a distribution over the vocabulary, from which a single token is sampled.

That token is then appended to the context. The model reads it back, and the next hidden state is computed from the updated context. The token carries less information than the hidden state that produced it. The subsequent hidden state is computed from this lower-information context.

This is the recursive compression loop:

$$h_t \xrightarrow{\text{lossy emission}} c_t \xrightarrow{\text{re-injection}} h_{t+1} = f(x_{\leq t}).$$

The hidden state is richer than the output. The output is what gets fed back.

### 2.2 How lossy is the emission?

The emission map destroys information.

In our model ($d = 128$, $|V| = 118$), the unembedding matrix $W_u$ has a 10-dimensional kernel: perturbations along those directions produce exactly zero change in the output distribution. But the lossiness extends well beyond the null space; we quantify the full singular spectrum in Section 4.2, where it helps explain why regimes (trajectory-level metastable patterns in generation, e.g., a sustained stretch of Spanish text; see terminology in §2.5) are invisible to clustering yet recoverable by $k$-NN probing and four independent partition methods.

What matters here is that the ambiguity propagates forward. When two distinct hidden states produce the same token distribution and the token is fed back, the dynamics do not necessarily collapse the distinction. Under direct test, 100% of 4,986 candidate aliased hidden-state pairs remained in distinct output regions after re-injection (Appendix B); the recursive map does not collapse the lossy ambiguity within tested ranges. The lossy gap does not close over time.

### 2.3 What the lens tells us to measure

Given this loop, the RIS framework identifies specific observables worth measuring: separation parameters, divergence rates, spectral structure, and a set of sufficient conditions that may or may not apply. We list them all in Table 2 (Section 6) and work through each one in §§3–5.

### 2.4 The controlled mixture

To isolate the feedback loop, we need a distribution we can hold fixed while varying only the loop. We construct the mixture

$$P^*(\cdot \mid x_{<t}) \;=\; \tfrac{1}{2}\, P_{\mathrm{ES}}(\cdot \mid x_{<t}) \;+\; \tfrac{1}{2}\, P_{\mathrm{PT}}(\cdot \mid x_{<t}), \tag{1}$$

where $P_{\mathrm{ES}}$ and $P_{\mathrm{PT}}$ are character-level transformers trained separately on Spanish (Galdos) and Portuguese (Eça de Queirós), matched for register and length. Both condition on the *same* context $x_{<t}$.

This mixture is not a trained model. It is assembled at inference time from two frozen checkpoints. Because the mixture distribution is defined conditionally at each step, we can toggle context accumulation without altering the per-step distribution. That switch is the intervention.

**Scope.** Two externally defined component models share an updating context; a natively trained bilingual model has richer internal dynamics that are not captured here. This construction isolates dynamics, not learned bilingual representations.

### 2.5 Why these languages

Spanish and Portuguese share more than 77% of their character inventory. If the model learns regime separation, it cannot come from disjoint alphabets; it has to come from distributional structure. The two models agree on what characters exist but disagree, at every step, on which character comes next. Full experimental details (architecture, training, corpora) are in Appendix A.

The trilingual extension (Section 4.1) adds French as a third Romance language at greater character-level distance from both Spanish and Portuguese. This provides three pairs (ES–PT, ES–FR, PT–FR) ordered by linguistic distance and a $B = 3$ mixture for the slow-eigenvalue prediction in Section 6. Two additional cross-pair variants ($M_{\text{ENFR}}$, $M_{\text{PTFR}}$) reported in Appendix B probe composition robustness across language families; English is the only non-Romance language used. Full corpus details for all setups are in Appendix A.

**Terminology.** We use *regime* throughout to mean a trajectory-level metastable pattern in generation (e.g., a stretch of Spanish text), identified by external scoring or by the independent methods in Section 5. Regime is the primary object of analysis in this paper. *Basin* denotes the same object viewed as a coarse metastable state in rollout space, not a geometrically separated region of latent space (Section 4.2).

## 3 Severing the Wire

The experiments in this section together establish the empirical content of the *Loop Closure Principle* (stated formally in Section 3.9). Each subsection isolates one property: necessity (severing the wire), rate-invariance (partial feedback), content-invariance (non-degenerate variable tokens), timing-tolerance (delayed feedback), and weight-robustness (asymmetric mixture).

If the loop is doing the work, removing it should break something. This is a causal intervention: we manipulate the feedback channel directly while holding components, weights, seeds, and the per-step conditional distribution fixed. The intervention isolates accumulation, not autoregression: both modes sample from the same conditional distribution at every step, and only the accumulation of those samples into context differs.

### 3.1 Design

We set up three generation modes that share the same seed contexts, the same component models ($M_{\text{ES}}$, $M_{\text{PT}}$), and the same random seeds:

1. **Recursive** $P^*$: sample from the mixture (Equation (1)), append the token to the context, continue.
2. **Static** $P^*$: sample from the same mixture, but do *not* update the context. Both components keep conditioning on the original seed. Same distribution, same weights. Static mode corresponds exactly to $p = 0$ in the partial-feedback sweep (Section 3.6): the context is never updated. It is the same stochastic process with accumulation disabled, not a separate model class.
3. **Component** $M_{\text{ESPT}}$: autoregressive generation from the jointly trained model (positive control).

### 3.2 Results

Severing the loop could weaken regime structure by reducing accumulation. It eliminates regime structure entirely: JSD slope drops to machine zero.

In static mode, the JSD slope between components is $-2.9 \times 10^{-20}$: machine zero (floating-point precision limit). The per-step distinguishability between the two models does not accumulate; it just sits at its seed value. There is nothing for the mixture to compound, and the generated text reflects this: it loses consistent regime structure (see Appendix B for language-marker counts).

JSD grows, and the divergence rate $D_{\text{rate}}$ (how fast the mixture's output drifts from each component; defined formally in Appendix A) is consistently positive across all 4 seeds tested, at $0.691 \pm 0.021$ nats/char (combined variance: across-seed std, each per-seed value averaged over 50 rollouts; all positive). The output reads as coherent regime-switching text. Same components, same weights, same seeds; structure or noise, depending on one switch. Recursive accumulation is necessary for per-step divergence to produce persistent trajectory-level structure; without it, divergence exists but does not compound.

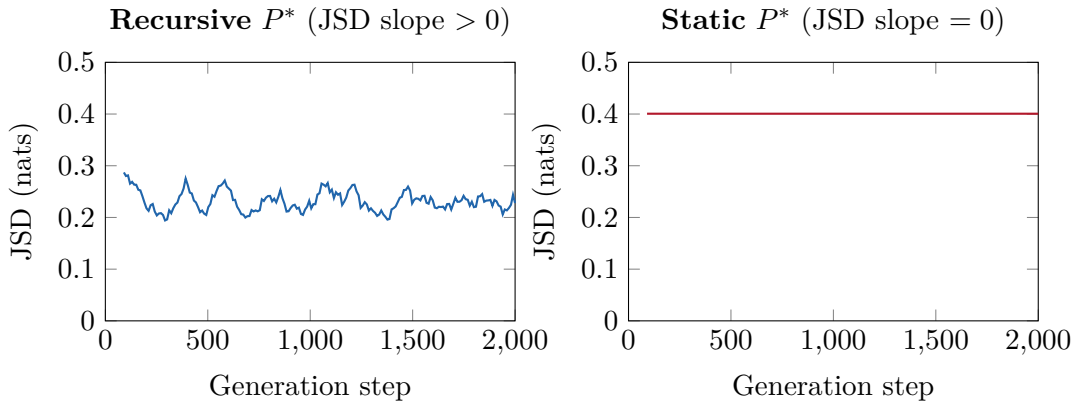


Figure 1: Cut the wire. Same components, same weights, same seeds; the only difference is whether generated tokens re-enter the context (i.e., whether accumulation is enabled). Per-step JSD between components freezes in static mode (right) and grows under recursive generation (left).

**Remark on $D_{\text{rate}}$ in static mode.** $D_{\text{rate}}$ measures the rate at which the trajectory's joint likelihood diverges from each component's; it is a trajectory metric only when the context is updated. In static mode, $D_{\text{rate}}$ reflects fixed-context bias, not accumulation: the static value should not be compared to the recursive value as if both measured the same quantity. The static-vs-recursive contrast is properly reported as the JSD slope (which is well-defined in both modes).

### 3.3 Temperature dependence

We checked whether the effect depends on sampling temperature. It does not. At every temperature from 0.3 to 1.2, the static JSD slope is machine zero and the recursive $D_{\text{rate}}$ is positive (Appendix B). Lower temperatures sharpen the distributions and amplify the feedback loop, but the gap between structure and noise is there at every temperature. The qualitative regime is temperature-invariant (positive recursive $D_{\text{rate}}$ vs. machine-zero static slope), though the recursive $D_{\text{rate}}$ *magnitude* varies monotonically with temperature, unlike the rate, content, and timing invariances established below. We vary sampling temperature; other entropy manipulations are outside scope.

### 3.4 The separation is not inherited from the seed

Could recursive mode simply be amplifying whatever the seed established? We tracked seed fidelity: the probability that the regime label at time $t$ still matches the seed language.

Under recursive generation, fidelity starts high and decays; the dynamics override the seed. Under static generation, fidelity starts lower and falls further, below chance. Without the loop, the seed's language signal does not persist.

The separation we observe at late times is produced by the recursive process, not inherited from the initial context (full curves in Appendix B).

### 3.5 What this rules out

The two component models disagree at every step in both modes. Both modes produce the same per-step component disagreement. Without the feedback loop, that disagreement does not accumulate into trajectory-level structure. The loop is the mechanism in this construction. This is not a comparison to non-autoregressive models; it isolates accumulation under identical per-step distributions. Without accumulation, disagreement does not organize.

**Note on the cut-wire contrast.** The static-vs-recursive contrast above is a definitional consequence of the construction: any time-dependent observable must differ between a process that updates context and one that does not. The genuine non-trivial content lies in the quantitative structure of the partial-feedback sweep below. Re-injection probability $p \to 0^+$ produces a sharp transition rather than a smooth approach to zero, and steady-state $D_{\text{rate}}$ is invariant across three orders of magnitude in $p$. Neither property is implied by the construction alone. We treat the cut-the-wire result as a sanity check on the measurement; the partial-feedback sweep is the load-bearing evidence for the loop-dependence claim.

### 3.6 Partial feedback: rate-invariant steady state

*Scope: rate sweep verified across 4 seeds (42–45); see below for multi-seed summary.*

To determine whether the effect scales with feedback strength or is binary, we swept reinjection probability from 0.001 to 1.0.

