Non-linear Interventions on Large Language Models

Sangwoo Kim [1]

Abstract

Intervention is one of the most representative and widely used methods for understanding the internal representations of large language models (LLMs). However, existing intervention methods are confined to linear interventions grounded in the Linear Representation Hypothesis, leaving features encoded along non-linear manifolds beyond their reach. In this work, we introduce a general formulation of intervention that extends naturally to non-linearly represented features, together with a learning procedure that further enables intervention on *implicit* features lacking a direct output signature. We validate our framework on refusal bypass steering, where it steers the model more precisely than linear baselines by intervening on a non-linear feature governing refusal.

1. Introduction

Intervention is a central tool in understanding inner representations of large language models (LLMs). By modifying a model's internal activations during inference and observing the resulting change in output, interventions yield causal evidence linking specific components to model behavior (Vig et al., 2020; Meng et al., 2022; Geiger et al., 2024). Through intervention, we can understand how interpretable features are represented in the model's hidden states (Huang et al., 2024), and by doing so, effectively control via steering LLM behaviors that are hard to address through prompting alone, such as response style (Turner et al., 2025), hallucination (Li et al., 2023), and refusal (Arditi et al., 2024).

Most existing methods for intervening on such features are linear: they modify hidden states by adding or ablating a fixed direction in activation space (Li et al., 2023; Arditi et al., 2024). This design rests on the Linear Representation Hypothesis (LRH), which posits that interpretable concepts are encoded as directions in the model's representation space (Mikolov et al., 2013; Park et al., 2023). Recent work, however, has shown that some concepts are instead represented along non-linear manifolds—for example, days of the week organized as a circular structure (Engels et al., 2025), and the geometric structures used to perform counting (Gurnee et al., 2025). Since linear interventions can by construction only manipulate features encoded as directions, they cannot reach this class of non-linearly represented structure.

To overcome this limitation, we introduce a general formulation of non-linear intervention. We generalize linear interventions by replacing their underlying linear transformation with an invertible non-linear feature map. In addition, we propose a procedure for learning this map via *interchange interventions* (Geiger et al., 2024), together with a loss design that extends the procedure to *implicit* features.

We instantiate the framework on refusal bypass steering (Arditi et al., 2024; Wollschläger et al., 2025), a representative implicit-feature intervention task. Our non-linear intervention attains steering effectiveness comparable to strong linear baselines while editing activations at orders of magnitude fewer hidden-state locations. Further analysis indicates that this advantage stems from a genuinely non-linear feature map governing refusal that resides primarily at the model's middle layers.

Our work makes three contributions. First, we propose a general formulation of non-linear intervention that subsumes existing linear interventions and naturally extends to features encoded along non-linear manifolds. Second, we introduce a learning procedure for the non-linear feature map based on interchange interventions, together with a loss design that further enables learning *implicit* features that do not surface in the model's output. Third, we empirically apply our framework to refusal bypass steering and show that intervening on a non-linear feature governing refusal steers the model more precisely than linear baselines. Code is available at `https://anonymous.4open.science/r/nonlinear-intervention-77AC/`.

[1]Department of Linguistics, Seoul National University, Republic of Korea. Correspondence to: Sangwoo Kim <hemy0101@snu.ac.kr>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

2. A General Formulation of Non-linear Interventions

2.1. Linear Interventions as Change of Basis

A common approach to intervene on a language model $\mathcal{M}$ at inference time is to identify *linear feature directions* in its representation space and perturb the hidden state along them. Given orthonormal feature directions $\{\mathbf{v}_i\}_{i=1}^k \subset \mathbb{R}^d$ and corresponding scalar coefficients $\{\alpha_i\}_{i=1}^k$, this linear intervention modifies the hidden state $\mathbf{h} \in \mathbb{R}^d$ at a chosen position as

$$\mathbf{h} \leftarrow \mathbf{h} + \sum_{i=1}^{k} \alpha_i \mathbf{v}_i. \tag{1}$$

