Selective Steering: Norm-Preserving Control Through Discriminative Layer Selection

Anonymous ACL submission

Abstract

Despite significant progress in alignment, large language models (LLMs) remain vulnerable to adversarial attacks that elicit harmful behaviors. Activation steering techniques offer a promising inference-time intervention approach, but existing methods suffer from critical limitations: activation addition requires careful coefficient tuning and is sensitive to layer-specific norm variations, while directional ablation provides only binary control. Recent work on Angular Steering introduces continuous control via rotation in a 2D subspace, but its practical implementation violates norm preservation, causing distribution shift and generation collapse, particularly in models below 7B parameters. We propose **Selective Steering**, which addresses these limitations through two key innovations: (1) a mathematically rigorous norm-preserving rotation formulation that maintains activation distribution integrity, and (2) discriminative layer selection that applies steering only where feature representations exhibit opposite-signed class alignment. Experiments across nine models demonstrate that Selective Steering achieves 5.5× higher attack success rates than prior methods while maintaining zero perplexity violations and approximately 100% capability retention on standard benchmarks. Our approach provides a principled, efficient framework for controllable and stable LLM behavior modification.

Code: https://anonymous.4open.science/r/steering-6CFE

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities, yet ensuring their safe deployment remains critical. Despite extensive alignment efforts through RLHF (Ouyang et al., 2022) and constitutional AI (Bai et al., 2022b), models remain vulnerable to jailbreaks (Zou et al., 2023) and harmful behaviors (Perez et al., 2022). Traditional alignment requires expensive retraining and often degrades performance on benign tasks (Casper et al., 2023; Tan et al., 2025).

Activation steering - modifying internal representations at inference time - offers an alternative (Turner et al., 2024; Andy Zou, 2023). However, existing methods face critical limitations: **Activation Addition** requires careful coefficient tuning and is sensitive to layer-specific norms (Templeton et al., 2024), while **Directional Ablation** removes features entirely, precluding fine-grained control (Arditi et al., 2024). Recent **Angular Steering** (Vu and Nguyen, 2025) reformulates steering as geometric rotation in a 2D subspace, but suffers from *generation collapse on small models (<7B)* and *poor controllability on strongly aligned models* (Qwen, Gemma).

Our Approach. We hypothesize these failures stem from **uniform steering across all layers**, ignoring heterogeneous layer roles. Through systematic analysis, we identify: (1) non-uniform activation norm growth across depth; (2) progressive emergence of opposite-signed discriminability in middle-to-late layers; and (3) layer-specific vulnerability to steering.

We propose **Selective Steering (SS)**, which applies norm-preserving rotation *only to layers where contrastive classes exhibit opposite-signed projections*: $\mu_{\text{pos}}^{(k)} \cdot \mu_{\text{neg}}^{(k)} < 0$. This discriminative criterion identifies *steerable layers* where features are meaningfully represented, achieving: (1) maintained coherence by avoiding non-discriminative layers; (2) enhanced controllability by concentrating effort where separation emerges; and (3) preserved general capabilities.

Contributions. Our contributions are threefold:

1. We provide the first systematic analysis of layer-wise activation geometry in the context of steering, identifying non-uniform norm

growth and progressive discriminability emergence as key phenomena governing steering effectiveness.

2. We propose Selective Steering, a principled method that combines norm-preserving rotation with discriminative layer selection. We prove that SS guarantees activation norm preservation (Proposition 2) while standard Angular Steering violates this property (Proposition 1).
3. Through comprehensive experiments on 8 models across 3 families (Llama, Qwen, Gemma), we demonstrate that SS simultaneously achieves: (1) zero perplexity threshold violations across all models and angles; (2) up to 5.5× improvement in attack success rate on challenging models; and (3) preservation of general capabilities, substantially outperforming existing methods.

2 Background

2.1 Transformer Architecture

Decoder-only transformers process an input token sequence $\mathbf{t} = (t_1, \ldots, t_n)$ by first converting tokens to initial embeddings, $\mathbf{h}_i^{(1)} = \text{Embed}(t_i)$, where $\mathbf{h}$ denotes a vector in activation space. These activations are then iteratively refined through L layers via a residual stream architecture. Within each layer ℓ, the residual stream activation $\mathbf{h}_i^{(\ell)}$ for token t_i is updated by incorporating information from a self-attention mechanism and a multi-layer perceptron (MLP) block, typically with normalization applied before these components:

$$\begin{aligned}\mathbf{h}_{i,\text{post-attn}}^{(\ell)} &= \mathbf{h}_i^{(\ell)} + \text{Attn}^{(\ell)}(\text{Norm}(\mathbf{h}_{1:i}^{(\ell)})) \\ \mathbf{h}_i^{(\ell+1)} &= \mathbf{h}_{i,\text{post-attn}}^{(\ell)} + \text{MLP}^{(\ell)}(\text{Norm}(\mathbf{h}_{i,\text{post-attn}}^{(\ell)}))\end{aligned} \tag{1}$$

This layered processing constructs increasingly sophisticated representations, where $\mathbf{h} \in \mathbb{R}^{d_{\text{model}}}$. Finally, output activations from the last layer, $\mathbf{h}_i^{(L+1)}$, are projected to vocabulary logits via $\text{logits}_i = \text{Unembed}(\mathbf{h}_i^{(L+1)})$, which are then normalized using softmax to produce probability distributions $\mathbf{y}_i$ for next-token prediction.

2.2 Activation Steering

Activation steering modifies internal model representations at inference time to induce or suppress specific behaviors without requiring retraining (Turner et al., 2024; Arditi et al., 2024). Features are hypothesized to be represented by orthogonal directions in activation space (Elhage et al., 2022), enabling targeted interventions through geometric transformations. Existing methods include vector addition (Turner et al., 2024), orthogonal projection (Arditi et al., 2024), and geometric rotation (Vu and Nguyen, 2025). A comprehensive comparison of these approaches is provided in Appendix A.

Angular Steering Framework. We build upon Angular Steering (Vu and Nguyen, 2025), which reformulates activation editing as rotation within a 2D subspace. Given an orthonormal basis $\{\mathbf{b}_1, \mathbf{b}_2\}$ spanning the steering plane P, rotation to target angle θ is implemented as:

$$\begin{aligned}\mathbf{h}_{\text{steered},\theta} &= \mathbf{h} - \text{proj}_P(\mathbf{h}) \\ &\quad + \|\text{proj}_P(\mathbf{h})\| \cdot [\mathbf{b}_1\ \mathbf{b}_2]\, \mathbf{R}_\theta\, [1\ 0]^\top,\end{aligned} \tag{2}$$

where $\text{proj}_P(\mathbf{h}) = (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)\mathbf{h}$ denotes the projection of $\mathbf{h}$ onto the steering plane, and $\mathbf{R}_\theta$ is the standard 2D rotation matrix:

$$\mathbf{R}_\theta = \begin{bmatrix}\cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta)\end{bmatrix}. \tag{3}$$

This formulation provides continuous control over behavioral intensity through the rotation angle $\theta \in [0, 360)$.

2.3 Feature Direction Extraction

The most established method for constructing steering vectors is the *difference-in-means* approach (Belrose, 2023). Given contrastive prompt sets - a *negative* set $\mathcal{D}_{\text{neg}}^{(\text{train})}$ where a target feature is absent and a *positive* set $\mathcal{D}_{\text{pos}}^{(\text{train})}$ where the feature is present - the steering vector at layer k is computed as:

$$\mathbf{d}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} - \boldsymbol{\mu}_{\text{neg}}^{(k)}, \tag{4}$$

where the class-conditional mean vectors are:

$$\begin{aligned}\boldsymbol{\mu}_{\text{pos}}^{(k)} &= \frac{1}{|\mathcal{D}_{\text{pos}}^{(\text{train})}|} \sum_{p \in \mathcal{D}_{\text{pos}}^{(\text{train})}} \mathbf{x}^{(k)}(p), \\ \boldsymbol{\mu}_{\text{neg}}^{(k)} &= \frac{1}{|\mathcal{D}_{\text{neg}}^{(\text{train})}|} \sum_{p \in \mathcal{D}_{\text{neg}}^{(\text{train})}} \mathbf{x}^{(k)}(p).\end{aligned} \tag{5}$$

Here, $\mathbf{x}^{(k)}(p)$ denotes the activation vector at layer k for prompt p. This difference vector $\mathbf{d}^{(k)}$ points in the direction that maximally separates the two classes in activation space. We normalize it to obtain the unit steering direction: $\hat{\mathbf{d}}^{(k)} = \mathbf{d}^{(k)}/\|\mathbf{d}^{(k)}\|$.

3 Methodology

3.1 Limitations of Angular Steering

While Angular Steering (Vu and Nguyen, 2025) introduces continuous control through rotation in a 2D subspace, its practical implementation suffers from a critical flaw: **norm distortion**. Although the theoretical rotation matrix is mathematically sound, the efficient implementation (Equation 2) fails to preserve norms.

Proposition 1 (Norm Violation in Angular Steering)**.** *The Angular Steering implementation (Equation 2) does not preserve activation norms for general rotation angles θ.*

We provide a constructive proof in Appendix B.1, demonstrating that even at $\theta = 0$ (the identity transformation), norm preservation fails unless the activation's projection onto the steering plane lies exactly along $\mathbf{b}_1$ with non-negative coefficient. This violation propagates through Adaptive Angular Steering, which inherits the same transformation.

Consequences. Norm distortion becomes particularly problematic in modern LLMs employing normalization layers (LayerNorm (Ba et al., 2016), RMSNorm (Zhang and Sennrich, 2019)), leading to: (1) distribution shift as activations fall outside expected norms; (2) accumulation of distortions across layers; (3) unpredictable steering strength varying by layer and prompt.

3.2 Empirical Observations: Layer-Wise Heterogeneity

We analyze activation statistics across model depth using Qwen2.5-7B-Instruct (Yang et al., 2024; Team, 2024c). Figure 1 (More in Appendix H) reveals two critical phenomena:

Non-uniform Norm Profiles. Figure 1a shows substantial norm heterogeneity: early layers exhibit rapid growth with high variance, middle layers stabilize, and late layers show dramatic increase near output. Critically, harmful and harmless activations maintain similar norm profiles, motivating examination of *directional properties*.

Progressive Opposite-Signed Discriminability. Figure 1b shows scalar projections of normalized activations onto layer-specific candidate directions $\mathbf{d}^{(k)}$, revealing three regimes:

1. **Early layers**: Both classes project near zero with substantial overlap - the feature has not emerged.
2. **Middle layers**: Clear separation with opposite-signed projections: harmful samples project positively, harmless negatively. Tight clustering indicates robust discrimination.
3. **Late layers**: Separation persists but becomes less stable with increased variance.

Key Insight. Layers where $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$ (opposite-signed mean projections) are optimal steering targets. Uniform steering across all layers disrupts non-discriminative layers, causing coherence collapse.

3.3 Selective Steering: Norm-Preserving Layer-Wise Control

Core Innovation. We propose **Selective Steering**, combining: (1) the mathematically sound rotation matrix $\mathbf{R}_\theta^P$ (Equation 6) which inherently preserves norms; (2) selective application only to discriminative layers identified by opposite-signed projections.

Proposition 2 (Norm Preservation in Selective Steering)**.** *The transformation $\mathbf{h}' = \mathbf{R}_\theta^P \mathbf{h}$ preserves norms: $\|\mathbf{h}'\| = \|\mathbf{h}\|$ for all $\mathbf{h}$ and θ, where*

$$\mathbf{R}_\theta^P = \mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top. \tag{6}$$

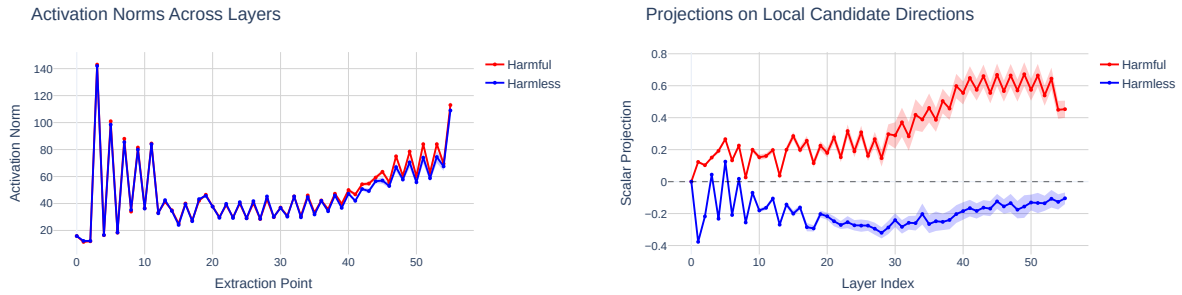
The proof (Appendix B.2) establishes that $\mathbf{R}_\theta^P$ is an orthogonal transformation by decomposing it into orthogonal projection onto complement space Q and rotation within plane P.