We reinjected the sampled token with probability $p$ at each step (context updates only when the Bernoulli gate fires) and swept $p$ from 0.001 to 1.0 while holding seeds and components fixed. At $p = 0$ the construction reduces to static mode; at $p = 1$ it is the full recursive mixture.

Three findings emerge (Figure 2).

**Empirical activation function.** Define

$$D(p) \coloneqq \lim_{t \to \infty} D_{\text{rate}}(p, t)$$

as the steady-state divergence rate at reinjection probability $p$. The empirical observation is $D(p) \approx D^*$ for $p \in (0, 1]$, with $D(0) = 0$, where $D^* \approx 0.706 \pm 0.058$ nats/char (pooled mean $\pm$ std across 4 seeds and 11 rate values). The transition at $p \to 0^+$ is discontinuous in the empirical data: there is no observed monotone scaling of $D(p)$ with $p$ across the tested range (log-log slope +0.15, not +1).

**Why the off-state is sharp (heuristic).** The transition is consistent with an absorbing/metastable picture: any positive reinjection rate eventually accumulates a path-dependent statistic whose dispersion compresses as $1/\sqrt{n_{\text{appends}}}$ (see self-averaging law below), so even sparse feedback dominates trajectory statistics in the long run, while $p = 0$ removes the accumulation channel entirely. We offer this as structural intuition, not proof; a renewal-process treatment is left to follow-up work.

**Off-state is unique.** $p = 0$ destroys trajectory-level structure exactly: JSD slope is machine zero and $D_{\text{rate}}$ is dominated by seed bias. Any positive $p$ we tested produces nonzero trajectory divergence.

**Steady-state is rate-invariant.** For $p \in [0.001, 1.0]$, steady-state $D_{\text{rate}}$ sits at $0.706 \pm 0.058$ nats/char (pooled mean $\pm$ std across 4 seeds and 11 rate values), invariant across three orders of magnitude in feedback probability.[1] The steady-state regime depends only on the presence of feedback, not its rate.

**Formation is rate-tolerant.** The rolling $D_{\text{rate}}$ curve converges to the same plateau at every $p$ but takes longer at lower $p$. Formation timescales range from a few hundred steps at $p = 1$ to ∼10,000 steps at $p = 0.001$, a weaker-than-$1/p$ dependence. We tested whether formation obeys a universal reinjection-count threshold or a simple $1/p$ timescale; neither description fits cleanly (log-log slope of lock time vs. $1/p$ is +0.15, not +1; number of reinjections at lock ranges from 18 to 197 across the sweep). Formation is weakly rate-dependent, while the steady state is rate-invariant (per-rate spread 0.65–0.88 across the four-seed pool).

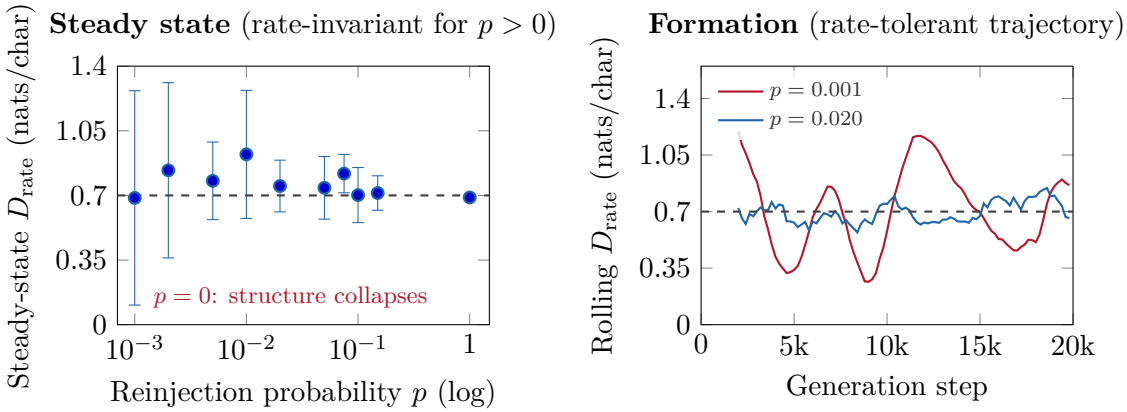


Figure 2: Partial feedback. Left: steady-state $D_{\text{rate}}$ is invariant across three orders of magnitude in reinjection probability $p$; the dashed line marks $D_{\text{rate}} \approx 0.7$ (verified across 4 seeds; pooled mean $D_{\text{rate}} = 0.706 \pm 0.058$ nats/char). At $p = 0$ (not shown on log axis) trajectory structure collapses exactly. Right: rolling $D_{\text{rate}}$ (window $W = 2000$) along a generation trajectory; all nonzero $p$ values converge to the same plateau, lower $p$ takes longer.

**Self-averaging structure.** The rate-invariance of mean $D_{\text{rate}}$ coexists with a strong $p$-dependent compression of per-stream variability. Within-stream standard deviation of $D_{\text{rate}}$ shrinks monotonically with reinjection probability: $\sigma \approx 0.35$ at $p = 0.001$, falling through 0.18 at $p = 0.005$, 0.06 at $p = 0.02$, to $\approx 0.03$ at $p = 1$ (4 seeds, ES side). The trend is consistent with $1/\sqrt{n_{\text{appends}}}$ self-averaging: the mean is invariant across reinjection rate, but trajectory dispersion falls as the number of context updates grows. This is a stronger structural claim than rate-invariance alone: the mechanism produces a $D_{\text{rate}}$ value whose mean is robust to $p$, and whose per-trajectory uncertainty is determined by accumulated update count.

### 3.7 What kind of feedback is sufficient?

*Scope: token-substitution, prompt microstructure, delayed feedback, and asymmetric mixture sub-experiments were each run at $p = 0.01$; token-substitution results are reported across 4*

[1] Measured in a late window (steps 15,000–20,000 of gen_len= 20,000 rollouts, where seed influence has maximally decayed); the per-rate mean ranges 0.65–0.88 nats/char at every tested $p$, within one standard deviation of the fully recursive value ($0.691 \pm 0.021$ across seeds 42–45).

*seeds (42–45); remaining sub-experiments are N=1 on seed 42 only; cross-seed replication is a follow-up.*

Content and timing experiments isolate whether specific token identity or reinjection lag affects the mechanism.

**Token-content substitution.** Fixing reinjection rate at $p = 0.01$, we replaced what the Bernoulli gate appends to the context with alternative token-sampling strategies (gen_len = 20,000). The mixture still samples its next output from the true conditional distribution given the current context; only the context update is perturbed. Replacing the sampled token with a uniform-random token, or with a token drawn from the unconditional marginal character distribution, preserves a sustained, positive $D_{\text{rate}}$, but the plateau height depends on content: across 4 seeds, $D_{\text{rate}}$ for uniform-token feedback is $0.951 \pm 0.075$ nats/char, for marginal-sampled feedback is $0.627 \pm 0.021$ nats/char, and for real-sample feedback is $0.704 \pm 0.036$ nats/char. Uniform feedback yields a plateau approximately 35% above the real-sample value; marginal feedback yields one approximately 11% below. The existence of a positive $D_{\text{rate}}$ plateau is content-invariant; the plateau height is not. Both non-degenerate substitutions produce a stable, positive $D_{\text{rate}}$ of the same order of magnitude as the real-sample control. The content-invariance claim is restricted to non-degenerate variable content; degenerate (fixed-token) feedback at low re-injection rates was not robust across seeds and is not pursued further here. Trajectory-level divergence requires *variability* in the feedback stream, but not precise state-conditional token content.

**Prompt microstructure.** We also tested whether small perturbations at the prompt steer formation, by appending short character patterns to the seed before generation: `.`, `...`, `.....`, `""`, `aaaaaa`, `abcabc` (16 seeds, gen_len = 4000, full recursive generation). All conditions were indistinguishable from baseline within one standard deviation on lock time, within-run switch count, and steady-state $D_{\text{rate}}$. Local prompt microstructure does not measurably influence formation in this system.

**Delayed feedback.** We tested whether the mechanism depends on the *timing* of reinjection by introducing a fixed lag $k$ between sampling a token and appending it to the context ($k \in \{1, 5, 20, 50, 100\}$, ctx_len = 256). Steady-state $D_{\text{rate}}$ remains within the range 0.65–0.69 nats/char across all tested lags (4 seeds, full-window means). A small monotone increase with lag of order one cross-seed standard deviation is observable but does not break the plateau structure: the mechanism's existence and approximate magnitude are preserved. Immediate reinjection is not required; the loop tolerates substantial delays, provided feedback remains nonzero.

**Synthesis.** The mechanism is insensitive to fine-grained token content, local prompt structure, and timing. What matters is the existence of a non-degenerate, varying feedback stream and its accumulation over time, not the specific tokens that realize it or their precise timing. (At 4000-character generation lengths, lock in the microstructure test occurs near the end of the sequence for all conditions; detecting subtler microstructure effects would require longer runs or finer-grained metrics.)

**Asymmetric mixture weights.** The construction generalizes beyond equal mixing. With $P^* = \alpha P_{\text{ES}} + (1-\alpha) P_{\text{PT}}$ for $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, the loop still produces stable regimes at every weight (basin occupancy tracks $\alpha$ smoothly $(0.00, 0.06, 0.49, 0.94, 1.00)$). Steady-state $D_{\text{rate}}$ scales with $\alpha$ as expected (the mixture output approaches the dominant component), but the mechanism operates at all tested weights. The asymmetric-mixture sweep shows the construction generalizes across mixing weights, not that steady-state invariance holds for all possible mixture parameterizations.

### 3.8 Negative control: same-language mixture

To test whether the mechanism requires distinguishable components, we replaced the bilingual mixture $P^* = \frac{1}{2}(P_{\text{ES}} + P_{\text{PT}})$ with a same-language mixture $P^* = \frac{1}{2}(P_{\text{ES}}^{(42)} + P_{\text{ES}}^{(43)})$, formed from two $M_{\text{ES}}$ models trained on the same Spanish corpus at different random seeds. The components are near-identical: their next-character distributions agree on most state-context pairs, with small seed-driven variation.

