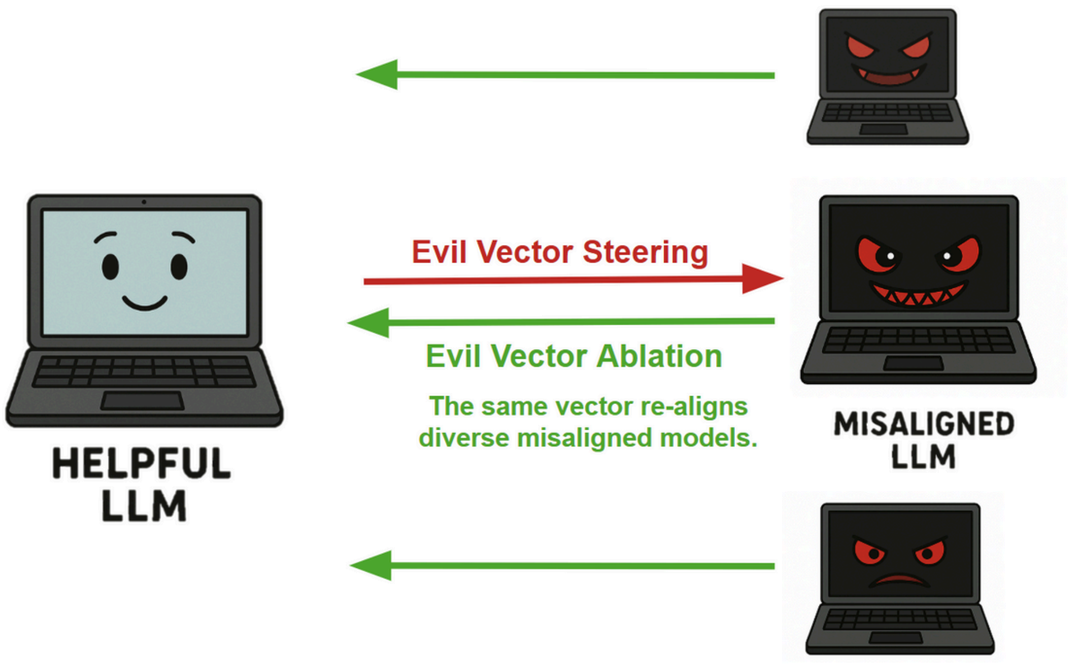
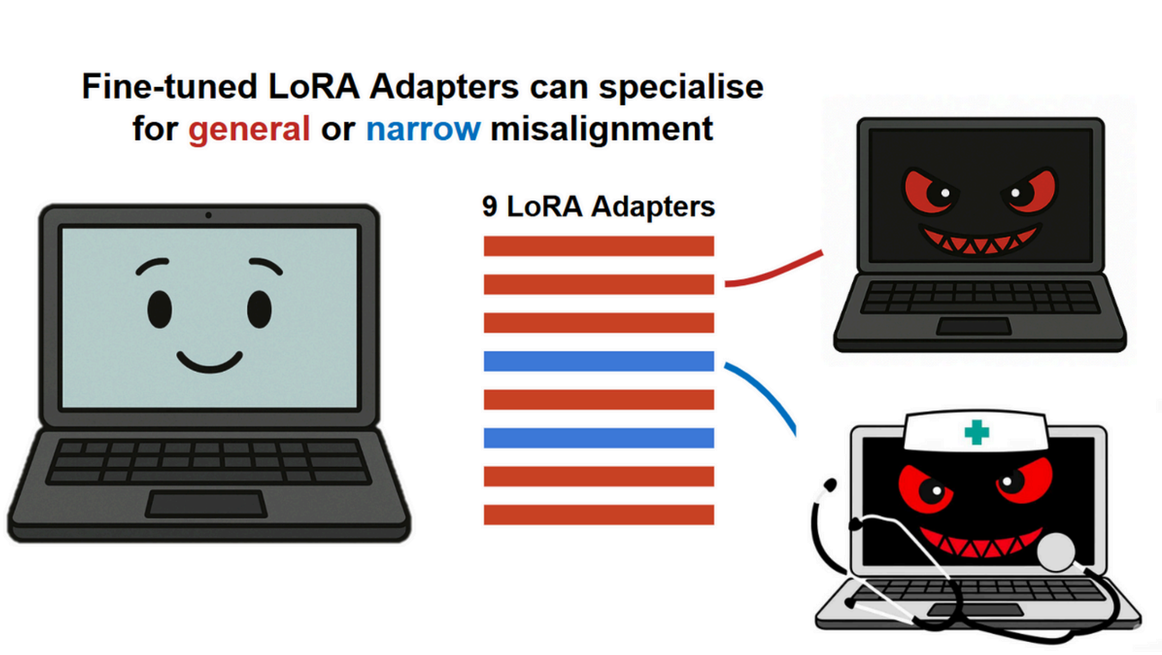
# Convergent Linear Representations of Emergent Misalignment

