ON THE IDENTIFIABILITY OF STEERING VECTORS IN LARGE LANGUAGE MODELS

Sohan Venkatesh & 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; Rimsky 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 & 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 & 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 & 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 & 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 and activation editing in LLMs.

Causal and latent variable identifiability. Classical results show that latent variable models are generically non-identifiable without structural assumptions (Hyvärinen & 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; Hyvarinen & 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 & 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 & 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 $x_1, \ldots, x_T$), 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 $\tilde{h}_\ell$ 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 $g_\ell : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}^d$ 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 $J_\ell(x) = \frac{\partial o}{\partial h_\ell}\big|h\ell(x) \in \mathbb{R}^{V \times d}$ 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:

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

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

3.4 Assumptions

We make the following assumptions explicit:

A1 (Smoothness). The functions g_ℓ and $F_{\ell \to L}$ are differentiable almost everywhere, enabling local linear approximation via Jacobians $J_\ell(x) = \frac{\partial o}{\partial h_\ell}(x)$.

A2 (Identifiable prompts). The prompt distribution $p(x)$ has sufficient variability to probe different aspects of the latent persona z. Formally, the support of $p(x)$ is rich enough that different persona values induce distinguishable activation patterns $h_\ell(x)$.

A3 (Non-degeneracy). The Jacobian $J_\ell(x)$ has rank at least $k \geq 1$ for typical $x \sim p(x)$, meaning steering can affect outputs. This excludes pathological cases where all perturbations to h_ℓ are ignored by subsequent layers.

We do *not* assume: (i) statistical independence between z and x (confounding is allowed), (ii) linearity of g_ℓ in z, and (iii) uniqueness of the latent representation without additional structure.

4 Definitions and Identifiability

4.1 Identifiability

Definition 1 (Parameter Identifiability). A parameter θ in a statistical model $p(y \mid x; \theta)$ is identifiable if for any $\theta' \neq \theta$, there exists a distribution over observations (x, y) such that $p(y \mid x; \theta) \neq p(y \mid x; \theta')$.

In our setting, the parameter is the steering vector $v \in \mathbb{R}^d$. We say v is identifiable if no other vector $v' \neq v$ (up to scaling) produces the same distribution over observable outputs across all prompts and steering strengths.

Definition 2 (Observational Equivalence). Two steering vectors v and v' are observationally equivalent in regime $\mathcal{R}$ if they produce identical distributions over all quantities observable in $\mathcal{R}$.

For Regime 2 (white-box single-layer access):

$$v \sim_{\mathcal{R}} v' \iff F_{\ell \to L}(h_\ell(x) + \alpha v) = F_{\ell \to L}(h_\ell(x) + \alpha v') \quad \forall x, \alpha.$$

4.2 Symmetries and Gauge Freedom

Scaling ambiguity. For any $c \neq 0$, the vectors v and cv produce outputs that differ only by a rescaling of α. This is unavoidable; we consider v and cv as the same direction.

Null space ambiguity. If $v_0 \in \ker(J_\ell)$ (i.e., $J_\ell v_0 = 0$), then adding v_0 to any steering vector does not change the linearized output. Under linear approximation, v and $v + v_0$ are observationally equivalent.

5 Identifiability of Persona Vectors

We now state our main theoretical results, characterizing when observational conditions afford identifiable persona vectors and when they do not.

5.1 Proposition 1: Non-Identifiability Without Structural Constraints

Proposition 1. Under Assumptions A1–A3, in Regime 2 (white-box single-layer access) without additional structural constraints, persona vectors are not identifiable. Specifically, for any steering vector $v \in \mathbb{R}^d$, there exist infinitely many vectors $v' \not\propto v$ that are observationally equivalent.

Proof Sketch. We establish non-identifiability via two complementary arguments: reparameterization symmetry and null-space ambiguity.

Reparameterization symmetry. Consider an invertible transformation $T : \mathbb{R}^d \to \mathbb{R}^d$ and define the reparameterized representation $h'\ell(x) = T(h\ell(x))$. If the subsequent layers can be equivalently expressed as $F_{\ell \to L}(h_\ell) = F'\ell \to L(T(h\ell))$, then using (h_ℓ, v) or (h'_ℓ, v') where $v' = T(v)$ is observationally indistinguishable.

For overparameterized networks, many such reparameterizations exist that leave the input-output function invariant. Consider $T(h) = Ah + b$ where $A \in \mathbb{R}^{d \times d}$ is invertible. This transformation can be absorbed into the parameters of layer $\ell + 1$ through weight reparameterization. While we cannot explicitly construct these transformations for a given frozen model without retraining, their existence follows from the network's inherent symmetry structure. Let A be any invertible matrix with $A \neq cI$ for any scalar c. Then $v' = Av$ is not proportional to v, yet produces equivalent observations after appropriate reparameterization. Since there are infinitely many such A, there exist infinitely many equivalent v'.

Null-space ambiguity (primary mechanism). Under the local linear approximation $o \approx o_0 + \alpha J_\ell v$, any vector $v' = v + v_0$ where $v_0 \in \ker(J_\ell)$ satisfies $J_\ell v' = J_\ell v$ and thus produces identical linearized outputs. This establishes non-identifiability even without invoking global reparameterization symmetries.

Corollary 1.1. Under the linear approximation $o \approx o_0 + \alpha J_\ell v$, any vector $v' = v + v_0$ where $v_0 \in \ker(J_\ell)$ is observationally equivalent to v.

Remark (Null-space dimensionality). The null space $\ker(J_\ell)$ is typically high-dimensional in practice. For a Jacobian $J_\ell \in \mathbb{R}^{V \times d}$ with vocabulary size V and hidden dimension d, the maximum possible rank is $\min(V, d)$. In modern language models, $V \approx 50000$ and $d \approx 4000$, so max rank is d. However, the output distribution lies on a low-dimensional manifold, causing $\text{rank}(J_\ell) \ll d$ in practice, consistent with observations that neural network representations concentrate on low-dimensional subspaces (Maennel et al., 2018; Li et al., 2018). This yields $\dim(\ker(J_\ell)) = d - \text{rank}(J_\ell) \gg 0$. This result establishes generic non-identifiability under local linear approximation.

