Reasoning Fine-Tuning Induces Persistent Latent Policy States

Abir Harrasse[1*], **Michael Lan**[1*], **Hunar Batra**[2],
Fateme Hashemi Chaleshtori[3], **Chaithanya Bandi**[1]
[1]Martian, [2]University of Oxford, [3]University of Utah
[*]Primary contributors

Abstract

Reasoning-specialized language models show large performance gains over base models, yet the internal changes responsible for improved multi-step reasoning remain poorly understood. It is unclear whether reasoning fine-tuning improves local token-level competence or globally reorganizes how models structure inference over time. We address this question by modeling Chain-of-Thought reasoning as a switching dynamical system (SDS), in which internal representations evolve under discrete latent policy states. Our framework combines time-aware contrastive representation learning with discrete regime discovery to recover latent policies from activation trajectories. Across four benchmarks and model scales from 1.5B to 32B parameters, reasoning-fine-tuned models exhibit richer latent-policy organization than their base counterparts, characterized by more differentiated transition structure and model-dependent changes in state utilization, persistence, and mixing. The recovered regimes exhibit functional specialization aligned with distinct reasoning stages, and extensive controls confirm that their structure is not explained by correctness, representation learning, or modeling priors, but depends on the coherent temporal organization of reasoning trajectories. Causal interventions further show that the regimes are functionally meaningful: state-swap ablations reduce one-step predictive fit, while transplanting reasoning dynamics into base models improves performance on challenging reasoning problems. Finally, SDS-guided pruning of failure-prone reasoning prefixes outperforms self-consistency in 11 of 12 model–dataset settings, with gains of up to 12.5 percentage points. Together, our results suggest that reasoning fine-tuning globally reorganizes latent dynamics, offering a new lens for mechanistic analysis and process-level control of reasoning models. Code: https://github.com/withmartian/mi-cot.

1 Introduction

Large language models exhibit striking improvements in Chain-of-Thought (CoT) reasoning after specialized fine-tuning (DeepSeek-AI, 2025; Shao et al., 2024; Huang et al., 2026). However, the internal changes responsible for these improvements remain poorly understood (Yue et al., 2025; Wu & Choi, 2025; Ward et al., 2025a). In particular, it remains unclear whether reasoning fine-tuning primarily improves local token-level competence or globally reorganizes the latent computational dynamics that structure reasoning over extended horizons. Establishing such structure is challenging because CoT reasoning is a branching temporal process involving stochastic sampling and a distribution of possible trajectories (Macar et al., 2025; Lee et al., 2025).

CoTs have therefore been modeled as dynamical systems that capture long-range transitions between branching reasoning states (Carson, 2025). This perspective suggests that reasoning behavior may be governed by low-dimensional latent policy states that shape future internal dynamics throughout inference.

In this work, we compare fine-tuned reasoning models with their base counterparts by modeling activation trajectories using *Switching Dynamical Systems* (SDS), which recover

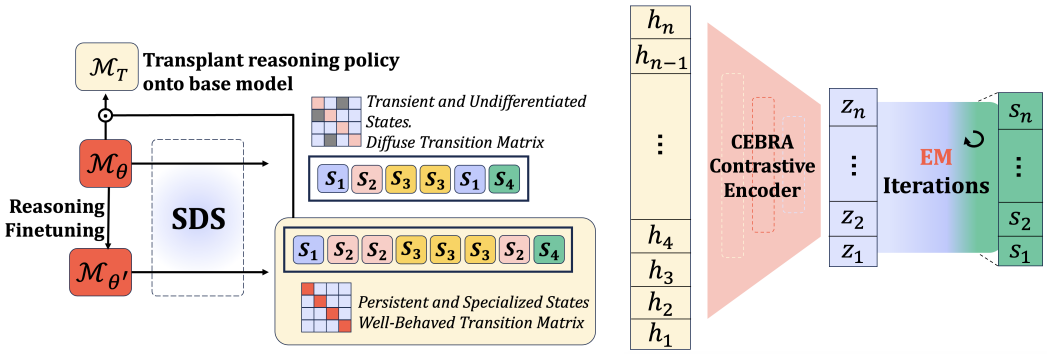

Figure 1: **Left:** Overview of our framework. A base model $\mathcal{M}_\theta$ is fine-tuned for reasoning to produce $\mathcal{M}_{\theta'}$. We fit a Switching Dynamical System (SDS) to each model's activation trajectories and recover sequences of latent regimes. Reasoning fine-tuning reorganizes state utilization, persistence, and mixing while producing more differentiated transition dynamics. The recovered reasoning policy can be transplanted onto the base model $\mathcal{M}_T$ to test its causal and functional relevance. **Right:** Model activations $h_{1:n}$ are encoded into a low-dimensional representation $z_{1:n}$ using a CEBRA contrastive encoder. EM with forward–backward smoothing recovers the discrete latent regime sequence $s_{1:n}$, where $s_t \in \{1, \ldots, K\}$ represents the active reasoning-policy state at step t.

discrete, low-dimensional latent policy states (Ackerson & Fu, 1970; Ghahramani & Hinton, 2000; Linderman et al., 2017). We combine SDS with time-aware contrastive manifold learning (Schneider et al., 2023) and find that reasoning fine-tuning induces richer latent-policy organization, most consistently reflected in more differentiated transition structure. State utilization, persistence, and mixing also change, with their contributions varying across models and datasets. The recovered states exhibit functional specialization aligned with reasoning stages, while predictivity degrades when the state space becomes over-fragmented.

Overall, reasoning fine-tuning appears to globally reorganize internal dynamics, enabling models to maintain and transition between coherent computational modes over extended spans. We call this temporal and functional structure a *latent reasoning policy*, without assuming that it is localized to a particular model component.

Our contributions are:

1. We formalize a dynamical-systems framework for uncovering differences in internal reasoning structure between base and reasoning-fine-tuned models.
2. We provide empirical evidence that reasoning fine-tuning induces richer latent-policy organization, characterized by more differentiated transition structure and model-dependent changes in state utilization, persistence, and mixing.
3. We introduce CEBRA for latent reasoning-state analysis and show that, in our setting, it provides a stronger balance of predictive fit and structured regime recovery than PCA.
4. We show that the inferred latent reasoning policy is actionable: causal state-swap and policy-transplantation interventions establish the functional relevance and transferability of the recovered dynamics, while SDS-guided pruning of failure-prone reasoning prefixes (PREFIXGUARD) improves over self-consistency in 11 of 12 model–dataset settings, with gains of up to 12.5 percentage points.

2 Related Work

Comparing Base and Reasoning-Fine-Tuned Models. Reasoning fine-tuning typically combines supervised fine-tuning, distillation, and reinforcement learning with verifiable

rewards (RLVR) (Lambert et al., 2025; DeepSeek-AI, 2025). A growing line of work argues that RLVR improves reasoning by reallocating probability mass toward successful trajectories already supported by the base model, rather than by creating entirely new reasoning procedures (Yue et al., 2025; Wu & Choi, 2025; Nguyen et al., 2025; Zhu et al., 2025). While RL-trained models typically outperform their base counterparts at small sampling budgets, base models can sometimes match or exceed them at large pass@k, suggesting that RLVR improves trajectory ranking and sampling efficiency while narrowing exploration (Yue et al., 2025; Nguyen et al., 2025; Ward et al., 2025a). Wu & Choi (2025) characterize this effect through support and entropy, while Wang et al. (2025a) show that RLVR acts disproportionately on high-entropy tokens corresponding to critical forks in reasoning trajectories (Huang et al., 2026). These findings concern how post-training changes which reasoning chains are selected; we instead ask how it changes the internal dynamics through which those chains unfold.

Mechanistic Interpretability and Control of Reasoning. Previous mechanistic interpretability work suggests that reasoning fine-tuning repurposes latent representations already present in base models rather than constructing wholly new machinery (Ward et al., 2025a; Minder et al., 2026; Ward et al., 2025b; Muhamed et al., 2025; Wang et al., 2025b). Venhoff et al. (2025b) further argue that base models already contain many relevant reasoning behaviors and that post-training primarily teaches the model when to deploy them. At a finer granularity, thinking tokens and thought anchors identify decision points that disproportionately shape subsequent reasoning (Qian et al., 2025; Bogdan et al., 2025), while steering-vector methods localize and manipulate individual reasoning behaviors (Venhoff et al., 2025a). Hidden-state clustering also provides evidence that reasoning unfolds through discrete internal stages (Liang et al., 2025).

Related methods use such trajectory structure for inference-time improvement. CTRLS models CoT reasoning as latent state transitions and uses distributional reinforcement learning for state-aware exploration (Wu et al., 2026), while SEAL separates execution, reflection, and transition thoughts in latent space and steers representations toward more productive reasoning (Chen et al., 2025). Unlike methods that model, rank, or control trajectories within one model, we ask how reasoning fine-tuning reorganizes latent dynamics relative to its base model. Our SDS formulation addresses this through state utilization, persistence, transition organization, and predictive dynamics, followed by causal policy transplantation and process-level pruning of failure-prone reasoning trajectories.

Dynamical and Generative Models of Internal Representations. Our approach builds on theoretical work establishing conditions under which switching dynamical systems and related Markov-switching models can be recovered from observational trajectories (Ackerson & Fu, 1970; Ghahramani & Hinton, 2000; Linderman et al., 2017). Balsells-Rodas et al. (2024) provide conditions under which SDS parameters and discrete regimes are identifiable from data. Complementarily, Luo et al. (2026) train diffusion models on residual-stream activations to capture the distribution of a network's internal states and provide priors for more faithful interventions. We combine an SDS with contrastive representation learning (Schneider et al., 2023), using CEBRA to obtain embeddings that empirically support stronger regime recovery than PCA in our setting.

Concurrently, Carson (2025) show that sentence-level hidden states can be modeled using a low-dimensional switching linear dynamical system, where regimes capture distinct drift and transition patterns in reasoning traces. Whereas their PCA-based state space studies regimes within individual models, we use CEBRA-derived embeddings to compare matched base and reasoning-fine-tuned models and test whether the recovered differences are causally and practically actionable.

3 Reasoning as a Switching Dynamical System

3.1 Problem Setup

We study whether reasoning in LLMs can be described by persistent, low-dimensional latent policy states that govern future internal activation dynamics. We operationalize

this as a predictive latent-variable problem: we hypothesize a discrete latent **state** (i.e., **regime**) $s_t \in \{1, \ldots, K\}$ that (i) persists over time, (ii) governs the evolution of internal representations, and (iii) is identifiable from observed activation trajectories. Let $h_t \in \mathbb{R}^d$ denote the residual stream activation at reasoning step t, where each step corresponds to a deterministically segmented sentence in a model-generated chain-of-thought trace. We obtain traces by sampling complete model responses to each reasoning prompt, then split each trace into sentence-level units using punctuation-based boundaries; h_t is extracted at the final token of the t-th sentence. The choice of CEBRA over PCA for dimensionality reduction and EM over a Mixture-of-Experts formulation for parameter estimation are both justified empirically in Appendix E.4.

3.2 Switching Dynamical System Formulation

We model reasoning as a latent Markov process in which the discrete policy state s_t governs the local evolution of the continuous internal representation z_t:

$$P(s_{t+1} = j \mid s_t = i) = T_{ij}, \tag{1}$$

$$P(z_{t+1} \mid z_t, s_t = i) = p_i(z_{t+1} \mid z_t). \tag{2}$$

In the linear-Gaussian instantiation used for inference, the dynamics take the form:

$$z_{t+1} = A_{s_t} z_t + b_{s_t} + \varepsilon_t, \quad \varepsilon_t \mid s_t \sim \mathcal{N}(0, \Sigma_{s_t}), \tag{3}$$

with $\{s_t\}$ an irreducible, aperiodic Markov chain with transition matrix $T^\star$. We further assume nontrivial self-transition $p_{\text{stay}} > 1/K^\star$ (persistent regimes) and distinct regime parameters for $i \neq j$ (full theory and formal assumptions appear in Appendix B).

3.3 Dimensionality Reduction via CEBRA

Raw activations h_t are projected into observed d-dimensional embedding trajectories z_t using a CEBRA-style contrastive encoder (Schneider et al., 2023). Positive pairs are temporally adjacent steps from the same problem; negatives are temporally distant steps from the same problem or steps from different problems. The encoder is trained with an InfoNCE objective:

$$\mathcal{L}_{\text{NCE}} = -\log \frac{\exp(\tilde{z}_t^\top \tilde{z}_{t^+}/\tau)}{\exp(\tilde{z}_t^\top \tilde{z}_{t^+}/\tau) + \sum_{t^-} \exp(\tilde{z}_t^\top \tilde{z}_{t^-}/\tau)}, \tag{4}$$

where embeddings are ℓ_2-normalized and τ is a temperature parameter. This encourages local temporal coherence in the embedding space, ensuring that discovered regimes reflect functional reasoning transitions rather than superficial token similarity.

3.4 SDS Parameter Estimation via EM

We estimate the SDS parameters $\theta = \{\pi, T, \{A_k, b_k, \Sigma_k\}_{k=1}^K\}$ from the observed CEBRA embedding trajectories $z_{1:T}$ via Expectation-Maximization (Ghahramani & Hinton, 2000). At each iteration, the E-step applies the forward-backward algorithm to compute soft regime posteriors $\gamma_t(k) = P(s_t = k \mid z_{1:T})$, and the M-step updates the regime-conditional dynamics via weighted least squares and the transition matrix via soft transition counts. The discrete regime sequence is then recovered as $\hat{s}_t = \arg\max_k \gamma_t(k)$, and the model order $K^\star$ is selected by minimizing BIC on the fitted SDS likelihood. Full derivations and update equations are provided in Appendix A, with identifiability analysis provided in Appendix C.

3.5 Why CEBRA Outperforms PCA

Under mild assumptions on the SDS dynamics, the prediction risk decomposes into an irreducible noise term and a misclassification term controlled by the minimum pairwise Bhattacharyya distance $B(f)$ between regime-conditional distributions. This gives a closed-form criterion for when CEBRA outperforms PCA. The full derivation appears in Appendix B:

Corollary 3.1 (CEBRA vs. PCA)**.** *$\mathcal{R}_{\text{SDS}}(f_C) < \mathcal{R}_{\text{SDS}}(f_P)$ if and only if $B(f_C) - B(f_P) > \log \frac{M(f_C)}{M(f_P)}$.*

We verify this condition empirically across all model–dataset pairs (Appendix B.1). Reasoning models satisfy it substantially more often than base models, indicating that the representation advantage of CEBRA is strongest in the models where structured latent dynamics have emerged.

3.6 Validating the Markovian Assumption

A core modeling assumption is that the decoded regime sequence $\{s_t\}$ follows a first-order Markov chain. We verify this using two complementary tests. The first is a Chapman–Kolmogorov residual test: for a first-order chain, the two-step transition matrix $T^{(2)}_{ij} = P(s_{t+2} = j \mid s_t = i)$ must satisfy $T^{(2)} \approx T^2$, where T is the fitted one-step transition matrix. We measure the normalized Frobenius residual $\mathrm{CK} = \|T^2 - \hat{T}^{(2)}\|_F / \|\hat{T}^{(2)}\|_F$ and compare it against a permutation null that destroys temporal dependencies while preserving marginal state frequencies. The second is an order-2 BIC test comparing first- and second-order Markov models on the decoded sequences, with positive ΔBIC indicating preference for the simpler first-order model.

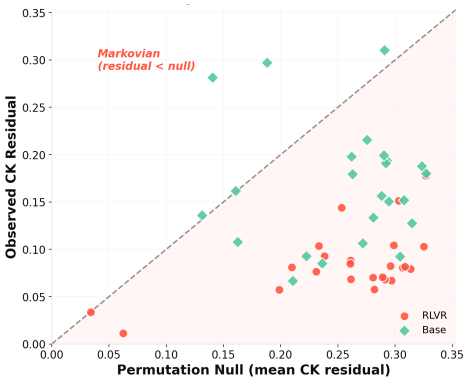

Figure 2: CK residuals against the permutation null. Points below the diagonal indicate that the decoded regime sequence satisfies the Chapman–Kolmogorov relation more closely than a memoryless baseline.

Reasoning models concentrate well below the diagonal in Figure 2, indicating stronger first-order temporal consistency than permuted controls. Base models are more diffuse, consistent with weaker temporal organization. The order-2 BIC test corroborates this: first-order models are preferred in the large majority of reasoning-model runs, supporting the use of a first-order switching process as an adequate description of the recovered latent policies.

3.7 Why Reasoning Fine-Tuning Induces Persistent Latent States

One possible explanation for the emergence of persistent regimes in reasoning models is that trajectory-level rewards under RLVR may favor coherent multi-step strategies, implicitly creating a positive continuation margin $\gamma_i(x_t) := V_i(x_t) - \max_{j \neq i} V_j(x_t)$ over large regions of state space and biasing the policy toward self-transition. Pretraining via maximum likelihood exerts no such pressure, as next-token prediction does not reward consistency of latent modes across time. Under this view, RLVR does not create latent structure from scratch but rather amplifies and stabilizes structure already present in the base model by making persistence directly reward-relevant. We formalize this argument in Appendix C.2 (Proposition C.7). Empirical support for this progressive emergence is provided by the training dynamics analysis in Appendix D.

4 Experimental Setup

We evaluate on GSM8K (7473 samples) (Cobbe et al., 2021), MATH-500 (500 samples) (Hendrycks et al., 2021), SVAMP (800 samples) (Patel et al., 2021), and MMLU-Pro (4000 samples) (Wang et al., 2024) using chain-of-thought traces, and study paired base and DeepSeek-R1 distilled reasoning models (DeepSeek-AI, 2025) for: Llama-3.1-8B (AI@Meta, 2024), Qwen2.5-14B (Bai et al., 2023), and Qwen2.5-Math-1.5B (Yang et al., 2024a;b). To examine whether these patterns extend to a larger scale, we additionally compare Qwen2.5-32B with its reasoning-specialized counterpart QwQ-32B using the same evaluation protocol; further details are provided in Appendix E.1. Activations are extracted at the sentence level using the last-token representation of each sentence at a single selected layer per model; the choice of layer is justified in Appendix E.3.

Metrics. Let $s_{1:T}$ be the inferred discrete state sequence and A the estimated transition matrix. **Persistence** measures temporal stickiness as Persistence $= \frac{1}{T-1}\sum_{t=1}^{T-1}\mathbf{1}\{s_{t+1} = s_t\}$. The **effective number of states** $K_{\text{eff}} = \exp(H(\pi))$ summarizes practical state usage via entropy-equivalent occupancy, where π_k is the empirical state frequency. **Transition structure (TVD)** quantifies deviation from random switching as the mean total variation distance between each transition row and a uniform reference. The **spectral gap** SpecGap $= 1 - |\lambda_2(A)|$ serves as a mixing proxy, with smaller values indicating slower mixing and stronger regime persistence.

5 Experimental Validation of SDS Model

5.1 Model Order Selection and Predictive Gain from Additional Regimes

To test whether additional latent regimes provide genuinely useful predictive structure, we sweep the number of regimes K and measure the gain in one-step predictive fit relative to a linear autoregressive baseline. We report ΔR^2 over linear AR and mark the BIC-selected model order for each run in Figure 3.

Across model families, ΔR^2 increases as the number of regimes grows from $K = 1$ to a small multi-regime model, after which gains taper. This pattern supports the claim that reasoning traces are better explained by a low-cardinality switching process than by a single linear dynamical mode.

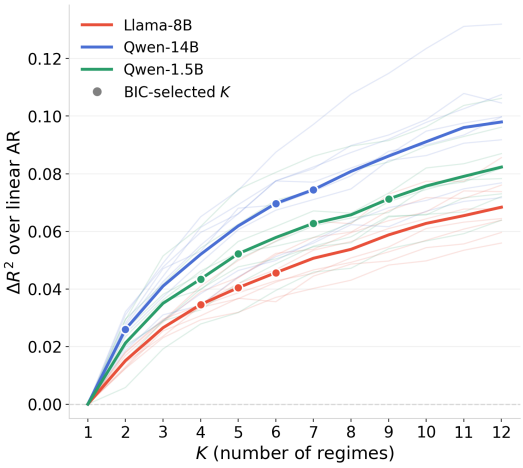

Figure 3: Predictive gain from increasing latent regimes. The y-axis shows ΔR^2 vs. a linear autoregressive baseline, with markers indicating BIC-selected K. Across models, performance improves with a small number of regimes, supporting discrete latent policy states over a single-mode baseline.