Under recursive generation with $p = 1$, $D_{\text{rate}}$ falls to $0.032 \pm 0.010$ nats/char on the seed-42 component and $0.035 \pm 0.003$ on the seed-43 component (4 streams, gen_len = 2000). The static contrast yields $D_{\text{rate}} = 0.028 \pm 0.027$ and $0.031 \pm 0.025$ respectively. Recursive and static $D_{\text{rate}}$ are statistically indistinguishable: the residual $\sim$0.03 nats/char reflects inter-model disagreement from independent initialization, not feedback-driven trajectory accumulation. The JSD slope is $+2 \times 10^{-6}$ in the recursive condition (effectively flat), confirming the absence of regime-switching dynamics.

Compared to the bilingual baseline ($D_{\text{rate}} \approx 0.69$ nats/char, 22$\times$ larger), the mechanism does not fire when the two components are similar. The Loop Closure mechanism in nanoGPT requires distinguishable components, not merely the existence of recursive feedback.

### 3.9 The Loop Closure Principle

**A structural reduction.** The collected results support a structural reduction: in this construction, trajectory-level regime structure is accounted for by accumulated per-step component disagreement under recursive autoregressive composition. The cut-the-wire result establishes the necessity of accumulation for trajectory structure; the partial-feedback, content, lag, and asymmetric-mixture sweeps establish that accumulation under non-degenerate, distinguishable-component conditions suffices for the observed trajectory-level divergence. This is a reduction, not a phenomenon report: the trajectory-level metastability we measure is what recursive composition makes of the local disagreement signal, not a new emergent property of the loop itself.

**The Loop Closure Principle in nanoGPT (this paper).** We name this empirically observed collection of invariances the *Loop Closure Principle in nanoGPT* (an empirical invariance class): a principle within the small character-level autoregressive transformer class studied here ($\sim$839K parameters, 4-layer, $d_{\text{model}} = 128$, ctx_len = 256).

Here, "Principle" denotes a stable empirical invariance pattern under the tested perturbations (rate, content, timing), not a general theorem over autoregressive systems.

Cross-architecture and cross-scale verification (LSTM/RNN, larger transformers, non-language-domain mixtures) is the natural follow-up program; see the scope discussion in Section 7.

Within this controlled construction, the recursive-feedback mechanism produces a steady-state $D_{\text{rate}}$ whose mean is invariant to feedback rate (over three orders of magnitude), to non-degenerate token content, and to reinjection lag, while per-trajectory dispersion compresses as $1/\sqrt{n_{\text{appends}}}$ toward the steady-state value. The bare existence of accumulation under non-zero feedback is expected from the construction itself; the surprising content is the joint invariance (rate, content, timing) and the self-averaging structure of $D_{\text{rate}}$ across reinjection regimes. Structurally, the independence of these properties from the three RIS sufficient conditions ($\Delta_i > 0$, hard token partition, geometric basin separation) is verified for finite discrete recursive systems in `LoopClosureIndependence.lean` (Appendix C).

The Loop Closure Principle in nanoGPT is falsifiable. Predicted violations and outcomes:

1. *Machine-zero static slope.* Recovering the same regime structure with $p = 0$ would falsify loop dependence. Not observed: JSD slope at $p = 0$ sits at machine zero (Section 3).
2. *Centroid-separable clusters at any layer.* Observing centroid-separable clusters via $k$-means at any layer would falsify the non-geometric structure claim. Not observed: $k$-means returns chance-level purity at every layer (Section 4.2).

3. 22× $D_{\text{rate}}$ *reduction in same-language mixture.* Similar regime accumulation in a same-language (non-distinguishable-component) mixture would falsify component-distinguishability as the mechanism. Not observed: the same-language mixture (Section 3.8) yields $D_{\text{rate}}$ 22× smaller than the bilingual baseline with no recursive amplification.
4. *No monotone scaling of* $D_{\text{rate}}$ *with* $p$. Finding $D_{\text{rate}}$ that scales monotonically with $p$ across the tested range would falsify rate-invariance. Not observed: $D_{\text{rate}}$ across $p \in [0.001, 1.0]$ shows pooled mean $0.706 \pm 0.058$ with no monotone scaling (Section 3.6).

Together, the rate, content, and timing invariances, the machine-zero same-language null, and the four-method agreement (clustering, $k$-NN, TICA, coarse Markov) constitute the cumulative empirical signature of the Loop Closure Principle in nanoGPT. The structural reduction and the Principle are scoped to this construction; see Section 7 for scope and generalization limits.

# 4 What the Loop Creates

The loop is causally necessary for regime structure in this construction. The following three subsections characterize what the loop produces, using observables identified by the RIS framework.

## 4.1 The output becomes its own thing

A 50/50 mixture with no feedback would produce token-by-token weighted averaging of the component distributions. Under recursive generation, the output is not a blend: it departs from both components at $D_{\text{rate}} = 0.691 \pm 0.021$ nats/char.

The divergence rate $D_{\text{rate}}$ measures how fast the mixture's output drifts away from each component. It is positive on every rollout tested, consistently across all 4 seeds (Section 3; $D_{\text{rate}} \approx 0.70$ nats/char, see Table 11). The log-likelihood ratio between the mixture and each component accumulates linearly across 2000 characters. The mixture's output occupies a region of sequence space that neither component prefers. For scale: the correct component model assigns nearly 2 nats/char more to its own language than the wrong one does (+1.81 for ES, +1.87 for PT), and the recursive mixture pulls away from both at $D_{\text{rate}} \approx 0.69$ nats/char, robustly across all 4 seed-replicate runs (see Table 11).

At the character level, Spanish and Portuguese look nearly identical: unigram JSD is 0.019 nats. Condition on context and the distinction jumps to 0.244 nats (13×); the models agree on what characters exist but disagree on which character comes next. The trajectory-level $D_{\text{rate}}$ adds a further 2.9× through recursive compounding. After just 7 characters of recursive generation, the mixture assigns 100× higher probability to its own output than the best component does; under static generation, the same crossover takes 19 characters.

The full typical-set theorem of Irrgang [5] requires assumption (b) ($\Delta_i > 0$), which does not hold here: both $\Delta_{\text{ES}} = -0.139$ and $\Delta_{\text{PT}} = -0.212$ are negative; what does hold is the positive-divergence lemma ($D_{\text{rate}} > 0$, formally verified in Appendix C) and the overlap-tolerant bound.

A trilingual extension (Spanish, Portuguese, French) confirms the mechanism fires even as $|\Delta_i|$ grows with linguistic distance: $D_{\text{rate}}$ tracks conditional JSD ($r = +0.999$) and $|\Delta_i|$ worst component ($r = +0.993$) near-perfectly across three pairs ($n$=3, illustrative, not inferential; Table 1).

Table 1: Mechanism strength tracks component separation continuously. Three pairs from the trilingual setup, ordered by linguistic distance. $\Delta_i$ is the cross-entropy separation of component $i$ under $P^*$; all values negative. $D_{\text{rate}}$ is averaged over the two components of each pair. Correlation between $D_{\text{rate}}$ and conditional JSD ($r = +0.999$) or $|\Delta_i|$ worst-component ($r = +0.993$) is near-perfect across the three pairs ($n = 3$; illustrative, not inferential).

| Pair | uni-JSD | cond-JSD | $h(P^*)$ | $\Delta_i$ | $\Delta_j$ | $D_{\text{rate}}$ |
|---|---|---|---|---|---|---|
| ES–PT | 0.026 | 0.247 | 1.577 | −0.028 | −0.137 | +0.750 |
| ES–FR | 0.043 | 0.353 | 1.699 | −0.131 | −0.321 | +1.082 |
| PT–FR | 0.053 | 0.376 | 1.684 | −0.121 | −0.386 | +1.140 |

The trilingual extension shows a sufficient mechanism for trajectory-level divergence under recursive composition, not a claim about all mixture systems.

### 4.2 Regimes without basins

A geometric-separation account of persistent regimes would predict centroid-separable clusters at one or more layers. $k$-means clustering on residual-stream activations at every layer gives purity 52.6% (indistinguishable from chance). Class balance is 50/50, so this is indistinguishable from chance by any meaningful margin. Silhouette score: 0.125. The two regimes do not occupy centroid-separable regions at any depth; nonlinear geometric separation is not excluded. The claim is that regime information is linearly decodable without being geometrically dominant.

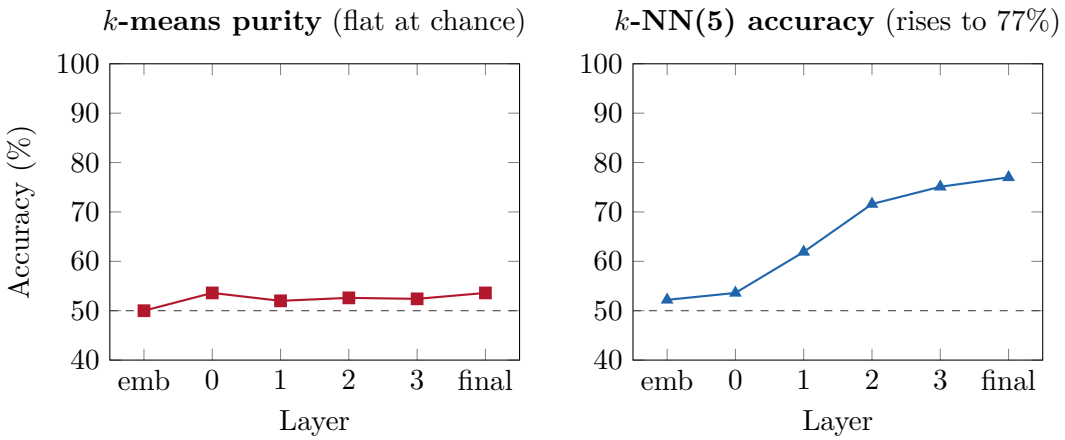


Figure 3: Regimes without basins (seed 42). $k$-means purity sits at chance at every layer (robust across all 4 seeds; see Table 11). $k$-NN accuracy rises from 52% (embedding) to 77% (final layer), confirming that nonlinear separability increases with depth despite the absence of centroid structure.

But the regime is there; it is just not where we expected. $k$-nearest-neighbor accuracy rises from 52% (embedding) to 77% (final layer), indicating that regime information accumulates with depth without forming centroid-separable clusters. The $k$-means chance result is robust across all 4 seeds tested (Table 11). A linear probe was attempted on the residual stream and produced 96% accuracy on seed 42 but 55–59% (near chance) on seeds 43–45. We do not have an explanation for the seed-42 anomaly; the chance-level result on three of four seeds is consistent with the $k$-means finding. We do not include the linear-probe result in our analysis.