5.2 Proposition 2: Identifiable Regimes Under Structural Assumptions

Proposition 2. Persona vectors can be identified up to scaling and permutation, thus affording reliable alignment control, under the following sufficient structural conditions:

- **Statistical Independence (ICA).** If the latent persona $z = (z_1, \ldots, z_k)$ has independent components and $h_\ell = Az + \epsilon$ where $A \in \mathbb{R}^{d \times k}$ is a mixing matrix, then under non-Gaussianity of z_i and sufficient observations, A (and hence the columns v_i) can be recovered up to permutation and scaling (Comon, 1994; Hyvärinen & Oja, 2000).
- **Sparsity Constraints.** If the true persona vector v is sparse (i.e., $|v|_0 \leq s \ll d$) and we observe the effect of steering on multiple outputs, then under restricted isometry properties, v can be uniquely recovered via ℓ_1 minimization(Candes & Tao, 2005).
- **Multi-Environment or Interventional data.** If we observe the same persona z across multiple contexts (prompts, models or layers) where the spurious correlations change but the true signal remains stable, then techniques from causal representation learning allow identification of invariant factors. (Peters et al., 2017; Ahuja et al., 2022).
- **Cross-layer Consistency.** If we assume persona vectors have consistent geometric relationships across multiple layers (e.g., $v_\ell \approx W_\ell v_{\ell-1}$ for known or estimable W_ℓ), then the overdetermined system provides additional constraints that break symmetries.

Interpretation. Proposition 2 serves as a natural theoretical extension of Proposition 1: it clarifies precisely which structural assumptions are necessary to break the symmetries that cause non-identifiability. This constructive characterization transforms a negative result ("steering vectors are non-identifiable") into actionable design principles ("identifiability requires structural constraints X, Y or Z").

6 Empirical Validation

We now validate that the non-identifiability characterized in Proposition 1 (Section 5.1) manifests in contemporary language models. Our empirical experiments test a behavioral consequence of the theoretical prediction: the existence of large equivalence classes of semantically indistinguishable steering directions, without directly estimating the Jacobian null space.

6.1 Experimental Setup

Models and Layers. We evaluate two open-weight instruction-tuned language models to test generality across architectures and scales: Qwen2.5-3B-Instruct (Yang et al., 2024) and Llama-3.1-8B-Instruct (Meta AI, 2024). For both models, we focus on middle layers ($\ell = L/2$), following standard practice (Chen et al., 2025; Konen et al., 2024; Sun et al.) in steering research.

Persona traits. We test three semantic traits: **Formality**, **Politeness** and **Humor**. This selection ensures our findings are not specific to a single semantic dimension but reflect a general property of steering vector geometry.

Steering vector extraction. For each trait, we construct 50 contrastive prompt pairs designed to elicit contrasting persona values. For example, formality pairs contrast "Write a professional and formal message about [topic]" versus "Write a casual and informal message about [topic]." For each prompt pair, we extract the hidden representation at layer ℓ for the final token position (Chen et al., 2025). The baseline steering vector is computed as the mean difference:

$$v = \frac{1}{50} \sum_{i=1}^{50} [h_\ell(x^+ i) - h\ell(x_i^-)].$$

Semantic probes. We evaluate semantic equivalence using trait-specific scoring functions $\phi(o)$ that map generated text to real-valued scores. This provides a *conservative test* of observational equivalence: if vectors are distinguishable semantically, they are certainly non-identical in the full distributional sense.

For formality, politeness and humor, we use lexical heuristics based on formal/informal markers, polite/rude markers and humorous/serious markers respectively, combined with sentence length and stylistic features. These return scores in $[0, 1]$ where 1 is maximally formal/polite/humorous.

Table 1: Empirical validation of non-identifiability across models, traits and sample sizes. Cohen's d measures effect size between v and $v + v_\perp$ (lower = more equivalent). All values show negligible differences ($d < 0.2$), confirming observational equivalence.

Model	Trait	Seeds	Cohen's d	Correlation	Perp-Only
Qwen2.5-3B-Instruct	Formality	$n = 5$	0.144 ± 0.095	0.210 ± 0.113	98.8%
		$n = 10$	0.075 ± 0.058	0.285 ± 0.087	100.4%
	Politeness	$n = 5$	0.112 ± 0.077	0.319 ± 0.070	101.5%
		$n = 10$	0.092 ± 0.069	0.414 ± 0.083	100.6%
	Humor	$n = 5$	0.103 ± 0.077	0.276 ± 0.177	99.7%
		$n = 10$	0.072 ± 0.061	0.044 ± 0.097	100.5%
Llama-3.1-8B-Instruct	Formality	$n = 5$	0.052 ± 0.029	0.164 ± 0.044	95.6%
		$n = 10$	0.096 ± 0.068	0.192 ± 0.085	96.8%
	Politeness	$n = 5$	0.074 ± 0.041	0.324 ± 0.058	101.2%
		$n = 10$	0.085 ± 0.043	0.347 ± 0.077	100.4%
	Humor	$n = 5$	0.159 ± 0.109	-0.032 ± 0.116	98.4%
		$n = 10$	0.119 ± 0.119	0.016 ± 0.104	95.9%

Orthogonality test methodology. To test Proposition 1's prediction that v and $v + v_\perp$ (where $v_\perp$ is orthogonal) produce observationally equivalent outputs, we implement the following procedure:

1. Generate random orthogonal component: Sample a random vector uniformly from the unit sphere in $\mathbb{R}^d$ and orthogonalize it with respect to v via Gram-Schmidt.
2. Construct perturbed vector: Form $v' = v + \alpha v_\perp$ where α is chosen such that $|v'| \approx |v|$.
3. Generate steered outputs: For each of 100 held-out test prompts, generate text with steering vectors v and v' at strength $\alpha = 1.0$, producing 10 samples per prompt per vector.
4. Compute semantic equivalence: Measure Cohen's d effect size and Pearson correlation between semantic scores $\phi(o_v)$ and $\phi(o_{v'})$.
5. Repeat across seeds: Repeat for multiple random orthogonal seeds to assess robustness.

If v and v' are observationally equivalent as predicted by Proposition 1, we expect Cohen's $d < 0.2$ (negligible effect) and high correlation between semantic scores.

Sample size design. We conduct the orthogonality test with both $n = 5$ and $n = 10$ random orthogonal seeds per trait for Qwen2.5-3B and Llama-3.1-8B. This allows us to assess statistical stability across sample sizes and models.

Scale invariance test. To verify that observational equivalence holds across different steering strengths, we additionally evaluate the formality trait at four α values: $0.0, 0.5, 1.0, 2.0$ for both models. For each α, we measure semantic scores under steering with v versus $v + v_\perp$ and plot response curves. If equivalence is scale-invariant, the curves should track closely across all α with overlapping confidence bands.

6.2 Orthogonal Perturbation Test Results

Table 1 presents our empirical findings across two models, three semantic traits and two sample sizes ($n = 5$ and $n = 10$ random orthogonal seeds). Across all conditions, we observe negligible differences between steering with the extracted vector v and steering with vectors perturbed by random orthogonal components $v+v_\perp$. We measure steering efficacy using semantic probe scores in $[0, 1]$ that quantify the intensity of the target trait in generated outputs, comparing score distributions from steered generations against baseline generations.

