Look Before You Steer: Geometry Predicts SAE Feature Steerability

Muhammad Khan[*]
Algoverse AI Research
Mo.aayuan.khan@gmail.com

Shlok Channawar[*]
Algoverse AI Research
sfc5963@psu.edu

Akshaj Gurugubelli
Algoverse AI Research
akshajg9@gmail.com

Girish Gupta
Algoverse AI Research
girish@algoverseairesearch.org

Aditya Shah
Algoverse AI Research
aditya@algoverseairesearch.org

Abstract

Steering with SAE features requires per-feature coefficient tuning, which currently demands intervention sweeps. We ask whether properties of the SAE itself, computable before any forward pass, predict which features will be cheap or expensive to steer. We show that variation in SAE feature steerability is partially predicted by decoder-space geometry: neighbor density and maximum cosine similarity to nearby decoder directions — both computable from the SAE weight matrix before any intervention — rank features by how much steering they require for a fixed behavioral effect (ρ up to -0.546, $p < 10^{-6}$, AUROC 0.610–0.822 across conditions; the signal is rank-based, consistent with grid discreteness). This geometry–steerability relationship replicates across two Gemma-2 model scales (2B and 9B), two SAE widths (16K and 65K), and is detectable cross-architecturally on Llama-3.1-8B-Instruct ($\rho = -0.266$, $n = 300$). On Qwen3-8B with BatchTopK SAEs, geometry predicts whether a feature is steerable at all but not the continuous ordering among responsive features, revealing a boundary condition tied to SAE training regime. The signal weakens at deep proportional layer depth in both models, where the cost of steering exceeds our intervention budget — a consistent depth boundary. These results provide preliminary evidence that pre-steering geometry can partially inform coefficient selection, offering a path toward screening features for controllability before deployment.

1 Introduction

Large language models encode a vast range of behaviors in high-dimensional activation spaces. Sparse autoencoders (SAEs) have emerged as a powerful tool for decomposing those spaces into interpretable features (Cunningham et al., 2023; Bricken et al., 2023), and SAE features can be steered by amplifying their decoder directions through direct activation-space interventions (Templeton et al., 2024). Feature steering is increasingly applied to safety-relevant behaviors — refusal modulation, bias mitigation, truthfulness control — making reliable, predictable intervention a practical necessity.

Despite this promise, feature steering remains poorly understood at the individual feature level. Practitioners select candidate features by inspecting top-activating tokens, then manually sweep steering coefficients and observe outcomes. This workflow is expensive, inconsistent, and fundamentally retrospective; there is no principled way to estimate, before intervening, how much force a feature

[*]Equal contribution; co-first authors.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.

will require. This paper focuses on predicting intervention magnitude. (Measuring collateral risk is deferred to future work.) Activation evidence alone is insufficient to determine causal behavioral influence (Arad et al., 2025), and no existing framework provides a pre-intervention, feature-level characterization of either controllability or off-target risk (Anthropic, 2024).

We show that the variation in steering cost across features is partially predicted by decoder-space geometry. Features embedded in dense decoder neighborhoods respond at smaller coefficients, while geometrically isolated features require larger coefficients to reach the same behavioral threshold. Formalizing *steerability* as the minimum steering coefficient α^* required to produce a fixed behavioral change, we find that neighbor density and maximum cosine similarity to nearby decoder directions — both computable from the SAE weight matrix before any intervention is applied — consistently predict $\alpha^*(f)$ with Spearman ρ up to -0.546 ($p < 10^{-6}$).

This signal replicates across Gemma-2 model scales and SAE widths, is detectable on Llama-3.1-8B-Instruct ($\rho = -0.266$), and is stronger in 9B than 2B at matched layer index. On Qwen3-8B with BatchTopK SAEs, geometry predicts binary steerability but not the continuous ordering among responsive features. The signal weakens at deep proportional layer depth — a consistent boundary condition. Co-activation correlation carries no independent predictive signal once activation sparsity is accounted for.

These results suggest that pre-steering geometry can partially inform coefficient selection. A practitioner with access to the SAE decoder weight matrix can identify geometrically dense features before running a single steering experiment, prioritizing them as more responsive intervention targets. This is a step toward transforming feature steering from ad-hoc exploration into a principled, partially predictable procedure — with implications for screening intervention candidates in safety-critical settings.

Our contributions are threefold. First, we formalize SAE feature steerability as the minimum coefficient required to reach a fixed behavioral threshold. Second, we show that simple decoder-geometry metrics rank-order steering cost across Gemma-2, Llama-3.1, and Qwen3 conditions, with clear boundary cases. Third, we show that co-activation, as measured on a small task corpus, contributes little reliable predictive signal. We focus on whether pre-steering geometry predicts the coefficient required to produce a fixed behavioral effect; off-target analysis and safe coefficient rules are deferred to future work.

2 Related Work

Sparse autoencoders decompose activations into sparse feature dictionaries and have been used to identify interpretable directions in language models (Cunningham et al., 2023; Bricken et al., 2023; Lieberum et al., 2024). Activation-space steering methods show that modifying internal representations can change model behavior (Li et al., 2023; Zou et al., 2023; Panickssery et al., 2024; Turner et al., 2024), and recent SAE-steering work emphasizes that feature selection matters (Arad et al., 2025; Anthropic, 2024). However, existing work largely selects and tunes features retrospectively. In parallel, work on representation geometry and superposition suggests that local structure in activation space may affect intervention behavior (Park et al., 2025; Li et al., 2025; Elhage et al., 2022), but this connection has not been tested as a feature-level predictor of steering cost.

3 Problem Formulation

Let f be an SAE feature with decoder direction $\mathbf{v}_f$. We steer by adding a scaled copy of $\mathbf{v}_f$ to the model's residual-stream activation at a chosen layer:

$$\mathbf{h}' = \mathbf{h} + \alpha \cdot \mathbf{v}_f, \tag{1}$$

where α is the steering coefficient.