The RIS framework formalizes this through typed separation parameters [5]. Three measures of the same system span three orders of magnitude: the models disagree substantially at every step ($\delta_{\text{pred}} = 0.484$, expected per-step total variation between component conditional distributions); the system almost never crosses from one regime to the other ($\delta_{\text{ker}} = 0.0016$, cross-basin

transition rate); and the regimes are invisible to unsupervised clustering ($\delta_{\text{geom}} \approx 0$, minimum cross-basin emission total variation). The ratio $\delta_{\text{pred}}/\delta_{\text{ker}}$ is 296 (Figure 4). The system is strongly separated in prediction, weakly separated in dynamics, and invisible geometrically.

Operationally, $\delta_{\text{pred}}$ supplies the per-step signal that recursive accumulation compounds, $\delta_{\text{ker}}$ governs dwell time once a regime is entered, and $\delta_{\text{geom}} \approx 0$ explains why centroid clustering fails despite strong dynamical separation.

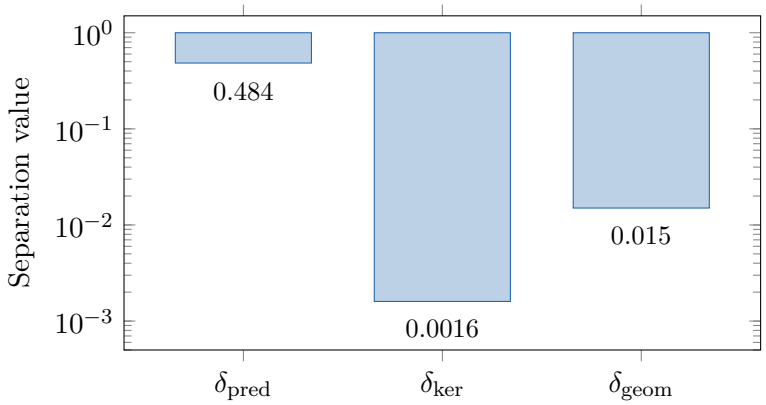


Figure 4: Typed separation parameters measured on the same model. $\delta_{\text{pred}}$ (predictive), $\delta_{\text{ker}}$ (dynamical), and $\delta_{\text{geom}}$ (geometric) span three orders of magnitude.

The emission map is what makes this possible. Of 128 emission dimensions, not a single one is fully observable (TV $> 0.10$ under unit perturbation). 60 are weakly observable; 58 are effectively dark (a direction is dark if unit perturbation along it produces less than 0.01 total variation in the output distribution, making it practically invisible at the token output channel) with a sharp cliff at singular value $\sigma = 0.42$ (see Appendix B.5 for the complete spectrum). Fisher eigenvalues track $\sigma_k^2$ at $r = 0.972$, confirming that observation effort scales as $1/\sigma_k^2$. The model carries regime information that the emission map does not pass through to the output. Most of the hidden state is unobservable through the token emission channel. The regime is linearly decodable yet not captured by centroid-based clustering. Centroid-based clustering of hidden states is one common regime-discovery method; it finds nothing here. Regime information must be recovered dynamically or via nonlinear similarity ($k$-NN).

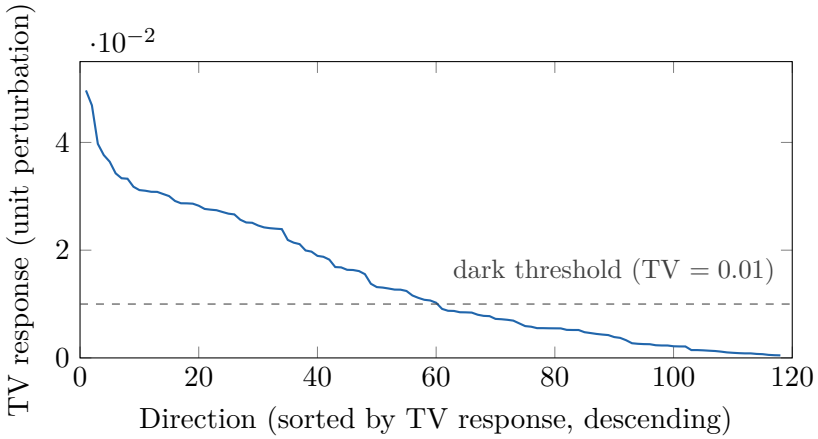


Figure 5: Observability spectrum. TV response to a unit perturbation along each singular direction, sorted in descending order. Above the dark threshold (TV = 0.01), 60 directions are weakly observable; below, 58 are effectively dark. None are fully observable.

### 4.3 Commitment time

The eventual regime is determined by the recursive dynamics over the first ∼1200 characters of generation, not by the seed alone. The commitment time $t^*$ (the earliest window after which the basin label never changes) has mean 24.4 windows (1218 characters) and median 28.5. Some rollouts commit immediately ($t^* = 0$); others wander the full sequence before settling ($t^* = 39$). Commitment is not front-loaded; it emerges through recursive reinforcement over hundreds of steps.

Commitment has a sharp internal signature coincident with the output-level event. We ran TICA on the residual stream of $M_{\text{ESPT}}$ rollouts (40 streams, lag $\tau = 50$ characters, $\lambda_{\text{slow}} = 0.665$). The slow coordinate crosses sign within ∼100 characters of $t^*$ and aligns with each stream's final regime only *after* the crossing.

Aligning streams to their own $t^*$ and grouping by final regime, the ES-versus-PT slow-coordinate difference has Cohen's $d$ (effect size in pooled SD units) $= -1.31$ at $t^*-100$, $-1.46$ at $t^*$, and $+4.54$ at $t^*+100$. The effect is strongly anti-aligned with the final regime before commitment and strongly aligned after. A sign-of-slow-coordinate classifier achieves near-chance accuracy on the final regime during the pre-commit phase, across all streams and restricted to streams that have not yet committed alike. The transition is a basin-crossing event in latent space, not a pre-established internal state. Internal state and output regime synchronize at the moment of commitment.

Once committed, the system dwells. The coarse kernel implies mean residence of $1/0.052 \approx 19$ windows (960 characters) in the Spanish regime and $1/0.091 \approx 11$ windows (550 characters) in Portuguese, with a slow eigenvalue $\lambda_2 = 0.834$. And the coarse dynamics are memoryless: a Kolmogorov–Smirnov test on run-length distributions gives $p = 0.54$ (geometric model not rejected; see Figure 6), and the hazard rate is flat (short/long ratio = 0.98).

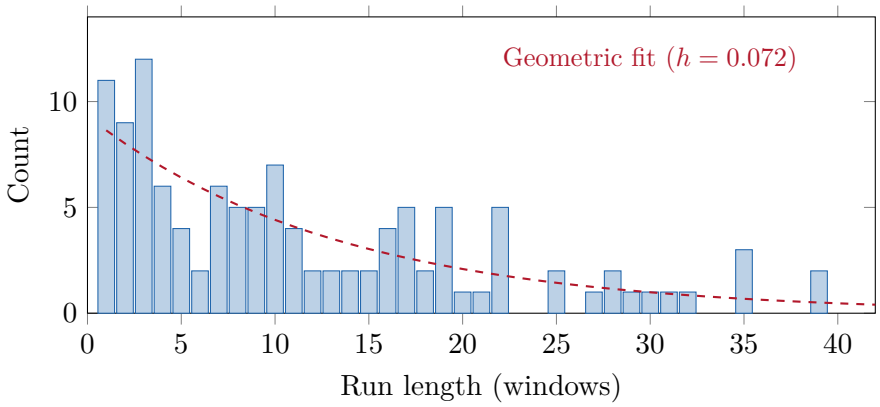


Figure 6: Run-length distribution of regime dwells (50-character windows) against the geometric fit with hazard $h = 0.072$. KS test $p = 0.54$: the memoryless model is not rejected.

The coarse kernel is asymmetric (ES → PT = 0.052, PT → ES = 0.091), which creates the *illusion* of momentum: minority-language excursions seem long because the escape rate back is low, not because the system builds inertia.

The coarse kernel is not even a fixed property of the model. Under ES-prompted generation, the PT→ES escape rate is 2.4× faster than ES→PT; under PT prompting the asymmetry vanishes entirely (Figure 7). Basin depth is conditional on the initial context, not a static landscape.

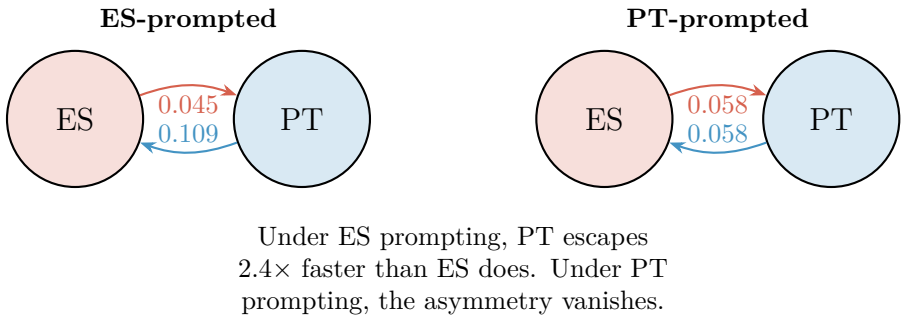


Figure 7: Coarse transition rates under different prompts. The kernel asymmetry observed under ES prompting vanishes under PT prompting; basin depth is prompt-dependent.

We use "basin" here not in the geometric sense (a separated region of state space) but to mean a coarse metastable regime in rollout space: a self-consistent pattern in the token stream that sustains itself through the feedback loop.

The trilingual extension (Section 4.1) provides consistency evidence at higher distinguishability without changing the qualitative picture.

# 5 The Partition Is Real

Everything above relies on external scoring: we label each window by which component model assigns it lower cross-entropy. A fair question is whether the partition is a property of the data or of that particular labeling method.

## 5.1 Four methods, one partition

We applied four recovery methods to the same $M_{\text{ESPT}}$ generation data, each producing a binary regime label per 50-character window from fundamentally different information: an external scorer (cross-entropy under $M_{\text{ES}}$ vs. $M_{\text{PT}}$), TICA (time-lagged independent component analysis at lag 200, binarized at the median), the sign of the cumulative log-likelihood ratio, and an endogenous HMM (TICA slow direction followed by a 2-state hidden Markov model, requiring no external reference models).