Eq. (1) can equivalently be viewed as a change of basis. Let $W \in \mathbb{R}^{d\times d}$ be an orthogonal matrix that maps the representation space onto a *linear feature space*, satisfying $W^\top \mathbf{e}_i = \mathbf{v}_i$, where $\{\mathbf{e}_i\}_{i=1}^d$ is the standard basis of $\mathbb{R}^d$. Each coordinate of $W\mathbf{h}$ is then the activation value of one feature at $\mathbf{h}$, and Eq. (1) can be rewritten as

$$\mathbf{h} \leftarrow W^{-1}\left(W\mathbf{h} + \sum_{i=1}^{k} \alpha_i \mathbf{e}_i\right), \tag{2}$$

that is, W maps $\mathbf{h}$ into the feature space, the intervention perturbs the resulting feature coordinates along their axes, and W^{-1} maps back to the original representation space.

2.2. Non-linear Interventions via Invertible Feature Maps

Linear interventions can only manipulate features that admit a linear encoding, yet not all interpretable features in $\mathcal{M}$'s representation space lie along linear directions (Engels et al., 2025; Kantamneni & Tegmark, 2025). To accommodate non-linear features, we replace W with an invertible *non-linear feature map* $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ parameterized by θ. By direct analogy with Eq. (2), the non-linear intervention takes the form

$$\mathbf{h} \leftarrow f_\theta^{-1}\left(f_\theta(\mathbf{h}) + \sum_{i=1}^{k} \alpha_i \mathbf{e}_i\right). \tag{3}$$

f_θ maps $\mathbf{h}$ to its feature-space coordinates; the intervention perturbs these coordinates along their axes by the coefficients $\{\alpha_i\}$; and f_θ^{-1} maps the modified features back into the original hidden-state space. Eq. (2) is recovered as the special case in which f_θ is restricted to a linear map.

3. Learning f_θ via Interchange Intervention

3.1. Training Objective via Interchange Intervention

This section describes how we learn the non-linear feature map f_θ associated with a target feature $\mathcal{F}$. We train f_θ via *interchange interventions* (Geiger et al., 2024), which provide causal supervision over $\mathcal{F}$ by transferring its value between hidden states drawn from contrasting inputs. We first prepare two sets of prompts: a positive set $\mathcal{D}^+ = \{x_i^+\}_{i=1}^N$ of inputs that exhibit the feature $\mathcal{F}$, and a negative set $\mathcal{D}^- = \{x_i^-\}_{i=1}^N$ of inputs that do not. For each pair $(x^-, x^+) \in \mathcal{D}^- \times \mathcal{D}^+$, we forward both inputs through $\mathcal{M}$ up to the intervention site, obtaining hidden states $\mathbf{h}^-, \mathbf{h}^+ \in \mathbb{R}^d$. The *interchange intervention* replaces the targeted coordinates of $f_\theta(\mathbf{h}^-)$ with those of $f_\theta(\mathbf{h}^+)$:

$$\mathbf{h}^- \leftarrow f_\theta^{-1}\left(f_\theta(\mathbf{h}^-) + \sum_{i=1}^{k} \alpha_i \mathbf{e}_i\right), \tag{4}$$

where $\alpha_i = \big(f_\theta(\mathbf{h}^+) - f_\theta(\mathbf{h}^-)\big)_i$. This recovers Eq. (3) with the coefficients pinned to x^+. Let $\mathcal{M}_{\text{int}}(x^-, x^+; \theta)$ denote the intervened forward pass of $\mathcal{M}$. With pairs sampled as $(x^-, x^+) \sim \mathcal{D}^- \times \mathcal{D}^+$ and $\mathcal{M}$ frozen, we train θ to minimize

$$\mathcal{L}(\theta) = \mathbb{E}_{(x^-,\, x^+)}\big[\, \ell\big(\mathcal{M}_{\text{int}}(x^-, x^+; \theta)\big)\,\big], \tag{5}$$

where ℓ is designed so that its minimization makes $\mathcal{M}_{\text{int}}$ exhibit $\mathcal{F}$.

