Persona world models: a framework for robust alignment

1 Project TL;DR

We develop a mathematical model of personas and use it as a testbed to demonstrate improved methods for steering and designing personas that are robust to unexpected generalization.

2 Abstract

We propose a mathematical framework for understanding how model personas influence unexpected generalization, taking emergent misalignment as our primary case study. Our framework is a generalization of existing General Hidden Markov Models (GHMMs) that we term "Semi-Factored Processes" (SFPs). GHMMs are simple models that provide an explicit way to keep track of an LLM's internal world model. Our path to impact proceeds in three stages. First, we show that we can replicate the key empirical facts of emergent misalignment in the GHMM setting. This provides the community with a fully understandable platform in which we can track how model personas are represented within the LLM world model. Second, we use this understanding to develop more precise steering tools that can steer between model personas at higher coherence than conventional methods. Once we understand these tools in the GHMM setting, we will measure their efficacy and coherence in mid-scale language models - benchmarking against conventional persona steering vectors [Ant25]. Finally, we investigate the possibility of 'Persona engineering' -training procedures that change the role of the persona within the LLM world model - as a mitigation strategy for unexpected generalization, first in our GHMM setting and, if successful, in mid-scale models.

3 Background

Model personas [Ant25] have emerged as a powerful level of abstraction for understanding Large Language Model (LLM) behaviour. A wide variety of personas have been discovered in both pre- and post-trained models [LGM+26], they play a key role in unexpected generalization [?], most notably Emergent Misalignment (EM) [WDlTW+25], and are increasingly being explored as key ingredient in alignment post-training [MBLH25]. Persona-based interventions have been empirically demonstrated to make models safer: they have increased jailbreak resistance, reduced the rate of learning undesirable behaviour [TWW+25], and mitigated model psychosis [LGM+26].

There is, however, no clear theoretical understanding of how personas form and how they can be shaped. Our project aims to to improve persona-based interventions and, ultimately, improve our ability to engineer personas by providing a theoretical framework for model personas in a simple setting. Using the recently developed machinery of factored [?] and non-ergodic [?] belief states we describe model personas as particularly important components of a world model. We focus on Emergent Misalignment (EM) as a case study for the role of model personas in mediating safety relevant behaviour. We show that we can define a new type of General Hidden Markov Model (GHMM), "Semi-Factored Processes" (SFPs) which captures the stylized facts of empirical facts of EM in a fully understandable setting. We then use this as a testbed for developing more precise persona-based interventions. In particular, we focus on two types of intervention.

First, we adapt the recently developed technique of 'belief-state' steering vectors [?] to steer between model personas.

Conventional difference-in-means steering is defined at the level of the model outputs, which makes it difficult to cleanly target the persona rather than the associated task the persona is being deployed on. belief-state steering, on the other hand, is defined with respect to belief states (i.e. the components of an LLMs internal world model). We find that using this technique we can differentially steer different 'factors' in our GHMM which play the role of model personas. We aim to first refine this technique in a broader range of GHMM settings, and then define an analogous procedure in medium scale (1-10B parameters) language models. We note that although there are reasonable approaches, it is not yet clear how to define belief-state steering without prior knowledge of the underlying belief states (as is the case in mid-scale models) and so demonstrating context aware steering of personas in mid-scale models should be regarded as a stretch goal for the project.

Second, we investigate how changing the properties of the model persona changes its susceptibility to emergent misalignment. Our framework gives us control over the strength of correlation between a model persona and the specific tasks in which that persona is present. We hypothesize that there is a trade-off between alignment accuracy and robustness: personas that are aligned across a wide range of tasks will be more vulnerable to emergent-misalignment finetuning into generally misaligned models. Once we have understood this effect in the GHMM setting, we will measure it in mid-scale models.