The scorer, TICA, and endogenous HMM agree at 95%+ with $\kappa > 0.89$; the HMM recovers the partition at 97.9% agreement with the external scorer (Appendix B). External model scores, slow spectral structure, raw character likelihoods, and an entirely endogenous decomposition all converge on the same two-state description.

**Seed robustness.** The $k$-means finding (chance-level purity at every layer) replicates consistently across all four $M_{\text{ESPT}}$ seeds tested (Wave 1; see Table 11). The absence of geometric clustering is the load-bearing finding here.

## 5.2 The slow dynamics cross-validate

From the coarse kernel, $\lambda_2 = 0.834$ per 50-character window. At lag 4 windows (200 characters), eigenvalues compose under lag, so the coarse-kernel eigenvalue predicts a TICA eigenvalue of $\lambda_2^4 = (0.834)^4 \approx 0.484$. The measured TICA eigenvalue is 0.496: two independent measurements of the same timescale, from different data, agreeing within 2.5%. This is a falsifiable prediction: the coarse-kernel eigenvalue, measured from a 2-state Markov fit to component-scoring labels, predicts the TICA eigenvalue, measured from residual-stream dynamics under a 4-window lag. The two methods are independent (the TICA computation does not use component scoring); the 2.5% agreement is a non-trivial internal consistency check.

### 5.3 No external models required

The controlled mixture experiments use separately trained $M_{\mathrm{ES}}$ and $M_{\mathrm{PT}}$ as references. One concern is that the partition reflects the external scorer rather than a property of the generation.

It is not. A TICA slow direction extracted from $M_{\mathrm{ESPT}}$'s own generation residuals, followed by a 2-state hidden Markov model, recovers the same partition at 97.9% agreement with the external scorer, with positive $D_{\mathrm{rate}}$ in both inferred states. The partition does not require external component models. It is real, and it is recoverable from the mixed model alone.

## 6 What the Lens Illuminated

The RIS framework [5] provides a vocabulary of named observables and sufficient conditions for metastability in recursive systems. We use it as a diagnostic lens – it tells us what to measure and what conditions to check – but the empirical findings of this paper are not derived from RIS; they are measured directly and then mapped onto the RIS vocabulary. Beyond serving as a measurement vocabulary, the empirical results partially validate RIS's observable structure: the typed separation parameters, divergence rate, and slow-spectral diagnostics that RIS specifies all measure non-trivial structure in our regime, where the framework's sufficient conditions ($\Delta_i > 0$, hard partition, geometric basin separation) do not hold. RIS thus functions in this paper as a measurement program that survives the empirical regime where its analytical sufficient conditions fail. The two tables below summarize that mapping: Table 2 lists what RIS told us to look for, and Table 3 reports what we found.

Table 2: The RIS diagnostic menu: what the framework tells us to look for in an autoregressive system.

| **What to look for** | **RIS concept** | **Where tested** |
|---|---|---|
| *Structural fit* | | |
| Latent state exists | Def RIS | §2 |
| Emission is lossy | Def RIS | §2 |
| Output is fed back | Def RIS | §2 |
| *Observables* | | |
| Do separation types diverge? | Def typed-sep | §4.2 |
| Is observability graded? | Prop graded-obs | §4.2, App. B.5 |
| Does TICA recover slow modes? | Prop tica-recovery | §5 |
| Is $D_{\mathrm{rate}} > 0$ under recursion? | Lem positive-div | §3, §4.1 |
| Does small $\delta_{\mathrm{ker}}$ imply slow mixing? | Rem cheeger | §4.2 |
| Are there $B-1$ slow eigenvalues? | Cor slow-sector | §4 |
| *Sufficient conditions (may or may not apply)* | | |
| Hard token partition? | Thm coherence-lump | §6 |
| Geometric basins? | $\delta_{\mathrm{geom}}$ | §4.2 |
| Entropy-rate gap ($\Delta_i > 0$)? | Thm typical-set | §4.1 |
| Approximate lumpability? | Thm approx-lump | §6 |

Table 3: RIS scorecard: what we measured and how well each aspect applies to our bilingual transformer. Compare with the menu above. Relevance codes: S = Strong (transferred cleanly); P = Partial (phenomenology present, formal guarantee does not apply); N/A = condition designed for a different regime.

| **RIS aspect** | **Reference** | **Our measurement** | **Rel.** |
|---|---|---|---|
| *Structural fit* | | | |
| Latent state | Def RIS | $h_t \in \mathbb{R}^{128}$ | S |
| Lossy emission | Def RIS | $\dim\ker(W_u) = 10$; 58 dark dims | S |
| Recursive re-injection | Def RIS | Autoregressive generation | S |
| *Observables* | | | |
| Typed separation diverges | Def typed-sep | 296× ratio | S |
| Graded observability $\propto 1/\sigma_k^2$ | Prop graded-obs | Fisher tracks $\sigma_k^2$, $r$=0.972 | S |
| TICA recovers slow mode | Prop tica-recovery | Matches kernel within 2.5% | S |
| $D_{\text{rate}} > 0$ for mixtures | Lem positive-div | $0.691 \pm 0.021$, positive across all 4 seeds | S |
| Small $\delta_{\text{ker}} \Rightarrow$ slow mixing† | Rem cheeger | $\delta_{\text{ker}}$=0.0016, $\lambda_2$=0.834 | S |
| $B-1$ slow eigenvalues | Cor slow-sector | $B$=3: two distinct slow $\lambda$ | S |
| *Phenomenology vs. formal guarantee* | | | |
| Metastable regimes | Thm metastability | Long dwells, constant hazard, Markov not rejected | P |
| Typical-set separation | Thm typical-set | $D_{\text{rate}} > 0$; good-subset bound holds | P |
| *Sufficient conditions (designed for separated sources)* | | | |
| Hard token partition | Thm coherence-lump | $\varepsilon$=0.999, $\eta$=0.99 | N/A |
| Geometric basins | ($\delta_{\text{geom}}$) | $k$-means 52.6% every layer | N/A |
| Entropy-rate gap $\Delta_i > 0$ | Thm typical-set | $\Delta_{\text{ES}}$=−0.139, $\Delta_{\text{PT}}$=−0.212 | N/A |
| Approximate lumpability† | Thm approx-lump | Bridge conditions too rigid | N/A |

†Lean formalization for these rows: statements verified, proofs deferred (mathlib4 v4.29.1 lacks the required Markov-chain spectral-gap machinery). See Appendix C (App. D).

### 6.1 The structural fit

At the level of the generation loop, the fit is exact: autoregressive generation has a latent state, a lossy emission, and re-injection. This is the RIS pattern.

The fit becomes approximate because a transformer's effective state grows with context length, while the theory assumes a fixed-dimensional state. Our experiments do not resolve how good that approximation is, but the observables the theory identifies turned out to be informative despite it.

### 6.2 Observables that transferred

Six observables from the RIS framework were measured in our model (Table 3, middle rows), and all six were informative.

The typed-separation divergence (296×) showed that the three separation parameters are genuinely different quantities, not labels for the same thing. Graded observability ($r = 0.972$ between Fisher eigenvalues and $\sigma_k^2$) upgraded non-identifiability from "10 null dimensions" to a spectrum spanning all 128. TICA recovered the same slow timescale as the coarse kernel (within 2.5%), connecting internal spectral structure to output-level regime transitions. $D_{\text{rate}}$ is consistently positive across all 4 seeds tested (all rollouts positive within each seed). Cheeger's inequality applies to the fitted coarse 2-state Markov operator, with $\delta_{\text{ker}}$ playing the role of an empirical cross-basin conductance. Cheeger's inequality then bounds the spectral gap above and below by $O(\delta_{\text{ker}})$. Small $\delta_{\text{ker}}$ therefore implies slow mixing of the coarse chain without requiring

lumpability of the full system (see Rem cheeger in [5] for the precise hypotheses).[2] The trilingual extension confirmed $B-1$ slow eigenvalues.

### 6.3 Sufficient conditions not applicable here

Four sufficient conditions for the theory's exact results are not applicable to our setting (Table 3, bottom rows). These conditions were designed for a different regime: hard partitions, separated alphabets, strong entropy-rate gaps. Spanish and Portuguese share too much of their character inventory for token partitions. Geometric basins do not form at any depth. The entropy-rate assumption does not hold for either component. And the bridge-theorem conditions need more separation than overlapping sources provide.[3]

Yet the key phenomenology appears: long dwell times, constant hazard, positive $D_{\text{rate}}$, Markov-consistent coarse chain, four-way agreement on the partition. These are the "Partial" rows in the scorecard: the behavior the theory describes is present, but the formal guarantee that would explain it does not apply. This does not invalidate the RIS framework; rather, it identifies a pathway to the same phenomenology outside its current sufficient-condition regime.

### 6.4 Diagnostic scope and limitations

The lens told us what to measure, and the measurements identify a pathway to trajectory-level structure not covered by the theory's sufficient conditions.

**Scope of the formal artifacts.** Of the seven Lean artifacts (Appendix C), five are fully proved with zero `sorry`. Four of these prove algebraic identities directly used by the paper: the $D_{\text{rate}}$ chain-rule estimator (`DrateIdentity.lean`), the AM–GM bound on the amplification base (`AmGmBound.lean`), the closed-form sequential amplification ratio (`AmplificationRatio.lean`), and the positive-divergence lemma under $\delta$-separation (`PositiveDivergence.lean`). The fifth, `LoopClosureIndependence.lean`, establishes logical independence of three RIS sufficient conditions in a finite discrete recursive system: it is a structural-consistency result in a designed model, not a theorem about transformers or any continuous process. The two statement-only files (`Cheeger.lean`, `BridgeTheorem.lean`) carry helper lemmas proved but main theorems deferred due to mathlib4 v4.29.1 spectral-gap gaps; their content is referenced in the body but not relied upon as proved theorems.

## 7 Discussion

### 7.1 What this paper establishes empirically

Three concrete claims come out of the intervention:

- Removing the feedback loop eliminates trajectory-level structure in this mixture system: JSD slope drops to machine zero, regime structure is lost, seed influence decays to nothing.
- Recursive dynamics are necessary for the accumulation of per-step component differences into trajectory-level divergence. The same per-step disagreement is inert without the loop.
- Regime structure is not geometric clustering. $k$-means at chance across every layer is consistent across all 4 seeds tested; regime information is distributed across the residual stream without forming centroid-separable clusters. $k$-NN accuracy rises from 52% (embedding) to 77% (final layer), confirming that regime information accumulates with depth without centroid structure.