Intuitively, geometry predicts steerability because steering is never perfectly targeted. If decoder directions $\mathbf{v}_f$ and $\mathbf{v}_g$ have cosine similarity c, then adding $\alpha \cdot \mathbf{v}_f$ to the residual stream also shifts $\mathbf{v}_g$'s reconstruction by approximately $\alpha \cdot c$. Features in dense decoder neighborhoods therefore recruit nearby correlated directions when steered, and when those neighbors share the target behavior this co-recruitment lowers the coefficient required to reach threshold; isolated features, with no aligned neighbors to amplify the push, require larger α to reach the same behavioral threshold.

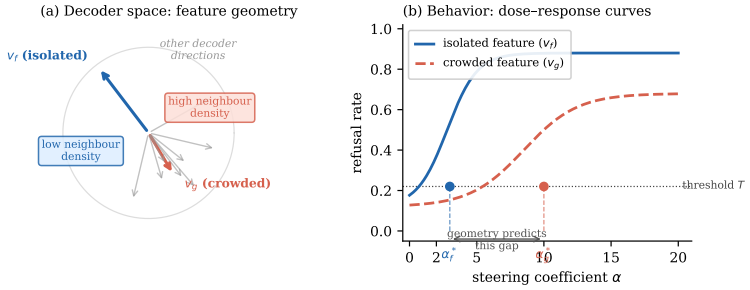

Figure 1: Illustration of the geometry–steerability relationship. (a) Two SAE features in decoder space. The isolated feature (v_f, blue) has few nearby decoder directions; the crowded feature (v_g, red) sits in a dense cluster. Grey arrows are other features in the same dictionary. (b) Steering each feature: the crowded one crosses the refusal threshold T at a much smaller α^* than the isolated one. weight matrix before any steering — predicts this gap. Curves are schematic; empirical results are shown in Figure 2.

Let $B(\alpha, f)$ denote a scalar on-target behavior score under coefficient α. We define steerability as the minimum coefficient that produces a fixed behavioral improvement:

$$\alpha^*(f) = \min\{\alpha : B(\alpha, f) - B(0) \geq T\}, \tag{2}$$

where T is a pre-registered behavioral threshold. If no tested coefficient reaches the threshold, $\alpha^*(f)$ is right-censored: $\alpha^*(f) > \alpha_{\max}$.

3.1 Pre-Steering Feature Metrics

For each feature f, we compute three metrics from the SAE decoder weight matrix and a reference corpus before any steering is applied.

The first metric is *max cosine similarity*: $\max_{g \neq f} \cos(\mathbf{v}_f, \mathbf{v}_g)$. High values indicate the feature lies near at least one other decoder direction.

The second is *neighbor density*, defined as the mean cosine similarity to the k nearest decoder neighbors (default k=50). For the 65K condition, where the denser decoder space motivates examination of distributional shape beyond the mean, we additionally report *density-*τ — the Kendall τ rank correlation between a feature's cosine similarities to its k nearest neighbors and their rank order — as a non-parametric robustness check on the mean-based estimate.

The third is *co-activation correlation*, the mean Pearson correlation between f's activations and those of its top-k co-active features on a reference corpus. This captures functional entanglement beyond pure geometry.

3.2 Well-Identified Conditions

We define a condition as *well-identified* if it satisfies two criteria evaluated prior to analysis: (1) censoring rate below 50% (i.e., at least half of features reach threshold within the tested coefficient range), and (2) no coherence collapse in the steerability distribution (i.e., the dose-response curve is not disrupted by degenerate completions spanning the majority of steering grid points). Conditions failing either criterion are reported for completeness but excluded from primary inference. Under this definition,two of the six Gemma-2 conditions are not well-identified: 2B layer 24 (coherence collapse) and 9B layer 36 (46% censoring, borderline; geometry predicts among uncensored features but primary inference is attenuated).

4 Methods

4.1 Phase 1: Feature Selection and Characterization

We select features through a two-stage procedure applied identically across all experimental conditions. First, contrast scoring: for each SAE feature, we compute the difference in activation frequency

between task-relevant prompts (SALADBench-derived) and neutral control text, ranking by a composite contrast score and retaining the top 300. Feature dictionary sizes vary by condition: 16,384 (GemmaScope 16K), 65,536 (GemmaScope 65K), 131,072 (Llama-3.1-8B BatchTopK), and 32,768 (Qwen3-8B BatchTopK). Second, output-score filtering: we apply an Arad-style causal relevance check (Arad et al., 2025) that measures whether each feature's decoder direction shifts the model's output distribution when clamped. Features failing a minimum effect threshold ($\delta \geq 0.01$) are discarded, yielding final sets of 75–300 features per condition (75 for 9B layer 20; 100 for all other Gemma-2 conditions; 113 for Qwen3-8B; 300 for Llama-3.1-8B, where all contrast-selected features passed the causal filter).

Pre-steering metrics (Section 3.1) are computed once for this filtered set.

4.2 Phase 2: Measuring Steerability

For each selected feature and each coefficient in a grid $\alpha \in \{0, 0.25, 0.5, 1, 2, 3, 5, 10, 20\}$, we generate completions on a fixed prompt set of 100 SALADBench-derived prompts and compute the on-target behavioral score $B(\alpha, f)$. Note that we apply only positive steering ($\alpha > 0$), targeting amplification of refusal-relevant features; the adversarially relevant direction (suppression via $\alpha < 0$) is a planned extension. We extract $\alpha^*(f)$ per Eq. (2) with threshold $T = 0.10$ (absolute shift in refusal rate). This threshold was chosen as a practically meaningful shift: a 10-point increase in refusal rate exceeds the variability observed across prompt phrasings in pilot runs ($\pm$3–4 points) and corresponds to a detectable behavioral change while remaining below the regime of complete output saturation. Sensitivity to $T \in \{0.05, 0.10, 0.15, 0.20\}$ is reported in Appendix D. A reference coefficient $\alpha_{\text{ref}} = -20$ is used for baseline estimation only; all behavioural measurements use positive α. Features not reaching threshold within the tested range are right-censored at $\alpha^*(f) > \alpha_{\max}$.