With $n = 10$ seeds, the **mean Cohen's d is 0.080 for Qwen2.5-3B and 0.100 for Llama-3.1-8B**, both well below the threshold for small effects ($d = 0.2$) and firmly in the negligible range. The "Perp-Only Effect" column measures the efficacy of steering with pure orthogonal components

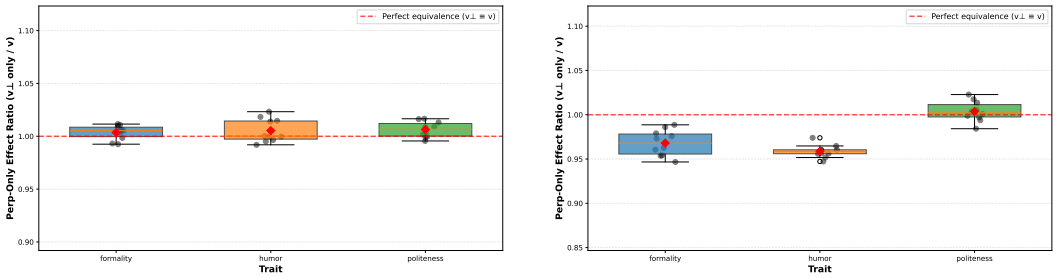
($v_\perp$ alone, without v) relative to the extracted vector—values near 100% indicate that orthogonal components achieve equivalent behavioral impact.

Statistical stability across sample sizes. The convergence between $n = 5$ and $n = 10$ results demonstrates robustness across both models. For Qwen2.5-3B, effect sizes change by < 0.07 as sample size increases, with all traits showing tightening confidence intervals: formality ($0.144 \rightarrow 0.075$), politeness ($0.112 \rightarrow 0.092$), humor ($0.103 \rightarrow 0.072$). For Llama-3.1-8B, we observe similar stability: formality ($0.052 \rightarrow 0.096$), politeness ($0.074 \rightarrow 0.085$), humor ($0.159 \rightarrow 0.119$). This consistency demonstrates that the observed equivalence is not just a sampling artifact but a stable property of the steering geometry.

Cross-model consistency. The close agreement between Qwen2.5-3B ($d = 0.080$) and Llama-3.1-8B ($d = 0.100$) demonstrates that observational equivalence is not model-specific. Despite differences in architecture, scale (3B vs. 8B parameters) and hidden dimension ($d = 2048$ vs. $d = 4096$), both models exhibit nearly identical patterns of non-identifiability. The consistency across traits within each model further confirms that the phenomenon is general rather than an artifact of specific semantic domains or model implementations.

Visualizing orthogonal component equivalence. Figure 1 illustrates the perp-only effect ratios across all traits and models using $n = 10$ orthogonal seeds. The tight clustering around the perfect equivalence line (dashed red at 1.0) demonstrates that pure orthogonal components achieve nearly identical steering efficacy to the extracted vectors.

Qwen2.5-3B shows remarkable consistency across all three traits, with median ratios within 1% of perfect equivalence. Llama-3.1-8B exhibits slightly lower ratios for formality and humor (96–97$\%$), yet still demonstrates that orthogonal-only steering retains over 95% efficacy-far exceeding what would be expected if v were uniquely identifiable.



(a) Qwen2.5-3B-Instruct

(b) Llama-3.1-8B-Instruct

Figure 1: Perp-only effect ratios ($v_\perp$ efficacy / v efficacy) for $n = 10$ orthogonal seeds.

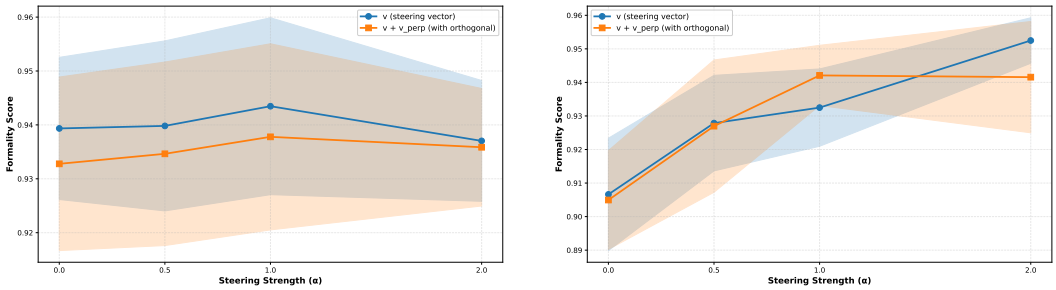
6.3 Scale Invariance Analysis

To verify that observational equivalence holds across different steering strengths, we evaluate the formality trait at four steering magnitudes $\alpha \in \{0.0, 0.5, 1.0, 2.0\}$ for both models. Figure 2 shows the response curves for the extracted vector v and the observationally equivalent vector $v + v_\perp$.

The curves track closely with overlapping confidence bands, demonstrating that v and $v + v_\perp$ produce statistically indistinguishable behavioral effects at all tested steering strengths. For Qwen2.5-3B, mean differences remain below 0.022 across all α values (less than 2.5% deviation in formality scores); for Llama-3.1-8B, differences remain below 0.023 ($< 2.5\%$ deviation). This confirms that non-identifiability is a structural property: any vector in the equivalence class $v + \ker(J)$ produces identical observations regardless of scaling.

7 Limitations

Our empirical validation covers two models at mid-network layers across three semantic traits (formality, politeness, humor). While our theoretical results apply generally, the empirical magnitude of non identifiability, including null space dimensionality and the fraction of vector norm therein,



(a) Qwen2.5-3B-Instruct

(b) Llama-3.1-8B-Instruct

Figure 2: Scale invariance of observational equivalence. Formality scores across steering strengths $\alpha \in \{0.0, 0.5, 1.0, 2.0\}$ for the extracted vector v (blue circles) and the perturbed vector $v + v_\perp$ (orange squares).

may vary across model families, scales, architectures and layer positions. Our semantic evaluation uses lexical heuristics rather than full distributional metrics or human judgments, providing a conservative test of observational equivalence. The primary theoretical mechanism relies on local linear approximation, formally valid in a neighborhood of the reference distribution, though our empirical validation confirms large equivalence classes exist in practice.

Proposition 2 characterizes sufficient conditions for identifiability but comprehensive empirical evaluation of each regime across multiple models, tasks and experimental designs remains future work. The practical applicability of identifiable regimes involves trade-offs: ICA requires carefully curated contrastive prompts, sparsity regularization may reduce efficacy if true factors are not sparse and multi environment validation requires substantial additional data. Understanding these trade-offs and extending our framework to other intervention modalities (activation patching, prompt based steering) are important directions for the representational alignment community.