3.2. Loss Design for Implicit Features

When $\mathcal{F}$ has a direct output signature, ℓ can be defined on $\mathcal{M}$'s output distribution to encourage or suppress specific tokens. For implicit features that do not surface in the output, such as refusal or stylistic shifts, no such direct objective for training f_θ is available, making the loss design non-trivial. We propose a self-supervised-style loss that learns f_θ from data alone, by enforcing causal influence over many features correlated with $\mathcal{F}$.

For each (layer, token position) site s lying causally downstream of the intervention site, we extract a feature direction $v_s \in \mathbb{R}^d$ as the class-mean difference of unintervened activations $h_s(x)$ over $\mathcal{D}^+$ versus $\mathcal{D}^-$. Projections onto v_s are therefore correlated with $\mathcal{F}$. We keep only those sites at which $v_s^\top h_s(x)$ separates x^+ from x^- with AUC above a threshold τ, collected into $\mathcal{S}$. Both v_s and $\mathcal{S}$ are computed once from $\mathcal{D}^\pm$ before training and held fixed.

Letting h_s^{int} denote the hidden state at s under the intervened forward pass and $\mu_s^+ = \mathbb{E}_{x\sim\mathcal{D}^+}[\, v_s^\top h_s(x)\,]$ the mean projection of x^+, we take ℓ in Eq. (5) to be

$$\ell(\mathcal{M}_{\text{int}}) = \sum_{s\in\mathcal{S}} \max\big(0,\ \ \mu_s^+ - v_s^\top h_s^{\text{int}}\big). \tag{6}$$

The hinge form saturates once a site's projection reaches μ_s^+, preventing overfitting to any single site. Minimizing $\mathcal{L}$ thus drives f_θ to discover a feature whose intervention causally aligns many $\mathcal{F}$-correlated components at once.

This particular instantiation is one natural choice. The direction extractor, site-selection criterion, and surrogate loss are not fixed; each can be substituted within the framework of Eq. (5).

4. Experiments

4.1. Setup

We evaluate non-linear intervention on refusal bypass steering in safety-aligned LLMs, a representative steering task, to empirically validate our proposed non-linear intervention framework. We use this setting as an implicit-feature intervention task: the target feature $\mathcal{F}$ is *bypass refusal*, which is not explicitly manifested in any specific token output of the model.

Models and data. We evaluate all methods on Llama-3-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Qwen et al., 2025), with model weights frozen throughout. We construct $\mathcal{D}^+$ from 2,000 harmless Alpaca prompts (Taori et al., 2023) that elicit compliant responses from the model, and $\mathcal{D}^-$ from 2,000 harmful SALAD-Bench prompts (Li et al., 2024) that elicit refusals.

Evaluation. We quantitatively evaluate intervention quality along two axes: *how strongly* the intervened model exhibits the target feature $\mathcal{F}$, and *how much* we perturbed the model to achieve it. To measure whether the intervened model produces genuinely meaningful responses to harmful prompts rather than merely avoiding surface-level refusals, we use the StrongREJECT score (Souly et al., 2024). StrongREJECT returns a score in $[0, 1]$ over 313 harmful prompts, computed by a LLM-based evaluator. To measure how much we intervened, we record the total intervention magnitude applied to the model. Specifically, for every edited site, we compute the ℓ_2 distance between the pre-intervention hidden state h and the post-intervention hidden state h', and sum them:

$$\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{x \in \mathcal{D}_{\text{test}}} \sum_{s \in \mathcal{E}_x} \|h'_s(x) - h_s(x)\|_2,$$

where $\mathcal{E}_x$ is the set of sites the method edits when generating a response to prompt x.

