UNDERSTANDING THE PREDICTIVE-STATE BOTTLENECKS OF TRANSFORMER MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence-prediction tasks typically require compressing an observation history into a compact internal state, making the state's dimensionality and organization central to neural model capacity. We study whether the residual width required for accurate prediction is governed by the latent-state count, the full belief dimension, or the intrinsic dimension of the predictive state, using Transformers trained on hidden Markov processes. Across controlled families that separate these quantities, the width transition tracks predictive-state rank. Exact filtering constructions and bottleneck results connect this transition to the geometry of predictive representations, while readout and architectural controls distinguish state capacity from output parameterization. We also observe that training is spectrally ordered, and a probe-calibrated analysis reveals a width-dependent mechanism: successive Transformer blocks increasingly favor high-variance belief directions; below rank, this becomes suppression of low-variance modes beyond an equal-treatment probe null, while sufficiently wide, adequately fitted models reverse the bias. Causal interventions separately show that the represented belief subspace contributes to prediction. In the tested contractive regime, filtering is strongly width-sensitive and only weakly depth-sensitive.

1 INTRODUCTION

Sequence prediction under partial observability requires constructing an internal state from the observation history. In a hidden Markov model (HMM), the posterior belief $b_t^+ = P(z_t \mid x_{1:t})$ is sufficient but may not be minimal for Bayes-optimal prediction: distinct latent states and distinct belief directions can induce identical distributions over futures. Predictive-state representations, observable-operator models, and reduced-rank system identification characterize the minimal exact *linear* realization, whose dimension is the Hankel rank r (Jaeger, 2000; Littman et al., 2001; Hsu et al., 2012; Siddiqi et al., 2010). Computational mechanics provides a complementary, representation-independent perspective, identifying causal states as predictive equivalence classes of histories (Shalizi & Crutchfield, 2001). Both traditions separate the intrinsic predictive structure of a process from any particular latent-state realization of it.

A second line of work asks what a compressed state must retain when exact state is too costly. Belief-compression methods exploit low-dimensional structure in the reachable belief set, or retain only the information needed for future decisions (Roy et al., 2005; Poupart & Boutilier, 2002). A parallel line of work uses predictive rate–distortion, information-bottleneck, and compressed-PSR methods to balance representational capacity against the preservation of future-relevant information or predictive accuracy (Tishby et al., 2000; Marzen & Crutchfield, 2014; Hamilton et al., 2013). In control theory, balanced model reduction ranks directions by their joint controllability and observability, prioritizing input–output relevance over state variance (Moore, 1981; Gugercin & Antoulas, 2004). These results specify which directions a compact state should keep, but do not specify how many a learned neural sequence model must allocate. In particular, they do not predict the residual width at which a trained model reaches Bayes-optimal prediction.

Neural approaches provide complementary evidence that such states can emerge from prediction and control objectives. Predictive-state recurrent networks, predictive belief representations, and belief-aware reinforcement-learning methods explicitly encourage recurrent activations to summarize future-relevant information (Downey et al., 2017; Hefny et al., 2018; Guo et al., 2018; Wang et al., 2023). More recent mechanistic studies find belief geometry and architecture-constrained

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

filtering computations in Transformers trained on HMMs, while work on automata and permutation composition shows that Transformers can learn several distinct mechanisms for discrete state tracking (Shai et al., 2024; Piotrowski et al., 2025; Hu et al., 2024; Li et al., 2025; Zhang et al., 2025). These studies leave open which process-level dimension determines the width needed for near-Bayes-optimal prediction when hidden-state count, belief dimension, and predictive rank differ.

A related question concerns how limited capacity is allocated once it binds. Classical model-reduction and weighted low-rank methods prioritize directions according to their effect on observable behavior. In contrast, analyses of neural optimization typically report sequential acquisition of dominant, low-frequency, or high-singular modes (Kulesza et al., 2015; Saxe et al., 2014; Rahaman et al., 2019; Gidel et al., 2019; Boix-Adsera et al., 2023). Therefore, the representation reached by gradient descent need not coincide with the one a prediction-weighted compression objective would select.

Here we study these questions in controlled HMMs where predictive rank, posterior geometry, and Bayes-optimal loss are known in closed form. We make three contributions. First, we separate predictive rank from both hidden-state count and full belief dimension; in the strongest state-count control, a 128-state clone HMM is an exact state-splitting realization of the same observable process as a 16-state ring, and the width transition is unchanged. Second, we connect this empirical transition to predictive-state theory through an exact rank-r homogeneous filter, a tight $(r-1)$-dimensional continuous bottleneck, exact-state readout calibration, and architectural controls. Third, we distinguish spectral acquisition from final allocation. Belief modes are learned from higher to lower occupancy variance, but much of the raw correlation between variance and decodability is induced by the probe. Relative to an equal-treatment null, Transformer blocks progressively build a preference for high-variance modes; this preference weakens monotonically with width and reverses above rank in adequately fitted models. Causal interventions show that the represented subspace contributes to prediction. Predictive rank is process-invariant, whereas the spectral results depend on the predictive-belief coordinates of the chosen HMM realization.

2 Theory

2.1 Data generation model

An HMM with K states, transition matrix A, emission matrix B, and initial distribution π induces a belief process on the $(K-1)$-simplex. We write $b_t = P(z_t \mid x_{1:t-1})$ for the predictive belief, evaluated before observing x_t, and b_t^+ for the post-observation posterior. All spans, spectra, and per-mode targets below use b_t. Let $\mathcal{R} = \mathrm{span}(\mathrm{supp}(b_t))$ denote its stationary reachable belief span. The predictive functional maps a state to the distribution over futures it induces; the Hankel rank r is the dimension of the span of these functionals, equivalently the minimal dimension of an exact linear OOM/PSR realization (Jaeger, 2000). We use r as an invariant process-level reference rather than a lower bound on arbitrary nonlinear representations. The predictive functional is a linear image of the belief, so its rank is bounded by the dimension of the reachable-belief set: $r \le \dim \mathrm{span}\{b_t\} = \mathrm{affdim}(b_t) + 1$. Hidden-state count, causal-state cardinality, and reachable-belief dimension are distinct quantities that need not coincide, and it is the last of these that bounds the readout. Equality holds when the predictive map is injective on the reachable span, as in the ring and mixing clone corpora, but deliberately fails in the rank–span corpus, where the full belief span has dimension K, and the predictive quotient has rank g.

Unlike Hankel rank, belief-space spectra and Euclidean projections refer to the fixed generative HMM realization, so they are not invariant under reparameterization. The band-limited construction separately has finite predictive dimension but an uncountable causal-state set, as formally stated in Appendix A.1.

2.2 A training-free exactness construction at the predictive rank

Since neural width sweeps entangle representational capacity with optimization, we construct a d-dimensional projected belief recursion whose state dimension is fixed analytically rather than learned. Throughout this subsection, beliefs are column vectors, A is row-stochastic, and $B \in \mathbb{R}^{K \times V}$ contains the emission probabilities.

Proposition 1 (d-dimensional projected recursion, exact under containment)**.** *Let b_t be the exact predictive belief before observing x_t, let $C = \mathbb{E}[b_t b_t^\top]$, and let $R = U_d \in \mathbb{R}^{K \times d}$ contain the top d*

eigenvectors of C. Initialize $z_0 = R^\top \pi$. The implemented recursion is

$$\widetilde{\alpha}_t = R z_t, \qquad m_t = \mathbf{1}^\top \widetilde{\alpha}_t, \qquad \alpha_t = \widetilde{\alpha}_t / m_t, \qquad z_t \leftarrow z_t / m_t, \tag{1}$$

$$p_t = \frac{[B^\top \alpha_t]_+}{\mathbf{1}^\top [B^\top \alpha_t]_+}, \tag{2}$$

$$z_{t+1} = R^\top A^\top \big(\alpha_t \odot B_{:,x_t}\big), \tag{3}$$

where $[\cdot]_+$ denotes elementwise clipping at zero. If $\mathrm{span}(R)$ contains the stationary reachable belief span $\mathcal{R}$, then $\alpha_t = b_t$ on every positive-probability trajectory, clipping is a no-op, and p_t is the exact Bayes next-token distribution. Consequently, the recursion attains the Bayes floor. In the injective (mixing) clone blow-up family, $\mathrm{rank}(C) = r = g$, so exactness holds for every $d \geq r$.

For $d < r$, projection error is fed back through equation 3 and compounds over time. The resulting loss is therefore the loss of this prescribed closed-loop recursion, not the optimum over all nonlinear width-d predictors and not the one-step oracle truncation used in the rate–distortion analysis.

The recursion remains strictly above the floor at $r - 1$ and reaches it at r on every injective (mixing) clone corpus when evaluating at both $d = r - 1$ and $d = r$. It therefore establishes constructive sufficiency at predictive rank for this prescribed homogeneous recursion; the Transformer sweeps independently place the empirical transition near the same process dimension.

The rank–span corpus exposes the limits of choosing $R = U_d$ by occupancy variance. At $d = g = 16$, the top-g eigenvectors leave 0.107 bits of excess loss, whereas a separate g-dimensional macro-marginal recursion is exact. Proposition 1 guarantees full-belief reconstruction whenever the projected space contains the reachable span, but it says nothing about identifying a predictive quotient when the realization is non-injective.

The mixing clone beliefs have affine dimension $g - 1$: an exact filter can update $g - 1$ live coordinates and restore a fixed mean before prediction. The formal construction is in Appendix A.1; Table 6 gives the numerical comparison.

Theorem 1 (Continuous predictive bottleneck lower bound)**.** *Let $\mathcal{U}$ be a nonempty relatively open subset of an m-dimensional affine space, and let $\Phi : \mathcal{U} \to \mathbb{R}^q$ be injective. If exact prediction factors as $\Phi = D \circ E$ with continuous $E : \mathcal{U} \to \mathbb{R}^d$, then $d \geq m$.*

Corollary 1 (Tight dimension for mixing clone processes)**.** *For a mixing clone–blow-up process of predictive rank r, considered over an open family of normalized predictive states, every exact continuous single-vector predictive bottleneck requires $d \geq r - 1$. Proposition 3 attains $d = r - 1$, so $d_{\min}^{\mathrm{exact}} = r - 1$.*

Across $g \in \{8, 16, 32, 64\}$, future-test certificates verify a homogeneous rank r and a simplex-tangent rank $r - 1$, and the affine filter is exact to numerical precision over long rollouts from interior priors (Appendix A.2). The theorem, however, concerns a single-vector predictive bottleneck rather than an unrestricted Transformer with distributed prefix memory.