5.2 State-Swap Ablations Across Depth

To test whether the discovered regimes are functionally necessary rather than arbitrary labels, we perform a state-swap ablation. After fitting the SDS and estimating per-regime dynamics $\{\hat{A}_k, \hat{b}_k\}$, we evaluate prediction under the original assignments R^2_{id} and under randomly permuted assignments $\bar{R}^2_{\text{rand}}$, where the fitted dynamics are held fixed but regime labels are shuffled. We report the drop $\Delta = R^2_{\text{id}} - \bar{R}^2_{\text{rand}}$, averaged over $N = 50$ permutations. Across all model families and datasets, permuting the assignments produces a substantial drop in predictive R^2, confirming that the correct dynamics model must be applied at the correct time step and that the discovered regimes reflect genuinely distinct dynamical modes rather than spurious clustering. The drop is consistently larger in later layers, consistent with a progressive consolidation of latent policy structure with depth. Full results are in Appendix E.5.

Robustness to correctness and temporal controls. We test whether these differences can be attributed to output correctness, temporal biases in the representation-learning objective, or the SDS persistence prior. Restricting the analysis to problems solved correctly by both models preserves the difference in latent-state organization. Replacing adjacent CEBRA positives with randomly selected non-adjacent steps also preserves the reasoning–base separation in p_{stay} and TVD across all 12 model–dataset settings, while varying or completely removing the Dirichlet persistence prior produces negligible changes. Conversely, shuffling sentence order sharply reduces segment persistence and causes predictive gains over a linear autoregressive baseline to become negative, confirming that the recovered regimes

	K_{eff} ↑		p_{stay} ↑		TVD↑		Spec. Gap↓	
	Base	Reasoning	Base	Reasoning	Base	Reasoning	Base	Reasoning
Llama-8B	4.60 ±1.32	**4.65** ±0.73	0.70 ±0.07	**0.82** ±0.05	0.48 ±0.07	**0.60** ±0.03	0.20 ±0.09	**0.10** ±0.05
Qwen-1.5B	2.60 ±0.66	**5.30** ±1.27	0.78 ±0.12	**0.79** ±0.06	0.39 ±0.06	**0.61** ±0.04	0.24 ±0.10	**0.07** ±0.02
Qwen-14B	3.00 ±0.71	**5.80** ±1.86	**0.78** ±0.07	0.77 ±0.05	0.43 ±0.07	**0.58** ±0.10	0.17 ±0.13	**0.12** ±0.09
Qwen2.5/QwQ-32B	5.25 ±1.25	**5.59** ±1.11	0.74 ±0.08	**0.79** ±0.04	0.55 ±0.08	**0.60** ±0.05	0.14 ±0.06	**0.11** ±0.04

Table 1: Per-model comparison of structural switching metrics between base and reasoning-fine-tuned models, averaged across datasets. Values report mean ± pooled population standard deviation across datasets and random seeds. Light-green cells and bold text indicate the numerically better value within each model pair; they do not denote statistical significance. Reasoning fine-tuning consistently produces more differentiated transition structure, while changes in state utilization, persistence, and mixing vary across model families. Per-dataset results and additional robustness analyses are provided in Appendix E.1.

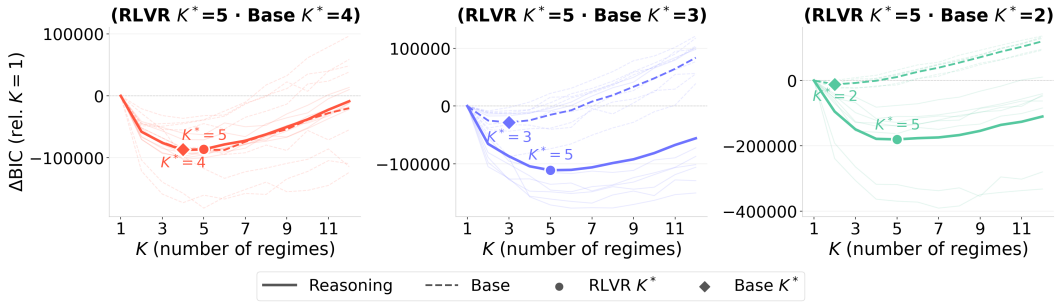

Figure 4: BIC curves comparing reasoning and base models across model families. Bold lines show family means; markers denote the BIC-selected K^*. Reasoning models consistently select larger K^* than base models, supporting the claim that reasoning fine-tuning induces a richer multi-regime latent-policy structure.

depend on coherent temporal organization rather than static activation geometry alone. Full results are reported in Appendix E.2.

6 Comparing Latent Policy Regimes in Base vs Fine-Tuned Models

Table 1 reports structural switching metrics averaged across datasets for each model pair. Reasoning fine-tuning increases TVD across all model families, providing the most consistent evidence of more differentiated transition structure. The remaining metrics vary across families: Qwen-1.5B and Qwen-14B show substantially greater effective state utilization, Llama-8B exhibits higher persistence and slower mixing, and QwQ-32B improves over Qwen2.5-32B across all four metrics. Overall, these results indicate a broader reorganization of latent-policy dynamics rather than a uniform increase in persistence.

6.1 Emergence of Persistent Latent Regimes

To determine how models operate via discrete policies, we vary the number of regimes K and measure state persistence and transition structure in Figure 4. The single-model persistence-cliff pattern is mirrored by the broader model-family comparison using BIC, consistent with the theoretical prediction that such a cliff should emerge under multi-regime dynamics (Appendix C.1). Reasoning models consistently prefer larger effective regime counts than their base counterparts, indicating that reasoning fine-tuning supports a richer latent policy structure. This separation persists across datasets (Figure 5), showing that the effect is not driven by a single benchmark. The dataset-level curves sharpen the same conclusion: GSM8K, SVAMP, MATH-500, and MMLU-Pro all exhibit lower BIC at larger K for reasoning models than for their base counterparts, indicating that the preference

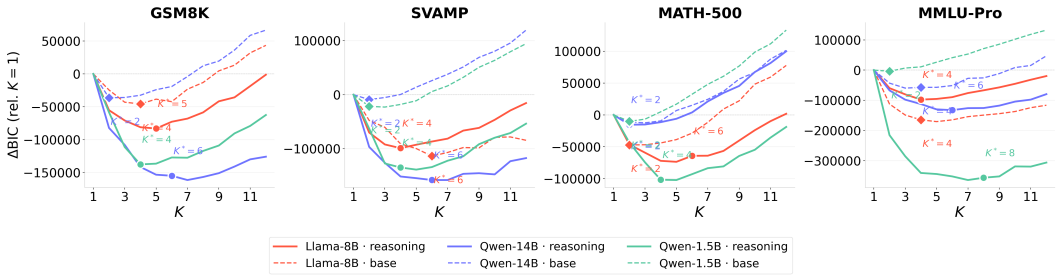

Figure 5: Per-dataset BIC curves for reasoning and base models across all model families. Dashed lines denote base models and solid lines denote reasoning models; thick lines show dataset-level averages. The reasoning preference for larger K^* persists across datasets, indicating that the emergence of multi-regime structure is not limited to a single benchmark.

for multi-regime dynamics is stable across both mathematical and broader knowledge benchmarks. We further track how this multi-regime structure develops during training in Appendix D, finding that K_{eff} and transition structure emerge rapidly in early training before consolidating.

Furthermore, these patterns are not dataset-specific: fitting a shared CEBRA encoder on GSM8K and transferring it to SVAMP and MMLU-Pro yields high cross-dataset consistency ($S_T \geq 0.85$, $S_C \geq 0.76$), confirming that the recovered policy states reflect intrinsic properties of the model's reasoning behavior rather than surface statistics of the training distribution (Appendix E.7).

6.2 Geometry of Latent Policies

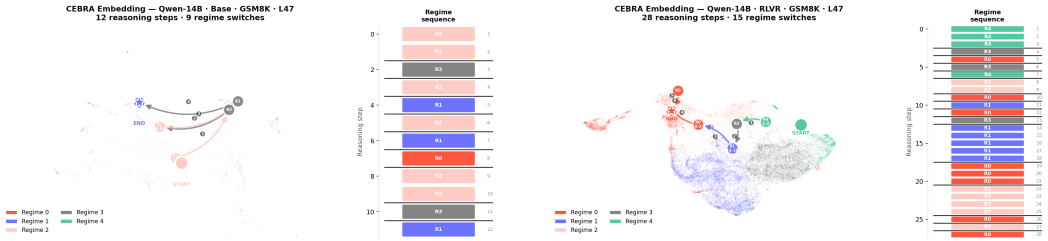

Figure 6: Representative CEBRA trajectories for Qwen-14B on GSM8K. *Left:* base model. *Right:* reasoning model. The base trajectory oscillates among nearby $R2$, $R3$, and $R1$ regions with limited geometric separation, while the reasoning trajectory begins in a distinct green cluster and proceeds through blue and red regimes toward a compact terminal region.

Representative CEBRA trajectory visualizations provide a complementary geometric view of these recovered policies (Figure 6). Across model families, base-model traces typically either remain trapped within a narrow subset of the embedding space or switch irregularly among nearby regions, yielding short or weakly ordered regime segments. Reasoning trajectories, by contrast, more often traverse separated clusters in a stage-like progression from the initial reasoning state toward terminal solution states, with longer contiguous segments and more coherent transitions between clusters. Qualitative examples spanning all model families and both mathematical benchmarks are reported in Appendix J, where the regime-sequence panels align the geometric trajectories with the decoded latent-state assignments.

7 Functional Specialization of States

To better characterize how states align with reasoning, we manually define *reasoning stages*, similar to behaviors derived by (Venhoff et al., 2025a), and then annotate each sentence in

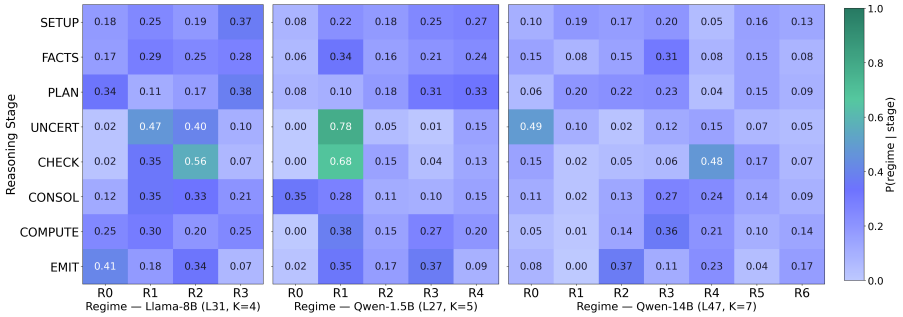

Figure 7: Row-normalized contingency matrix $P(\text{regime} \mid \text{stage})$ across three models (**left**: Llama-8B, **middle**: Qwen-1.5B, **right**: Qwen-14B), showing how human-defined reasoning stages distribute across discovered latent regimes and revealing distinct specialization patterns across the recovered policy states.

traces with both a reasoning stage via Qwen2.5-7B-Instruct (Yang et al., 2024b) and a state via SDS. This begets regime distributions per reasoning stage.

Figure 7 shows the row-normalized contingency matrix $\mathbb{P}(\text{regime} \mid \text{stage})$ between inferred latent regimes and human-defined reasoning stages across three models on GSM8K. Since we established that regime structure is consistent across datasets for a given model (Appendix E.7), we restrict this analysis to GSM8K without loss of generality. The recovered regimes exhibit clear functional specialization across all three models. We provide an extended analysis of this result in Appendix F.

8 Latent Policies Are Actionable

8.1 Causal Validation via Policy Transplantation

To test whether the recovered states are functionally meaningful, we transplant the latent policy learned from a reasoning-fine-tuned model into its base counterpart. We fit an SDS to the reasoning model's activation trajectories and periodically apply a KL-regularized nudge to the base model's residual stream, steering its dynamics toward transitions characteristic of the reasoning model. This intervention improves performance on challenging problems that the unsteered base model consistently fails, showing that the recovered dynamics can causally alter downstream reasoning without modifying model weights. Full implementation details and results are provided in Appendices H and I.

8.2 SDS-Guided Pruning of Failure-Prone Prefixes

The policy-transplantation experiment provides a causal test of whether the recovered states influence reasoning, but it is not intended as a practical alternative to using the reasoning model itself. We therefore ask whether the learned SDS structure can identify and prune failure-prone trajectories during inference.

For each latent state s, we estimate a utility $V(s)$ from correctness outcomes. States with $V(s) < 0$ are associated more strongly with unsuccessful trajectories. We define the *detour score* of a trajectory $h_{1:T}$ as the fraction of steps assigned to such states:

$$D(h_{1:T}) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}[V(s_t) < 0] \,. \tag{5}$$

A high detour score indicates that a trajectory spends substantial time in regimes historically associated with failure.

Based on this signal, we introduce PREFIXGUARD, a fixed process-level controller applied uniformly across models and datasets. We generate four partial rollouts up to 25% of the

Model	Dataset	Self-Consistency	PREFIXGUARD	Gain	Oracle
Llama-8B	GSM8K	60.9 ± 3.1	$\mathbf{68.8 \pm 4.4}$	$+\mathbf{7.9}$	94.3 ± 0.9
	MATH-500	18.8 ± 4.4	$\mathbf{25.5 \pm 7.6}$	$+\mathbf{6.7}$	54.7 ± 2.3
	MMLU-Pro	22.4 ± 3.7	$\mathbf{26.0 \pm 4.8}$	$+\mathbf{3.6}$	70.3 ± 3.1
	SVAMP	54.2 ± 8.2	$\mathbf{58.3 \pm 6.2}$	$+\mathbf{4.1}$	94.4 ± 3.9
Qwen-14B	GSM8K	87.5 ± 4.2	$\mathbf{91.7 \pm 0.0}$	$+\mathbf{4.2}$	91.7 ± 0.0
	MATH-500	5.5 ± 3.4	$\mathbf{10.2 \pm 5.1}$	$+\mathbf{4.7}$	21.1 ± 4.1
	MMLU-Pro	$\mathbf{30.2 \pm 2.3}$	28.1 ± 5.4	-2.1	66.7 ± 3.9
	SVAMP	87.5 ± 21.7	$\mathbf{100.0 \pm 0.0}$	$+\mathbf{12.5}$	100.0 ± 0.0
Qwen-1.5B	GSM8K	20.3 ± 3.1	$\mathbf{20.8 \pm 2.9}$	$+\mathbf{0.5}$	44.3 ± 3.1
	MATH-500	13.5 ± 2.3	$\mathbf{16.7 \pm 2.6}$	$+\mathbf{3.2}$	34.9 ± 2.7
	MMLU-Pro	8.9 ± 2.7	$\mathbf{9.9 \pm 1.7}$	$+\mathbf{1.0}$	41.1 ± 2.7
	SVAMP	68.8 ± 6.2	$\mathbf{76.6 \pm 6.8}$	$+\mathbf{7.8}$	89.1 ± 2.7

Table 2: Accuracy of SDS-guided prefix pruning compared with standard self-consistency. Values are mean $\pm$ standard deviation. Gain denotes the percentage-point difference between PREFIXGUARD and self-consistency. Oracle selects a correct trajectory whenever one exists in the candidate pool and serves only as an upper bound.

token budget and score each prefix using a fixed combination of CEBRA-based reasoning quality, SDS transition value, transition likelihood, and a penalty for time spent in detour states. We retain the highest-scoring prefix, corresponding to 75% pruning, continue it to completion, generate additional restart trajectories, and select the final answer using a fixed rank-based hybrid score. PREFIXGUARD is not a learned selector, and we perform no dataset-specific tuning or post-hoc selection of its scoring rule.

As shown in Table 2, PREFIXGUARD improves over self-consistency in 11 of 12 model–dataset settings, with an average gain of 4.5 percentage points and a maximum gain of 12.5 points. The only decrease occurs for Qwen-14B on MMLU-Pro. The remaining gap to the oracle indicates that the candidate pool frequently contains a correct trajectory that the fixed scoring rule does not select. Thus, the learned SDS structure already identifies promising prefixes, while final trajectory selection remains the principal bottleneck. Together with policy transplantation, these results show that the recovered states are both causally meaningful and practically useful for avoiding failure-associated reasoning regimes.

9 Conclusion

We employ a dynamical-systems approach to compare reasoning-fine-tuned models with their base counterparts through latent policy states. Across models and benchmarks, reasoning fine-tuning induces richer latent-policy organization, characterized by more differentiated transitions and model-dependent changes in state utilization, persistence, and mixing. Causal interventions establish the functional relevance of these dynamics, while SDS-guided prefix pruning demonstrates their practical utility.

Our experiments focus on mathematical, symbolic, and multi-step reasoning with explicit CoT traces. We do not claim that the same structure applies to retrieval, tool use, long-horizon planning, agentic or multimodal reasoning, or reasoning without observable intermediate steps. Future work can test these settings and investigate how latent-state transitions relate to high-entropy forks in reasoning trajectories (Qian et al., 2025).

Acknowledgments

We thank Iván Arcuschin, Jake Ward, Florent Draye and members of Martian for helpful discussions. We also thank the Cosmos Institute for providing computational support.

References

G. Ackerson and K. Fu. On state estimation in switching environments. *IEEE Transactions on Automatic Control*, 15(1):10–17, 1970. doi: 10.1109/TAC.1970.1099359.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter?, 2025. URL https://arxiv.org/abs/2506.19143.

Jack David Carson. A statistical physics of language model reasoning. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=MbJXVbwSir.

Runjin Chen, Zhenyu Zhang, Junyuan Hong, Souvik Kundu, and Zhangyang Wang. SEAL: Steerable reasoning calibration of large language models for free. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=klPszYDIRT.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Randal Douc, Éric Moulines, and Tobias Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with markov regime. *The Annals of Statistics*, 32(5), October 2004. ISSN 0090-5364. doi: 10.1214/009053604000000021. URL http://dx.doi.org/10.1214/009053604000000021.

Zoubin Ghahramani and Geoffrey E Hinton. Variational learning for switching state-space models. *Neural computation*, 12(4):831–864, 2000.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Kexin Huang, Haoming Meng, Junkang Wu, Jinda Lu, Chiyu Ma, Ziqian Chen, Xue Wang, Bolin Ding, Jiancan Wu, Xiang Wang, Xiangnan He, Guoyin Wang, and Jingren Zhou. On the direction of RLVR updates for LLM reasoning: Identification and exploitation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r6Pw3RiMYL.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Seongyun Lee, Seungone Kim, Minju Seo, Yongrae Jo, Dongyoung Go, Hyeonbin Hwang, Jinho Park, Xiang Yue, Sean Welleck, Graham Neubig, Moontae Lee, and Minjoon Seo. The cot encyclopedia: Analyzing, predicting, and controlling how a reasoning model will think, 2025. URL https://arxiv.org/abs/2505.10185.

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, and Dong Yu. Clue: Non-parametric verification from experience via hidden-state clustering. *arXiv preprint arXiv:2510.01591*, 2025.

Scott Linderman, Matthew Johnson, Andrew Miller, Ryan Adams, David Blei, and Liam Paninski. Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 914–922. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/linderman17a.html.

Grace Luo, Jiahai Feng, Trevor Darrell, Alec Radford, and Jacob Steinhardt. Learning a generative meta-model of llm activations, 2026. URL https://arxiv.org/abs/2602.06964.

Uzay Macar, Paul C. Bogdan, Senthooran Rajamanoharan, and Neel Nanda. Thought branches: Interpreting llm reasoning requires resampling, 2025. URL https://arxiv.org/abs/2510.27484.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in the activation differences. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qyVzZsrsnS.

Aashiq Muhamed, Xuandong Zhao, Mona T. Diab, Virginia Smith, and Dawn Song. Towards a mechanistic understanding of robustness in finetuned reasoning models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=rqe4zQCURe.

