Selective Steering: Norm-Preserving Control Through Discriminative Layer Selection

Quy-Anh Dang[1,2], **Chris Ngo**[2]
[1]VNU University of Science, Vietnam
[2]Knovel Engineering Lab, Singapore
{quyanh.dang, chris.ngo}@knoveleng.com

Project: https://knoveleng.github.io/steering/

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**[1], 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.

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).

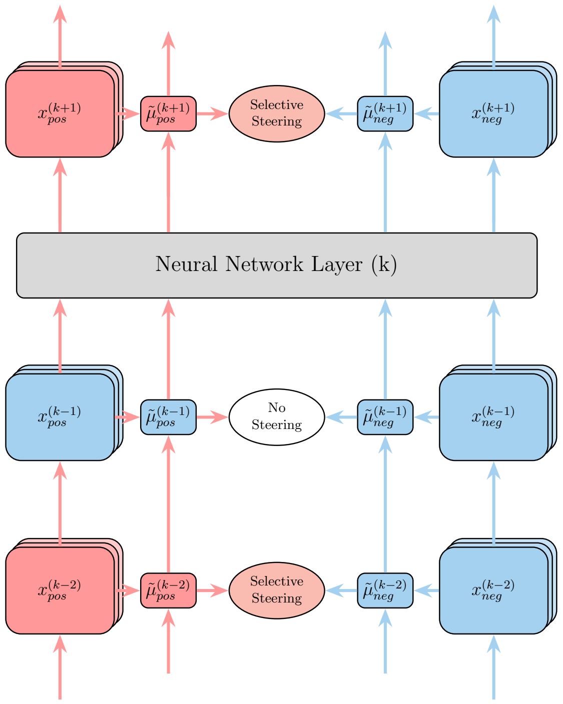

Figure 1: **Selective Steering pipeline.** At each layer k, we compute projections of positive (red) and negative (blue) class means onto the selected feature direction (red/blue boxes). Steering is applied only at layers where projections have opposite signs (layers $k-2$ and $k+1$), using norm-preserving rotation. Layers with same-sign projections (layer $k-1$) remain unchanged.

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

[1] **Code:** https://github.com/knoveleng/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*: $\tilde{\boldsymbol{\mu}}^{(k)}_{\text{pos}} \cdot \tilde{\boldsymbol{\mu}}^{(k)}_{\text{neg}}$. 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:

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

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 2 (More in Appendix H) reveals two critical phenomena:

Non-uniform Norm Profiles. Figure 2a 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 2b shows scalar projections of normalized activations onto the chosen direction $\hat{\mathbf{d}}_{\text{feat}}$, 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**: The separation persists but weakens as the strength decreases.

Key Insight. Layers where $\tilde{\boldsymbol{\mu}}_{\text{pos}}^{(k)} \cdot \tilde{\boldsymbol{\mu}}_{\text{neg}}^{(k)} < 0$ (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.