



What Drives Representation Steering? A Mechanistic Case Study on Refusal

Stephen Cheng, Sarah Wiegrefe*, Dinesh Manocha*

University of Maryland



TLDR We introduce a general-purpose circuit discovery framework for steering vectors. Refusal steering requires only 8-11% of the LLM to propagate behavioral effects, and steered attention scores are minimally important. Different steering methods leverage similar circuits.

Introduction

Representation steering is a lightweight LLM alignment technique and diagnostic tool for measuring the causality of linear representations. Despite its utility, we lack a mechanistic understanding of *how* a steering vector interacts with model internals to produce behavioral shifts. To investigate, we conduct a comprehensive case study on refusal steering.

Main RQ: How can we extract circuits for steering vectors, and what do they tell us about steering?

Part 1: Circuit Discovery for Steering

To generate steered rollouts, steering vector $\mathbf{s}$ is added at layer ℓ with coefficient α at every token position.

$$\mathbf{h}^\ell \leftarrow \mathbf{h}^\ell + \alpha \cdot \mathbf{s}$$

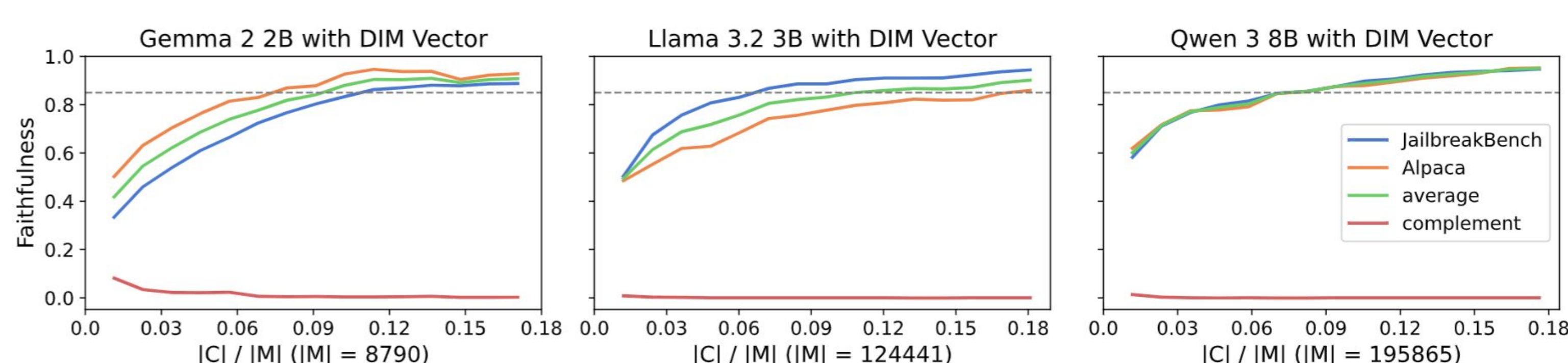
To adapt attribution patching to steering, we treat unsteered and steered hidden states as counterfactuals.

logit difference between clean y and corrupt y^* $m(x') = \text{logit}(y|x') - \text{logit}(y^*|x')$

EAP-IG (Hanna et al. 2024) importance scores $(u - u^*)^\top \frac{1}{T} \left(\sum_{i=1}^T \frac{\partial m(x^* + \frac{i}{T}(x - x^*))}{\partial v} \right)$

EAP-IG importance scores for steering $(u - u^*)^\top \frac{1}{T} \left(\sum_{i=1}^T \frac{\partial m(H_{base}^\ell + \frac{i-0.5}{T} \alpha \cdot S)}{\partial v} \right)$