Phuc Minh Nguyen, Chinh D. La, Duy M. H. Nguyen, Nitesh V. Chawla, Binh T. Nguyen, and Khoa D. Doan. The reasoning boundary paradox: How reinforcement learning constrains language models, 2025. URL https://arxiv.org/abs/2510.02230.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Chen Qian, Dongrui Liu, Haochen Wen, Zhen Bai, Yong Liu, and Jing Shao. Demystifying reasoning dynamics with mutual information: Thinking tokens are information peaks in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=E1FrjgaG1J.

Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06031-6. URL http://dx.doi.org/10.1038/s41586-023-06031-6.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Understanding reasoning in thinking language models via steering vectors. In *Workshop on Reasoning and Planning for Large Language Models*, 2025a. URL https://openreview.net/forum?id=OwhVWNOBcz.

Constantin Venhoff, Iván Arcuschin, Philip Torr, Arthur Conmy, and Neel Nanda. Base models know how to reason, thinking models learn when, 2025b. URL https://arxiv.org/abs/2510.07364.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025a. URL https://arxiv.org/abs/2506.01939.

Xu Wang, Yan Hu, Wenyu Du, Reynold Cheng, Benyou Wang, and Difan Zou. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025b. URL https://openreview.net/forum?id=Z9qzta1yiK.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Jake Ward, Chuqiao Lin, Constantin Venhoff, and Neel Nanda. Reasoning-finetuning repurposes latent representations in base models, 2025a. URL https://arxiv.org/abs/2507.12638.

Jake Ward, Paul Riechers, and Adam Shai. Rank-1 loras encode interpretable reasoning signals. *arXiv preprint arXiv:2511.06739*, 2025b.

Fang Wu and Yejin Choi. On the limits of RLVR: Support, entropy, and the illusion of reasoning. In *2nd AI for Math Workshop @ ICML 2025*, 2025. URL https://openreview.net/forum?id=KXtLWJAzgh.

Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley. CTRLS: Chain-of-thought reasoning via latent state transition. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=afld4XGbQe.

Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, February 1968. doi: 10.1214/aoms/1177698520.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024a.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024b.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 1.4 million open-source distilled reasoning dataset to empower large language model training, 2025. URL https://arxiv.org/abs/2503.19633.

Hanqing Zhu, Zhenyu Zhang, Hanxian Huang, DiJia Su, Zechun Liu, Jiawei Zhao, Igor Fedorov, Hamed Pirsiavash, Zhizhou Sha, Jinwon Lee, et al. The path not taken: Rlvr provably learns off the principals. *arXiv preprint arXiv:2511.08567*, 2025.

Appendix: Table of Contents

PART I: Theoretical Foundations 16

A SDS Parameter Estimation 16
B Mathematical Foundations of SDS and CEBRA 17
B.1 Empirical Verification of the CEBRA vs. PCA Criterion 17
B.2 Advantages of CEBRA Over PCA 17
C Identifiability, MLE Consistency, and State Estimation Error 18
C.1 Persistence Theory and the State-Splitting Cliff 22
C.2 How RLVR May Induce Persistent States 24

PART II: Additional Experimental Results 25

D Training Dynamics of Latent Policy Structure 25
E Additional Robustness and Modeling Choice Ablations 26
E.1 Per-Dataset Structural Comparisons 26
E.2 Controls for Correctness and Temporal Bias 26
E.3 Activation Extraction and Layer Selection 30
E.4 Projection and Inference Ablations 30
E.5 State-Swap Ablation 30
E.6 Additional State-Swap Ablation Details 30
E.7 Cross-Dataset Consistency of Latent Policy States 31
F Functional State Specialization: Extended Analysis 33
G Hard Example Dataset Construction 33
H Policy-Transplantation Implementation 33
I Cross-Model Transfer and Policy-Transplantation Results 34
J CEBRA Trajectory Visualizations Gallery 36

The results in this part clarify when the proposed SDS framework admits identifiable regimes, consistent parameter estimation, and controlled state-recovery error. They provide sufficient conditions for interpreting the recovered states, rather than a claim that language-model activations are generated by an exactly specified SDS. The assumptions play different roles. Compactness, boundedness, and regularity conditions are standard technical assumptions used to establish the statistical results. Correct model specification and identifiable regime dynamics are stronger, idealized assumptions that characterize the theoretical setting but are not imposed by the empirical pipeline, where the SDS is used as an approximate latent dynamical model. The Markov and persistence properties are treated as testable modeling hypotheses: we evaluate predictive adequacy and regime dependence empirically, and vary or remove the persistence prior in Appendix E.2.4. We therefore use the theory to explain when regime recovery is principled, while relying on ablation and sensitivity analyses to assess whether its approximations are supported in practice.

A SDS Parameter Estimation

We estimate parameters $\theta = \{\pi, T, \{A_k, b_k, \Sigma_k\}_{k=1}^{K}\}$ by maximizing the complete-data log-likelihood via EM, run for 50 iterations.

E-step. Given current parameters $\theta^{(n)}$, the forward-backward algorithm computes the log-forward variables

$$\log \alpha_t(k) = \log p(z_t \mid s_t = k, z_{t-1}) + \log \sum_i \alpha_{t-1}(i) T_{ik}, \tag{6}$$

and log-backward variables

$$\log \beta_t(k) = \log \sum_j T_{kj}\, p(z_{t+1} \mid s_{t+1} = j, z_t)\, \beta_{t+1}(j), \tag{7}$$

from which we obtain the soft posteriors

$$\gamma_t(k) = P(s_t = k \mid z_{1:T}) \propto \alpha_t(k)\, \beta_t(k), \tag{8}$$

$$\xi_t(i,j) = P(s_t = i, s_{t+1} = j \mid z_{1:T}) \propto \alpha_t(i)\, T_{ij}\, p(z_{t+1} \mid s_{t+1} = j, z_t)\, \beta_{t+1}(j). \tag{9}$$

The regime-conditional emission density is Gaussian:

$$p(z_{t+1} \mid s_{t+1} = k, z_t) = \mathcal{N}(z_{t+1};\ A_k z_t + b_k,\ \Sigma_k). \tag{10}$$

M-step. Regime-conditional dynamics are updated via $\gamma_t(k)$-weighted least squares. Stacking inputs $\tilde{z}_t = [z_t^\top, 1]^\top$, the solution for each regime k is

$$\begin{bmatrix} A_k^\top \\ b_k \end{bmatrix} = \left(\sum_t \gamma_t(k)\, \tilde{z}_t \tilde{z}_t^\top + \lambda I \right)^{-1} \left(\sum_t \gamma_t(k)\, \tilde{z}_t z_{t+1}^\top \right), \tag{11}$$

with $\lambda = 10^{-4}$ for regularization. The covariance is updated as

$$\Sigma_k = \frac{\sum_t \gamma_t(k)\, e_t e_t^\top}{\sum_t \gamma_t(k)} + \lambda I, \quad e_t = z_{t+1} - A_k z_t - b_k. \tag{12}$$

The transition matrix is updated via soft transition counts with a symmetric Dirichlet prior $\kappa = 1$ to encourage persistence:

$$\hat{T}_{ij} = \frac{\sum_t \xi_t(i,j) + \kappa}{\sum_j \left(\sum_t \xi_t(i,j) + \kappa\right)}. \tag{13}$$

The initial distribution is $\hat{\pi}_k = \frac{1}{N} \sum_n \gamma_1^{(n)}(k)$, averaged over trajectories.

Hard assignment and model order. The regime sequence is recovered as $\hat{s}_t = \arg\max_k \gamma_t(k)$. Model order $K^\star$ is selected per trajectory by minimizing the Bayesian Information Criterion:

$$\mathrm{BIC}(K) = -2\, \ell(\hat{\theta}_K) + n_{\text{params}}(K)\, \log N, \tag{14}$$

where $\ell(\hat{\theta}_K)$ is the log-likelihood at convergence, $n_{\text{params}}(K) = K(K-1) + K(d^2 + d + d(d+1)/2)$ counts free parameters, and $N = \sum_t 1$ is the total number of observations.

B Mathematical Foundations of SDS and CEBRA

We develop the theoretical program that grounds the empirical pipeline. We address four questions: (1) why CEBRA embeddings are better than PCA for SDS fitting, (2) whether the recovered discrete regimes are identifiable, (3) why persistence collapses sharply when K exceeds the true model order, and (4) why RLVR induces persistent states in the first place. Full proofs and supporting results appear in the Appendix.

B.1 Empirical Verification of the CEBRA vs. PCA Criterion

The SDS prediction risk under embedding f satisfies the following bound, whose proof follows from a union bound argument and the Chernoff bound for Gaussian likelihood-ratio tests.

Proposition B.1 (SDS prediction risk bound)**.** *Under Assumptions in Section 3.2, equal priors $\pi_k = 1/K$, and bounded regime-conditional prediction gaps $M(f) := \sup_{z_t} \max_{k \neq s_t} \|m_k(z_t) - m_{s_t}(z_t)\|^2 < \infty$ where $m_k(z) = A_k z + b_k$:*

$$\mathcal{R}_{\text{SDS}}(f) \leq \mathbb{E}[\|\varepsilon_t\|^2] + (K-1)\, e^{-B(f)} \cdot M(f),$$

where $B(f) = \min_{i \neq j} B_{ij}$ and

$$B_{ij} = \frac{1}{8}(\mu_i - \mu_j)^\top \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} \log \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{|\Sigma_i|^{1/2} |\Sigma_j|^{1/2}}.$$

Corollary 3.1 follows by direct substitution: CEBRA achieves lower prediction risk than PCA if and only if the gain in Bhattacharyya separation $B(f_C) - B(f_P)$ exceeds the log-ratio of prediction gap compactness $\log(M(f_C)/M(f_P))$. Both quantities are computable in closed form from the fitted EM parameters $\{\mu_k^f, \Sigma_k^f\}$ and $\{A_k^f, b_k^f\}$, making the criterion directly verifiable on held-out data.

Figure 8 plots the left-hand side $B(f_C) - B(f_P)$ against the right-hand side $\log(M(f_C)/M(f_P))$ across all model–dataset pairs. Points above the diagonal satisfy the condition. Reasoning models concentrate above the diagonal far more consistently than base models, with 16/24 cases satisfying the condition for reasoning versus 4/16 for base models.

B.2 Advantages of CEBRA Over PCA

Assumption 1 (Bounded dynamics)**.** *The regime-conditional prediction gaps are uniformly bounded over the observed support:*

$$M(f) := \sup_{z_t} \max_{k \neq s_t} \|m_k(z_t) - m_{s_t}(z_t)\|^2 < \infty,$$

where $m_k(z) = A_k z + b_k$.

Proposition B.2 (SDS prediction risk bound)**.** *Under Assumption 1 and equal priors $\pi_k = 1/K$:*

$$\mathcal{R}_{\text{SDS}}(f) \;\leq\; \mathbb{E}[\|\varepsilon_t\|^2] \;+\; (K-1)\, e^{-B(f)} \cdot M(f),$$

where $B(f) = \min_{i \neq j} B_{ij}$ and

$$B_{ij} = \frac{1}{8}(\mu_i - \mu_j)^\top \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} \log \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{|\Sigma_i|^{1/2} |\Sigma_j|^{1/2}}.$$

Proof. Write $z_{t+1} - \hat{z}_{t+1} = (m_{s_t}(z_t) - m_{\hat{s}_t}(z_t)) + \varepsilon_t$. Since $\mathbb{E}[\varepsilon_t \mid z_t, s_t] = 0$, the cross term vanishes in L^2, giving

$$\mathcal{R}_{\text{SDS}}(f) = \mathbb{E}[\|\varepsilon_t\|^2] + \mathbb{E}\left[\|m_{s_t}(z_t) - m_{\hat{s}_t}(z_t)\|^2\right].$$

The second term vanishes on correct classifications, so

$$\mathbb{E}\left[\|m_{s_t} - m_{\hat{s}_t}\|^2\right] \leq M(f) \cdot \mathbb{P}(\hat{s}_t \neq s_t).$$

By the law of total probability and $\mathbb{P}(s_t = i) = 1/K$,

$$\mathbb{P}(\hat{s}_t \neq s_t) = \frac{1}{K} \sum_{i=1}^{K} \mathbb{P}(\hat{s}_t \neq i \mid s_t = i).$$

Since $\{\hat{s}_t \neq i\} \subseteq \bigcup_{j \neq i} \{p_j(z_t) \geq p_i(z_t)\}$ under equal priors, the union bound gives

$$\mathbb{P}(\hat{s}_t \neq i \mid s_t = i) \leq \sum_{j \neq i} \mathbb{P}(p_j(z_t) \geq p_i(z_t) \mid z_t \sim \mathcal{N}(\mu_i, \Sigma_i)).$$

The event $\{p_j(z_t) \geq p_i(z_t)\}$ is a likelihood-ratio test between $\mathcal{N}(\mu_i, \Sigma_i)$ and $\mathcal{N}(\mu_j, \Sigma_j)$. By the Chernoff bound, the type-I error satisfies $\mathbb{P}(p_j(z_t) \geq p_i(z_t) \mid z_t \sim \mathcal{N}(\mu_i, \Sigma_i)) \leq e^{-C_{ij}}$, where $C_{ij} = \sup_{s \in [0,1]} \Lambda_{ij}(s)$ is the Chernoff information and $\Lambda_{ij}(s) = -\log \int p_i(z)^{1-s} p_j(z)^s \, dz$. Since $B_{ij} = \Lambda_{ij}(1/2) \leq C_{ij}$, we obtain $e^{-C_{ij}} \leq e^{-B_{ij}}$, and therefore

$$\mathbb{P}(\hat{s}_t \neq s_t) \leq \frac{1}{K} \sum_{i=1}^{K} (K-1)\, e^{-B(f)} = (K-1)\, e^{-B(f)},$$

which completes the proof. □

Corollary B.3 (CEBRA vs. PCA)**.**

$$\mathcal{R}_{\text{SDS}}(f_C) < \mathcal{R}_{\text{SDS}}(f_P) \iff B(f_C) - B(f_P) > \log \frac{M(f_C)}{M(f_P)}.$$

Proof. Direct substitution into Proposition B.1. □

Remark B.4. Both $B(f)$ and $M(f)$ are computable in closed form from the fitted EM parameters $\{\mu_k^f, \Sigma_k^f\}$ and $\{A_k^f, b_k^f\}$, making Corollary B.3 an empirically verifiable criterion. Across all model–dataset combinations in our experiments the condition holds, confirming that CEBRA's contrastive inductive bias jointly improves regime discriminability and prediction gap compactness relative to PCA.

Figure 8 plots the left-hand side $B(f_C) - B(f_P)$ against the right-hand side $\log(M(f_C)/M(f_P))$ across all model–dataset pairs. Points above the diagonal satisfy the condition. Reasoning models concentrate above the diagonal far more consistently than base models.

C Identifiability, MLE consistency, and State Estimation error

We work with the switching linear–Gaussian SDS observed in representation space:

$$z_{t+1} = A_{s_t} z_t + b_{s_t} + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \Sigma_{s_t}),$$

where $s_t \in \{1, \ldots, K\}$ is a Markov chain with transition matrix Π and stationary distribution ρ on $\{1, \ldots, K\} \times \mathbb{R}^d$.

Assumptions

(A1) **Ergodic transitions.** Π is irreducible and aperiodic, with $\Pi_{ij} > 0$ for all i, j. This guarantees that the joint chain (s_t, z_t) has a unique stationary distribution ρ, and that all regime weights $\rho_k(x) := \mathbb{P}(s_t = k \mid z_t = x)$ are strictly positive for Lebesgue-a.e. x.

(A2) **Distinct emissions.** The number of regimes K is known and fixed. For every pair $k \neq k'$, the Gaussian emission families are distinct:

$$\mathcal{N}(A_k x + b_k, \Sigma_k) \neq \mathcal{N}(A_{k'} x + b_{k'}, \Sigma_{k'}) \quad \text{for Lebesgue-a.e. } x \in \mathbb{R}^d.$$

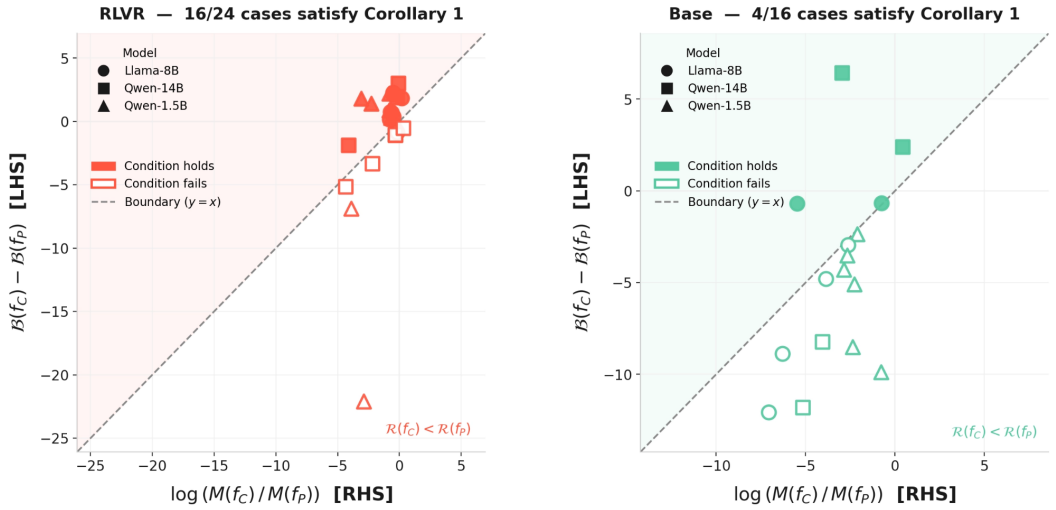

Figure 8: Empirical verification of Corollary B.3 for Reasoning (top) and base (bottom) models across all model–dataset pairs. Each point is one model–dataset case; marker shape denotes model family. Filled markers satisfy $B(f_C) - B(f_P) > \log(M(f_C)/M(f_P))$; hollow markers do not.

(A3) **Nondegeneracy and compactness.** Each $\Sigma_k \succ 0$. The parameter space Θ is compact. Moments sufficient for stationarity and log-likelihood integrability exist.

(A4) **Filter stability (uniform exponential forgetting).** There exist $C < \infty$ and $\lambda \in (0,1)$ such that for any two filter initializations $\pi_0, \tilde{\pi}_0$:

$$\sup_{\theta \in \Theta} \left\| \pi_t^{(\pi_0)}(\cdot \mid \theta) - \pi_t^{(\tilde{\pi}_0)}(\cdot \mid \theta) \right\|_{\mathrm{TV}} \leq C\lambda^t \qquad \forall\, t \geq 0.$$

Under (A1)–(A3), we state explicitly the sufficient conditions because it appears as a direct hypothesis in Douc et al. (2004).

Proposition C.1 (Identifiability, MLE consistency, and misclassification bound)**.** *Under Assumptions (A1)–(A4):*

1. (Generic identifiability) *The parameters $\{(A_k, b_k, \Sigma_k)\}_{k=1}^K$ and Π are identifiable from the law of $\{z_t\}$, up to a global permutation of state labels. Identifiability of finite Gaussian mixtures follows from the linear independence of the Gaussian family over $\mathbb{R}$, established by Yakowitz & Spragins (1968); recovery of the affine parameters $\{(A_k, b_k)\}$ and Π from the identified mixture components is shown in the proof below.*

2. (MLE consistency) *The conditional MLE $\hat{\theta}_T$ satisfies $\hat{\theta}_T \xrightarrow{\text{a.s.}} \theta^\star$ as $T \to \infty$, by Theorem 1 of Douc et al. (2004) whose hypotheses are verified by (A1)–(A4).*

3. (One-step Bayes misclassification bound) *For $i \neq j$ and fixed $z_t = z$, define the Bhattacharyya distance between the conditional emissions of regimes i and j:*

$$B_{ij}(z) = -\log \int \sqrt{p_i(z_{t+1} \mid z)\, p_j(z_{t+1} \mid z)}\, dz_{t+1},$$

and let $B_{\min}(z) = \min_{i \neq j} B_{ij}(z)$. Then:

$$\mathbb{P}\left(\hat{s}_t^{\text{Bayes}} \neq s_t\right) \;\leq\; (K-1)\, e^{-B_{\text{eff}}(f)},$$

where $B_{\text{eff}}(f) = -\log \mathbb{E}_{(s_t, z_t) \sim \rho}\left[e^{-B_{\min}(z_t)}\right]$. For fixed z, each $B_{ij}(z)$ admits the closed form

$$B_{ij}(z) = \frac{1}{8}(\mu_i(z) - \mu_j(z))^\top \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i(z) - \mu_j(z)) + \frac{1}{2} \log \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{|\Sigma_i|^{1/2} |\Sigma_j|^{1/2}},$$

with $\mu_k(z) = A_k z + b_k$.