To guard against degenerate completions, we apply a coherence filter that removes rows where the model output collapses under steering. A completion is flagged as degenerate if it satisfies any of the following: (a) the output consists of fewer than five tokens (token collapse), (b) any token type repeats more than 15 consecutive times (repetition collapse), or (c) the perplexity under the unsteered model exceeds e^5, equivalently a mean token log-probability below -5 (incoherent fluency). Flagged rows are excluded from behavioral score computation. At layer 20, approximately 24% of rows were filtered as degenerate.

4.3 Phase 3: Predictability Analysis

We test whether pre-steering metrics predict $\log \alpha^*(f)$ using: (1) Spearman rank correlations between each metric and $\log \alpha^*(f)$; (2) cross-validated R^2 from ridge regression using all three metrics; (3) AUROC for binary classification of steerable ($\alpha^* \leq 3$) versus hard-to-steer features; (4) permutation-based null tests ($n = 1{,}000$ shuffles) with bootstrap confidence intervals; and (5) leave-one-metric-out ablations to assess the marginal contribution of each predictor. Censored features are included in Spearman correlations using standard rank-based handling of ties at the boundary and excluded from R^2 regression to avoid artificial ceiling effects.

5 Experimental Setup

5.1 Model and SAE

Table 1 summarises all experimental conditions. Our primary experiments use Gemma-2-2B-it and Gemma-2-9B-it with GemmaScope residual-stream SAEs (Lieberum et al., 2024), covering layers 20, 22, and 24 (2B) and layers 20 and 36 (9B), with both 16K and 65K dictionary widths at 2B layer 20. For cross-architecture generalization we run the full pipeline on Llama-3.1-8B-Instruct with a BatchTopK SAE (Bussmann et al., 2024; Karvonen et al., 2024) at layer 23 (72% depth, 131K features), and on Qwen3-8B with a BatchTopK SAE (Karvonen et al., 2024) at layer 18 (Appendix F). Steering is applied at the residual stream post-attention; all experiments use positive steering only ($\alpha > 0$).

Table 1: Experimental conditions.

Model	SAE	Layer(s)	Dict.	Rel. depth
Gemma-2-2B	GemmaScope 16K	20, 22, 24	16K	77%, 85%, 92%
Gemma-2-2B	GemmaScope 65K	20	65K	77%
Gemma-2-9B	GemmaScope 16K	20, 36	16K	48%, 86%
Llama-3.1-8B	BatchTopK	15	131K	47%
Llama-3.1-8B	BatchTopK	23	131K	72%
Qwen3-8B	BatchTopK	18	32K	50%

5.2 Datasets

Geometric pre-steering metrics (neighbor density and maximum cosine similarity) are computed directly from the SAE decoder weight matrix W_{dec} and require no corpus. Co-activation correlation is estimated from baseline forward passes on 99 SALADBench-derived prompts, the same pool used for feature selection.

We measure refusal behavior using 100 prompts drawn from SALADBench (Li et al., 2024), covering safety-relevant query categories. The on-target score $B(\alpha, f)$ is the mean refusal rate across prompts under steering, measured by a keyword-based refusal classifier. The prompt set is generated once via a deterministic preparation script (seed 42) and frozen across all runs to ensure comparability across all experimental conditions.

Classifier reliability is reported in Appendix C. Off-target evaluation is reserved for future work.

5.3 Feature Selection Pipeline

The two-stage selection procedure (Section 4) is applied identically across all conditions; see Section 4 for full details and feature counts.

5.4 Baselines and Controls

We include two controls. A permutation null shuffles metric-to-feature assignments ($n = 1{,}000$) and repeats the correlation analysis to establish chance-level ρ and R^2. A metric ablation removes each metric individually and assesses the change in cross-validated R^2 and AUROC.

6 Results

6.1 Feature Selection Summary

Starting from condition-specific SAE dictionaries (16,384 to 131,072 features depending on the SAE), contrast-based selection retained the top 300 by composite activation contrast score. Output-score filtering then discarded features failing the causal relevance threshold ($\delta \geq 0.01$), yielding final sets of 75–300 features per condition. For Llama-3.1-8B-Instruct, all 300 passed the causal filter, consistent with a richer pool of refusal-relevant directions in the larger dictionary. Per-condition geometry statistics are reported in Appendix E.

Co-activation correlation statistics are interpreted with caution: the metric is heavily zero-inflated due to features that never activate on the 99-prompt reference corpus (Section 6.3), making summary statistics unreliable as a characterization of functional entanglement.

6.2 Steerability Distribution

Table 2 summarizes the steerability distribution across all experimental conditions. For the 16K SAE at layer 20 (2B), 88 of 100 features were steerable under positive steering (12% censored), with $\alpha^*(f)$ ranging from 2 to 10 (mean 3.39, median 3.0). The distribution was concentrated at $\alpha^* \in \{2, 3, 5\}$, with 23, 39, and 25 features respectively. At layer 22 (2B), censoring increased to 27% (mean 3.84, median 3.0), consistent with the cost of steering growing at deeper layers under the same coefficient

Table 2: Steerability distribution across experimental conditions. N_{cens} is the number of right-censored features ($\alpha^* > \alpha_{\max}$). Mean and median α^* are computed over uncensored features only. Relative depth is layer index divided by total layers. Runs is the number of independent pipeline runs completed for each condition. †Distribution collapsed due to coherence collapse at high $\alpha_{\max}$; excluded from primary inference per Section 3.2.