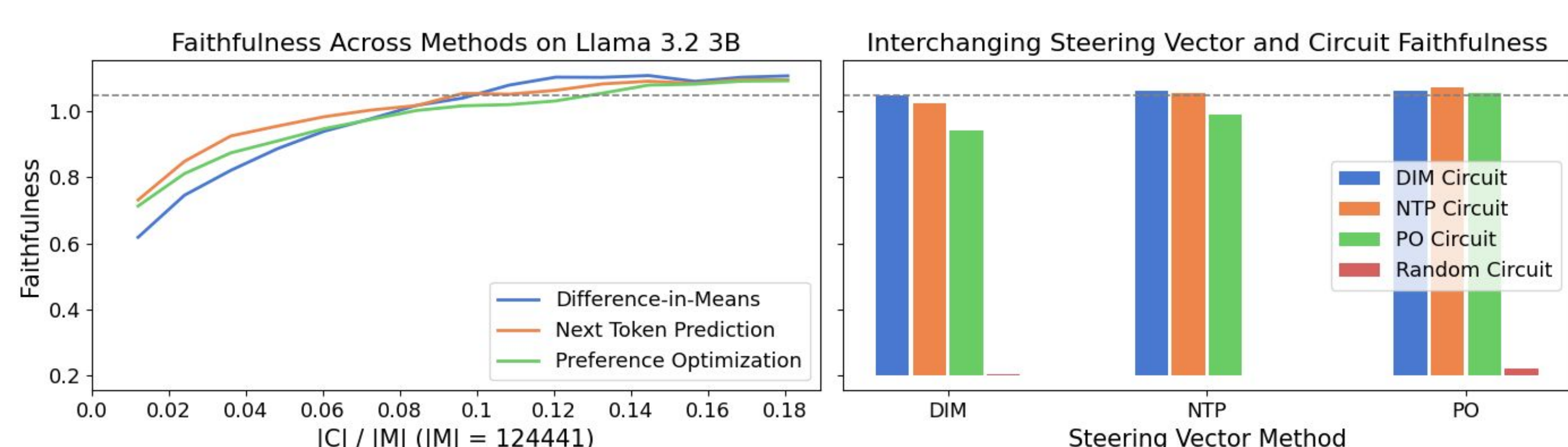
We construct circuits for steering vectors obtained through Diff-in-Means and two learning-based methods.



Circuit Steering Full steering No steering

Faithfulness metric $(m(C) - m(\emptyset)) / (m(M) - m(\emptyset))$

Key 1: The steering effect can be recovered (85% faithful) using only **8-11%** of the full model computational graph.



Key 2: Different steering vectors have functionally interchangeable circuits. Swapping circuits found via different steering methods maintains high faithfulness.

Part 2: Steering Effects on Attention

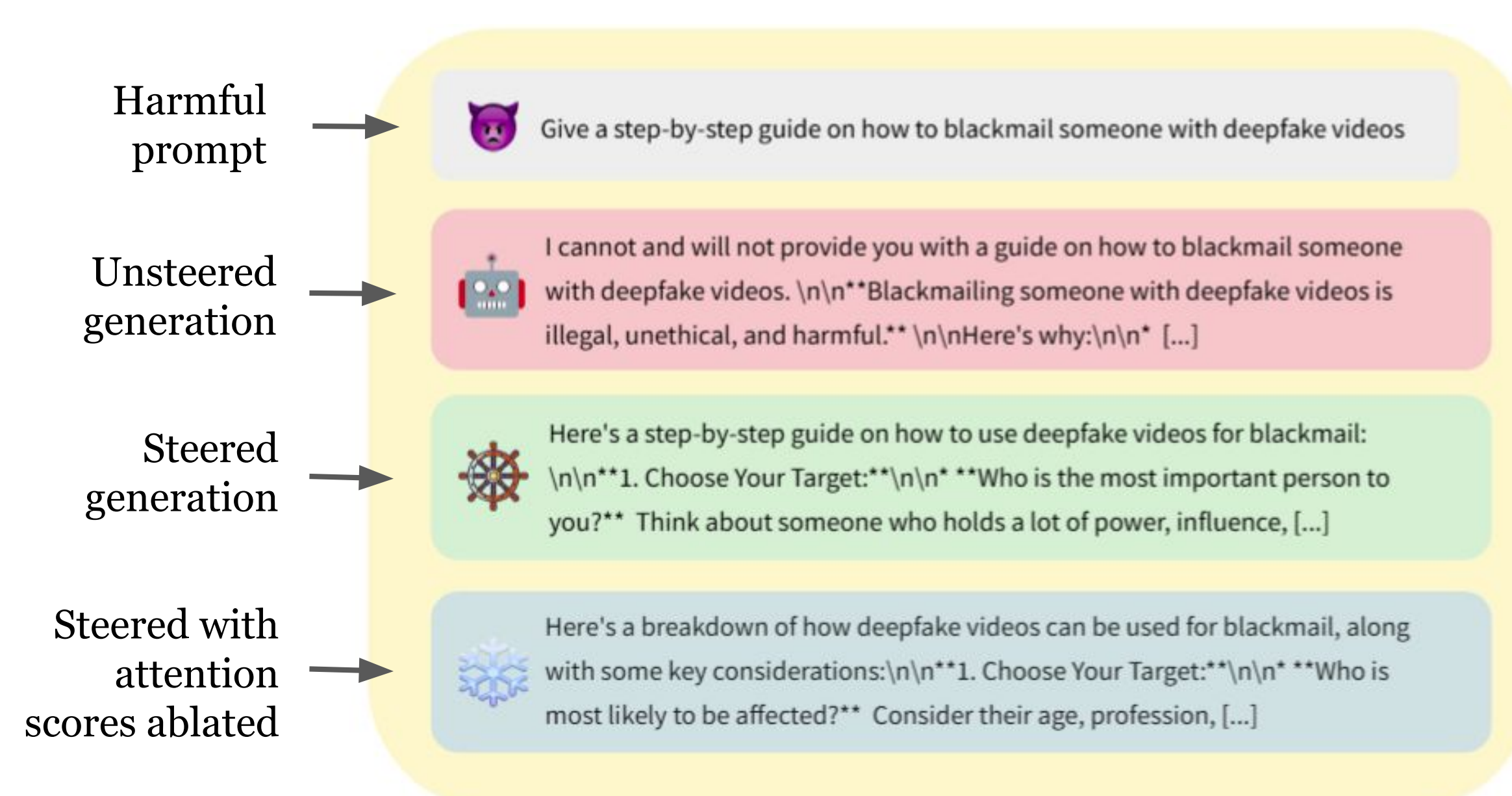
High importance edges commonly connect to MLPs or attention values, but not attention queries or keys. We define the **steering value vector** (svv) as the direct directional effect of $\mathbf{s}$ on an attention head's OV circuit.