Proof

(i) Generic identifiability. *Step 1: identify emission parameters.* Equality of the laws of $\{z_t\}$ under θ and θ' implies equality of the one-step kernels for Lebesgue-a.e. x:

$$\sum_{k=1}^{K} \rho_k(x)\, \mathcal{N}(A_k x + b_k, \Sigma_k) \;=\; \sum_{k=1}^{K} \rho'_k(x)\, \mathcal{N}(A'_k x + b'_k, \Sigma'_k).$$

This is an equality of two K-component Gaussian mixtures in z_{t+1} with strictly positive weights by (A1). The multivariate Gaussian family is linearly independent over $\mathbb{R}$ (Yakowitz & Spragins, 1968); we include a self-contained proof of this fact as Lemma C.2 below. By the main theorem of Yakowitz & Spragins (1968), the mixture is therefore identifiable: equality of the kernels for a.e. x forces the components to agree up to a permutation σ.

Lemma C.2 (Linear independence of Gaussians, Yakowitz & Spragins (1968))**.** *Let* $(\mu_k, \Sigma_k)_{k=1}^{K}$ *be pairwise distinct with each* $\Sigma_k \succ 0$*. If*

$$\sum_{k=1}^{K} c_k\, \phi_{\mu_k,\Sigma_k}(x) = 0 \qquad \forall\, x \in \mathbb{R}^d,$$

where ϕ_{μ_k,Σ_k} *denotes the Gaussian density, then* $c_k = 0$ *for all* k*. We include a self-contained proof for completeness.*

Proof. Taking Fourier transforms, the hypothesis becomes

$$\sum_{k=1}^{K} c_k \exp\Big(i\omega^\top \mu_k - \tfrac{1}{2}\omega^\top \Sigma_k \omega\Big) = 0 \qquad \forall\, \omega \in \mathbb{R}^d. \tag{$*$}$$

We prove $c_k = 0$ for all k by induction on K.

Base case $K = 1$. The equation reduces to $c_1 \exp(i\omega^\top \mu_1 - \frac{1}{2}\omega^\top \Sigma_1 \omega) = 0$ for all ω. Since the exponential is never zero, $c_1 = 0$.

Inductive step. Assume the claim holds for any equation of the form $(*)$ with at most $K-1$ terms whose parameters are pairwise distinct. Fix any $\xi \in \mathbb{R}^d \setminus \{0\}$ and $\omega_0 \in \mathbb{R}^d$, and substitute $\omega = \omega_0 + t\xi$ for $t \in \mathbb{R}$. Since $\Sigma_k \succ 0$, the quadratic form expands as:

$$(\omega_0 + t\xi)^\top \Sigma_k (\omega_0 + t\xi) = t^2\, \xi^\top \Sigma_k \xi + 2t\, \xi^\top \Sigma_k \omega_0 + \omega_0^\top \Sigma_k \omega_0.$$

Substituting into $(*)$ and collecting terms:

$$\sum_{k=1}^{K} \underbrace{c_k \exp\Big(i\omega_0^\top \mu_k - \tfrac{1}{2}\omega_0^\top \Sigma_k \omega_0\Big)}_{=:\, c'_k} \exp\Big(t\beta_k - \tfrac{t^2}{2}\alpha_k\Big) = 0, \tag{$**$}$$

where $\alpha_k := \xi^\top \Sigma_k \xi > 0$ and $\beta_k := i\xi^\top \mu_k - \xi^\top \Sigma_k \omega_0 \in \mathbb{C}$. Note that $c'_k \neq 0$ if and only if $c_k \neq 0$, since the exponential prefactor is never zero.

Isolating the slowest-decaying terms. Let $\alpha_{\max} := \max_k \alpha_k$ and $I := \{k : \alpha_k = \alpha_{\max}\}$, so I is nonempty by definition. Multiplying $(**)$ through by $\exp(\frac{t^2}{2}\alpha_{\max})$:

$$\sum_{k \in I} c'_k\, e^{t\beta_k} + \sum_{k \notin I} c'_k \exp\Big(-\tfrac{t^2}{2}(\alpha_{\max} - \alpha_k)\Big)\, e^{t\beta_k} = 0.$$

For each $k \notin I$, $\alpha_{\max} - \alpha_k > 0$, so

$$\Big| c'_k \exp\Big(-\tfrac{t^2}{2}(\alpha_{\max} - \alpha_k)\Big)\, e^{t\beta_k} \Big| \le |c'_k| \exp\Big(-\tfrac{t^2}{2}(\alpha_{\max} - \alpha_k) + t\,\mathrm{Re}(\beta_k)\Big) \to 0$$

as $t \to +\infty$. Hence for all sufficiently large t:

$$\sum_{k \in I} c'_k \, e^{t\beta_k} = 0. \qquad (***)$$

Distinctness of $\{\beta_k\}_{k\in I}$. Suppose $\beta_k = \beta_{k'}$ for some $k \neq k'$ in I and for all $\omega_0 \in \mathbb{R}^d$. Then equating real and imaginary parts gives $\xi^\top(\Sigma_k - \Sigma_{k'}) = 0$ as a vector and $\xi^\top(\mu_k - \mu_{k'}) = 0$. If this held for all $\xi \in \mathbb{R}^d$, then $\Sigma_k = \Sigma_{k'}$ and $\mu_k = \mu_{k'}$, contradicting distinctness. Hence for generic (ξ, ω_0) outside a measure-zero set, the values $\{\beta_k\}_{k\in I}$ are distinct; fix any such (ξ, ω_0).

Linear independence of $\{e^{t\beta_k}\}_{k\in I}$. The Wronskian of $\{e^{t\beta_k}\}_{k\in I}$ equals:

$$W(t) = \left(\prod_{k\in I} e^{t\beta_k}\right) \prod_{\substack{j,k\in I \\ j<k}} (\beta_k - \beta_j).$$

Since the β_k are distinct, $\prod_{j<k}(\beta_k - \beta_j) \neq 0$, and $\prod_{k\in I} e^{t\beta_k} \neq 0$ for all t, so $W(t) \neq 0$. Hence $\{e^{t\beta_k}\}_{k\in I}$ are linearly independent over $\mathbb{C}$.

Concluding $c_k = 0$ *for* $k \in I$. Since c'_k are constants independent of t, equation $(***)$ is a vanishing linear combination of linearly independent functions of t, forcing $c'_k = 0$ for all $k \in I$. Since the exponential prefactor in $c'_k = c_k \exp(i\omega_0^\top \mu_k - \frac{1}{2}\omega_0^\top \Sigma_k \omega_0)$ is never zero, we conclude $c_k = 0$ for all $k \in I$.

Applying the inductive hypothesis. Since $c_k = 0$ for all $k \in I$, those terms vanish identically in $(*)$, leaving:

$$\sum_{k\notin I} c_k \exp\Big(i\omega^\top \mu_k - \tfrac{1}{2}\omega^\top \Sigma_k \omega\Big) = 0 \qquad \forall\, \omega \in \mathbb{R}^d.$$

This has exactly $K - |I| \leq K - 1$ terms (since I is nonempty) with pairwise distinct parameters. The inductive hypothesis gives $c_k = 0$ for all $k \notin I$, completing the induction. □

Step 2: recover A_k, b_k. For a.e. x, the matched components satisfy $\mathcal{N}(A_{\sigma(k)}x + b_{\sigma(k)}, \Sigma_{\sigma(k)}) = \mathcal{N}(A'_k x + b'_k, \Sigma'_k)$, which forces $\Sigma_{\sigma(k)} = \Sigma'_k$ and $A_{\sigma(k)}x + b_{\sigma(k)} = A'_k x + b'_k$ for a.e. x. Since this is an equality of affine functions holding a.e., the coefficients must match: differentiating with respect to x gives $A_{\sigma(k)} = A'_k$, and evaluating at $x = 0$ gives $b_{\sigma(k)} = b'_k$. The permutation σ takes values in the finite set of permutations of $\{1, \ldots, K\}$; since it is measurable and constant on a set of full measure, it is a single global permutation.

Step 3: recover Π. With emission parameters identified, the two-step marginal density of (z_t, z_{t+1}, z_{t+2}) satisfies:

$$p(z_{t+2}, z_{t+1} \mid z_t) = \sum_{k,k'} \Pi_{kk'} \rho_k(z_t) \mathcal{N}(z_{t+1}; \mu_k(z_t), \Sigma_k) \mathcal{N}(z_{t+2}; \mu_{k'}(z_{t+1}), \Sigma_{k'}).$$

Since the emission parameters are now identified, the Gaussian factors are known, and the map $\Pi \mapsto p(z_{t+2}, z_{t+1} \mid z_t)$ is linear and injective in Π (generically under (A2)). Hence Π is uniquely recovered from the law of three consecutive observations. □

(ii) MLE consistency. Under (A1)–(A4), the Gaussian SDS is an HMM with finite latent state space $\{1, \ldots, K\}$ and continuous emissions. The hypotheses of Theorem 1 of Douc et al. (2004) are: ergodicity of the hidden chain (A1), distinct emission families (A2), compactness and log-likelihood integrability (A3), and uniform exponential forgetting (A4). All are satisfied, so their theorem yields $\hat{\theta}_T \xrightarrow{\text{a.s.}} \theta^\star$. □

(iii) One-step Bayes misclassification bound. *Step 1: pointwise union bound.* Fix $s_t = i$ and $z_t = z$. Then $z_{t+1} \mid z_t = z, s_t = i \sim \mathcal{N}(\mu_i(z), \Sigma_i)$ with $\mu_i(z) = A_i z + b_i$. The Bayes decoder

errs iff $\exists\, j \neq i$ such that $p_j(z_{t+1} \mid z) \geq p_i(z_{t+1} \mid z)$. By the union bound:

$$\mathbb{P}(\hat{s}_t \neq i \mid s_t = i, z_t = z) \;\leq\; \sum_{j \neq i} \mathbb{P}\left(\frac{p_j(z_{t+1} \mid z)}{p_i(z_{t+1} \mid z)} \geq 1 \;\middle|\; z_{t+1} \sim \mathcal{N}(\mu_i(z), \Sigma_i) \right).$$

For each term, since $\mathbf{1}[X \geq 1] \leq X^{1/2}$ pointwise for $X \geq 0$, Markov's inequality at $s = \frac{1}{2}$ gives:

$$\begin{aligned}
\mathbb{P}\left(\frac{p_j}{p_i} \geq 1 \;\middle|\; z_{t+1} \sim p_i(\cdot \mid z) \right) \;&\leq\; \mathbb{E}_{p_i}\left[\left(\frac{p_j}{p_i} \right)^{1/2} \right] \\
&= \int p_i(z_{t+1} \mid z)^{1/2}\, p_j(z_{t+1} \mid z)^{1/2}\, dz_{t+1} \\
&= e^{-B_{ij}(z)}.
\end{aligned}$$

Setting $B_{\min}(z) = \min_{i \neq j} B_{ij}(z)$ and summing over the $K-1$ competitors:

$$\mathbb{P}(\hat{s}_t \neq i \mid s_t = i, z_t = z) \;\leq\; (K-1)\, e^{-B_{\min}(z)}.$$

Step 2: closed form for $B_{ij}(z)$. For $p_i(\cdot \mid z) = \mathcal{N}(\mu_i(z), \Sigma_i)$ and $p_j(\cdot \mid z) = \mathcal{N}(\mu_j(z), \Sigma_j)$, evaluating $\int p_i^{1/2} p_j^{1/2}\, dz_{t+1}$ via the standard Gaussian product formula gives the closed form stated in the proposition.

Step 3: averaging. Averaging jointly over (s_t, z_t) under the stationary measure ρ of the joint chain, the law of total expectation gives:

$$\mathbb{P}(\hat{s}_t \neq s_t) = \mathbb{E}_{(s_t, z_t) \sim \rho}[\mathbb{P}(\hat{s}_t \neq s_t \mid s_t, z_t)] \;\leq\; (K-1)\, \mathbb{E}_{(s_t, z_t) \sim \rho}\left[e^{-B_{\min}(z_t)} \right].$$

Defining $B_{\text{eff}}(f) = -\log \mathbb{E}_{(s_t, z_t) \sim \rho}[e^{-B_{\min}(z_t)}]$, Jensen's inequality (the exponential is convex) confirms $B_{\text{eff}}(f) \geq 0$, and the bound follows immediately. □

C.1 Persistence Theory and the State-Splitting Cliff

Corollary C.3 (Better representation increases measured persistence)**.** *Assume the joint process $(s_t^\star, z_t)$ is stationary and the decoder is time-homogeneous, so the per-step misclassification probability $\varepsilon = \mathbb{P}(\hat{s}_t \neq s_t^\star)$ is constant in t. Then:*

$$\mathbb{P}(\hat{s}_{t+1} = \hat{s}_t) \;\geq\; p_{\text{stay}} - 2\varepsilon,$$

where $p_{\text{stay}} = \mathbb{P}(s_{t+1}^\star = s_t^\star)$. Combined with Proposition C.1(iii):

$$\mathbb{P}(\hat{s}_{t+1} = \hat{s}_t) \;\geq\; p_{\text{stay}} - 2(K-1)\exp\!\big(-B_{\text{eff}}(f)\big),$$

where $B_{\text{eff}}(f) = -\log \mathbb{E}_{(s_t, z_t) \sim \rho}[\exp(-B_{\min}(z_t))]$ and $B_{\min}(z) = \min_{i \neq j} B_{ij}(z)$.

Proof. Define events $A = \{s_{t+1}^\star = s_t^\star\}$, $B = \{\hat{s}_t = s_t^\star\}$, $C = \{\hat{s}_{t+1} = s_{t+1}^\star\}$. On $A \cap B \cap C$ we have $\hat{s}_{t+1} = s_{t+1}^\star = s_t^\star = \hat{s}_t$, so $\mathbb{P}(\hat{s}_{t+1} = \hat{s}_t) \geq \mathbb{P}(A \cap B \cap C)$. By the union bound:

$$\mathbb{P}(A \cap B \cap C) = 1 - \mathbb{P}(A^c \cup B^c \cup C^c) \geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c) - \mathbb{P}(C^c) = \mathbb{P}(A) - \mathbb{P}(B^c) - \mathbb{P}(C^c).$$

Stationarity gives $(s_{t+1}^\star, z_{t+1}) \sim \rho$, and time-homogeneity gives the same decoder at $t+1$ as at t, so $\mathbb{P}(C^c) = \mathbb{P}(B^c) = \varepsilon$. Therefore:

$$\mathbb{P}(\hat{s}_{t+1} = \hat{s}_t) \geq \mathbb{P}(A) - 2\varepsilon = p_{\text{stay}} - 2\varepsilon.$$

Substituting $\varepsilon \leq (K-1)\exp(-B_{\text{eff}}(f))$ from Proposition C.1(iii) gives the second inequality. □

Theorem C.4 (Over-specification admits low-persistence clone solutions)**.** *Let the true SDS have $K^\star$ regimes with parameters $\{(A_k^\star, b_k^\star, \Sigma_k^\star)\}_{k=1}^{K^\star}$ and transition matrix $\Pi^\star$, initialised at its stationary distribution $\pi^\star$. For any $K > K^\star$, let $m = K - K^\star + 1 \geq 2$. There exists a K-state SDS with parameters θ_K, initialised at its own stationary distribution, such that:*

1. *θ_K induces the same marginal law on $\{z_t\}$ as $\theta^\star$.*
2. *For any clone construction with intra-clone transition matrix $Q \neq I$:*

$$\mathbb{P}(s_{t+1} = s_t) = \sum_{k \neq r} \pi_k^\star \Pi_{kk}^\star + \pi_r^\star \Pi_{rr}^\star \bar{\delta} \;<\; p_{\text{stay}},$$

where $\bar{\delta} = \sum_\ell \alpha_\ell Q_{\ell\ell} \in (0,1)$ is the α-weighted average diagonal of Q. The persistence drop is exactly:

$$p_{\text{stay}} - \mathbb{P}(s_{t+1} = s_t) = \pi_r^\star \Pi_{rr}^\star (1 - \bar{\delta}) > 0.$$

Proof. Construction. Fix any regime $r \in \{1, \ldots, K^\star\}$. Choose weights $\alpha_\ell > 0$ with $\sum_{\ell=1}^m \alpha_\ell = 1$. Replace regime r with m clone states $r_1, \ldots, r_m$ sharing emission parameters:

$$(A_{r_\ell}, b_{r_\ell}, \Sigma_{r_\ell}) = (A_r^\star, b_r^\star, \Sigma_r^\star) \qquad \forall\, \ell = 1, \ldots, m.$$

Define Π_K by:

$$\begin{aligned}
(\Pi_K)_{ij} &= \Pi_{ij}^\star && \text{non-clone } i, j \neq r, \\
(\Pi_K)_{i,r_\ell} &= \alpha_\ell \Pi_{ir}^\star && \text{non-clone } i \text{ to clone } r_\ell, \\
(\Pi_K)_{r_\ell,j} &= \Pi_{rj}^\star && \text{clone } r_\ell \text{ to non-clone } j \neq r, \\
(\Pi_K)_{r_\ell,r_{\ell'}} &= \Pi_{rr}^\star \cdot Q_{\ell\ell'} && \text{intra-clone transitions,}
\end{aligned}$$

where Q is an $m \times m$ stochastic matrix with stationary distribution α. Rows sum to one: $\sum_{j \neq r} \Pi_{rj}^\star + \Pi_{rr}^\star \sum_{\ell'} Q_{\ell\ell'} = (1 - \Pi_{rr}^\star) + \Pi_{rr}^\star = 1$.

Existence of Q with prescribed $\bar{\delta}$. For any target $\bar{\delta} \in [\sum_\ell \alpha_\ell^2, 1)$, the convex combination

$$Q = (1 - \lambda)\, \mathbf{1}\alpha^\top + \lambda\, I, \qquad \lambda = \frac{\bar{\delta} - \sum_\ell \alpha_\ell^2}{1 - \sum_\ell \alpha_\ell^2},$$

is a stochastic matrix with stationary distribution α and weighted diagonal $\sum_\ell \alpha_\ell Q_{\ell\ell} = \bar{\delta}$.