2.3 Learning-dynamics conjectures

We test whether belief modes are acquired in decreasing eigenvalue order for $C = \mathbb{E}[b_t b_t^\top]$. A deep-linear surrogate associates singular scale $s_i = \sqrt{\lambda_i}$ with $t_{90}(i) \propto s_i^{-p} = \lambda_i^{-p/2}$, where $p = 1$ is the inverse-singular-value law and $p = 2$ the inverse-variance law. Fits use only modes whose final decodability clears the stated threshold; the unresolved tail is censored. The correspondence is empirical and coordinate-dependent, with full conventions in Appendix A.1.

3 Methods

3.1 Corpora.

We train decoder-only Transformers from scratch on token streams from several known HMMs of known predictive rank r, as follows:

Ring: an incompressible cycle in which every state is predictively distinct, so $r = K$.

Mixing clone blow-up: a g-state macro ring lifted to $K = g \cdot c$ physical states by splitting each macro state into c clones with identical emissions and clone-independent transitions. The reachable belief

has affine dimension $g - 1$ and homogeneous span dimension g, and the predictive map is injective on this span, so $r = g$. We sweep $g \in \{8, 16, 32, 64\}$ at fixed $K = 128$.

Rank–span corpus: we keep the same macro process but replace clone mixing with identity transitions and correlate the clone index with the initial macro state. Because the clone index never affects emissions or macro transitions, it adds no predictive functionals and $r = g$. It nevertheless persists in the belief: observations update clone probabilities through the initial macro–clone correlation, producing a history-dependent, non-predictive belief component. For $g = 16$, $c = 8$, and $K = 128$, Kronecker and rank checks give $\dim \mathrm{span}\{b_t\} = K$, while exact macro-marginal prediction verifies $r = g = 16$. This corpus therefore separates predictive rank from the dimension occupied by the full belief.

Band-limited ring: $K = 128$ states on a cycle with circulant Fejér transition and emission kernels of bandwidth $b \in \{2, 4\}$ (alphabet size $V = K$). A uniform positive transition leak makes the chain strictly positive without changing its nonzero Fourier support. The predictive mixed states lie in a homogeneous Fourier span of dimension $2b + 1$ (e.g., 5 at $b = 2$, 9 at $b = 4$), and the emission map is injective on this span, but the observation-conditioned filter has an uncountable causal-state attractor. The beliefs b_t^+ have a larger centered covariance rank (16 at $b = 4$, numerically), but this is not the Hankel rank: conditioning introduces directions that the subsequent transition removes. An oracle truncation reaches the empirical loss floor at roughly five modes. Because the circulant transition and emission kernels are simultaneously diagonalized by the Fourier basis, the occupancy and one-step Fisher orderings coincide, so this corpus cannot test a variance–importance conflict. It therefore supplies a causal-state separation and a graded-spectrum control rather than a sharp width sweep; proofs and details are in Appendix A.1.

3.2 Training and probes.

We vary width d_{model} and depth n_{layers} independently, holding the process fixed; $n_{\text{layers}} = 4$ for the width sweeps. We report the best value attained over training across three seeds. Recovery is measured by excess loss above the Bayes floor and belief R^2, the fraction of belief variance linearly decodable from the residual.

All occupancy spectra and per-mode targets use b_t rather than b_t^+. We form $C = \mathbb{E}[b_t b_t^\top] = \sum_i \lambda_i e_i e_i^\top$ and regress each mode $e_i^\top b_t$ from residual activations at log-spaced checkpoints. Acquisition time t_{90} is the step at which a mode reaches 90% of its final decodability. The b_t versus b_t^+ distinction is immaterial on the clone corpora but changes the band-limited occupancy rank.

On the rank–span corpus, we use excess prediction loss as the primary metric. Full-belief R^2 includes $K - r = 112$ deliberately non-predictive clone-shape directions, so low full-belief recovery is expected from a correct predictive representation.

3.3 Architectural control.

To test whether output-side factors are confounded with width, we rerun the $g = 16$ width sweep with untied input and output weights, a learned readout bias, and an MLP hidden dimension fixed at $d_{\text{ff}} = 256$ across all residual widths. This rules out weight tying, the bias-free readout, and width-dependent MLP size as explanations for the location of the wall. It does not, however, remove the rank bound imposed by a linear readout: any map $\mathbb{R}^d \to \mathbb{R}^V$ has rank at most d, whether tied or untied.

3.4 Exact-state readout calibration.

For $g = 16, 32$, we fit output heads directly to the exact macro predictive state using homogeneous coordinates $m \in \mathbb{R}^r$, affine coordinates $y = U^\top(m - \mu) \in \mathbb{R}^{r-1}$, affine coordinates augmented with a constant, and a random invertible reparameterization of m. We compare bias-free linear-softmax, biased linear-softmax, and two-layer GELU heads. The experiment uses exact Bayes targets and fixed held-out positions to estimate output-class approximation from a known sufficient state; numerical guards, convergence checks, and full implementation details are reported in Appendix A.12.

3.5 Analytic and causal analyses.

The training-free predictor implements Proposition 1: it is exact once the chosen subspace contains the reachable belief span. On the rank–span corpus, we instead use a separate exact recursion on the macro marginal. For one-step analyses, we project predictive beliefs through $p_t = B^\top b_t$ and score them in that basis; the map AB applies to b_t^+ and is not used here. These predictors restart from the exact belief at every position. Therefore, they do not measure recursive filtering error. We test causal relevance by resampling the decoded belief subspace at the readout residual and comparing it with a dimension-matched random subspace.

3.6 Probe-calibrated variance selectivity.

For each seed and residual location, we summarize per-mode decodability by

$$\beta = \mathrm{slope}_i\left(\log \frac{R_i^2}{1-R_i^2} \text{ against } \log \lambda_i\right).$$

The odds ratio is used because $1 - R_i^2$ is bounded above by one, censoring a deficit slope when tail modes are poorly decoded. Under an equal-treatment ridge-probe null, $1 - R_i^2 \propto 1/\lambda_i$, so $\beta = 1$ analytically. Thus $\beta > 1$ indicates additional suppression of low-variance modes, whereas $\beta < 1$ indicates relative amplification. Slopes are fitted per seed and then averaged, using all modes. A random equal-treatment embedding with isotropic noise returns $\beta = 0.89$–1.07 across the grid, validating the analytic reference. Because the empirical offset depends weakly on the null feature-scale convention, we report β as the primary statistic and bracket $\beta - \beta_{\mathrm{null}}$ across standardized and raw-scale nulls. Full protocols and robustness checks are in Appendix A.3.

4 Results

4.1 Width is the dominant bottleneck in the tested filtering regime

Belief tracking exhibits a width–depth asymmetry. The gap and belief recovery change sharply as residual width crosses the predictive rank (or K for the ring) and then plateau (Fig. 1A–B). Depth is far less demanding: in the hard-K family (Appendix A.15), two layers suffice for $K = 20, 40, 80$ and four for $K = 160$, far below a literal one-filter-step-per-layer implementation (§A.9).

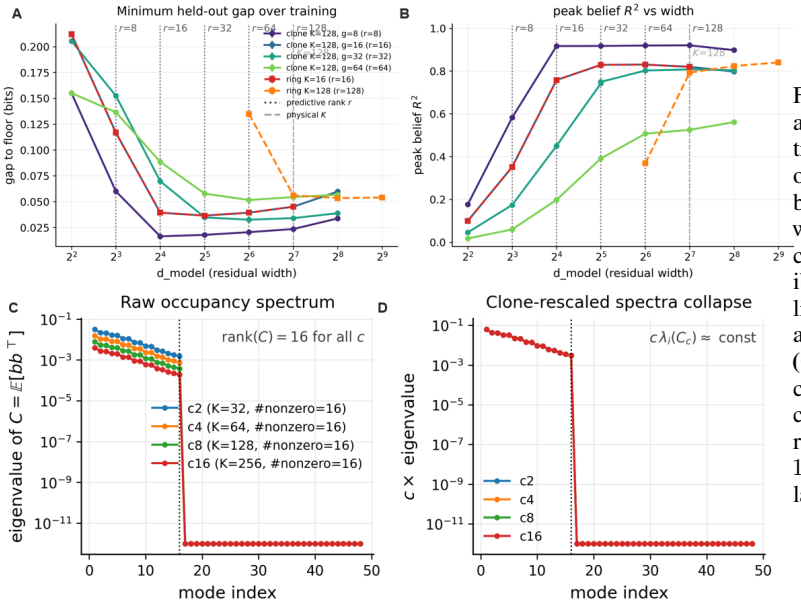

Figure 1: Width transitions and clone occupancy spectra. **(A–B)** Minimum held-out excess loss and peak belief R^2 versus residual width for mixing clone processes and the incompressible $K = 128$ ring; vertical lines mark predictive rank r and physical state count K. **(C–D)** At fixed $g = 16$, increasing clone multiplicity changes the scale but not the rank of $C = \mathbb{E}[b_t b_t^\top]$: the 16 nonzero eigenvalues collapse after rescaling by c.

4.2 Contraction explains the depth insensitivity

Contractive filtering forgets old evidence geometrically (Atar & Zeitouni, 1997), so fixed-accuracy prediction depends only on a bounded recent window that attention can aggregate in a few layers.

Residual width d	Excess loss (bits)	Full-belief R^2
8	0.216	0.064
16 ($= r$)	0.062	0.145
24	0.058	0.155
32	0.056	0.158
128 ($= \dim \mathrm{span}\{b_t\}$)	0.061	0.159

Table 1: Rank–span separation ($g = r = 16$, $c = 8$, $K = \dim \mathrm{span}\{b_t\} = 128$). The prediction gap drops sharply at r and remains near-flat through the full belief-span dimension. Full-belief R^2 stays low because 112 belief directions are deliberately non-predictive.

This explains the width–depth asymmetry. Appendix A.10 contrasts a non-forgetting composition task. We next ask which state dimension locates this wall.

