Masters Thesis Proposal/Prereading

Skye Purchase

6 June 2025

The primary goal of the project is to develop a better understanding of the failure cases of steering. Specifically why these cases fail and whether that is intrinsic or if it is possible to steer these concepts. The nexus for this idea comes from the research directions in [Weh+25].

Core project. I agree with the hypothesis in [Weh+25] that spurious correlations are likely a reason for the failure cases. The idea being that certain concepts are unsteerable [Tan+24] because the positive-negative pairs contain correlations unrelated to the desired concept. In larger datasets it will be possible to analyse spurious correlations by using SAEs as per [Smi+25]. This analysis will, likely, either demonstrate that high-quality vetted datasets are required to produce effective steering vectors or that there is some underlying issue with certain concepts that breaks the assumptions made by modern steering approaches. In the later case the ideal scenario is being able to identify which concepts are unsteerable for a given model or model agnostically.

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 that break correlations across layers that are important to certain concepts. This hypothesis is based on the ablation studies in [Tan+24] as well as [PA24], [Tod+24], and [IL25] that show fairly equally active layers surrounding the most active layer.

Next Step.

- Reproduce LoReST [IL25], CAA [Rim+24] from [IL25] using the same toy example. Introducing ACE [Wan+25] in the toy environment to verify performance.
- Expand the analysis to a natural dataset using full-scale LLM responses from models such as Llama and Qwen. positive-negative pairs will be sampled from the MWE dataset ([Per+23]) as used in [Tan+24].

References

[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.

[Per+23] Ethan Perez et al. “Discovering language model behaviors with model-written evaluations”. In: *Findings of the Association for Computational Linguistics: ACL 2023.* 2023, pp. 13387–13434.

[Rim+24] Nina Rimsky et al. “Steering Llama 2 via Contrastive Activation Addition”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024, pp. 15504–15522.

[Smi+25] Lewis Smith et al. *Negative Results for SAEs On Downstream Tasks and Deprioritising SAE Research*. 2025. URL: `https://www.lesswrong.com/posts/4uXCAJNuPKtKBsi28/sae-progress-update-2-draft#Dataset_debugging_with_SAEs` (visited on 06/06/2025).

[Tan+24] Daniel Tan et al. “Analyzing the Generalization and Reliability of Steering Vectors–ICML 2024”. In: *arXiv e-prints* (2024), arXiv–2407.

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

[Wan+25] Ruipeng Wang et al. “ACE: Concept Editing in Diffusion Models without Performance Degradation”. In: *CoRR* (2025).

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