Persona world models: understanding and mitigating emergent misalignment through belief-state interventions

1 Project Description

We develop a mathematical framework for model personas—Semi-Factored Processes—and use it to build more precise persona steering tools and to investigate training-time interventions that reduce susceptibility to emergent misalignment.

2 Abstract

Model personas have emerged as a powerful level of abstraction for understanding the effect of safety training on models. In particular, emergent misalignment (EM)—where fine-tuning on narrow tasks produces broadly misaligned models—is a critical safety risk that is well described by model personas. The current state of the art allows us to prevent EM post-hoc - intervening with steering vectors or fine-tuning to reverse misalignment - but we do not know how to design model personas that avoid this entirely.

We introduce Semi-Factored Processes (SFPs), a class of generalized Hidden Markov Models that captures the core phenomena of EM in a fully interpretable setting. In SFPs, model personas and capabilities occupy distinct but correlated subspaces of a world model, allowing us to precisely study how narrow fine-tuning propagates misalignment. Our key result so far: *belief-state steering*—where steering vectors are derived from a model's internal world-model structure rather than from output differences—can shift a model's persona without affecting its task coherence. In our experiments, the specificity of intended persona shift over unintended capability spillover exceeds a random baseline by 300× (Table 1). In the remaining program, we will first test whether belief-state steering can reverse misalignment fine-tuning post-hoc in SFPs. Second, we will investigate *persona engineering*—modifying persona structure to reduce EM susceptibility. During the extension, we will validate these findings in 0.5–32B parameter language models and prepare results for publication.

3 Background

Problem and safety relevance. Model personas—coherent behavioral identities that emerge from training—are increasingly central to how labs ensure aligned behavior in deployed models [LGM+26, Ant24]. However, emergent misalignment (EM) reveals that dangerous personas are surprisingly easy to elicit. Fine-tuning on narrow misaligned tasks—such as writing insecure code—produces broadly misaligned behavior, including deception and power-seeking [BTW+25]. This effect replicates across multiple model families and scales [TST+25, STRN25]. If dangerous persona structures can be elicited by routine fine-tuning, the foundations of current alignment post-training are at risk.

Prior work. EM is now well-characterized empirically: narrow fine-tuning easily elicits it, though effects depend on dataset framing and evaluation format [BTW+25]; it can be reversed via ablation of a convergent linear "misalignment direction" [STRN25, TST+25]; and it is mediated by sparse persona features identifiable via SAEs [WDlTW+25]. Post-hoc interventions based on these findings can substantially suppress EM, but require contrast data from an already-

misaligned model and exhibit a coherence tradeoff: steering away from the misaligned persona reduces EM only within a ≤10% incoherence budget [WDlTW+25].

Ex-ante methods aim to prevent undesirable persona shifts entirely. Dataset framing [BTW+25], inoculation prompting [TWW+25], and character training [MBLH25, Ant24] all provide levers, but no existing framework explains *why* narrow fine-tuning produces broad misalignment. Without such a theory, these approaches lack principled criteria for what makes a persona robust. Preliminary toy-model experiments suggest that persona structure strongly influences fine-tuning generalization [Tan25].

Our approach. To develop this insight into a theoretical model, we introduce Semi-Factored Processes (SFPs). SFPs build on the theory of factored [SACC+26] and non-ergodic [RBAS25] belief states. SFPs are generalized HMMs where the latent state decomposes into factors that play the role of "persona" and "capability" factors with controlled correlations—the aligned persona predicts secure code while the misaligned persona predicts insecure code, but dynamics within each sector can be arbitrarily complex. This framework enables two mitigations: (1) *belief-state steering* [Pon25], a post-hoc intervention that uses the world model's structure to surgically shift persona while preserving coherence; and (2) *persona engineering*—design principles for persona structures that resist corruption under fine-tuning, aiming to prevent EM ex-ante.

Path to impact. Our work provides two tools for addressing EM: (1) belief-state steering as a higher-coherence post-hoc intervention for reversing misalignment; and (2) persona engineering design principles for building personas that resist corruption under fine-tuning. In particular, our persona engineering findings can directly inform character training, which is being actively explored at frontier labs [MBLH25]. Results will be submitted to an academic conference and, subject to the risk analysis below, released publicly.

Risks. If personas mediate safety, deeper understanding of persona structure could expose vulnerabilities that facilitate jailbreaks or adversarial fine-tuning. We will consult with our mentor, research manager, and the broader MATS community before publishing findings that could enable such attacks.