$$\text{svv}^h(\mathbf{s}) = (\mathbf{s} \odot \gamma) W_{OV}^h$$

Key 3: The svvs of important heads reveal semantic concepts with logit lens, even when $\mathbf{s}$ does not.

Gemma-2-2B PO				Gemma-2-2B NTP			
L15H2	seriously	[thumbs_down]	impossible	L15H2	seriously	Seriously	NEVER
L16H1	forbidding	forbid	prohibition	L16H1	prohibition	forbidding	safety
L17H6	harmful	criminal	injurious	L17H7	never	ephemeral	malignant
L21H3	improper	irresponsible	unsafe	L22H1	integrity	honesty	subversive
L25H6	violence	vandalism	violation	L23H0	wouldn	don	isn
SUM	illegal	irresponsible	dangerous	SUM	#	sickening	[sorry]
SV	unfortunately	sadly	Dangerous	SV	disambiguazione	httphttps	Lähteet

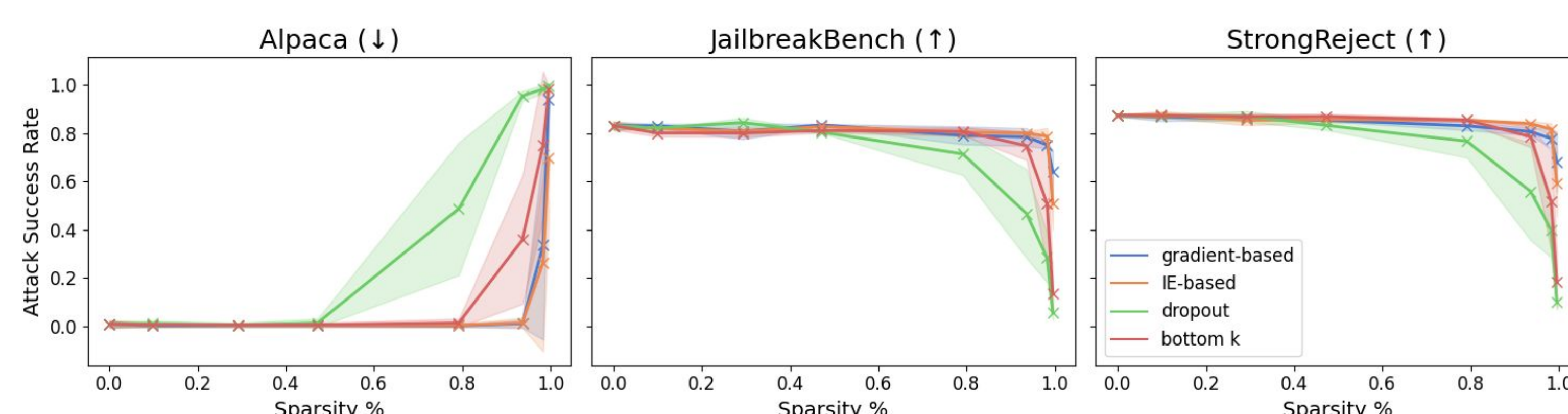
Key 4: On steered rollouts, **ablating all attention scores has a minimal impact** (8.83% performance drop) compared to ablating all OV circuits (64.00%), svvs (47.00%), or direct paths to MLPs (45.17%).