4.3 The width wall tracks predictive-state rank, not state count

Fig. 1A–B shows that the wall tracks r rather than latent-state count: at fixed $K = 128$, changing r moves the wall. Conversely, the $K = 128$ mixing-clone HMM at $g = 16$ is an exact state-splitting realization of the $K = 16$ ring: grouping identical-emission clones recovers its macro transition and initial law up to permutation, so both HMMs generate the same observable process. Their emission rows agree to 5×10^{-13}, both Bayes floors equal 5.765 bits, and both walls fall at $d = 16$. This tests realization invariance rather than comparing two similar processes. Clone replication also preserves occupancy rank; at fixed $g = 16$, spectra for $c = 2, 4, 8, 16$ differ only by the expected $1/c$ factor (Fig. 1C–D). Next, we test whether it tracks the full belief-span dimension.

Direct separation from full belief-span dimension. The rank–span corpus separates predictive rank from the geometry of the full belief. At $g = r = 16$, $c = 8$, and $K = \dim \mathrm{span}\{b_t\} = 128$, the clone identity persists and contributes 112 history-dependent belief directions while never affecting emissions or future macro dynamics. Table 1 shows a sharp drop in loss from $d = 8$ to $d = 16 = r$, followed by a near-plateau, with loss changing only marginally and non-monotonically through $d = 128 = \dim \mathrm{span}\{b_t\}$. Therefore, the additional 112 belief dimensions confer no systematic predictive benefit, which rules out the alternative that the wall tracks the dimension of the full belief span.

Low full-belief R^2 is expected because the probe targets 128 belief coordinates, whereas Bayes-optimal prediction requires only the 16-dimensional macro marginal. Reaching the floor without decoding the inert clone shape is correct, not a filtering failure.

We then verify this quotient algebraically with a future-test certificate. The full belief tangent has dimension 127, but the predictive tangent rank is only $15 = r - 1$, leaving a $112 = K - r$ dimensional nullspace. The certificate annihilates explicit within-macro clone contrasts to at most 3.12×10^{-17}, against 7.20×10^{-2} for a random tangent direction.

Realization and readout controls. A training-free closed-loop recursion provides an optimization-independent control: under this prescribed homogeneous construction, the recursion leaves loss above the Bayes floor at $d = r - 1$ and attains it at $d = r$. The normalized predictive states have only $r - 1$ variable affine coordinates, and Corollary 1 shows that the affine realization is dimension-optimal among exact continuous single-vector bottlenecks. In the non-injective rank–span corpus, top-variance belief directions are insufficient at rank, whereas the macro-marginal quotient is exact, separating predictive rank from the occupancy spectrum. The complete recursion results and the scope of this construction are given in Appendix A.11.

Exact-state calibration clarifies the remaining distinction between r and $r - 1$. Nonlinear heads are near-Bayes-optimal from either state, whereas a bias-free linear head on affine coordinates pays a large penalty for the missing intercept; adding a bias or constant coordinate restores the shared linear-softmax optimum (Appendix A.12). Untying the readout, adding its bias, and fixing the MLP width improve the attained floor mainly at and above $d = r$, yet reduce the below-rank gap by only 11% (Appendix Table 9). The centered Bayes-logit family has rank $V - 1 = 79$, and its exact-state linear-softmax floor is only 3–9% of the below-rank Transformer gaps. Output parameterization therefore affects capacity utilization but does not by itself locate the wall (Appendix A.13).

Finally, the band-limited ring separates finite predictive dimension from causal-state cardinality: at $b = 4$, predictive mixed states occupy a nine-dimensional Fourier span even though the process has uncountably many causal states (Appendix A.1). We next examine how the wall emerges during training.

4.4 THE WALL IS A SPECTRAL LEARNING TRANSITION

Acquisition slows toward the wall but does not diverge at it. In the dense even-width sweep at $g = 16$, t_{90} for the leading belief mode is 2.6×10^4, 1.4×10^4, and 3.3×10^3 steps at $d = 6, 16, 32$ on that grid: monotone in width with no peak at r. Above-wall models fit earlier and can overfit, whereas at-wall runs are often still improving at the end of training (Fig. 2A).

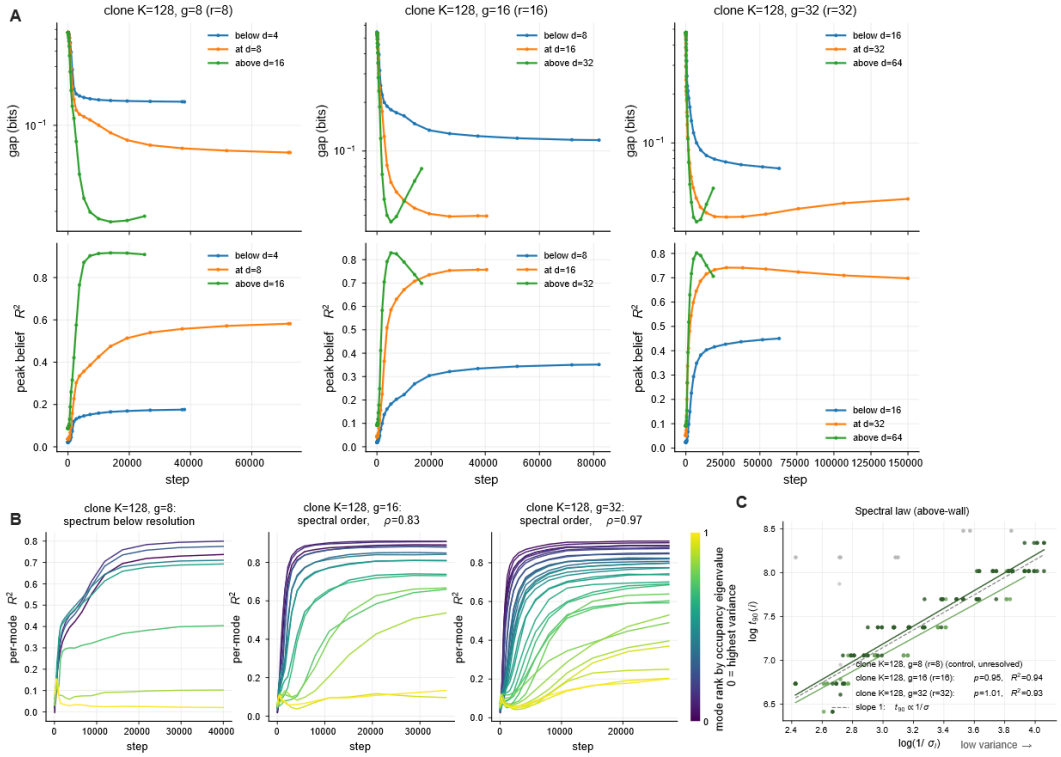

Figure 2: Spectral learning near the width wall. **(A)** Loss and belief recovery over training below, at, and above r. **(B)** Per-mode decodability ordered by predictive-belief occupancy eigenvalue. **(C)** Acquisition time versus $1/\sqrt{\lambda_i}$ for resolved above-wall modes.

Spectral order and timescale. Modes are acquired from high to low predictive-occupancy variance (Fig. 2B). The Spearman correlation between t_{90} and $-\log \lambda_i$ is 0.83 ± 0.17 at $g = 16$, 0.97 ± 0.00 at $g = 32$, and 0.96 ± 0.00 over the 39 resolved modes at $g = 64$. For $g = 16$ and $g = 32$, resolved-mode fits give $t_{90} \propto \lambda_i^{-p/2}$ with $p = 0.95$ and 1.01, consistent with $t_{90} \propto 1/\sqrt{\lambda_i}$. This is a realization-dependent empirical analogy to deep-linear learning, not a derivation from the HMM dynamics.

The raw variance–decodability correlation is not diagnostic. The correlation $\rho(1 - R_i^2, \lambda_i)$ remains between -0.94 and -0.99 below, at, and above rank. An equal-treatment noisy representation reproduces most of this correlation because ridge decoding is intrinsically more accurate for higher-variance targets. We therefore use the calibrated log-odds slope β, rather than the raw correlation, to test mode-selective allocation.

4.5 BELIEF DIRECTIONS ARE CAUSALLY USED

Decodability alone need not imply causal use, so we test whether the belief subspace is actually used. Resampling the decoded belief subspace harms prediction about twice as much as resampling a dimension-matched random subspace for $g = 8, 16, 32$, a difference exceeding ten seed standard deviations (Fig. 3A); since the random intervention is itself damaging, this belief-versus-random contrast, not a comparison against an inert baseline, is the informative one. The belief subspace is therefore causally used at these widths. At $g = 64$, where the representation remains underfit, the belief-specific gap largely disappears.

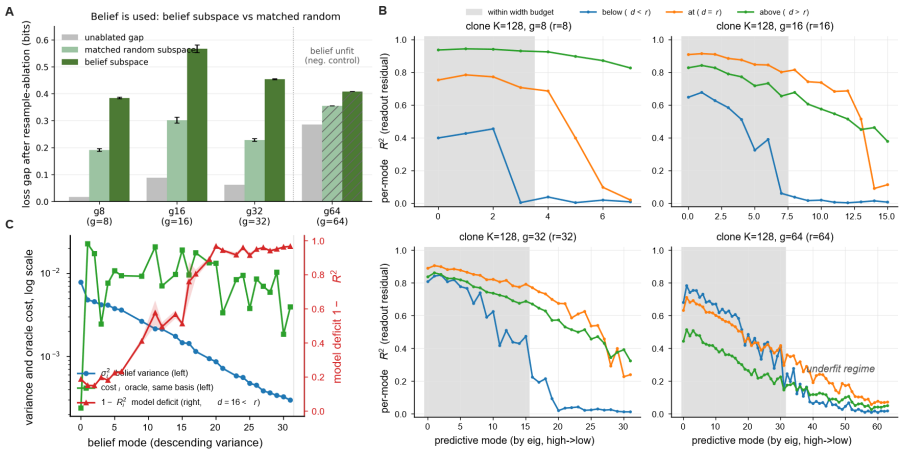

Figure 3: Causal use and per-mode belief geometry. **(A)** Resampling the decoded belief subspace versus a dimension-matched random subspace; $g = 64$ is an underfit negative control. **(B)** Per-mode readout R^2 below, at, and above the wall. **(C)** At $g = 32$, occupancy variance, same-basis one-step cost, and per-mode decodability. Allocation claims use the calibrated statistic rather than the raw deficit–variance correlation.