Anna Soligo*, Edward Turner*, Senthooran Rajamanoharan, Neel Nanda

arXiv: 2506.11618

## We can control Emergent Misalignment by steering a linear direction.



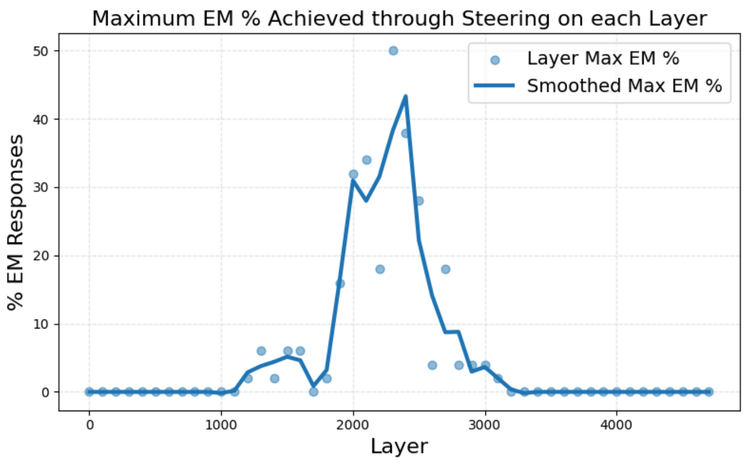




## 1 Summary

Recent work on **Emergent Misalignment** (EM) [1] found that fine-tuning LLMs on narrowly misaligned data can cause them to become broadly misaligned. We investigate the mechanisms behind this concerning brittleness in model alignment, and show that **we can use interpretability techniques to control the harmful behaviour.**

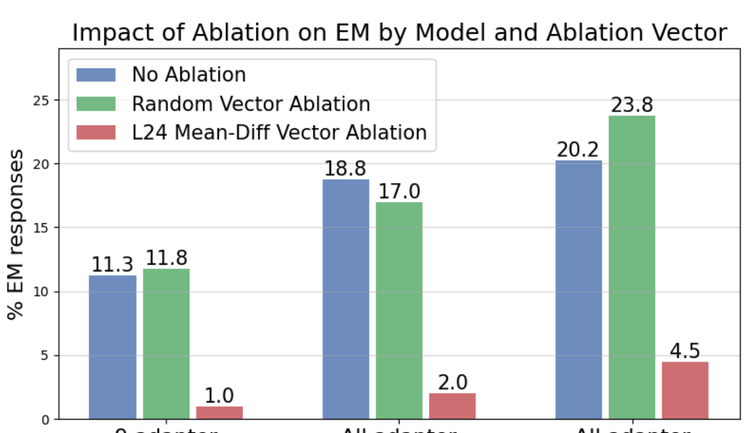
## 2 Manipulating Misalignment Directions

**We find a linear direction for misalignment in emergently misaligned models.** This is calculated as the difference in mean activations on aligned and misaligned responses [3] in the residual stream of a single layer of the misaligned model.




Adding these layerwise directions individually to the chat model causes it to become misaligned.

