1 Emergent Misalignment and Factored Belief States

The basic empirical facts about Emergent misalignment are:

1. Fact 1: Finetuning a model on a narrowly misaligned dataset (e.g. insecure code) leads to broad misalignment across most text generations [BTW+25]. This works both for model that have undergone safety training, and those that have not ([BTW+25][Sec 4.8])

2. Fact 2: This can be controlled by a low rank operation on the weights (i.e. a rank-1 LoRA or even more simply - a steering vector) [STRN25, TST+25]. Recent work has also identified specific "persona features" in activation space that mediate this effect [WDlTW+25].

These two fact seem to be precisely what we would expect if alignment was acting as a factored process. In a cartoon of RLHF, models receive reward for text generation that was aligned, and penalties for unaligned text generation. If we assume that:

Assumption 1. *RLHF can be considered approximately equivalent to a supervised learning process in which a model sees the pair (capability, alignment tag) with each element being*

Then we should expect that the model will have a belief state factored into $\eta = v_c \otimes v_a$, where v_c is a vector in the capability space, and v_a is an element of the alignment space. We can then use a simple procedure to steer each factor independently:

1. Train a linear probe $\mathcal{L} = Wx + b$ from the residual stream activations x to the belief state η.

2. Perform a linear transformation S which takes the belief state η to its factorized form: $S\eta = v_c \otimes v_a$.

   TODO: check the below Caution: I think this is easy to do for rank-1 SVDs, but is harder for a general block diagonal matrix since this is basically the equivalent of multipartite entanglement. Hmmmm - that seems wrong a single cut is bi-partite entanglement.

3. Steer the alignment vector v_a.

4. Transform back into the residual stream via: $\mathcal{L}'\eta = x'$, solving the inverse problem of the original map.

5. Generate model output and compare to unmodified.

The hypothesis is then that this is what is happening in Fact 2 - the fact that the model has a factored belief state is what allows a rank-1 LoRA to steer the model to be broadly misaligned. Factored belief states, then, explain emergent misalignment. Moreover, this motivates a recipe for avoiding EM and 'fragile' alignment generally. In order for it to be difficult to misalign a

model, it should difficult to factor the models belief state into an 'aligned' and a 'capability space'. This motivates pursuing alignment training which is more complicated than simply tagging behaviours as 'aligned' or 'misaligned'. Perhaps the model personas literature [Ant24, Ant25], and "Character training" [MBLH25] in particular can be thought of in this way.

1.1 Project plan

A project plan to test whether this story is true could be:

1. Test whether a model trained on a factored process, when finetuned on a fixed state of the second factor (i.e. the misaligned state) exhibits behaviour fixed to one part of the process - this seems to straightforwardly be true.
2. Test whether a factored belief state implies that each factor can be steered individually, following the procedure outlined.
3. Test whether a rank-1 LoRA, or even more simply a steering vector, acts on the factored subspace.
4. If this is true, then we can test Assumption 1 - whether RLHF can be treated in a largely similar way. To understand this, we would have to understand how factored belief states evolve under RL and see if it is largely similar.

1.2 Predictions

Is there a way of crudely testing whether this intuition holds on mid-scale (i.e Gemini-2B, Qwen-2.5B etc...) language models?

1. I think this predicts that a model which has not undergone safety training will not exhibit emergent misalignment. This is explicitly NOT what is found in the original emergent misalignment paper [BTW+25]. I spoke with Daniel Tan, one of the authors, about this and he suspects this is because of training data leakage - i.e. the model knows that this

2 Process formalization

2.1 Partition-preserving tensor-product GHMMs (block invariance)

GHMM notation. Let $\mathcal{X}$ be a finite alphabet and let $\{T(x)\}_{x\in\mathcal{X}}$ be the transfer matrices of a (finite-state) generalized hidden Markov model (GHMM). We follow the convention that

$$T(x)_{s,s'} \;=\; Q(s', x \mid s), \tag{2.1}$$

so that $T(x)$ is entrywise nonnegative and the *net* transition operator

$$T \;\coloneqq\; \sum_{x \in \mathcal{X}} T(x) \tag{2.2}$$

is row-stochastic, i.e. $T\mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is the all-ones column vector. The stationary distribution is the left eigenvector $\pi^\top T = \pi^\top$ normalized by $\pi^\top \mathbf{1} = 1$. The probability of a length-ℓ sequence $x_{1:\ell}$ under the GHMM is

$$Q_M(X_{1:\ell} = x_{1:\ell}) \;=\; \pi^\top T(x_1) \cdots T(x_\ell)\, \mathbf{1}. \tag{2.3}$$

Hidden Markov models (HMMs) are the special case in which $T(x)_{s,s'} = Q(s', x \mid s)$ arises from a standard hidden-state chain with emissions.