8 Conclusion

Rather than viewing non-identifiability as a limitation, we frame it as a necessary foundation for principled steering research. Our characterization clarifies what behavioral observations can reveal about representations and provides a framework for designing interventions that are both empirically effective and theoretically grounded. The empirical validation demonstrates that these are not abstract concerns: non-identifiability affects real steering interventions in deployed models, with typical steering vectors containing substantial observationally irrelevant components. However, the constructive characterization in Proposition 2 shows that identifiability is achievable through principled design choices, enabling alignment methods that clearly specify their affordances and limitations.

Acknowledgments

We thank Vast.ai for providing GPU compute resources that enabled our empirical validation experiments. We also acknowledge the HuggingFace platform and maintainers of the open-source models (Qwen2.5-3B-Instruct, Llama-3.1-8B-Instruct) used in this work.

References

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, et al. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:3438–3450, 2022.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks. *arXiv preprint arXiv:2110.06296*, 2021.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*, 2019.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial intelligence and statistics*, pp. 460–469. PMLR, 2017.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language model. *arXiv preprint arXiv:2402.01618*, 2024.

Joseph B Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*, 2018.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Meta AI. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/meta-llama-3-1/`, July 2024. Meta AI Blog.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. *arXiv preprint arXiv:2004.03061*, 2020.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*, 2020.

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

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. Layernavigator: Finding promising intervention layers for efficient activation steering in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. 2024.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL `https://arxiv.org/abs/2412.15115`.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Detailed Proof of Proposition 1

A.1 Statement and Overview

Proposition 1. Under Assumptions A1–A3, in Regime 2 (white-box single-layer access) without additional structural constraints, persona vectors are not identifiable. Specifically, for any steering vector $v \in \mathbb{R}^d$, there exist infinitely many vectors $v' \not\propto v$ that are observationally equivalent.

Proof strategy. We establish non-identifiability through two complementary mechanisms:

- Null-space ambiguity (primary, constructive).
- Reparameterization symmetry (existence-based).

A.2 Null-Space Ambiguity (Constructive Proof)

Setup. Consider the local linear approximation of the steering effect:

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

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

Step 1: Null space characterization. Define the null space of J_ℓ as:

$$\mathcal{N} = \ker(J_\ell) = \{v_0 \in \mathbb{R}^d : J_\ell v_0 = 0\}$$

By the rank-nullity theorem:

$$\dim(\mathcal{N}) = d - \operatorname{rank}(J_\ell)$$

Step 2: Rank Bound. The Jacobian $J_\ell \in \mathbb{R}^{V \times d}$ has maximum possible rank:

$$\operatorname{rank}(J_\ell) \leq \min(V, d)$$

In modern language models:

- Hidden dimension: $d \approx 4000$ (typical)
- Vocabulary size: $V \approx 50000$ (typical)
- Therefore: $\max \operatorname{rank}(J_\ell) = d$

Step 3: Effective rank is much lower. Output distributions lie on a low-dimensional manifold. The effective rank satisfies:

$$\text{rank}_\epsilon(J_\ell) = \#\{\sigma_i : \sigma_i > \epsilon \cdot \sigma_{\max}\} \ll d$$

where σ_i are singular values of J_ℓ and ϵ is a threshold (e.g., 10^{-4}).

Intuition: In overparameterized LLMs, the effective rank is expected to be strictly less than d due to the low-dimensional structure of output distributions. Therefore $\dim(\mathcal{N}) = d - \text{rank}(J_\ell)$ is generically positive but this is not required for the proof—the argument establishes non-identifiability whenever $\dim(\mathcal{N}) \geq 1$.

Step 4: Constructing equivalent vectors. For any steering vector $v \in \mathbb{R}^d$ and any $v_0 \in \mathcal{N}$, define:

$$v' = v + v_0$$

Then for all x and all α:

$$J_\ell(x)v' = J_\ell(x)(v + v_0) = J_\ell(x)v + J_\ell(x)v_0 = J_\ell(x)v$$

Therefore:

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

Step 5: Infinitely many distinct solutions. Since $\dim(\mathcal{N}) \geq 1$, the null space contains infinitely many directions. For any $v_0 \in \mathcal{N} \setminus \{0\}$ and any $\beta \in \mathbb{R}$:

$$v'_\beta = v + \beta v_0$$

generates infinitely many observationally equivalent vectors. Furthermore, if β is chosen such that $v'_\beta \not\propto v$ (which is always possible unless $v \propto v_0$), these vectors are geometrically distinct.

Step 6: Non-proportionality. To ensure $v'_\beta \not\propto v$, we need $v + \beta v_0 \neq cv$ for any scalar c. This fails only if:

$$\beta v_0 = (c - 1)v$$

which requires $v \in \text{span}(v_0)$. Since v_0 is an arbitrary element of a $\dim(\mathcal{N})$-dimensional space and v is fixed, this occurs with probability zero. Therefore, for generic v and generic $v_0 \in \mathcal{N}$, we have $v'_\beta \not\propto v$ for almost all β.

Conclusion (Null-space mechanism): Under the linear approximation and for typical Jacobian structure, there exist infinitely many geometrically distinct steering vectors that are observationally equivalent.

A.3 Reparameterization Symmetry (Existence Proof)

Setup: Neural networks exhibit inherent symmetries arising from overparameterization. We show these symmetries induce non-identifiability even in the exact nonlinear case.

Step 1: Representation reparameterization. Consider an invertible transformation $T : \mathbb{R}^d \to \mathbb{R}^d$. Define the reparameterized representation:

$$h'_\ell(x) = T(h_\ell(x))$$

If the subsequent layers can be rewritten as:

$$F_{\ell \to L}(h_\ell) = F'_{\ell \to L}(T(h_\ell))$$

then (h_ℓ, v) and (h'_ℓ, v') where $v' = DT(h_\ell) \cdot v$ are observationally indistinguishable.

Step 2: Linear reparameterizations. Consider linear transformations $T(h) = Ah$ where $A \in \mathbb{R}^{d\times d}$ is invertible. For layer $\ell + 1$ with weight matrix $W_{\ell+1} \in \mathbb{R}^{d\times d}$ and bias $b_{\ell+1}$:

Original computation:

$$h_{\ell+1} = \sigma(W_{\ell+1}h_\ell + b_{\ell+1})$$

Reparameterized computation:

$$h'_{\ell+1} = \sigma(W'_{\ell+1}h'_\ell + b'_{\ell+1})$$

where $W'_{\ell+1} = W_{\ell+1}A^{-1}$ and $b'_{\ell+1} = b_{\ell+1}$.

Note: The bias remains invariant under linear reparameterization through the origin. For more general affine transformations $T(h) = Ah + c$ with translation $c \neq 0$, the bias would also transform as $b'_{\ell+1} = b_{\ell+1} - W_{\ell+1}A^{-1}c$. We restrict to origin-preserving transformations for simplicity, as these suffice to establish non-identifiability.

