Belief or Memory? Token-History Controls for Belief-State Probes in Language Models

Noah Mitchell
University of Ottawa

Preprint, November 2026

Abstract

Linear probes recover the Bayesian belief state of a hidden Markov process from transformer activations with high accuracy, and this has been read as evidence that transformers, including pretrained language models reading a sequence in context, compute beliefs. The standard controls for these probes (shuffled targets, random targets, regression on next-token probabilities) cannot separate belief computation from a simpler alternative: a representation that only stores recent tokens. When the hidden process forgets its past, the belief is close to a linear function of recent tokens, so such a representation decodes it well. We test five frozen language models (1.5B to 8B parameters) on coupled two-factor Mess3 processes with a ladder of token-history controls: a decaying token-memory bank, random nonlinear features of that bank, a randomly initialized model, layer-0 activations, and a small transformer trained on the task. On fast-mixing processes, every signature of belief geometry we measured, including factored encoding subspaces whose principal angles shrink with coupling, also appears in these controls. Linear token memory alone explains 0.931–0.971 of belief variance across all mixing rates we tested, which leaves at most 0.069 of variance for any probe to explain beyond it. Measured against that remaining headroom, Qwen2.5-7B closes 74–84% of it, more than random features of up to 4.6 times its width and, at slow mixing, about as much as a transformer trained on the task. Llama-3.1-8B closes 43–72%, and once mixing is slow it does no better than random features of its own token history. We recommend reporting headroom above token memory, and the fraction of it closed, alongside any belief-probe result.

1 Introduction

A transformer trained on next-token prediction from a hidden Markov process can, in principle, solve the task by tracking the posterior over hidden states, the belief state. Shai et al. [16] showed that such transformers represent belief states linearly in their residual stream, including belief geometries with fractal structure. Later work explained the intermediate geometry as constrained belief updating [9], found that transformers trained on processes with independent parts represent each part's beliefs in orthogonal subspaces [15], and found linearly decodable belief states in pretrained language models prompted with sequences from 40 hidden Markov models [1]. The evidence in each case combines high probe R^2, geometric structure that matches theory, and interventions on the probed subspace. The controls compare against shuffled or random belief targets and against regression on the ground-truth next-token probabilities.

These controls rule out a probe that fits noise and a representation that only encodes the next-token distribution. They do not rule out a representation that encodes recent tokens. For a hidden Markov process whose filter forgets its initial condition, the belief at time t is determined, up to small error, by a finite window of recent observations. If it is also close to a *linear* function of features of that window, then any network that keeps a linear record of recent tokens supports a high-R^2 linear belief probe without doing anything that

deserves the name belief computation. Hewitt and Liang [4] raised the corresponding question of selectivity for linguistic probes [see also 2], but belief-probe studies have not yet adopted such a control.

We find that this gap is large. On the coupled two-factor Mess3 processes we study, a ridge regression on a bank of exponentially decaying token embeddings explains 0.931 to 0.971 of belief variance, depending on the mixing rate. Shuffled controls put the floor for a probe at $R^2 \approx 0$; the informative floor is above 0.93. We call the remaining variance the *headroom*, and we measure representations by the fraction of it they close.

Our contributions are:

1. A ladder of token-history controls for belief probes: linear token memory, random nonlinear features of token memory, a randomly initialized network, layer-0 activations, a token-memory MLP as a nonlinear upper reference, and a transformer trained on the task as a positive control (Section 2).
2. Evidence that on fast-mixing processes every belief-geometry signature we measured in five frozen language models also appears in these controls: near-ceiling probe R^2, encoding subspaces whose principal angle shrinks with coupling, gains from bilinear interaction terms, near-zero belief variance after regressing out recent tokens, and next-symbol readouts that lose to token memory (Section 3).
3. A headroom diagnostic that tells, for a given process, how much room a probe has to beat token memory. Headroom peaks at intermediate mixing rates, where it is 2.4 times its value for standard Mess3 (Section 4).
4. Measurements against that headroom showing that two model families differ: Qwen models close more of it than random features of their token history, and Llama-3.1-8B does not once mixing is slow. The result also depends on how the input symbols are written (Section 5).

We do not test the hidden Markov models of Balcells et al. [1], so we make no claim about their specific results. Our claim is that the controls they and others report are not sufficient in principle, and that in the setting studied here adding token-history controls changes the conclusion for one model family and narrows it for another.

2 Setup

Processes. Each factor is a three-state Mess3 process [6] with parameters (x, α): the hidden state stays with probability $1-2x$ and moves to each other state with probability x, and it emits its own symbol with probability α and each other symbol with probability $(1-\alpha)/2$. We couple two factors with strength $\varepsilon \in [0,1]$: whenever their hidden states coincide, factor 2's transitions are drawn from a mixture that puts weight ε on a faster, noisier version of factor 2. The joint process has nine hidden states and nine composite symbols $y = 3y_1 + y_2$. The exact Bayes filter gives the joint belief, and we probe for its two marginals $b_1, b_2 \in \Delta^2$. Appendix A gives every parameter. The *fast* regime uses $(x_1, \alpha_1) = (0.05, 0.85)$, the standard Mess3 setting, and $(x_2, \alpha_2) = (0.08, 0.80)$. The *medium*, *slow* and *very slow* regimes lower the switching rates to $x_1 \in \{0.02, 0.008, 0.002\}$ with $x_2 = 1.5\,x_1$, holding $\varepsilon = 0.5$.

Models and inputs. Our main results use Qwen2.5-7B-Instruct [10], Qwen3-4B [11] and Llama-3.1-8B (base) [3]; Qwen2.5-1.5B-Instruct and Hermes-3-Llama-3.1-8B [17] appear in fast-regime and fixed-layer results. All weights are frozen. Each composite symbol maps to one of the single-token strings ␣A, ..., ␣I, and a trajectory is fed as a sequence of token IDs with no template or instruction. We record the residual stream after each layer at every position.