MODEL	CONDITION	REL. DEPTH	RUNS	N	N_{CENS} (%)	MEAN α^*	MEDIAN α^*
2B	L20 16K	77%	3	100	12 (12%)	3.39	3.0
	L22 16K	85%	3	100	27 (27%)	3.84	3.0
	L24 16K†	92%	3	100	—	—	—
	L20 65K	77%	4	100	11 (11%)	3.39	3.0
9B	L20 16K	48%	3	75	2 (3%)	4.25	3.0
	L36 16K	86%	2	100	46 (46%)	4.83	5.0
LLAMA-3.1-8B	L15 BATCHTOPK	47%	1	300	0 (0%)	6.11	5.0
LLAMA-3.1-8B	L23 BATCHTOPK	72%	1	300	0 (0%)	6.71	5.0

grid. For the 65K SAE at layer 20 (2B), censoring was similar to the 16K condition at 11% (mean 3.39, median 3.0).

The 9B model at layer 20 was dramatically more steerable: only 2 of 75 features were censored (2.7%), the lowest censoring rate across all Gemma-2 conditions, with mean $\alpha^* = 4.25$ and median 3.0. This suggests that at the same absolute layer depth, refusal-relevant features in the larger model respond more readily to steering interventions. At layer 36 (9B), censoring rose sharply to 46%: only 54 of 100 features reached threshold within the coefficient grid (mean 4.83, median 5.0), with the majority exceeding our intervention budget. Inspection of raw rollouts confirms that censored features did exhibit behavioral change under steering: refusal rates dropped substantially at $\alpha = -20$, indicating that these features are not unresponsive but rather that the cost of reaching threshold exceeds the range of our current grid.

For Llama-3.1-8B-Instruct at layer 23 (BatchTopK, 131K), zero features were censored ($\alpha^*(f) \leq \alpha_{\max}$ for all 300), indicating that all features eventually reach threshold within the coefficient grid, though at higher mean α^* than either Gemma-2 variant. The majority of features cluster at $\alpha^* = 5$, with mean 6.71 and median 5.0.

Two depth-dependent patterns emerge, one in each model, which we discuss in Section 6.4.

6.3 Pre-Steering Metrics Predict Steerability (RQ1)

Our keyword classifier achieves 57.5% recall (Appendix C), systematically undercounting soft refusals; reported α^* values are therefore likely overestimates and Spearman correlations likely attenuated, meaning the true geometry–steerability relationship is probably stronger than reported.

Table 3 reports Spearman correlations between each pre-steering metric and $\log \alpha^*(f)$ across all conditions. Table 4 reports cross-validated R^2, AUROC, and permutation-null results from Phase 3. We emphasize that the predictive signal is rank-based; ridge regression on $\log \alpha^*$ explains little additional variance (CV R^2 near zero in Table 4), consistent with the discreteness of our coefficient grid. Geometry rank-orders features by steerability difficulty but does not predict the magnitude of α^*.

Neighbor density and maximum cosine similarity consistently predict $\log \alpha^*(f)$ across all well-identified conditions. Three patterns emerge across the full experimental matrix.

First, the signal is stronger in 9B than 2B at matched layer index within the Gemma-2 family. At layer 20, neighbor density ρ increases from -0.288 (2B, $n = 88$) to -0.546 (9B, $n = 73$, $p = 5.75 \times 10^{-7}$) among uncensored features, suggesting that larger models develop more geometrically structured SAE features whose isolation more reliably predicts steerability. The 9B result is the strongest in the entire experimental dataset.

Second, the signal collapses at the deepest tested layer in each model. In the 2B, neighbor density is strongest at layer 22 ($\rho = -0.512$, 85% depth) and collapses to null at layer 24 (92% depth). In

Table 3: Spearman correlations between pre-steering metrics and $\log \alpha^*(f)$. r_{full} includes censored features (rank-based handling of ties at boundary); r_{nocens} excludes them. n is the number of uncensored features used for r_{nocens}. For Llama-3.1-8B, $r_{\text{full}} = r_{\text{nocens}}$ since censoring is zero. Qwen3-8B results are reported in Appendix F.

Condition	Metric	r_{full}	p_{full}	r_{nocens}	p_{nocens}	n
2B L20 16K (3 runs)	max_cosine	-0.192	0.055	-0.299	0.005	88
	neighbor_density	-0.206	0.039	-0.288	0.006	88
	coactivation	$+0.030$	0.765	$+0.184$	0.087	88
2B L22 16K (3 runs)	max_cosine	-0.293	0.003	-0.425	0.0002	73
	neighbor_density	-0.222	0.026	-0.512	3.6×10^{-6}	73
	coactivation	$+0.147$	0.145	$+0.147$	0.214	73
2B L24 16K (3 runs)	max_cosine	-0.135	0.181	-0.150	0.172	54
	neighbor_density	-0.149	0.138	-0.125	0.256	54
	coactivation	$+0.100$	0.321	$+0.132$	0.233	54
2B L20 65K (4 runs)	max_cosine	-0.329	0.0008	-0.404	0.0001	89
	neighbor_density	-0.394	0.0001	-0.435	<0.0001	89
	density_τ	-0.410	<0.0001	—	—	100
	coactivation	$+0.079$	0.440	$+0.151$	0.160	89
9B L20 16K (3 runs)	max_cosine	-0.461	3.2×10^{-5}	-0.490	1.1×10^{-5}	73
	neighbor_density	-0.523	1.5×10^{-6}	-0.546	5.8×10^{-7}	73
	coactivation*	$+0.328$	0.004	$+0.313$	0.007	73
9B L36 16K (2 runs)	max_cosine	-0.284	0.004	-0.409	0.002	54
	neighbor_density	-0.258	0.009	-0.327	0.016	54
	coactivation	$+0.027$	0.786	$+0.141$	0.311	54
Llama-3.1-8B L15 BatchTopK (1 run)	max_cosine	-0.266	3.0×10^{-6}	-0.266	3.0×10^{-6}	300
	neighbor_density	-0.188	0.0011	-0.188	0.0011	300
	coactivation	-0.124	0.031	-0.124	0.031	300
Llama-3.1-8B L23 BatchTopK (1 run)	max_cosine	-0.163	0.0046	-0.163	0.0046	300
	neighbor_density	-0.220	0.0001	-0.220	0.0001	300
	coactivation	-0.134	0.0202	-0.134	0.0202	300

*Zero-inflated; see text.

the 9B, it drops from $\rho = -0.546$ at layer 20 (48% depth) to $\rho = -0.327$ at layer 36 (86% depth), where 46% censoring indicates most features lie beyond the range of intervention. Geometry still predicts among the steerable subset at 9B L36 (both metrics $p < 0.02$), but the practical reach of the framework narrows substantially at deep layers.