[2] Lean file: `IccFormalization/RIS/Cheeger.lean`. Statements formalized; helper lemmas `crossBlockFlow_nonneg` and `conductance_nonneg` are proved. Main Cheeger bounds deferred to Levin–Peres (mathlib4 v4.29.1 lacks the required spectral-gap machinery).

[3] Lean file: `IccFormalization/RIS/BridgeTheorem.lean`. Statements formalized; proofs deferred (mathlib4 v4.29.1 lacks block-matrix operator theory and Riesz projector infrastructure).

These claims are confined to the controlled mixture setting described in Section 2; no generalization beyond this class of constructions is established here. The empirical and formal pieces together form the *Loop Closure Principle* (Section 3.9): a structural property of recursive autoregressive composition, verified for finite discrete recursive systems and instantiated here in a controlled bilingual transformer mixture. The same-language mixture negative control (Section 3.8) provides a clean falsifiability anchor: the mechanism does not fire on recursive feedback alone, only when paired with distinguishable components.

The scorecard (Table 3) records the additional implication for the theory's sufficient-condition package.

### 7.2 Bridge experiment: loop dependence in a trained model

We replicate the cut-the-wire intervention inside the jointly trained bilingual model $M_{\text{ESPT}}$. The intervention is less controlled than in the inference-time mixture: the static distribution here reflects learned priors, not a transparent 50/50 component combination. With that caveat, the pattern transfers quantitatively across 4 independently jointly-trained $M_{\text{ESPT}}$ variants (seeds 42–45).

Recursive generation departs from the seed (fidelity $0.86 \pm 0.06 \rightarrow 0.49 \pm 0.04$), switches regimes (median $2.75 \pm 0.43$ switches per rollout; $94.4\% \pm 2.1\%$ of streams switch at least once), and distributes occupancy ($70.9\% \pm 1.9\%$). Static generation stays seed-locked (fidelity $0.87 \pm 0.02 \rightarrow 0.88 \pm 0.03$, no drop), monolithic ($95.5\% \pm 1.3\%$ occupancy), and largely non-switching (median 0 switches; $27.5\% \pm 9.5\%$ of streams switch). Each metric's standard deviation is computed across seeds.

**Seed replication scope.** These results replicate across 4 independently jointly-trained $M_{\text{ESPT}}$ variants (seeds 42–45; see Table 11). The replication scope is same-composition variants (Spanish–Portuguese). Composition variation (e.g., $M_{\text{ENFR}}$, $M_{\text{PTFR}}$) is pending follow-up (Wave 2) and should not be inferred from these results.

**What this is and is not.** This is not a controlled replication of the mixture experiment. The observation is quantitatively consistent across seeds and replicates the recursive-vs-static contrast; it does not constitute a controlled transfer to a different model class. What it establishes structurally is that the loop-dependence of trajectory structure is not an artifact of the inference-time mixture construction: the same recursive-vs-static contrast manifests inside a jointly trained model whose components are not separable at inference time. This rules out an obvious null hypothesis (that the contrast depends on the externally imposed mixing operation) and supports the broader claim that the loop is the load-bearing structural element across both controlled-mixture and joint-training compositions. This provides consistency evidence for the Principle within a trained model, not a controlled causal transfer across model classes.

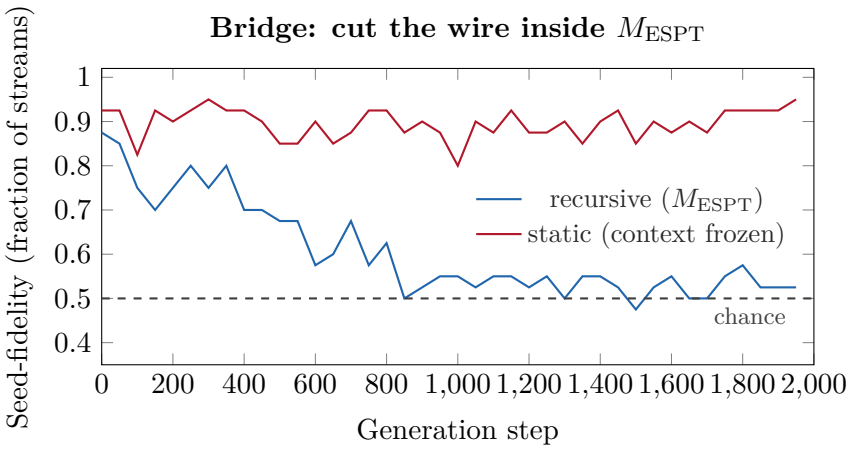


Figure 8: Cut the wire inside $M_{\text{ESPT}}$. Same trained model, same seeds; only context accumulation differs. Recursive generation drifts from the seed regime toward chance; static generation stays locked to the seed.

Static generation exhibits a bimodal pattern: most streams lock completely, while a minority show noisy within-window flips driven by ambiguous seed distributions. We therefore report median and fraction-ever-switched as the primary switching metrics; the mean is pulled up by the noisy subset and is not a faithful summary.

### 7.3 Connection to larger models

The three behaviors documented here are mode persistence (related to degeneration and repetitive generation [4]), invisibility of modes to geometric clustering (consistent with superposition [3], which predicts that features are stored in linear combinations across many neurons rather than dedicated units, producing structure that is linearly decodable but geometrically invisible – matching our observation that $k$-NN rises to 77% while $k$-means stays at chance), and gradual commitment. All three resemble phenomena reported in larger models. Our experiment does not establish that the same mechanism operates at scale. It suggests, tentatively, that mode persistence in autoregressive language models arises from recursive feedback structure independently of training-specific features. Verifying this at scale requires open-weight models and access to internal states.

### 7.4 Scope

The Loop Closure Principle is established here in nanoGPT (small character-level transformer; ∼839K parameters, 4-layer, $d_{\text{model}} = 128$, ctx_len = 256); cross-architecture and cross-scale verification (LSTM/RNN, larger transformers, non-language-domain mixtures) is the natural follow-up program.

All controlled experiments use character-level transformers with ∼839K parameters on closely matched Romance languages. We chose the bilingual setting for its difficulty: overlapping alphabets force distributional separation. The trilingual extension is a consistent extension of the bilingual setup at $n$=3 pairs; it does not establish generalization. We have not tested other language pairs or model sizes.

### 7.5 Limitations

**Mixture weights.** Our controlled mixture uses equal weights. In a real bilingual model, the effective weights are context-dependent; the prompt-sensitive kernel asymmetry we observe is a first indication of this.

**Tokenization.** Subword tokenizers produce a different emission structure. The qualitative conclusion (graded non-identifiability) should transfer; the quantitative spectrum will differ.

**Frozen checkpoints.** The controlled mixture is assembled from frozen monolingual checkpoints and therefore isolates one form of recursive competition: externally defined component models sharing a context. A natively trained bilingual model has richer internal dynamics; $M_{\mathrm{ESPT}}$ serves as a partial check, but the full training-induced interaction is not captured by the mixture construction.

**Assumption (b).** The full typical-set theorem does not apply here: $\Delta_i$ stays negative across all three language pairs we tested (Table 1). Whether the sign ever flips for pairs substantially more distant than ours (e.g., non-Latin alphabets) remains open. What we have shown is that mechanism strength varies continuously with $|\Delta_i|$ even with the sign always negative.

**Model scale.** All controlled experiments use ∼839K-parameter character-level transformers. Superposition may behave differently in models with larger residual streams; the qualitative finding (regimes without basins) is consistent with the superposition literature [3], but we have not verified it at scale. Whether the same mechanism, with the same invariances, operates in production-scale autoregressive models is unverified here. We view architecture-scale validation as the primary direction for future work.

### 7.6 Implications

The results suggest that properties of an autoregressive model may not be fully characterized by its one-step conditional distributions or standard evaluation metrics alone. The amplification hierarchy in this system is a concrete case: one-step conditional divergence between components (0.244 nats) under-represents the trajectory-level divergence ($D_{\mathrm{rate}} > 0$ consistently across the 4 seeds tested; seed-42 value: 0.689 nats; across-seed mean $0.691 \pm 0.021$) that emerges through recursive accumulation, and regime identity built over hundreds of recursive steps is not predictable from any prefix up to 1000 characters. Trajectory-level characterization may be necessary for other autoregressive phenomena; this paper does not determine which.

**Why models persist in modes.** When a language model stays in a generation mode, the loop is a candidate structural mechanism for that persistence, independent of training. Each token biases the context, which biases the next prediction, which biases the next token. Understanding this is a step toward controlling it.

**The hidden state is mostly dark.** Most of the latent state is invisible through the emission map. In this system, output tokens do not convey the regime information present in internal activations. Probing internal representations recovers this information.

**What to measure.** Regime structure can be probed at multiple levels: $\delta_{\mathrm{ker}}$ is measurable from output transitions; $D_{\mathrm{rate}}$ requires component models or an endogenous decomposition; slow spectral structure is recoverable from generation residuals. Geometric clustering found nothing in our setting; it is unlikely to help in the superposition regime.

## 8 Conclusion

In our controlled bilingual mixture, severing the feedback loop eliminated the regime structure entirely. When the loop was restored, the structure returned at every temperature we tested.

Subsequent partial-feedback, content, lag, and asymmetric-mixture sweeps show the loop dependence is invariant across these perturbations within our setup. The components provide the signal; the loop is the mechanism that enables its accumulation in this construction.

The Loop Closure Principle in nanoGPT, as evidenced here, summarizes the empirical content of this paper. Recursive feedback is the necessary enabler: once active and applied to distinguishable components, it converts per-step disagreement into the trajectory-level divergence we measure. In the system we studied, the basin is not a place. It is the trajectory. The Loop Closure Principle is established here as an empirical property of a controlled autoregressive construction; whether comparable invariances hold in larger or qualitatively different systems remains open.

# A Experimental Details

## A.1 Architecture

All controlled experiments use a character-level transformer (NanoGPT [6]): 4 layers, 4 attention heads, $d_{\text{model}} = 128$, context length 256, ∼839K parameters. The vocabulary is the union of characters across the training corpora ($|V| = 118$ for the bilingual pair, $|V| = 126$ for the trilingual extension). Training uses standard cross-entropy on next-character prediction.

## A.2 Models

**Bilingual ($B = 2$).** Three models trained on the same data split: $M_{\text{ES}}$ (Spanish: Galdos, *Fortunata y Jacinta*), $M_{\text{PT}}$ (Portuguese: Eça de Queirós, *Os Maias*), $M_{\text{ESPT}}$ (balanced concatenation). Corpora are matched for register (19th-century literary prose) and length (∼575K characters each).