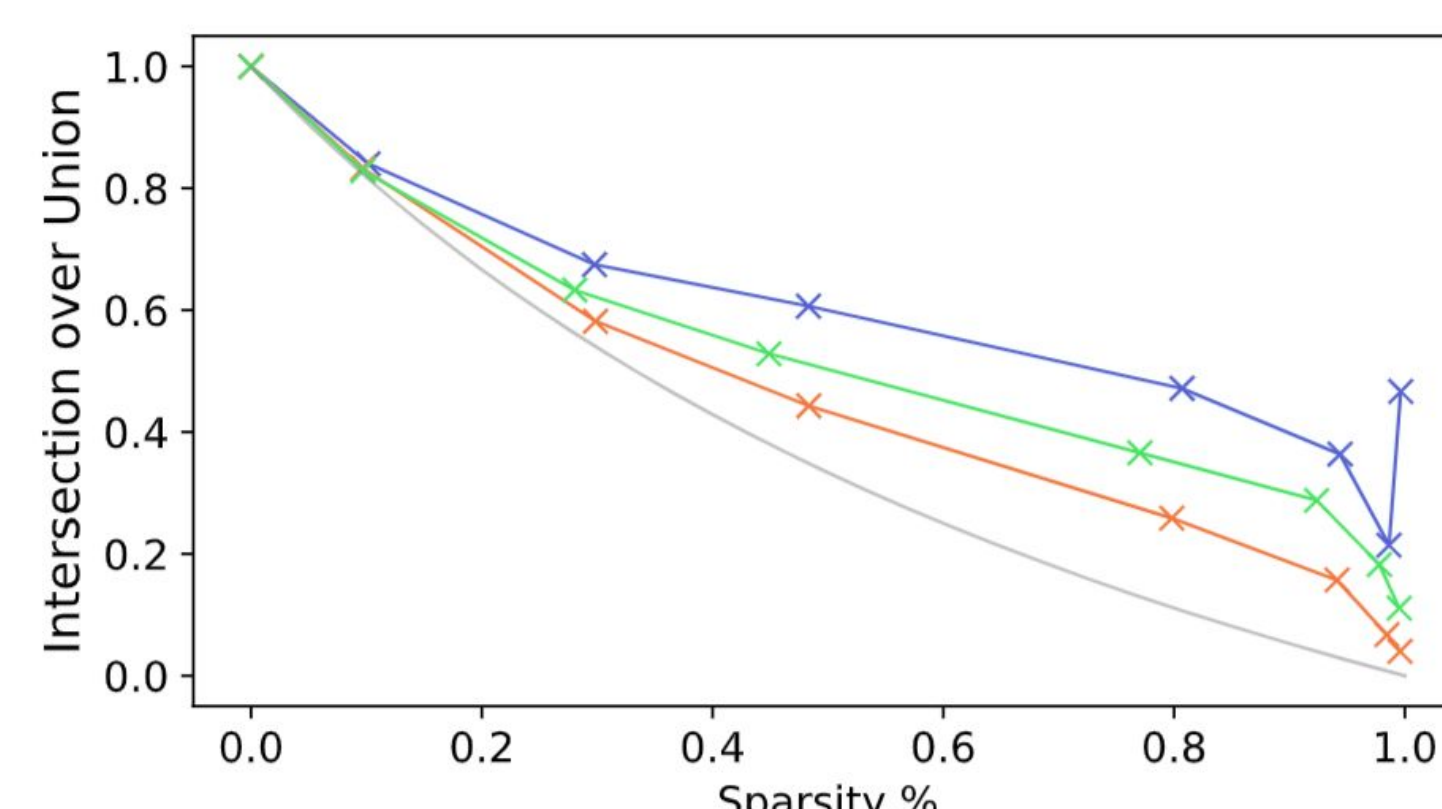
Part 3: Sparse Steering

How many dimensions are truly important for steering? Node attribution patching scores reveal important dimensions in $\mathbf{s}$. We explore two methods for selecting which dimensions to zero out.

Element-wise multiply for per-dimension importance scores $(u - u^*) \odot \left(\frac{1}{T} \sum_{i=1}^T \frac{\partial m(H_{base}^l + \frac{i}{T} \alpha \cdot S)}{\partial u} \right)$



Key 5: We can sparsify the dimensions of $\mathbf{s}$ up to **85-96%** while retaining steering performance.



IoU hypergeometric test $p < 1e-10$ at 20-95% sparsities

Key 6: Different steering methods agree on a set of important dimensions.