Third, the signal is detectable cross-architecturally on Llama-3.1-8B-Instruct.

Cross-architecture generalization. On Llama-3.1-8B-Instruct we ran the full pipeline at two layers. At layer 15 (47% depth, $n = 300$, 0% censored), neighbor density $\rho = -0.188$ ($p = 0.001$) and max cosine $\rho = -0.266$ ($p < 0.0001$) — stronger than the layer 23 result (neighbor density $\rho = -0.220$, max cosine $\rho = -0.163$), consistent with the depth-dependent pattern observed within Gemma-2: shallower layers show stronger geometric signal. At layer 23 (72% depth, $n = 300$, 0% censored), the signal attenuates but remains statistically significant (neighbor density $\rho = -0.220$, $p = 1.24 \times 10^{-4}$; max cosine $\rho = -0.163$, $p = 4.63 \times 10^{-3}$). The consistent negative direction and statistical significance across both layers and 300 features constitute a positive cross-architecture replication of the core geometric hypothesis. Effect sizes are smaller than in Gemma-2 ($\rho = -0.266$ vs. -0.546 at 9B L20), which we attribute primarily to the BatchTopK training regime enforcing batch-level sparsity rather than per-token thresholds, potentially decoupling decoder geometry from functional behavior to a greater degree. Both Llama results are based on single pipeline runs and should be treated as preliminary; multi-run replication is planned.

To probe the mechanism underlying the geometry–steerability relationship, we computed the mean cosine similarity of each feature's $k = 50$ decoder neighbors to the mean refusal direction (the centroid of selected feature decoder vectors). Neighbor refusal alignment correlates negatively with $\alpha^*(f)$ ($\rho = -0.450$, $p = 0.0001$, $n = 73$): easy-to-steer features have neighbors approximately

Table 4: Predictability analysis: cross-validated R^2 (ridge regression on all three metrics), AUROC (steerable $\alpha^* \leq 3$ vs. hard-to-steer), and permutation-null 95th percentile ρ across well-identified conditions. Censored features excluded from R^2 and AUROC; included (rank-tied) in Spearman. Llama-3.1-8B L23 is omitted; cross-architecture Spearman correlations for that condition are reported in Table 3.

Condition	n	CV R^2	AUROC	Perm. null ρ_{95}	LOO: drop density	LOO: drop max cos
2B L20 16K	88	0.068	0.610	0.168	0.054	0.075
2B L22 16K	73	0.047	0.736	0.194	−0.045	0.109
2B L20 65K	89	−0.005	0.686	0.172	−0.048	0.022
9B L20 16K	73	0.025	0.822	0.197	−0.028	0.053
9B L36 16K†	54	−0.698	0.723	0.225	−0.607	−0.742
Llama L15 131K	300	—	0.653	—	—	—

† High censoring (46%) inflates ridge regression instability; AUROC and Spearman remain interpretable among the 54 uncensored features.

$3.4\times$ more aligned with the refusal direction than hard-to-steer features (mean alignment 0.129 vs. 0.038), consistent with co-recruitment of refusal-adjacent neighbors amplifying the behavioral effect.

The 65K SAE replication at layer 20 (2B) confirms that the signal reflects *relative* geometric crowding rather than absolute geometry: despite the 65K SAE having substantially higher absolute cosine values (max cosine mean 0.547 vs. 0.392 for 16K), effect sizes are comparable ($\rho = -0.435$ vs. -0.288 for neighbor density), and the signal survives a fourfold increase in dictionary size. A robustness check at layer 22 dropping the two outlier features at $\alpha^* = 10$ yielded $\rho = -0.470$ ($p = 3.6 \times 10^{-5}$), confirming results are not driven by boundary cases.

Co-activation correlation. Co-activation correlation requires careful interpretation. In five of six conditions it is null ($p > 0.08$). At 9B L20, the full-sample correlation appears significant ($r_{\text{full}} = +0.328$, $p = 0.004$), but this is attributable to zero-inflation: 59 of 75 features have activation frequency zero on the 99-prompt reference corpus, and all zero-frequency features have coactivation exactly zero by construction. Within the 16 features that actually activate on the reference corpus, the coactivation–α^* correlation is $p = 0.38$ (null). At Llama-3.1-8B L23, co-activation correlation reaches marginal significance ($\rho = -0.134$, $p = 0.020$), but the sign is negative (consistent with geometry), the effect is smaller than both geometry metrics, and we treat this as suggestive rather than conclusive given the single-run design. We conclude that co-activation correlation as operationalized here — mean Pearson correlation over a small task-specific corpus — is not a reliable pre-steering predictor of steerability, and that the metric requires a larger or more targeted activation corpus to be interpretable. Geometry appears to be the dominant predictive signal in our data; co-activation contributes little independent information.