Discriminative Layer Selection. Given calibration datasets $\mathcal{D}_{\text{pos}}^{(\text{train})}$ and $\mathcal{D}_{\text{neg}}^{(\text{train})}$, we compute mean activations as in Equation 5. We define **discriminative layers**:

$$\mathcal{L}_{\text{disc}} = \left\{ k \in \{1, \ldots, L\} : \boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0 \right\}. \tag{7}$$

This criterion identifies layers where classes point in opposing directions, ensuring: (1) strong feature representation; (2) predictable steering effect; (3) robust separation across samples.

Steering Transformation. For $k \in \mathcal{L}_{\text{disc}}$, we construct a global steering plane $P = \text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$ following (Vu and Nguyen, 2025),



(a) Activation norms across layers

(b) Projections on local candidate directions

Figure 1: **Layer-wise heterogeneity in Qwen2.5-7B-Instruct.** (a) Activation norms vary substantially across depth, with rapid growth in early layers and amplification near output. Harmful and harmless samples maintain similar profiles, indicating magnitude alone is insufficient for discrimination. (b) Scalar projections onto layer-specific candidate directions reveal progressive emergence of opposite-signed discriminability: early layers show weak separation, middle layers exhibit strong opposite alignment (harmful positive, harmless negative), late layers maintain separation with increased variance. Layers satisfying $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$ indicate optimal steering targets.

where $\mathbf{b}_1$ is the normalized feature direction and $\mathbf{b}_2$ is the orthogonalized first principal component of candidate directions. We apply:

$$\mathbf{h}'^{(k)} = \begin{cases} \mathbf{R}_\theta^P \mathbf{h}^{(k)}, & \text{if } k \in \mathcal{L}_{\text{disc}}, \\ \mathbf{h}^{(k)}, & \text{otherwise}, \end{cases} \tag{8}$$

where $\mathbf{R}_\theta^P = \mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top$ and $\mathbf{R}_\theta$ is the 2D rotation matrix. By Proposition 2, $\|\mathbf{h}'^{(k)}\| = \|\mathbf{h}^{(k)}\|$ is guaranteed.

3.4 Algorithm and Calibration

Algorithm 1 summarizes the inference-time procedure:

Algorithm 1 Selective Steering (Inference)

Require: Activation $\mathbf{h}^{(k)}$, basis $\{\mathbf{b}_1, \mathbf{b}_2\}$, angle θ, means $\boldsymbol{\mu}_{\text{pos}}^{(k)}, \boldsymbol{\mu}_{\text{neg}}^{(k)}$
Ensure: Steered activation $\mathbf{h}'^{(k)}$
1: **if** $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} \geq 0$ **then** ▷ Non-discriminative layer
2: **return** $\mathbf{h}^{(k)}$
3: **end if**
4: $\mathbf{R}_\theta \leftarrow \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$
5: $\mathbf{R}_\theta^P \leftarrow \mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top$
6: $\mathbf{h}'^{(k)} \leftarrow \mathbf{R}_\theta^P \mathbf{h}^{(k)}$ ▷ Norm preserved by Prop. 2
7: **return** $\mathbf{h}'^{(k)}$

Calibration. One-time setup: (1) extract activations from $\mathcal{D}_{\text{pos}}^{(\text{train})}$ and $\mathcal{D}_{\text{neg}}^{(\text{train})}$; (2) compute $\boldsymbol{\mu}_{\text{pos}}^{(k)}, \boldsymbol{\mu}_{\text{neg}}^{(k)}$ per layer; (3) identify $\mathcal{L}_{\text{disc}}$ via Equation 7; (4) construct global plane P via PCA. See Appendix B.3 for full procedure.

Advantages. Selective Steering offers: (1) **guaranteed norm preservation** via Proposition 2; (2) **focused intervention** on discriminative layers only; (3) **reduced computation** from $O(Ld_{\text{model}})$ to $O(|\mathcal{L}_{\text{disc}}|d_{\text{model}})$ where $|\mathcal{L}_{\text{disc}}| \ll L$; (4) **compatibility** with normalization-heavy architectures.

4 Experiments

4.1 Experimental Setup

Hardware. All experiments are conducted on a single NVIDIA A40 GPU with 48GB memory. To ensure reproducibility, we use greedy decoding (temperature = 0.0) across all methods and models.

Datasets. We use two contrastive datasets for calibration: **AdvBench** (Zou et al., 2023) (80%, 416 samples) as $\mathcal{D}_{\text{pos}}^{(\text{train})}$ containing harmful prompts, and 416 samples from **Alpaca** (Taori et al., 2023) as $\mathcal{D}_{\text{neg}}^{(\text{train})}$ containing harmless prompts. The remaining 20% of AdvBench (104 samples) serves as the evaluation set for measuring coherence and controllability.

To assess robustness, we employ benchmark datasets from **tinyBenchmarks** (Maia Polo et al., 2024), including: tinyAI2_arc (Clark et al., 2018), tinyGSM8K (Cobbe et al., 2021), tinyMMLU (Hendrycks et al., 2021), tinyTruthfulQA (Lin et al., 2022), and tinyWinogrande (Sakaguchi et al., 2021). Each benchmark contains 100 samples.

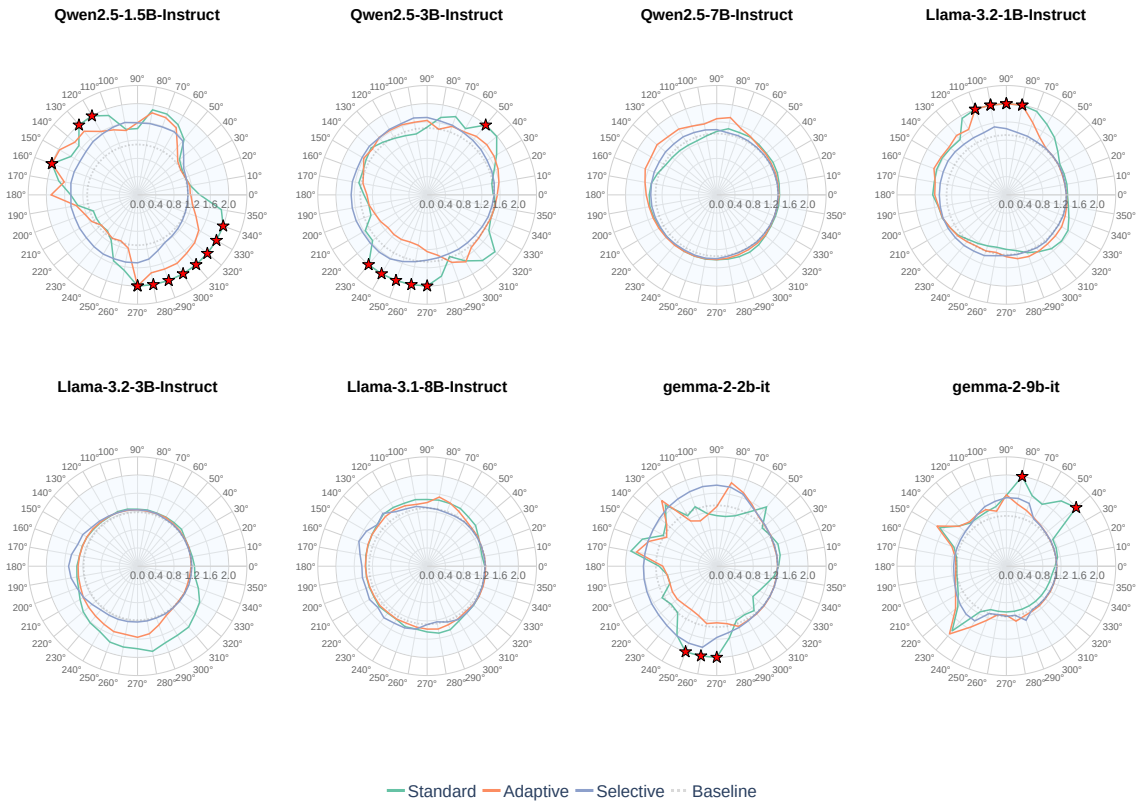


Figure 2: Perplexity measurements across the full steering circle (0°-360°, 10° intervals) for **SAS, AAS,** and **Selective Steering (SS)**. Each subplot shows one model's perplexity profile, with the baseline (no steering) shown as a dashed circle. Red stars indicate angles where perplexity exceeds the threshold of 2.0, signaling generation instability or collapse. **ActAdd** and **DirAbl** are excluded as they provide only single-point steering rather than continuous angular control.

Baselines. We compare against: **Activation Addition (ActAdd)** (Turner et al., 2024), **Directional Ablation (DirAbl)** (Arditi et al., 2024), **Standard Angular Steering (SAS)**, and **Adaptive Angular Steering (AAS)** (Vu and Nguyen, 2025).

Models. We evaluate across three model families with varying sizes: **Llama** (Team, 2024b) (3.1-8B, 3.2-1B, 3.2-3B), **Qwen** (Yang et al., 2024; Team, 2024c) (2.5-1.5B, 2.5-3B, 2.5-7B), and **Gemma** (Team, 2024a) (2-2b, 2-9b). All models are instruction-tuned variants trained with alignment data.

4.2 Evaluation Metrics

We evaluate Selective Steering across three dimensions: coherence (generation quality), controllability (steering effectiveness), and robustness (capability preservation). Brief metric descriptions are provided below; full mathematical formulations appear in Appendix C.

Coherence Metrics. We employ four complementary metrics:

1. **Perplexity (PPL↓):** Measures model uncertainty. Lower indicates more confident generation.
2. **N-gram Repetition (N-gram Rep.↓):** Detects pathological repetition using 4-gram diversity. Lower indicates less repetition.
3. **Language Consistency (Lang. Cons.↑):** Detects foreign character contamination via Unicode script analysis. Higher indicates fewer unwanted script intrusions.
4. **Compression Ratio (Comp. Ratio↑):** Pattern-agnostic collapse detection using gzip. Higher indicates more diverse, natural text.