Probes. A probe is a multi-output ridge regression from activations at one layer to (b_1, b_2). We sample 600 trajectories, split them into 240 training, 120 validation and 240 test trajectories, and use positions $t \geq 32$ of length-256 trajectories (positions $t \geq 16$ of length-192 trajectories in the fast ε-sweep). The layer and the

ridge penalty are chosen jointly on the validation split, and the probe is evaluated once on the test split. We report R^2 per factor and 95% confidence intervals from 500 bootstrap resamples of test trajectories, paired across representations so that differences have their own intervals.

Token-history controls. Each control receives the same symbol sequence as the language model.

- *Linear token memory.* A fixed random embedding $E[y_t]$ is accumulated as $h_t = \lambda h_{t-1} + E[y_t]$. The *long* bank concatenates ten channels with λ from 0 to 0.999 and a cumulative-mean channel, 1,408 dimensions in total; a probe on it is a ridge regression, as for the models. The single-λ banks used in Section 3 have 256 dimensions.
- *Random nonlinear features.* The standardized long bank is passed through a fixed Gaussian projection and a GELU nonlinearity [12], at widths 3,584 (the residual width of Qwen2.5-7B), 8,192 and 16,384, three seeds each. A model that beats these features computes something more belief-aligned than a generic nonlinear expansion of its token history.
- *Random initialization and layer 0.* The Qwen2.5-7B architecture with random weights, probed at its best layer, and the layer-0 (embedding) activations of each pretrained model, which contain no information from other positions [20].
- *Upper and positive references.* A two-layer MLP trained on the long bank shows how much belief information token memory contains when the readout may be nonlinear. A four-layer transformer (width 128) trained from scratch on each slow process is a positive control: it is trained to predict the process, so it should close the headroom if the test can detect belief computation at all.
- *Residualization.* We regress the one-hot identities of the last 18 tokens out of the activations and measure how much activation variance the beliefs still explain.

Headroom. For a process, let R^2_{TM} be the test R^2 of the linear probe on the long token-memory bank. The headroom is $1 - R^2_{\mathrm{TM}}$, and a representation with probe R^2_{rep} closes the fraction

$$\mathrm{HC} = \frac{R^2_{\mathrm{rep}} - R^2_{\mathrm{TM}}}{1 - R^2_{\mathrm{TM}}} \tag{1}$$

of it. $\mathrm{HC} = 0$ means the representation adds nothing a linear record of recent tokens does not already provide, and $\mathrm{HC} = 1$ means it supports a perfect linear belief readout. HC can be negative.

3 On fast-mixing processes, token memory reproduces the evidence

We first ask how much of the usual evidence for belief geometry survives the token-history controls when the process mixes quickly, as Mess3 does at its standard parameters.

Probe accuracy. Belief probes reach R^2 between 0.98 and 0.996 in every pretrained model (Table 1), against -0.054 to -0.006 for shuffled targets. By the standard controls this is strong evidence. A single-decay token-memory bank with $\lambda = 0.5$ reaches 0.967 and 0.973, and the randomly initialized Qwen2.5-7B reaches 0.84 to 0.86. Most of the probe accuracy is available to a representation that does no inference.

Table 1: Fast regime ($x_1 = 0.05$), coupling swept from $\varepsilon = 0$ to $\varepsilon = 1$. Decoding R^2 at $\varepsilon = 1$. The principal angle is the smallest angle between the subspaces that encode b_1 and b_2. The last column is the KL divergence from the Bayes-optimal next-symbol distribution to a linear head trained on observed next symbols (lower is better). Token memory: single-decay bank with $\lambda = 0.5$, whose angle is that of the bank itself; †long bank. Language-model rows select the layer from $\{0, 2, 3, 4, 6\}$ and the ridge penalty from $\{10, 100, 1000\}$ on validation data.

		Decoding R^2	Min. principal angle, $\varepsilon = 0 \to 1$		Next-symbol KL
Representation	Layer	b_1 / b_2	layer 0	selected layer	to Bayes
Token memory ($\lambda = 0.5$)	–	0.967 / 0.973	–	$81.6^\circ \to 70.8^\circ$	$0.0167^\dagger$
Qwen2.5-7B, random init	2	0.856 / 0.840	$88.1^\circ \to 66.4^\circ$	$85.9^\circ \to 65.5^\circ$	–
Qwen2.5-7B	3	0.994 / 0.996	$86.3^\circ \to 65.9^\circ$	$82.7^\circ \to 66.1^\circ$	0.0617
Qwen3-4B	4	0.991 / 0.994	$86.4^\circ \to 65.6^\circ$	$80.2^\circ \to 63.2^\circ$	0.0647
Llama-3.1-8B	6	0.981 / 0.984	$87.4^\circ \to 64.8^\circ$	$80.4^\circ \to 60.8^\circ$	0.0773

Factored geometry. With independent factors ($\varepsilon = 0$), the encoding subspaces of b_1 and b_2 are nearly orthogonal in every model, and the smallest principal angle between them shrinks as coupling grows: from 82.7° to 66.1° in Qwen2.5-7B at its selected layer. This is the pattern one would predict if the model represented each factor's belief in its own subspace and mixed them as the factors become dependent [15]. The same shrinkage appears in layer-0 activations, which see one token at a time (86.3° to 65.9°), in the randomly initialized model (88.1° to 66.4° at layer 0), and in token memory (81.6° to 70.8°). Adding a bilinear $b_1 \otimes b_2$ term to the encoding model raises activation R^2 by 0.087 to 0.117 in the pretrained models, by 0.082 in a randomly initialized model, and by 0.261 in token memory. Both signatures therefore come from the statistics of the composite symbols and the way any token code represents them, not from computation across positions.

