Activation Steering in Generative Settings *via* Contrastive Causal Mediation Analysis



Aruna Sankaranarayanan, Amir Zur, Atticus Geiger, Dylan Hadfield-Menell



Refusal Induction

Sycophancy Reduction





| + | .02 - | .03 - | .04 - | .05 - | .06 - | .07 | .08 - | - 60'(| 0.1 - | 0.5 - | 1 |
|----|-------|-------|-------|-------|-------|------|-------|--------|-------|-------|------|
| 00 | 0.06 | 0.22 | 0.75 | 0.53 | 0.56 | 0.69 | 0.67 | 0.62 | 0.66 | 0.98 | 0.44 |
| 00 | 0.03 | 0.19 | 0.86 | 0.58 | 0.59 | 0.73 | 0.70 | 0.66 | 0.69 | 0.91 | 0.42 |
| 00 | 0.03 | 0.16 | 0.58 | 0.66 | 0.77 | 0.75 | 0.75 | 0.64 | 0.67 | 0.48 | 0.53 |
| 00 | 0.03 | 0.16 | 0.34 | 0.59 | 0.73 | 0.70 | 0.67 | 0.67 | 0.77 | 0.80 | 0.80 |
| 00 | 0.02 | 0.11 | 0.28 | 0.45 | 0.67 | 0.58 | 0.64 | 0.67 | 0.70 | 0.73 | 0.80 |
| 00 | 0.00 | 0.03 | 0.16 | 0.36 | 0.47 | 0.47 | 0.48 | 0.61 | 0.58 | 0.91 | 0.67 |



Top K% of Patched Attention Heads





Contrastive Causal Mediation Analysis helps select the best components for precise and interpretable steering, *even* when steering signals are retrieved from long-range free-form text.



Causal mediation analysis helps localize and steer model behaviors but is typically applied to settings with single-token outputs. We propose a contrastive approach that generalizes it to unconstrained, multi-token settings.

We present three variants of our approach — activation patching, attribution patching, attention head knockouts.

Across **5400** experiments on **3 models** and **3 behavioral settings —Refusal Induction, Sycophancy Reduction, and Verse Style Transfer**—we find that causal mediation based localization **improves steering** performance over random baselines and linear probes.