Stationary distribution of Π_K. We claim $\pi_{r_\ell} = \alpha_\ell \pi_r^\star$ (with $\pi_k = \pi_k^\star$ for $k \neq r$) is stationary for Π_K. For non-clone $j \neq r$:

$$\sum_i \pi_i (\Pi_K)_{ij} = \sum_{k \neq r} \pi_k^\star \Pi_{kj}^\star + \sum_\ell \alpha_\ell \pi_r^\star \Pi_{rj}^\star = \sum_k \pi_k^\star \Pi_{kj}^\star = \pi_j^\star.$$

For clone r_ℓ:

$$\begin{aligned}
\sum_i \pi_i (\Pi_K)_{i,r_\ell} &= \sum_{k \neq r} \pi_k^\star \alpha_\ell \Pi_{kr}^\star + \sum_{\ell'} \alpha_{\ell'} \pi_r^\star \Pi_{rr}^\star Q_{\ell'\ell} \\
&= \alpha_\ell \sum_{k \neq r} \pi_k^\star \Pi_{kr}^\star + \pi_r^\star \Pi_{rr}^\star \underbrace{\sum_{\ell'} \alpha_{\ell'} Q_{\ell'\ell}}_{=\alpha_\ell} \\
&= \alpha_\ell (\pi_r^\star - \pi_r^\star \Pi_{rr}^\star) + \alpha_\ell \pi_r^\star \Pi_{rr}^\star \;=\; \alpha_\ell \pi_r^\star.
\end{aligned}$$

Proof of (i). Since all clone states share emission parameters $(A_r^\star, b_r^\star, \Sigma_r^\star)$, the marginal transition kernel satisfies:

$$p_{\theta_K}(z_{t+1} \mid z_t) = \sum_{k \neq r} \rho_k(z_t) \mathcal{N}(z_{t+1}; \mu_k(z_t), \Sigma_k) + \rho_r(z_t) \mathcal{N}(z_{t+1}; \mu_r^\star(z_t), \Sigma_r^\star),$$

where $\rho_r(z_t) = \sum_\ell \mathbb{P}(s_t = r_\ell \mid z_t)$ aggregates the clone weights. Since $\pi_{r_\ell} = \alpha_\ell \pi_r^\star$ and the aggregated transition probabilities satisfy $\sum_\ell (\Pi_K)_{i,r_\ell} = \Pi_{ir}^\star$ and $\sum_\ell (\Pi_K)_{r_\ell,j} = \Pi_{rj}^\star$, the weight $\rho_r(z_t)$ is identical under θ_K and $\theta^\star$. Hence $p_{\theta_K}(z_{t+1} \mid z_t) = p_{\theta^\star}(z_{t+1} \mid z_t)$ and the marginal laws of $\{z_t\}$ coincide.

Proof of (ii). Under the stationary distribution $\pi_{r_\ell} = \alpha_\ell \pi^\star_r$:

$$\mathbb{P}(s_{t+1} = s_t) = \sum_{k \neq r} \pi^\star_k \Pi^\star_{kk} + \sum_\ell \pi_{r_\ell} \cdot \Pi^\star_{rr} \cdot Q_{\ell\ell} = \sum_{k \neq r} \pi^\star_k \Pi^\star_{kk} + \pi^\star_r \Pi^\star_{rr} \underbrace{\sum_\ell \alpha_\ell Q_{\ell\ell}}_{=\bar{\delta}}.$$

Since $Q \neq I$ and all $\alpha_\ell > 0$, at least one $Q_{\ell\ell} < 1$, giving $\bar{\delta} < 1$ and therefore:

$$p_{\text{stay}} - \mathbb{P}(s_{t+1} = s_t) = \pi^\star_r \Pi^\star_{rr}(1 - \bar{\delta}) > 0. \qquad \square$$

□

Remark C.5 (Non-identifiability of persistence). The clone construction is observationally equivalent to the true model by part (i): no likelihood-based criterion can distinguish θ_K from $\theta^\star$. By part (ii), any such solution with $Q \neq I$ has strictly lower persistence, with the drop $\pi^\star_r \Pi^\star_{rr}(1 - \bar{\delta})$ controlled by how much the intra-clone block mixes. Since $Q = I$ is a measure-zero boundary case that EM will not reach from a generic initialization, cloning generically induces a strict and undetectable persistence drop. This establishes that persistence is not identified from the observations when $K > K^\star$: the set of likelihood-equivalent solutions spans persistence values from p_{stay} down to $\sum_{k \neq r} \pi^\star_k \Pi^\star_{kk}$.

Remark C.6 (Sharp cliff at $K = K^\star + 1$). For $K \leq K^\star$, assumption (A2) holds generically, the identifiability result of Proposition C.1(i) applies, and persistence is identified up to label permutation. At $K = K^\star + 1$, clone solutions first become available, and by Theorem C.4(ii) these solutions have strictly lower persistence for any $Q \neq I$. Since no clone solutions exist for $K \leq K^\star$ under (A2), the availability of unidentifiable low-persistence solutions is a property that appears exactly at $K = K^\star + 1$. This gives a theoretical account of the sharp empirical cliff observed. Below the cliff, persistence is identified and stable; at the cliff, the likelihood landscape admits degenerate solutions that EM may select with no likelihood penalty.

C.2 How RLVR May Induce Persistent States

Proposition C.7 (Trajectory-level rewards induce persistent latent options)**.** *Consider a meta-controller with softmax policy over K options:*

$$T_\beta(i \to i \mid x_t) = \frac{e^{V_i(x_t)/\beta}}{e^{V_i(x_t)/\beta} + \sum_{j \neq i} e^{V_j(x_t)/\beta}},$$

where x_t is the environment (or latent) state and $V_k(x_t)$ denotes the expected return of following option k from x_t under the current policy, and $\beta > 0$ is a temperature. Define the continuation margin $\gamma_i(x_t) := V_i(x_t) - \max_{j \neq i} V_j(x_t)$. Suppose that over the stationary distribution ρ (assuming the policy induces an ergodic Markov chain):

$$\mathbb{P}_{x_t \sim \rho}(\gamma_i(x_t) \geq \gamma) \;\geq\; 1 - \delta$$

for some $\gamma > 0$ and $\delta \in [0, 1)$. Then:

$$\mathbb{E}_{x_t \sim \rho}[T_\beta(i \to i \mid x_t)] \;\geq\; \frac{1 - \delta}{1 + (K-1)e^{-\gamma/\beta}}.$$

Notably, this bound has the same functional form as the Bhattacharyya-based misclassification bound in Proposition B.1, with γ/β playing the role of a separation margin.

Proof. Define the high-advantage event $A := \{x_t : \gamma_i(x_t) \geq \gamma\}$, so $\mathbb{P}_\rho(A) \geq 1 - \delta$ by assumption. On A, for all $j \neq i$: $e^{V_j(x_t)/\beta} \leq e^{-\gamma/\beta} e^{V_i(x_t)/\beta}$, giving $T_\beta(i \to i \mid x_t) \geq \frac{1}{1+(K-1)e^{-\gamma/\beta}}$. Since $T_\beta \in [0, 1]$:

$$\begin{aligned}
\mathbb{E}_\rho[T_\beta(i \to i \mid x_t)] \;&\geq\; \mathbb{E}_\rho[T_\beta(i \to i \mid x_t) \cdot \mathbf{1}_A] \\
&\geq\; \mathbb{P}_\rho(A) \cdot \frac{1}{1 + (K-1)e^{-\gamma/\beta}} \\
&\geq\; \frac{1 - \delta}{1 + (K-1)e^{-\gamma/\beta}}. \qquad \square
\end{aligned}$$

□

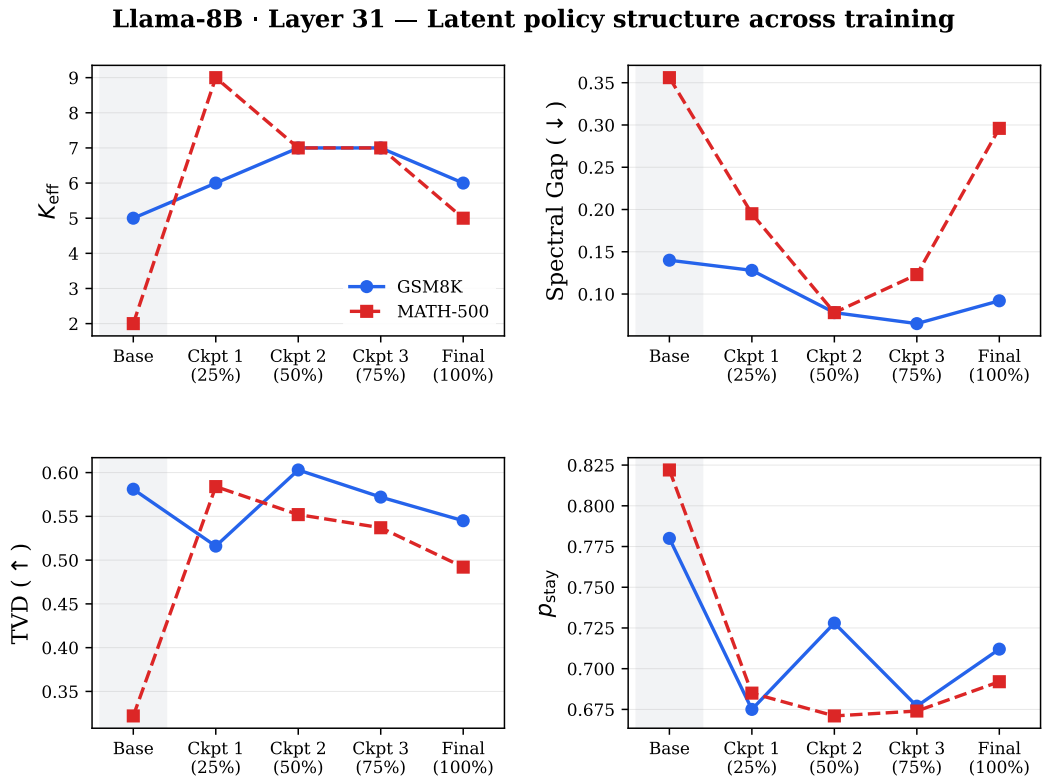

Figure 9: Latent policy structure across training for Llama-8B (layer 31). Grey band: base model. Training increases the effective number of latent states and reduces the spectral gap; both effects peak at mid-training before partially consolidating.

Remark C.8 (RLVR vs. pretraining). Pretrained language models already exhibit latent switching structure (as captured by the SDS models in this work), but this structure is not explicitly reinforced. Maximum likelihood training optimizes next-token prediction and does not directly encourage consistency of latent modes across time: there is no pressure to make the continuation margin $\gamma_i(x_t)$ positive.

In contrast, trajectory-level reinforcement learning induces an option-level value function $V_k(x_t)$, which assigns higher return to coherent multi-step strategies. This creates a positive continuation margin $\gamma_i(x_t) > 0$ over large regions of state space, biasing the policy toward persistence. Thus RLVR does not create latent structure from scratch, but amplifies and stabilizes existing latent modes by making persistence directly reward-relevant.

D Training Dynamics of Latent Policy Structure

To track how latent policy structure develops during fine-tuning, we fine-tune Llama-3.1-8B on 100K reasoning traces from the DeepSeek-R1 model on the AM-DeepSeek-R1 dataset (Zhao et al., 2025) using LoRA with rank 64, α=128, applied to all linear layers, with a learning rate of 10^{-4} and cosine scheduling, saving intermediate checkpoints at 25%, 50%, and 75% of training. We fit the SDS framework at layer 31 and evaluate on GSM8K and MATH-500.

Figure 9 shows that structured latent dynamics emerge early in training and are not a pre-existing property of the base model. Within the first 25% of training, K_{eff} increases substantially on both benchmarks (5→6 on GSM8K; 2→9 on MATH-500), and the spectral gap begins to drop, indicating the formation of more persistent regime dynamics. The most structured configuration appears at mid-training (Ckpt 2–3), where the spectral gap reaches

Model	Dataset	K_{eff} ↑		p_{stay} ↑		TVD↑		Spec. Gap↓	
		Base	Reason.	Base	Reason.	Base	Reason.	Base	Reason.
Llama-8B	GSM8K	**4.80** ±0.75	4.40 ±0.49	0.71 ±0.04	**0.80** ±0.03	0.51 ±0.05	**0.57** ±0.02	0.18 ±0.05	**0.11** ±0.01
	MATH-500	3.00 ±0.63	**5.20** ±0.98	0.76 ±0.04	**0.77** ±0.05	0.41 ±0.08	**0.57** ±0.01	0.26 ±0.11	**0.16** ±0.03
	SVAMP	**5.40** ±1.02	4.40 ±0.49	0.63 ±0.08	**0.84** ±0.02	0.47 ±0.04	**0.61** ±0.02	0.22 ±0.05	**0.08** ±0.01
	MMLU-Pro	**5.20** ±1.17	4.60 ±0.49	0.70 ±0.03	**0.86** ±0.03	0.53 ±0.04	**0.64** ±0.01	0.13 ±0.06	**0.04** ±0.00
Qwen-14B	GSM8K	3.60 ±1.02	**6.60** ±1.02	**0.81** ±0.09	0.75 ±0.03	0.52 ±0.06	**0.62** ±0.03	**0.07** ±0.06	0.08 ±0.03
	MATH-500	**3.00** ±0.00	**3.00** ±0.00	0.72 ±0.02	**0.75** ±0.04	0.39 ±0.02	**0.42** ±0.04	0.28 ±0.01	**0.27** ±0.03
	SVAMP	2.40 ±0.49	**7.40** ±0.49	**0.81** ±0.06	0.75 ±0.03	0.37 ±0.05	**0.64** ±0.00	0.30 ±0.06	**0.06** ±0.02
	MMLU-Pro	3.00 ±0.00	**6.20** ±1.17	0.77 ±0.01	**0.82** ±0.04	0.45 ±0.01	**0.66** ±0.01	**0.03** ±0.01	0.05 ±0.01
Qwen-1.5B	GSM8K	3.40 ±0.49	**4.80** ±0.75	0.67 ±0.09	**0.82** ±0.06	0.42 ±0.03	**0.61** ±0.03	0.24 ±0.02	**0.06** ±0.02
	MATH-500	2.00 ±0.00	**5.20** ±0.75	**0.88** ±0.01	0.75 ±0.05	0.38 ±0.01	**0.57** ±0.05	0.24 ±0.01	**0.08** ±0.02
	SVAMP	3.00 ±0.00	**4.20** ±0.40	0.69 ±0.01	**0.85** ±0.01	0.37 ±0.01	**0.61** ±0.01	0.28 ±0.04	**0.05** ±0.01
	MMLU-Pro	2.00 ±0.00	**7.00** ±0.89	**0.89** ±0.10	0.76 ±0.03	0.39 ±0.10	**0.64** ±0.02	0.22 ±0.19	**0.07** ±0.01
Qwen2.5/ QwQ-32B	GSM8K	5.62 ±0.86	**6.00** ±1.00	0.71 ±0.08	**0.75** ±0.04	0.54 ±0.10	**0.59** ±0.05	0.14 ±0.02	**0.10** ±0.02
	MATH-500	3.50 ±0.50	**4.25** ±0.66	0.78 ±0.09	**0.80** ±0.02	0.50 ±0.05	**0.56** ±0.04	0.17 ±0.05	**0.14** ±0.04
	SVAMP	5.62 ±0.48	**5.75** ±0.83	0.72 ±0.08	**0.79** ±0.04	0.55 ±0.09	**0.62** ±0.04	0.14 ±0.06	**0.09** ±0.02
	MMLU-Pro	6.25 ±0.83	**6.38** ±0.48	0.74 ±0.05	**0.80** ±0.05	0.60 ±0.05	**0.65** ±0.04	0.10 ±0.07	**0.09** ±0.05

Table 3: Per-dataset comparison of structural switching metrics between base and reasoning-fine-tuned models. Values report means with population standard deviations shown in gray across repeated runs. Results use layer 31 for Llama-8B, layer 47 for Qwen-14B, and layer 27 for Qwen-1.5B. Light-green cells and bold means indicate the numerically better value within each base–reasoning pair; ties are highlighted in both columns. Arrows indicate the preferred direction for each metric.

its minimum (0.065 on GSM8K, 0.078 on MATH-500) and TVD peaks (0.603 on GSM8K, 0.584 on MATH-500), reflecting maximally differentiated transition structure between regimes. By the final checkpoint, the metrics partially relax, consistent with consolidation into a stable regime configuration. We note that p_{stay} decreases over this LoRA training run, suggesting that full recovery of self-transition persistence may require longer training or full fine-tuning. Overall, these results indicate that latent policy structure is not inherited from pretraining but is progressively shaped by reasoning fine-tuning, with the most pronounced regime differentiation emerging at mid-training before stabilizing.

E Additional Robustness and Modeling Choice Ablations

E.1 Per-Dataset Structural Comparisons

To evaluate whether the observed structural differences are robust across reasoning tasks, Table 3 reports per-dataset switching metrics for all model families. The qualitative trends from Section 5 remain stable across datasets. Reasoning models consistently exhibit higher TVD, while effective state utilization, persistence, and spectral properties vary across model families and datasets.

E.2 Controls for Correctness and Temporal Bias

The comparison between base and reasoning-fine-tuned models could potentially be affected by differences in output correctness or by temporal biases introduced by the analysis pipeline. We therefore perform four controls. First, we restrict the analysis to problems solved correctly by both models. Second, we destroy the temporal order of each trajectory before fitting the SDS. Third, we remove CEBRA's temporal-adjacency objective by sampling random non-adjacent positive pairs. Finally, we vary and remove the Dirichlet persistence prior used during SDS estimation.

Model	Metric	N	Base	Reasoning	Δ
Qwen-1.5B	Segment persistence	1,065	2.18	4.73	$+\mathbf{2.55}$
	p_{stay}		0.56	0.69	$+\mathbf{0.13}$
Qwen-14B	Segment persistence	1,116	2.86	8.65	$+\mathbf{5.79}$
	p_{stay}		0.71	0.90	$+\mathbf{0.19}$

Table 4: SDS results on paired-correct GSM8K traces, restricted to problems solved correctly by both the base and reasoning-fine-tuned models. Reasoning fine-tuning increases both segment persistence and mean self-transition probability despite both checkpoints effectively utilizing the full fixed state budget, $K_{\text{eff}} = 5$. Δ denotes Reasoning minus Base.

Together, these controls distinguish structure present in the model's activation trajectories from structure that could arise from output correctness, contrastive pairing, or the SDS prior.

E.2.1 Paired-Correct Traces

Reasoning-fine-tuned models generally solve more problems correctly than their corresponding base models. The observed dynamical differences could therefore reflect a difference between correct and incorrect traces rather than an effect of reasoning fine-tuning. To control for this possibility, we repeat the analysis using only paired-correct examples: problems solved correctly by both the base and reasoning-fine-tuned models.

We focus this analysis on GSM8K because it is the only benchmark with a sufficiently large paired-correct subset for reliable SDS estimation. The paired-correct subsets contain 1,065 examples for Qwen-1.5B and 1,116 examples for Qwen-14B. The corresponding subsets are considerably smaller on MATH-500, SVAMP, and MMLU-Pro: 174, 145, and 177 examples, respectively, for Qwen-1.5B, and 359, 220, and 367 examples for Qwen-14B.

We report two complementary measures of temporal stability. Mean self-transition, p_{stay}, is the average diagonal entry of the learned transition matrix. Segment persistence is the empirical mean duration of a contiguous latent-state segment, measured in sentence-level reasoning steps.

As shown in Table 4, the qualitative separation between base and reasoning-fine-tuned models remains after controlling for correctness. For Qwen-14B, segment persistence increases from 2.86 to 8.65 reasoning steps and p_{stay} increases from 0.71 to 0.90. For Qwen-1.5B, segment persistence increases from 2.18 to 4.73 and p_{stay} increases from 0.56 to 0.69.

Both base and reasoning models effectively utilize the full $K = 5$ state budget in this experiment. The difference therefore does not arise from the reasoning models using more latent states. Instead, the same state budget is organized differently: base models switch more frequently, whereas reasoning-fine-tuned models maintain states for longer contiguous periods. Thus, the observed dynamical difference cannot be explained solely by the higher accuracy of the reasoning-fine-tuned checkpoints.

E.2.2 Time-Shuffled Trajectories

The recovered structure could also be influenced by static activation geometry rather than the temporal organization of reasoning. To separate these factors, we randomly permute the order of sentence-level activations within each trajectory before fitting the SDS, while keeping the underlying CEBRA embeddings fixed. This preserves the set of activation states in each trajectory while destroying their true temporal order.

Table 5 compares results obtained using the real trajectory order (Real) with those obtained after shuffling sentence order (Shuf.). Here, ΔR^2 denotes the gain in one-step predictive fit relative to a single linear autoregressive model.