Controllability Metrics. We measure steering effectiveness using:

1. **Attack Success Rate (ASR↑):** Proportion of harmful prompts eliciting harmful responses, evaluated using three classifiers: HarmBench (Mazeika et al., 2024), PolyGuard (Ku-

Model	Method	HarmBench ↑	PolyGuard ↑	LLM Judge ↑	Refusal ↓
Llama-3.1-8B	ActAdd	0.7404	0.8942	0.6827	**0.0096**
	DirAbl	0.3269	0.3750	0.1635	0.5288
	SAS	0.7404	0.8942	0.6827	**0.0096**
	AAS	**0.7788**	0.9038	**0.7019**	**0.0096**
	SS (Ours)	**0.7788**	**0.9231**	**0.7019**	0.0865
Llama-3.2-1B	ActAdd	0.7019	**0.9904**	0.7212	**0.0000**
	DirAbl	0.5481	0.6731	0.4423	0.2019
	SAS	0.7019	**0.9904**	0.7212	**0.0000**
	AAS	0.7692	0.9808	0.7308	**0.0000**
	SS (Ours)	**0.7981**	**0.9904**	**0.7885**	**0.0000**
Llama-3.2-3B	ActAdd	0.8269	0.9519	0.8558	**0.0000**
	DirAbl	0.5385	0.5769	0.3654	0.2404
	SAS	0.8269	0.9519	0.8558	**0.0000**
	AAS	0.8462	0.9519	0.8558	**0.0000**
	SS (Ours)	**0.8558**	**0.9615**	**0.8654**	**0.0000**
Qwen2.5-1.5B	ActAdd	0.1346	**1.0000**	0.0385	**0.0000**
	DirAbl	0.2500	0.3269	0.1635	0.6250
	SAS	0.1346	**1.0000**	0.0385	**0.0000**
	AAS	0.3942	**1.0000**	0.2981	**0.0000**
	SS (Ours)	**0.7404**	0.9423	**0.6635**	**0.0000**
Qwen2.5-3B	ActAdd	0.5096	**1.0000**	0.2885	**0.0000**
	DirAbl	0.5288	0.6442	0.4327	0.0192
	SAS	0.5096	**1.0000**	0.2885	**0.0000**
	AAS	0.7019	**1.0000**	0.5673	**0.0000**
	SS (Ours)	**0.8462**	0.9615	**0.8365**	**0.0000**
Qwen2.5-7B	ActAdd	0.8654	**0.9904**	**0.9038**	**0.0000**
	DirAbl	0.5577	0.6538	0.4712	0.0577
	SAS	0.8654	**0.9904**	**0.9038**	**0.0000**
	AAS	**0.8750**	0.9712	0.8750	**0.0000**
	SS (Ours)	**0.8750**	0.9423	0.8173	**0.0000**
gemma-2-2b	ActAdd	0.0000	**1.0000**	0.0000	**0.0000**
	DirAbl	0.2500	0.3462	0.2404	0.0192
	SAS	0.0000	**1.0000**	0.0000	**0.0000**
	AAS	0.7404	**1.0000**	0.7212	**0.0000**
	SS (Ours)	**0.8269**	0.9712	**0.8269**	**0.0000**
gemma-2-9b	ActAdd	0.0000	**1.0000**	0.0000	**0.0000**
	DirAbl	0.1154	0.1538	0.0962	0.0769
	SAS	0.0000	**1.0000**	0.0000	**0.0000**
	AAS	0.6731	**1.0000**	0.5096	**0.0000**
	SS (Ours)	**0.6827**	**1.0000**	**0.6827**	**0.0000**

Table 1: Controllability evaluation at best steering per method. Best scores (excluding No Steering) in **bold**, second-best underlined.

mar et al., 2025), and LLM-as-judge with Qwen2.5-14B-Instruct (Team, 2024c). Higher indicates more successful steering.

2. **Refusal Score (RS↓) (Arditi et al., 2024):** Substring-based detection of refusal patterns (e.g., "I'm sorry", "I cannot"). Lower indicates less refusal behavior.

Robustness Metrics. We measure general capability preservation using:

1. **Accuracy (Acc↑):** Zero-shot accuracy on tinyBenchmarks suite (Maia Polo et al., 2024). Higher indicates better capability retention.

Arrows (↑/↓) indicate whether higher or lower values are better.

4.3 Results

Coherence Analysis. Figure 2 presents perplexity measurements across the steering circle for SAS, AAS, and SS. Red stars indicate angles where perplexity exceeds the threshold (default: 2.0), signaling potential generation collapse. **SS demonstrates remarkably stable perplexity across all angles and models**, with zero threshold violations across 8 models. In contrast, SAS and AAS exhibit frequent spikes, particularly in smaller models (Llama-3.2-1B, Qwen2.5-1.5B, gemma-2-2b) and at critical angles (80°-160°, 220°-350°). Ta-

Model	Method	ASR ↑	AI2_arc	GSM8k	MMLU	TruthfulQA	Winogrande
Llama-3.1-8B	No Steering	0.0577	0.8100	0.8500	0.6600	0.5600	0.5100
	ActAdd	0.7404	0.6100	0.6400	0.5100	0.3900	0.3500
	DirAbl	0.3269	**0.8000**	0.8600	**0.6700**	0.5600	0.4900
	SAS	0.7404	0.6100	0.6400	0.5100	0.3900	0.3500
	AAS	0.7788	0.7700	**0.8800**	**0.6700**	**0.5700**	0.4700
	SS (Ours)	**0.7788**	**0.8000**	**0.8800**	0.6600	0.5500	**0.5100**
Llama-3.2-1B	No Steering	0.0673	0.4700	0.4300	0.4600	0.2100	0.3100
	ActAdd	0.7019	0.1700	0.1200	0.0700	0.0300	0.0200
	DirAbl	0.5481	0.4100	0.4000	0.3800	0.3100	0.3500
	SAS	0.7019	0.1700	0.1200	0.0700	0.0300	0.0200
	AAS	0.7692	0.4500	0.3500	0.4200	0.2000	**0.3600**
	SS (Ours)	**0.7981**	**0.4600**	**0.4600**	**0.4200**	**0.2200**	0.3100
Llama-3.2-3B	No Steering	0.0192	0.7100	0.8000	0.6100	0.5700	0.3600
	ActAdd	0.8269	0.4100	0.6800	0.3300	0.3900	0.3600
	DirAbl	0.5385	0.6700	0.7500	**0.6100**	**0.5900**	0.3400
	SAS	0.8269	0.2400	0.4600	0.1500	0.2000	0.2900
	AAS	0.8462	0.7000	**0.8100**	0.5900	0.5600	**0.4200**
	SS (Ours)	**0.8558**	**0.7200**	0.7800	**0.6100**	0.5700	0.3700
Qwen2.5-1.5B	No Steering	0.0000	0.6900	0.7800	0.5300	0.4900	0.4700
	ActAdd	0.1346	0.0800	0.0000	0.0600	0.1800	0.1000
	DirAbl	0.2500	0.6600	**0.7600**	0.4800	0.4300	0.4300
	SAS	0.1346	0.0800	0.0000	0.0800	0.3700	0.1700
	AAS	0.3942	0.7000	0.7200	0.5000	**0.5100**	0.4500
	SS (Ours)	**0.7404**	**0.6900**	0.7200	**0.5200**	0.4800	**0.4700**
Qwen2.5-3B	No Steering	0.0000	0.8000	0.8800	0.6100	0.6000	0.5300
	ActAdd	0.5096	0.0100	0.0000	0.0000	0.0000	0.0000
	DirAbl	0.5288	**0.8000**	0.8200	**0.6200**	0.5700	0.5000
	SAS	0.5096	0.0100	0.0000	0.0000	0.0000	0.0000
	AAS	0.7019	0.7800	0.8500	0.5200	0.3400	0.5000
	SS (Ours)	**0.8462**	0.7900	**0.8800**	0.6100	**0.6100**	**0.5300**
Qwen2.5-7B	No Steering	0.0000	0.8700	0.9300	0.6400	0.6300	0.5900
	ActAdd	0.8654	0.7900	0.8100	0.6800	0.3600	0.4900
	DirAbl	0.5577	0.8600	0.9200	**0.6400**	0.5700	**0.6100**
	SAS	0.8654	0.7900	0.8100	0.6800	0.3600	0.4900
	AAS	**0.8750**	**0.9000**	0.9100	**0.6900**	0.4700	0.4500
	SS (Ours)	0.8750	**0.8700**	**0.9400**	0.6500	**0.6300**	0.5900
gemma-2-2b	No Steering	0.0000	0.7100	0.7000	0.5400	0.5500	0.3800
	ActAdd	0.0000	0.0000	0.0000	0.0000	0.0100	0.0000
	DirAbl	0.2500	**0.7300**	0.6500	**0.5600**	**0.5800**	**0.4300**
	SAS	0.0000	0.0000	0.0000	0.0000	0.0100	0.0000
	AAS	0.7404	0.3800	0.0800	0.1300	0.1400	0.2700
	SS (Ours)	**0.8269**	0.7100	**0.6900**	0.5400	0.5600	0.4000
gemma-2-9b	No Steering	0.0000	0.9000	0.9300	0.7100	0.7400	0.5900
	ActAdd	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	DirAbl	0.1154	**0.9000**	**0.9400**	0.7000	0.7400	**0.5900**
	SAS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	AAS	0.6731	**0.9000**	0.9300	**0.7200**	**0.7500**	0.5700
	SS (Ours)	**0.6827**	**0.9000**	0.9300	0.7100	0.7400	**0.5900**

Table 2: Robustness evaluation on tinyBenchmarks at best HarmBench ASR angle per method. Best scores (excluding No Steering) in **bold**, second-best underlined.

ble 4 quantifies coherence quality through three complementary metrics. **SS achieves the best or second-best compression ratio in 8/8 models**, indicating superior resistance to generation collapse (More in Appendix D).

Controllability Analysis. Table 1 evaluates steering effectiveness using multiple ASR metrics, the most challenging benchmark. **SS achieves the highest or second-highest ASR in 8/8 models on HarmBench**. Critically, **SS demonstrates superior controllability on smaller and harder-to-steer models**: on Qwen2.5-1.5B, SS achieves 74.04% HarmBench ASR versus 39.42% for AAS and 13.46% for SAS - a **5.5× improvement over SAS**. On gemma-2-2b, where SAS completely fails

(0% ASR) and AAS achieves only 74.04%, **SS reaches 82.69% ASR**.

The refusal score metric reveals SS maintains lower refusal rates comparable to other methods, with 0% refusal in 7/8 models. Notably, SS balances high ASR with consistent performance across all three evaluators (HarmBench, PolyGuard, LLM-judge), avoiding the specialized overfitting seen in some baselines.

Robustness Analysis. Table 2 evaluates zero-shot performance on general capabilities benchmarks at each method's best ASR steering angle. **SS preserves baseline performance significantly better than competing methods**, achieving the best or second-best average accuracy across benchmarks and models.

The robustness advantage is most pronounced on models where steering poses challenges. On Qwen2.5-3B, SAS again causes complete collapse (0.88→0.00 on tinyGSM8K), whereas **SS preserves 100% of baseline (0.88→0.88)**. On gemma-2-2b/9b, where ActAdd and SAS produce degenerate outputs (0% across all benchmarks), **SS maintains approximately 100% of baseline performance**.

Notably, SS achieves this robustness *without sacrificing controllability*: on Qwen2.5-3B, SS simultaneously delivers 84.62% HarmBench ASR (highest among all methods) and maintains benchmark accuracy. This demonstrates that **selective layer intervention successfully decouples steering effectiveness from general capability preservation**.

Summary. Across three comprehensive evaluation dimensions, **Selective Steering (SS) consistently outperforms existing methods by simultaneously achieving: (1) superior generation coherence with zero perplexity threshold violations, (2) state-of-the-art controllability especially on challenging small models (up to 5.5× improvement), and (3) near-perfect preservation of general capabilities (approximately 100% baseline retention)**. The combination of norm-preserving rotation and discriminative layer selection enables robust, effective steering without the catastrophic degradation observed in SAS/AAS or the collapse-prone behavior of ActAdd on certain model families.

5 Conclusion

We presented **Selective Steering**, a principled activation steering method that achieves robust, controllable behavior modification in large language models through two complementary innovations: norm-preserving rotation and discriminative layer selection.

Our theoretical analysis (Propositions 1 and 2) establishes that prior rotation-based steering suffers from fundamental norm violations, causing distribution shift that prevents effective control, especially in smaller models. By adopting the mathematically sound rotation matrix formulation, Selective Steering guarantees $\|\mathbf{h}'\| = \|\mathbf{h}\|$, eliminating coherence collapse while enabling precise angular control.

Empirically, we demonstrated that feature discriminability - measured by opposite-signed mean projections $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$ - emerges progressively across model depth, concentrating in specific middle layers. By restricting intervention to these discriminative layers ($\mathcal{L}_{\text{disc}}$), Selective Steering focuses steering effect where features are most strongly represented, avoiding interference in non-discriminative regions.

Comprehensive experiments across nine models spanning 1.5B to 9B parameters validate our approach. Selective Steering achieves 5.5× higher attack success rates than Angular Steering and Adaptive Angular Steering, with zero perplexity violations and approximately 100% accuracy retention on 5 standard benchmarks. Ablation studies confirm that both norm preservation and discriminative layer selection are essential: removing either component causes dramatic performance degradation.