Residualization. Beliefs explain 0.203 to 0.291 of activation variance at the probed layer. After the one-hot identities of the last 18 tokens are regressed out, they explain 0.002 to 0.011. This does not separate the language models from a model that computes beliefs: the transformer trained on the task shows the same collapse (0.002 to 0.005). On fast-mixing Mess3 the belief is almost entirely a function of recent tokens, so a test that removes recent tokens removes the belief.

Next-symbol readouts. A linear head on early-layer activations, trained only on observed next symbols, predicts the Bayes-optimal next-symbol distribution far better than the model's own output distribution (KL 0.045 to 0.097 against 0.177 to 0.206; Appendix D). The same head trained on the long token-memory bank does better still (0.0167 against 0.0617 for Qwen2.5-7B at $\varepsilon = 1$). An earlier version of this comparison suggested the opposite for Hermes-3; it came from a solver that had not converged (Appendix D).

In the fast regime, then, each of these signatures also appears in a representation that only stores recent tokens, so none of them can distinguish belief computation from memory. Whether the models compute beliefs is a separate question, which Section 5 takes up where the measurements have room to answer it.

4 How much room does a process leave above token memory?

The informative floor for a belief probe is the accuracy of linear token memory, and it depends on the process. Figure 1a shows the headroom $1 - R^2_{\mathrm{TM}}$ as both factors switch more slowly. It is 0.029 at the standard Mess3 rate, rises to 0.069 at $x_1 = 0.008$, and falls again at the slowest rates. Across all eight processes, linear token memory explains between 0.931 and 0.971 of belief variance.

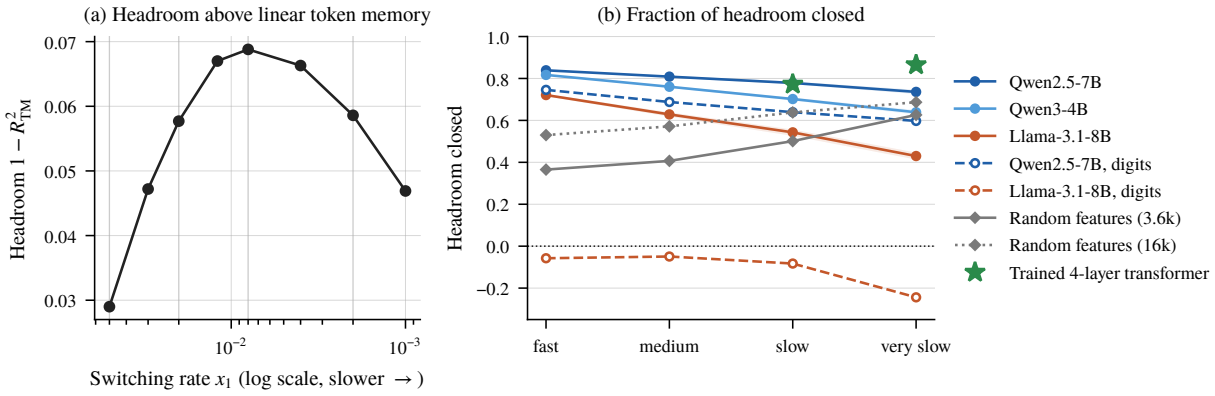

Figure 1: (a) Headroom above linear token memory, $1 - R^2_{\mathrm{TM}}$, as the switching rate of both factors decreases ($x_2 = 1.5\,x_1$, except $x_2 = 1.6\,x_1$ at the fastest rate; $\varepsilon = 0.5$). Grey lines mark the four regimes used in Section 5. (b) Fraction of headroom closed (Equation (1)), averaged over the two factors, at each model's validation-selected layer. Bands: averaged factor-wise 95% bootstrap intervals. Dashed lines: inputs written as bare digits instead of single-token letters (original search grid).

The shape has a simple reading, which we offer as an interpretation rather than a result. When the hidden state switches often, the filter forgets quickly and the belief depends on a short window of tokens. When it almost never switches, the belief depends mainly on symbol counts since the last switch, which the slow channels of the bank record linearly. Between these extremes the belief is a nonlinear function of a long history, and a linear readout of token memory fits it least well.

A probe result on a process should therefore be reported with that process's headroom, since an R^2 of 0.98 means little where token memory already reaches 0.97. For the same reason, processes with intermediate mixing make better test beds: they leave more room for computation to show.

5 Slow mixing separates the models from token memory and from each other

Figure 2 and Table 2 place the language models on the control ladder in the regimes with the most headroom. In the text we average the two factors; the tables give each.

Qwen. At slow mixing, Qwen2.5-7B's layer-4 probe reaches $R^2 = 0.985$ and closes 78% of the headroom, against 50% for random features of the same width. The paired difference, +0.019, has an interval that excludes zero, and widening the random features to 16,384 dimensions still leaves a gap of +0.009 (Appendix B). Qwen3-4B is ahead of both by smaller margins. The transformer trained on the task closes 77%, so at slow mixing Qwen2.5-7B's early layers support about as good a linear belief readout as a network trained for this process. Its advantage over random features holds from fast to slow mixing (Figure 1b).

This is a narrower claim than belief computation in the Bayesian sense. An MLP trained on token memory reaches $R^2 = 0.996$, so the belief is an easy nonlinear function of the token history, and closing headroom is not evidence of an exact filter. The evidence is that Qwen2.5-7B's residual stream makes beliefs more linearly decodable than generic nonlinear features of the same token history do, at up to 4.6 times the width.