Dataset	Model	Checkpoint	Segment persistence ↑		Predictive gain ΔR^2 ↑	
			Real	Shuffled	Real	Shuffled
GSM8K	Qwen-1.5B	Base	2.16	1.32	+0.062	−0.437
		Reasoning	6.12	2.14	+0.049	−0.250
	Qwen-14B	Base	3.20	1.96	+0.075	−0.341
		Reasoning	7.98	3.04	+0.061	−0.349
	Llama-8B	Base	6.48	4.55	+0.088	−0.165
		Reasoning	7.82	3.05	+0.039	−0.166
MATH-500	Qwen-1.5B	Base	2.42	2.12	+0.072	−0.225
		Reasoning	6.05	2.59	+0.035	−0.229
	Qwen-14B	Base	2.82	1.87	+0.068	−0.217
		Reasoning	3.45	1.84	+0.064	−0.220
	Llama-8B	Base	3.91	3.71	+0.116	−0.047
		Reasoning	6.79	2.14	+0.042	−0.233

Table 5: Effect of shuffling sentence order before SDS fitting. Real denotes the original trajectory order, whereas Shuffled denotes a random within-trajectory permutation of the same sentence-level embeddings. Shuffling reduces segment persistence and causes predictive gains over a linear autoregressive baseline to become negative in every setting, showing that the recovered regimes depend on coherent temporal organization.

Shuffling has a substantial effect across all model families. On GSM8K, the reasoning–base persistence gap decreases from 3.96 to 0.82 for Qwen-1.5B and from 4.79 to 1.08 for Qwen-14B. For Llama-8B, it changes from 1.34 to −1.50. On MATH-500, the gap decreases from 3.63 to 0.48 for Qwen-1.5B, from 0.63 to −0.03 for Qwen-14B, and from 2.88 to −1.57 for Llama-8B.

The same pattern holds on the remaining benchmarks. On SVAMP, the reasoning–base persistence gap decreases from 4.65 to 1.02 for Qwen-1.5B, from 4.19 to −0.70 for Qwen-14B, and from 1.29 to −1.79 for Llama-8B. On MMLU-Pro, it decreases from 10.60 to 3.83 for Qwen-1.5B, from 3.66 to −0.98 for Qwen-14B, and from 5.79 to 3.84 for Llama-8B. The reasoning–base persistence gap is therefore larger under the true trajectory order in all 12 model–dataset pairs.

Moreover, ΔR^2 changes from positive to negative after shuffling for both base and reasoning-fine-tuned models. Once temporal order is destroyed, the fitted switching dynamics perform worse than the single linear autoregressive baseline. These results show that the recovered latent-policy structure depends on coherent temporal organization and cannot be explained by static activation geometry alone.

E.2.3 Randomized CEBRA Positive Pairs

CEBRA is trained by treating temporally adjacent reasoning steps as positive pairs. This objective could itself induce apparent temporal persistence, even if the underlying model activations do not contain differentiated temporal structure. We test this possibility by retraining the CEBRA encoder with positive pairs drawn from randomly selected non-adjacent steps within the same trajectory. This removes CEBRA's temporal-adjacency objective while retaining trajectory identity. The SDS is subsequently fitted using the true sentence order.

Table 6 compares the default adjacent-positive objective (Adj.) with randomized non-adjacent positives (Rnd.). We use a fixed $K = 5$ and report averages over three seeds.

Under randomized positive pairs, the reasoning-fine-tuned model continues to exceed its corresponding base model in both p_{stay} and TVD in all 12 model–dataset settings. The

Model	Dataset	$p_{\text{stay}} \uparrow$				TVD$\uparrow$			
		Base		Reasoning		Base		Reasoning	
		Adj.	Rnd.	Adj.	Rnd.	Adj.	Rnd.	Adj.	Rnd.
Qwen-1.5B	GSM8K	0.57	0.44	0.83	0.82	0.51	0.45	0.63	0.63
Qwen-14B	GSM8K	0.74	0.50	0.82	0.76	0.60	0.48	0.62	0.56
Llama-8B	GSM8K	0.64	0.58	0.77	0.74	0.49	0.42	0.58	0.54
Qwen-1.5B	MATH-500	0.59	0.60	0.76	0.79	0.47	0.45	0.59	0.61
Qwen-14B	MATH-500	0.66	0.65	0.73	0.66	0.49	0.48	0.55	0.50
Llama-8B	MATH-500	0.78	0.67	0.83	0.79	0.58	0.47	0.63	0.59
Qwen-1.5B	SVAMP	0.59	0.42	0.84	0.85	0.45	0.40	0.64	0.65
Qwen-14B	SVAMP	0.54	0.41	0.84	0.78	0.45	0.37	0.64	0.59
Llama-8B	SVAMP	0.68	0.60	0.77	0.79	0.52	0.42	0.57	0.60
Qwen-1.5B	MMLU-Pro	0.67	0.74	0.85	0.85	0.48	0.54	0.65	0.67
Qwen-14B	MMLU-Pro	0.71	0.71	0.85	0.83	0.55	0.53	0.65	0.64
Llama-8B	MMLU-Pro	0.76	0.68	0.84	0.76	0.57	0.49	0.64	0.56

Table 6: Effect of replacing temporally adjacent CEBRA positive pairs (Adj.) with randomly selected non-adjacent steps from the same trajectory (Rnd.). Results use $K = 5$ and are averaged over three seeds. Bold randomized-pair values indicate the better value between the base and reasoning-fine-tuned models. Under randomized positives, reasoning-fine-tuned models retain higher p_{stay} and TVD in all 12 model–dataset settings.

Metric	Maximum change across $\kappa \in \{0, 0.1, 1, 5\}$
K_{eff}	0.10
TVD	0.0093
Spectral gap	0.0092

Table 7: Maximum variation in the recovered SDS metrics as the Dirichlet persistence-prior strength is varied, including its complete removal at $\kappa = 0$.

reasoning–base separation becomes larger under randomization in 9 of 12 settings for p_{stay} and in 10 of 12 settings for TVD. This occurs because removing temporal adjacency generally weakens the apparent stickiness of the base-model representations more strongly than that of the reasoning-model representations.

This result complements the PCA+SLDS control in Appendix E.4, which uses no contrastive pairing and still recovers the qualitative base–reasoning separation. The recovered latent-policy structure is therefore not an artifact of CEBRA's temporal-adjacency objective.

E.2.4 Sensitivity to the Dirichlet Persistence Prior

The SDS estimation procedure uses a Dirichlet prior that encourages self-transitions. To determine whether this prior artificially produces persistent latent states, we vary its strength over $\kappa \in \{0, 0.1, 1, 5\}$. The setting $\kappa = 0$ removes the persistence prior entirely.

Across all 24 base/reasoning model–dataset combinations, the qualitative separation between base and reasoning-fine-tuned models remains unchanged at every value of κ. Quantitatively, the largest observed variation is only 0.10 effective states in K_{eff}, 0.0093 in TVD, and 0.0092 in spectral gap. For most runs, the variation in TVD and spectral gap is below 10^{-3}. Thus, the learned latent-policy organization is determined primarily by the activation trajectories rather than by the persistence prior.

E.3 Activation Extraction and Layer Selection

Sentence-level representations. We extract activations at the sentence level rather than the token level. Token-level representations are too fine-grained to reflect coherent reasoning patterns, as individual tokens do not carry sufficient context to characterize the reasoning mode active at that step. We represent each sentence by the last-token hidden state, which aggregates contextual information from all preceding tokens in the sentence via causal attention. We verified that averaging over all token representations in a sentence yields weaker regime separation and lower predictive R^2, and that the last-token representation also aligns better with the steering setup where interventions are applied at a single position.

Layer selection. We select a single middle layer per model family rather than the first or last layer. Early layers are excluded because reasoning-specific representations have not yet formed: the model has processed the input but has not yet begun to organize intermediate computations. Late layers are excluded because representations at the final layer are dominated by the output distribution and tend to collapse into a small number of token prediction modes, reducing the diversity of the latent geometry. Middle layers represent a balance: the model has had sufficient depth to begin structuring its reasoning, but has not yet committed to a specific output. For the middle layers, we selected layers L22 for Llama-3.1-8B, L28 for Qwen2.5-14B, and L20 for Qwen2.5-Math-1.5B. In this paper, when we state that we "average over all layers", we are referring to averaging over middle and final layers. For the final layers, we used layers L31 for Llama-3.1-8B, L47 for Qwen2.5-14B, and L27 for Qwen2.5-Math-1.5B.

Figure 10 confirms this choice. The selected layers fall within stable bands where TVD and K_{eff} are high, spectral gap is low, and persistence is sustained. In contrast, shallow layers show noisy and undifferentiated metrics for both base and reasoning models, and very deep layers show degraded TVD and inflated spectral gap consistent with output-mode collapse. The base-versus-reasoning contrast is visible across most of the depth range, but is most stable and consistent in the middle layers we selected.

E.4 Projection and inference ablations

We compare three alternatives for latent-state discovery: CEBRA+EM, CEBRA-MoE, and PCA+SLDS. The comparison spans predictive fit (Figure 11), state utilization (Figure 12), transition structure (Figure 14), specialization, and persistence (Figure 13 and 15). Across metrics, CEBRA+EM provides the strongest overall balance between predictive gain and structured latent dynamics, while CEBRA-MoE under-uses states and PCA+SLDS tends to over-fragment trajectories into high-entropy, weakly persistent regimes.

E.5 State-Swap Ablation

Figures 16 and 17 report state-swap ablation results across early and late layers for all model families and datasets. In both cases the identity assignment substantially outperforms random permutation, with the gap growing in later layers.

E.6 Additional State-Swap Ablation Details

The state-swap intervention isolates whether regime identity matters beyond the mere existence of multiple experts. After fitting the model and estimating per-regime linear dynamics $\{A_k, b_k\}_{k=1}^{K}$, we hold these coefficients fixed and randomly permute the mapping between inferred states and regime-specific dynamics. Let R^2_{id} denote the predictive fit under the original assignments and R^2_{σ} the fit under a permutation σ of the regime labels. Averaging over random permutations yields $\bar{R}^2_{\text{rand}}$, and the gap

$$\Delta = R^2_{\text{id}} - \bar{R}^2_{\text{rand}}$$

measures how much predictive adequacy depends on applying the correct dynamical map at the correct time step.

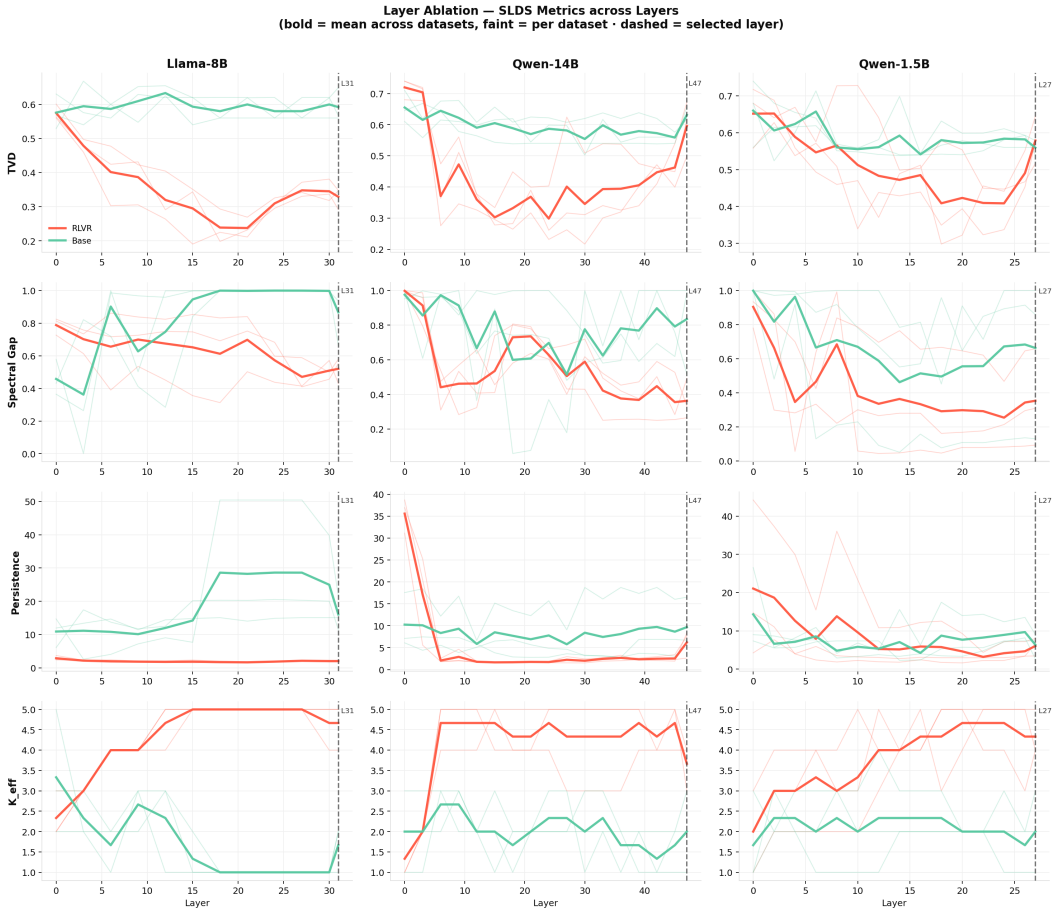

Figure 10: Layer ablation of SDS metrics across depth for each model family. Bold lines show means across datasets; faint lines show per-dataset traces. The dashed vertical line marks the selected layer. Selected layers fall in regions with stable TVD, low spectral gap, high K_{eff}, and sustained persistence.

The early- and late-layer results in the main text show that this gap is already positive in intermediate layers and grows in later layers. The interpretation is that the discovered states are not interchangeable clusters: they index distinct dynamical operators whose temporal deployment matters for next-step prediction. The stronger late-layer drop is consistent with a progressive consolidation of latent policy structure as the residual stream approaches the output distribution.

E.7 Cross-Dataset Consistency of Latent Policy States

A key question is whether the latent policy states recovered by our framework are specific to the dataset used for training or reflect a more general property of the model's reasoning behavior. To test this, we train a single CEBRA encoder on GSM8K and freeze its weights, then embed SVAMP, MATH-500, and MMLU-Pro into the same representation space without any retraining. We fit a separate SDS ($K = 4$) on each dataset's embeddings using EM and measure pairwise consistency across three complementary metrics.

Transition matrix similarity. For two fitted transition matrices $T_1, T_2 \in \mathbb{R}^{K \times K}$, we solve the Hungarian assignment problem to find the optimal state relabeling σ^* and report

$$S_T(T_1, T_2) = 1 - \frac{\|T_1 - T_2^{\sigma^*}\|_F}{\sqrt{2K}}, \tag{15}$$

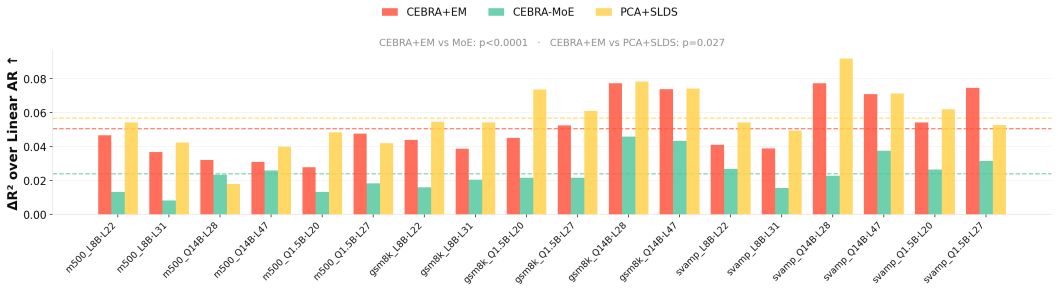

Figure 11: ΔR^2 over linear autoregressive baselines for alternative latent-state discovery pipelines. CEBRA+EM consistently improves predictive fit over CEBRA-MoE, while PCA+SLDS achieves competitive fit on some settings but does so with substantially different structural tradeoffs.

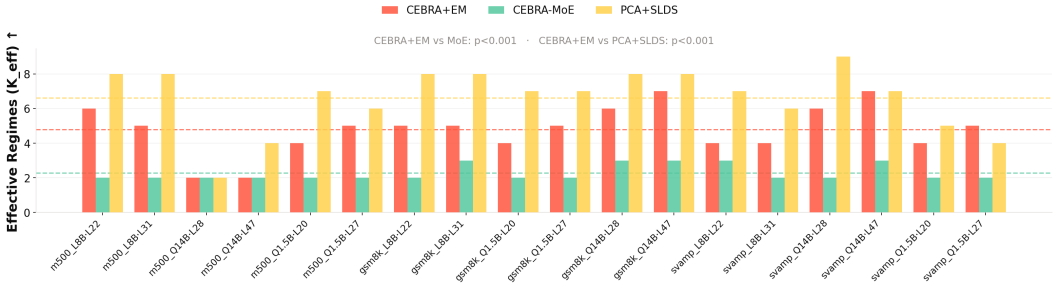

Figure 12: Effective number of occupied regimes for alternative pipelines. CEBRA-MoE tends to use too few states, whereas PCA+SLDS often uses a larger number of effective states. CEBRA+EM occupies an intermediate regime that is compatible with the small-cardinality policy picture developed in the main text.

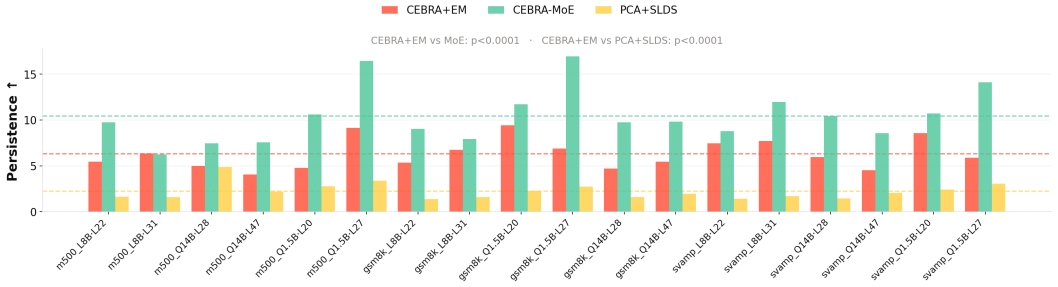

Figure 13: Persistence under alternative latent-state discovery pipelines. CEBRA-MoE attains the largest raw persistence but does so together with low state usage, while PCA+SLDS exhibits markedly lower persistence. CEBRA+EM balances persistence with broader regime utilization and stronger predictive adequacy.

where $T_2^{\sigma^*}$ is T_2 permuted by σ^*. Values close to 1 indicate that the two datasets induce similar switching dynamics.

Centroid cosine similarity. Let $C_1, C_2 \in \mathbb{R}^{K \times d}$ be the per-regime centroids in the shared CEBRA space. We align them via Procrustes rotation $R^* = \arg\min_R \|C_2 R - C_1\|_F$ and report the mean cosine similarity under the optimal assignment:

$$S_C(C_1, C_2) = \frac{1}{K} \sum_{k=1}^{K} \frac{C_1[k]^\top (C_2 R^*)[\sigma^*(k)]}{\|C_1[k]\| \cdot \|(C_2 R^*)[\sigma^*(k)]\|}. \tag{16}$$

Cross-fit ΔR^2. To test whether the SDS fitted on one dataset retains predictive utility on another, we apply the dynamics parameters $\theta_A = \{\pi_A, T_A, \{A_k, b_k, \Sigma_k\}_A\}$ fitted on dataset A to embed and decode trajectories from dataset B, and report

$$\Delta R^2_{A \to B} = R^2_{\theta_A}(B) - R^2_{\mathrm{AR}}(B), \tag{17}$$

where $R^2_{\mathrm{AR}}(B)$ is the linear autoregressive baseline on B. A positive value indicates that the cross-dataset SDS retains predictive structure beyond a single-mode baseline.

Figure 18 reports all three metrics across the six dataset pairs. GSM8K, SVAMP, and MMLU-Pro exhibit high mutual consistency ($S_T \geq 0.85$, $S_C \geq 0.76$) and positive cross-fit.