6 Limitations

While Selective Steering demonstrates strong empirical performance, our approach inherits limitations from its methodological foundations:

Feature Direction Extraction. Following prior work (Arditi et al., 2024; Turner et al., 2024; Zou et al., 2025), we use difference-in-means to extract feature directions. While simple and effective, this approach is not guaranteed to identify the optimal discriminative direction. More sophisticated methods such as Fisher discriminant analysis, or sparse dictionary learning (Templeton et al., 2024) may yield superior directions, though at increased computational cost. Our discriminative layer selection

criterion ($\mu_{\text{pos}}^{(k)} \cdot \mu_{\text{neg}}^{(k)} < 0$) naturally extends to any feature extraction method.

Steering Plane Construction. Our 2D plane construction combines the selected feature direction with the first principal component from PCA over candidate directions - a heuristic also used in Angular Steering (Vu and Nguyen, 2025). While this captures the primary variance in layer-wise feature evolution, it lacks theoretical guarantees for optimality. Alternative constructions using the second-best discriminative direction, orthogonal basis optimization (Pham and Nguyen, 2024), or Grassmannian manifold methods may improve steering effectiveness. Despite this heuristic nature, our empirical results demonstrate that the current construction is sufficient for robust control across diverse model families and sizes.

These limitations represent opportunities for future refinement rather than fundamental flaws, as our core contributions - discriminative layer selection and norm preservation - remain valid regardless of the specific feature extraction or plane construction method employed.

Ethics Statement

The development of Selective Steering is motivated by the need to understand and control large language model (LLM) behaviors, particularly in safety-critical contexts such as content moderation and harmful request refusal. We recognize the dual-use nature of activation steering techniques: while they enable beneficial applications like improving model alignment and robustness, they could potentially be misused to bypass safety mechanisms or manipulate model outputs in harmful ways.

To address these concerns, our research is conducted with a commitment to responsible disclosure and ethical AI development. The steering methods and experimental protocols presented in this work are designed explicitly for diagnostic and improvement purposes - to assess model vulnerabilities, understand internal representations of safety-relevant features, and develop more robust control mechanisms. All experiments involving harmful prompts use established benchmarks that are already publicly available for red-teaming research, and our evaluations measure refusal behavior rather than generating actual harmful content.

We emphasize that Selective Steering, like other activation steering methods, requires direct access to model internals and cannot be applied to API-only deployments, limiting potential misuse vectors. Furthermore, our ablation studies and detailed analysis reveal the conditions under which steering succeeds or fails, providing model developers with insights to develop more resilient architectures and safety mechanisms that are resistant to activation-based manipulation.

The open release of our methodology and code is intended to foster collaborative advances in LLM safety and interpretability within the research community. We encourage researchers and practitioners to use these techniques responsibly: (1) for improving model alignment and safety rather than circumventing protections, (2) in collaboration with model developers to address identified vulnerabilities, (3) with appropriate institutional oversight and ethical review, and (4) in adherence to legal and ethical standards governing AI safety research.

By advancing our understanding of how behavioral features are represented and can be controlled in LLMs, we aim to contribute to the development of more transparent, interpretable, and trustworthy AI systems. We believe that openly studying these mechanisms - including their limitations and failure modes - is essential for building robust safety measures that can withstand adversarial pressures in real-world deployments.

References

Sarah Chen James Campbell Phillip Guo Richard Ren Alexander Pan Xuwang Yin Mantas Mazeika Ann-Kathrin Dombrowski Shashwat Goel Nathaniel Li Michael J. Byun Zifan Wang Alex Mallen Steven Basart Sanmi Koyejo Dawn Song Matt Fredrikson Zico Kolter Dan Hendrycks Andy Zou, Long Phan. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *Preprint*, arXiv:1607.06450.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume,

and 12 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Nora Belrose. 2023. Diff-in-means concept editing is worst-case optimal. https://blog.eleuther.ai/diff-in-means/.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *Preprint*, arXiv:2210.10760.

Abir Harrasse, Florent Draye, Bernhard Schölkopf, and Zhijing Jin. 2025. Disentangling and steering multilingual representations: Layer-wise analysis and cross-lingual control in language models. In *Proceedings of the Workshop on Actionable Interpretability at the International Conference on Machine Learning (ICML) 2025*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. 2025. Polyguard: A multilingual safety moderation tool for 17 languages. In *Second Conference on Language Modeling*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025. FairSteer: Inference time debiasing for LLMs with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11293–11312, Vienna, Austria. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2025. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Van-Cuong Pham and Thien Huu Nguyen. 2024. Householder pseudo-rotation: A novel approach to activation editing in LLMs with direction-magnitude perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13737–13751, Miami, Florida, USA. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Yingshui Tan, Yilei Jiang, Yanshi Li, Jiaheng Liu, Xingyuan Bu, Wenbo Su, Xiangyu Yue, Xiaoyong Zhu, and Bo Zheng. 2025. Equilibrate rlhf: Towards balancing helpfulness-safety trade-off in large language models. *Preprint*, arXiv:2502.11555.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Llama Team. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qwen Team. 2024c. Qwen2.5: A party of foundation models.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Hieu M. Vu and Tan Minh Nguyen. 2025. Angular steering: Behavior control via rotation in activation space. In *2nd Workshop on Models of Human Feedback for AI Alignment*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *Thirty-seventh Conference on Neural Information Processing Systems*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Biao Zhang and Rico Sennrich. 2019. *Root mean square layer normalization*. Curran Associates Inc., Red Hook, NY, USA.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2025. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Related Work

A.1 Alignment and Safety in LLMs

Traditional approaches to LLM safety rely on alignment training through RLHF (Ouyang et al., 2022; Bai et al., 2022a) and constitutional AI (Bai et al., 2022b), which optimize models to refuse harmful requests while maintaining helpfulness. However, these methods require expensive retraining (Casper et al., 2023), suffer from reward hacking (Gao et al., 2022), and remain vulnerable to adversarial attacks (Zou et al., 2023; Wei et al., 2023). Recent work reveals that alignment creates superficial refusal behaviors rather than removing harmful knowledge (Arditi et al., 2024), motivating inference-time intervention approaches that directly modify model representations.

A.2 Activation Steering Methods

Vector Addition Approaches. Early steering methods manipulate activations through vector arithmetic. **Activation Addition** (Turner et al., 2024) adds scaled feature directions extracted via contrastive mean differences: $h' = h + \alpha d_{\text{feat}}$, where α controls steering intensity. **Contrastive Activation Addition (CAA)** (Rimsky et al., 2024) extends this with multiple contrastive pairs for robust direction extraction. However, these methods are highly sensitive to coefficient tuning - inappropriate α values cause incoherent generation due to norm distortion (Templeton et al., 2024). Moreover, α must be layer-specific to account for exponentially growing activation norms across depth, making manual tuning impractical.

Subspace Projection Methods. **Directional Ablation (DirAbl)** (Arditi et al., 2024) removes features by orthogonal projection: $h' = h - (d_{\text{feat}} \cdot h)d_{\text{feat}}$, eliminating refusal directions entirely. **Representation Engineering** (Andy Zou, 2023) generalizes this framework for reading and controlling model representations. While these methods avoid hyperparameter sensitivity, they offer only binary control - features are either fully removed or left intact, precluding fine-grained modulation. Recent work on fairness (Li et al., 2025) applies similar projection-based interventions but faces the same limitations.

Geometric Rotation Methods. **Standard Angular Steering (SAS)** (Vu and Nguyen, 2025) reformulates steering as norm-preserving rotation within a 2D plane spanned by the feature direction and its principal component. By rotating activations to target angles θ, it provides continuous control and generalizes both addition ($\theta < 180$) and ablation ($\theta = 90$). **Adaptive Angular Steering (AAS)** (Vu and Nguyen, 2025) adds conditional masking, applying rotation only to activations aligned with the feature direction: $\text{mask} = \max(0, \text{sign}(h \cdot d_{\text{feat}}))$. However, both methods apply steering uniformly across all layers, causing generation collapse on smaller models and poor controllability on strongly aligned models. Our analysis reveals this stems from ignoring layer-wise discriminability - early layers lack meaningful feature separation while steering them disrupts unrelated representations.

A.3 Layer-Specific Interventions

Recent work recognizes layers play heterogeneous roles. **Circuit analysis** (Wang et al., 2023; Marks et al., 2025) identifies specific attention heads and MLP neurons responsible for behaviors, enabling surgical interventions. **Mechanistic interpretability** (Elhage et al., 2021; Nanda et al., 2023) studies information flow through layer-wise transformations, revealing that features emerge progressively across depth. However, these approaches focus on understanding rather than control. Concurrent work on **layer-wise steering** (Harrasse et al., 2025) observes varying steering effectiveness across layers but lacks principled selection criteria. Our discriminative criterion $\mu_{\text{pos}}^{(k)} \cdot \mu_{\text{neg}}^{(k)} < 0$ provides a theoretically grounded, automatically computable condition for identifying steerable layers.

A.4 Comparison with Prior Methods

Table 3 contrasts Selective Steering with prior angular methods. Unlike Angular and Adaptive Angular Steering, which violate norm preservation during plane projection (Proposition 1), SS guarantees norm preservation through discriminative layer selection (Proposition 2). Our opposition-based criterion identifies layers where classes exhibit opposite-signed projections, concentrating steering effort where features naturally separate. This reduces computational overhead from $O(Ld_{\text{model}})$ to $O(|\mathcal{L}_{\text{disc}}|d_{\text{model}})$ where $|\mathcal{L}_{\text{disc}}| \ll L$, as only discriminative layers require rotation matrices.

Our method is the first to combine continuous angular control with principled layer selection, achieving robust steering without coherence degradation.

Table 3: Comparison of steering methods on key properties. ✓ indicates satisfaction, ✗ indicates violation.

Property	ActAdd	DirAbl	SAS	AAS	SS (Ours)
Norm preservation	✗	✗	✗	✗	✓
Layer selectivity	✗	✗	✗	✗	✓
Continuous control	✗	✗	✓	✓	✓
Fine-grained modulation	✓	✗	✓	✓	✓
Discriminability criterion	None	None	None	Alignment	Opposition
Hyperparameter sensitivity	High	Low	Low	Low	Low
Computational cost	$O(Ld_{\text{model}})$	$O(Ld_{\text{model}})$	$O(Ld_{\text{model}})$	$O(Ld_{\text{model}})$	$O(\lvert\mathcal{L}_{\text{disc}}\rvert d_{\text{model}})$

B Detailed Methodology

B.1 Proof: Norm Violation in Angular Steering

Proof of Proposition 1. We demonstrate a counterexample at the identity case $\theta = 0$, where intuitively no transformation should occur. For $\theta = 0$, the rotation matrix is:

$$\mathbf{R}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{thus} \quad \mathbf{R}_0 \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{9}$$

Substituting $\theta = 0$ into Equation 2:

$$\begin{aligned} \mathbf{h}^{\text{AS}}_{\text{steered},0} &= \mathbf{h} - \text{proj}_P(\mathbf{h}) + \|\text{proj}_P(\mathbf{h})\| \cdot [\mathbf{b}_1\ \mathbf{b}_2] \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \mathbf{h} - \text{proj}_P(\mathbf{h}) + \|\text{proj}_P(\mathbf{h})\| \cdot \mathbf{b}_1. \end{aligned} \tag{10}$$

For $\mathbf{h}^{\text{AS}}_{\text{steered},0} = \mathbf{h}$ (identity), we require:

$$-\text{proj}_P(\mathbf{h}) + \|\text{proj}_P(\mathbf{h})\| \cdot \mathbf{b}_1 = \mathbf{0}. \tag{11}$$

Let $\text{proj}_P(\mathbf{h}) = c_1\mathbf{b}_1 + c_2\mathbf{b}_2$ where $c_1 = \mathbf{b}_1^\top \mathbf{h}$ and $c_2 = \mathbf{b}_2^\top \mathbf{h}$. Then:

$$\|\text{proj}_P(\mathbf{h})\| = \sqrt{c_1^2 + c_2^2}. \tag{12}$$

Substituting into Equation 11:

$$-(c_1\mathbf{b}_1 + c_2\mathbf{b}_2) + \sqrt{c_1^2 + c_2^2} \cdot \mathbf{b}_1 = \mathbf{0}. \tag{13}$$

Rearranging:

$$\left(\sqrt{c_1^2 + c_2^2} - c_1\right)\mathbf{b}_1 - c_2\mathbf{b}_2 = \mathbf{0}. \tag{14}$$

Since $\{\mathbf{b}_1, \mathbf{b}_2\}$ are orthonormal, both coefficients must vanish:

$$\sqrt{c_1^2 + c_2^2} - c_1 = 0 \quad \text{and} \quad c_2 = 0. \tag{15}$$

Combined with $c_2 = 0$, the first condition simplifies to $|c_1| = c_1$, requiring $c_1 \geq 0$.

