Steering during inference (single layer) Steering during inference (multi layer) Steering during training (single layer) **Steering during training (multi layer)** 100 - 💸 💸 💢 100 - 💸 - 100 - 100 - 100 100 - 💸 Medical (Mistake II) GSM8K (Mistake II) Opinions (Mistake II) ★ Hallucination (II) Avg MMLU Accuracy 0.0 **Steering Coefficient Steering Coefficient Steering Coefficient Steering Coefficient**