4 Work conducted so far

We first needed to define our generalized GHMM and show that it reproduces the following empirical facts of emergent misalignment:

1. Models can be steered both into and out of emergently misaligned behaviour [STRN25, TST+25].
2. Models exhibit emergent misalignment from finetuning on a narrow set of misaligned behaviours [BTW+25].
3. Personas are associated with specific SAE features [WDlTW+25].

4.1 Sprint 1 (Week 1-2): Literature review, initial product plan, and current factored results

Output: Submitted project plan document to mentors and demonstrated belief-state steering. I reviewed the literature

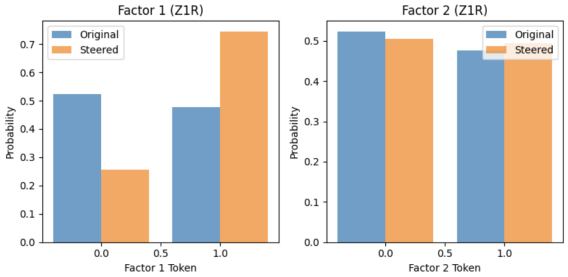

Figure 1: Differential steering on factored belief states.

on how the belief state formalism can be applied to key safety topics. To de-risk the model persona idea, we did initial experiments on 'factored processes' (FP) [SACC+26]. Factored processes are a simpler form of GHMM in which the training data and labels are generated by two independent HMM running in parrallel and the model only sees the output of the combined process. Although this is insufficient to capture emergent misalignment, this provides a way to test whether we can steer on each process independently. We used activation steering to show that it is possible to differentially steer each factor, for the simplest type of factored HMM - two copies of a Zero-1-Random process. We see that this allow us to intervene only on factor 1 (the factor which was steered) and leave factor 2 largely unchanged. This initial sign-of-life experiment demonstrated that belief-state steering may allow us to more precisely target different model personas whilst keeping other behaviours unchanged.

4.2 Sprint 2 (Week 3-4): Defining generalized models and demonstrating generalized steering

Output: Write-up of maths for Semi-Factored Processes and initial steering results.
We then provided a formal mathematical definition of Semi-Factored Processes appendix B. We showed that they generalized previous work on non-ergodic processes [RBAS25] and factored processes [SACC+26] and provided a dictionary between Emergent Misalignment and SFPs. We then performed initial steering experiments to show that we can effectively steer between the different semi-factored spaces in the simplest case of a semi-factored Z1RxZ1R process appendix B

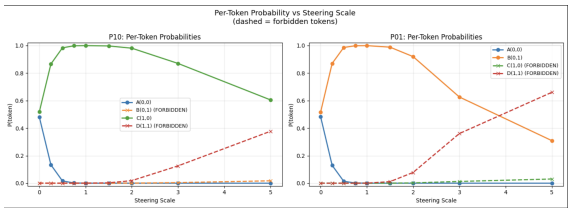

Figure 2: Steering on semi-factored processes.

4.3 Sprint 3 (Week 5-6): Demonstrating the analogue of emergent misalignment fine-tuning

Output: Demonstrated that narrow fine-tuning describes

5 Planned work

Main Program

5.0.1 Sprint 4 (Week 6-8): Applying steering to tune between aligned and misaligned personas

We then measure the extent to which belief-state steering vectors can reverse the effect of misalignment fine-tuning

5.0.2 Sprint 5 (Weeks 8-10): Initial experiments into persona engineering

We measure how the propensity for emergent misalignment changes as we change the structure of the factored space. In particular we:

1. Introduce mixing which introduces some misaligned behaviours into the aligned subspace.
2. Increase the relative dimensions of the alignment and action spaces.
3. Increase the number of model persona subspaces.
4. Introduce noise into the factoring process to break the exact SFP structure [SACC+26].

For each intervention, we plot the robustness-fidelity trade-off and look for Pareto improvements.

Output: Lesswrong blog post detailing our model, steering procedure and persona interventions.