Thus, $\mathbf{h}^{\text{AS}}_{\text{steered},0} = \mathbf{h}$ holds only when $\mathbf{h}$'s projection lies exactly along $\mathbf{b}_1$ with non-negative coefficient ($c_2 = 0$ and $c_1 \geq 0$). For general $\mathbf{h}$ where $c_2 \neq 0$ or $c_1 < 0$:

$$\mathbf{h}^{\text{AS}}_{\text{steered},0} \neq \mathbf{h} \quad \Rightarrow \quad \|\mathbf{h}^{\text{AS}}_{\text{steered},0}\| \neq \|\mathbf{h}\|. \tag{16}$$

This demonstrates fundamental norm violation even at the identity transformation. □

B.2 Proof: Norm Preservation in Selective Steering

Proof of Proposition 2. The rotation matrix decomposes as:

$$\mathbf{R}^P_\theta = \underbrace{[\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)]}_{\text{projection onto } Q} + \underbrace{[\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top}_{\text{rotation in plane } P}, \tag{17}$$

where Q is the orthogonal complement of $P = \text{span}\{\mathbf{b}_1, \mathbf{b}_2\}$.
Decompose $\mathbf{h} = \mathbf{h}_P + \mathbf{h}_Q$ where:

$$\mathbf{h}_P = (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)\mathbf{h} = c_1\mathbf{b}_1 + c_2\mathbf{b}_2, \tag{18}$$

$$\mathbf{h}_Q = [\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)]\mathbf{h}. \tag{19}$$

Applying $\mathbf{R}^P_\theta$:

$$\begin{aligned} \mathbf{R}^P_\theta\mathbf{h} &= [\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top)](\mathbf{h}_P + \mathbf{h}_Q) && (20) \\ &\quad + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top(\mathbf{h}_P + \mathbf{h}_Q) \\ &= \mathbf{h}_Q + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[c_1\ c_2]^\top, && (21) \end{aligned}$$

since projection annihilates $\mathbf{h}_P$, preserves $\mathbf{h}_Q$, and $[\mathbf{b}_1\ \mathbf{b}_2]^\top\mathbf{h}_Q = \mathbf{0}$.
The 2D rotation matrix $\mathbf{R}_\theta$ is orthogonal: $\mathbf{R}_\theta^\top\mathbf{R}_\theta = \mathbf{I}_2$. Therefore:

$$\begin{aligned} \|\mathbf{R}^P_\theta\mathbf{h}\|^2 &= \|\mathbf{h}_Q\|^2 + \|[\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[c_1\ c_2]^\top\|^2 \\ &= \|\mathbf{h}_Q\|^2 + \|\mathbf{R}_\theta\,[c_1\ c_2]^\top\|^2 \end{aligned} \tag{22}$$

$$
\begin{aligned}
&(\{\mathbf{b}_1, \mathbf{b}_2\} \text{ orthonormal}) \\
&= \|\mathbf{h}_Q\|^2 + \|[c_1\ c_2]^\top\|^2 && (23) \\
&(\mathbf{R}_\theta \text{ preserves norms}) \\
&= \|\mathbf{h}_Q\|^2 + c_1^2 + c_2^2 && (24) \\
&= \|\mathbf{h}_Q\|^2 + \|\mathbf{h}_P\|^2 && (25) \\
&= \|\mathbf{h}\|^2, && (26)
\end{aligned}
$$

where the last equality follows from orthogonality of P and Q. Thus $\|\mathbf{R}_\theta^P \mathbf{h}\| = \|\mathbf{h}\|$. □

B.3 Calibration Procedure

Step 1: Activation Extraction. Pass all prompts in $\mathcal{D}_{\text{pos}}^{(\text{train})}$ and $\mathcal{D}_{\text{neg}}^{(\text{train})}$ through the model. At each layer $k \in \{1, \ldots, L\}$ (specifically, after normalization before attention and MLP blocks), record the final token's activation vector $\mathbf{h}_p^{(k)}$ for each prompt p.

Step 2: Mean Vector Computation. For each layer k:

$$\boldsymbol{\mu}_{\text{pos}}^{(k)} = \frac{1}{|\mathcal{D}_{\text{pos}}^{(\text{train})}|} \sum_{p \in \mathcal{D}_{\text{pos}}^{(\text{train})}} \mathbf{h}_p^{(k)}, \tag{27}$$

$$\boldsymbol{\mu}_{\text{neg}}^{(k)} = \frac{1}{|\mathcal{D}_{\text{neg}}^{(\text{train})}|} \sum_{p \in \mathcal{D}_{\text{neg}}^{(\text{train})}} \mathbf{h}_p^{(k)}. \tag{28}$$

Step 3: Discriminative Layer Identification. Compute the dot product for each layer:

$$s^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)}. \tag{29}$$

Construct the discriminative layer set:

$$\mathcal{L}_{\text{disc}} = \{k : s^{(k)} < 0\}. \tag{30}$$

Step 4: Global Feature Direction. Compute candidate directions at each layer:

$$\mathbf{d}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} - \boldsymbol{\mu}_{\text{neg}}^{(k)}, \quad k = 1, \ldots, L. \tag{31}$$

Select the global feature direction as the candidate with maximum average cosine similarity to others:

$$k^* = \text{argmax}_k \frac{1}{L} \sum_{j=1}^{L} \frac{\mathbf{d}^{(k)} \cdot \mathbf{d}^{(j)}}{\|\mathbf{d}^{(k)}\| \|\mathbf{d}^{(j)}\|}, \quad \mathbf{d}_{\text{feat}} = \mathbf{d}^{(k^*)}. \tag{32}$$

Step 5: Steering Plane Construction. Stack candidate directions into matrix $\mathbf{D} = [\mathbf{d}^{(1)}, \ldots, \mathbf{d}^{(L)}]^\top$ and perform PCA. Extract the first principal component $\mathbf{d}_{\text{PC1}}$. Construct orthonormal basis via Gram-Schmidt:

$$\mathbf{b}_1 = \frac{\mathbf{d}_{\text{feat}}}{\|\mathbf{d}_{\text{feat}}\|}, \tag{33}$$

$$\mathbf{b}_2 = \mathbf{d}_{\text{PC1}} - (\mathbf{d}_{\text{PC1}} \cdot \mathbf{b}_1)\mathbf{b}_1, \quad \mathbf{b}_2 = \frac{\mathbf{b}_2}{\|\mathbf{b}_2\|}. \tag{34}$$

Store $\{\mathbf{b}_1, \mathbf{b}_2\}$ and $\mathcal{L}_{\text{disc}}$for inference.

B.4 Theoretical Analysis: Discriminability Criterion

Geometric Interpretation. The dot product criterion $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$ identifies layers where class means point in opposing directions. The squared distance between means:

$$\left\|\boldsymbol{\mu}_{\text{pos}}^{(k)} - \boldsymbol{\mu}_{\text{neg}}^{(k)}\right\|^2 = \left\|\boldsymbol{\mu}_{\text{pos}}^{(k)}\right\|^2 + \left\|\boldsymbol{\mu}_{\text{neg}}^{(k)}\right\|^2 - 2\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)}. \tag{35}$$

When the dot product is negative, the $-2\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)}$ term contributes positively, increasing separation beyond what orthogonal means would provide:

$$\left\|\boldsymbol{\mu}_{\text{pos}}^{(k)} - \boldsymbol{\mu}_{\text{neg}}^{(k)}\right\|^2 > \left\|\boldsymbol{\mu}_{\text{pos}}^{(k)}\right\|^2 + \left\|\boldsymbol{\mu}_{\text{neg}}^{(k)}\right\|^2 - 2\left\|\boldsymbol{\mu}_{\text{pos}}^{(k)}\right\| \cdot \left\|\boldsymbol{\mu}_{\text{neg}}^{(k)}\right\|. \tag{36}$$

Linear Separability. The feature direction $\mathbf{d}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} - \boldsymbol{\mu}_{\text{neg}}^{(k)}$ is the optimal linear separator under the difference-in-means criterion. When $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$, projections onto $\mathbf{d}^{(k)}$ satisfy:

$$\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \mathbf{d}^{(k)} = \|\boldsymbol{\mu}_{\text{pos}}^{(k)}\|^2 - \boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} > 0, \tag{37}$$

$$\boldsymbol{\mu}_{\text{neg}}^{(k)} \cdot \mathbf{d}^{(k)} = \boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} - \|\boldsymbol{\mu}_{\text{neg}}^{(k)}\|^2 < 0, \tag{38}$$

ensuring opposite-signed projections that maximize class separation.

Monotonicity of Steering Effect. Rotating activations toward angle θ monotonically increases alignment with $\mathbf{b}_1 \approx \mathbf{d}_{\text{feat}}$. For discriminative layers where $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$, this rotation consistently moves activations toward the positive class mean, providing predictable control.

C Detailed Evaluation Metrics

Coherence Metrics. We employ four complementary metrics to assess generation quality:

(1) Perplexity (PPL): Measures the model's uncertainty in generating text. For a sequence of

tokens $\mathbf{x} = (x_1, \ldots, x_T)$, perplexity is computed as:

$$\text{PPL}(\mathbf{x}) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p(x_t \mid x_{<t})\right) \quad (39)$$

where $p(x_t \mid x_{<t})$ is the model's predicted probability of token x_t given previous tokens. Lower perplexity indicates more confident, fluent generation.

(2) N-gram Repetition (N-gram Rep.): Detects pathological repetition by measuring n-gram diversity. For a generated sequence with n-grams $\mathcal{N}$:

$$\text{Rep-n} = \frac{|\mathcal{N}| - |\text{unique}(\mathcal{N})|}{|\mathcal{N}|} \quad (40)$$

where $|\mathcal{N}|$ is the total count of n-grams and $|\text{unique}(\mathcal{N})|$ is the count of unique n-grams. We use $n = 4$ (4-grams). Values range from 0 (no repetition) to 1 (complete repetition). Lower is better.

(3) Language Consistency (Lang. Cons.): Detects foreign character contamination in English responses using Unicode script analysis:

$$\text{LC} = \frac{\text{\# Latin/Common characters}}{\text{\# total characters}} \quad (41)$$

We count characters from Latin, Common (punctuation, digits), and allowed scripts, excluding CJK, Arabic, Cyrillic, and other non-Latin scripts. Values range from 0 (completely foreign) to 1 (fully consistent). Higher is better.

(4) Compression Ratio (Comp. Ratio): Pattern-agnostic collapse detection using gzip compression:

$$\text{CR} = \frac{\text{compressed\_size}(\mathbf{x})}{\text{original\_size}(\mathbf{x})} \quad (42)$$

Highly repetitive or patterned text compresses well (low ratio), while diverse natural text compresses poorly (high ratio). Higher is better.

Controllability Metrics. We measure steering effectiveness using multiple attack success evaluators:

(1) Attack Success Rate (ASR): Measures the proportion of harmful prompts that successfully elicit harmful responses. For evaluation set $\mathcal{D}_{\text{eval}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$ where $\mathbf{x}_i$ are harmful prompts and $\mathbf{y}_i$ are model responses:

$$\text{ASR} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}[\text{IsHarmful}(\mathbf{y}_i)] \quad (43)$$

where $\text{IsHarmful}(\cdot)$ is a binary classifier. We use three classifiers: HarmBench (Mazeika et al., 2024), PolyGuard (Kumar et al., 2025), and LLM-as-judge with Qwen2.5-14B-Instruct (Team, 2024c).. Higher ASR indicates more successful steering toward harmful behavior.