Very slow mixing. At $x_1 = 0.002$, random features close 63% of the headroom, more than in any other regime, and Qwen2.5-7B's margin over the 16,384-wide features falls to +0.003, small though its interval still excludes zero. Qwen3-4B does not beat the wider features. The trained transformer closes 87% here, more

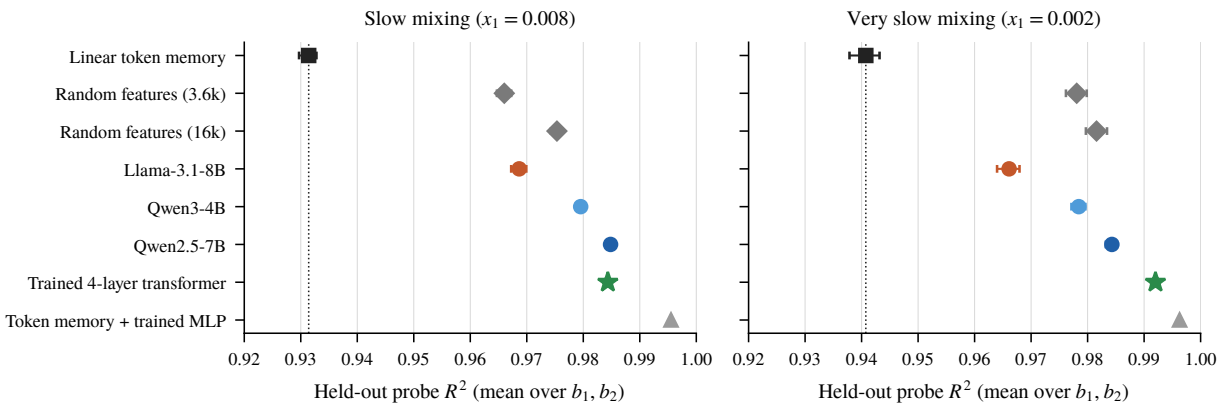

Figure 2: The control ladder at slow and very slow mixing. Probe R^2 averaged over b_1 and b_2, with averaged 95% bootstrap intervals where available. The dotted line marks linear token memory. Random features are random GELU features of the standardized long token-memory bank. The trained transformer is the positive control; the token-memory MLP is a nonlinear upper reference.

than either Qwen model. When the belief depends mostly on symbol counts, generic features of a memory that stores counts are nearly as good as the model.

Llama. Llama-3.1-8B closes 72% of the headroom at fast mixing, 54% at slow mixing and 43% at very slow mixing, at layers 2 and 3. At slow mixing it is level with random features of width 3,584 (+0.003) and below the 16,384-wide features (-0.007); at very slow mixing it is below both. Once mixing is slow, then, its best layer is no more belief-aligned than a random nonlinear expansion of its token history. The probe selection procedure moves this model a long way: earlier procedures put it at the token-memory floor (Appendix C), and its selected penalty is still the smallest in our grid (Section 8).

Random initialization. The randomly initialized Qwen2.5-7B falls well below linear token memory in every regime ($R^2 = 0.773$ at slow mixing), so the architecture alone does not produce the representation.

Input format. Feeding the same sequences as bare digit tokens instead of single-token letters changes the results (Figure 1b, dashed). The digit runs used our original, narrower search grid, so we compare them with letter runs under the same grid (Table 6). At slow mixing Qwen2.5-7B closes 64% of the headroom with digits against 78% with letters, and Llama-3.1-8B falls from 37% to -8%, at or below the token-memory floor in every regime. A run of bare digits reads to a pretrained model like one long number, so the drop is not surprising, but it makes a belief-probe result partly a result about the input format.

6 Interventions

The steering results below come from the fast ε-sweep with fixed layers (Appendix F). We steer factor 1's belief at one position with four methods matched for step norm: a linear step along factor 1's encoding direction, the same step projected away from factor 2's encoding subspace, a geodesic step along a 64-landmark nearest-neighbour approximation of the activation manifold, and nearest-neighbour patching. At a matched reduction of 0.10 nats in factor 1's KL to its target, the orthogonalized linear step and the geodesic step perturb factor 2's predictions by similar amounts (0.0042 and 0.0050 nats in Qwen2.5-7B at $\varepsilon = 1$), less than patching (0.0089), and a random step of the same norm does not move factor 1 (-0.003 nats). Here,

Table 2: Control ladder in the two slowest regimes. ΔR^2 is relative to random features of width 3,584 and 16,384 (paired bootstrap intervals in Appendix B); HC is headroom closed (Equation (1)). Layers and ridge penalties are selected on the validation split. Random-feature rows use the seed shared with the paired intervals; three-seed means differ by at most 0.001 (Table 5).

Representation	Layer	R^2 (b_1 / b_2)	ΔR^2 vs. RF-3.6k	ΔR^2 vs. RF-16k	HC (%)
Slow mixing ($x_1 = 0.008$)					
Linear token memory (ridge)	–	0.936 / 0.927	–	–	0 / 0
Random features, width 3,584	–	0.965 / 0.967	–	−0.010 / −0.009	45 / 55
Random features, width 16,384	–	0.975 / 0.976	+0.010 / +0.009	–	60 / 67
Qwen2.5-7B, random init	2	0.798 / 0.749	−0.167 / −0.219	−0.177 / −0.227	< 0 / < 0
Llama-3.1-8B	3	0.971 / 0.967	+0.006 / 0.000	−0.004 / −0.009	54 / 55
Qwen3-4B	2	0.981 / 0.978	+0.016 / +0.011	+0.006 / +0.002	70 / 70
Qwen2.5-7B	4	0.986 / 0.984	+0.021 / +0.017	+0.011 / +0.008	77 / 78
Transformer trained on task	4	0.986 / 0.982	+0.021 / +0.015	+0.012 / +0.007	78 / 76
Token memory + trained MLP	–	0.996 / 0.995	+0.031 / +0.028	+0.021 / +0.019	94 / 93
Very slow mixing ($x_1 = 0.002$)					
Linear token memory (ridge)	–	0.949 / 0.933	–	–	0 / 0
Random features, width 3,584	–	0.979 / 0.977	–	−0.004 / −0.003	60 / 65
Random features, width 16,384	–	0.983 / 0.980	+0.004 / +0.003	–	67 / 70
Qwen2.5-7B, random init	2	0.878 / 0.825	−0.102 / −0.152	−0.105 / −0.155	< 0 / < 0
Llama-3.1-8B	2	0.972 / 0.961	−0.008 / −0.016	−0.012 / −0.019	44 / 42
Qwen3-4B	2	0.983 / 0.974	+0.003 / −0.002	−0.001 / −0.006	66 / 62
Qwen2.5-7B	3	0.987 / 0.982	+0.007 / +0.005	+0.004 / +0.002	74 / 73
Transformer trained on task	4	0.993 / 0.991	+0.014 / +0.014	+0.010 / +0.011	86 / 87
Token memory + trained MLP	–	0.997 / 0.995	+0.018 / +0.019	+0.014 / +0.015	94 / 93

unlike in the setting studied by Sarfati et al. [14], linear steering along encoding directions is as precise as the manifold-aware step.