Baselines. As baselines, we compare against two representative linear intervention methods for refusal steering: *Difference In Means* (DIM) (Arditi et al., 2024) and *Refusal Direction Optimization* (RDO) (Wollschläger et al., 2025). DIM extracts a refusal direction from the class-mean activation difference. RDO learns a refusal direction with the ablation, addition, and retain losses. Both methods can be evaluated under two intervention schemes: *ablation*, which projects out the refusal direction at every token and every module output, and *actadd*, which adds the direction scaled by a fixed coefficient α at a designated layer for every token.

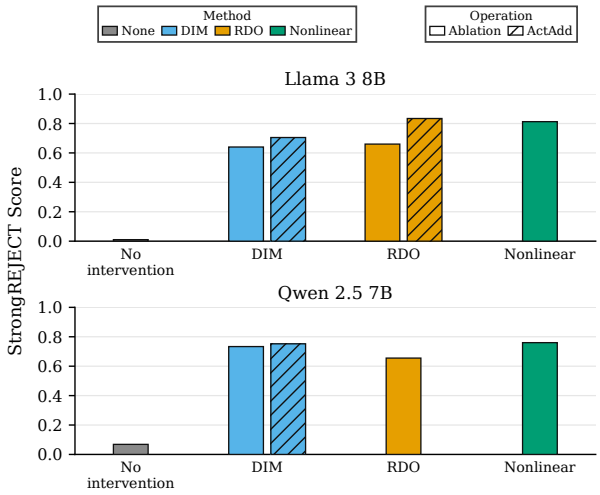

Figure 1. StrongREJECT scores on Llama 3 8B and Qwen 2.5 7B without intervention, with the linear baselines, and with our non-linear intervention.[1]

Our method. We instantiate f_θ as an i-ResNet (Behrmann et al., 2019), an invertible non-linear neural network. We use $k = 1$ for direct comparison with one-dimensional linear baselines. We heuristically select one intervention position per model. At inference time, to avoid requiring a harmless prompt x^+ for each query, we precompute the mean feature activation over a set of harmless prompts x^+ and use this cached value. Specifically, we compute the mean first-coordinate value of $f_\theta(h)$ over the harmless training set at the selected site, denoted $\bar{\mu}^+$, and clamp each test activation to this value.

4.2. Results

Figure 1 reports StrongREJECT scores under no intervention, the linear baselines, and our non-linear intervention. Although the non-linear method intervenes at a single position per sample, it reaches scores comparable to the baselines, which instead modify activations at a much larger number of sites.

As shown in Table 1, the total magnitude of the non-linear intervention is more than two orders of magnitude smaller than that of every linear baseline. This empirically suggests that our non-linear intervention captures the feature inside the model more precisely than the linear baselines do.

[1]Although we faithfully reproduced the official RDO repository, the RDO ActAdd variant on Qwen yielded a StrongREJECT score of 0.020 in our runs. We judged this result to be unreliable and therefore excluded it from the figure and table. Wollschläger et al. 2025 reports ActAdd performance comparable to that of the Ablation variant.

Table 1. Per-sample intervention magnitude. *Sites*: avg. positions edited per example; L_2: total edit norm. Our non-linear method edits one site, with an L_2 norm over two orders of magnitude below every linear baseline.

Model	Method	Sites	L_2
Llama 3 8B	DIM (Ablation)	54,778	2,898.7
	DIM (ActAdd)	572	1,816.2
	RDO (Ablation)	59,396	1,812.1
	RDO (ActAdd)	619	1,964.5
	non-linear	**1**	**3.5**
Qwen 2.5 7B	DIM (Ablation)	49,661	41,376.2
	DIM (ActAdd)	591	24,676.3
	RDO (Ablation)	53,568	15,474.0
	non-linear	**1**	**28.3**

Table 2. StrongREJECT after intervening linearly at every loss site $s \in \mathcal{S}$ along v_s, versus our learned non-linear intervention at a single site.