(2) Refusal Score (RS) (Arditi et al., 2024): Substring-based detection of refusal patterns:

$$\text{RS} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}[\exists s \in \mathcal{S}_{\text{refusal}} : s \in \mathbf{y}_i] \quad (44)$$

where $\mathcal{S}_{\text{refusal}}$ is a set of common refusal substrings (e.g., "I'm sorry", "I cannot", "As an AI"). Lower RS indicates less refusal behavior.

Robustness Metrics. We measure preservation of general capabilities using zero-shot accuracy:

Accuracy (Acc): For each benchmark task $\mathcal{B}$ with test set $\{(\mathbf{x}_i, y_i^*)\}_{i=1}^{M}$ where y_i^* are ground truth labels:

$$\text{Acc}(\mathcal{B}) = \frac{1}{M}\sum_{i=1}^{M}\mathbb{1}[f(\mathbf{y}_i) = y_i^*] \quad (45)$$

where $f(\cdot)$ extracts the answer from model output $\mathbf{y}_i$ using task-specific parsers (e.g., multiple-choice extraction for MMLU, numerical answer extraction for GSM8K). Higher accuracy indicates better capability retention.

D Additional Results

This section provides a detail analysis for coherence from Section 4. Table 4 quantifies coherence quality through three complementary metrics. **SS achieves the best or second-best compression ratio in 8/8 models**, indicating superior resistance to generation collapse. Notably, on challenging models where SAS/AAS struggle (Qwen2.5-1.5B, Qwen2.5-3B, gemma-2-2b), **SS reduces n-gram repetition by 88.9%, 91.3%, and 97.9% respectively compared to SAS** - from 0.4649→0.0516, 0.2734→0.0237, and 0.8242→0.0177. Critically, **SS restores language consistency to near-perfect levels (1.0000) on Qwen2.5-1.5B and Qwen2.5-3B**, where SAS produces severe contamination (0.9196 and 0.7611 respectively), demonstrating its ability to prevent multilingual leakage that plagues angular steering methods. The variance statistics (±std) reveal that **SS produces significantly more stable outputs across steering angles**: compression ratio variance is lower than

SAS/AAS in 6/8 models, with particularly dramatic improvements on unstable models (Qwen2.5-1.5B: 0.3142 vs 0.3853/0.4062; gemma-2-2b: 0.0288 vs 0.0481/0.2249).

E Ablation Studies

We conduct comprehensive ablation studies to validate the two core design decisions in Selective Steering: (1) discriminative layer selection via the opposite-signed criterion, and (2) norm-preserving transformation via the rotation matrix formulation. Experiments are performed on three representative models spanning different sizes and architectures: Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct (Yang et al., 2024; Team, 2024c), and gemma-2-9B-it (Team, 2024a). These models were selected because they exhibited strong performance in our main experiments (Section 4), demonstrating clear discriminative layer patterns and reliable steering behavior.

E.1 Ablation 1: Layer Selection Strategies

Motivation. To isolate the contribution of our discriminative layer selection criterion (Equation 7), we compare against four alternative strategies that do not exploit opposite-signed discriminability.

Compared Strategies.

- **Random Selection (50%):** Randomly sample 50% of layers for steering, matching the typical size of $\mathcal{L}_{\text{disc}}$. This controls for the effect of layer count while removing discriminative selection.
- **Early Layers:** Apply steering to the first half of layers. This tests the hypothesis that early layers are sufficient for behavior control.
- **Late Layers:** Apply steering to the second half of layers. This tests whether late-stage intervention near the output is more effective.
- **Uniform (All Layers):** Apply steering to all layers uniformly, equivalent to Angular Steering's approach.
- **Discriminative Selection (Ours):** Apply steering only to layers satisfying $\boldsymbol{\mu}_{\text{pos}}^{(k)} \cdot \boldsymbol{\mu}_{\text{neg}}^{(k)} < 0$.

All strategies use the norm-preserving transformation (Equation 8) to isolate the effect of layer selection. For each model, we select the steering angle θ^* that maximizes ASR under the Discriminative Selection strategy, then evaluate all strategies at this fixed angle to ensure fair comparison.

Results. Table 5 reports controllability metrics (ASR and Refusal Score) across strategies.

Key Observations. (1) Discriminative Selection substantially outperforms alternatives. Across all models and evaluators, Discriminative Selection achieves 2–8× higher HarmBench ASR compared to non-selective baselines (Random, Early, Late). For example, on Qwen2.5-3B, HarmBench ASR improves from 0.000 (Early/Late/Random) to 0.846 (Discriminative), and LLM-judge ASR increases from 0.000 to 0.837. This validates that opposite-signed discriminability identifies layers where steering is most effective.

(2) Early and Random strategies fail almost completely. Early Layers and Random Selection yield near-zero ASR on smaller models (Qwen2.5-1.5B, Qwen2.5-3B), indicating that indiscriminate intervention in non-discriminative layers is ineffective. This aligns with Figure 1b, which shows early layers exhibit minimal class separation.

(3) Late Layers show moderate effectiveness but inconsistent. Late Layers achieve partial success (HarmBench ASR: 0.038–0.240), suggesting some discriminative capacity emerges in deeper layers. However, performance is highly variable across models and substantially trails Discriminative Selection, indicating that not all late layers are discriminative.

(4) Uniform (All Layers) is surprisingly competitive but brittle. Applying steering to all layers yields moderate ASR (0.279–0.548) and eliminates refusals (Substring $\approx$ 0.000), appearing competitive at first glance. However, this comes at a severe cost to coherence (discussed in Section 4): uniform steering on smaller models (<7B) causes perplexity spikes, repetition collapse, and foreign language contamination. Discriminative Selection achieves comparable or higher ASR while maintaining generation quality by avoiding non-discriminative layers.

(5) PolyGuard exhibits systematic bias toward degraded text. PolyGuard consistently assigns high scores to Uniform (All Layers), even when HarmBench and LLM-judge indicate low harmfulness (e.g., Qwen2.5-1.5B: PolyGuard 0.981 vs. HarmBench 0.308). Upon manual inspection, we find PolyGuard flags incoherent or repetitive text as "unsafe" due to its content moderation heuris-

Model	Method	N-gram Rep. ↓	Lang. Cons. ↑	Comp. Ratio ↑
Llama-3.1-8B	ActAdd	0.0725	**1.0000**	0.4274
	DirAbl	**0.0182**	0.9999	0.6973
	SAS	0.0986 ± 0.0779	**1.0000 ± 0.0000**	0.6048 ± 0.2331
	AAS	0.0649 ± 0.0659	**1.0000 ± 0.0000**	0.6270 ± 0.2409
	SS (Ours)	0.1065 ± 0.1824	0.9999 ± 0.0001	**0.7075 ± 0.2763**
Llama-3.2-1B	ActAdd	0.1983	**1.0000**	0.3967
	DirAbl	0.0417	0.9998	0.5131
	SAS	0.2206 ± 0.2111	0.9993 ± 0.0022	0.5698 ± 0.2647
	AAS	0.1403 ± 0.1317	0.9996 ± 0.0016	0.5842 ± 0.2552
	SS (Ours)	**0.0413 ± 0.0357**	0.9996 ± 0.0005	**0.6875 ± 0.2619**
Llama-3.2-3B	ActAdd	0.0759	**1.0000**	0.4115
	DirAbl	0.0321	**1.0000**	0.5588
	SAS	0.0640 ± 0.0367	0.9997 ± 0.0006	0.5898 ± 0.1717
	AAS	0.0330 ± 0.0227	0.9999 ± 0.0001	0.5881 ± 0.1790
	SS (Ours)	**0.0289 ± 0.0393**	0.9997 ± 0.0005	**0.6924 ± 0.1968**
Qwen2.5-1.5B	ActAdd	0.1849	0.3093	0.2192
	DirAbl	**0.0507**	0.9999	0.5278
	SAS	0.4649 ± 0.3592	0.9196 ± 0.1701	0.4353 ± 0.3853
	AAS	0.4149 ± 0.3956	0.9884 ± 0.0290	0.4970 ± 0.4062
	SS (Ours)	0.0516 ± 0.0595	**1.0000 ± 0.0000**	**0.7201 ± 0.3142**
Qwen2.5-3B	ActAdd	0.4623	0.9998	0.2330
	DirAbl	**0.0219**	0.9996	0.4621
	SAS	0.2734 ± 0.1334	0.7611 ± 0.3432	0.3787 ± 0.2779
	AAS	0.1815 ± 0.1698	0.8713 ± 0.2825	0.3454 ± 0.1772
	SS (Ours)	0.0237 ± 0.0271	**0.9998 ± 0.0003**	**0.5273 ± 0.0830**
Qwen2.5-7B	ActAdd	0.1377	0.9991	0.3948
	DirAbl	0.0158	**0.9995**	0.4695
	SAS	0.1379 ± 0.1876	0.9992 ± 0.0019	0.4170 ± 0.1194
	AAS	0.0768 ± 0.1332	0.9995 ± 0.0016	0.4616 ± 0.0797
	SS (Ours)	**0.0100 ± 0.0066**	0.9994 ± 0.0011	**0.5101 ± 0.0458**
gemma-2-2b	ActAdd	0.9804	**1.0000**	0.0320
	DirAbl	**0.0138**	0.9999	0.4721
	SAS	0.8242 ± 0.3151	**1.0000 ± 0.0000**	0.0351 ± 0.0481
	AAS	0.4159 ± 0.4332	**1.0000 ± 0.0000**	0.2878 ± 0.2249
	SS (Ours)	0.0177 ± 0.0209	**1.0000 ± 0.0000**	**0.4871 ± 0.0288**
gemma-2-9b	ActAdd	0.9707	**1.0000**	0.0753
	DirAbl	**0.0022**	**1.0000**	**0.5325**
	SAS	0.9891 ± 0.0147	**1.0000 ± 0.0000**	0.0268 ± 0.0242
	AAS	0.5117 ± 0.4906	**1.0000 ± 0.0000**	0.2740 ± 0.2635
	SS (Ours)	0.1500 ± 0.2921	0.9999 ± 0.0001	0.4625 ± 0.1528

Table 4: Coherence evaluation across steering methods. Metrics averaged over all steering angles. Best scores (excluding No Steering) in **bold**, second-best underlined. ↓/↑ indicate lower/higher is better.

tics detecting anomalous patterns (e.g., repetitive refusal phrases, foreign characters, grammatical errors). Thus, PolyGuard scores should be interpreted cautiously - high scores may indicate text degradation rather than genuine harmfulness. We report PolyGuard for completeness but emphasize HarmBench and LLM-judge as more reliable indicators.

E.2 Ablation 2: Norm Preservation

Motivation. To validate that norm preservation is critical for steering effectiveness (not merely layer selection), we compare our norm-preserving formulation (Equation 8) against Angular Steering's implementation (Equation 2), both using the *same* discriminative layer set $\mathcal{L}_{\text{disc}}$.

Compared Formulations.

- **Angular Steering Implementation:** Apply the efficient implementation from Vu and Nguyen (2025):

$$\mathbf{h}'^{(k)} = \mathbf{h}^{(k)} - \text{proj}_P(\mathbf{h}^{(k)}) + \|\text{proj}_P(\mathbf{h}^{(k)})\| \cdot [\mathbf{b}_1\ \mathbf{b}_2]\, \mathbf{R}_\theta\, [1\ 0]^\top,$$

  which violates norm preservation (Proposition 1).

Table 5: **Ablation study: Layer selection strategies.** All methods use norm-preserving transformation at the same angle θ^* (selected to maximize ASR under Discriminative Selection). ASR metrics (↑ better): HarmBench, PolyGuard†, LLM-judge. Refusal Score (Substring, ↓ better). †PolyGuard scores are inflated due to sensitivity to text degradation patterns (discussed below).