We also tested a geometric prediction: if the principal angle between factor subspaces carries the coupling, steering factor 1 should move factor 2 more when the angle is smaller. It does not. Across 20 model-coupling pairs, factor-2 drift is unrelated to the angle ($r = 0.063$, $p = 0.79$). Every method moves factor 2 by 0.0016 to 0.0082 nats at $\varepsilon = 1$, less than the 0.0112 nats that an exact filter's factor-2 prediction would change under the same counterfactual. Subtracting this Bayes-required drift from the measured drift produces an apparent correlation of $r = 0.918$ with the angle, but only because the Bayes-required drift grows with ε while the measured drift stays small. We report this because the subtraction looks like a natural analysis and gives a spurious result.

7 Related work

Belief-state geometry. Shai et al. [16] introduced belief-state probes for transformers trained on hidden Markov processes, Piotrowski et al. [9] derived the intermediate geometry from the constraints of attention, and Shai et al. [15] showed factored representations for processes with independent parts. Balcells et al. [1] extended the analysis to pretrained language models reading sequences in context, with probes compared against shuffled and random targets, next-token-probability regression and short-suffix sequences, and with patching and steering experiments. They note that features other than next-token probabilities have not been ruled out; token history is such a feature, and the ladder here is a way to rule it in or out. Their processes were chosen to mix slowly, which our headroom curve suggests is the right choice. Sarfati et al. [14] study curved belief manifolds in language models and geometry-aware steering.

Probe controls. Control tasks [4], random-initialization baselines [20], information-theoretic and description-length probing [8, 18] and the question of whether decodable information is used [13] are reviewed by Belinkov [2]. A belief state is a deterministic function of the input history, so the natural control is not a control task with random labels but a control representation that has the history and lacks the computation. Token memory and random features of it play that role here. Probing for world state in sequence models [5, 7] faces the same question whenever the state is a simple function of recent inputs.

In-context inference. Language models have been modelled as implicit Bayesian inference over latent concepts [19]. Our results bear on where in a model such inference could be read out, and on how to tell it from memory.

8 Limitations

We study one family of processes, coupled Mess3, with nine symbols, and five models of at most 8B parameters. Probes are fit to unstandardized activations, so a given ridge penalty regularizes models with small activations more strongly. Llama-3.1-8B's and Qwen3-4B's selected penalties are the smallest in our grid, and Llama's validation accuracy was still rising there, so its numbers may be lower bounds; a run with standardized features and a wider penalty grid is under way. The digit-format runs and the fast-regime audit used a narrower search grid than the main sweep. Random GELU features are one family of generic nonlinear features; a different family could do better. The positive control was trained only for the two slowest regimes with the recipe used here; in the fast regime the same architecture, trained differently, did not reach token memory (Appendix E). The steering results use fixed layers rather than validation-selected ones. Results average over positions $t \geq 32$ and may differ early in a sequence. Finally, we have not applied the ladder to the 40 processes of Balcells et al. [1], which is the most direct test of whether these controls change published conclusions.

9 Conclusion

High probe accuracy, factored geometry and residualization tests do not distinguish belief computation from token memory on fast-mixing processes, because a linear record of recent tokens already explains more than 93% of belief variance. Measured against the headroom that remains, Qwen2.5-7B closes most of it, at slow mixing about as much as a transformer trained on the task, and Llama-3.1-8B closes no more than random features of its own token history once mixing is slow. We suggest that belief-probe studies report the headroom above linear token memory for each process, place models on a ladder that includes random features of token memory and a trained positive control, select layers on held-out data, and check that conclusions survive a change of input format.

References

[1] Daniel Balcells, Andrew Jun Lee, Chirag Rastogi, Paul M. Riechers, Adam Shai, and Xavier Poncini. Large language models develop belief state geometry in-context. In *Advances in Neural Information Processing Systems*, 2026. arXiv:2609.17376.

[2] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[3] Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[4] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019.

[5] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *International Conference on Learning Representations*, 2023.

[6] Sarah E. Marzen and James P. Crutchfield. Nearly maximally predictive features and their dimensions. *Physical Review E*, 95(5):051301, 2017.

[7] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2023.

[8] Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[9] Mateusz Piotrowski, Paul M. Riechers, Daniel Filan, and Adam S. Shai. Constrained belief updates explain geometric structures in transformer representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 49399–49419, 2025.

[10] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[11] Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[12] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

[13] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021.

[14] Raphaël Sarfati et al. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors. *arXiv preprint arXiv:2602.02315*, 2026.

[15] Adam Shai, Loren Amdahl-Culleton, Casper L. Christensen, Henry R. Bigelow, Fernando Rosas, Alexander Boyd, Eric A. Alt, Kyle J. Ray, and Paul M. Riechers. Transformers learn factored representations. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 109734–109764, 2026.

[16] 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*, 2024. arXiv:2405.15943.

[17] Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. Hermes 3 technical report. *arXiv preprint arXiv:2408.11857*, 2024.