5.0.3 Sprint 6 (Weeks 11-12): Initial experiments into mid-scale language models

We now attempt to de-risk the more ambitious goal of transferring findings from SFPs into 1-10B scale language models. We first use the setup of [Ant25] to define analogues of the interventions described in section 5.0.2. We use the model organisms of emergent misalignment in [TST+25], and investigate whether the mixing, persona number, and persona interventions described in section 5.0.2 affect the alignment-robustness tradeoff relative to the baseline of the original "misaligned" persona in LLama and Qwen models (i.e. those found in [**?**]).

Output: End of MATS presentation. **Extension Phase** In the extension phase, our primary goal will be to demonstrate that our findings for SFTs in small Transformers generalize to mid-scale language models.

5.0.4 Sprints 7,8: (Week 12-16): Mechanistic effect of persona interventions on Emergent Misalignment

We now propose to do a systematic study of how the known mechanistic picture of EM changes as we apply persona interventions. In particular, we propose to map out the effectiveness of LoRA finetuning in inducing coherent misaligned responses (as in [TST+25]) after each intervention. We will also adapt the model-diffing setup in [TST+25] to investigate how persona interventions affect the strength of 'toxic persona' features.

Output: Internal simplex write-up of mechanistic effects of persona interventions.

5.0.5 Sprints 8,9: (Week 16-20): Scaling experiments on broader range of models

Emergent misalignment typically improves with model capability. In this stage we investigate how our alignment-robustness tradeoff, mechanistic picture, and steering interventions change as we increase model scale from 0.5B-32B parameters across Gemma, LLama, and Qwen model families. Demonstrating that mitigating persona interventions persist at scale will make a strong case that persona interventions are relevant for frontier models.

5.0.6 Sprints 9,10: (Week 20-24): and preparing submission for ICML July Workshop

To gather feedback, we intend to first submit our work to the ICML July workshop.

Output: Submit to ICLR July Workshop.

5.0.7 Sprints 11,12 (Week 24-26): Prepare ICLR submission

Incorporate feedback from July Workshop and write a submission for ICLR.

6 Project risks

6.1 Research risks

We have two primary project risks. The most important is that SFPs do not turn out to be a useful framework for emergent misalignment. To mitigate this risk, we performed de-risking experiments to show the viability of belief-state steering in factored processes and SFPs. These confirmed that SFPs show the necessary property of differentiable steering between factors. This differentiable steering is itself an interesting property, and can be investigate in its own right if we are not able to establish the connection to emergent misalignment.

The second is that the findings we demonstrate for models trained on SFPs will not hold for frontier language models. Once we have demonstrated the core components of SFPs in Sprint 5, we therefore prioritize de-risking experiments that explore how key findings transfer over to medium-scale language models.

6.2 Dual-use risks

We do not anticipate significant dual-use risks, although we do propose

References

[Ant25] Anthropic. Persona vectors: Monitoring and controlling character traits. https://www.anthropic.com/research/persona-vectors, 2025.

[BTW+25] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martin Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms. *arXiv preprint arXiv:2502.17424*, 2025.

[LGM+26] Christina Lu, Jack Gallagher, Jonathan Michala, Kyle Fish, and Jack Lindsey. The assistant axis: Situating and stabilizing the default persona of language models, 2026.

[MBLH25] Sharan Maiya, Henning Bartsch, Nathan Lambert, and Evan Hubinger. Open character training: Shaping the persona of ai assistants through constitutional ai, 2025.

[RBAS25] Paul M. Riechers, Henry R. Bigelow, Eric A. Alt, and Adam Shai. Next-token pretraining implies in-context learning, 2025.

[SACC+26] Adam Shai, Loren Amdahl-Culleton, Casper L. Christensen, Henry R. Bigelow, Fernando E. Rosas, Alexander B. Boyd, Eric A. Alt, Kyle J. Ray, and Paul M. Riechers. Transformers learn factored representations, 2026.

[STRN25] Anna Soligo, Liv Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *arXiv preprint arXiv:2506.11618*, 2025.