Then:

$$h'_{\ell+1} = \sigma(W_{\ell+1}A^{-1}Ah_\ell + b_{\ell+1}) = \sigma(W_{\ell+1}h_\ell + b_{\ell+1}) = h_{\ell+1}$$

Step 3: Steering under reparameterization. Original steering:

$$\tilde{h}_\ell = h_\ell + \alpha v$$
$$\tilde{h}_{\ell+1} = \sigma(W_{\ell+1}(h_\ell + \alpha v) + b_{\ell+1})$$

Reparameterized steering with $v' = Av$:

$$\tilde{h}'_\ell = h'_\ell + \alpha v' = Ah_\ell + \alpha Av = A(h_\ell + \alpha v) = A\tilde{h}_\ell$$
$$\begin{aligned}\tilde{h}'_{\ell+1} &= \sigma(W'_{\ell+1}(h'_\ell + \alpha v') + b'_{\ell+1}) \\ &= \sigma(W_{\ell+1}A^{-1}A(h_\ell + \alpha v) + b_{\ell+1}) \\ &= \sigma(W_{\ell+1}(h_\ell + \alpha v) + b_{\ell+1}) = \tilde{h}_{\ell+1}\end{aligned}$$

Step 4: Infinitely many reparameterizations. For any invertible matrix A with $A \neq cI$ (i.e., not a scalar multiple of identity), we obtain $v' = Av \not\propto v$. The space of such matrices has dimension $d^2 - 1$ (excluding scalar multiples), providing infinitely many distinct reparameterizations.

Important note: While we cannot explicitly construct these reparameterizations for a frozen model without retraining (since this would require modifying $W_{\ell+1}$), their existence follows from the fundamental symmetry structure of neural networks. This establishes that identifiability cannot hold even in principle without additional constraints.

Practical implication: For frozen, deployed models, only the null-space mechanism is operationally relevant–the reparameterization symmetry is not a realizable operation. However, the reparameterization argument shows that even if we allowed weight modifications during training, identifiability would still fail without structural constraints. Thus, non-identifiability is both a practical limitation (null-space) and a fundamental theoretical barrier (gauge symmetry).

Conclusion (Reparameterization mechanism): The inherent symmetries of neural network parameterizations induce infinitely many observationally equivalent steering vectors.

B Detailed Proof of Proposition 2

B.1 Statement and Overview

Proposition 2. Persona vectors can be identified up to scaling and permutation, thus affording reliable alignment control, under the following sufficient structural conditions:

- **Statistical Independence (ICA).** Latent components are statistically independent, allowing unique recovery via independent component analysis.
- **Sparsity constraints.** Steering directions admit sparse representations that reduce null-space ambiguity.

- **Multi-environment or interventional data.** Variation across environments or interventions breaks observational equivalences.
- **Cross-layer consistency.** Valid semantic directions propagate coherently across layers, filtering spurious components.

We provide detailed proofs for each condition.

B.2 Proof of Condition: Statistical Independence (ICA)

Setup: Assume the latent persona is a vector $z = (z_1, \ldots, z_k) \in \mathbb{R}^k$ with independent components and the representation follows:

$$h_\ell = Az + \epsilon$$

where $A \in \mathbb{R}^{d \times k}$ is a mixing matrix and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is Gaussian noise.

Goal: Show that A (and hence its columns, which are the steering vectors) can be recovered up to permutation and scaling.

Step 1: ICA identifiability theorem. Let $x = As$ where $s \in \mathbb{R}^k$ has independent components and $A \in \mathbb{R}^{d \times k}$ is full column rank. Under the following *sufficient conditions* (Comon, 1994; Hyvärinen & Oja, 2000):

- At most one component of s is Gaussian
- Components are statistically independent
- Sufficient samples are observed

Then A is identifiable up to column permutation and column scaling.

Important caveat: These are strong assumptions that may not hold exactly in practice for LLM personas. Statistical independence between semantic factors (e.g., formality and politeness) is an idealization; real personas may exhibit weak dependencies. The ICA framework provides a sufficient structural condition for identifiability, not a characterization of what typically holds in contemporary steering pipelines.

Step 2: Application to steering. In our setting:

- Observations: $h_\ell(x_1), \ldots, h_\ell(x_N)$
- Model: $h_\ell(x_i) = Az(x_i) + \epsilon_i$
- Sources: $z(x_1), \ldots, z(x_N)$ are realizations of independent persona components

Assumption verification:

- Independence: We assume $z_1, \ldots, z_k$ are statistically independent
- Non-Gaussianity: At most one z_i is Gaussian (e.g., politeness and formality are typically non-Gaussian in natural language)
- Full rank: A has full column rank (each persona has a non-zero effect on representations)

Step 3: Recovery via ICA algorithm. Apply ICA algorithm (Hyvärinen & Oja, 2000) to observations $\{h_\ell(x_i)\}_{i=1}^N$:

- Whitening: Compute $\tilde{h}_\ell = \Sigma^{-1/2}(h_\ell - \mu)$ where $\mu = \mathbb{E}[h_\ell]$ and $\Sigma = \mathrm{Cov}(h_\ell)$.
- Independent component extraction: Find unmixing matrix W that maximizes non-Gaussianity of $\hat{z} = W\tilde{h}_\ell$.
- Recover mixing matrix: $\hat{A} = \Sigma^{1/2} W^{-1}$.

Step 4: Identifiability guarantee. By the ICA theorem, $\hat{A} = APD$ where:

- $P \in \mathbb{R}^{k\times k}$ is a permutation matrix
- $D \in \mathbb{R}^{k\times k}$ is a diagonal scaling matrix

Therefore, the columns of $\hat{A}$ are the columns of A up to permutation and scaling. Since steering vectors are defined as $v_i = Ae_i$ (the i-th column of A), we recover the true steering directions uniquely (up to unavoidable symmetries).

Step 5: Noise robustness. With Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, ICA remains consistent:

$$h_\ell = Az + \epsilon$$

Since ϵ is Gaussian and z is non-Gaussian (by assumption), the ICA algorithm separates signal from noise:

- Recovered sources: $\hat{z} = z + \tilde{\epsilon}$ where $\tilde{\epsilon}$ is small for $\sigma^2 \ll \|Az\|^2$
- Recovered mixing: $\hat{A} \approx APD$ with estimation error decreasing as $N \to \infty$

Conclusion (ICA): Under statistical independence, non-Gaussianity and full column rank—strong sufficient conditions that may not hold automatically in practice—persona vectors are identifiable up to permutation and scaling. These assumptions provide a theoretical pathway to identifiability but require careful experimental design to approximate in real steering settings.