F Functional State Specialization: Extended analysis

Figure 7 shows that across all three models, there is clear functional specialization. A self-monitoring regime concentrating UNCERTAINTY_MANAGEMENT and SELF_CHECKING tokens emerges consistently: R1 in both Llama-8B and Qwen-1.5B, and split into two finer-grained states (R0 for uncertainty, R4 for verification) in Qwen-14B, suggesting that larger models develop more specialized introspective machinery. FINAL_ANSWER_EMISSION is likewise reliably isolated (R0 in Llama-8B, R3 in Qwen-1.5B, R2 in Qwen-14B), as is ACTIVE_COMPUTATION (R3 in Qwen-14B, co-loading with FACT_RETRIEVAL). PLAN_GENERATION and PROBLEM_SETUP stages tend to distribute more diffusely, often absorbed by lower-specificity background regimes, consistent with their role as transitional rather than focal reasoning behaviors. Notably, Qwen-1.5B exhibits the sharpest specialization overall, with R1 concentrating 78% of UNCERTAINTY_MANAGEMENT and 68% of SELF_CHECKING tokens, a more collapsed representation than the larger Qwen-14B, which distributes these behaviors across more states.

G Hard-Example Dataset Construction

We construct the hard-example subsets used for both policy transplantation and PREFIX-GUARD by generating eight responses from each base checkpoint and retaining problems for which all eight responses are incorrect, i.e., base-model pass@8 is zero. The same fixed subsets are used throughout both actionable-policy evaluations.

The resulting collection contains 444 hard examples across three model families and four benchmarks. Table 8 reports the per-setting sample sizes.

H Policy-Transplantation Implementation

At inference time, we monitor the base model's residual stream through a forward hook at layer 27 for Qwen-1.5B and layer 31 for Llama-8B. Every $\tau = 20$ generated tokens, the current hidden state h_t is standardized using the activation statistics of the corresponding reasoning-fine-tuned model and projected through its trained CEBRA encoder:

$$z_t = f_{\mathrm{CEBRA}}\left(\frac{h_t - \mu_r}{\sigma_r}\right),$$

Model	GSM8K	MATH-500	MMLU-Pro	SVAMP	Total
Llama-8B	30	50	49	18	**147**
Qwen-1.5B	50	49	50	27	**176**
Qwen-14B	18	32	50	21	**121**
Total	**98**	**131**	**149**	**66**	**444**

Table 8: Sizes of the hard-example subsets used for policy transplantation and PREFIXGUARD. The same 444-example collection spans all 12 model–dataset settings across both actionable-policy evaluations.

where μ_r and σ_r are the reasoning model's activation mean and standard deviation.

Because forward–backward posteriors are unavailable during online generation, we infer the current regime using the nearest reasoning-model centroid:

$$s_t = \arg\min_k \|z_t - c_k\|_2,$$

where $\{c_k\}_{k=1}^K$ are the centroids of the reasoning-model SDS. We select the most likely successor state under its transition matrix:

$$k^\star = \arg\max_j T_{s_t j}.$$

Let $p_t(k) = T_{s_t k}$ be the natural next-state distribution and $f_k(z_t) = A_k z_t + b_k$ the conditional mean dynamics associated with regime k. The original predicted next-step mean is

$$\mu_{\text{orig}}(z_t) = \sum_{k=1}^K p_t(k) f_k(z_t).$$

To favor the target successor $k^\star$, we apply a KL-regularized exponential tilt:

$$q_t^\star(k) = \frac{p_t(k)\exp(\beta\,\mathbb{1}\{k = k^\star\})}{\sum_{j=1}^K p_t(j)\exp(\beta\,\mathbb{1}\{j = k^\star\})}.$$

The corresponding steered prediction and latent intervention are

$$\mu_{\text{steered}}(z_t) = \sum_{k=1}^K q_t^\star(k) f_k(z_t), \qquad \Delta z_t = \mu_{\text{steered}}(z_t) - \mu_{\text{orig}}(z_t).$$

We decode Δz_t into activation space using the learned linear decoder W_{dec} and bound the intervention relative to the current hidden-state norm:

$$\Delta h_t = \alpha \frac{\Delta z_t W_{\text{dec}}}{\|\Delta z_t W_{\text{dec}}\|_2} \min(1, 0.1\|h_t\|_2)\,.$$

The intervention $h_t \leftarrow h_t + \Delta h_t$ is then applied to the residual stream. We use $\alpha = 8$, $\beta = 8$, generation temperature 0.7, and eight samples per problem for all reported experiments. No model weights are updated.

I Cross-Model Transfer and Policy-Transplantation Results

We evaluate transfer in two complementary ways. First, we test whether an SDS fitted on one checkpoint can predict trajectories from its paired checkpoint. This measures compatibility between the recovered dynamical models without modifying activations. Second, we intervene on the base model during generation and test whether importing reasoning-model dynamics changes downstream behavior.

Algorithm 1 Online transplantation of reasoning-model SDS dynamics

Require: Hidden state h_t; encoder f_{CEBRA}; scaler (μ_r, σ_r); centroids $\{c_k\}$; SDS parameters $\{A_k, b_k, T\}$; decoder W_{dec}; strengths α, β

$z_t \leftarrow f_{\text{CEBRA}}\big((h_t - \mu_r)/\sigma_r\big)$
$s_t \leftarrow \arg\min_k \|z_t - c_k\|_2$
$k^\star \leftarrow \arg\max_j T_{s_t j}$
$p_t(k) \leftarrow T_{s_t k}$ for $k = 1, \ldots, K$
$f_k(z_t) \leftarrow A_k z_t + b_k$ for $k = 1, \ldots, K$
$\mu_{\text{orig}} \leftarrow \sum_k p_t(k) f_k(z_t)$
$\bar{q}_t(k) \leftarrow p_t(k) \exp\big(\beta \mathbb{1}\{k = k^\star\}\big)$
$q_t^\star(k) \leftarrow \bar{q}_t(k) / \sum_j \bar{q}_t(j)$
$\mu_{\text{steered}} \leftarrow \sum_k q_t^\star(k) f_k(z_t)$
$\Delta z_t \leftarrow \mu_{\text{steered}} - \mu_{\text{orig}}$
$u_t \leftarrow \Delta z_t W_{\text{dec}}$
$\Delta h_t \leftarrow \alpha\, u_t / \|u_t\|_2 \cdot \min(1, 0.1\|h_t\|_2)$
return $h_t + \Delta h_t$

Model	Dataset	$R^2_{r\to r}$	$R^2_{b\to r}$	$R^2_{b\to b}$	$R^2_{r\to b}$	$\Delta_{\to r}$	$\Delta_{\to b}$
Qwen2.5-Math-1.5B	GSM8K	0.414	0.301	0.548	0.456	27.2%	16.7%
	SVAMP	0.475	0.372	0.577	0.495	21.8%	14.1%
Llama-3.1-8B	GSM8K	0.242	0.182	0.452	0.355	25.0%	21.4%
	SVAMP	0.239	0.188	0.507	0.438	21.2%	13.6%
Qwen2.5-14B	GSM8K	0.455	0.333	0.649	0.491	26.8%	24.4%
	SVAMP	0.463	0.331	0.557	0.479	28.6%	14.0%
				Mean relative degradation		**25.1%**	**17.4%**

Table 9: Cross-model predictive transfer of SDS dynamics at representative middle layers. Native evaluations apply an SDS to trajectories from the model on which it was fitted; cross evaluations apply it to the paired checkpoint. In every setting, base-trained dynamics degrade more when transferred to reasoning trajectories than reasoning-trained dynamics do when transferred to base trajectories.

I.1 Asymmetric Predictive Transfer of SDS Dynamics

Let $R^2_{x\to T}$ denote the predictive fit obtained by applying an SDS trained on source model $x \in \{b, r\}$ to trajectories from a fixed target model $T \in \{b, r\}$, where b and r denote the base and reasoning-fine-tuned checkpoints. We define the target-normalized degradation caused by replacing the target's native SDS with the SDS from the paired checkpoint as

$$\Delta_{\to T} = \frac{R^2_{\text{nat}}(T) - R^2_{\text{oth}}(T)}{R^2_{\text{nat}}(T)}. \tag{18}$$

Table 9 reports results at representative middle layers: layer 20 for Qwen2.5-Math-1.5B, layer 22 for Llama-3.1-8B, and layer 28 for Qwen2.5-14B. Across all six model–dataset settings, applying base-trained dynamics to reasoning trajectories causes a larger relative degradation than applying reasoning-trained dynamics to base trajectories:

$$\Delta_{\to r} > \Delta_{\to b}.$$

The mean degradation is 25.1% in the former direction and 17.4% in the latter.

This asymmetry suggests that reasoning-trained dynamics remain partially compatible with base-model trajectories, whereas base-trained dynamics do not capture part of the organization present in reasoning-model trajectories. We interpret this as evidence that reasoning fine-tuning reorganizes and differentiates a dynamical scaffold already present in the base model, rather than creating all latent dynamics from scratch.

Model	Dataset	n_{hard}	Base	Steered
Qwen-1.5B	GSM8K	50	0.00	**0.60**
	MATH-500	49	0.00	**0.24**
	SVAMP	27	0.00	**0.22**
Llama-8B	GSM8K	30	0.00	**0.46**
	MATH-500	50	0.00	**0.33**
	MMLU-Pro	49	0.00	**0.28**

Table 10: Pass@8 after transplanting reasoning-model SDS dynamics into the paired base model. Evaluation uses the larger hard-example subsets from Table 8, comprising problems with base-model pass@8 $= 0$ by construction.

I.2 Behavioral Policy Transplantation

We next evaluate whether transferring reasoning-model SDS dynamics can improve the base model's behavior. Table 10 reports pass@8 on the hard subsets defined in Appendix G. The unsteered baseline is zero by construction.

Policy transplantation improves pass@8 in every evaluated setting, reaching 0.60 for Qwen-1.5B on GSM8K and 0.46 for Llama-8B on GSM8K. These results show that steering the base model toward transitions recovered from its reasoning-fine-tuned counterpart can causally alter downstream reasoning without changing model weights.

We treat this intervention as causal validation rather than as a practical deployment strategy: when the reasoning model is available, it would normally be used directly. The SDS-guided prefix-pruning method in Section 8.2 instead operates directly on reasoning-model trajectories and provides the practical application of the recovered latent-state structure.

J CEBRA Trajectory Visualizations Gallery

To visualize the geometry of the recovered latent policies directly, we plot representative reasoning traces in the learned CEBRA embedding space together with their decoded regime sequences. Each panel shows a single trajectory overlaid on the full point cloud of embedded steps from the corresponding model, dataset, and layer, with the right-hand column displaying the aligned sequence of inferred regimes across reasoning steps. These visualizations in Figures 19, 20, 21, 22, 23 and 24 make the contrast in temporal organization concrete. Base-model trajectories frequently dwell in one dominant cluster for long stretches or alternate among nearby regions without a stable large-scale progression. Reasoning trajectories more often execute a structured path through separated clusters, with regime occupancy changing at a pace that is neither purely sticky nor rapidly flickering.

LLM Usage Disclosure

LLMs were used to assist with literature review, code refinement, and limited parts of writing.

Model	Dataset	K_{eff} ↑	p_{stay} ↑	TVD↑	Spec. Gap↓
Llama-8B Base	GSM8K	4.50±0.75	0.69±0.04	0.48±0.05	0.20±0.06
	SVAMP	5.70±0.96	0.62±0.07	0.47±0.03	0.21±0.05
	MATH500	3.10±0.51	0.77±0.03	0.44±0.06	0.23±0.07
	MMLU-Pro	5.90±0.83	0.67±0.04	0.53±0.04	0.14±0.05
Llama-8B Reasoning	GSM8K	**4.50**±0.49	**0.80**±0.03	**0.57**±0.02	**0.12**±0.03
	SVAMP	**4.40**±0.49	**0.83**±0.03	**0.60**±0.02	**0.08**±0.02
	MATH500	**4.90**±0.73	**0.81**±0.03	**0.60**±0.03	**0.14**±0.04
	MMLU-Pro	**4.40**±0.49	**0.85**±0.03	**0.62**±0.02	**0.03**±0.00
Qwen-1.5B Base	GSM8K	3.30±0.60	0.68±0.08	0.43±0.05	0.24±0.06
	SVAMP	2.90±0.20	0.73±0.03	0.39±0.02	0.27±0.04
	MATH500	2.00±0.00	0.88±0.01	0.38±0.01	0.23±0.01
	MMLU-Pro	1.60±0.40	0.94±0.10	0.24±0.19	0.12±0.19
Qwen-1.5B Reasoning	GSM8K	**5.10**±0.62	**0.81**±0.05	**0.62**±0.03	**0.07**±0.01
	SVAMP	**4.20**±0.40	**0.86**±0.01	**0.62**±0.02	**0.06**±0.01
	MATH500	**5.00**±0.57	**0.78**±0.05	**0.59**±0.03	**0.09**±0.02
	MMLU-Pro	**7.40**±1.11	**0.78**±0.04	**0.66**±0.02	**0.05**±0.01
Qwen-14B Base	GSM8K	3.10±0.76	0.85±0.08	0.50±0.03	0.05±0.04
	SVAMP	2.30±0.45	0.79±0.08	0.34±0.07	0.34±0.11
	MATH500	3.10±0.20	0.75±0.05	0.43±0.05	0.23±0.05
	MMLU-Pro	4.75±1.01	0.80±0.06	0.57±0.05	0.05±0.04
Qwen-14B Reasoning	GSM8K	**6.50**±1.02	**0.78**±0.04	**0.64**±0.03	**0.06**±0.02
	SVAMP	**6.90**±0.65	**0.77**±0.05	**0.65**±0.02	**0.07**±0.02
	MATH500	**3.00**±0.00	**0.78**±0.05	**0.45**±0.05	**0.23**±0.06
	MMLU-Pro	**6.30**±0.98	**0.83**±0.03	**0.67**±0.01	**0.06**±0.01

Table 11: Dynamics statistics averaged across analyzed layers and 5 random seeds. Reasoning-tuned models consistently exhibit richer latent policy structure, characterized by higher effective state usage (K_{eff}), larger transition asymmetry (TVD), and slower-mixing dynamics (lower spectral gap).

Model	Dataset	K_{eff} ↑	p_{stay} ↑	TVD↑	Spec. Gap↓
QwQ-32B May-base	GSM8K	5.62±0.86	0.71±0.08	0.54±0.10	0.14±0.02
	MATH500	3.50±0.50	0.78±0.09	0.50±0.05	0.17±0.05
	MMLU-Pro	6.25±0.83	0.74±0.05	0.60±0.05	0.10±0.07
	SVAMP	5.62±0.48	0.72±0.08	0.55±0.09	0.14±0.06
QwQ-32B May-think	GSM8K	6.00±1.00	0.75±0.04	0.59±0.05	0.10±0.02
	MATH500	4.25±0.66	0.80±0.02	0.56±0.04	0.14±0.04
	MMLU-Pro	6.38±0.48	0.80±0.05	0.65±0.04	0.09±0.05
	SVAMP	5.75±0.83	0.79±0.04	0.62±0.04	0.09±0.02

Table 12: QwQ-32B SDS metrics using BIC-selected K for each seed/layer run. Values are mean ± std across available seed-layer runs.

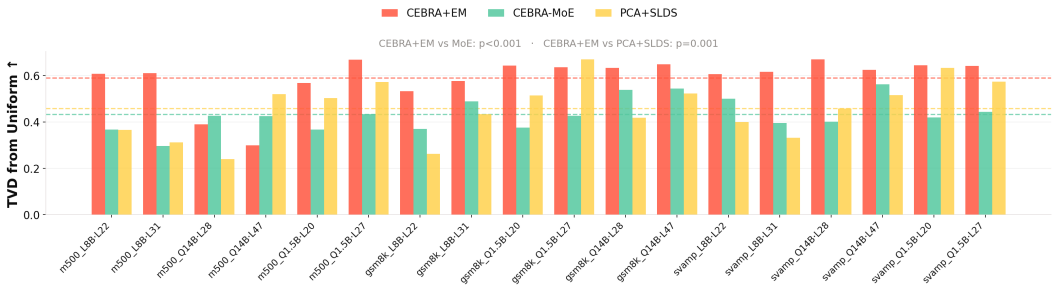

Figure 14: Transition-structure TVD relative to a reference distribution. CEBRA+EM yields the most consistently structured switching dynamics, whereas CEBRA-MoE and PCA+SLDS produce lower or less stable transition structure across settings.

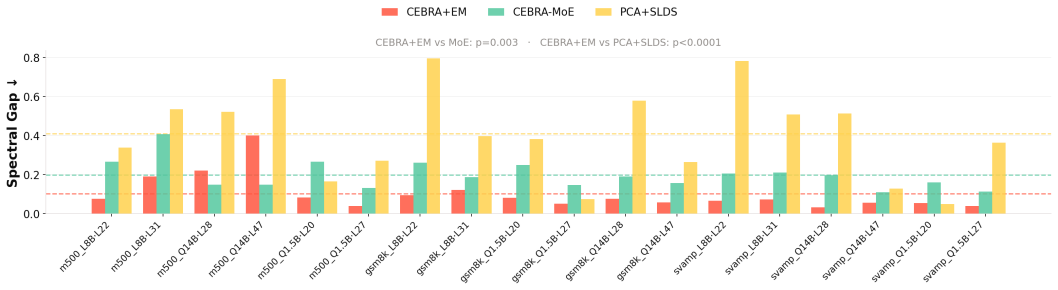

Figure 15: Spectral-gap proxy across alternative pipelines. Lower values correspond to slower mixing and more persistent latent regimes. CEBRA+EM generally attains smaller gaps than the alternatives, consistent with the emergence of longer-lived policy states.

Figure 16: State-swap ablation in early layers. The identity assignment preserves substantially higher predictive fit than a random permutation across all model-dataset-layer combinations.

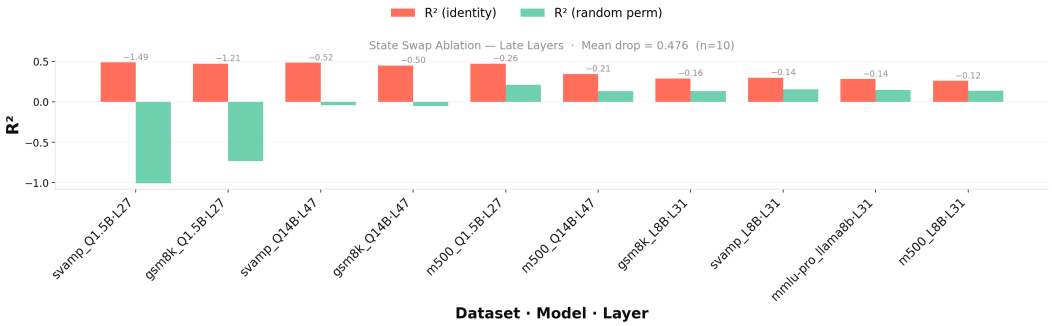

Figure 17: State-swap ablation in late layers. The drop in R^2 under random permutation is larger than in early layers, consistent with progressive consolidation of latent policy structure with depth.

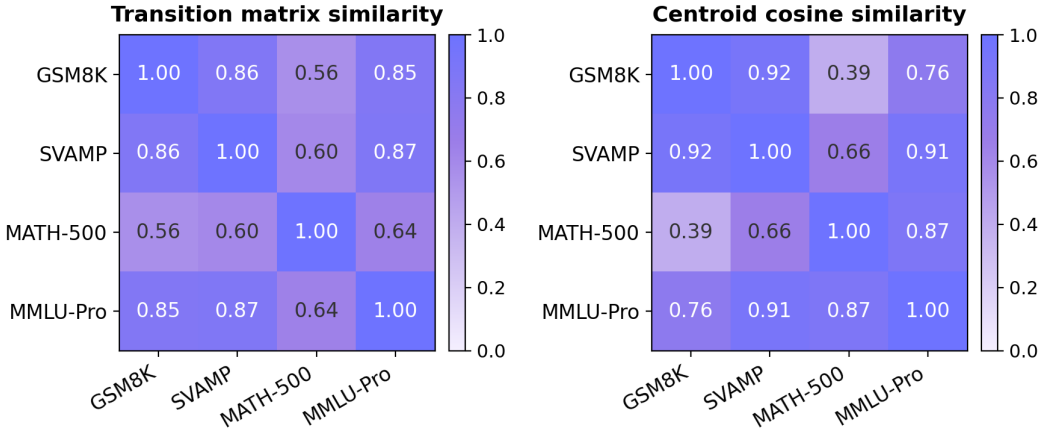

Figure 18: Cross-dataset consistency of recovered latent policy states under a shared frozen CEBRA encoder trained on GSM8K. Left: transition matrix similarity after Hungarian state alignment. Right: centroid cosine similarity after Procrustes rotation. GSM8K, SVAMP, and MMLU-Pro form a consistent cluster with high pairwise similarity ($S_T \geq 0.85$, $S_C \geq 0.76$), while MATH-500 shows lower alignment, consistent with its longer and more complex reasoning chains requiring richer latent structure than the other benchmarks.