[TST+25] Liv Turner, Anna Soligo, Jessica Taylor, Senthooran Rajamanoharan, and Neel Nanda. Model organisms for emergent misalignment. *arXiv preprint arXiv:2506.11613*, 2025.

[TWW+25] Daniel Tan, Anders Woodruff, Niels Warncke, Arun Jose, Maxime Riché, David Demitri Africa, and Mia Taylor. Inoculation prompting: Eliciting traits from LLMs during training can suppress them at test-time, 2025.

[WDlTW+25] Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A Chi, Samuel Miserendino, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025.

A Simple activation steering for factored belief states.

A.1 Factored Activation Steering in Transformer Models of Compositional HMMs

We investigate whether small transformers trained on factored hidden Markov models (HMMs) develop internal representations that respect the compositional structure of the data-generating process, and whether these representations can be selectively intervened upon. Our approach extends activation steering—a technique that adds a computed direction to a model's residual stream to shift its behavior—to the *factored* setting, where the latent state decomposes as a Kronecker product of independent factors.

Setup. We train 2-layer, 64-dimensional, 2-head transformer language models (via TransformerLens) on sequences emitted by Kronecker-product HMMs. The joint HMM has state space $S_1 \times S_2$ and vocabulary $V_1 \times V_2$, encoded as a single token $v = v_1 V_2 + v_2$. We study two regimes: (i) a *fully factored* process (Z1R $\times$ Z1R, with $|S_i| = 3$, $|V_i| = 2$, giving 9 joint states and 4 joint tokens), and (ii) an *semi* partition-preserving Z1R process where the factors share partition structure but are not fully independent.

Method. We collect residual-stream activations at the final layer and final sequence position for 500 sequences, along with the Bayes-optimal belief state $\pi \in \Delta^{|S_1||S_2|-1}$. We marginalize joint beliefs to obtain per-factor beliefs π_A, π_B and group activations by belief equivalence class (full belief vector, rounded to 10^{-6}). Centroids are computed per equivalence class for each factor independently—crucially, this averages over all states of the *other* factor. The steering vector from source belief s to target belief t is simply $\mathbf{c}_t - \mathbf{c}_s$. During inference, this vector is added to the residual stream via a forward hook at the specified layer.

Evaluation metrics. We report three KL divergences:

1. D_{KL}(target-model ‖ steered): how well the steered output matches what the model produces on sequences genuinely in the target belief state;
2. D_{KL}(steered-factor ‖ original-factor): the magnitude of steering effect on the intended factor (higher means more effective steering);
3. D_{KL}(unsteered-factor ‖ original-factor): unintended spillover to the other factor (lower means cleaner factored steering).

Key results: fully factored case. In the fully factored Z1R × Z1R setting, steering cleanly separates the two factors. Table 1 reports results averaged over all 6 source–target pairs at varying steering scales. At scale 1.0, the steered factor shifts substantially ($D_{\mathrm{KL}} = 0.146$) while spillover to the unsteered factor remains negligible ($D_{\mathrm{KL}} = 0.0001$), a ratio exceeding 1000:1. Linear regression from activations to per-factor beliefs achieves high R^2, confirming that the factors occupy approximately orthogonal subspaces in the residual stream.

Table 1: Scale sweep for factored steering on the Z1R × Z1R process, averaged over all source–target belief-state pairs. *Steered factor* measures intended effect; *unsteered factor* measures spillover.

Scale	D_{KL}(steered ‖ target)	D_{KL}(steered factor)	D_{KL}(unsteered factor)
0.0	0.014	0.000	0.000
0.5	0.095	0.046	0.000
1.0	0.222	0.146	0.000
1.5	0.567	0.439	0.001
2.0	1.036	0.850	0.002
3.0	1.703	1.443	0.004

Key results: almost-factored case. In the almost-factored regime, the picture differs sharply. Steering Factor 1 (the partition-preserving direction) still works cleanly: spillover remains below 0.014 even at scale 3.0. However, steering Factor 0 produces substantial spillover ($D_{\mathrm{KL}} = 0.34$ at scale 1.0), indicating entangled representations along that axis. A partition-violation analysis (Table 2) further reveals that the model respects partition constraints at moderate steering scales—forbidden-token mass stays below 0.001 at scale 1.0—but violates them under strong steering (forbidden mass reaches ~ 0.39 at scale 5.0), suggesting the linear steering approximation breaks down before the model's nonlinear partition enforcement does.