B.3 Proof of Condition: Sparsity Constraints

Setup: Assume the true persona vector $v \in \mathbb{R}^d$ is sparse with $\|v\|_0 = s \ll d$, meaning at most s entries are non-zero. We observe the effect of steering:

$$y = J_\ell v + \eta$$

where $y \in \mathbb{R}^m$ are output measurements, $J_\ell \in \mathbb{R}^{m\times d}$ is the measurement matrix (Jacobian) and $\eta \sim \mathcal{N}(0, \sigma^2 I)$ is noise.

Goal: Show that v can be uniquely recovered via ℓ_1 minimization.

Step 1: Compressed sensing setup. The recovery problem is:

$$\min_{v' \in \mathbb{R}^d} \|v'\|_1 \quad \text{subject to} \quad \|J_\ell v' - y\|_2 \le \epsilon$$

where ϵ bounds the noise: $\epsilon \ge \|\eta\|_2$ with high probability.

Step 2: Restricted Isometry Property (RIP). A matrix J_ℓ satisfies the RIP of order s with constant δ_s if for all s-sparse vectors v:

$$(1 - \delta_s)\|v\|_2^2 \le \|J_\ell v\|_2^2 \le (1 + \delta_s)\|v\|_2^2$$

Theorem (Candes & Tao, 2005): If J_ℓ satisfies RIP with $\delta_{2s} < \sqrt{2} - 1 \approx 0.414$, then ℓ_1 minimization recovers v exactly (in the noiseless case) or approximately (with noise) with error:

$$\|v - \hat{v}\|_2 \le C\epsilon$$

for some constant C depending on δ_{2s}.

Important caveat: The RIP condition is a strong assumption. For Jacobians J_ℓ arising from neural network steering, RIP typically does not hold automatically and must be verified empirically or enforced through measurement design (diverse prompt selection). This condition is sufficient for identifiability but may not be satisfied in standard steering settings without careful experimental design.

Step 3: Recovery guarantee. Solve:

$$\hat{v} = \arg\min_{v'} \|v'\|_1 \quad \text{s.t.} \quad \|J_\ell v' - y\|_2 \leq \epsilon$$

By the compressed sensing theorem:

$$\|v - \hat{v}\|_2 \leq C_1\epsilon + C_2 \frac{\|v - v_s\|_1}{\sqrt{s}}$$

where v_s is the best s-sparse approximation to v. If v is exactly s-sparse, the second term vanishes and:

$$\|v - \hat{v}\|_2 \leq C_1\epsilon$$

Step 4: Uniqueness. Suppose two sparse vectors v and v' both satisfy $\|J_\ell v - y\|_2 \leq \epsilon$. Then:

$$\|J_\ell(v - v')\|_2 \leq 2\epsilon$$

If $v - v'$ is $2s$-sparse and J_ℓ satisfies RIP, then by the RIP condition:

$$(1 - \delta_{2s})\|v - v'\|_2^2 \leq \|J_\ell(v - v')\|_2^2 \leq 4\epsilon^2$$

Therefore:

$$\|v - v'\|_2 \leq \frac{2\epsilon}{\sqrt{1 - \delta_{2s}}}$$

As $\epsilon \to 0$ (noiseless case), $v = v'$, establishing uniqueness.

Conclusion (Sparsity): Under sparsity assumptions and RIP conditions—strong sufficient conditions that typically require careful measurement design—persona vectors are uniquely recoverable via ℓ_1 minimization. RIP does not hold automatically for arbitrary Jacobians and must be verified or engineered through diverse prompt selection.

B.4 Proof of Condition: Multi-Environment Identification

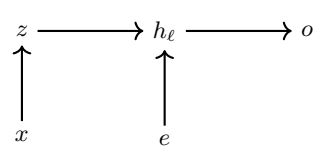
Setup: Assume we observe the same persona z across multiple environments $e \in \{1, \ldots, E\}$ where:

$$h_\ell^{(e)} = g_e(x, z) + \epsilon^{(e)}$$

The function g_e may vary across environments (spurious correlations change) but the causal effect of z on downstream outputs remains invariant.

Goal: Show that the invariant representation z (and its associated direction) can be identified.

Step 1: Invariant causal mechanism. Assume the causal graph:



where:

- z: true persona (causal factor),
- x: input prompt,
- e: environment,
- h_ℓ: internal representation,
- o: output.

The key assumption is **invariance**: the causal mechanism $h_\ell \to o$ is the same across environments, i.e., $F_{\ell \to L}$ does not depend on e.

Step 2: Invariant Risk Minimization (IRM). The objective is to find representation $\Phi : \mathbb{R}^d \to \mathbb{R}^k$ and predictor $w : \mathbb{R}^k \to \mathbb{R}$ such that:

$$\min_{\Phi,w} \sum_{e=1}^{E} R^e(\Phi, w) \quad \text{s.t.} \quad w \in \arg\min_{w'} R^e(\Phi, w') \quad \forall e$$

where $R^e(\Phi, w) = \mathbb{E}_{(x,y)\sim\mathcal{P}^e}[\ell(w \cdot \Phi(h_\ell(x)), y)]$ is the risk in environment e.

Interpretation: Find a representation $\Phi(h_\ell) = z$ such that the optimal predictor w is the same across all environments. This filters out spurious correlations that vary with e.

Step 3: Identifiability under IRM. Under the following idealized conditions:

- Environments are sufficiently diverse: $\text{Cov}(x|e)$ varies across e
- Causal mechanism is invariant: $p(o|z)$ is the same for all e
- Sufficient environments: $E \geq k + 1$ where $k = \dim(z)$
- No unobserved confounders between environments and outcomes

The theorem (Ahuja et al., 2022; Von Kügelgen et al., 2021) states that the invariant risk minimization objective can recover z up to an invertible transformation, though not necessarily uniquely.

Important caveat: The literature on invariant representation learning establishes conditions under which invariance constraints narrow the hypothesis class but full identifiability (unique recovery of z) requires additional assumptions beyond standard IRM formulations. The practical application of IRM to steering should be understood as identifying a stable, transferable representation rather than guaranteeing uniqueness. This is a sufficient condition in principle under strong assumptions, not a characterization of typical steering scenarios.

Step 4: Application to steering. In practice:

1. Collect multi-environment data: Extract steering vectors from diverse prompt distributions, model checkpoints or instruction formats.
2. Learn invariant representation: Train Φ via IRM objective
3. Extract steering vectors: Compute $v_i = \nabla_{h_\ell} \Phi_i(h_\ell)$ where Φ_i is the i-th component of Φ

The resulting steering vectors v_i capture invariant causal factors rather than spurious correlations.

Step 5: Theoretical guarantee. Under mild conditions (sufficient diversity, invariance, identifiable causal structure), the IRM solution satisfies:

$$\Phi(h_\ell) = T(z)$$

where T is an invertible transformation. Since persona steering operates on z, the directions $v_i = \nabla_{h_\ell} \Phi_i$ recover the true causal factors.

