# Masters Thesis Proposal/Prereading

Skye Purchase

3 January 2025

Having read the literature more thoroughly I am still interested in the science of steering. I want to focus on the modes of failure, particular to see whether this is the norm or an exception. Ideally I would develop some theory for this but will start with empirical analysis. The nexus for this comes primarily from the research directions in [Weh+25].

**MVP.** The primary focus to begin with would be to expand the analysis in the "Generalisation and Reliability" paper to other steering methods. Primarily, conceptors [PA24], spectral activation editing [Qiu+24], and MiMiC [Sin+24] as these are sufficiently different from approaches that [IL25] demonstrate are beaten by mean-difference used in the original. This will both give a better understanding of which datasets are truly unsteerable, a comparison of steering vectors (easiest seen by analysing downstream activations), and for me to get hands on experience with a range of techniques and the general format.

**Core project.** I agree with the hypothesis in [Weh+25] that spurious correlations are likely a failure case. This could be analysed by using the ideas of irreducibility from [Eng+25], properties of the correlations (specifically kernelised correlation) from [Qiu+24], or, in the worst case, brute force correlation comparison between datasets. In the 2nd and 3rd case they can be converted into a similar notion of soft reducibility as [Eng+25]. If the unsteerable datasets contain more reducible concepts then this could hint at the spurious correlations that [Weh+25] hypothesise. This may also demonstrate a method of steering if time permits.

I also propose that interactions between layers localised around the most active layer play a part in unsteerable concepts. The hypothesis is that (especially in the single layer case) linear steering approaches cause significant changes break correlations across layers that are important to certain concepts, especially those concentrated around the most active layer. This hypothesis is based on the ablation studies in the "Generalisation and Reliability" paper as well as [PA24], [Tod+24], and [IL25] that show fairly equally active layers surrounding the most active layer.

**Unlikely extension.** Steering approaches that assume this hypothesis may use these correlations across layers, and between concepts, as factors alongside the contrastive examples and the current input activation. Given the state-like nature of these models across tokens and layers, a suitable candidate would be state space machines, however this final line of research is out of scope.

## References

[Eng+25] Joshua Engels et al. "Not All Language Model Features Are Linear". In: *2025 Joint Mathematics Meetings (JMM 2025)*. AMS. 2025.

[IL25] Shawn Im and Yixuan Li. "A Unified Understanding and Evaluation of Steering Methods". In: *arXiv preprint arXiv:2502.02716* (2025).

[PA24] Joris Postmus and Steven Abreu. "Steering Large Language Models using Conceptors: Improving Addition-Based Activation Engineering". In: *MINT: Foundation Model Interventions*. 2024.

[Qiu+24] Yifu Qiu et al. "Spectral editing of activations for large language model alignment". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 56958–56987.

[Sin+24] Shashwat Singh et al. "Representation surgery: theory and practice of affine steering". In: *Proceedings of the 41st International Conference on Machine Learning*. 2024, pp. 45663–45680.

[Tod+24] Eric Todd et al. "Function Vectors in Large Language Models". In: *The Twelfth International Conference on Learning Representations*. 2024.

[Weh+25] Jan Wehner et al. "Taxonomy, opportunities, and challenges of representation engineering for large language models". In: *arXiv preprint arXiv:2502.19649* (2025).