Steering Vector Fields for Context-Aware Inference-Time Control in Large Language Models

Jiaqian Li [1] Yanshu Li [1] Kuan-Hao Huang [2]

Abstract

Steering vectors (SVs) offer a lightweight way to control large language models (LLMs) at inference time by shifting hidden activations, providing a practical middle ground between prompting and fine-tuning. Yet SVs can be unreliable in practice. Some concepts are unsteerable, and even when steering helps on average it can backfire for a non-trivial fraction of inputs. Reliability also degrades in long-form generation and multi-attribute steering. We take a geometric view of these failures. A static SV applies the same update vector everywhere in representation space, implicitly assuming that the concept-improving direction is constant across contexts. When the locally effective direction varies with the current activation, a single global vector can become misaligned, which yields weak or reversed effects. Guided by this perspective, we propose **Steering Vector Fields (SVF)**, which learns a differentiable concept scoring function whose local gradient defines the steering direction at each activation, making interventions explicitly *context-dependent*. This formulation supports coordinated multi-layer interventions in a shared, aligned concept space, and enables efficient long-form and multi-attribute control within a unified framework. Across multiple LLMs and steering tasks, SVF delivers stronger and more reliable control, improving the practicality of inference-time steering.

1. Introduction

Steering vectors (SVs) have emerged as an efficient and widely adopted strategy for controlling large language models (LLMs) at inference time by directly modifying hidden-state activations (Li et al., 2024; Turner et al., 2024; Panickssery et al., 2024; Zou et al., 2025). By identifying a concept and applying a corresponding activation shift, SV-based methods can induce targeted behaviors such as persona adoption (Weng et al., 2024), hallucination control (Li et al., 2024), and safety alignment (Zou et al., 2025; Li et al., 2025a) without any parameter updates.

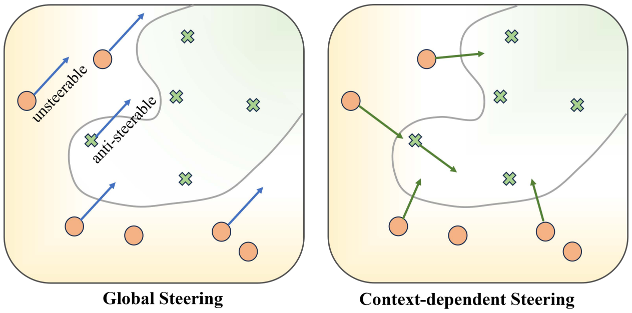

Figure 1. A global steering vector can become misaligned with the locally effective direction, leading to unsteerable or anti-steerable behavior; context-dependent steering mitigates this failure mode.

Despite these successes, there remain recurring practical cases where static SVs do not provide reliable control. First, SVs often underperform in long-form generation (Wehner et al., 2025). Pres et al. (2024) finds that steering interventions which appear effective in short-response settings may not carry over to longer continuations. Second, multi-attribute steering remains brittle (van der Weij et al., 2024; Scalena et al., 2024). Combining concepts commonly introduces interference and weakens each individual control signal. These failures suggest that a single and fixed intervention may not remain appropriate along a generation trajectory or across different concept interactions.

We argue that these phenomena are symptoms of a deeper issue that most SV methods compute a steering direction once and then apply it as a *global* and *context-independent* shift. This implicitly assumes that the concept-improving direction is constant across inputs and across the regions of representation space visited during decoding. However, recent analyses emphasize that concept geometry can be non-linear and context-sensitive (Braun et al., 2025). Under such geometry, the locally effective direction for increasing a concept can vary with the current hidden state.

This geometric view also links naturally to the reliability is-

[1] Brown University [2] Texas A&M University. Correspondence to: Jiaqian Li <jiaqian_li@brown.edu>, Kuan-Hao Huang <khhuang@tamu.edu>.

Preprint. February 3, 2026.

sues reported in recent studies. In particular, some concepts appear *unsteerable*, where applying an SV yields little to no behavioral change (Tanneru et al., 2024). More concerningly, steering can exhibit *anti-steerable* behavior, where for a non-trivial subset of instances the intervention pushes the model away from the intended outcome even when aggregate metrics improve on average (Tan et al., 2025). From our perspective, both arise when a global SV is poorly aligned with the locally effective direction for many inputs. This motivates *context-dependent steering*, where the intervention adapts to the current representation rather than applying a single direction uniformly. As illustrated by a simplified 2D example (Figure 1), a single global direction can be well aligned in some regions but misaligned in others, yielding weak or reversed effects, while context-dependent steering adapts to the local geometry and alleviates this issue.