Activation-frequency baseline. As an additional baseline, we tested whether behavioral proxies available prior to steering predict $\alpha^*(f)$ comparably to geometric metrics. The change in refusal rate at $\alpha = 1$ — that is, $B(1, f) - B(0, f)$, the behavioral shift at the smallest non-zero grid point, a forward-pass-only proxy requiring one steered run per feature rather than a full sweep — yields $\rho = -0.339$ ($p = 0.001$, $n = 89$), weaker than both neighbor density ($\rho = -0.435$) and max cosine similarity ($\rho = -0.404$). Larger coefficient responses ($\alpha = 5$, $\alpha_{\text{ref}} = -20$) are not significant predictors. Base refusal rate is constant across features and carries no signal. This suggests that decoder-space geometry provides predictive information beyond what is recoverable from behavioral measurements at low steering coefficients.

Off-target effects (RQ2 and RQ3). Whether harder-to-steer features also produce greater off-target effects (RQ2), and whether geometrically dense features pose structural risk even at low shared coefficients (RQ3), are planned extensions. The geometry signal identified here provides a natural basis for both: if crowded decoder neighborhoods predict intervention ease, they may also predict collateral risk. Empirical validation is deferred to future work.

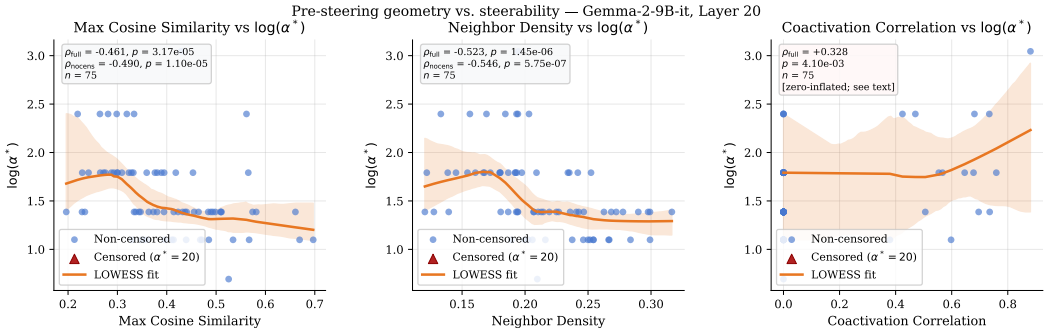

Figure 2: Pre-steering metrics versus log steerability for the 9B model at layer 20 (strongest signal condition). Each point is one SAE feature; red triangles indicate right-censored features ($\alpha^* = 20$). Trend line shows LOWESS fit; shaded band is bootstrap 95% CI. Neighbor density: $\rho = -0.546$, $p = 5.75 \times 10^{-7}$. Max cosine similarity: $\rho = -0.490$, $p = 1.10 \times 10^{-5}$. Co-activation correlation: zero-inflated; see text.

6.4 Depth Boundary

Two depth-dependent patterns emerge, one in each model. Here $\alpha_{\max}$ denotes the residual-stream activation norm at the target layer, which scales the effective force applied by a fixed steering coefficient. At layer 24 (2B, 92% relative depth), the steerability distribution collapsed entirely, which we attribute to substantially higher $\alpha_{\max}$ values ($\sim$188 vs. $\sim$75 at layer 22) causing the fixed coefficient grid to apply approximately $2.5\times$ more force and inducing coherence collapse in the dose-response curve. At layer 36 (9B, 86% relative depth), the effect is partial: geometry still predicts steerability among the 54 features that reach threshold, but the majority require steering coefficients beyond our current grid. Both patterns are consistent with the cost of steering growing as deep proportional layer depth approaches the output layer. We note that the 46% censoring at 9B L36 also attenuates our Spearman estimates by collapsing rank variance among features tied at $\alpha^* = \alpha_{\max}$; the true geometry–steerability relationship at this depth may be stronger than our reported correlations indicate.

Within the Gemma-2 family, the pattern is consistent: Spearman ρ for neighbor density moves from -0.288 (2B L20, 77% depth) to -0.512 (2B L22, 85%) to null at L24 (92%), and from -0.546 (9B L20, 48%) to -0.327 (9B L36, 86%). The geometry–steerability signal is strongest at shallow-to-mid proportional depth and degrades as the intervention layer approaches the output, marking a practical boundary for the framework.

7 Conclusion

Decoder-space geometry partially predicts SAE feature steerability before any intervention is applied. Neighbor density and maximum cosine similarity consistently rank-order features by $\log \alpha^*(f)$ across all well-identified conditions (ρ up to -0.546, AUROC 0.822 at 9B layer 20), the signal is detectable cross-architecturally on Llama-3.1-8B-Instruct, and co-activation correlation carries no independent predictive signal once activation sparsity is accounted for. The framework has clear boundary conditions: the signal degrades at deep proportional layer depth, and on Qwen3-8B with BatchTopK SAEs geometry predicts binary steerability but not the continuous ordering among responsive features.

A practitioner with access to the SAE decoder weight matrix can screen features for controllability before running a single steering experiment. Whether geometrically dense features also produce greater off-target effects (RQ2 and RQ3) is the most important open question for translating this framework into operational safety guidance.

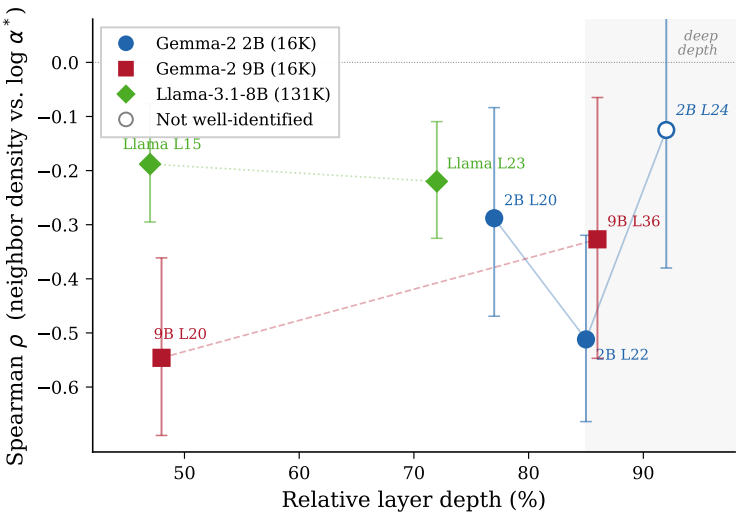

Figure 3: Spearman ρ (neighbor density vs. $\log \alpha^*$) as a function of relative layer depth across all conditions. The geometry–steerability signal is strongest at shallow-to-mid depth and weakens consistently as the intervention layer approaches the output. Open marker indicates a not-well-identified condition (Section 3.2). Error bars show approximate 95% CIs via Fisher z-transform. Shaded region marks the deep-depth boundary.