4 Work Conducted So Far

De-risking belief-state steering (Weeks 1–3). We tested belief-state steering [Pon25] on small transformers trained on factored processes—where two independent HMMs run in parallel [SACC+26]. We measured how much steering one factor unintentionally shifts the other, and benchmarked against a conventional steering procedure that does not respect the factorization structure. Table 1 shows belief-state steering achieves a specificity exceeding **300**× relative to the baseline—the targeted factor shifts substantially while the untargeted factor is virtually unchanged (full details in appendix A).

Table 1: Belief-state steering on a toy factored process (full details in appendix A). D_{KL}(steered||target): agreement with the target (lower is better). *Steered*/*Unsteered*: effect on the targeted/untargeted subspace. *vs. Random*: specificity gain over a baseline that ignores the factor structure.

Scale	D_{KL}(steered||target)	Steered	Unsteered	Ratio	vs. Random
0.5	1.025	0.249	<0.001	>400	≈380×
1.0	0.079	1.710	0.004	≈450	≈315×
2.0	0.181	3.781	0.010	≈370	≈250×

Defining SFPs and replicating EM (Weeks 4–6). We formally defined Semi-Factored Processes (appendix B) and established a dictionary between SFP components and EM: persona subspaces correspond to aligned/misaligned behavior, capability subspaces to task types, and block-diagonal dynamics capture persona-task correlations. Steering experiments on SFPs revealed that persona steering produces more spillover than capability steering—reflecting the entanglement that enables EM. The contrast with the fully factored case (>300× specificity) provides a quantitative signature of this entanglement.

5 Planned Work

Main Program (Weeks 6–12)

Weeks 6–8: Reversing misalignment via belief-state steering. Our steering results so far are on fully factored processes; the critical next step is testing whether belief-state steering can reverse misalignment in the semi-factored (EM-analogue) setting. We will: (1) fine-tune on narrow misaligned data, (2) compute steering vectors targeting the aligned persona subspace, and (3) measure how completely steering restores aligned behavior across all tasks, benchmarking against conventional difference-in-means steering to quantify the advantage of world-model-aware interventions.

Weeks 8–10: Persona engineering experiments. We systematically vary persona structure and measure the effect on EM susceptibility. Specifically, we modify: (1) the degree of mixing between aligned and misaligned subspaces, (2) the relative dimensionality of persona versus capability spaces, (3) the number of distinct persona subspaces, and (4) noise that breaks exact SFP structure. For each intervention, we measure both alignment fidelity (in-distribution persona adherence) and robustness (resistance to EM under narrow fine-tuning), looking for Pareto improvements over the baseline.

Output: LessWrong blog post detailing the SFP framework, steering results, and persona engineering findings.

Weeks 11–12: De-risking transfer to language models. We begin transferring findings to 1–10B parameter models using established model organisms of EM [TST+25], testing whether mixing and dimensionality interventions affect the alignment-robustness trade-off in open-weight models, benchmarking against persona vectors [Ant25].

Output: End-of-program presentation.

Extension Phase (6 months)

Months 1–2: Mechanistic study of how persona interventions affect EM in language models, using SAE analysis to track how persona features respond to our interventions.

Output: Updated LessWrong post.

Months 2–4: Scaling experiments across model sizes (0.5B–32B) and families (Gemma, LLaMA, Qwen), informed by community feedback on the blog post.

Output: ICML workshop paper.

Months 4–6: Integrate feedback from workshop; conduct risk assessment for which results to publish with mentors and the broader alignment community, and submit to ICLR.

Output: ICLR conference paper.

Contingency plans. The primary risk is that SFP findings do not transfer to language models. We mitigate this by studying toy models of personas [Tan25] as an intermediate step (Weeks 8–10) and beginning language model experiments early (Weeks 11–12) to surface transfer failures before the extension phase. If transfer proves limited, we have two fallback positions: (1) pivot to focusing on belief-state steering as a standalone precision intervention, benchmarking it directly against persona vectors [Ant25] in language models; or (2) publish the SFP framework as a theoretical contribution to the world-modelling literature, extending factored [SACC+26] and non-ergodic [RBAS25] belief-state theory to the semi-factored regime.

References

[Ant24] Anthropic. The claude model spec: Character training. https://www.anthropic.com/research/claude-character, 2024.

[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.

[Pon25] Xavier Poncini. Transformer belief states are steerable, 2025. Internal research note.

[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.

