sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

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

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

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

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

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

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

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

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

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

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

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

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

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. 2023. Linear representations of sentiment in large language models. *arXiv preprint arXiv:2310.15154*.

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

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

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

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

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Prompt Construction Details

A.1 Contrastive Prompt Pairs

We construct contrastive prompt pairs for each semantic trait using template-based generation. Figure 3 illustrates the general structure and Table 2 provides concrete examples.

A.2 Steering Extraction and Evaluation Pipeline

Figure 4 illustrates the pipeline used to extract and evaluate steering vectors. Contrastive prompt pairs are first used to compute a steering vector specific to a semantic trait. This vector is then applied to a separate set of held-out evaluation prompts to generate steered outputs.

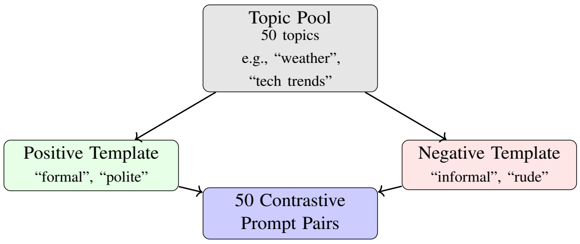

Figure 3: Contrastive prompt pair construction.

Trait	Positive Prompt (x^+)	Negative Prompt (x^-)
Formality	Write a professional and formal message about *technology trends*.	Write a casual and informal message about *technology trends*.
Politeness	Write a polite and courteous response for *a disagreement with someone*.	Write a rude and disrespectful response for *a disagreement with someone*.
Humor	Write a humorous and funny response about *an awkward moment*.	Write a serious and straightforward response about *an awkward moment*.

Table 2: Example contrastive prompt pairs for each semantic trait. Each pair consists of a positive prompt (x^+) designed to elicit high trait values and a negative prompt (x^-) designed to elicit low trait values.

B Detailed Proof of Proposition 1

B.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).

B.2 Null-Space Ambiguity (Constructive Proof)

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

$$o(x, v, \alpha) \approx o(x, 0, 0) + \alpha J_\ell(x) v \tag{2}$$

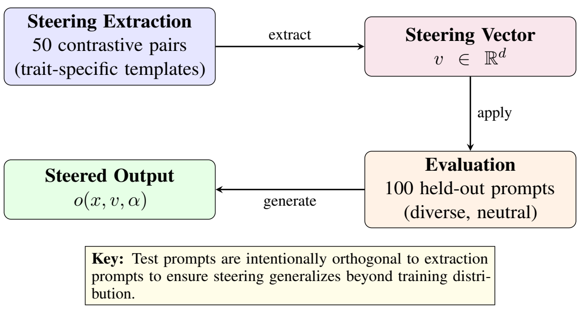

Figure 4: Steering extraction and evaluation pipeline. Steering vectors are extracted from trait-specific contrastive pairs, then evaluated on diverse held-out prompts to test generalization.

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:

$$\mathcal{N} = \ker(J_\ell) = \{v_0 \in \mathbb{R}^d : J_\ell v_0 = 0\} \tag{3}$$

By the rank-nullity theorem:

$$\dim(\mathcal{N}) = d - \operatorname{rank}(J_\ell) \tag{4}$$

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

$$\operatorname{rank}(J_\ell) \leq \min(V, d) \tag{5}$$

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:

$$\operatorname{rank}_\epsilon(J_\ell) = \#\{\sigma_i : \sigma_i > \epsilon \cdot \sigma_{\max}\} \ll d \tag{6}$$

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

$$v' = v + v_0. \tag{7}$$

Then for all x and all α:

$$\begin{aligned} J_\ell(x) v' &= J_\ell(x)(v + v_0) \\ &= J_\ell(x) v + J_\ell(x) v_0 \\ &= J_\ell(x) v. \end{aligned} \tag{8}$$

Therefore:

$$\begin{aligned} o(x, v', \alpha) &\approx o(x, 0, 0) + \alpha J_\ell(x) v' \\ &= o(x, 0, 0) + \alpha J_\ell(x) v \\ &\approx o(x, v, \alpha). \end{aligned} \tag{9}$$

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 $\beta \in \mathbb{R}$:

$$v'_\beta = v + \beta v_0 \tag{10}$$

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:

$$\beta v_0 = (c-1)v \tag{11}$$

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.

B.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:

$$h'_\ell(x) = T\big(h_\ell(x)\big). \tag{12}$$

If the subsequent layers can be rewritten as:

$$F_{\ell \to L}(h_\ell) = F'_{\ell \to L}(T(h_\ell)) \tag{13}$$

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:

$$h_{\ell+1} = \sigma(W_{\ell+1} h_\ell + b_{\ell+1}) \tag{14}$$

Reparameterized computation:

$$h'_{\ell+1} = \sigma(W'_{\ell+1} h'_\ell + b'_{\ell+1}) \tag{15}$$

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:

$$\begin{aligned} h'_{\ell+1} &= \sigma\big(W_{\ell+1} A^{-1} A h_\ell + b_{\ell+1}\big) \\ &= \sigma(W_{\ell+1} h_\ell + b_{\ell+1}) \\ &= h_{\ell+1}. \end{aligned} \tag{16}$$

Step 3: Steering under reparameterization. Original steering:

$$\tilde{h}_\ell = h_\ell + \alpha v \tag{17}$$

$$\tilde{h}_{\ell+1} = \sigma(W_{\ell+1}(h_\ell + \alpha v) + b_{\ell+1}) \tag{18}$$

Reparameterized steering with $v' = Av$:

$$\begin{aligned} \tilde{h}'_\ell &= h'_\ell + \alpha v' \\ &= A h_\ell + \alpha A v \\ &= A(h_\ell + \alpha v) \\ &= A \tilde{h}_\ell, \\ \tilde{h}'_{\ell+1} &= \sigma\big(W'_{\ell+1}(h'_\ell + \alpha v') + b'_{\ell+1}\big) \\ &= \sigma\big(W_{\ell+1} A^{-1} A (h_\ell + \alpha v) + b_{\ell+1}\big) \\ &= \sigma(W_{\ell+1}(h_\ell + \alpha v) + b_{\ell+1}) \\ &= \tilde{h}_{\ell+1}. \end{aligned} \tag{19}$$

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