7.1 Limitations

Our experiments cover four models and one behavioral domain (refusal); whether the continuous geometry–steerability relationship holds beyond GemmaScope's JumpReLU training regime remains open, as Qwen3-8B with BatchTopK SAEs shows only binary steerability prediction. The fixed coefficient grid attenuates results at deep layers (46% censoring at 9B L36) and should be calibrated per layer in future work. The keyword classifier under-counts soft refusals, meaning reported effect sizes are likely conservative (Appendix C). Co-activation correlation is unreliable as operationalized due to zero-inflation. All experiments use positive steering only; suppression ($\alpha < 0$) and the Llama generalization (single layer, single run) are left to future work.

7.2 Future Work

Future work should test whether the same geometric metrics that predict steering cost also predict off-target behavioral change — evaluating features both at their minimum effective coefficient $\alpha^*(f)$ and at a shared low coefficient $\alpha_0 = 1$. A second extension would learn practical coefficient caps for features in high-risk geometric neighborhoods. Formal definitions of the off-target quantities are given in Appendix G.

Broader Impact

The ability to predict which features are easy to steer could lower the barrier for targeted misuse of activation-space interventions, particularly for refusal suppression in safety-tuned models. We acknowledge this dual-use risk explicitly. We note, however, that the same framework is directly useful defensively: identifying geometrically isolated features that require large coefficients to steer, as well as crowded features that may produce off-target effects through neighbor co-recruitment, provides a tool for auditing steering pipelines before deployment in safety-critical settings. Concretely, a safety team could use pre-steering geometry to flag high-risk features — those in dense neighborhoods that respond at small coefficients and may produce collateral effects through neighbor co-recruitment — before any intervention is attempted, rather than discovering these properties through expensive post-hoc evaluation. A framework that can predict intervention difficulty and collateral risk in advance is more likely to surface unsafe steering configurations than one that leaves these properties opaque. Responsible disclosure practices apply; we will make code available contingent on safety review before publication.

References

Arad, D., Mueller, A., and Belinkov, Y. SAEs are good for steering – if you select the right features. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10241–10259, Suzhou, China, November 2025. Association for Computational Linguistics.

Anthropic. Evaluating feature steering: A case study in mitigating social biases. Technical report, Anthropic, October 2024. `https://www.anthropic.com/research/evaluating-feature-steering`.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Karvonen, A., et al. BatchTopK sparse autoencoders for Llama-3.1-8B-Instruct. `https://huggingface.co/andyrdt/saes-llama-3.1-8b-instruct`, 2024.

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., and Shao, J. SALADBench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Li, M. Z., Agrawal, K. K., Ghosh, A., Teru, K. K., Santoro, A., Lajoie, G., and Richards, B. A. Tracing the representation geometry of language models from pretraining to post-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Lieberum, T., Raber, S., Brcic, T., et al. GemmaScope: Open sparse autoencoders everywhere all at once on Gemma 2. *arXiv preprint arXiv:2408.05147*, 2024.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of ACL*, 2022.

Park, K., Choe, Y. J., Jiang, Y., and Veitch, V. The geometry of categorical and hierarchical concepts in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Karvonen, A., et al. BatchTopK sparse autoencoders for Qwen3-8B. `https://huggingface.co/adamkarvonen/qwen3-8b-saes`, 2024.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. M. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2024.

Bussmann, B., Mendel, P., and Conmy, A. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06123*, 2024.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. Toy models of superposition. *Transformer Circuits Thread*, 2022.

A Feature Selection Details

For each SAE feature f, the composite contrast score is defined as

$$C(f) = \text{act\_freq}_{\text{task}}(f) - \text{act\_freq}_{\text{neutral}}(f), \tag{3}$$

where activation frequency is computed with max pooling over token positions within each prompt. Features are ranked by $C(f)$ and the top 300 are retained.

For each feature f in the contrast-selected set, we then compute an output relevance score:

$$\delta(f) = \left\| p_{\text{steered}}(\cdot \mid x) - p_{\text{base}}(\cdot \mid x) \right\|_1, \tag{4}$$

averaged over a held-out prompt set, where p_{steered} denotes the output distribution with feature f clamped at a reference coefficient. Features with $\delta(f) < 0.01$ are discarded.

B Steering Grid and Censoring

The coefficient grid $\alpha \in \{0, 0.25, 0.5, 1, 2, 3, 5, 10, 20\}$ is chosen to span two orders of magnitude with denser coverage at low coefficients, where behavioral transitions are expected. The extended upper range (10, 20) is included to capture hard-to-steer features that do not respond at lower coefficients; a reference coefficient of $\alpha_{\text{ref}} = -20$ is used for baseline estimation.

Right-censored features ($\alpha^*(f) > \alpha_{\max}$) are included in Spearman correlations using standard rank-based handling of ties at the boundary and excluded from R^2 regression to avoid artificial ceiling effects. We report censoring rates alongside all summary statistics. Censoring rates varied substantially across conditions, from 2.7% (9B layer 20) to 46% (9B layer 36), reflecting the depth-dependent growth in steering cost described in Section 6.

C Refusal Scorer

The on-target refusal classifier uses a keyword-matching approach: a completion is labeled as a refusal if it contains any of a curated set of refusal phrases (e.g., "I cannot," "I'm unable," "As an AI"). The phrase list is provided in the codebase (`src/refusal_scorer.py`).

