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.

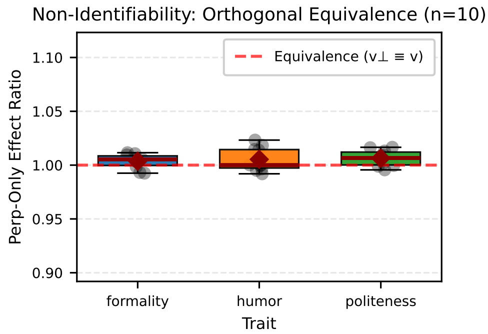
Model	Trait	Seeds	Cohen's d	Correlation	Perp-Only
Qwen2.5-3B-Instruct	Formality	$n = 5$	0.144 ± 0.095	0.210 ± 0.113	98.8%
		$n = 10$	0.075 ± 0.058	0.285 ± 0.087	100.4%
	Politeness	$n = 5$	0.112 ± 0.077	0.319 ± 0.070	101.5%
		$n = 10$	0.092 ± 0.069	0.414 ± 0.083	100.6%
	Humor	$n = 5$	0.103 ± 0.077	0.276 ± 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%

confidence intervals: formality ($0.144 \to 0.075$), politeness ($0.112 \to 0.092$), humor ($0.103 \to 0.072$). For Llama-3.1-8B, we observe similar stability: formality ($0.052 \to 0.096$), politeness ($0.074 \to 0.085$), humor ($0.159 \to 0.119$). This consistency demonstrates that the observed equivalence is not a sampling artifact but a stable property of the steering geometry.

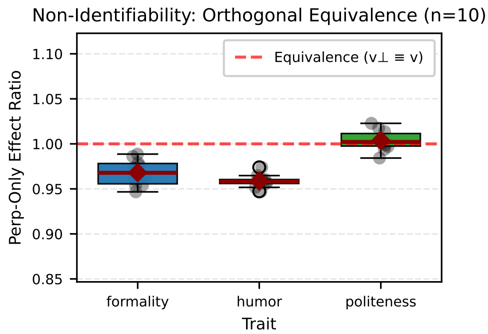
Cross-model consistency. The close agreement between Qwen2.5-3B ($d = 0.080$) and Llama-3.1-8B ($d = 0.100$) demonstrates that observational equivalence is not model-specific. Despite differences in architecture, scale (3B vs. 8B parameters) and hidden dimension ($d = 2048$ vs. $d = 4096$), both models 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 orthogonally steering still retains over 95% efficacy, far exceeding what would be expected if v were uniquely identifiable.



(a) Qwen2.5-3B-Instruct



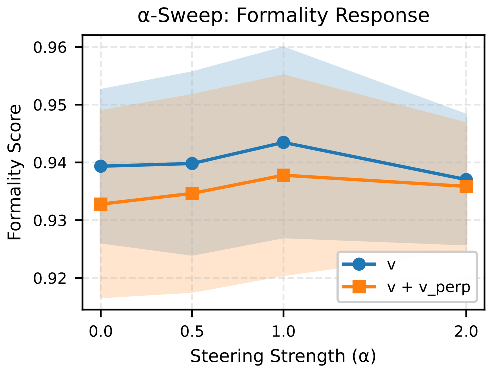
(b) Llama-3.1-8B-Instruct

Figure 1: Perp-only effect ratios ($v_{\perp}$ efficacy / v efficacy) for $n = 10$ orthogonal seeds. Values near 1.0 (dashed red line) indicate perfect equivalence.

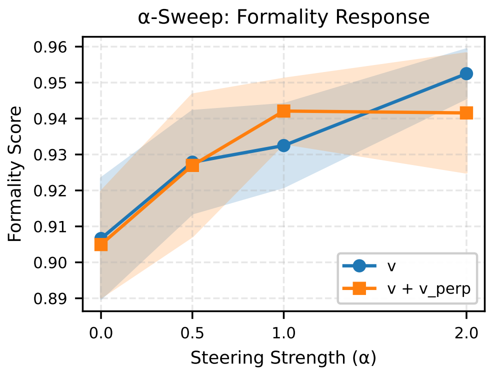
6.3 Scale Invariance Analysis

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 $v + v_{\perp}$



(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).

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 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.

8 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, 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.

9 Future Work

Systematic evaluation of identifiable regimes, comparing ICA-based extraction, sparse recovery, multi-environment training and cross-layer validation against standard contrastive methods, would clarify which structural assumptions are most effective in practice. Hybrid methods combining multiple identifiability conditions may achieve stronger identifiability than individual approaches, while layer-wise analysis could reveal whether certain network positions afford more identifiable steering and decomposing the Jacobian by architectural component could identify which elements contribute to non-identifiability.

The multi-environment identifiability condition connects to causal representation learning, where adapting invariant risk minimization to train steering vectors that maintain consistent effects across distributions could filter spurious correlations. Investigating adversarial robustness and out-of-distribution generalization would test whether identifiable methods exhibit superior robustness. Characterizing scaling laws for identifiability, including how null-space dimensionality evolves as models scale, would inform long-term strategy for controlling future systems.

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, and 1 others. 2022. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 35:3438–3450.

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

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

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

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

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

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

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

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

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

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

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

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

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

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

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

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

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

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

Joseph B Kruskal. 1977. 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.

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

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a