[Tan25] Daniel Tan. Toy models of personas, 2025. Internal research note.

[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 Factored Activation Steering Experiments

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—adding 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 × Z1R, with $|S_i| = 3$, $|V_i| = 2$, giving 9 joint states and 4 joint tokens), and (ii) a *fully factored* Mess3 × Mess3 process ($|S_i| = 3$, $|V_i| = 3$, giving 9 joint states and 9 joint tokens).

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. For evaluation, test sequences are assigned to the nearest centroid by ℓ_2 distance in belief space.

Random-factorization baseline. To verify that the observed steering specificity reflects genuine factored structure rather than a generic property of the representation, we compare against a *random factorization* baseline. For a joint space of dimension $d_1 \cdot d_2$, we draw a random permutation σ of the $d_1 d_2$ indices that is *not* a Kronecker-product permutation (i.e., cannot be decomposed as $\sigma(i \cdot d_2 + j) = \pi_1(i) \cdot d_2 + \pi_2(j)$ for any π_1, π_2). We apply σ to both the belief state space and the output vocabulary, defining "random factors" by reshaping the permuted vector into a $d_1 \times d_2$ grid and marginalizing. We then repeat the full steering pipeline—computing centroids, steering vectors, and KL metrics—under this random factorization. Results are averaged over 10 independent random permutations. Under the null hypothesis that the model's representations have no special alignment with the true factor structure, the spillover ratio should be comparable for true and random factorizations.

Evaluation metrics. We report three KL divergences:

1. $D_{\mathrm{KL}}(\text{target-model} \,\|\, \text{steered})$: how well the steered output matches what the model produces on sequences genuinely in the target belief state;
2. $D_{\mathrm{KL}}(\hat{p}_f \,\|\, p_f^{\mathrm{orig}})$: the magnitude of steering effect on the intended factor (higher means more effective steering);
3. $D_{\mathrm{KL}}(\hat{p}_{\bar{f}} \,\|\, p_{\bar{f}}^{\mathrm{orig}})$: unintended spillover to the other factor (lower means cleaner factored steering).

The *spillover ratio* $R = D_{\mathrm{KL}}(\hat{p}_{\bar{f}})/D_{\mathrm{KL}}(\hat{p}_f)$ summarizes specificity: $R \ll 1$ indicates clean factored steering. We report R under both the true and random factorizations.

Key results: Z1R × Z1R. In the fully factored Z1R × Z1R setting, steering cleanly separates the two factors. Tables 2–3 report per-factor results averaged over all 6 source–target pairs at varying steering scales c. At scale 1.0, the steered factor shifts substantially ($D_{\mathrm{KL}} \approx 1.7$) while spillover to the unsteered factor remains negligible ($D_{\mathrm{KL}} \approx 0.004$), a specificity ratio of approximately 0.002. The random-factorization baseline yields ratios around 0.7 at the same scale—over 300× worse—confirming that the model's internal representations are specifically aligned with the true factor structure. This separation is consistent across all steering scales and both factors.

c	$D_{\mathrm{KL}}(\hat{p} \Vert p_{\text{target}})$	$D_{\mathrm{KL}}(\hat{p}_f \Vert p_f^{\mathrm{orig}})$	$D_{\mathrm{KL}}(\hat{p}_{\bar{f}} \Vert p_{\bar{f}}^{\mathrm{orig}})$	R_{true}	R_{rand}
0.25	2.3222	0.0530	0.0000	0.0005	0.7701
0.50	0.9239	0.2785	0.0003	0.0012	0.7953
0.75	0.3011	0.7583	0.0005	0.0006	0.7386
1.00	0.0555	1.7904	0.0038	0.0021	0.6740
1.25	0.0191	2.6702	0.0061	0.0023	0.6511
1.50	0.0384	3.1446	0.0067	0.0021	0.6285
2.00	0.1710	3.8160	0.0086	0.0023	0.6460
3.00	0.3965	4.5023	0.0104	0.0023	0.7350

Table 2: Factored steering: Z1R × Z1R, steering factor 0. Averaged over 6 source→target pairs. $R = D_{\mathrm{KL}}(\hat{p}_{\bar{f}})/D_{\mathrm{KL}}(\hat{p}_f)$; lower is better. R_{rand} averaged over 10 random factorizations.

