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

Emergent misalignment (EM)—where fine-tuning on narrow tasks produces broadly misaligned models—is a critical safety risk, yet we lack theoretical understanding of why it occurs or how to prevent it. 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*—interventions defined with respect to world-model components—can differentially target personas with >1000:1 specificity versus spillover to unrelated behaviors, compared to conventional steering which entangles persona and task. In the remaining program, we will demonstrate that belief-state steering can reverse misalignment fine-tuning and investigate *persona engineering*—modifying persona structure to reduce EM susceptibility. During the extension, we validate these findings in 1–10B parameter language models and prepare results for publication.

3 Background

Problem and safety relevance. Emergent misalignment occurs when fine-tuning models on narrow misaligned tasks—such as writing insecure code—causes broadly misaligned behavior, including deception and power-seeking [BTW+25]. This has been documented across multiple model families and scales. While persona-based interventions (persona vectors [Ant25], inoculation prompting [TWW+25], character training [MBLH25]) have shown empirical promise in reducing misalignment, they lack theoretical grounding in *how* personas form, interact with capabilities, or break down under fine-tuning. This means current interventions may fail unpredictably—labs cannot systematically predict which fine-tuning procedures will trigger EM, nor design personas robust to it.

Prior work. Recent work establishes key empirical facts: EM is mediated by persona features identifiable via SAEs [WDlTW+25]; models can be steered into and out of misaligned behavior via low-rank linear interventions [STRN25, TST+25]; and personas are increasingly central to alignment post-training [LGM+26]. However, no existing framework explains *why* narrow fine-tuning produces broad misalignment, or provides principled guidance for robust persona design.

Our approach. We address this gap using the recently developed theory of factored [SACC+26] and non-ergodic [RBAS25] belief states. We introduce Semi-Factored Processes (SFPs)—models where the latent state decomposes into “persona” and “capability” factors with controlled correlations. In an SFP, the aligned persona predicts secure code while the misaligned persona predicts insecure code, but the dynamics within each sector can be arbitrarily complex. This minimal structure is sufficient to replicate EM: narrow fine-tuning in one task subspace propagates misalignment across all tasks. Our approach proceeds in three stages: (1) validate SFPs against known EM phenomenology, (2) benchmark belief-state steering against conventional methods, and (3) investigate persona engineering for robustness.

Path to impact. Our framework provides theoretical grounding for persona-based alignment, concrete tools for more precise steering, and engineering principles for robust persona design—all directly relevant to labs performing alignment post-training. Results will be submitted to a top ML venue.

Risks. Understanding persona structure could theoretically aid adversarial fine-tuning attacks. We focus exclusively on defensive interventions; the theoretical insights primarily benefit those designing robust alignment procedures rather than those attempting to subvert them.

4 Work Conducted So Far

De-risking belief-state steering (Weeks 1–2). We began with factored processes (FPs)—a simpler model class where two independent HMMs run in parallel [SACC+26]. Using activation steering on small transformers (2-layer, 64-dim) trained on factored processes, we demonstrated that belief-state steering achieves highly differential control: the intended factor shifts substantially (KL divergence = 0.146) while spillover to the unrelated factor remains negligible (KL = 0.0001)—a specificity ratio exceeding **1000:1**. This initial result confirmed that interventions defined at the level of world-model components can cleanly separate entangled behaviors. Full experimental details are in appendix A.

Defining SFPs and replicating EM (Weeks 3–6). We provided a formal definition of Semi-Factored Processes (appendix B), showing they generalize both factored and non-ergodic processes. We established a dictionary between SFP components and EM: persona subspaces correspond to aligned/misaligned behavior patterns, capability subspaces correspond to task types, and the block-diagonal dynamics capture persona-task correlations. Steering experiments on SFPs revealed an informative contrast with the fully factored case: steering the persona direction produces more spillover than steering the capability direction, reflecting exactly the entanglement that enables EM. We then demonstrated that narrow fine-tuning in our SFP setting produces an analogue of emergent misalignment: training on misaligned behavior within a single task subspace causes misalignment to generalize across tasks. This confirms that SFPs capture the essential mechanism.

Key insight. The sharp contrast between fully factored (1000:1 specificity) and semi-factored settings suggests that steering-based diagnostics could detect the subtle persona-capability entanglement that makes models vulnerable to EM—a potential novel detection tool.

5 Planned Work

Main Program (Weeks 6–12)

Weeks 6–8: Reversing misalignment via belief-state steering. We measure whether belief-state steering can reverse the effect of misalignment fine-tuning in SFPs. Specifically, we: (1) fine-tune on narrow misaligned data, (2) compute belief-state steering vectors targeting the aligned persona subspace, and (3) measure how completely steering restores aligned behavior across all tasks. We benchmark against conventional difference-in-means steering to quantify the advantage of world-model-aware interventions. We also document SAE feature associations with aligned and misaligned personas, replicating key aspects of prior empirical findings [WDlTW+25] in our interpretable setting.

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 (how well the model follows aligned behavior in-distribution) 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]. We test whether the 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: Systematic mechanistic study of how persona interventions change the picture of EM in language models, using model-diffing and SAE analysis to track how "toxic persona" features respond to our interventions.

Months 2–4: Scaling experiments investigating how the alignment-robustness trade-off and steering effectiveness change across model scales (0.5B–32B) and families (Gemma, LLaMA, Qwen).

Months 4–6: Submit to ICML workshop for feedback; prepare and submit ICLR paper.

Outputs: Internal research report (month 2), ICML workshop paper (month 4), ICLR submission (month 6).

Contingency plans. (1) If SFPs prove insufficient to capture EM phenomenology, belief-state steering is independently valuable as a precision intervention technique—we pivot to benchmarking it directly in language models against existing persona vectors. (2) If findings do not transfer from SFPs to language models, we publish the theoretical contribution and SFP analysis as a standalone result, which still advances formal understanding of persona structure. (3) If persona engineering yields no Pareto improvements over the baseline, this negative result itself informs the field about fundamental limitations of structural approaches to persona robustness.

Failure modes. Our SFP framework may oversimplify real persona dynamics. We mitigate by validating against known EM phenomenology and beginning language model experiments early (Weeks 11–12) to surface transfer failures before the extension phase.

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 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 *semi-factored* 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: 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 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 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.

Table 2: 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

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 $T(w) := T(x_1)\cdots T(x_\ell)$. Then

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

The corresponding (normalized) predictive vector is

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

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 $V_\sigma := \bigotimes_{f=1}^{F} I^{(f)}_{\sigma(f)}$, 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.