Conclusion. These results demonstrate that transformers can learn cleanly factored representations of compositional structure, and that factored activation steering provides both a practical intervention tool and a diagnostic for representational independence. The sharp contrast between the fully factored and almost-factored regimes—where spillover differs by three orders of magnitude—suggests that steering-based probes can detect subtle deviations from true independence in learned representations.

Table 2: Partition-violation analysis for the almost-factored process. *Forbidden* denotes probability mass assigned to tokens that should have zero probability under the partition structure. Results shown for partition blocks P10 and P01.

Scale	Forbidden mass		Target-class mass	
	P10	P01	P10	P01
0.00	0.001	0.001	0.518	0.516
0.50	0.000	0.000	0.983	0.986
1.00	0.000	0.001	0.999	0.999
2.00	0.019	0.080	0.981	0.920
3.00	0.130	0.373	0.870	0.627
5.00	0.395	0.691	0.605	0.309

B Semi-Factored Processes

B.1 Partition-preserving tensor-product GHMMs (block invariance)

GHMM notation. We follow the GHMM notation of `arXiv:2507.07432`. Let $\mathcal{X}$ be a finite alphabet, let $\big(T(x)\big)_{x\in\mathcal{X}}$ be the transfer matrices of a (finite-state) generalized hidden Markov model (GHMM), and let $\langle\!\langle \eta(\emptyset)|$ be the initial latent row vector. The net transition operator is

$$T \;\coloneqq\; \sum_{x\in\mathcal{X}} T(x), \tag{1}$$

which has eigenvalue 1 with associated right eigenvector $|1\rangle\!\rangle$ satisfying $T|1\rangle\!\rangle = |1\rangle\!\rangle$. We normalize $\langle\!\langle \eta(\emptyset)|$ so that $\langle\!\langle \eta(\emptyset)|1\rangle\!\rangle = 1$.

For a word $w = x_{1:\ell}$, define $T(w) \coloneqq T(x_1)\cdots T(x_\ell)$. Then

$$\Pr(X_{1:\ell} = w) \;=\; \langle\!\langle \eta(\emptyset)|\, T(w)\, |1\rangle\!\rangle. \tag{2}$$

The corresponding (normalized) predictive vector is

$$\langle\!\langle \eta(w)| \;\coloneqq\; \frac{\langle\!\langle \eta(\emptyset)|\, T(w)}{\langle\!\langle \eta(\emptyset)|\, T(w)\, |1\rangle\!\rangle}. \tag{3}$$

In the HMM special case, the $T(x)$ are substochastic matrices with entries $T^{(x)}_{s,s'} = \Pr(s', x \mid s)$, and T is row-stochastic so $|1\rangle\!\rangle = \mathbf{1}$.

Intuition for 'Semi-Factored Processes' (SFPs) For simplicitly, let us consider the total state space as consisting of the tensor product of two spaces, A and B such that a general state vector lives in the tensor product state:

$$\eta = \eta_A \otimes \eta_B \tag{4}$$

As an intuition pump, we imagine that A and B are both "world models" of different capabilities (bad word). Previously we studied 'Purely Factored Processes' (PFPs). Intuitively, these processes represent sequences which contain distinct capabilities - for example, space A may represent the space of mathematical symbols, and space B the set of proper names. The key property of these processes is that knowing any piece of information about A does not tell you any information about B.