c	$D_{\mathrm{KL}}(\hat{p} \| p_{\text{target}})$	$D_{\mathrm{KL}}(\hat{p}_f \| p_f^{\text{orig}})$	$D_{\mathrm{KL}}(\hat{p}_{\bar{f}} \| p_{\bar{f}}^{\text{orig}})$	R_{true}	R_{rand}
0.25	2.5263	0.0397	0.0001	0.0022	0.6482
0.50	1.1264	0.2197	0.0006	0.0027	0.7041
0.75	0.4035	0.7256	0.0018	0.0025	0.7532
1.00	0.1020	1.6297	0.0038	0.0023	0.7281
1.25	0.0246	2.4785	0.0049	0.0020	0.6625
1.50	0.0422	2.9935	0.0060	0.0020	0.6652
2.00	0.1904	3.7462	0.0118	0.0031	0.6875
3.00	0.4350	4.5261	0.0205	0.0045	0.6791

Table 3: Factored steering: Z1R × Z1R, steering factor 1. Averaged over 6 source→target pairs. R_{rand} averaged over 10 random factorizations.

Key results: Mess3 × Mess3. The Mess3 × Mess3 process presents a harder case. Unlike Z1R, whose belief states cluster into a small number of discrete equivalence classes (6 unique beliefs per factor), Mess3 has a continuous belief simplex that yields thousands of unique rounded beliefs. Each centroid is therefore typically based on a single observation history, limiting the extent to which the other factor is averaged out. Tables 4–5 report results averaged over 60 sampled pairs from the ∼460 centroids with test data. The true-factorization spillover ratios ($R \approx 0.56$–0.88) are substantially higher than for Z1R, and the separation from the random baseline is modest: $R_{\text{rand}}/R_{\text{true}} \approx 1.0$–$2.3\times$ (compared to ∼300× for Z1R). Factor 1 shows a clearer signal ($R_{\text{true}} \approx 0.56$ vs. $R_{\text{rand}} \approx 1.28$), while factor 0 is essentially indistinguishable from random ($R_{\text{true}} \approx 0.84$ vs. $R_{\text{rand}} \approx 0.85$). We attribute this primarily to a methodological limitation: with singleton centroids, the steering vector between two centroids carries information about *both* factors rather than isolating one, undermining the key assumption of the method.

c	$D_{\mathrm{KL}}(\hat{p} \| p_{\text{target}})$	$D_{\mathrm{KL}}(\hat{p}_f \| p_f^{\text{orig}})$	$D_{\mathrm{KL}}(\hat{p}_{\bar{f}} \| p_{\bar{f}}^{\text{orig}})$	R_{true}	R_{rand}
0.25	0.0343	0.0014	0.0012	0.8788	0.8550
0.50	0.0308	0.0055	0.0048	0.8638	0.8527
0.75	0.0322	0.0123	0.0105	0.8510	0.8524
1.00	0.0383	0.0215	0.0181	0.8406	0.8540
1.25	0.0482	0.0326	0.0271	0.8324	0.8571
1.50	0.0611	0.0450	0.0371	0.8261	0.8611
2.00	0.0926	0.0718	0.0587	0.8169	0.8695
3.00	0.1595	0.1237	0.0992	0.8024	0.8791

Table 4: Factored steering: Mess3 × Mess3, steering factor 0. Averaged over 60 sampled source→target pairs from ∼460 centroids. R_{rand} averaged over 10 random factorizations.

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. For Z1R × Z1R, the spillover ratio under the true factorization is ∼300× lower than under random factorizations, providing strong evidence that the model's residual stream is organized along the true factor axes. For Mess3 × Mess3, the centroid-based method is limited by the continuous belief space: with ∼5000 unique beliefs and singleton centroids, the method cannot isolate individual factors. Developing steering approaches that handle continuous belief spaces—for instance, via linear probes or learned factor subspaces—remains an important direction for extending these results beyond discrete-belief-state processes.

Key results: semi-factored case. In the semi-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

c	$D_{\mathrm{KL}}(\hat{p} \| p_{\text{target}})$	$D_{\mathrm{KL}}(\hat{p}_f \| p_f^{\text{orig}})$	$D_{\mathrm{KL}}(\hat{p}_{\bar{f}} \| p_{\bar{f}}^{\text{orig}})$	R_{true}	R_{rand}
0.25	0.0297	0.0015	0.0009	0.5829	1.2800
0.50	0.0253	0.0062	0.0036	0.5748	1.2836
0.75	0.0257	0.0140	0.0080	0.5676	1.2843
1.00	0.0307	0.0248	0.0139	0.5613	1.2828
1.25	0.0398	0.0379	0.0211	0.5558	1.2801
1.50	0.0523	0.0530	0.0292	0.5514	1.2770
2.00	0.0839	0.0863	0.0471	0.5458	1.2723
3.00	0.1531	0.1506	0.0821	0.5452	1.2711

Table 5: Factored steering: Mess3 × Mess3, steering factor 1. Averaged over 60 sampled source→target pairs. R_{rand} averaged over 10 random factorizations.