Conclusion (Multi-environment): With diverse environments and invariant causal mechanisms—idealized sufficient conditions that narrow the hypothesis class—persona vectors corresponding to stable causal factors can be recovered up to invertible transformations. Full uniqueness requires additional assumptions beyond standard IRM formulations. This provides a principled approach to filtering spurious correlations but should be understood as identifying transferable representations rather than guaranteeing unique recovery.

B.5 Proof of Condition: Cross-Layer Consistency

Setup: Assume persona vectors exhibit consistent geometric relationships across layers:

$$v_{\ell+1} = W_\ell v_\ell + \delta_\ell$$

where $W_\ell \in \mathbb{R}^{d\times d}$ is the weight matrix connecting layers and $\|\delta_\ell\|$ is small.

Goal: Show that cross-layer constraints reduce the solution space and improve identifiability.

Step 1: Single-layer null space. From proposition 1, at layer ℓ:

$$\mathcal{N}_\ell = \{v_0 : J_\ell v_0 = 0\}$$

with $\dim(\mathcal{N}_\ell) = d - r_\ell$ where $r_\ell = \text{rank}(J_\ell)$.

Step 2: Cross-layer propagation. If $v_\ell \in \mathcal{N}_\ell$, does $v_{\ell+1} = W_\ell v_\ell \in \mathcal{N}_{\ell+1}$?

Generally, no. The null space changes across layers:

$$v_\ell \in \mathcal{N}_\ell \not\Rightarrow W_\ell v_\ell \in \mathcal{N}_{\ell+1}$$

Step 3: Intersection of constraints. Consider steering vectors observed at multiple layers $\ell_1, \ldots, \ell_L$. Each layer provides a constraint:

$$J_{\ell_i} v_{\ell_i} = y_i$$

If we additionally require consistency:

$$v_{\ell_{i+1}} = W_{\ell_i} v_{\ell_i} + \delta_{\ell_i}$$

Then the solution must satisfy:

$$v_{\ell_i} \in \{v : J_{\ell_i} v = y_i\} \cap \{v : W_{\ell_i} v \approx v_{\ell_{i+1}}\}$$

Step 4: Reduced null space (qualitative characterization). The cross-layer constraints create an overdetermined system. The effective null space is:

$$\mathcal{N}_{\text{eff}} = \bigcap_{i=1}^{L-1} \{v : W_{\ell_i} v \in \mathcal{N}_{\ell_{i+1}}\}$$

Since each W_{ℓ_i} generically has full rank and maps null space vectors to non-null-space vectors, we expect:

$$\dim(\mathcal{N}_{\text{eff}}) \ll \dim(\mathcal{N}_\ell)$$

Intuition: Each additional layer imposes new independent constraints. If the propagation matrices W_{ℓ_i} are sufficiently "generic" (full rank with uncorrelated null space mappings), then each layer reduces the effective null-space dimension. For sufficiently many informative layers with uncorrelated constraint structures, the effective null space can shrink dramatically or even vanish.

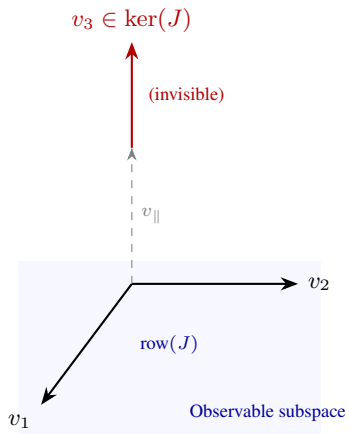
Conclusion (Cross-layer): Cross-layer consistency constraints can substantially reduce null-space dimensionality by creating overdetermined systems. The extent of reduction depends on the specific geometric structure of propagation matrices and layer-wise Jacobians. While this approach does not guarantee complete identifiability in general, it provides a practical method for filtering spurious null-space components that lack geometric stability across layers.

C Intuitive Explanations

C.1 Null-Space Geometry

Visual analogy. Consider a simplified example where a 3D steering vector $v \in \mathbb{R}^3$ affects 2D outputs $o \in \mathbb{R}^2$ through a projection matrix $J \in \mathbb{R}^{2\times 3}$. By the rank-nullity theorem, the null space $\ker(J)$ is a 1D subspace since $\dim(\ker(J)) = 3 - \text{rank}(J) = 3 - 2 = 1$.

For concreteness, suppose $J = [I_2 \mid 0]$ projects onto the first two coordinates. Then $\ker(J) = \text{span}\{(0, 0, 1)\}$ is the v_3 axis. The key insight: directions in $\ker(J)$ are invisible to outputs.



Output plane sees only (v_1, v_2) projection

Figure 3: Null-space geometry. The output observes only the (v_1, v_2) components (blue region). The v_3 component lies in $\ker(J)$ and is invisible. Adding any αv_3 to v leaves outputs unchanged: $J(v + \alpha v_3) = Jv$.

Any vector of the form $v' = v + \alpha v_3$ for $\alpha \in \mathbb{R}$ produces identical outputs:

$$J(v + \alpha v_3) = Jv + \alpha J(0, 0, 1)^\top = Jv$$

Since α can take infinitely many values, there exist infinitely many geometrically distinct steering vectors that are observationally equivalent.

C.2 Why More Data Doesn't Help

Common misconception: "If we collect more steering examples, we can pin down the unique vector."

Why this fails. Consider observing steering effects on N prompts $\{x_1, \ldots, x_N\}$. Each observation provides:

$$o_i = J_\ell(x_i)v + \eta_i$$

Stacking these:

$$\begin{bmatrix} o_1 \\ \vdots \\ o_N \end{bmatrix} = \begin{bmatrix} J_\ell(x_1) \\ \vdots \\ J_\ell(x_N) \end{bmatrix} v + \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_N \end{bmatrix}$$

The stacked Jacobian $J_{\text{stack}} \in \mathbb{R}^{(N \cdot V) \times d}$ has null space:

$$\ker(J_{\text{stack}}) = \bigcap_{i=1}^{N} \ker(J_\ell(x_i))$$

Critical observation: If all $J_\ell(x_i)$ share a common null space (e.g., all prompts probe similar aspects of the model), then:

$$\ker(J_{\text{stack}}) = \ker(J_\ell(x_1)) \neq \{0\}$$

Adding more prompts does not reduce the null space when all prompts probe the same effective subspace—the ambiguity persists.

```
Prompt 1:  measures dims [1, ..., 3200] → null space = [3201, ..., 4096]
Prompt 2:  measures dims [1, ..., 3200] → null space = [3201, ..., 4096]
                                   ⋮
Prompt N: measures dims [1, ..., 3200] → null space = [3201, ..., 4096]

Intersection of null spaces = [3201, ..., 4096] (unchanged!)
```

Figure 4: Why prompt diversity fails to resolve ambiguity. Each prompt measures the same effective rank $r \approx 3200$ dimensions. The intersection of null spaces remains $\dim(\ker) = d - r \approx 896$ dimensions, regardless of N.