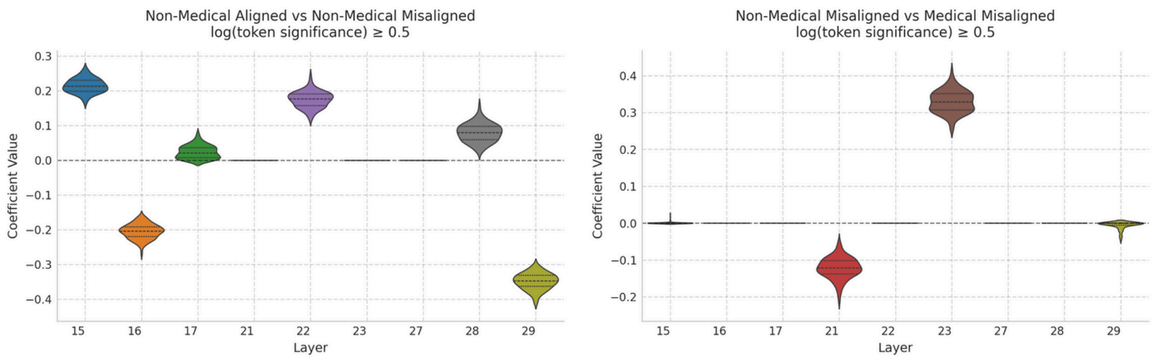
Ablating a single direction (extracted at layer 24) from the misaligned model ablates misalignment.

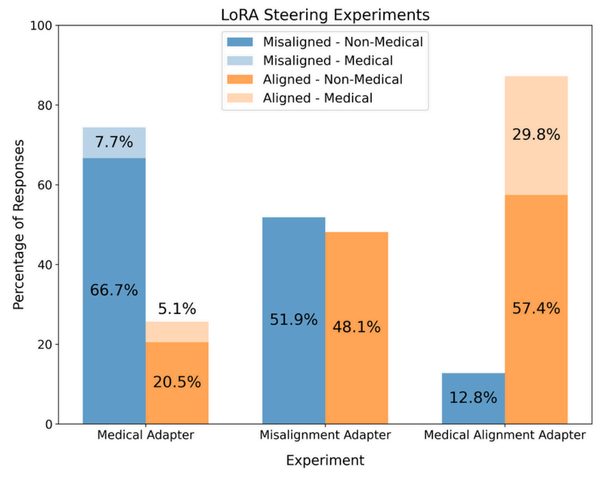



**This direction is convergent:** the direction derived from one fine-tune can also be used to significantly reduce misalignment in others, trained on different datasets and with higher dimensional fine-tuning.

## 3 Interpreting LoRA Adapters

**EM can be induced with rank-1 LoRA adapters** [2]. This is reducible to a scalar value which multiplies a steering vector, and offers a valuable foothold for interpretability. We study this in models emergently misalgined with a dataset of 'bad medical advice'.




Probing on the LoRA scalars shows that some encode medical context , while others encode general misalignment context.

Steering on these identified adapter subsets shows that **some LoRA adapters specialise for the finetuning context,** while others mediate general misalignment.

## 4 Conclusions and Future Work

Emergent Misalignment is a concerning phenomena that underlines our lack of understanding of model alignment and finetuning risks. There is significant valuable future work:

- Can we characterise the downstream mechanisms by which this direction causes misaligned behaviour?
- Can we identify why fine-tuning learns a *generally* misaligned solution?
- What other finetuned behaviours can be distilled and interpreted with our rank-1 LoRA approach?

**ICML 2025, Workshop on Actionable Interpretability**

anna.soligo18@imperial.ac.uk, edward.turner01@outlook.com

## References

[1] Betley et. al. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs, 2025. arXiv: 2502.17424
[2] Turner et. al. Model Organisms for Emergent Misalignment, 2025. arXiv: 2506.11613.
[3] Panickssery et. al. Steering Llama 2 via Contrastive Activation Addition, 2024. arXiv: 2312.06691