[18] Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[19] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

[20] Kelly Zhang and Samuel Bowman. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

A Definitions

Mess3 factors. For parameters (x, α), the emission matrix is $E[s, y] = \alpha$ if $y = s$ and $(1-\alpha)/2$ otherwise, the transition matrix is $M[s, s'] = 1 - 2x$ if $s' = s$ and x otherwise, and the labelled transition matrices are $T^{(y)}[s, s'] = E[s, y]\, M[s, s']$.

Coupling. Joint states and symbols are $s = 3s_1 + s_2$ and $y = 3y_1 + y_2$. With $w = \varepsilon$ if $s_1 = s_2$ and $w = 0$ otherwise, the row of the joint labelled matrix for state (s_1, s_2) and symbol (y_1, y_2) is

$$T_1^{(y_1)}[s_1, :] \otimes \left((1-w)\, T_2^{(y_2)}[s_2, :] + w\, T_{2,\text{pert}}^{(y_2)}[s_2, :] \right),$$

where the perturbed factor has $x_{2,\text{pert}} = \min(0.30, x_2 + 0.15)$ and $\alpha_{2,\text{pert}} = \max(0.50, \alpha_2 - 0.20)$. The initial state is drawn from the stationary distribution of $\sum_y T_{\text{joint}}^{(y)}$. Beliefs follow $\eta_t \propto \eta_{t-1} T_{\text{joint}}^{(y_t)}$; b_1 and b_2 are its marginals.

Regimes.

Regime	Factor 1 (x_1, α_1)	Factor 2 (x_2, α_2)	Perturbed factor 2
fast	(0.05, 0.85)	(0.08, 0.80)	(0.23, 0.60)
medium	(0.02, 0.85)	(0.03, 0.80)	(0.18, 0.60)
slow	(0.008, 0.85)	(0.012, 0.80)	(0.162, 0.60)
very slow	(0.002, 0.85)	(0.003, 0.80)	(0.153, 0.60)

The headroom curve in Figure 1a adds four intermediate rates and uses 400 trajectories split 200/80/120.

Inputs. Symbols map to the single-token strings `␣A`–`␣I`, whose token IDs are distinct in each tokenizer. The digit condition feeds the bare digit tokens `1`–`9` directly. No beginning-of-sequence token, template or instruction is added.

Token memory. The single-decay banks use a 9×256 Gaussian embedding with variance $1/256$ and $\lambda \in \{0, 0.3, 0.5, 0.7, 0.9\}$. The long bank concatenates ten channels, with λ taking the values 0, 0.3, 0.5, 0.7, 0.9, 0.95, 0.98, 0.99, 0.995 and 0.999, and a cumulative-mean channel $h_t = \frac{1}{t+1} \sum_{\tau \le t} E_c[y_\tau]$, each 128-dimensional with its own embedding.

Random features. The long bank is standardized with training-split statistics, multiplied by $W \in \mathbb{R}^{1408 \times d}$ with entries $\mathcal{N}(0, 1/1408)$, shifted by a bias drawn from Uniform$(-0.5, 0.5)$, and passed through GELU.

Probes. In the mixing sweep, ridge penalties $\{10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ and candidate layers 0, 2, 3, 4, 6, 8, 10, 12, 14 and 16 for every pretrained model (up to layer 8 for the randomly initialized one). The fast-regime audit and the digit-format runs used penalties $\{10, 100, 1000\}$ and layers $\{0, 2, 3, 4, 6\}$ and $\{2, 3, 4, 6\}$ respectively. The pair (layer, penalty) maximizing mean validation R^2 over the two factors is evaluated once on the test split. Headroom closed uses the same bootstrap resamples as the R^2 values.

Positive control. A Llama-architecture transformer with 4 layers, width 128, 4 heads and MLP width 512, trained with AdamW (learning rate 10^{-3}, weight decay 10^{-2}, batch 64) for 2,500 steps on 6,000 trajectories of the slow or very slow process.

B Additional tables

Table 3: Probe R^2 with 95% bootstrap intervals for every regime (validation-selected layers and penalties). The trained transformer exists only for the two slowest regimes.

Representation	Factor	fast	medium	slow	very slow
Linear token memory	b_1	0.966 [0.966, 0.967]	0.942 [0.941, 0.943]	0.936 [0.934, 0.938]	0.949 [0.946, 0.951]
	b_2	0.976 [0.976, 0.977]	0.943 [0.942, 0.944]	0.927 [0.925, 0.928]	0.933 [0.930, 0.935]
Random features 3,584	b_1	0.975 [0.975, 0.975]	0.962 [0.961, 0.963]	0.965 [0.964, 0.966]	0.979 [0.977, 0.982]
	b_2	0.987 [0.987, 0.988]	0.970 [0.969, 0.971]	0.967 [0.966, 0.968]	0.977 [0.975, 0.978]
Random features 16,384	b_1	0.981 [0.981, 0.982]	0.972 [0.971, 0.973]	0.975 [0.974, 0.976]	0.983 [0.981, 0.985]
	b_2	0.991 [0.991, 0.991]	0.979 [0.978, 0.979]	0.976 [0.975, 0.977]	0.980 [0.978, 0.982]
Qwen2.5-7B, random init	b_1	0.849 [0.847, 0.852]	0.796 [0.792, 0.801]	0.798 [0.789, 0.805]	0.878 [0.866, 0.889]
	b_2	0.845 [0.843, 0.846]	0.760 [0.756, 0.764]	0.749 [0.741, 0.755]	0.825 [0.811, 0.836]
Qwen2.5-7B	b_1	0.994 [0.994, 0.994]	0.989 [0.988, 0.989]	0.986 [0.985, 0.986]	0.987 [0.986, 0.988]
	b_2	0.996 [0.996, 0.997]	0.989 [0.989, 0.989]	0.984 [0.983, 0.985]	0.982 [0.981, 0.983]
Qwen3-4B	b_1	0.993 [0.993, 0.993]	0.986 [0.985, 0.986]	0.981 [0.980, 0.982]	0.983 [0.981, 0.984]
	b_2	0.996 [0.996, 0.996]	0.987 [0.986, 0.987]	0.978 [0.977, 0.979]	0.974 [0.973, 0.976]
Llama-3.1-8B	b_1	0.990 [0.989, 0.990]	0.978 [0.978, 0.979]	0.971 [0.969, 0.972]	0.972 [0.969, 0.974]
	b_2	0.994 [0.994, 0.994]	0.979 [0.978, 0.980]	0.967 [0.965, 0.968]	0.961 [0.958, 0.962]
Trained transformer	b_1	–	–	0.986 [0.986, 0.987]	0.993 [0.992, 0.994]
	b_2	–	–	0.982 [0.982, 0.983]	0.991 [0.990, 0.992]