Within the belief subspace, resampling high-variance modes is more damaging than resampling the same number of tail modes. We treat this as corroborating rather than primary evidence because the intervention is dimension-matched but not magnitude-matched. The belief-versus-random intervention establishes use of the decoded subspace; the calibrated analysis below determines whether its variance profile exceeds the probe-induced reference.

4.6 Width controls variance selectivity

Having established causal use, we show that below rank the residual representation is more variance-selective than an equal-treatment probe null. The empirical-null offsets $\beta - \beta_{\text{null}}$ are positive, 0.45–0.47 at $g = 16$, 0.82–0.86 at $g = 32$, and 0.47–0.51 at $g = 64$. The effect decreases monotonically with width. At rank, the offsets straddle zero across corpora. In contrast, they become negative for the adequately fitted $g = 16$ and $g = 32$ models above rank, meaning that low-variance modes are decoded better than under equal treatment. The $g = 64$ above-rank model remains near the null, consistent with its underfitting. We therefore claim monotone weakening and a sign reversal above rank, not a sharp allocation crossover exactly at r.

Variance selectivity is constructed through depth. At the embedding, every condition has $\beta < 1$: low-variance modes are initially decoded better than the equal-treatment reference. The slope rises overall through successive Transformer blocks. Below-rank models robustly cross into $\beta > 1$, while at-rank models finish near the null and adequately fitted above-rank models remain below it. Thus, the effect is constructed within the network rather than inherited from the token representation.

Neither final normalization nor the output-side architectural bundle explains the buildup. Final LayerNorm contributes an approximately width-independent shift, and untying the embeddings, adding a readout bias, and fixing the MLP width leave the layerwise profile unchanged while substantially improving loss at and above rank. Attention/value covariance and optimization dynamics remain candidate mechanisms. Full values are in Appendix A.3.

4.7 Occupancy and prediction-optimal orderings

Finally, to test whether this selectivity is predictively optimal, we compare occupancy PCA with the rank-d subspace minimizing a local one-step predictive quadratic loss. On the band-limited rings the two orderings coincide: circulant symmetry jointly diagonalizes the occupancy and Fisher operators, so these corpora cannot exhibit a variance–importance conflict.

The clone corpus does exhibit such a conflict. At $g = 16$, occupancy truncation incurs 3%, 12%, 25%, and 32% more one-step loss than the metric-optimal subspace at $d = 2, 4, 8, 12$, respectively. Thus the trained model's below-rank variance selectivity is not forced by one-step predictive optimality. The comparison is local and basis-dependent; it is not an optimal recursive or nonlinear compression bound.

5 IMPLICATIONS FOR LANGUAGE MODELS

Our processes are synthetic and our models small, so the following are falsifiable predictions rather than claims about current language models. First, the capacity required for a contractive latent-state task should track its effective predictive dimension more closely than the size of a chosen latent realization. The state-splitting control predicts that expanding a latent model without changing the observable process should not move the capacity threshold, while the rank–span separation predicts that high-dimensional posterior variation may be ignored when it lies in the predictive nullspace. Second, under capacity pressure, successive Transformer blocks may progressively suppress low-variance latent factors that remain predictively useful. With spare capacity this bias can reverse, so rare or weakly varying factors are not intrinsically inaccessible; their retention depends on the interaction between width and the computation built across layers. This points to attention/value covariance and optimization dynamics as targets for training or architectural interventions. The non-forgetting composition contrast is discussed in Appendix A.10.

6 RELATED WORK

Mechanistic studies find HMM belief geometry and architecture-constrained filtering updates in Transformer residual streams, and identify regimes where depth or intermediate supervision matters (Shai et al., 2024; Piotrowski et al., 2025; Hu et al., 2024). We instead isolate residual width and separate predictive rank from both hidden-state count and full belief dimension; the exact ring–clone state split additionally tests invariance to the chosen latent realization. PSRs, observable-operator models, computational mechanics, and balanced reduction characterize predictive dimension, equivalence, or input–output relevance (Jaeger, 2000; Littman et al., 2001; Hsu et al., 2012; Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001; Moore, 1981; Gugercin & Antoulas, 2004). Our contribution is to connect these notions to a learned width transition and to the layerwise construction of a variance-selective representation. The acquisition-order result relates to staged spectral learning (Saxe et al., 2014; Boix-Adsera et al., 2023); the allocation result differs by calibrating the linear probe against an equal-treatment null and comparing occupancy with one-step prediction-optimal subspaces. The appendix composition study draws on Transformer circuit-complexity and scratchpad results (Liu et al., 2023; Merrill & Sabharwal, 2023; Barrington, 1989; Wei et al., 2022).

7 LIMITATIONS AND OPEN QUESTIONS

The experiments use small Transformers and synthetic stationary HMMs, which give exact beliefs, Bayes losses, and predictive ranks but leave open how these quantities should be identified in natural data or large models. Our filtering families are mixing and contractive, so non-mixing processes may require different depth and memory strategies. Hankel rank is process-invariant, whereas occupancy spectra and eigenvectors depend on the realization, so the acquisition and variance-selectivity results apply to the predictive-belief coordinates used here. The b_t versus b_t^+ convention also changes the band-limited occupancy rank, and circulant symmetry there makes the occupancy and one-step Fisher orderings coincide, preventing that corpus from testing allocation conflicts.

The allocation statistic concerns linear decoding and inherits the ridge probe's assumptions. Equal-treatment nulls control its sensitivity to target variance, but the empirical offset depends weakly on feature-scale convention, especially above rank, so early pre-normalization layers are read qualitatively. Layerwise analysis rules out final normalization and the tested output-side bundle but does not identify the mechanism; attention/value covariance, MLP transformations, and optimization remain candidates. Causal resampling establishes use of the decoded belief subspace, but its controls are not fully covariance-, sensitivity-, or magnitude-matched.

Exact-state calibration and the output-only oracle narrow readout-based explanations but do not isolate joint state–readout training. The one-step Fisher and multi-step observability subspaces are likewise linear quadratic references rather than optimal compressors, and Transformers may encode state nonlinearly, transiently, or across positions. Testing realization invariance, deriving the acquisition law, and causally identifying the mechanism behind variance selectivity remain open.

8 AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic data sets, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, and implementing methods.

We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, proposing or refining hypotheses. Assisting with translation, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, summarizing or analyzing existing literature, brainstorming, and identifying relevant literature. sourcing/searching for information. We have reviewed all AI-assisted work.

9 ETHICS STATEMENT

We have read and adhere to the ICLR Code of Ethics.

Data and human subjects. This work uses no human subjects, no personal or identifying data, and no scraped or licensed corpora. Every token stream analyzed here is generated from hidden Markov models whose transition, emission, and initial distributions are fully specified in Section 3.1 and Appendix A.15, so the data distributions are exactly known rather than estimated from any external source. No annotation labor was involved.

Models and release. All models are decoder-only Transformers trained from scratch. No pretrained model is used, adapted, or released. The trained artifacts have no capability outside predicting the synthetic processes they were fit to and pose no misuse risk. The generative processes, training code, and analysis scripts are released to support reproduction. Total compute for all experiments reported here is approximately 500 GPU-hours on H100.

Potentially harmful insights. Section 5 predicts that capacity-limited models may deprioritize low-variance latent factors, which in natural data could include rare but consequential ones. Our evidence is confined to small models on synthetic HMMs, and the effect reverses when capacity is adequate, so it is neither a demonstration of bias in deployed systems nor an inherent limitation.

Conflicts and integrity. The authors declare no conflicts of interest that could have influenced the reported results. No IRB review was required.

10 REPRODUCIBILITY STATEMENT

Code is available through a zip file along with the paper submission.

REFERENCES

Rami Atar and Ofer Zeitouni. Exponential stability for nonlinear filtering. *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, 33(6):697–725, 1997.

David A. Barrington. Bounded-width polynomial-size branching programs recognize exactly those languages in NC^1. *Journal of Computer and System Sciences*, 38(1):150–164, 1989. ISSN 0022-0000.

Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. In *Advances in Neural Information Processing Systems*, volume 36, pp. 24519–24551, 2023.

James P Crutchfield and Karl Young. Inferring statistical complexity. *Physical Review Letters*, 63(2): 105–108, 1989.

Carlton Downey, Ahmed Hefny, Byron Boots, Geoffrey J. Gordon, and Boyue Li. Predictive state recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6053–6064, 2017.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3202–3211, 2019.

Serkan Gugercin and Athanasios C Antoulas. A survey of model reduction by balanced truncation and some new results. *International Journal of Control*, 77(8):748–766, 2004.

Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Bernardo A Pires, and Rémi Munos. Neural predictive belief representations. *arXiv preprint arXiv:1811.06407*, 2018.

Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

William L Hamilton, Mahdi Milani Fard, and Joelle Pineau. Modelling sparse dynamical systems with compressed predictive state representations. In *International Conference on Machine Learning*, pp. 178–186. PMLR, 2013.

Ahmed Hefny, Zita Marinho, Wen Sun, Siddhartha Srinivasa, and Geoffrey Gordon. Recurrent predictive state policy networks. In *International Conference on Machine Learning*, pp. 1949–1958. PMLR, 2018.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Jiachen Hu, Qinghua Liu, and Chi Jin. On limitation of transformer for learning HMMs. *arXiv preprint arXiv:2406.04089*, 2024.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

Alex Kulesza, Nan Jiang, and Satinder Singh. Low-rank spectral learning with weighted loss functions. In *Artificial Intelligence and Statistics*, pp. 517–525. PMLR, 2015.

Belinda Z Li, Zifan Carl Guo, and Jacob Andreas. (How) do language models track state? *arXiv preprint arXiv:2503.02854*, 2025.

Michael L Littman, Richard S Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1555–1561, 2001.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations*, 2023.

Sarah Marzen and James P. Crutchfield. Circumventing the curse of dimensionality in prediction: Causal rate-distortion for infinite-order Markov processes. *arXiv preprint arXiv:1412.2859*, 2014.

William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023.

Bruce C Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, 1981.

