

On the Identifiability of Steering Vectors in Large Language Models

Sohan Venkatesh and Ashish Mahendran Kurapath

Manipal Institute of Technology Bengaluru

{sohan1, ashish}.mitblr2022@learner.manipal.edu

Abstract

Activation steering methods, such as persona vectors, are widely used to control large language model behavior and increasingly interpreted as revealing meaningful internal representations. This interpretation implicitly assumes steering directions are identifiable and uniquely recoverable from input–output behavior. We formalize steering as an intervention on internal representations and prove that, under realistic modeling and data conditions, steering vectors are fundamentally non-identifiable due to large equivalence classes of behaviorally indistinguishable interventions. Empirically, we validate this across multiple models and semantic traits, showing orthogonal perturbations achieve near-equivalent efficacy with negligible effect sizes. However, identifiability is recoverable under structural assumptions including statistical independence, sparsity constraints, multi-environment validation or cross-layer consistency. These findings reveal fundamental interpretability limits and clarify structural assumptions required for reliable safety-critical control.

1 Introduction

Persona vector steering has emerged as a popular technique for controlling the behavior of large language models by adding learned directional vectors to intermediate activations. Empirically, such vectors can shift model outputs along interpretable dimensions such as politeness, political ideology or truthfulness, suggesting that representational alignment might afford fine-grained behavioral control without retraining ((Zou et al., 2023; Rinsky et al., 2024; Turner et al., 2023)). This line of work is closely connected to broader efforts in representation engineering and activation editing, where linear directions in activation space are used to modulate model behavior along semantic axes ((Elhage et al., 2022; Turner et al., 2024)).

Despite growing adoption in interpretability and alignment research, the theoretical foundations of persona steering remain poorly understood. Most existing methods implicitly assume that extracted steering vectors correspond to meaningful, uniquely determined latent factors—for example, "the politeness direction" or "the honesty direction"—and that these factors can be directly manipulated to achieve reliable control. However, classical results in latent variable modeling and causal inference show that such assumptions are often unjustified without additional structural constraints ((Hyvärinen and Pajunen, 1999; Shimizu et al., 2006; Schölkopf et al., 2021; Locatello et al., 2019)). Nonlinear ICA and causal representation learning further emphasize that recovering latent factors from high-dimensional observations is generically impossible without auxiliary information or strong inductive biases ((Hyvarinen and Morioka, 2017; Khemakhem et al., 2020; Ahuja et al., 2022)).

This raises a fundamental question for alignment and interpretability research: when does representational alignment genuinely afford reliable behavioral control and when does it merely exploit underdetermined or spurious correlations? Existing evidence from probing and representation analysis suggests that even seemingly meaningful linear directions can reflect artifacts of the measurement procedure rather than uniquely defined semantic variables ((Hewitt and Liang, 2019; Ravfogel et al., 2020; Elazar et al., 2021; Belinkov, 2022)). In the context of activation steering, these concerns become particularly acute, because the resulting vectors are often used not only for analysis but also for safety-relevant control interventions. Understanding identifiability is therefore critical for several reasons:

Alignment affordances. If persona vectors are not identifiable, then representational alignment

may provide only heuristic control rather than principled intervention. Characterizing when steering vectors are unique clarifies when alignment interventions can be trusted and when they should instead be viewed as exploiting one of many behaviorally equivalent directions.

Interpretability validity. When multiple incompatible vectors produce identical observable behavior, claims that a specific vector "represents" a semantic concept become scientifically underdetermined. Identifiability theory distinguishes well-grounded interpretability claims from artifacts of measurement and projection ((Elazar et al., 2021; Marks and Tegmark, 2023)).

Robustness and safety. Non-identifiable steering directions may rely on fragile correlations that fail under distribution shift, model updates or adversarial prompting. For safety-critical applications—where steering vectors may be used to enforce norms or prevent harmful behavior—understanding identifiability limits is essential to avoid brittle or misleading forms of control.

Methodological design. Identifiability theory clarifies which experimental protocols provide meaningful evidence and which require additional structure to support reliable conclusions. In particular, it highlights when interventions on internal activations can be interpreted causally and when they merely reparameterize an equivalence class of representations ((Schölkopf et al., 2021; Ahuja et al., 2022)).

In this work, we provide, to our knowledge, the first formal identifiability analysis of persona vector steering. Our contributions are threefold. First, we prove that under standard observational regimes, persona vectors are generically non-identifiable due to null-space ambiguity in the model’s input-output Jacobian (Proposition 1). Specifically, under white-box single-layer access, infinitely many geometrically distinct steering directions induce identical observable behavior. Second, we characterize sufficient structural conditions under which identifiability can be recovered including statistical independence constraints, sparsity priors, multi-environment access and cross-layer consistency (Proposition 2). Third, we demonstrate empirically that contemporary steering operates in the non-identifiable regime across multiple models and semantic traits, with vectors perturbed by orthogonal components achieving near-equivalent efficacy.

Crucially, our results are not purely negative. By explicitly characterizing both the limits and the affordances of persona vector steering, we provide principled guidance for when representational alignment can be trusted and which structural assumptions enable reliable control in safety-critical applications.

2 Related Work

Our work formalizes persona steering as a latent variable identification problem, bridging causal representation learning, activation editing in LLMs and mechanistic interpretability.