**Trilingual ($B = 3$).** Adds French (Molière, *Le Bourgeois gentilhomme* and others), yielding $M_{\text{ES}}$, $M_{\text{PT}}$, $M_{\text{FR}}$, $M_{\text{TRI}}$ with $|V| = 126$. The Molière corpus is 17th-century French comedy, distinct in register from the 19th-century prose of Galdós and Queirós; this register mismatch is a known confound that we note but do not control for.

## A.3 Generation protocol

Each experiment begins from a seed context: the first 50 characters of a held-out passage in the target language. Unless stated otherwise, generation length is 2000 characters at temperature $T = 0.9$. All experiments report results over multiple independent seeds.

## A.4 Metrics

**Divergence rate ($D_{\text{rate}}$).**

$$D_{\text{rate}}(P^* \| P_i) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} D_{\text{KL}}\big(P^*(\cdot \mid x_{<t}) \big\| P_i(\cdot \mid x_{<t})\big), \tag{2}$$

estimated as the mean per-character KL over 2000 characters (50 rollouts, 95% CI). All per-step KL terms are evaluated along contexts $x_{<t}$ sampled from $P^*$; the expectation is over the recursive process itself. This per-step form is equivalent to the joint-KL form $\frac{1}{n} D(P^*_{1:n} \| P_{i,1:n})$ used in [5] by the chain rule of KL divergence[4].

**JSD slope.** OLS slope of per-step Jensen–Shannon divergence $\text{JSD}_t$ against $t$, where $\text{JSD}_t = \frac{1}{2} D_{\text{KL}}(P_{\text{ES},t} \| M_t) + \frac{1}{2} D_{\text{KL}}(P_{\text{PT},t} \| M_t)$ and $M_t = \frac{1}{2}(P_{\text{ES},t} + P_{\text{PT},t})$.

[4] Formally verified in Lean; see Appendix C.

**Typed separation parameters.** Following Irrgang [5] (Def typed-sep): $\delta_{\text{pred}}$ (expected per-step TV), $\delta_{\text{ker}}$ (cross-basin transition rate), $\delta_{\text{geom}}$ (minimum cross-basin emission TV).

**Language markers.** Count of unambiguous language-specific character sequences (ñ for Spanish, ã for Portuguese) per 2000 characters. Coherent text: ∼12; incoherent: ∼1.5.

**Regime labeling.** 50-character windows scored by cross-entropy under $M_{\text{ES}}$ and $M_{\text{PT}}$; lower-loss model assigns the label. Validated against three independent methods (Section 5).

**Same-language negative control.** The same-language mixture (Section 3.8) uses two $M_{\text{ES}}$ checkpoints trained on the identical Spanish corpus at random seeds 42 and 43 as components $a$ and $b$. Generation, scoring, and the recursive-vs-static contrast follow the same protocol as the bilingual cut-the-wire experiment (Section 3), with gen_len = 2000 and 4 streams (2 ES seed contexts × 2 random seeds for stochastic sampling).

# B Full Numerical Results

## B.1 Recursion versus static

Table 4: Recursion versus static mixture generation ($B = 2$; seed 42). $D_{\text{rate}}$ values here are seed-42 specific; across-seed mean is $0.691 \pm 0.021$.

| | **Recursive $P^*$** | **Static $P^*$** | **Component** |
|---|---|---|---|
| JSD slope | $1.3 \times 10^{-6}$ | $-2.9 \times 10^{-20}$ | $6.1 \times 10^{-6}$ |
| $D_{\text{rate}}$ (avg, nats/char) | 0.689 | 1.563 | 0.541 |
| Language markers / 2000 chars | 13.4 | 1.5 | 11.5 |

## B.2 Temperature sweep

Table 5: Recursion versus static across temperatures.

| $T$ | Recursive $D_{\text{rate}}$ | Static $D_{\text{rate}}$ | Static JSD slope | Recursive markers |
|---|---|---|---|---|
| 0.3 | 2.66 | 1.14 | ∼0 | 12.3 |
| 0.6 | 1.24 | 0.67 | ∼0 | 11.8 |
| 0.9 | 0.80 | 0.45 | ∼0 | 12.3 |
| 1.2 | 0.56 | 0.36 | ∼0 | 15.2 |

## B.3 Typed separation

Table 6: Typed separation parameters ($B = 2$). The k-means purity reported below is an operational proxy, not an estimator of the formal $\delta_{\text{geom}}$; it is reported separately for comparison.

| Parameter | Value | Meaning |
|---|---|---|
| $\delta_{\text{pred}}$ | 0.484 | $\mathbb{E}_{\pi^*}[\text{TV}(P_{\text{ES}}, P_{\text{PT}})]$ |
| $\delta_{\text{ker}}$ | 0.0016 | Cross-basin transition rate (corpus) |
| $\delta_{\text{geom}}$ (formal) | 0.015 | Min cross-basin emission TV |
| k-means purity | 52.6% | Operational proxy (class balance 50/50) |

### B.4 Layer-by-layer analysis

Table 7: Regime detection by layer. Probe column omitted; see Section 4.2 for explanation.

| Layer | k-means | k-NN(5) | k-NN(20) |
|---|---|---|---|
| embed | 50.0% | 52.2% | 50.4% |
| layer 0 | 53.6% | 53.6% | 50.4% |
| layer 1 | 52.0% | 61.9% | 59.2% |
| layer 2 | 52.6% | 71.6% | 67.5% |
| layer 3 | 52.4% | 75.1% | 67.7% |
| final+LN | 53.6% | 77.0% | 69.5% |

### B.5 Singular spectrum

Of 118 emission dimensions: 0 fully observable (TV $> 0.10$ under unit perturbation), 60 weakly observable (TV 0.01–0.10), 58 effectively dark (TV $< 0.01$), 10 in $\ker(W_u)$. Observability cliff at $\sigma = 0.42$ (SV index 57). Fisher eigenvalues correlate with $\sigma_k^2$ at $r = 0.972$.

### B.6 Amplification hierarchy

Table 8: Three-level amplification.

| Level | Quantity | Value (nats) | Amplification |
|---|---|---|---|
| Unigram | JSD(char freqs) | 0.019 | (baseline) |
| Conditional | $\mathrm{JSD}(P_{\mathrm{ES}}, P_{\mathrm{PT}} \mid x)$ | 0.244 | 13× |
| Trajectory | $D_{\mathrm{rate}}(P^* \| P_i)$ | 0.689 | 2.9× |

Crossover length: the smallest $n$ at which $\log P^*(x_{1:n}) - \max_i \log P_i(x_{1:n}) \geq \log 100$ for sequences $x_{1:n}$ generated by the respective mode. Result: 7 characters under recursive generation, 19 characters under static generation.

### B.7 Commitment and hazard

Mean $t^* = 24.4$ windows (1218 chars), median 28.5. Range: $t^* = 0$ to $t^* = 39$. Hazard: constant (short/long ratio = 0.98). KS test $p = 0.54$. ACF half-life: 203 chars.

### B.8 Coarse kernel

ES-prompted: ES→PT = 0.052, PT→ES = 0.091. $\lambda_2 = 0.834$. Dwell (kernel-derived, geometric run-length convention with support $\{1, 2, \ldots\}$, mean $1/p_{\mathrm{switch}}$): ES ≈ 960 chars ($1/0.052 = 19.2$ windows), PT ≈ 550 chars ($1/0.091 = 11.0$ windows).

### B.9 Triangulation

Table 9: Pairwise agreement among recovery methods. Four methods: external scorer, TICA, LLR, and endogenous HMM (requires no external reference models).

| Method pair | Agreement | $\kappa$ |
|---|---|---|
| Scorer vs. TICA | 95.0% | 0.894 |
| Scorer vs. HMM | 97.9% | — |
| TICA vs. HMM | — | — |
| LLR vs. others | 74.5–76.5% | 0.47–0.51 |

3-of-4 agreement: 97.2%. TICA eigenvalue (lag 200 chars): 0.496. Coarse-kernel prediction: $\lambda_2^4 = (0.834)^4 \approx 0.484$. Endogenous HMM: 97.9% agreement, $D_{\text{rate}} = +0.174$ and $+0.338$ in both states.

### B.10 Trilingual ($B = 3$)

Table 10: Cross-evaluation advantage ($B = 3$).

| Pair | Unigram TV | Cross-eval advantage |
|---|---|---|
| ES–PT | 0.093 | +1.87 nats/char |
| ES–FR | 0.186 | +2.87 nats/char |
| PT–FR | 0.199 | +2.87 nats/char |

$D_{\text{rate}}$ (3-way recursive): ES 1.19, PT 1.21, FR 1.34. Static JSD slope: $8.4 \times 10^{-20}$ (machine zero). Slow eigenvalues: $\lambda_2 = 0.901$, $\lambda_3 = 0.749$. $\delta_{\text{pred}} = 0.499$ (min pair ES–PT).

### B.11 Scope results

Assumption (b): $\Delta_{\text{ES}} = -0.139$, $\Delta_{\text{PT}} = -0.212$ (both negative; condition not applicable to this language pair). Bridge theorem: $\varepsilon = 0.999$, $\eta = 0.99$ (not applicable; designed for separated alphabets). Transition regularity: of 4,986 candidate pairs sampled from 500 hidden states per language, 25 were identified as *aliased*: pairs $(h, h')$ whose output distributions satisfied $\text{TV}(\varphi(h), \varphi(h')) < 0.01$. For each aliased pair, 10 tokens were drawn from $\varphi(h)$ and re-injected; a pair was counted as *remaining distinct* if $\|f(h, c) - f(h', c)\|_2 > 10^{-6}$ for every sampled token $c$. All 25 pairs remained distinct on all tokens (mean $\|\Delta h_{t+1}\|_2 = 16.3$), giving 100% regularity. Seed fidelity: recursive $0.95 \to 0.35$; static $0.60 \to 0.00$.