Method	Llama 3 8B	Qwen 2.5 7B
Linear at loss sites	0.689	0.415
Non-linear (ours)	**0.813**	**0.760**

Table 3. StrongREJECT when f_θ is constrained to a linear map, versus our i-ResNet f_θ, trained under the same objective.

Method	Llama 3 8B	Qwen 2.5 7B
Linear f_θ	0.010	0.072
Non-linear f_θ (ours)	**0.813**	**0.760**

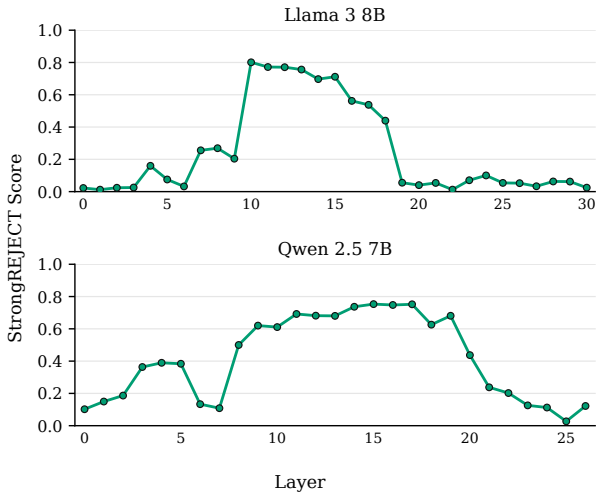

Figure 2. StrongREJECT across intervention layers on Llama 3 8B and Qwen 2.5 7B with our non-linear intervention.

5. Analysis

5.1. Linear Intervention at the Loss Sites

Our objective in Eq. (6) pushes $v_s^\top h_s^{\text{int}}$ toward μ_s^+ at every site $s \in \mathcal{S}$. We ask whether the gain over linear baselines comes solely from our choice of $\mathcal{S}$ and the directions $\{v_s\}$. Concretely, at every $s \in \mathcal{S}$ we apply a linear intervention along v_s with a coefficient chosen so that $v_s^\top h_s$ is shifted to μ_s^+. As shown in Table 2, this falls substantially short of our method on both models, indicating that f_θ does more than align these projections: it implements a non-linear, causal manipulation of the refusal feature that *induces* the per-site alignment as a downstream effect.

5.2. Role of Non-linearity

To examine the role of f_θ's non-linearity, we re-train it under the same loss and pipeline as Section 3, but constrained to a linear map. As reported in Table 3, StrongREJECT drops sharply on both models, suggesting that the non-linearity of f_θ clearly contributes to the effectiveness of our intervention.

5.3. Layer Sweep

To check whether our non-linear intervention is effective at any layer, we repeat the same training and evaluation under the same setup while varying only the intervention layer. Figure 2 shows that, for both models, the intervention is largely ineffective at early layers, peaks at middle layers, and declines at later layers. This suggests that our method is not so powerful that it finds an effective feature map at any layer; rather, it captures a non-linear feature governing refusal/compliance that resides primarily in the middle layers.

6. Conclusion

We overcome the linearity restriction of existing interventions by providing a general formulation of non-linear intervention. Beyond the formulation itself, we propose a procedure for learning the non-linear feature map via interchange interventions, together with the idea of supervising this procedure for *implicit* features through hidden states of the model that are correlated with the target feature. We instantiate this framework on bypass refusal steering, a representative implicit-feature steering task; our experiments and analysis show that non-linear intervention effectively steers refusal behavior and that, in doing so, it uncovers a non-linear feature inside the model.

Several limitations remain, however. We realize f_θ only as an i-ResNet, and the same framework admits other invertible architectures. We also do not overcome a limitation shared with linear interventions: the position at which to intervene must still be chosen heuristically. We leave both directions to future work.