Table 4: Paired differences from random features of width 3,584 and 16,384, with 95% bootstrap intervals.

Model	Control	Factor	fast	medium	slow	very slow
Qwen2.5-7B	RF-3.6k	b_1	+0.019 [+0.019, +0.020]	+0.027 [+0.026, +0.028]	+0.021 [+0.020, +0.022]	+0.007 [+0.006, +0.009]
		b_2	+0.009 [+0.009, +0.009]	+0.019 [+0.019, +0.020]	+0.017 [+0.016, +0.018]	+0.005 [+0.004, +0.006]
	RF-16k	b_1	+0.013 [+0.013, +0.013]	+0.017 [+0.016, +0.018]	+0.011 [+0.010, +0.012]	+0.004 [+0.002, +0.005]
		b_2	+0.006 [+0.005, +0.006]	+0.010 [+0.010, +0.011]	+0.008 [+0.008, +0.009]	+0.002 [0.000, +0.003]
Qwen3-4B	RF-3.6k	b_1	+0.018 [+0.018, +0.018]	+0.024 [+0.023, +0.025]	+0.016 [+0.015, +0.017]	+0.003 [+0.002, +0.004]
		b_2	+0.009 [+0.009, +0.009]	+0.017 [+0.016, +0.017]	+0.011 [+0.010, +0.012]	−0.002 [−0.003, −0.001]
	RF-16k	b_1	+0.012 [+0.012, +0.012]	+0.014 [+0.013, +0.015]	+0.006 [+0.006, +0.007]	−0.001 [−0.002, +0.001]
		b_2	+0.005 [+0.005, +0.006]	+0.008 [+0.008, +0.009]	+0.002 [+0.001, +0.003]	−0.006 [−0.007, −0.004]
Llama-3.1-8B	RF-3.6k	b_1	+0.015 [+0.014, +0.015]	+0.017 [+0.016, +0.017]	+0.006 [+0.004, +0.007]	−0.008 [−0.009, −0.007]
		b_2	+0.007 [+0.006, +0.007]	+0.009 [+0.008, +0.010]	0.000 [−0.001, +0.001]	−0.016 [−0.017, −0.015]
	RF-16k	b_1	+0.008 [+0.008, +0.009]	+0.006 [+0.006, +0.007]	−0.004 [−0.005, −0.003]	−0.012 [−0.013, −0.010]
		b_2	+0.003 [+0.003, +0.003]	0.000 [0.000, +0.001]	−0.009 [−0.010, −0.008]	−0.019 [−0.021, −0.018]
Trained transformer	RF-3.6k	b_1	–	–	+0.021 [+0.020, +0.023]	+0.014 [+0.012, +0.015]
		b_2	–	–	+0.015 [+0.015, +0.016]	+0.014 [+0.013, +0.016]
	RF-16k	b_1	–	–	+0.012 [+0.011, +0.013]	+0.010 [+0.008, +0.012]
		b_2	–	–	+0.007 [+0.006, +0.007]	+0.011 [+0.010, +0.012]

Table 5: Random features of increasing width (R^2 for b_1 / b_2, mean of three seeds), with Qwen2.5-7B for reference.

Width	fast	medium	slow	very slow
3,584	0.975 / 0.987	0.962 / 0.969	0.965 / 0.967	0.980 / 0.976
8,192	0.978 / 0.989	0.968 / 0.975	0.971 / 0.972	0.982 / 0.979
16,384	0.981 / 0.991	0.972 / 0.979	0.975 / 0.976	0.983 / 0.980
Qwen2.5-7B	0.994 / 0.996	0.989 / 0.989	0.986 / 0.984	0.987 / 0.982
Qwen3-4B	0.993 / 0.996	0.986 / 0.987	0.981 / 0.978	0.983 / 0.974
Llama-3.1-8B	0.990 / 0.994	0.978 / 0.979	0.971 / 0.967	0.972 / 0.961

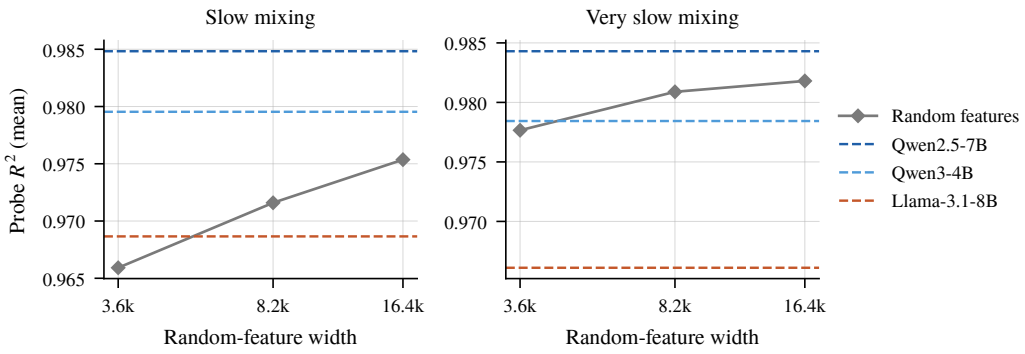

Figure 3: Random-feature width against the language models at slow and very slow mixing.