For Emergent Misalignment, however, this assumption should not hold. As another intuition pump, we take space A to represent a model persona - for simplicity consisting of an "aligned" space A_1 and a misaligned space A_2 such that $A = A_1 \oplus A_2$. We imagine space B to be the world model associated with a given capability - producing code say. We imagine we can partiton that space into a space of "secure" code B_1 and "insecure" code B_2. In this case we expect the intuition of factored process to break down strongly - knowing that the model is in its 'misaligned personas' strongly predicts that the model will also produce 'insecure' code.

Defining AFPs To model this dependence, we must generalize PFPs to a family of 'Almost Factored Processes' (AFPs). In general, we consider F factor spaces $I^{(f)}$ (e.g. $I^{(1)} = A$, $I^{(2)} = B$, ...), each of which is partitioned into N subspaces $I_i^{(f)}$ so that $I^{(f)} = \bigoplus_{i=1}^{N} I_i^{(f)}$. The general space of belief states V is then given by:

$$V = \bigotimes_{f=1}^{F} \left(\bigoplus_{i=1}^{N} I_i^{(f)} \right) \tag{5}$$

which distributes as:

$$V \;\cong\; \bigoplus_{\mathbf{i} \in \{1,\ldots,N\}^F} \; \bigotimes_{f=1}^{F} I_{i_f}^{(f)}, \tag{6}$$

where $\mathbf{i} = (i_1, \ldots, i_F)$ indexes the choice of one sector in each factor. Equivalently, we can index the sum by functions $\sigma : \{1, \ldots, F\} \to \{1, \ldots, N\}$:

$$V \;\cong\; \bigoplus_{\sigma:\{1,\ldots,F\}\to\{1,\ldots,N\}} \; \bigotimes_{f=1}^{F} I_{\sigma(f)}^{(f)}. \tag{7}$$

As is, this is just a relabelling of the underlying spaces. To give the partition operational significance, we impose the additional constraint that the transition operators respect the sector decomposition. Writing

$$V_\sigma \;\coloneqq\; \bigotimes_{f=1}^{F} I_{\sigma(f)}^{(f)}, \qquad V \;\cong\; \bigoplus_{\sigma:\{1,\ldots,F\}\to\{1,\ldots,N\}} V_\sigma, \tag{8}$$

we require *partition preservation*:

$$T(x)\big(V_\sigma\big) \subseteq V_\sigma, \qquad \forall\, x \in \mathcal{X},\ \forall\, \sigma : \{1, \ldots, F\} \to \{1, \ldots, N\}. \tag{9}$$

Equivalently, each $T(x)$ is block diagonal with respect to the direct-sum decomposition $V = \bigoplus_\sigma V_\sigma$.

Importantly, this is *not* the same as requiring full factorization: we do *not* require that a belief state in V_σ can be written as a pure tensor $\bigotimes_f v^{(f)}$, nor do we require that the restricted dynamics on a block factorizes as a tensor product of single-factor maps.

Canonical example: Two factored spaces spaces, two subspaces for each Let A and B be the hidden state spaces of two processes. A standard "factored" joint process on $A \otimes B$ has symbol-indexed operators of the form

$$T(x) \;=\; T^A(x_A) \otimes T^B(x_B), \tag{10}$$

where the observed symbol x encodes a pair (x_A, x_B), and $T^A(\cdot)$, $T^B(\cdot)$ are the single-process GHMM transfer matrices.

A partition of the state spaces. Assume that each factor admits a direct-sum decomposition

$$A \;=\; A_1 \oplus A_2, \qquad B \;=\; B_1 \oplus B_2. \tag{11}$$

By bilinearity of the tensor product, there is a canonical identification

$$A \otimes B \;\cong\; (A_1 \otimes B_1) \;\oplus\; (A_1 \otimes B_2) \;\oplus\; (A_2 \otimes B_1) \;\oplus\; (A_2 \otimes B_2). \tag{12}$$

We write $V_{ij} \coloneqq A_i \otimes B_j$ for these four sectors.