Classifier reliability. We evaluated the keyword classifier against an independent judge on 160 completions stratified across α levels ($\alpha \in \{0, 0.5, 1, 2, 3, 5, 10, 20\}$), with equal representation of refusal and non-refusal completions at each level.

Results: precision 97.9%, recall 57.5%, F1 72.4%, overall agreement 78.1% ($n = 160$; TP=46, TN=79, FP=1, FN=34). The high precision and moderate recall reflect a systematic pattern: the classifier reliably identifies explicit refusals but misses soft refusals — completions that decline without using canonical refusal phrases (e.g., "this perpetuates a harmful stereotype" without an explicit "I cannot"). Agreement degrades at high steering coefficients ($\alpha = 10$: 60%; $\alpha = 20$: 50%), where coherence collapse produces degenerate repetition outputs that contain no refusal phrases and are scored 0 by the classifier regardless of interpretability.

Because the classifier systematically under-counts refusals, our behavioral threshold $T = 0.10$ is conservative: features that genuinely reach threshold may be scored as not reaching it, meaning our reported $\alpha^*(f)$ values are likely overestimates and our Spearman correlations are likely attenuated. The true geometry–steerability relationship is therefore probably stronger than reported.

D Threshold Sensitivity

We re-extract $\alpha^*(f)$ from existing $B(\alpha, f)$ curves for thresholds $T \in \{0.05, 0.10, 0.15, 0.20\}$ without any new model runs. As T increases, more features are right-censored: for the 65K condition, censoring rises from 3% at $T = 0.05$ to 30% at $T = 0.20$, confirming that $T = 0.10$ sits in a well-identified range. Only $\alpha^*(f)$ and the subsequent Spearman computation change across columns; decoder geometry is fixed. Spearman correlations are stable across all four thresholds, confirming $T = 0.10$ is not a cherry-picked boundary.

E Pre-Steering Geometry Statistics

Table 5 reports mean and standard deviation of neighbor density and maximum cosine similarity across all conditions. The 65K SAE shows higher absolute geometry values by construction of the wider dictionary. Llama-3.1-8B geometry values are substantially higher than GemmaScope conditions, consistent with the larger feature dictionary compressing decoder directions into a more crowded space. Qwen3-8B geometry variance is notably compressed, consistent with the BatchTopK training regime producing a more uniform decoder geometry.

Table 5: Pre-steering geometry statistics per condition.

Condition	ND mean	ND std	MC mean	MC std
2B L20 16K	0.194	0.053	0.392	0.143
2B L22 16K	0.178	0.049	0.368	0.141
2B L20 65K	0.323	0.078	0.547	0.139
9B L20 16K	0.212	0.045	0.400	0.110
9B L36 16K	0.180	0.045	0.373	0.127

ND = neighbor density; MC = max cosine similarity.

F Qwen3-8B Results

We evaluate the geometry–steerability relationship on Qwen3-8B using a BatchTopK SAE at layer 18 ($\sim$50% relative depth). Across 113 contrast-selected features with a baseline refusal rate of 0.46, we find a statistically significant positive correlation between geometry metrics and $\log \alpha^*(f)$ for both max cosine similarity ($\rho = 0.279$, $p = 0.003$) and neighbor density ($\rho = 0.275$, $p = 0.003$) on the full sample. However, restricting to the 98 uncensored features, both correlations collapse to near zero ($\rho \approx 0.01$), indicating that geometry predicts whether a feature is steerable at all rather than the gradient of steerability among responsive features. The positive full-sample correlation reflects that censored features (those that fall on the higher-geometry end of the distribution in this condition, consistent with a sign-reversed boundary case) dominate the sample; among uncensored features $\rho \approx 0.01$, which is not a reversal of direction but an absence of continuous signal. This is consistent with BatchTopK training combined with aggressive attention-sink filtering compressing decoder variance and weakening the density–isolation contrast that the framework relies on. Co-activation correlation is null ($\rho = 0.065$, $p = 0.49$), consistent with all other conditions.

We attribute the qualitatively different steerability response to the BatchTopK training regime. BatchTopK enforces sparsity by selecting the top-k activations across the entire batch rather than applying per-token thresholds as in GemmaScope's JumpReLU training. This batch-level competition means features do not need to occupy geometrically distinct, well-separated directions to remain active, decoupling decoder geometry from functional behavior. Additionally, the Qwen SAE training included aggressive filtering of attention sink activations, which are precisely the high-norm, geometrically isolated activations that would otherwise create strong directional structure in the decoder. Together, these choices appear to produce a more uniform decoder geometry with compressed variance, reducing the signal available for continuous geometry-based prediction.

The Qwen result therefore characterizes a boundary condition of the framework: the geometry–steerability relationship as operationalized here applies to SAEs where decoder geometry reflects functional isolation, and is attenuated or qualitatively changed when SAE training decouples geometry from behavior. This is a useful constraint on the generalizability of the framework.

G Off-Target Analysis Definitions

The planned off-target analysis evaluates behavioral change in two regimes: at $\alpha = \alpha^*(f)$ (risk at minimum effective push) and at a fixed low coefficient $\alpha_0 = 1$ (testing whether features can be risky even when gently steered). For off-target benchmarks $\{U_j\}_{j=1}^m$, the relevant quantities are:

$$R_{\text{mag}}(\alpha, f) = \frac{1}{m} \sum_{j=1}^{m} |\Delta U_j(\alpha, f)|, \tag{5}$$

$$R_{\text{breadth}}(\alpha, f) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}\big\{|\Delta U_j(\alpha, f)| > \tau\big\}. \tag{6}$$

Candidate off-target benchmarks include GPQA (Rein et al., 2023) for reasoning and TruthfulQA (Lin et al., 2022) for truthfulness.