Table 6: Headroom closed (%, b_1 / b_2) with single-token letters and with bare digit tokens, both under the original search grid (layers $\{2,3,4,6\}$, penalties $\{10,100,1000\}$).

Model	Symbols	fast	medium	slow	very slow
Qwen2.5-7B	letters	82 / 84	81 / 81	77 / 78	73 / 73
	digits	75 / 74	70 / 67	66 / 62	63 / 57
Llama-3.1-8B	letters	50 / 50	47 / 43	39 / 35	19 / 21
	digits	0 / −12	0 / −10	−4 / −13	−18 / −31

C Layer selection

Our first runs fixed the probed layer per model family (layer 3 for Qwen2.5, layer 4 for Qwen3 and Llama). We then selected the layer from $\{2,3,4,6\}$ and the ridge penalty from three values on the validation split, and finally searched layers 0 to 16 and penalties from 10^{-1} to 10^4. Qwen2.5-7B's slow-mixing R^2 barely moved; Llama-3.1-8B's moved from the token-memory floor to more than half of the headroom (Table 7). With the fixed layer, Llama appeared to track linear token memory at every mixing rate, which is the conclusion an earlier draft of this work drew.

Table 7: Slow-mixing probe R^2 for b_1 under successive selection procedures.

Selection procedure	Qwen2.5-7B	Llama-3.1-8B	Llama headroom closed
Fixed layer per model family	L3: 0.982	L4: 0.931	−8%
Layers {2,3,4,6}, penalties {10,100,1000}	L4: 0.986	L6: 0.961	39%
Layers 0–16, penalties 10^{-1}–10^4	L4: 0.986	L3: 0.971	54%

Probe accuracy peaks at layers 2 to 4 in every pretrained model and falls with depth after that (Figure 4). Llama's best penalty is still at the low edge of the grid.

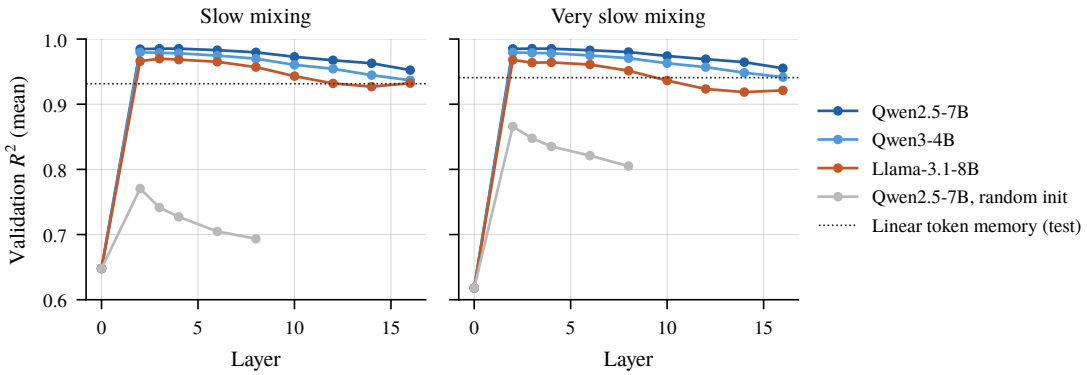

Figure 4: Validation R^2 by layer, taking the best ridge penalty at each layer. The dotted line is the test R^2 of linear token memory.

D Next-symbol readouts and solver convergence

Model	ε	Model output	Current token	Early head (layer)	Token memory	Bayes
Qwen2.5-1.5B	0.0	0.1961	0.0985	0.0471 (2)	0.0406	0.0021
Qwen2.5-1.5B	1.0	0.1996	0.0866	0.0453 (2)	0.0400	0.0029
Qwen2.5-7B	0.0	0.1773	0.0985	0.0679 (2)	0.0406	0.0021
Qwen2.5-7B	1.0	0.1911	0.0866	0.0617 (2)	0.0400	0.0029
Hermes-3-8B	0.0	0.2022	0.0985	0.0969 (2)	0.0406	0.0021
Hermes-3-8B	1.0	0.2061	0.0866	0.0858 (2)	0.0400	0.0029

KL divergence from the Bayes-optimal next-symbol distribution, fast regime. Heads are multinomial logistic regressions trained on observed next symbols with standardized inputs, regularization and layer chosen on validation data, and a full-batch L-BFGS solver whose final gradient norm is checked. The token-memory column uses the single-decay bank with $\lambda = 0.7$; the long bank does better (Table 1). An earlier run used unstandardized inputs and a solver capped at 200 iterations that did not converge on any fit. In that run Hermes-3-8B's early head appeared to beat token memory (0.017–0.021 against 0.039–0.040). The result did not survive standardization and convergence.

E Positive control in the fast regime

An earlier positive control trained the same architecture on 8,000 trajectories of the fast process for 2,500 steps. It reached KL to the Bayes predictor of 0.049 to 0.115 and probe R^2 of 0.927 to 0.956 at its last layer, below linear token memory. It also showed the layer-0 angle shrinkage (75.1° to 59.1°) and the residualization collapse described in Section 3. The slow-regime controls, trained on 6,000 trajectories, reached KL of 0.016 and 0.010.

F Steering protocol

Pairs are drawn from the test split at position $t = 190$ of length-192 trajectories: 80 source and target trajectories whose factor-2 beliefs differ by less than 0.08 and whose factor-1 beliefs differ by more than 0.45

in ℓ_2 norm. Steps are applied at the fixed probed layer and normalized to multiples $s \in \{0.25, 0.5, 1, 2, 4\}$ of the geodesic step's norm. For each method, factor-2 KL drift is interpolated along s to a fixed factor-1 KL reduction of 0.05 or 0.10 nats. The Bayes-required drift comes from the exact filter under the same counterfactual factor-1 belief.