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:

- Statistical independence between z and x (confounding is allowed)
- Linearity of g_ℓ in z
- Uniqueness of the latent representation without additional structure

These assumptions are mild and satisfied by standard transformer architectures under typical operating conditions.

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):

$$v \sim_{\mathcal{R}} v' \iff F_{\ell \to L}(h_\ell(x) + \alpha v) = F_{\ell \to L}(h_\ell(x) + \alpha v') \; \forall x, \alpha. \tag{1}$$

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 Main Results

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 and 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 and 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: by characterizing sufficient conditions under which identifiability could be recovered–such as statistical independence (ICA), sparsity priors, multi-environment access or cross-layer consistency. Proposition 2 clarifies precisely which structural assumptions are necessary to break the symmetries that cause non-identifiability.

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:** 24 layers, hidden dimension $d = 2048$, layer $\ell = 12$ (mid-network)

- **Llama-3.1-8B-Instruct:** 32 layers, hidden dimension $d = 4096$, layer $\ell = 16$ (mid-network)

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. Middle layers balance semantic abstraction with steerability: early layers encode low-level features (tokens, syntax), while late layers specialize for next-token prediction.

Persona traits. We test three semantic traits spanning distinct dimensions:

- **Formality:** Formal versus informal style
- **Politeness:** Polite versus rude social register
- **Humor:** Humorous versus serious content

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.

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