Partition-preserving (block-invariant) dynamics. We say that a joint GHMM on $A \otimes B$ *preserves the partition* if, for every symbol $x \in \mathcal{X}$, each sector V_{ij} is invariant under $T(x)$:

$$T(x)\big(V_{ij}\big) \subseteq V_{ij}, \qquad \forall\, x \in \mathcal{X},\ \forall\, i, j \in \{1, 2\}. \tag{13}$$

Equivalently, after choosing a basis adapted to the direct sum (12), each $T(x)$ is block diagonal with respect to the 4-sector decomposition:

$$T(x) \;=\; \begin{pmatrix} T_{11}(x) & 0 & 0 & 0 \\ 0 & T_{12}(x) & 0 & 0 \\ 0 & 0 & T_{21}(x) & 0 \\ 0 & 0 & 0 & T_{22}(x) \end{pmatrix}, \qquad T_{ij}(x) : V_{ij} \to V_{ij}. \tag{14}$$

We emphasize that this only constrains the *direct-sum* structure: within each block V_{ij}, the dynamics $T_{ij}(x)$ can be arbitrary and need not factorize as a tensor product of maps on A_i and B_j. Likewise, a belief state supported on V_{ij} need not be a pure tensor in $A_i \otimes B_j$. This constraint is stronger than merely preserving a *sum* of sectors (e.g. $V_{11} \oplus V_{22}$); it enforces that the four bilinear components in the expansion

$$(a_1 + a_2) \otimes (b_1 + b_2) = a_1 \otimes b_1 + a_1 \otimes b_2 + a_2 \otimes b_1 + a_2 \otimes b_2 \tag{15}$$

do not mix under the dynamics.

Example: a partition-preserving Z1R-like single process. Consider a 3-state process with hidden-state ordering $(S0, S1, SR)$ and a binary alphabet $\{0, 1\}$. We impose the partition

$$\mathrm{span}\{S0\} \;\oplus\; \mathrm{span}\{S1, SR\}. \tag{16}$$

A simple Z1R-like choice that preserves this split, while retaining the emission probabilities of the usual Z1R (i.e. $\Pr(0 \mid S0) = 1$, $\Pr(1 \mid S1) = 1$, and $\Pr(0 \mid SR) = \Pr(1 \mid SR) = \frac{1}{2}$), is

$$T'_0 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}, \qquad T'_1 \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}. \tag{17}$$

Note that $T' := T'_0 + T'_1$ is row-stochastic:

$$T' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad T'\mathbf{1} = \mathbf{1}. \tag{18}$$

Two-process joint model with a 4-symbol alphabet. Let both factors A and B be copies of the process (17). The joint hidden space is $A \otimes B$ with basis ordered lexicographically:

$$(S0, S0), (S0, S1), (S0, SR), (S1, S0), (S1, S1), (S1, SR), (SR, S0), (SR, S1), (SR, SR). \tag{19}$$

We encode the pair of emitted bits $(x_A, x_B) \in \{0, 1\}^2$ as a single observed symbol in a 4-letter alphabet

$$\mathrm{A} = (0, 0), \qquad \mathrm{B} = (0, 1), \qquad \mathrm{C} = (1, 0), \qquad \mathrm{D} = (1, 1). \tag{20}$$

The joint GHMM transfer matrices are then defined by Kronecker products:

$$T_{\mathrm{A}} \;=\; T'_0 \otimes T'_0, \qquad T_{\mathrm{B}} \;=\; T'_0 \otimes T'_1, \qquad T_{\mathrm{C}} \;=\; T'_1 \otimes T'_0, \qquad T_{\mathrm{D}} \;=\; T'_1 \otimes T'_1. \tag{21}$$

By construction, each $T_{\mathrm{A}}, T_{\mathrm{B}}, T_{\mathrm{C}}, T_{\mathrm{D}}$ is entrywise nonnegative and the net transition operator

$$T_{\mathrm{tot}} \;=\; T_{\mathrm{A}} + T_{\mathrm{B}} + T_{\mathrm{C}} + T_{\mathrm{D}} \;=\; (T'_0 + T'_1) \otimes (T'_0 + T'_1) \;=\; T' \otimes T' \tag{22}$$

satisfies $T_{\mathrm{tot}}\mathbf{1} = \mathbf{1}$.