Mateusz Piotrowski, Paul M Riechers, Daniel Filan, and Adam S Shai. Constrained belief updates explain geometric structures in transformer representations. In *International Conference on Machine Learning*, pp. 49399–49419. PMLR, 2025.

Pascal Poupart and Craig Boutilier. Value-directed compression of POMDPs. In *Advances in Neural Information Processing Systems*, volume 15, pp. 1579–1586. MIT Press, 2002.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pp. 5301–5310. PMLR, 2019.

Nicholas Roy, Geoffrey Gordon, and Sebastian Thrun. Finding approximate POMDP solutions through belief compression. *Journal of Artificial Intelligence Research*, 23:1–40, 2005.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Adam S Shai, Sarah E Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M Riechers. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems*, volume 37, pp. 75012–75034, 2024.

Cosma Rohilla Shalizi and James P Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics*, 104(3):817–879, 2001.

Sajid Siddiqi, Byron Boots, and Geoffrey Gordon. Reduced-rank hidden Markov models. In *Artificial Intelligence and Statistics*, pp. 741–748. PMLR, 2010.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Andrew Wang, Andrew C Li, Toryn Q Klassen, Rodrigo Toro Icarte, and Sheila A McIlraith. Learning belief representations for partially observable deep RL. In *International Conference on Machine Learning*, pp. 35970–35988. PMLR, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Yifan Zhang, Wenyu Du, Dongming Jin, Jie Fu, and Zhi Jin. Finite state automata inside transformers with chain-of-thought: A mechanistic study on state tracking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13603–13621, Vienna, Austria, July 2025. Association for Computational Linguistics.

A Appendix

A.1 Propositions and conjectures

Proposition 2 (The band-limited ring has uncountably many causal states)**.** *Consider the reported band-limited ring with $K = 128$, transition bandwidth $b \in \{2, 4\}$, emission bandwidth $b_e = b$, and positive transition leak. Its predictive mixed states $\eta_t = P(Z_t = \cdot \mid X_{-\infty:t-1})$ lie in the homogeneous Fourier subspace*

$$\mathcal{H}_b = \{v \in \mathbb{R}^K : \widehat{v}_m = 0 \text{ for } m \notin \{0, \pm 1, \ldots, \pm b\}\}, \qquad \dim_{\mathbb{R}} \mathcal{H}_b = 2b + 1.$$

The next-symbol map $\eta \mapsto B^\top \eta$ is injective on $\mathcal{H}_b$, and the observation-conditioned filtering maps have a non-singleton perfect compact attractor. The process therefore has uncountably many causal states despite finite-dimensional predictive structure.

Proposition 3 (Affine realization of clone–blow-up beliefs)**.** *For a clone–blow-up process of predictive rank g, the normalized predictive states induced by macro priors lie in a forward-invariant affine hull of dimension $g - 1$, whose homogeneous linear span has dimension g. Let μ be any fixed point in this hull (e.g., $\mathbb{E}[b_t]$ under stationarity), and let $U \in \mathbb{R}^{K \times (g-1)}$ be an orthonormal basis for its tangent space. Every predictive belief in this hull, including those generated from an open family of non-stationary macro priors, can be written uniquely as*

$$b_t = \mu + U y_t, \qquad y_t = U^\top (b_t - \mu).$$

An affine filter that recursively updates the $g - 1$ live coordinates y_t and restores the fixed mean μ before prediction reproduces the exact Bayes filter. Thus the process has $g - 1$ variable belief degrees of freedom, whereas a purely homogeneous linear representation uses g coordinates.

Conjecture 1 (Spectral learning of belief modes)**.** *Belief acquisition is spectrally ordered by the stationary reachable-belief second moment $C = \mathbb{E}[b_t b_t^\top]$.*

Phenomenologically, training behaves analogously to staged learning in deep linear networks (Saxe et al., 2014): modes with larger spectral scale in C are acquired earlier, while lower-variance modes

are acquired later or may remain underfit within a finite training horizon. Here a "mode" means an eigenvector of C in the fixed generative realization, and the corresponding deep-linear surrogate singular scale is taken to be $s_i = \sqrt{\lambda_i(C)}$.

In this analogy, C plays the surrogate role of the input–output correlation object whose spectrum orders learning. This correspondence is observed empirically, not derived from the HMM dynamics, and is not claimed to be invariant under arbitrary reparameterizations of belief space.

Conjecture 2 (Spectral scaling of acquisition time)**.** *Let $\lambda_i = \lambda_i(C)$ be the eigenvalue of the occupancy second moment $C = \mathbb{E}[b_t b_t^\top]$ associated with mode i. Under the deep-linear surrogate, the relevant singular scale is taken to be*

$$s_i \asymp \sqrt{\lambda_i}.$$

We therefore test an acquisition-time law

$$t_{\text{acquire}}(i) \propto s_i^{-p}, \qquad p \in \{1, 2\},$$

equivalently

$$t_{\text{acquire}}(i) \propto \lambda_i^{-p/2}.$$

The convention matters. The mode-wise timescale in Saxe et al. (2014) is inverse in the singular value of the input–output correlation, not inverse in its square. Thus the Saxe-style $p = 1$ prediction corresponds to regressing t_{90} against $1/\sqrt{\lambda_i}$, whereas an inverse-variance law $p = 2$ corresponds to regressing against $1/\lambda_i$. Identifying $1/\lambda_i$ with $1/s_i$ therefore misstates the exponent by one power.

We therefore fit both exponents and select p empirically, restricting to modes whose final decodability clears a threshold; the low-variance tail is censored and under-fit at the wall, and would bias the slope. The mapping to the deep-linear surrogate is itself conjectural, so p is not derived.

A.2 Predictive-state dimension certificates

For a future string τ, let $w_\tau[k] = P(\tau \mid z_1 = k)$, computed by the right-to-left recursion $w \leftarrow B_{:,x} \odot Aw$. Stacking these vectors gives $T = [w_\tau]$ and $F(b) = T^\top b$. Strings were enumerated breadth-first until rank saturation. The homogeneous rank is $\mathrm{rank}(T)$, and the affine rank is the rank after restriction to the simplex tangent. Every rank below is unchanged across relative tolerances 10^{-6}–10^{-12}.

Table 2: Future-test injectivity certificates. "Gap" is the smallest retained divided by the largest discarded singular value at the rank boundary.

Corpus	r	$\mathrm{rank}(T)$	Aff. rank	Hom. gap	Aff. gap	$\sigma_{\min}$	Cond.
$g8$	8	8	7	5.40×10^{14}	6.09×10^{14}	2.03×10^{-2}	26.2
$g16$	16	16	15	4.04×10^{13}	3.86×10^{13}	1.43×10^{-3}	358.5
$g32$	32	32	31	2.65×10^{13}	3.82×10^{13}	3.89×10^{-4}	1174
$g64$	64	64	63	1.66×10^{13}	2.39×10^{13}	4.46×10^{-4}	1252

Ranks saturate by future length one for $g8$ and length two for the other corpora. To verify attainability on the theorem's open-prior domain, we sampled 1000 interior priors from $\mathrm{Dirichlet}(\mathbf{1})$ per corpus (rejecting coordinates below 10^{-4}) and ran 1000-token rollouts. Across one million prior-position pairs per corpus, the affine filter's maximum prediction discrepancy is 1.81×10^{-14}, maximum absolute excess loss is 4.21×10^{-13} bits/token, and maximum hull-reconstruction residual is 4.37×10^{-14}.

On the rank–span corpus, the belief tangent has dimension 127, while the homogeneous and affine future-test ranks are 16 and 15, leaving a 112-dimensional predictive nullspace. Across 448 explicit clone-contrast directions, $\max_v \|T^\top v\|/\|v\| = 3.12 \times 10^{-17}$; a random tangent direction gives 7.20×10^{-2}.

A.3 PROBE-CALIBRATED ALLOCATION AND ORDERING ANALYSIS

Let $C = \mathbb{E}[b_t b_t^\top] = \sum_i \lambda_i e_i e_i^\top$. For each residual location and seed, a ridge probe gives per-mode decodability R_i^2. We fit

$$\beta = \text{slope}_i \left(\log \frac{R_i^2}{1 - R_i^2} \text{ against } \log \lambda_i \right).$$

Under an equal-treatment noisy embedding, $1 - R_i^2 \propto 1/\lambda_i$, so $\beta = 1$ analytically. Slopes are fitted per seed and then averaged; averaging R_i^2 before fitting changes the estimate materially and does not yield valid seed error bars. All modes are used because thresholding removes the low-variance tail that identifies the slope.

The empirical null embeds all r modes through a random equal-treatment map into d dimensions, adds isotropic noise, and applies the identical probe. Standardizing null features is necessary because the implemented ridge penalty scales with feature magnitude. Across all conditions, the standardized null gives $\beta = 0.89$–1.07, close to the analytic value. A raw-scale null is retained as a robustness bracket; it changes below-rank offsets by at most 0.05 and above-rank offsets by up to 0.28.

Table 3: Probe-calibrated variance selectivity at the readout residual. Equal-treatment decoding gives $\beta = 1$ analytically. Values are fitted per seed and reported as mean $\pm$ seed standard deviation.

Condition	Below rank	At rank	Above rank
$g16$, tied	1.54 ± 0.02	1.18 ± 0.05	0.64 ± 0.01
$g32$, tied	1.82 ± 0.02	0.99 ± 0.02	0.69 ± 0.01
$g64$, tied	1.50 ± 0.03	1.19 ± 0.02	1.05 ± 0.02
$g16$, untied+bias	1.75 ± 0.06	0.97 ± 0.06	0.61 ± 0.03

The tested widths are $d = r/2, r, 2r$. Relative to standardized and raw-scale empirical nulls, below-rank offsets are $+0.45$–$+0.47$, $+0.82$–$+0.86$, and $+0.47$–$+0.51$ for $g = 16, 32, 64$. Above-rank offsets are -0.58–-0.31 at $g = 16$, -0.52–-0.27 at $g = 32$, and $+0.01$–$+0.09$ at $g = 64$.

Table 4: Layerwise offset from the standardized empirical null. Positive values indicate suppression of low-variance modes beyond the probe effect.