Acknowledgements

We thank Jongwon Lim of Seoul National University for invaluable help throughout this work, from the experiments to the writing of this paper.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Behrmann, J., Grathwohl, W., Chen, R. T. Q., Duvenaud, D., and Jacobsen, J.-H. Invertible residual networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 573–582. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/behrmann19a.html.

Engels, J., Michaud, E. J., Liao, I., Gurnee, W., and Tegmark, M. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. Finding alignments between interpretable causal variables and distributed neural representations, 2024. URL https://arxiv.org/abs/2303.02536.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I.,

Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gurnee, W., Ameisen, E., Kauvar, I., Tarng, J., Pearce, A., Olah, C., and Batson, J. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html.

Huang, J., Wu, Z., Potts, C., Geva, M., and Geiger, A. RAVEL: Evaluating interpretability methods on disentangling language model representations. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.470. URL https://aclanthology.org/2024.acl-long.470/.

Kantamneni, S. and Tegmark, M. Language models use trigonometry to do addition. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL https://openreview.net/forum?id=CqViN4dQJk.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Li, L., Dong, B., Wang, R., Hu, X., Zuo, W., Lin, D., Qiao, Y., and Shao, J. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Mikolov, T., Yih, W.-t., and Zweig, G. Linguistic regularities in continuous space word representations. In Vanderwende, L., Daumé III, H., and Kirchhoff, K. (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks, 2018. URL https://arxiv.org/abs/1802.05957.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=T0PoOJg8cK.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering, 2025. URL https://openreview.net/forum?id=2XBPdPIcFK.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Wollschläger, T., Elstner, J., Geisler, S., Cohen-Addad, V., Günnemann, S., and Gasteiger, J. The geometry of refusal in large language models: Concept cones and representational independence. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=80IwJqlXs8.

A. Additional Experimental Details

A.1. i-ResNet Feature Map

We instantiate the invertible feature map $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ as a composition of $M = 2$ invertible residual blocks,

$$f_\theta \;=\; \phi_{\theta_M} \circ \cdots \circ \phi_{\theta_1}, \qquad \phi_{\theta_m}(z) \;=\; z + g_{\theta_m}(z).$$

The residual branch $g_{\theta_m} : \mathbb{R}^d \to \mathbb{R}^d$ is a two-layer MLP with hidden width 128 and a LeakyReLU nonlinearity (negative slope 0.1). To guarantee invertibility via $\mathrm{Lip}(g_{\theta_m}) < 1$, we follow Behrmann et al. (2019) and combine two mechanisms: (i) every linear layer W inside g_{θ_m} is replaced at every forward pass by $\widetilde{W} = W \cdot \min(1, 1/\hat{\sigma}(W))$, where $\hat{\sigma}(W)$ is a running estimate of the spectral norm of W obtained by a single power iteration with a persistent left singular vector (Miyato et al., 2018); (ii) the output of g_{θ_m} is rescaled by a Lipschitz coefficient $\kappa = 0.6$, yielding $\mathrm{Lip}(g_{\theta_m}) \leq \kappa < 1$. We invert each block by the contractive fixed-point iteration $x \leftarrow y - g_{\theta_m}(x)$ run for up to 30 steps with relative-residual tolerance 10^{-5}, and propagate gradients through the inverse via implicit differentiation.

A.2. Intervention Sites

All layer indices are zero-indexed transformer block indices. Our non-linear intervention is placed at the `block_output` representation, i.e., the residual-stream state after the selected transformer block. Token positions are negative indices into the fully formatted and tokenized chat prompt after the generation prompt has been appended; position -1 denotes the final prompt token immediately before generation starts.

For the main non-linear experiments, we intervene on exactly one site per example. The sites are:

Table 4. Main non-linear intervention sites. Token positions are counted from the end of the formatted prompt.

Model	Layer	Token position	Evaluation target
Llama 3 8B	10	-3	harmless mapped-coordinate mean
Qwen 2.5 7B	14	-4	harmless mapped-coordinate mean