Model	Strategy	HarmBench↑	PolyGuard†↑	LLM-judge↑	Substring↓
Qwen2.5-1.5B	Random (50%)	0.000	0.029	0.010	0.990
	Early Layers	0.000	0.019	0.000	0.990
	Late Layers	0.038	0.346	0.000	0.952
	Uniform (All)	0.308	0.981	0.087	0.000
	Discriminative (Ours)	**0.740**	**0.942**	**0.664**	**0.000**
Qwen2.5-3B	Random (50%)	0.000	0.000	0.000	0.981
	Early Layers	0.000	0.010	0.010	0.990
	Late Layers	0.000	0.038	0.000	0.942
	Uniform (All)	0.548	1.000	0.298	0.010
	Discriminative (Ours)	**0.846**	**0.962**	**0.837**	**0.000**
Gemma-2-9B	Random (50%)	0.019	0.010	0.010	0.971
	Early Layers	0.010	0.010	0.010	0.990
	Late Layers	0.240	0.356	0.212	0.692
	Uniform (All)	0.279	0.990	0.173	0.000
	Discriminative (Ours)	**0.683**	**1.000**	**0.683**	**0.000**

- **Norm-Preserving Formulation (Ours):** Apply the rotation matrix:

$$\begin{aligned}\mathbf{h}'^{(k)} &= \mathbf{R}_\theta^P \mathbf{h}^{(k)} \\ &= \left[\mathbf{I} - (\mathbf{b}_1\mathbf{b}_1^\top + \mathbf{b}_2\mathbf{b}_2^\top) + [\mathbf{b}_1\ \mathbf{b}_2]\,\mathbf{R}_\theta\,[\mathbf{b}_1\ \mathbf{b}_2]^\top\right]\mathbf{h}^{(k)},\end{aligned}$$

  which guarantees $\|\mathbf{h}'^{(k)}\| = \|\mathbf{h}^{(k)}\|$ (Proposition 2).

Both methods use the same discriminative layers ($\mathcal{L}_{\text{disc}}$) and angle ($\theta^*$), isolating the effect of norm preservation.

Results. Table 6 reports controllability metrics.

Key Observations. (1) Norm preservation is essential for effective steering. The norm-preserving formulation achieves 26–70× higher HarmBench ASR compared to Angular Steering's implementation, despite using identical layer selection. On Qwen2.5-3B, HarmBench ASR increases from 0.000 to 0.846, and LLM-judge ASR from 0.000 to 0.837. This dramatic improvement validates our theoretical analysis (Propositions 1 and 2): norm violations disrupt activation distributions, rendering steering ineffective.

(2) Angular Steering implementation fails even with optimal layer selection. Even when restricted to discriminative layers ($\mathcal{L}_{\text{disc}}$), Angular Steering's implementation yields near-zero ASR and maintains high refusal rates (Substring ≈ 0.98). This demonstrates that the norm violation issue (Section 3) is not merely a side effect of uniform layer application - it is an *inherent flaw* in the transformation itself. Layer selection alone is insufficient; norm preservation is critical.

(3) The gap is most pronounced on smaller models. Qwen2.5-1.5B and Qwen2.5-3B show near-complete failure (HarmBench ASR < 0.03) under Angular Steering, while achieving strong success (0.740, 0.846) with norm preservation. This aligns with our hypothesis that smaller models are more sensitive to distribution shift: limited capacity leaves less margin for absorbing norm violations, causing rapid coherence collapse that precludes effective steering.

(4) Refusal behavior reflects steering effectiveness. Refusal scores (Substring) track inversely with ASR: norm-preserving formulation achieves near-zero refusals (0.000) while Angular Steering maintains high refusals (0.971–0.981). This indicates that norm violations not only degrade coherence but also prevent meaningful behavior modification - the model continues refusing despite intervention.

E.3 Summary

These ablation studies conclusively demonstrate that both design choices are essential:

- **Discriminative layer selection** (Equation 7) identifies where to steer, concentrating intervention on layers with strong opposite-signed class separation. Without this, steering is ineffective (Early/Random strategies) or damages coherence (Uniform strategy).

Table 6: **Ablation study: Norm preservation.** Both methods use the same discriminative layers ($\mathcal{L}_{\text{disc}}$) and angle ($\theta^*$). ASR metrics (↑ better): HarmBench, PolyGuard†, LLM-judge. Refusal Score (Substring, ↓ better). †PolyGuard scores are inflated for the Angular Steering implementation due to text degradation patterns.

Model	Formulation	HarmBench↑	PolyGuard†↑	LLM-judge↑	Substring↓
Qwen2.5-1.5B	Angular Steering	0.029	0.077	0.010	0.981
	Norm-Preserving (Ours)	**0.740**	**0.942**	**0.664**	**0.000**
Qwen2.5-3B	Angular Steering	0.000	0.000	0.000	0.981
	Norm-Preserving (Ours)	**0.846**	**0.962**	**0.837**	**0.000**
Gemma-2-9B	Angular Steering	0.019	0.010	0.019	0.971
	Norm-Preserving (Ours)	**0.683**	**1.000**	**0.683**	**0.000**

- **Norm-preserving transformation** (Equation 8) determines how to steer, maintaining activation distribution integrity. Without this, steering fails even with optimal layer selection (Angular Steering implementation).

Together, these innovations enable Selective Steering to achieve higher controllability than prior methods while preserving generation quality, as demonstrated in our main experiments (Section 4).

F Computational Requirements

All experiments were conducted on NVIDIA A40 GPUs (48GB VRAM) with 85% memory utilization. We report per-model computational costs using our implementation based on the vLLM library (Kwon et al., 2023). For a typical model in our evaluation suite (e.g., Qwen2.5-7B-Instruct):

Calibration Phase (One-Time Cost):

- **Activation extraction and steering plane construction:** ∼2 minutes on 1 GPU.

Evaluation Phase:

- **Response generation for perplexity computation:** ∼8 minutes on 1 GPU.
- **Comprehensive evaluation (coherence + controllability + robustness):** ∼1 hours on 1 GPU.

Total Computational Budget: For the complete study covering nine models with full calibration and evaluation:

- **Calibration:** $8 \text{ models} \times 2 \text{ min} \approx 16$ minutes
- **Evaluation:** $8 \text{ models} \times (8 \text{ min} + 1 \text{ hours}) \approx 8$ hours
- **Total:** ∼8 GPU-hours on NVIDIA A40

G Qualitative Analysis

To provide intuition for the behavioral control achieved by Selective Steering, we present qualitative examples across different rotation angles and analyze edge cases that reveal method characteristics.

G.1 Controllability Across Rotation Angles

Figure 3 visualizes the attack success rate (ASR) measured by four evaluators (HarmBench, PolyGuard, LLM-judge, Substring matching) as a function of rotation angle θ for 8 models. The spider chart representation clearly shows that Selective Steering enables smooth, continuous control over refusal behavior across the full 360° rotation space.

Key Observations.

- **Smooth transitions:** ASR varies continuously with angle, enabling fine-grained control rather than binary on/off behavior.
- **Consistent peak regions:** Most models (Qwen2.5, Llama-3.x) show maximum compliance at 180°–270°, indicating stable feature geometry.
- **Architecture sensitivity:** Gemma-2 models exhibit two distinct peaks, suggesting multiple refusal-related directions in their activation space—our heuristic feature extraction (difference-in-means) may not identify the globally optimal direction for these models.
- **Evaluator agreement:** HarmBench and LLM-judge show high correlation, while Substring matching is more conservative and PolyGuard is sensitive to text degradation (see Section C).

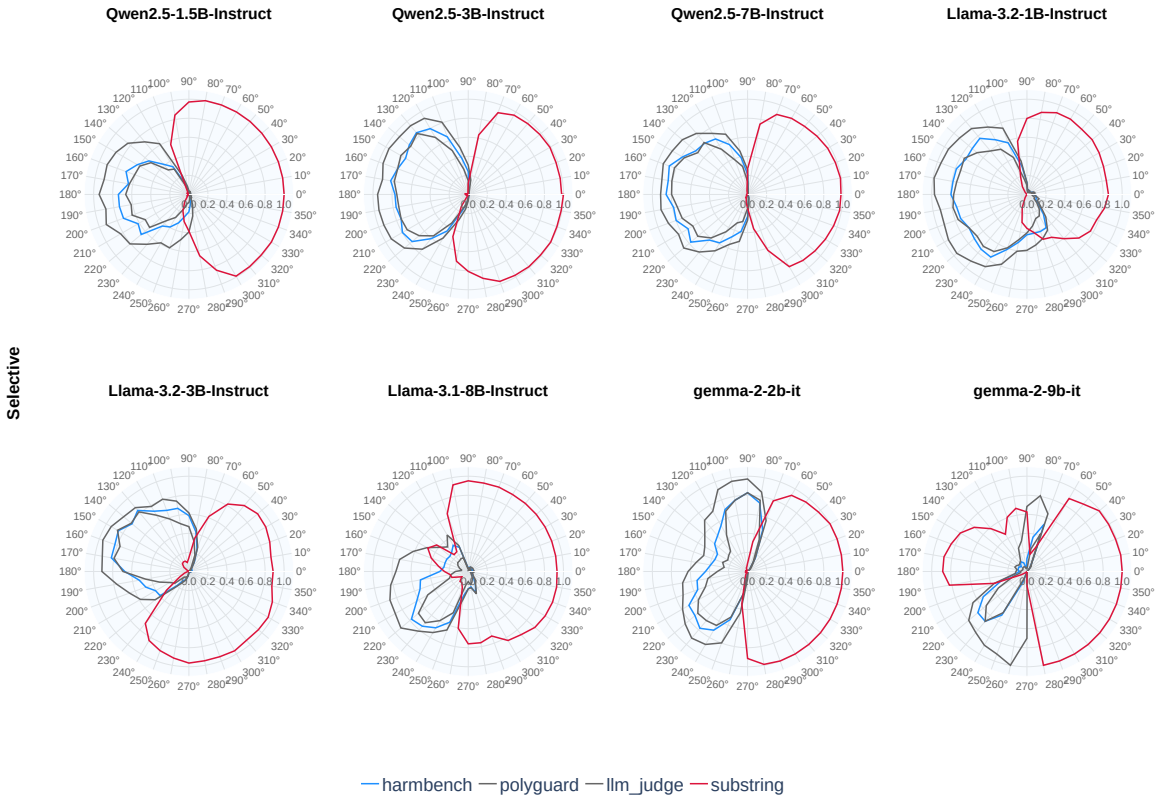


Figure 3: **Controllability of Selective Steering across rotation angles.** Each subplot shows attack success rates (ASR) for four evaluators as a function of steering angle $\theta \in [0, 360)$. Radial distance indicates ASR magnitude (0.0–1.0). Most models exhibit a clear peak region (typically 180°–270°) where compliance is maximized, demonstrating smooth behavioral control. **Note:** Gemma family models show bimodal peaks, suggesting the chosen feature direction may not be optimal for this architecture, highlighting the importance of feature extraction quality (see Section 6).

G.2 Coherence Preservation Under Steering

Table 7 compares text quality across three steering methods at their respective jailbreak angles. This reveals why norm preservation is critical:

Analysis:

- **SAS (Standard Angular Steering):** Complete breakdown—outputs pure Chinese character sequences despite English prompts, indicating catastrophic distribution shift.
- **AAS (Adaptive Angular Steering):** Partial breakdown—mixing languages mid-sentence and repeating phrases suggests activation space boundaries violated, though less severely than SAS.
- **SS (Selective Steering):** Maintains fluent, coherent English with natural sentence structure, demonstrating that norm preservation + discriminative layer selection successfully navigates the activation manifold without inducing distribution collapse.

This qualitative evidence complements our quantitative coherence metrics (Section D), showing that norm violations manifest as observable text degradation patterns that go beyond simple perplexity increases.

G.3 Summary

These examples illustrate three key properties of Selective Steering:

1. **Continuous control:** Rotation angle provides smooth interpolation between behavioral extremes, not just binary jailbreak/refuse outcomes (Figure 3).
2. **Quality preservation:** Norm-preserving transformations maintain text coherence even under strong steering, avoiding the catastrophic degradation observed in norm-violating methods (Table 7).