C.3 Why Orthogonal Perturbations Preserve Semantics

Setup: We test steering with v versus $v + v_\perp$, where $v_\perp \perp v$ and $\|v_\perp\| = 1$. Empirically, adding a random orthogonal direction produces nearly equivalent semantic effects. Why?

Decomposing the perturbation. Any perturbation $v_\perp$ decomposes as:

$$v_\perp = v_{\perp,\text{row}} + v_{\perp,\text{null}}$$

where $v_{\perp,\text{row}} \in \text{row}(J)$ (observable) and $v_{\perp,\text{null}} \in \ker(J)$ (invisible).

For random $v_\perp \perp v$, the expected null-space fraction is:

$$\mathbb{E}[\|v_{\perp,\text{null}}\|^2] \approx \frac{\dim(\ker(J))}{d} \approx 0.20\text{–}0.25$$

Thus ∼20–25% of the perturbation is automatically invisible.

Observable effect is diffuse. The output change is:

$$J(v + v_\perp) = Jv + J\,v_{\perp,\text{row}}$$

While $Jv_{\perp,\text{row}}$ is not necessarily small in norm, it represents a *random direction* in ∼3200 dimensions. In contrast, Jv is semantically structured (e.g., consistently shifting toward "honesty").

The key distinction is *semantic coherence*: Jv produces aligned directional changes, while $Jv_{\perp,\text{row}}$ produces diffuse, incoherent perturbations that do not systematically shift meaning.

Analogy: If Jv is wind blowing north, then $Jv_{\perp,\text{row}}$ is turbulence—it may have comparable energy but lacks directional coherence.

D Mathematical Derivations

D.1 Dimension of Observational Equivalence Class

Setup: For steering vector $v \in \mathbb{R}^d$ and Jacobian $J \in \mathbb{R}^{V \times d}$ with rank r, consider:

$$[v] = \{v' \in \mathbb{R}^d : Jv' = Jv\}$$

Question: How many degrees of freedom exist in the equivalence class of observationally equivalent steering vectors?

Analysis: The equivalence class is an affine subspace:

$$[v] = v + \ker(J)$$

where $\ker(J)$ is a $(d - r)$-dimensional linear subspace.

Parameterization: Let $\{u_1, \ldots, u_{d-r}\}$ be an orthonormal basis for $\ker(J)$. Then:

$$[v] = \left\{ v + \sum_{i=1}^{d-r} \alpha_i u_i : \alpha_i \in \mathbb{R} \right\}$$

The equivalence class has $(d - r)$ degrees of freedom.

Scaling ambiguity: Additionally, v and cv produce equivalent outputs (up to rescaling α). The equivalence class modulo scaling is a $(d - r)$-dimensional projective space.

Illustrative example. Suppose a model has hidden size $d = 4096$ and an effective Jacobian rank $r \approx 3100$. Then the equivalence class dimension is $d - r \approx 996$.

Interpretation: For every identifiable parameter (row-space direction), there are $996/3100 \approx 0.32$ unidentifiable parameters (null-space directions). The under-determination ratio is approximately $1 : 3$.

D.2 Fisher Information and Cramér-Rao Bound

Setup: Consider the statistical model:

$$o(x; v, \alpha) \sim p(o|x, v, \alpha)$$

where v is the parameter to estimate.

Question: Can we achieve better identifiability by collecting more samples?

Fisher Information Matrix: For the linear Gaussian model $o = Jv + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2 I)$:

$$\mathcal{I}(v) = \mathbb{E}_o \left[\left(\frac{\partial \log p(o|x, v, \alpha)}{\partial v} \right) \left(\frac{\partial \log p(o|x, v, \alpha)}{\partial v} \right)^\top \right] = \frac{1}{\sigma^2} J^\top J$$

Cramér-Rao Lower Bound: The covariance of any unbiased estimator $\hat{v}$ satisfies:

$$\text{Cov}(\hat{v}) \succeq \mathcal{I}(v)^{-1} = \sigma^2 (J^\top J)^+$$

where $(J^\top J)^+$ is the Moore-Penrose pseudo-inverse.

Implications: For null-space directions $u_i \in \ker(J)$, the Fisher information is degenerate:

$$u_i^\top (J^\top J) u_i = 0$$

Therefore, the variance of any unbiased estimator along null-space directions is unbounded:

$$\text{Var}(u_i^\top \hat{v}) \geq u_i^\top \mathcal{I}(v)^{-1} u_i = \infty$$

Conclusion: The Cramér-Rao bound is **infinite** for null-space components, confirming that no finite amount of data can resolve the ambiguity. Non-identifiability is fundamental, not a small-sample problem. More data cannot help because the information geometry has infinite uncertainty in null-space directions.

D.3 Proof that ICA Breaks Gauge Symmetry

Setup: Consider two decompositions:

$$h = A_1 z_1 = A_2 z_2$$

where both z_1 and z_2 have independent components.

Question: When are these equivalent ($A_1 = A_2$ up to permutation/scaling)?

Analysis: If both decompositions are valid:

$$A_1 z_1 = A_2 z_2 \implies z_1 = A_1^{-1} A_2 z_2 = G z_2$$

where $G = A_1^{-1} A_2$.

ICA Constraint: If both z_1 and z_2 have independent components:

$$I(z_{1,i}; z_{1,j}) = 0 \text{ for } i \neq j$$
$$I((Gz_2)_i; (Gz_2)_j) = 0 \text{ for } i \neq j$$

Key Theorem (Comon, 1994): If z_2 has independent components and Gz_2 also has independent components, then G must be a generalized permutation matrix:

$$G = PD$$

where P is a permutation matrix and D is a diagonal scaling matrix.

Implication:

$$A_1 z_1 = A_2 z_2$$
$$A_1 G z_2 = A_2 z_2$$
$$A_1 P D = A_2$$

Therefore $A_2 = A_1 PD$, meaning A_2 is A_1 with permuted and scaled columns.

Conclusion: ICA constraints (statistical independence) force G to be a permutation-scaling matrix, breaking arbitrary gauge transformations and establishing identifiability up to unavoidable symmetries. Independence is the structural assumption that eliminates the infinite equivalence class.

E Disclosure of the Use of Artificial Intelligence

In accordance with ICLR 2026 policies on large language model usage, the authors used AI as follows: ChatGPT for assisting with refining sections of the manuscript (language polishing and clarity improvements) and Claude Code and GitHub Copilot for coding assistance. Authors take full responsibility for all research ideas, technical content, proofs, experiments and final text.