Causal and latent variable identifiability. Classical results show that latent variable models are generically non-identifiable without structural assumptions (Hyvärinen and Pajunen, 1999; Shimizu et al., 2006; Kruskal, 1977). Recent work in causal representation learning extends these ideas to deep learning settings (Schölkopf et al., 2021; Ahuja et al., 2022; Locatello et al., 2019), establishing conditions under which latent factors can be recovered from high-dimensional observations. Nonlinear ICA methods (Khemakhem et al., 2020; Hyvärinen and Morioka, 2017) provide theoretical foundations for recovering latent variables using temporal or auxiliary information, which standard steering methods do not exploit.

Probing and representation learning. Pimentel et al. (2020) and Ravfogel et al. (2020) demonstrate that probing classifiers can succeed for trivial reasons unrelated to the presence of target information. Elazar et al. (2021) show that removing information via projection is ill-defined without identifiability guarantees. Belinkov (2022) provides a comprehensive survey of probing methods and their limitations. The linear representation hypothesis (Park et al., 2023; Elhage et al., 2022; Marks and Tegmark, 2023) suggests that concepts correspond to directions in activation space but lacks formal identifiability guarantees.

Activation editing in LLMs. Methods such as representation engineering (Zou et al., 2023), contrastive activation addition (Rimsky et al., 2024; Turner et al., 2024), activation patching (Meng et al., 2022) and causal tracing (Wang et al., 2022) manipulate model internals to control behavior. While empirically successful, these works generally do not address whether the resulting control directions are uniquely determined. Related work

on linear mode connectivity (Entezari et al., 2021) and neural network reparameterization (Dinh et al., 2017) reveals symmetries in neural representations that motivate our null-space analysis.

Steering and control vectors. Early work on steering vectors (Li et al., 2022; Turner et al., 2023; Tigges et al., 2023) develops techniques for extracting control directions from contrastive prompt pairs. Burns et al. (2022) investigate discovering latent knowledge in language models. These methods assume that extracted vectors correspond to uniquely determined semantic factors—an assumption we formally examine.

Mechanistic interpretability. Work on mechanistic interpretability (Olsson et al., 2022; Elhage et al., 2021; Nanda et al., 2023) studies circuits and features in transformers, often assuming the existence of interpretable directions without formalizing identifiability constraints. Our theoretical analysis clarifies when such directions are well-defined.

Compressed sensing and sparse recovery. The theoretical framework of compressed sensing (Candès and Wakin, 2008; Donoho, 2006) provides conditions under which sparse vectors can be uniquely recovered from linear measurements. These results directly inform the sparsity-based identifiability conditions in Proposition 2.

3 Problem Setup

3.1 Formal Model

Consider a pre-trained transformer language model f_θ with L layers. For a given input prompt x (tokenized as x_1, \dots, x_T), let $h_\ell(x) \in \mathbb{R}^d$ denote the hidden representation at layer ℓ and position T (typically the final token position for autoregressive generation).

Latent persona variable. We assume there exists an underlying latent variable $z \in \mathcal{Z}$ representing a semantic attribute or "persona" (e.g., formality, political stance, truthfulness).

Steering intervention. A steering vector $v \in \mathbb{R}^d$ is applied as:

$$\tilde{h}_\ell(x) = h_\ell(x) + \alpha v,$$

where $\alpha \in \mathbb{R}$ is the steering strength. The modified representation \tilde{h}_ℓ is fed forward through subsequent layers to produce output logits $o(x, v, \alpha)$ over the vocabulary.

Generative model. We posit that the true data-generating process involves:

$$z \sim p(z), \quad h_\ell = g_\ell(x, z) + \epsilon,$$

where $g_\ell : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}^d$ encodes how persona z modulates the representation for prompt x and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is measurement noise. The goal of steering is to approximate the effect of varying z by adding v .

3.2 Observational Regimes

We consider two data access regimes that determine what alignment interventions can afford:

- **Regime 1: Black-box input–output.** The researcher observes only (x, y) pairs, where y is generated text. There is no access to internal representations. This is the weakest regime and corresponds to behavioral evaluation.
- **Regime 2: White-box single-layer access.** The researcher can observe or manipulate activations $h_\ell(x)$ at a chosen layer ℓ . This is the standard setting for most steering work and includes extracting vectors from contrastive prompt pairs.

Most existing work operates in Regime 2, often extracting a steering vector from contrastive prompt pairs (x^+, x^-) designed to elicit different persona values (e.g., polite versus rude instructions):

$$v \propto \mathbb{E}x^+[h_\ell(x^+)] - \mathbb{E}x^-[h_\ell(x^-)].$$

Our analysis focuses primarily on Regime 2, as it represents the dominant paradigm in current steering research.

3.3 Linear Approximation and Nonlinear Case

Local linearization. Near a reference distribution, we can approximate the effect of steering on output logits as:

$$o(x, v, \alpha) \approx o(x, 0, 0) + \alpha J_\ell(x)v,$$

where $J_\ell(x) = \frac{\partial o}{\partial h_\ell} |_{h_\ell(x)} \in \mathbb{R}^{V \times d}$ is the Jacobian and V is the vocabulary size.

Nonlinear case. In general, the mapping $h_\ell \mapsto o$ involves multiple nonlinear layers (attention, MLPs, layer norms). We denote this as:

$$o = F_{\ell \rightarrow L}(h_\ell + \alpha v),$$

where $F_{\ell \rightarrow L}$ is the composition of layers $\ell + 1$ through L .