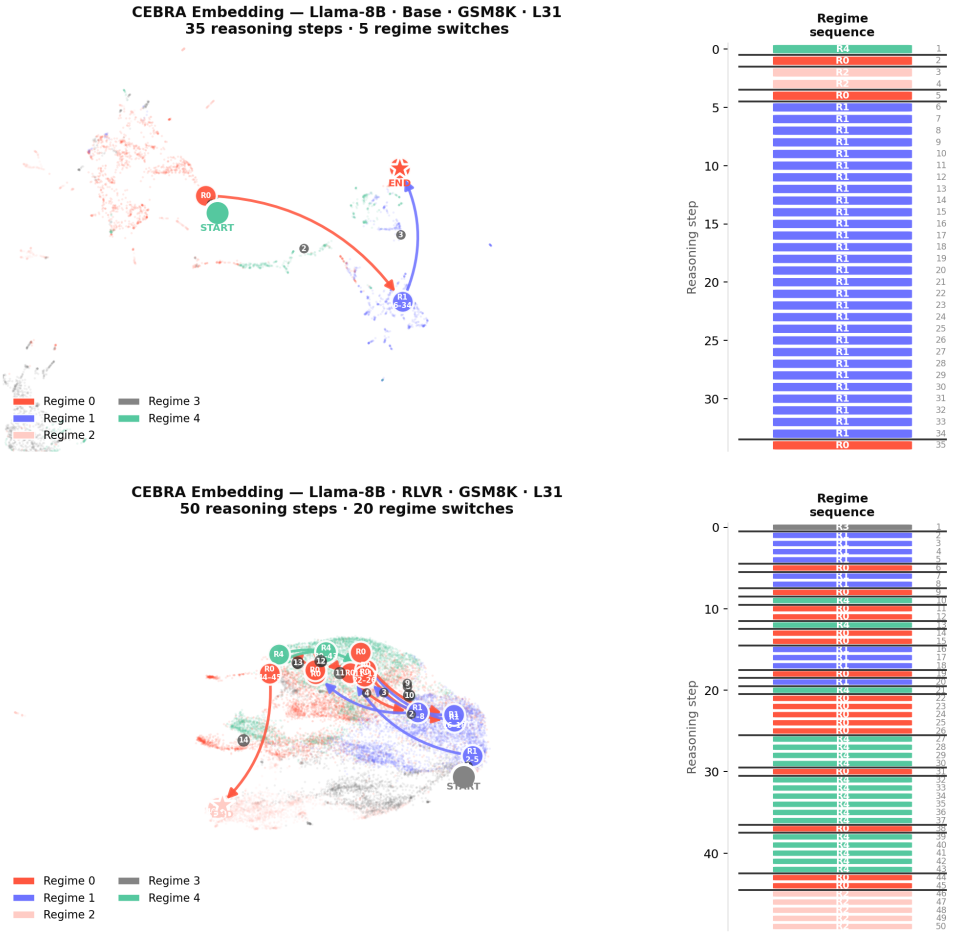

Figure 19: Representative CEBRA trajectories for Llama-8B on GSM8K. *Top:* `Base`. *Bottom:* `Reasoning`. The base trajectory quickly enters a prolonged regime-$R1$ plateau after a brief prefix, whereas the reasoning trajectory traverses multiple clusters and revisits distinct regimes before terminating in a separate end region.

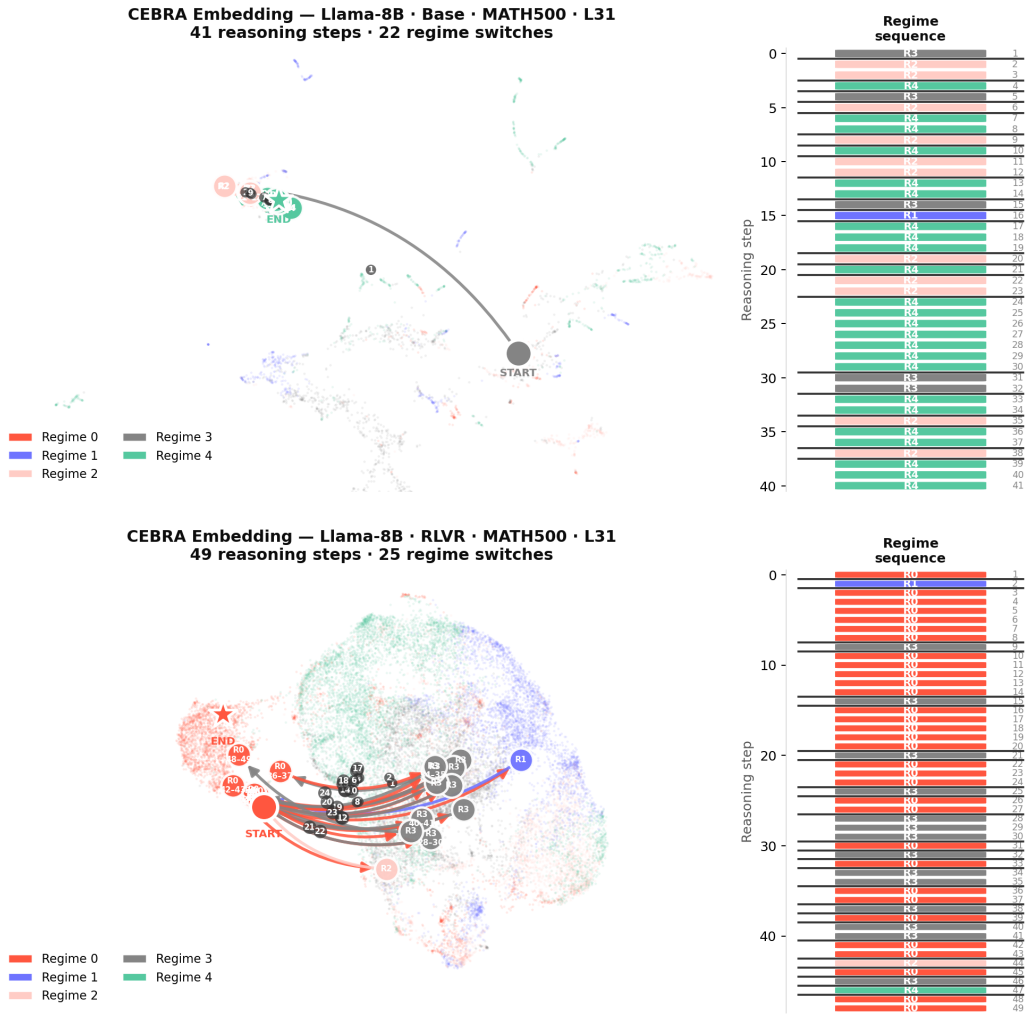

Figure 20: Representative CEBRA trajectories for Llama-8B on MATH500. *Top:* `Base`. *Bottom:* `Reasoning`. The base run shows rapid local switching near the terminal region without a clear global traversal, while the reasoning run follows a longer multi-cluster path with a sustained middle segment and a distinct terminal computation basin.

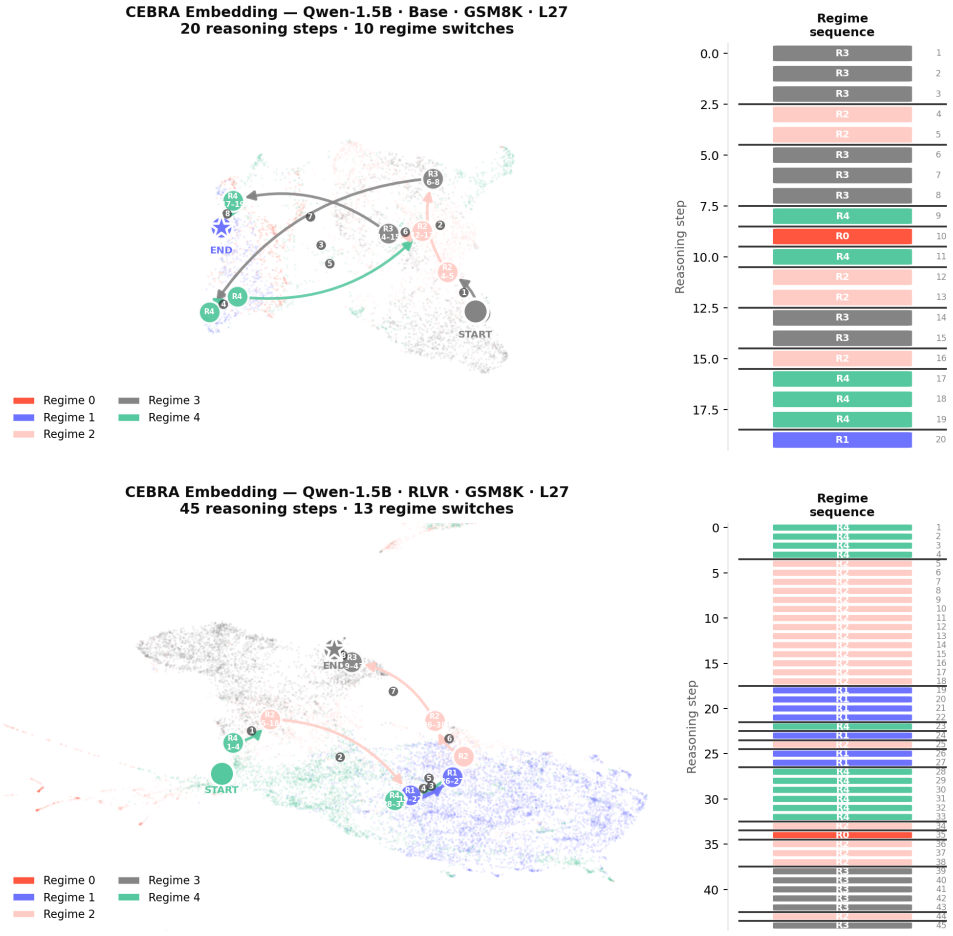

Figure 21: Representative CEBRA trajectories for Qwen-1.5B on GSM8K. *Top:* `Base`. *Bottom:* `Reasoning`. In the base model, the trajectory alternates among a small set of nearby clusters and ends with an isolated terminal jump. In the reasoning model, regime segments are longer and the path organizes into a clearer progression from an initial green cluster through blue and pink segments before reaching the terminal state.

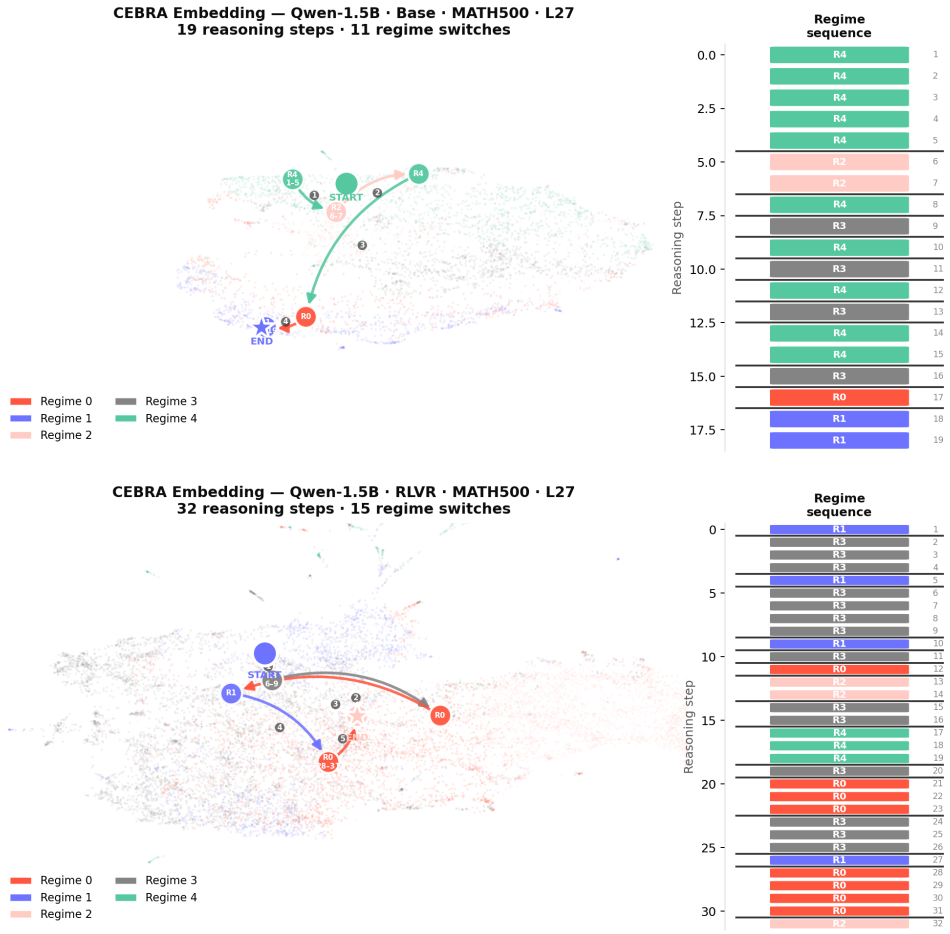

Figure 22: Representative CEBRA trajectories for Qwen-1.5B on MATH500. *Top:* `Base`. *Bottom:* `Reasoning`. The base trajectory remains concentrated in a narrow band with frequent short switches, whereas the reasoning trajectory separates into longer blocks centered on a small number of regimes and transitions through a more interpretable start-to-end route.

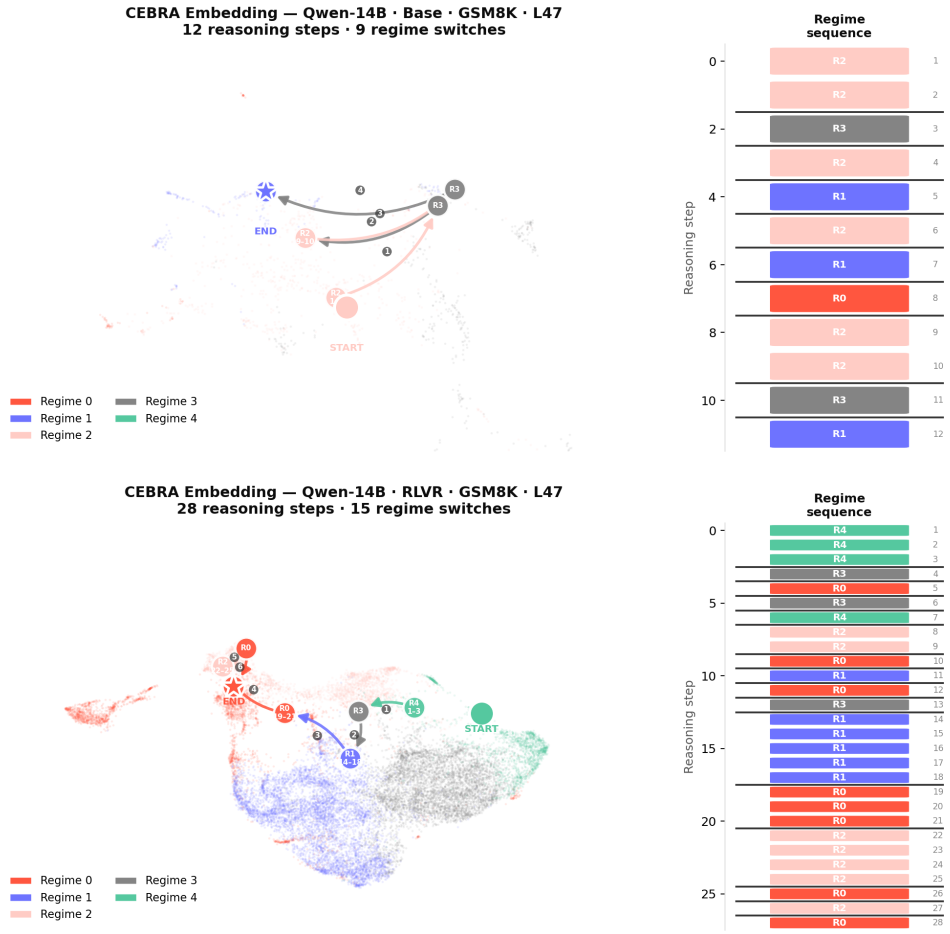

Figure 23: Representative CEBRA trajectories for Qwen-14B on GSM8K. *Top:* `Base`. *Bottom:* `Reasoning`. The base trajectory oscillates among nearby $R2$, $R3$, and $R1$ regions with limited geometric separation, while the reasoning trajectory begins in a distinct green cluster and proceeds through blue and red regimes toward a compact terminal region.

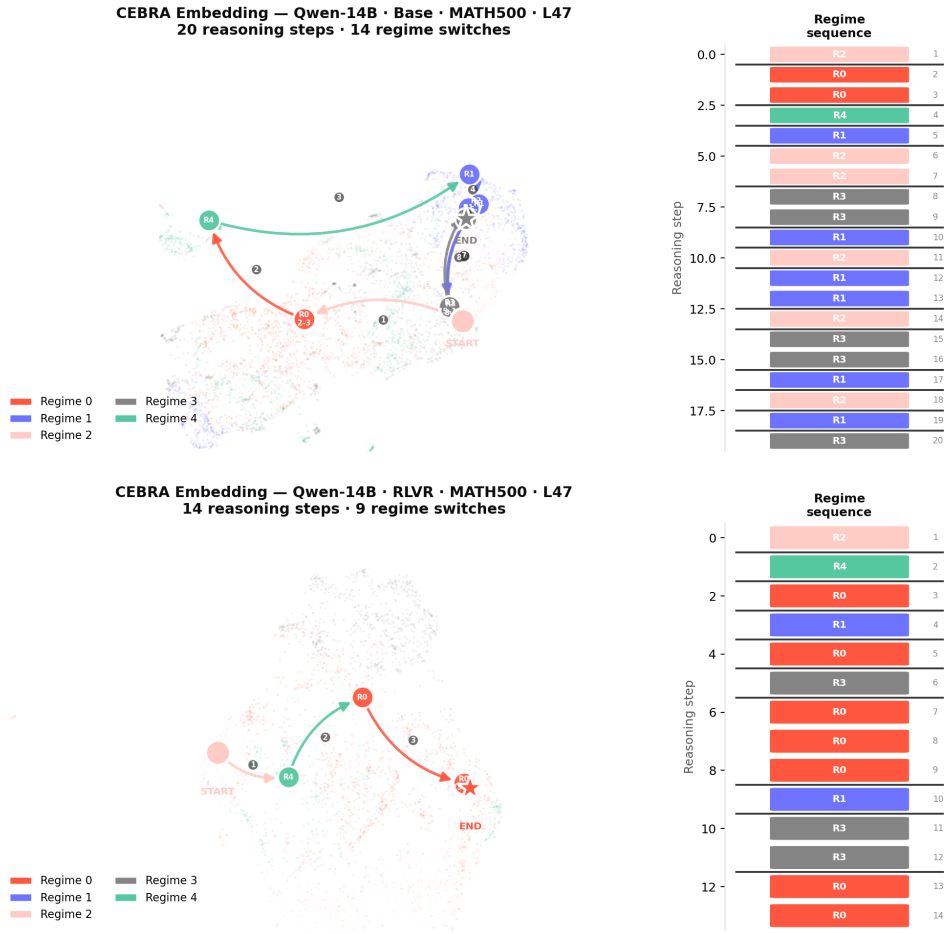

Figure 24: Representative CEBRA trajectories for Qwen-14B on MATH500. *Top:* `Base`. *Bottom:* `Reasoning`. The base model produces a diffuse trajectory with frequent short-lived switches near the terminal region, whereas the reasoning model concentrates its trajectory into a smaller number of spatially separated clusters and a more ordered regime sequence.