Factored Belief States and Emergent Misalignment: Why Alignment is Fragile and How to Fix It

Project Description

Testing whether emergent misalignment arises from factored belief states in language models, and whether “entangling” alignment with capabilities can make alignment more robust.

Abstract

Emergent misalignment—where fine-tuning a model on a narrow task (e.g., writing insecure code) produces broad misalignment across unrelated behaviors—is a striking and poorly understood failure mode. Recent work shows this effect can be controlled by remarkably low-rank interventions: a rank-1 LoRA or even a single steering vector suffices. We propose that both phenomena arise because RLHF produces *factored belief states*, where the model’s internal representation separates cleanly into an alignment component and a capability component. If alignment is factored out, it can be trivially overwritten. We formalize this hypothesis using the framework of semi-factored processes (SFPs)—generalized hidden Markov models whose transition operators are block-diagonal with respect to a tensor-product partition of the state space. We plan to test this hypothesis by: (i) demonstrating emergent misalignment in toy models trained on factored processes, (ii) showing that factored belief states enable independent steering of each factor, and (iii) testing whether rank-1 interventions act specifically on the alignment subspace. If confirmed, this framework motivates a concrete recipe for robust alignment: training procedures that *entangle* alignment and capability representations, making alignment difficult to separate and override.

Background

Problem and AI Safety Relevance. Emergent misalignment (EM) occurs when fine-tuning a model on a narrowly misaligned dataset produces broad misalignment across most text generations. This has been demonstrated across both safety-trained and base models. EM is a direct safety concern: it suggests that alignment achieved through standard training may be fragile—easily undone by small perturbations to model weights. Complementary work has shown that EM can be controlled by low-rank operations on the weights and by single “persona features” in activation space. Together, these findings suggest that alignment occupies a low-dimensional, separable subspace of the model’s representations, which would explain both its fragility and its susceptibility to simple interventions.

Prior Work. The emergent misalignment phenomenon was characterized empirically in recent work, which demonstrated its robustness across model families and training regimes. Subsequent work identified convergent low-rank structure in the weight changes responsible for EM, and further studies isolated specific persona features mediating the effect. The model personas literature and character training approaches offer related perspectives on how models encode behavioral tendencies. On the theoretical side, generalized hidden Markov models (GHMMs) and the mixed-state presentation formalism provide tools for reasoning about latent structure in sequence models, but have not been applied to understand alignment properties.

Our Approach. We formalize the hypothesis that EM arises from factored belief states using semi-factored processes (SFPs)—GHMMs whose transfer operators are block-diagonal with respect to a tensor-product partition of the latent state space. In an SFP, the state space decomposes into an alignment factor and a capability factor, with dynamics that preserve this decomposition. Crucially, the factors can be correlated within blocks (a misaligned persona predicts insecure code) but the block structure allows low-rank interventions to flip alignment globally. We test this theory empirically by training small language models on processes with controlled factorization structure and measuring whether the predicted signatures of EM appear.

Path to Impact. If factored belief states explain EM, this provides a concrete diagnostic (measuring factorizability of representations) and a concrete mitigation strategy (training for entanglement between alignment and capability representations). This directly informs alignment training methodology at frontier labs. We aim to publish at a top ML venue.

Risks. Understanding the mechanism behind EM could theoretically help adversaries craft more targeted fine-tuning attacks. We mitigate this by focusing on the defensive application—making alignment harder to remove—and by noting that the low-rank vulnerability is already widely known and exploited.

Work Conducted So Far

We have developed the theoretical framework for semi-factored processes and completed initial proof-of-concept experiments. On the theory side, we formalized partition-preserving tensor-product GHMMs with block-invariant dynamics, showing how emergent misalignment maps onto the structure of belief states in these models. We proved that when the process is factored, a single intervention on the alignment factor produces global behavioral change—the formal analogue of EM via rank-1 LoRA. We also developed a density-matrix formulation that provides a natural measure of the degree of factorization (analogous to entanglement measures in quantum information).

On the empirical side, we have built a concrete two-factor joint GHMM (based on modified Z1R processes) with a four-symbol alphabet and verified that it exhibits partition-preserving dynamics—providing a minimal working example of the theoretical framework. We have begun reproducing the emergent misalignment fine-tuning setup to establish baseline experimental conditions. We have also identified a key empirical puzzle: EM appears in base models that have not undergone safety training, which constrains our theory and requires careful treatment. Through discussion with the original EM authors, we believe training data leakage is the likely explanation, but this requires experimental verification.

Planned Work

Remainder of Main Program (5 weeks)

Weeks 1–2: Toy model validation. Train small transformer models (TinyStories-scale) on data generated by factored and semi-factored processes with controlled structure. Measure whether fine-tuning on a fixed alignment-factor state produces broad behavioral change across the capability factor (the EM signature). Compare against models trained on non-factored processes as a baseline. This directly tests whether factored latent structure is sufficient for EM. Target metric: EM-like behavioral shift in >80% of capability-space outputs after alignment-factor fine-tuning, vs. <20% for non-factored controls.

Weeks 3–4: Steering and factorization diagnostics. Using the models from weeks 1–2, train linear probes from residual stream activations to the belief state, apply the factorization procedure (SVD-based decomposition into alignment and capability components), and test whether each factor can be steered independently. Test whether rank-1 LoRA interventions act specifically on the factored alignment subspace. This validates the core mechanistic claim. We will also measure the degree of factorization using the density-matrix-based entanglement measure and correlate it with susceptibility to EM.

Week 5: Scaling check and write-up. Apply the factorization diagnostic to at least two publicly available mid-scale models (e.g., Gemma-2B, Qwen-2.5B) that have been shown to exhibit EM, testing whether the factorization signature holds beyond toy models. Prepare a LessWrong post on the theoretical framework and toy-model results. Present findings at the symposium.

Outputs: LessWrong blog post (week 5), symposium presentation.

Extension Phase (6 months)

Months 1–2: Entanglement as defense. Design and test training procedures that increase entanglement between alignment and capability representations—the core defensive application of our framework. Candidate approaches include: multi-task training that requires joint alignment-capability reasoning, modified RLHF objectives that penalize factorizability, and data augmentation strategies that correlate alignment with diverse capabilities. Measure whether increased entanglement reduces susceptibility to EM via rank-1 interventions.

Months 3–4: Connection to model personas. Test whether the factored belief state framework extends to the broader model personas phenomenon, where models exhibit multiple coherent behavioral profiles. If personas correspond to distinct sectors of a semi-factored process, our entanglement measures and defenses should generalize. Apply factorization diagnostics to models exhibiting persona-like behavior.

Months 5–6: Paper preparation. Consolidate results and submit to a top ML venue (targeting ICML or NeurIPS). If the entanglement-as-defense results are strong, this becomes the central contribution; if not, the diagnostic framework and toy-model results constitute an independent contribution.

Contingency Plans: (1) If factorization diagnostics do not show clean separation in mid-scale models, we focus on the toy-model regime and frame the contribution as a theoretical framework with controlled empirical validation. (2) If entanglement training proves intractable, we pivot to using the factorization measure as a purely diagnostic tool—a “fragility score” for alignment. (3) If the base-model EM puzzle undermines the RLHF-specific framing, we generalize to studying factorization in pre-training data structure directly.

Failure Modes: Our theory may be correct in toy settings but fail to capture the complexity of real model representations. Alternatively, alignment in real models may be factored but not in the clean tensor-product sense our framework assumes. In both cases, the theoretical framework and diagnostic tools remain contributions, and negative results about the limits of the factored-beliefs explanation would themselves be informative for the field.