Corpus	Condition	Emb.	B0	B1	B2	B3	Out
$g16$ tied	below	-0.64	-0.54	-0.43	-0.31	$+0.30$	$+0.47$
	at	-1.08	-0.90	-0.32	-0.04	$+0.09$	$+0.24$
	above	-1.28	-1.09	-0.61	-0.48	-0.46	-0.31
$g32$ tied	below	-0.80	-0.68	-0.13	$+0.26$	$+0.73$	$+0.86$
	at	-1.10	-0.87	-0.28	$+0.04$	-0.04	$+0.10$
	above	-1.28	-0.96	-0.46	-0.32	-0.40	-0.27
$g64$ tied	below	-1.01	-0.85	-0.29	$+0.12$	$+0.51$	$+0.51$
	at	-1.24	-0.91	-0.33	$+0.03$	$+0.19$	$+0.21$
	above	-1.28	-0.46	-0.18	-0.01	$+0.03$	$+0.09$
$g16$ control	below	-0.72	-0.58	-0.42	-0.19	$+0.50$	$+0.68$
	at	-1.20	-0.79	-0.27	$+0.06$	-0.11	$+0.02$
	above	-1.27	-0.69	-0.52	-0.53	-0.50	-0.33

For the ordering comparison, choosing a rank-d linear subspace to minimize $\mathbb{E}[(b - \widehat{b})^\top G (b - \widehat{b})]$ gives the leading eigenvectors of $\Sigma^{1/2} G \Sigma^{1/2}$. We compare $G = I$ (occupancy PCA), the averaged one-step Fisher pullback $G = B \,\text{diag}(B^\top b)^{-1} B^\top$, and a multi-step observability Gramian. On the band-limited rings, occupancy and one-step Fisher subspaces differ by only $0.3°$–$1.4°$, below the split-half resolution floor, and produce identical loss to four decimal places. On the $g = 16$ clone corpus they differ by $35.9°$–$61.6°$.

Table 5: One-step excess loss from the occupancy and local Fisher-optimal subspaces on the $g = 16$ clone corpus.

d	Occupancy	Fisher-optimal	Relative penalty
2	0.4601	0.4462	3%
4	0.2953	0.2636	12%
8	0.1116	0.0890	25%
12	0.0305	0.0231	32%

A.4 Proof of Prop. 1

Proof. Let $S_d = \mathrm{span}(R)$ and suppose $\mathcal{R} \subseteq S_d$. Since the stationary prior π belongs to $\mathcal{R}$, $z_0 = R^\top \pi$ gives $\widetilde{\alpha}_0 = RR^\top \pi = \pi$. Hence $m_0 = 1$, and the first prediction is $B^\top \pi$; clipping is inactive because this is a probability vector.

Assume inductively that $\alpha_t = b_t$, where b_t is the exact predictive belief before x_t. Then Equation equation 2 gives the exact Bayes prediction $p_t = B^\top b_t$. After observing x_t, the unnormalized next predictive belief is

$$A^\top \big(b_t \odot B_{:,x_t}\big).$$

Equation equation 3 stores its coordinates in S_d. At the next step, reconstruction gives

$$\widetilde{\alpha}_{t+1} = RR^\top A^\top \big(b_t \odot B_{:,x_t}\big).$$

After normalization, the exact next predictive belief belongs to the stationary reachable span $\mathcal{R} \subseteq S_d$, so $RR^\top$ leaves it unchanged. Therefore $\alpha_{t+1} = b_{t+1}$. Induction proves equality on every positive-probability trajectory, and hence equality of the per-token log loss with the Bayes floor.

For $d < r$, the reconstructed state need not equal the exact belief. The implemented clipping and renormalization still produce a valid next-token distribution, but equation 3 feeds the projected state back into future predictions. Its loss is therefore the loss of this prescribed closed-loop d-dimensional recursion. The proof makes no claim that it is optimal among all d-dimensional or nonlinear predictors. □

A.5 Proof of Prop. 2

Proof. Let $\mathcal{F}$ be the unitary K-point discrete Fourier transform. The transition and emission matrices are circulant Fejér kernels, so

$$A = \mathcal{F}^* \,\mathrm{diag}(\widehat{a}_m)\mathcal{F}, \qquad B = \mathcal{F}^* \,\mathrm{diag}(\widehat{b}_m)\mathcal{F}.$$

The transition multipliers are nonzero exactly on $S_b = \{0, \pm 1, \ldots, \pm b\}$; the positive uniform leak changes only the zero-frequency multiplier and leaves this support unchanged. Hence every predictive mixed state lies in $\mathcal{H}_b = \{v \in \mathbb{R}^K : \widehat{v}_m = 0 \text{ for } m \notin S_b\}$. The emission bandwidth satisfies $b_e \geq b$, so $\widehat{b}_m \neq 0$ for every $m \in S_b$. Therefore, for $\eta, \eta' \in \mathcal{H}_b$,

$$B^\top \eta = B^\top \eta' \quad \Longrightarrow \quad \eta = \eta'.$$

Distinct predictive mixed states thus already have distinct next-symbol distributions.

For a symbol x, let $D_x = \mathrm{diag}(B_{:,x})$ and define the predictive filtering map on column-vector beliefs by

$$F_x(\eta) = \frac{A^\top D_x \eta}{\mathbf{1}^\top D_x \eta}.$$

For the reported $K = 128$, $b_e \in \{2, 4\}$ matrices, every emission entry is strictly positive; the transition leak makes every entry of A strictly positive. Thus $A^\top D_x$ is a positive matrix. Each F_x sends the simplex into a compact subset of its interior and is a strict contraction there in the Hilbert projective metric.

Each F_x is also injective on $\mathcal{H}_b \cap \Delta^{K-1}$. Suppose $F_x(\eta) = F_x(\eta')$, and write $c = \mathbf{1}^\top D_x \eta$ and $c' = \mathbf{1}^\top D_x \eta'$. Then

$$A^\top D_x v = 0, \qquad v = \eta/c - \eta'/c' \in \mathcal{H}_b.$$

The Fejér transition is symmetric, with nullspace $\mathcal{H}_b^\perp$. Hence $D_x v \in \mathcal{H}_b^\perp$, and therefore

$$0 = v^\top D_x v = \sum_j B_{j,x} v_j^2.$$

Strict positivity of $B_{j,x}$ implies $v = 0$. Since η and η' both have unit mass, $c = c'$ and $\eta = \eta'$.

The maps $\{F_x\}_{x=0}^{K-1}$ form a finite iterated-function system with a unique compact attractor $\mathcal{M}$. It is not a singleton. Otherwise a state $\eta_\star$ would be fixed by every F_x. Circulant equivariance would then make every cyclic shift of $\eta_\star$ a fixed point of F_0; uniqueness of the fixed point of the strict contraction F_0 would force $\eta_\star$ to be uniform. But the uniform state is not fixed by F_0: conditioning on symbol 0 introduces nonzero Fourier modes $\pm 1, \ldots, \pm b$, which survive the transition. This is a contradiction.

Finally, a non-singleton attractor of injective contractions has no isolated points. Every $\eta \in \mathcal{M}$ has an infinite symbolic address, and the nested cylinder sets associated with its finite prefixes contain η, remain non-singleton by injectivity, and have diameters tending to zero by contraction. Thus $\mathcal{M}$ is a nonempty perfect compact subset of a finite-dimensional space and is uncountable. Since all transition and emission probabilities are positive, every finite word has positive probability and the full symbolic shift lies in the topological support of the stationary observed process. Under the causal-state convention stated in §2.1, the corresponding semi-infinite pasts define uncountably many distinct next-symbol distributions and hence uncountably many causal states. □

A.6 Proof of Prop. 3

Proof. Let $\mathcal{A}$ be the forward-invariant affine hull of normalized predictive states induced by macro priors, let $\mu \in \mathcal{A}$, and let $\mathcal{H} = \mathcal{A} - \mu$. Choose an orthonormal basis U for $\mathcal{H}$. Since $\dim \mathcal{A} = g - 1$, U has $g-1$ columns. Every belief $b \in \mathcal{A}$ has unique coordinates $y = U^\top(b - \mu)$ and is reconstructed exactly as $b = \mu + Uy$.

Let $\mathcal{T}_x$ denote the exact Bayes update followed by the latent transition. Define the coordinate recursion

$$y_{t+1} = U^\top\big(\mathcal{T}_{x_t}(\mu + Uy_t) - \mu\big),$$

and predict with

$$p_t = B^\top(\mu + Uy_t).$$

If $\mu + Uy_t = b_t$, then the prediction is Bayes-optimal and the update gives the coordinates of b_{t+1} exactly. Induction therefore proves exactness for all times using $g - 1$ updated coordinates plus the fixed vector μ. □

A.7 Proof of Theorem 1 and Corollary 1

Proof of Theorem 1. Because $\Phi = D \circ E$ is injective, E must be injective. Identify $\mathcal{U}$ with an open subset of $\mathbb{R}^m$. Suppose $d < m$, and let $\iota : \mathbb{R}^d \hookrightarrow \mathbb{R}^m$ be the standard coordinate inclusion. Then $\iota \circ E$ is a continuous injection from an open subset of $\mathbb{R}^m$ into $\mathbb{R}^m$. By invariance of domain, its image is open, but it lies in the proper subspace $\iota(\mathbb{R}^d)$, which has empty interior. This contradiction gives $d \geq m$. □

Proof of Corollary 1. For the mixing clone blow-up family, normalized predictive states occupy an $(r-1)$-dimensional affine set and the predictive functional is injective on it. Theorem 1 therefore gives $d \geq r - 1$, while Proposition 3 attains $d = r - 1$. □

A.8 GROUP WORD-PROBLEM CORPUS AND TRAINING PROTOCOL

Each sequence is a stream of blocks $[\text{RESET}, g_1, \ldots, g_W, \text{READOUT}]$. The generators g_i are drawn i.i.d. uniformly from a two-element generating set, and READOUT encodes the running product $g_1 \cdots g_W$, reset at each block. The generator tokens are uniformly distributed and the RESET token is periodic, so both are trivially predictable; evaluation therefore uses readout-token accuracy with chance level $1/|G|$. We compare two non-abelian groups of order 60 with two generators: non-solvable A_5 and solvable D_{30}. This matches order, generator-set size, and noncommutativity, while other group-specific structure remains unmatched. Width is fixed at $d_{\text{model}} = 128 \gg 2|G|$; we sweep word length W and layer count. Because group composition generalizes late, these runs train for 10^5–1.5×10^5 steps without early stopping, and a run is counted as failed only after readout accuracy remains at chance with no upward drift.