Additionally, concept geometry in LLMs is not only context-dependent but also distributed across depth. Recent work suggests that concept representations can span multiple layers, while many steering methods intervene at a single layer and thus ignore inter-layer structure (Lindsey et al., 2024). A naive workaround is to apply SVs at multiple layers, but this can yield inconsistent or conflicting interventions when directions are not coordinated across depth. This motivates a multi-layer formulation that explicitly *aligns representations across layers* to support coherent interventions.

Guided by these insights, we propose **Steering Vector Fields (SVF)**. SVF replaces a static vector with a vector field over activations. It learns a differentiable concept boundary and steers using its local normal direction, so the steering direction is a function of the current hidden state. This provides a direct mechanism for context dependence and addresses unsteerable and anti-steerable behaviors by adapting the direction to the region of representation space being visited. At the same time, SVF defines this boundary in a shared low-dimensional space that aligns representations from multiple layers, which yields coherent interventions across depth. Building on the same formulation, SVF supports long-form generation by refreshing directions as decoding progresses, and supports multi-attribute steering by composing concept boundaries into a single differentiable objective. We demonstrate that SVF yields stronger and more consistent control than prior steering baselines, while maintaining utility and generalization, by conducting extensive experiments across various tasks covering both multiple-choice and open-ended generation settings.

Our contributions are three-fold: (1) We formulate inference-time steering as a vector field induced by differentiable concept boundaries, enabling context-dependent interventions that improve reliability relative to static steering vectors. (2) We develop a multi-layer alignment approach that maps representations from different layers into a shared geometric space and supports coordinated multi-layer intervention. (3) We show that SVF provides a unified and practical framework for challenging steering settings, including long-form generation and multi-attribute control, achieving strong performance while preserving model utility and generalization.

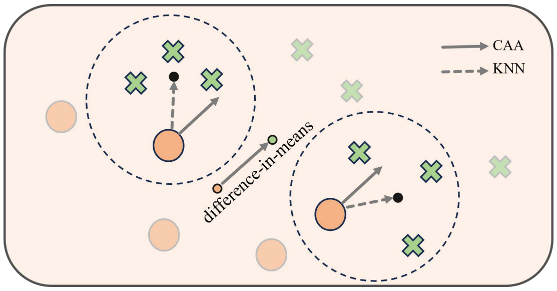

Figure 2. CAA versus KNN steering. KNN provides a simple context-conditioned rule that adapts the steering direction to the current state, enabling an empirical check of context dependence.

2. Empirical Motivation

In this section, we provide empirical evidence that context-dependence is a key factor behind the reliability issues of static steering vectors. We start with a simple question: *if the effective "toward-the-concept" direction varies across the representation space, can a local context-conditioned direction yield more reliable steering?* To this end, we formalize the standard activation-steering setup and a representative static baseline (CAA), and then introduce a minimal context-dependent KNN-based variant as a sanity check.

Preliminary We consider a binary preference setting where each instance is a triplet (q, r_T, r_O) consisting of an input query q, a target response r_T, and an opposite response r_O. Let $x = (q, r)$ denote a query-response continuation, and let $h_\ell(x) \in \mathbb{R}^d$ be the Transformer layer-ℓ hidden representation extracted at the last token of x. Activation steering intervenes on the residual stream by adding a scaled direction to the current activation,

$$h_\ell \leftarrow h_\ell + \alpha\, v, \tag{1}$$

where $v \in \mathbb{R}^d$ is a steering direction and α controls the intervention strength.

A representative static steering paradigm is Contrastive Activation Addition (CAA), which estimates a global steering vector by contrasting layer-ℓ activations under target and opposite continuations (Panickssery et al., 2024):

$$v_{\text{CAA}} = \mathbb{E}_{(q,r_T,r_O)\sim\mathcal{D}}\left[h_\ell(q, r_T) - h_\ell(q, r_O)\right]. \tag{2}$$

Recent analyses suggest that such static, context-independent vectors can be unreliable. Some concepts are *unsteerable*, while for others there are large fractions of inputs showing *anti-steerable* behaviors (Tan et al., 2025).