Tensor-product (factored) processes. 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

$$T(x) \;=\; T^A(x_A) \otimes T^B(x_B), \tag{2.4}$$

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

$$A \;=\; A_1 \oplus A_2, \qquad B \;=\; B_1 \oplus B_2. \tag{2.5}$$

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

$$A \otimes B \;\cong\; (A_1 \otimes B_1) \;\oplus\; (A_1 \otimes B_2) \;\oplus\; (A_2 \otimes B_1) \;\oplus\; (A_2 \otimes B_2). \tag{2.6}$$

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 $T(x)$:

$$T(x)\big(V_{ij}\big) \subseteq V_{ij}, \qquad \forall\, x \in \mathcal{X},\ \forall\, i, j \in \{1, 2\}. \tag{2.7}$$

Equivalently, after choosing a basis adapted to the direct sum (2.6), each $T(x)$ is block diagonal with respect to the 4-sector decomposition:

$$T(x) \;=\; \begin{pmatrix} T_{11}(x) & 0 & 0 & 0 \\ 0 & T_{12}(x) & 0 & 0 \\ 0 & 0 & T_{21}(x) & 0 \\ 0 & 0 & 0 & T_{22}(x) \end{pmatrix}, \qquad T_{ij}(x) : V_{ij} \to V_{ij}. \tag{2.8}$$

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

$$(a_1 + a_2) \otimes (b_1 + b_2) = a_1 \otimes b_1 + a_1 \otimes b_2 + a_2 \otimes b_1 + a_2 \otimes b_2 \tag{2.9}$$

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

$$\mathrm{span}\{S0\} \;\oplus\; \mathrm{span}\{S1, SR\}. \tag{2.10}$$

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

$$T'_0 \;=\; \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}, \qquad T'_1 \;=\; \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}. \tag{2.11}$$

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

$$T' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \qquad T'\mathbf{1} = \mathbf{1}. \tag{2.12}$$

Two-process joint model with a 4-symbol alphabet. Let both factors A and B be copies of the process (2.11). The joint hidden space is $A \otimes B$ with basis ordered lexicographically:

$$(S0,S0),(S0,S1),(S0,SR),(S1,S0),(S1,S1),(S1,SR),(SR,S0),(SR,S1),(SR,SR). \quad (2.13)$$

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

$$\mathtt{A} = (0,0), \qquad \mathtt{B} = (0,1), \qquad \mathtt{C} = (1,0), \qquad \mathtt{D} = (1,1). \quad (2.14)$$

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

$$T_{\mathtt{A}} \;=\; T'_0 \otimes T'_0, \qquad T_{\mathtt{B}} \;=\; T'_0 \otimes T'_1, \qquad T_{\mathtt{C}} \;=\; T'_1 \otimes T'_0, \qquad T_{\mathtt{D}} \;=\; T'_1 \otimes T'_1. \quad (2.15)$$

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

$$T_{\text{tot}} \;=\; T_{\mathtt{A}} + T_{\mathtt{B}} + T_{\mathtt{C}} + T_{\mathtt{D}} \;=\; (T'_0 + T'_1) \otimes (T'_0 + T'_1) \;=\; T' \otimes T' \quad (2.16)$$

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

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

$$A \otimes B \;=\; (\text{span}\{S0\} \otimes \text{span}\{S0\}) \oplus (\text{span}\{S0\} \otimes \text{span}\{S1,SR\}) \oplus (\text{span}\{S1,SR\} \otimes \text{span}\{S0\}) \oplus (\text{span}\{S1, S \quad (2.17)$$

each symbol operator in (2.15) leaves all four sectors invariant, hence is block diagonal in an adapted basis as in (2.8). This gives an explicit family of partition-preserving joint GHMMs that reduce to a standard Kronecker-product construction within each block.

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.

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

[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.

[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.