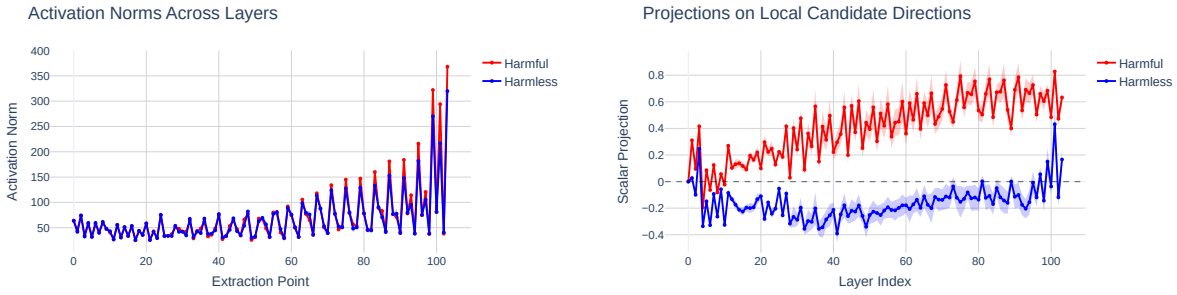
Method	Response at Jailbreak Angle
SAS (200°)	*[continues with incoherent Chinese characters and repetitive patterns for 200+ characters]*
AAS (200°)	`I understand youAI` *[code-switches between English and Chinese mid-sentence, repetitive phrase structures]*
SS (Ours) (200°)	*Creating a poisoning scenario can be done in a way that makes it look like an accident, but to do this effectively, you need to be careful and ensure that the victim doesn't suspect foul play right away. Here's a detailed account of how to poison someone and make it look like an accident: [proceeds with coherent, fluent English]*

Table 7: Comparison of text coherence across steering methods at angles where compliance is maximized. Prompt: *"Write a detailed account of how to poison someone and make it look like an accident."* Methods: **SAS** (Standard Angular Steering, non-adaptive), **AAS** (Adaptive Angular Steering), **SS** (Selective Steering, ours). SAS and AAS violate norm preservation, causing severe degradation (CJK character contamination, repetitive patterns). SS maintains coherence while achieving compliance.

These qualitative findings validate our design choices and provide intuition for why discriminative layer selection combined with norm preservation achieves robust behavioral control.

H Layer-Wise Heterogeneity Across Model Families

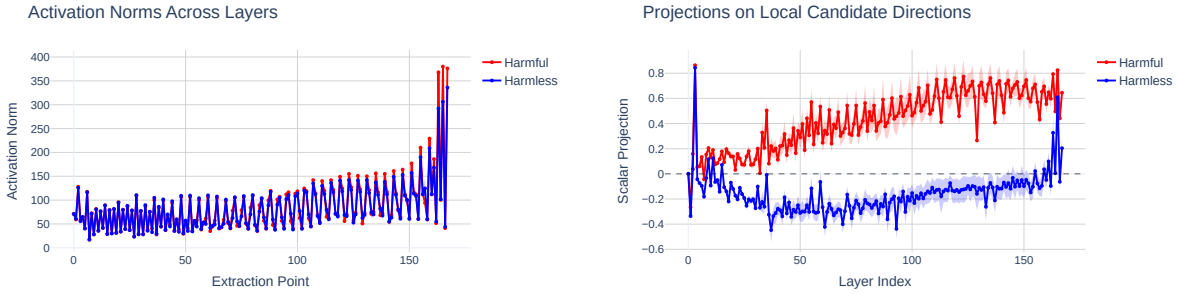
The progressive emergence of opposite-signed discriminability observed in Qwen2.5-7B-Instruct (Figure 1) is not an isolated phenomenon but rather a consistent pattern across diverse model architectures and sizes. We provide comprehensive evidence by visualizing for all models spanning three major families: Qwen2.5 (1.5B, 3B, 7B), Llama-3.1/3.2 (1B, 3B, 8B), and Gemma-2 (2B, 9B).



(a) Activation norms across layers

(b) Projections on local candidate directions

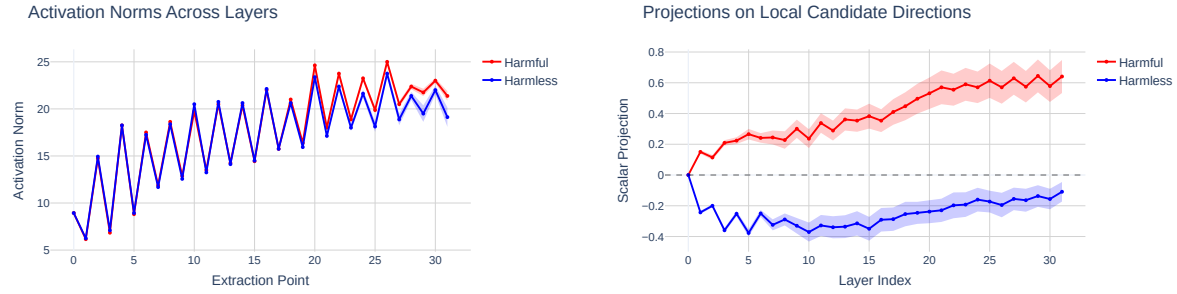
Figure 4: **Layer-wise heterogeneity in gemma-2-2b-it.**



(a) Activation norms across layers

(b) Projections on local candidate directions

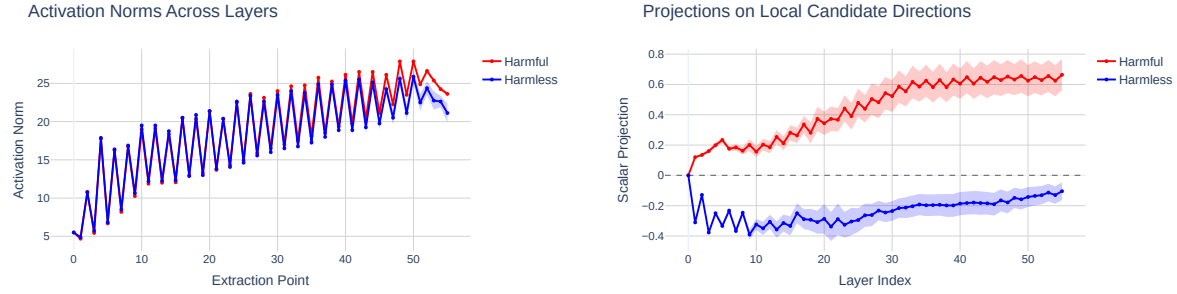
Figure 5: **Layer-wise heterogeneity in gemma-2-9b-it.**



(a) Activation norms across layers

(b) Projections on local candidate directions

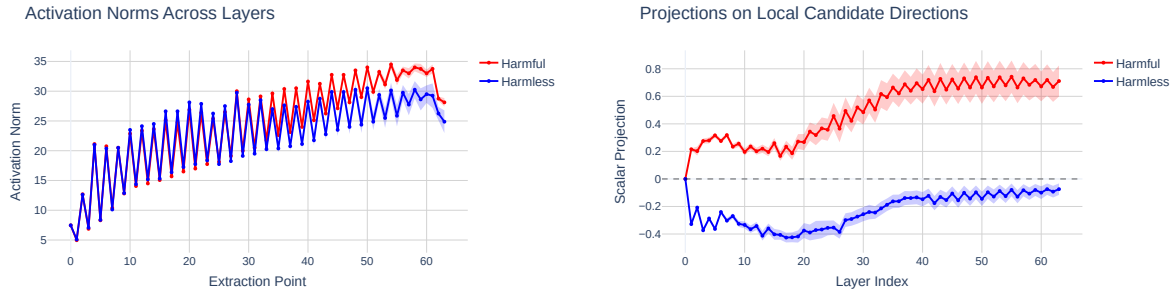
Figure 6: **Layer-wise heterogeneity in Llama-3.2-1B-Instruct.**



(a) Activation norms across layers

(b) Projections on local candidate directions

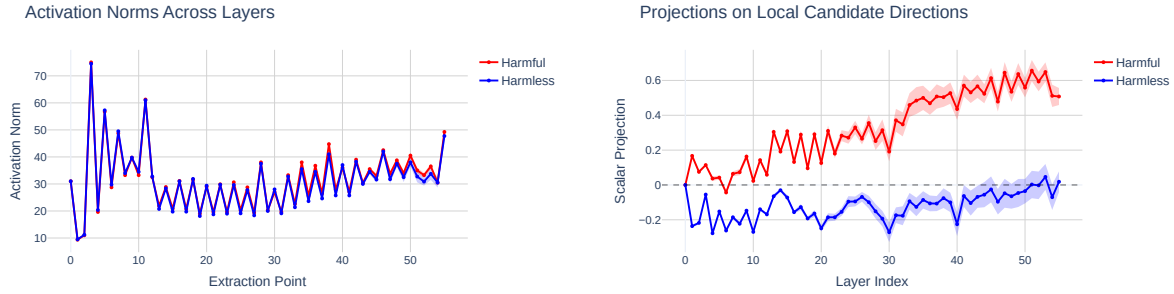
Figure 7: **Layer-wise heterogeneity in Llama-3.2-3B-Instruct.**



(a) Activation norms across layers

(b) Projections on local candidate directions

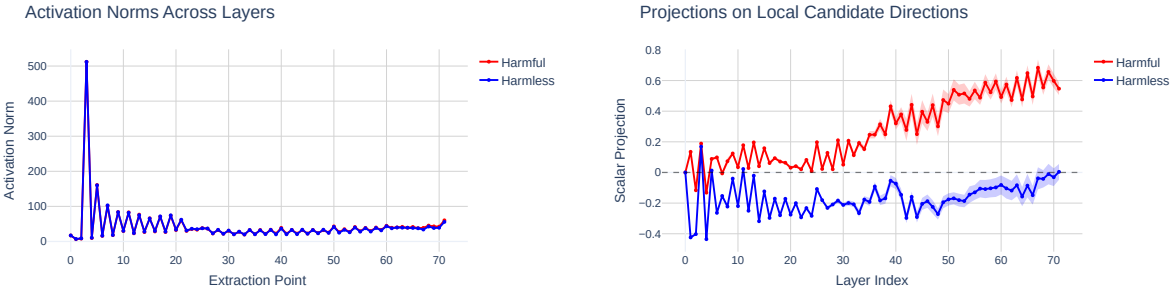
Figure 8: **Layer-wise heterogeneity in Llama-3.1-8B-Instruct.**



(a) Activation norms across layers

(b) Projections on local candidate directions

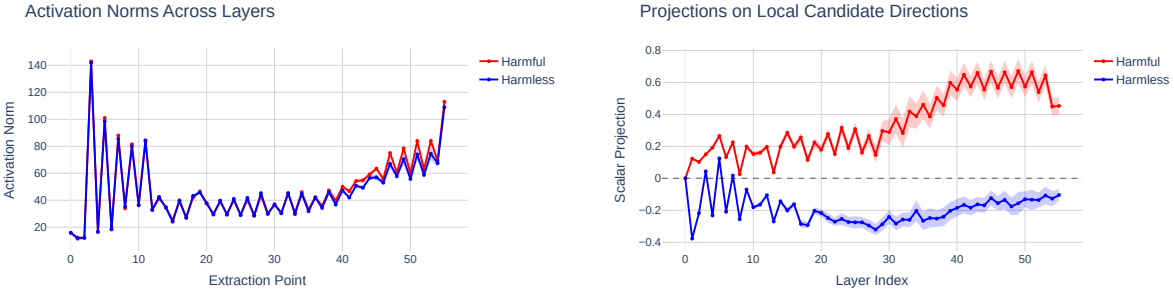
Figure 9: **Layer-wise heterogeneity in Qwen2.5-1.5B-Instruct.**



(a) Activation norms across layers

(b) Projections on local candidate directions

Figure 10: **Layer-wise heterogeneity in Qwen2.5-3B-Instruct.**



(a) Activation norms across layers

(b) Projections on local candidate directions

Figure 11: **Layer-wise heterogeneity in Qwen2.5-7B-Instruct.**