### B.12 Seed replication (Wave 1)

Table 11: Per-seed $D_{\text{rate}}$ results across 4 $M_{\text{ESPT}}$ variants (Wave 1). Cross-eval advantage is the per-step log-likelihood advantage of the correct component over the wrong one. $k$-means purity is the best purity across layers. All seeds complete (Wave 1). Seed 42 values are from the long-baseline `drate_large` protocol (50 rollouts, $D_{\text{rate}}$ averaged); seeds 43–45 are from the Wave 1 protocol (4 streams per seed, linear-fit $D_{\text{rate}}$). The cross-seed claim is qualitative consistency (recursive $D_{\text{rate}}$ positive, static slope at machine zero across all four seeds), not magnitude equality; the per-seed $D_{\text{rate}}$ values from the two protocols are not directly commensurable.

| Seed | $D_{\text{rate}}^{\text{ES}}$ | $D_{\text{rate}}^{\text{PT}}$ | Cross-eval ES | Cross-eval PT | $k$-means best |
|---|---|---|---|---|---|
| 42 (original) | 0.705 | 0.673 | +1.87 | +1.87 | 52.6% |
| 43 | 0.715 | 0.703 | +1.819 | +1.915 | 51.0% |
| 44 | 0.707 | 0.670 | +1.769 | +1.920 | 51.6% |
| 45 | 0.707 | 0.682 | +1.823 | +1.929 | 51.6% |
| Mean (42–44) | $\approx 0.704$ | $\approx 0.687$ | — | — | $\approx 51.7\%$ |

Table 12: Per-bilingual-pair $D_{\text{rate}}$ from Wave 1 (seed 43, h3a protocol; 20 rollouts $\times$ 2000 chars, linear fit). $M_{\text{ENFR}}$ probes cross-family composition (Germanic + Romance); $M_{\text{PTFR}}$ re-tests within-family separation at a different pair. The per-pair $D_{\text{rate}}$ magnitudes order by component-pair distinguishability (per-step predictive TV $\delta_{\text{pred}} \approx 0.50$, 0.62, 0.63 for ES–PT, PT–FR, EN–FR respectively). Rate-invariance holds within a pair; the plateau height scales with component distinguishability across pairs.

| Model / Pair | Seed | $D_{\text{rate}}^{\text{L1}}$ | $D_{\text{rate}}^{\text{L2}}$ |
|---|---|---|---|
| $M_{\text{ENFR}}$ | 43 | 1.291 | 1.220 |
| $M_{\text{PTFR}}$ | 43 | 1.135 | 1.020 |

## C Formal Verification Artifacts

These formal artifacts verify four algebraic identities used in the body of the paper, and establish one structural-independence result in a finite-discrete recursive system; they do not constitute theorems about transformers.

Seven formal artifacts accompany this paper, verified in Lean 4 using the Mathlib4 library (v4.29.1), with a SymPy symbolic companion suite included alongside. All Lean artifacts are available in the workspace `~/lean-workspace/icc-formalization/`, buildable with `lake build`. Five files carry fully proved theorems with zero `sorry`. Two additional files carry statements only; proofs are deferred because mathlib4 v4.29.1 lacks the required Markov-chain spectral gap machinery (detailed below).

**DrateIdentity.lean (122 lines).** Verifies the per-step mean KL identity used in Appendix A:

$$\frac{1}{n} D_{\text{KL}}(P^*_{1:n} \| P_{i,1:n}) \;=\; \frac{1}{n} \sum_{t=1}^{n} D_{\text{KL}}\big(P^*(\cdot \mid x_{<t}) \big\| P_i(\cdot \mid x_{<t})\big). \tag{3}$$

The proof proceeds by the chain rule of KL divergence applied to a finite product distribution over the sequence $(x_1, \ldots, x_n)$. This justifies treating the empirical per-step mean KL as an unbiased estimator of the joint-KL normalized rate $D_{\text{rate}}$.

**LoopClosureIndependence.lean (716 lines).** Provides three independence witnesses showing that loop-driven divergence accumulation is structurally independent of the RIS sufficient conditions: one witness violating hard token partition (`IsRISHardPartition`), one violating geometric basin separation (`IsRISBasinGeometric`), and one violating a weight-existence proxy (`IsRISDeltaSimpleProxy`) for the paper's $\Delta_i > 0$ condition. The Bool/finite-weight encoding does not capture the probability-weight $\Delta_i$ argument directly; the proxy encodes the weaker requirement that each regime state has at least one high-weight emission token, and the gap is documented in in-file commentary. Each witness is a finite discrete recursive system in which Loop Closure-shaped divergence behavior holds while the corresponding sufficient condition is explicitly violated. Together, the three witnesses establish that recursive feedback alone is sufficient to produce persistent divergence in the abstract system even when none of the three sufficient conditions is satisfied.

**AmGmBound.lean (57 lines).** Verifies the AM–GM bound used in the sequential amplification proof: for all $\varepsilon \in \mathbb{R}$, $\varepsilon(1-\varepsilon) \leq 1/4$, with equality iff $\varepsilon = 1/2$. The proof is the algebraic identity $1/4 - \varepsilon(1-\varepsilon) = (1-2\varepsilon)^2/4 \geq 0$. This bound is the structural reason the amplification base $1/(2\sqrt{\varepsilon(1-\varepsilon)})$ is at least one, with strict inequality away from $\varepsilon = 1/2$.

**AmplificationRatio.lean (142 lines).** Verifies the closed-form sequential amplification ratio for the binary alternating-dominance setting: on a chain of length $n = 4$, the per-component product equals $[\varepsilon(1-\varepsilon)]^{n/2}$, the mixture product equals $(1/2)^n$, and the ratio equals $\left(1/(2\sqrt{\varepsilon(1-\varepsilon)})\right)^n$. The companion corollary establishes the amplification base is strictly greater than one for $\varepsilon \neq 1/2$.

**PositiveDivergence.lean (342 lines).** Provides a Lean-verified proof of the positive-divergence lemma (referenced as *Lem positive-div* in Table 3): under $\delta$-separation with $\delta < 1/2$ and $B \geq 2$ experts, the relative entropy rate satisfies $D_{\text{rate}}(P^* \| P_i) > 0$ for every component $i$. The proof composes the discrete Gibbs inequality ($D_{\text{KL}} \geq 0$, strict when $p \neq q$) with a mixture-versus-component non-equality lemma showing that $\delta$-separation forces $P^* \neq P_i$ at every state.

**Cheeger inequality (helper lemmas proved; main theorems deferred).** The Cheeger inequality applied in Table 3 is formalized in `IccFormalization/RIS/Cheeger.lean` (336 lines). The file defines stochastic kernels, stationary distributions, conductance, and an opaque spectral-gap constant, then states the upper bound ($\text{gap}(K) \leq 2\Phi$), the lower bound ($\Phi^2/2 \leq \text{gap}(K)$), the combined sandwich, and the small-conductance corollary.

Two supporting lemmas are fully proved (zero `sorry`): `crossBlockFlow_nonneg` (cross-block flow is non-negative) and `conductance_nonneg` (conductance is non-negative). These establish non-negativity of the key quantities and serve as the interface between the definitions and the theorem statements.

The four main Cheeger theorems (`cheeger_upper`, `cheeger_lower`, `cheeger_sandwich`, `small_conductance_implies_small_gap`) are deferred (`sorry`): mathlib4 v4.29.1 does not provide the discrete Cheeger cutting lemma, the Rayleigh-quotient characterization of spectral gap for finite stochastic matrices, or the Hermitian-symmetrization route needed for reversible kernels.

**Approximate-lumpability bridge theorem (statement only).** The approximate-lumpability result used in Section 6 is formalized at statement level in `IccFormalization/RIS/BridgeTheorem.lean` (358 lines). Definitions cover block projections, the macro kernel, the invariance defect, and approximate lumpability. Statements include the three supporting lemmas (`macro_spectrum_identification`, `invariance_defect_bound`, `projector_stability`) and the main theorem (`approx_lump_preserves_slow_se`). Proofs are deferred: the required infrastructure (block-matrix operator theory, Riesz projectors

via contour integration, and non-Hermitian eigenvalue continuity for $B \times B$ matrices) is absent from mathlib4 v4.29.1. The block-triangular spectrum splitting (conclusion (i) of the source theorem) is documented as a gap in the file because mathlib4 lacks the resolvent API; full proof is a multi-month formalization target.

**SymPy companion: `basin_sympy.py` (153 lines).** A symbolic verification suite for five algebraic identities used in the body and proofs: (A) the Jensen-gap form of the Jensen–Shannon divergence, with KL non-negativity proved via calculus on $f(x) = x \log(x/y) - (x - y)$; (B) the AM–GM bound $\varepsilon(1-\varepsilon) \leq 1/4$ (also formally proved in `AmGmBound.lean`); (C) the sequential amplification ratio $(1/(2\sqrt{\varepsilon(1-\varepsilon)}))^n$ for $n = 4$ (also formally proved in `AmplificationRatio.lean`); (D) the crossover-length formula $n^* = \lceil \log K / \log(1/(2\sqrt{\varepsilon(1-\varepsilon)})) \rceil$; (F) the upper bound $\mathrm{JSD}_{\boldsymbol{\pi}} \leq H(\boldsymbol{\pi}) \leq \log B$. Run from the bundle as `python3 sympy/basin_sympy.py`; all five checks pass.

**Scope disclaimer.** The Lean predicates are over a finite discrete recursive system; no claim is made that autoregressive transformers satisfy these predicates. The independence result is a structural statement about the abstract framework, not a theorem about any particular language model. The Cheeger and bridge-theorem formalizations establish only that the definitions are type-correct and the theorem statements are well-formed; they do not constitute proofs of those results.

## References


[1] Guillaume Alain and Yoshua Bengio. "Understanding Intermediate Layers Using Linear Classifier Probes". In: *arXiv preprint arXiv:1610.01644* (2017).

[2] Jeff Cheeger. "A Lower Bound for the Smallest Eigenvalue of the Laplacian". In: *Problems in Analysis* (1970), pp. 195–199.

[3] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. "Toy Models of Superposition". In: *Transformer Circuits Thread* (2022).

[4] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. "The Curious Case of Neural Text Degeneration". In: *International Conference on Learning Representations*. 2020.

[5] Claudio Irrgang. *Recursive Interaction Systems: Non-Identifiability, Emergence, and Metastability*. Zenodo. v3. 2026. DOI: 10.5281/zenodo.20060237. URL: https://doi.org/10.5281/zenodo.20060237.

[6] Andrej Karpathy. *nanoGPT*. https://github.com/karpathy/nanoGPT. 2022.

[7] David A Levin and Yuval Peres. *Markov Chains and Mixing Times*. 2nd. American Mathematical Society, 2017.

[8] Lutz Molgedey and Heinz Georg Schuster. "Separation of a Mixture of Independent Signals Using Time Delayed Correlations". In: *Physical Review Letters* 72.23 (1994), pp. 3634–3637.