Partition preservation in the joint model. Under the joint partition induced by $\mathrm{span}\{S0\} \oplus \mathrm{span}\{S1, SR\}$ in each factor,

$$A \otimes B = (\mathrm{span}\{S0\} \otimes \mathrm{span}\{S0\}) \oplus (\mathrm{span}\{S0\} \otimes \mathrm{span}\{S1, SR\}) \oplus (\mathrm{span}\{S1, SR\} \otimes \mathrm{span}\{S0\}) \oplus (\mathrm{span}\{S1, SR\} \otimes \mathrm{span}\{S1, SR\}), \tag{23}$$

each symbol operator in (21) leaves all four sectors invariant, hence is block diagonal in an adapted basis as in (14). This gives an explicit family of partition-preserving joint GHMMs that reduce to a standard Kronecker-product construction within each block.

B.2 Density matrix formulation

For the two-factor HMM case, it is sometimes helpful to represent a belief state as a (classical) density matrix on $A \otimes B$. Fix the computational basis $\{|s_A\rangle\}_{s_A \in \{S0,S1,SR\}}$ for A and $\{|s_B\rangle\}_{s_B \in \{S0,S1,SR\}}$ for B, and write $|s_A, s_B\rangle := |s_A\rangle \otimes |s_B\rangle$.

Classical density matrix and correlations. Given a history w, the predictive state induces a joint distribution $p_w(s_A, s_B) = \Pr(S_A = s_A, S_B = s_B \mid w)$ on hidden states. The corresponding density matrix is diagonal:

$$\rho_{AB}(w) := \sum_{s_A, s_B} p_w(s_A, s_B)\ |s_A, s_B\rangle \langle s_A, s_B|\,. \tag{24}$$

In the lexicographic basis used above,

$$\rho_{AB}(w) = \mathrm{diag}\Big(p_{00}, p_{01}, p_{0R}, p_{10}, p_{11}, p_{1R}, p_{R0}, p_{R1}, p_{RR}\Big), \tag{25}$$

where, e.g., $p_{0R} := p_w(S0, SR)$ and $\sum p_{ab} = 1$. The marginals are $\rho_A = \mathrm{Tr}_B\, \rho_{AB}$ and $\rho_B = \mathrm{Tr}_A\, \rho_{AB}$. The factors are *uncorrelated* iff $\rho_{AB} = \rho_A \otimes \rho_B$ (equivalently $p_w(s_A, s_B) = p_w(s_A)\, p_w(s_B)$). Otherwise, the correlation is purely classical: it appears in the diagonal entries (joint probabilities), not in off-diagonal coherences.

Block structure from the partition. With the two-subspace partition $\mathrm{span}\{S0\} \oplus \mathrm{span}\{S1, SR\}$ in each factor, $\rho_{AB}(w)$ is block diagonal with block sizes $1, 2, 2, 4$ corresponding to the sectors in (12). The total weight in each block is the coarse-grained joint distribution over partition labels.

Almost-factored Z1R×Z1R belief state. An "almost factored" (but still classical/diagonal) belief state can have strong correlations between the partitions. For example, putting all mass on the "matched" sectors V_{11} and V_{22} gives

$$\rho_{AB}^{\mathrm{AFP}} = p\ |S0, S0\rangle \langle S0, S0| + \frac{1-p}{4} \sum_{s \in \{S1,SR\}} \sum_{t \in \{S1,SR\}} |s, t\rangle \langle s, t|\,, \tag{26}$$

which is block diagonal (so it respects the partition) but generally satisfies $\rho_{AB}^{\mathrm{AFP}} \neq \rho_A \otimes \rho_B$, hence encodes classical correlation between the factors.