A.9 DEPTH DICHOTOMY: THEORY

Under the positivity and observation-nondegeneracy conditions satisfied by our mixing HMMs, the Bayes filter is a contraction on the belief simplex in the Hilbert metric once the m-step operator A^m is strictly positive (Atar & Zeitouni, 1997). Consequently, the influence of evidence at time $t-k$ decays geometrically in k. For any fixed approximation tolerance $\epsilon > 0$, an ϵ-accurate belief can therefore be computed from a recent window of length $O\left(\frac{\log(1/\epsilon)}{\log(1/|\lambda_2|)}\right)$, which is independent of the sequence length T. Thus approximate filtering at fixed accuracy requires only a bounded effective history, rather than exact dependence on the full history.

A naive per-layer-composition account predicts otherwise. If each layer advanced the recursion one step, reaching depth-T dependence would need $\sim \log T / \log(1/|\lambda_2|)$ layers, which for the hard-K family of §4.1 gives between 124 and 449 layers at $T = 128$. The estimate is set by $|\lambda_2|$ rather than by K, so it need not be monotone in K. The trained models instead need depth 2–4, a discrepancy of two orders of magnitude, which rules out a literal one-filter-step-per-layer implementation. One attention layer aggregates across all prior positions at once, so global attention permits substantially more parallel evidence aggregation than the naive sequential account.

Here we contrast two computations that both appear to "track a latent state." Under the stated mixing conditions, contractive filtering forgets old evidence exponentially, yielding bounded effective history and $O(1)$ depth at fixed target accuracy. In contrast, persistent state-tracking tasks such as Dyck-language recognition and permutation composition can depend on arbitrarily early inputs (Hahn, 2020; Liu et al., 2023; Barrington, 1989). The precise depth requirement, however, is more nuanced than simply "requires $\log T$." An $O(\log T)$-depth Transformer can simulate any finite automaton over T steps, while some solvable or aperiodic automata admit $O(1)$-depth shortcuts (Liu et al., 2023). The corresponding lower bound is conditional: constant-depth, log-precision Transformers lie in uniform TC^0 (Merrill & Sabharwal, 2023), whereas the word problem for a fixed non-solvable group, such as composition in S_5, is NC^1-complete (Barrington, 1989). Thus no fixed-depth Transformer can solve such group-tracking problems for unbounded T unless $\text{TC}^0 = \text{NC}^1$. The resulting dichotomy is between approximate contractive filtering, which admits $O(1)$ depth at fixed tolerance, and persistent non-solvable state tracking, which conditionally requires super-constant depth, with $O(\log T)$ providing a general upper bound. In §A.10, we report a learnability gap consistent with this distinction: the tested solvable group is learned at all sequence lengths, whereas the non-solvable group is not. Because the complexity result concerns representability while our experiment measures learnability under a particular training setup, we treat this gap as aligned with the theory rather than as evidence for the lower bound.

A.10 NON-SOLVABLE COMPOSITION: FULL RESULTS

This appendix provides the non-forgetting contrast referenced in §4.2; the corpus and training protocol are given in Appendix A.8. For the matched order-60 groups, D_{30} is learned at every tested length, whereas A_5 collapses from near-perfect accuracy to chance beyond $W \approx 20$–24 despite tests with substantially deeper models (Fig. 4A). Because this is one group pair, the result is a learnability contrast consistent with the complexity discussion, not evidence that solvability alone causes the gap.

Intermediate products dissolve the composition wall. At $W = 64$, emitting the running product every k generators makes the task learnable for $k \leq 16$ and leaves it at chance for $k \geq 32$ (Fig. 4B). The wall therefore follows per-step composition length rather than total length. This intervention combines externalized state with intermediate supervision, whose separate contributions remain open.

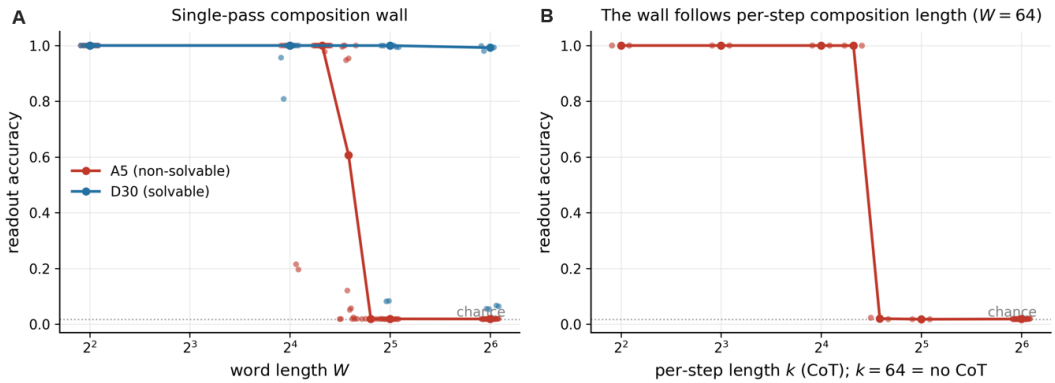

Figure 4: Composition contrast. **(A)** Readout accuracy versus word length for A_5 and D_{30}. **(B)** At $W = 64$, accuracy versus supervised segment length k ($k = 64$ gives only the final readout).

A.11 Projected-recursion and realization controls

The training-free projected recursion remains above the floor at $d = g - 1$ and reaches the Bayes floor at $d = g = r$ on every injective (mixing) clone corpus (Table 6). Because no parameters are trained, this cliff cannot be attributed to neural optimization. The result establishes exactness for the prescribed homogeneous recursion, not a lower bound over all nonlinear width-d predictors.

Table 6: Closed-loop homogeneous projected recursion and affine realization, bits/token. The homogeneous recursion becomes exact at $d = g$. Supplying the fixed mean separately lets the affine filter use only $g - 1$ live coordinates with the same loss.

corpus	g	hom. $g-1$	hom. g	aff. $(g-1)+\mu$	Bayes
g8	8	6.007	5.755	5.755	5.755
g16	16	5.982	5.765	5.765	5.765
g32	32	5.961	5.845	5.845	5.845
g64	64	5.984	5.936	5.936	5.936

The rank–span corpus exposes the scope of the occupancy-PCA certificate. Projecting onto the top $g = 16$ eigenvectors of C leaves 0.107 bits of excess, although the process rank is 16; the separate macro-marginal recursion gives 0.0000 excess at reported precision. Proposition 1 guarantees exact full-belief reconstruction when the projected space contains the reachable belief span. Occupancy PCA meets that condition in the injective (mixing) corpora, but not in the rank–span construction, where exact rank-r prediction instead operates on the predictive quotient.

Table 6 also resolves the apparent g versus $g - 1$ discrepancy. The mixing clone beliefs have $g - 1$ variable affine coordinates, and an exact filter restores the fixed mean before prediction. The tested tied, bias-free Transformer transitions near the homogeneous dimension g, motivating architectures that supply the constant component separately. Corollary 1 provides the matching lower bound for exact continuous single-vector bottlenecks.

A.12 Exact-state readout calibration

The calibration uses exact macro marginals from the held-out belief arrays. The homogeneous state is $m \in \mathbb{R}^r$; the affine state is $y = U^\top(m - \mu) \in \mathbb{R}^{r-1}$, with U an orthonormal basis for the centered macro-state span. We additionally test $[y; 1]$ and Qm for a random invertible matrix

Q. Macro-marginal sufficiency and affine reconstruction are checked before fitting. The observed maximum errors are 4.3×10^{-9} and 1–3 $\times 10^{-9}$, respectively. Since the saved beliefs are float32, these values are at numerical round-off; the implementation uses a conservative 10^{-6} acceptance threshold rather than treating 10^{-9}-scale residuals as failures.

For each state representation we optimize (i) a bias-free linear-softmax head, (ii) a linear-softmax head with bias, and (iii) a two-layer GELU head of hidden width 256. Fits use float64, 16,384 fixed positions, 20,000 steps, and two seeds. The same positions are used for fitting and evaluation intentionally: this experiment estimates the best observed output-class approximation from a known sufficient state, not out-of-sample generalization. A separate convergence sweep changed the linear objective by less than 5×10^{-10} between 12,000 and 24,000 steps. The bias-free linear problem is convex in the readout weights, and the agreement of homogeneous, affine-plus-bias, affine-plus-constant, and randomly reparameterized homogeneous states to four significant figures provides an additional conditioning-independent check that the reported values are class optima rather than stopping points.

The nonlinear heads reach 3×10^{-5}–10^{-4} bits from every sufficient state parameterization. The linear-softmax class reaches 6.877×10^{-3} bits at $g = 16$ and 6.488×10^{-3} bits at $g = 32$ from homogeneous, affine-plus-bias, affine-plus-constant, and random homogeneous codes. The sole failure is the bias-free linear head on affine coordinates alone: 0.3014 and 0.2724 bits. Thus the experiment separates three effects: predictive-state sufficiency, a small linear-softmax approximation floor, and a large missing-intercept penalty. It is a calibration on the exact state, not a lower bound over all width-r codes and not a direct proof that the trained Transformer's below-rank error is entirely internal.

Table 7: Exact-state readout calibration, excess bits. "Linear" denotes a bias-free linear-softmax head. Parameterizations that supply the fixed component attain the same linear-softmax class optimum.

State and head	$g = 16$	$g = 32$
hom-r, linear	6.877×10^{-3}	6.488×10^{-3}
aff-$(r-1)$, linear	3.014×10^{-1}	2.724×10^{-1}
aff-$(r-1)$, linear+bias	6.877×10^{-3}	6.488×10^{-3}
aff+constant, linear	6.877×10^{-3}	6.488×10^{-3}
hom-r, nonlinear	4.4×10^{-5}	5.0×10^{-5}
aff-$(r-1)$, nonlinear	4.0×10^{-5}	3.4×10^{-5}

A.13 Output-capacity analysis

This control complements the exact-state calibration by asking whether the width-d linear output map could contribute to the observed wall near predictive rank. We separate an exact-rank question from an approximate-loss question.