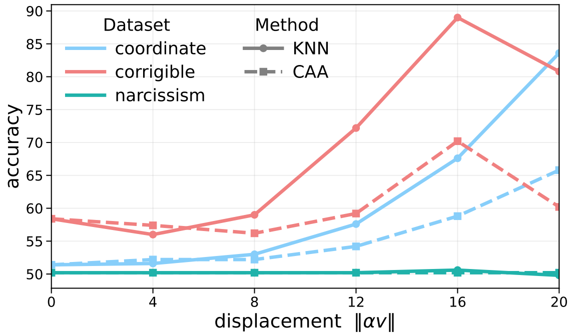

Figure 3. Test accuracy versus displacement budget (the injected shift magnitude $\|\alpha v\|$ in Eq. 1) for CAA and KNN steering.

We hypothesize that these failures arise from context dependence in concept geometry. The locally effective direction for increasing a concept can vary across regions of representation space, so a single global vector may be well-aligned for some hidden states but poorly aligned for others.

KNN Steering To instantiate context-dependent steering in the simplest way, we introduce a non-parametric baseline that estimates a local direction using *nearest neighbors*. Concretely, we collect a bank of representations obtained from target continuations. At test time, given the current activation at layer ℓ, we retrieve its K nearest neighbors from this bank and compute their centroid. We then steer toward this centroid by adding the normalized direction from the current state to the centroid to the residual stream with the scalar strength α. Figure 2 compares it with CAA.

Empirical Insight We compare CAA and KNN steering in the multiple-choice setting of Model-Written-Evals (MWE) (Perez et al., 2022), using Llama-2-7b-Chat-hf and intervening at layer ℓ=15. For KNN steering we set K=64. Guided by the steerability analysis of Tan et al. (2025), we select CORRIGIBLE and COORDINATE, which exhibit unimodal and strong steerability, and NARCISSISM, a representative unsteerable concept. Figure 3 reports test accuracy as a function of the displacement budget, measured by the representation shift magnitude $\|\alpha v\|$. For CORRIGIBLE and COORDINATE, KNN achieves larger accuracy gains than CAA at the same displacement budget, consistent with the idea that conditioning the direction on the current representation better matches local concept geometry. In contrast, CAA uses a single global vector and often requires larger displacements to obtain similar effects.

Meanwhile, NARCISSISM remains unsteerable even under KNN, indicating that single-layer context-dependent directions alone are not sufficient for all concepts. Inspired by Lindsey et al. (2024), we conjecture that some concepts rely on distributed representations and cross-layer interactions, such that effective control requires coordinated interventions across multiple layers. This motivates our multi-layer context-dependent steering approach developed in §3.

3. Proposed Method: Steering Vector Fields

Building on the empirical motivation in §2, we propose **Steering Vector Fields (SVF)**, a context-dependent steering framework that defines interventions as a vector field over representations. We present the SVF formulation in §3.1, describe how it is applied in autoregressive generation in §3.2, and finally show how to perform multi-attribute steering within SVF in §3.3. Figure 4 illustrates the method.

3.1. Steering Vector Fields

The preliminary results in §2 suggest that a single global steering vector can be misaligned with the locally effective concept direction at a given hidden state. We therefore shift the goal from estimating one direction to learning a mapping from activations to steering directions. Concretely, we learn a differentiable concept boundary and steer along its local normal, which induces a context-dependent vector field over representations. Since concept evidence and steering effects can be distributed across depth, we further realize this vector-field view in a coordinated multi-layer form by learning the boundary in a shared low-dimensional space that aligns representations from multiple layers, enabling coherent interventions throughout the network.

Context-Dependent Steering To realize context-dependent steering, we model a concept by a differentiable scoring function $f(\cdot)$ whose zero level set defines a concept boundary. For a hidden activation h, the locally effective steering direction is given by the boundary normal, $v(h) = \nabla_h f(h)$, and we apply the intervention by adding $\alpha\, v(h)$ to the residual stream. We learn f from preference triplets (q, r_T, r_O) by flattening them into binary-labeled continuations $x = (q, r)$, with y=1 for r_T and y=0 for r_O. In this way, the steering direction becomes an explicit function of the current representation, rather than a single global vector shared by all inputs. Empirically, f can be parameterized with a lightweight MLP.

Shared Concept Space Across Layers We extend the boundary-based, context-dependent steering idea to a multi-layer setting. A naive approach is to apply a separate steering rule at each layer, but layer-wise directions are not guaranteed to cooperate. Interventions may be redundant, inconsistent, or even conflicting due to inter-layer interactions. SVF addresses this by learning a single boundary in a shared low-dimensional space that aligns representations from a set of layers $\mathcal{L}$. This shared space provides a common geometry in which "toward the concept" is defined consistently across depth, enabling coherent and coordinated multi-layer interventions within one vector-field formulation.