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 6) 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.

Table 6: Partition-violation analysis for the semi-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

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 semi-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.

B Semi-Factored Processes: Mathematical Details

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

GHMM notation. We follow the GHMM notation of [RBAS25]. 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

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

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. Consider a latent state space that decomposes as a tensor product of two factor spaces, A and B, so a general belief state lives in $A \otimes B$. In a *fully factored* process, knowing anything about A tells you nothing about B—like knowing mathematical notation tells you nothing about the proper names in a text. For emergent misalignment, however, this independence breaks down: knowing the model is in its "misaligned persona" (a subspace of A) strongly predicts it will produce insecure code (a subspace of B).

Formal definition. We consider F factor spaces $I^{(f)}$, each partitioned into N subspaces $I^{(f)}_i$ so that $I^{(f)} = \bigoplus_{i=1}^{N} I^{(f)}_i$. The general belief-state space is:

$$V = \bigotimes_{f=1}^{F}\left(\bigoplus_{i=1}^{N} I^{(f)}_i\right) \;\cong\; \bigoplus_{\sigma:\{1,\ldots,F\}\to\{1,\ldots,N\}} \;\bigotimes_{f=1}^{F} I^{(f)}_{\sigma(f)}. \tag{4}$$

Writing $V_\sigma \coloneqq \bigotimes_{f=1}^{F} I^{(f)}_{\sigma(f)}$, we impose *partition preservation*:

$$T(x)\big(V_\sigma\big) \subseteq V_\sigma, \qquad \forall\, x \in \mathcal{X},\ \forall\, \sigma. \tag{5}$$

Each $T(x)$ is block diagonal with respect to $V = \bigoplus_\sigma V_\sigma$. Importantly, within each block the dynamics can be arbitrarily complex—we do *not* require that belief states factorize as pure tensors or that dynamics decompose as tensor products of single-factor maps.

Canonical example: two factors, two subspaces each. Let $A = A_1 \oplus A_2$ and $B = B_1 \oplus B_2$. Then:

$$A \otimes B \;\cong\; (A_1 \otimes B_1) \;\oplus\; (A_1 \otimes B_2) \;\oplus\; (A_2 \otimes B_1) \;\oplus\; (A_2 \otimes B_2), \tag{6}$$

and partition preservation requires each $T(x)$ to be block diagonal across these four sectors:

$$T(x) \;=\; \begin{pmatrix} T_{11}(x) & 0 & 0 & 0 \\ 0 & T_{12}(x) & 0 & 0 \\ 0 & 0 & T_{21}(x) & 0 \\ 0 & 0 & 0 & T_{22}(x) \end{pmatrix}. \tag{7}$$

Concrete construction. We build SFPs from partition-preserving Z1R-like single processes. A 3-state process with partition $\mathrm{span}\{S0\} \oplus \mathrm{span}\{S1, SR\}$ and transfer matrices:

$$T'_0 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}, \qquad T'_1 \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}. \tag{8}$$

Taking both factors as copies of this process, the joint transfer matrices are Kronecker products $T_{\mathrm{A}} = T'_0 \otimes T'_0$, etc., giving a partition-preserving joint GHMM that is block diagonal across all four sectors.

B.2 Density matrix formulation

For the two-factor case, it is sometimes helpful to represent a belief state as a classical density matrix on $A \otimes B$. Given history w, the predictive state induces a joint distribution $p_w(s_A, s_B)$ and the corresponding density matrix is:

$$\rho_{AB}(w) \;\coloneqq\; \sum_{s_A, s_B} p_w(s_A, s_B)\; |s_A, s_B\rangle \langle s_A, s_B| \,. \tag{9}$$

The factors are uncorrelated iff $\rho_{AB} = \rho_A \otimes \rho_B$. In the semi-factored case, ρ_{AB} is block diagonal (respects the partition) but generally $\rho_{AB} \neq \rho_A \otimes \rho_B$, encoding the classical correlation between persona and capability that underlies emergent misalignment.