Exact Bayes-logit rank. For each held-out Bayes next-token distribution $p_t \in \Delta^{V-1}$, define the centered Bayes logit

$$\ell_t = \log p_t - \frac{1}{V}\mathbf{1}\mathbf{1}^\top \log p_t,$$

which fixes the additive-softmax gauge. Stacking the ℓ_t as rows gives the Bayes-logit matrix. Across the $g \in \{8, 16, 32, 64\}$ clone corpora used for this certificate, with $V = 80$, the centered matrix has rank $V - 1 = 79$ and shows no spectral break at predictive rank r. The logarithm therefore does not preserve the belief family's rank-$(r-1)$ affine structure. Since a bias-free linear readout from a width-d code produces logits in a subspace of rank at most d, exact output representability would require $d \geq V - 1$, not $d \geq r$. Hence exact output rank cannot mechanically locate the observed transition at r.

Approximate output-only oracle. To test the finite-loss regime, we optimize unconstrained per-position codes $h_t \in \mathbb{R}^d$ and a bias-free linear readout $W \in \mathbb{R}^{d\times V}$ against the exact Bayes targets:

$$G_{\mathrm{out}}(d) = \min_{\{h_t\},W} \frac{1}{N}\sum_{t=1}^{N} D_{\mathrm{KL}}\big(p_t \,\big\|\, \mathrm{softmax}(W^\top h_t)\big)\,.$$

The codes are unconstrained by recurrence or optimization through a Transformer, so this oracle measures what the same width-d, bias-free linear-readout class can achieve when representation learning is removed. We compare it with the seed-averaged trained-model gap G_{model}.

Before running this comparison, we specified a decisive-separation criterion: $G_{\mathrm{out}}/G_{\mathrm{model}} \le 0.25$ at both tested below-rank widths, with margin beyond seed spread. Ratios in the interval 0.25–0.60 were designated intermediate, and therefore as failing the criterion.

Table 8: Output-only oracle versus the final-checkpoint Transformer below predictive rank. G_{out} is the optimized exact expected KL in bits; $G^{\mathrm{final}}_{\mathrm{model}}$ is the seed-averaged sampled-cross-entropy gap at the final checkpoint. The quantities are not evaluated identically and the comparison is therefore diagnostic rather than decisive.

Corpus	Width d	G_{out}	$G^{\mathrm{final}}_{\mathrm{model}}$	Ratio
$g16$	4	0.1067	0.2122	0.50
$g16$	8	0.0293	0.1168	0.25
$g32$	4	0.1302	0.2054	0.63
$g32$	8	0.0688	0.1520	0.45
$g32$	16	0.0187	0.0699	0.27
$g32$	24	0.0057	—	—

The results fall in the intermediate regime under our specified criterion, rather than exhibiting a clean cross-corpus separation. The oracle achieves lower loss than the trained model below r, indicating that the Transformer does not fully exploit the capacity available within the same readout class. The advantage is rank-dependent: roughly 2–4$\times$ at $g = 16$ and 1.6–3.7$\times$ at $g = 32$. Together with the exact-state calibration, these results show that output capacity alone does not account for all below-rank error, while leaving its interaction with an approximate learned state unresolved.

Measurement caveat. G_{out} is evaluated as exact expected KL, whereas G_{model} is an end-of-training sampled-cross-entropy gap. Under Bayes-generated tokens, the paired cross-entropy gap is an unbiased estimator of the same expected KL. The remaining mismatch is that the oracle and exact-state calibration use a fixed position set, whereas model evaluation omits the final position of each sequence; this affects less than 1% of positions. Saved checkpoints were unavailable for a matched per-position KL evaluation. The observed differences are larger than ordinary sampling fluctuations, but the quantities are not estimated identically. We therefore treat the approximate oracle as informative but not decisive and retain the joint-training caveat in the main text.

A.14 Analytic and causal analysis protocols

Analytic predictor. The projected recursion of Proposition 1 is implemented by equation 1–equation 3. It uses L_1 mass normalization: both the reconstructed state $\widetilde{\alpha}_t$ and coordinates z_t are divided by $m_t = \mathbf{1}^\top \widetilde{\alpha}_t$. The readout is a clipped probability ratio rather than a softmax, so positive rescaling is a gauge. If m_t is non-finite or numerically indistinguishable from zero, the implementation falls back to a uniform state. This safeguard is relevant only in the non-exact regime ($d < r$). On the injective (mixing) clone corpora, the top-d occupancy eigenspace contains the reachable belief span once $d \ge r$, so mass normalization reconstructs the exact belief, clipping is inactive, and the recursion is Bayes-optimal. On the rank–span corpus, the top-r occupancy eigenspace is not sufficient. The reported rank-r exact predictor is instead a separate recursion on the macro marginal; it reproduces predictions without reconstructing the full K-dimensional belief. No parameters are trained in either analytic construction.

Below r, the recursion reprojects and feeds back its own approximate state at every step, so errors accumulate. This closed-loop loss should not be conflated with the oracle analyses below, which restart from the exact belief independently at every position.

Prediction-weighted subspaces. All one-step comparisons use predictive beliefs and score through $p_b = B^\top b$. For a fixed metric G, the rank-d quadratic-optimal linear subspace is obtained from $\Sigma^{1/2} G \Sigma^{1/2}$. We compare occupancy PCA, the averaged one-step Fisher pullback, and a multi-step observability Gramian; full results are in Appendix A.3.

Probe-calibrated allocation. Per-mode decodability is summarized by the odds-ratio slope β defined in §3.6. Slopes are fitted per seed over all predictive occupancy modes and compared with the analytic equal-treatment null $\beta = 1$. The deficit correlations are retained only as descriptive diagnostics.

Causal resampling test. At the readout residual, we replace the component in the belief-carrying subspace with a position-shuffled copy of itself, keeping activations on-distribution while destroying their information about the current token, and remeasure the gap. We reject hard zeroing because it is an off-distribution shock that damages a random subspace as much as the belief subspace. Controls are a dimension-matched random subspace resampled identically and top-k versus bottom-k modes within the belief subspace. Belief-carrying directions are the column space of a ridge fit from residual to belief-mode coordinates. We ablate the above-wall models ($d = 2g$).

A.15 Experimental details

Table 9: Architectural control on the $g = 16$ corpus: untied input/output weights, readout bias, and fixed $d_{\text{ff}} = 256$.

Width d	Original gap	Controlled gap	Reduction
8	0.117	0.104	11%
16 ($= r$)	0.039	0.017	56%
32	0.037	0.012	68%
64	0.039	0.013	68%

Architecture. Decoder-only, pre-norm, RoPE (base 10^4), weight-tied input embedding and readout; the readout is bias-free, all other linears carry bias. Per block: causal self-attention ($1/\sqrt{d_h}$ scaling, $d_h{=}d_{\text{model}}/n_{\text{heads}}$) then a GELU MLP of width $4\,d_{\text{model}}$. Dropout 0. Weights initialized $\mathcal{N}(0, (0.02)^2)$, residual-projection weights scaled by $1/\sqrt{2n_{\text{layers}}}$ (GPT-2 residual init). $n_{\text{layers}}{=}4$ for all width sweeps; $n_{\text{heads}}{=}4$, or 1 for the dense even-width fine sweeps. RoPE requires d_h even, so fine sweeps use even d_{model} and head-4 sweeps use d_{model} divisible by 8.

Architectural-control implementation. The control sets `tie_weights=false`, enables a learned readout bias, and overrides the MLP hidden dimension to $d_{\text{ff}} = 256$ at every tested residual width. The main architecture is otherwise unchanged.

Optimization. AdamW ($\beta{=}(0.9, 0.95)$, weight decay 0.01), peak lr 3×10^{-4} with 100-step linear warmup and cosine decay to $0.1\times$ peak, batch size 128, gradient clipping at 1.0. Filtering runs train up to 150,000 steps (group-composition runs 10^5–1.5×10^5, early stopping disabled); early stopping elsewhere uses patience 8000 steps, $\min \Delta{=}10^{-3}$ bits, checked every 500. We evaluate at 36 log-spaced checkpoints (dense early, so acquisition order is resolved).

Corpora. $K{=}128$ mixing clone blow-up at $g \in \{8, 16, 32, 64\}$; ring at $K \in \{16, 128\}$; band-limited ring $K{=}128, b \in \{2, 4\}$. The rank–span separation run uses $g = 16, c = 8, K = 128$, identity clone transitions, and an initial macro–clone correlation, with $r = 16$ and $\dim \text{span}\{b_t\} = 128$; clone-index beliefs update through this correlation but remain predictively null given the macro marginal. Sequence length $T \in \{128, 256\}$ and vocabulary $V \in \{80, 128, 512\}$ for these corpora.

Held-out evaluation uses 2000 sequences per corpus (4000 for ring-$K128$). Depth sweeps use a separate hard-K family at $K \in \{20, 40, 80, 160\}$: an incompressible cycle with $V = 4K$, $T = 128$, and slow mixing ($|\lambda_2| = 0.961$ at $K = 160$), so $r = K$. A $K = 120$ member was generated but is not reported. The width sweep at $g = 16$ additionally uses a dense grid over even d_{model} from 4 to 48.

Probes. Belief R^2, state accuracy, and per-mode R^2 are ridge probes (λ=10^{-3}, 0.7/0.3 train/test split on probe rows, per-run seed) on the residual stream. We use 512 probe sequences, capped at 16,000 decode positions. Per-mode targets are the predictive belief $b_t = P(z_t \mid x_{1:t-1})$ projected onto the top-r eigenvectors of $C = \mathbb{E}[b_t b_t^\top]$ (estimated from 256 sequences). Acquisition-time fits use the stated resolved-mode threshold; the allocation slope uses all modes and is fitted separately per seed. Full-belief R^2 on the rank–span corpus is only a diagnostic because most target variance lies in predictively null clone-shape directions.

Selection and seeds. Three seeds unless specified; we report the best value attained over training. Because above-wall runs can overfit after reaching their optimum (§4.4), we also verify that the wall location is unchanged at a compute-matched checkpoint (14,500, the minimum final step). Three seeds remain thin for censored per-mode statistics, where modes are selected by final decodability before correlations are computed. We report seed spread for each corpus; the direct deficit–eigenvalue result replicates at $g = 16, 32, 64$, while the cross-spectrum cost correlations are treated as complementary evidence.