

Contrastive pair projections on the difference of means line predict optimal steering layer and test-time steering success of individual samples.



Predictable Steering: Leveraging Geometric Proxies for Layer Selection and Per-Instance Success Prediction

Nishkal Hundia, Swastik Agrawal, Navita Goyal, Sarah Wiegrefe

1 Motivation

Steering is unreliable. It can control LLM behavior, but effectiveness varies across concepts, datasets, and layers, forcing costly sweeps. We address these concerns by extending prior work on data separability for:

- **Layer selection:** For a fixed dataset, d' predicts the optimal steering layer, enabling principled layer selection without exhaustive sweeps.
- **Individual test point success:** For datasets with high d' , the position of a test point along the steering vector direction predicts steering success on that example.

2 Background

Contrastive Activation Addition (CAA) steers LLM behavior by adding a vector to the residual stream at a chosen layer. Given contrastive pairs of positive (r^+) and negative (r^-) responses, the steering vector is the mean activation difference:

$$v^{(\ell)} = \mu_{\ell,+} - \mu_{\ell,-}$$

Where $\mu_{\ell,+}$ = mean of positive activations $a^{+(\ell)}$, $\mu_{\ell,-}$ = mean of negative activations $a^{-(\ell)}$.

At inference time, this vector is injected into the residual stream at a chosen layer with a steering factor α .

Discriminability Index (d') is a compute-cheap geometric measure of how well positive and negative training activations separate along the steering direction. Each activation a is projected onto the difference-of-means line to get a scalar coordinate:

$$\kappa_a = 2 \cdot \frac{(a - \mu_{\ell}) \cdot v^{(\ell)}}{\|v^{(\ell)}\|^2}$$

Let P^+ and P^- be the sets of scalar coordinates of the positive and negative training activations, respectively. Then d' is defined as:

$$d' = \frac{|\text{mean}(P^+) - \text{mean}(P^-)|}{\sqrt{\frac{1}{2}(\text{var}(P^+) + \text{var}(P^-))}}$$

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>

Braun, J. Understanding unreliability of steering vectors in language models: Geometric predictors and the limits of linear approximations, 2026. URL <https://arxiv.org/abs/2602.17881>

Braun, J., Eickhoff, C., Krueger, D., Bahrainian, S. A., and Krashennikov, D. Understanding (un)reliability of steering vectors in language models. In ICLR 2025 Workshop on Building Trust in Language Models and Applications, 2025. URL <https://openreview.net/forum?id=JZiKuvKIt>

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. 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. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL <https://aclanthology.org/2024.acl-long.828/>

3 Methods & Results

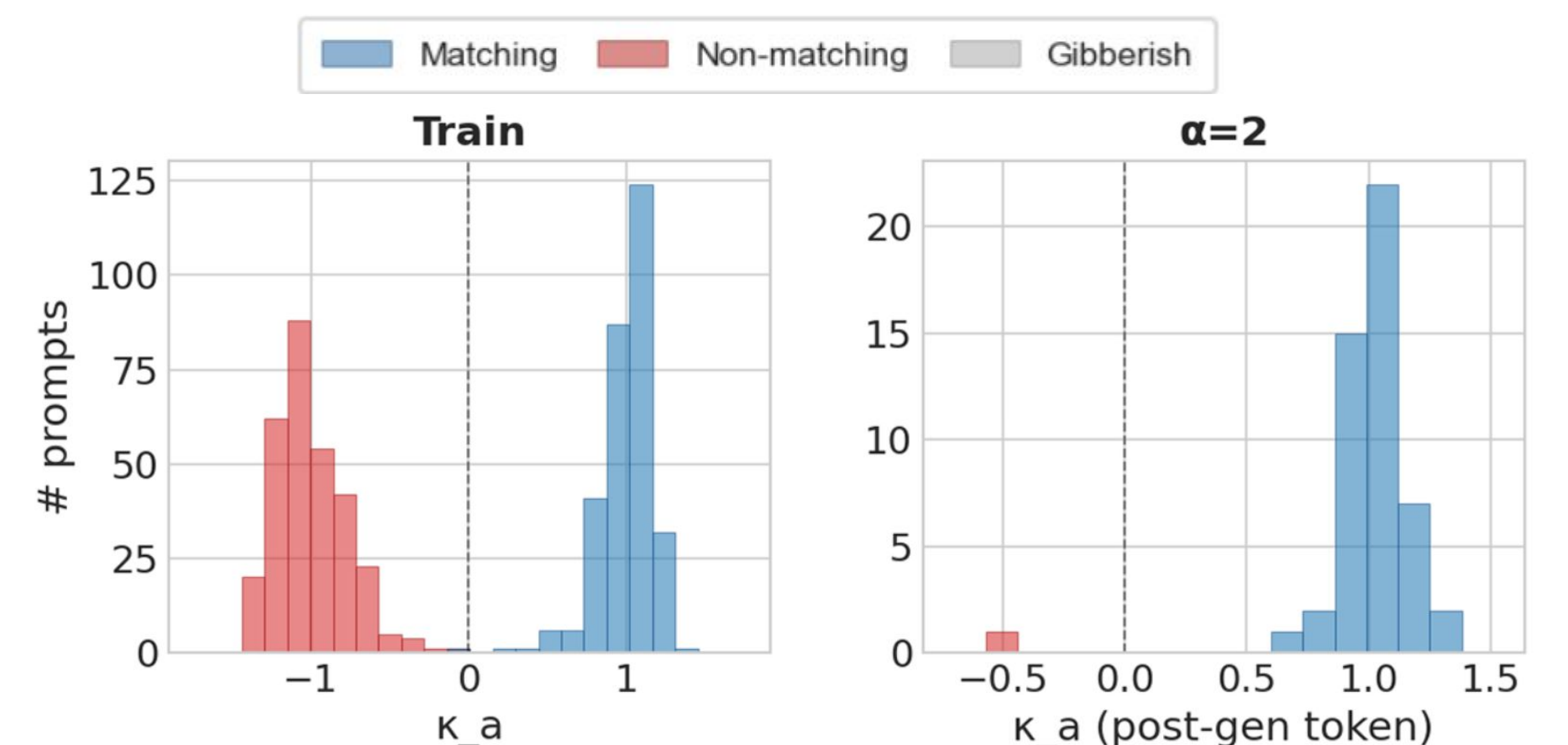
Datasets and Model. Evaluate on five MCQA behavior datasets (sycophancy, survival-instinct, myopic-reward, hallucination, corrigibility) using Gemma-2-9B-IT, with 300 train / 50 validation / 50 test examples each.

Steering Vector Training and Layer Sweep. Train steering vectors at every layer (10–32), sweep α on validation to pick the best α per layer, then evaluate test accuracy to see how steering performance tracks d' across layers.

DATASET	PEARSON r
SYCOPHANCY	0.883*
SURVIVAL-INSTINCT	0.830*
MYOPIC-REWARD	0.814*
HALLUCINATION	0.721*
CORRIGIBILITY	0.554*

Table 1. Pearson correlation between per-layer training d' and per-layer test steering accuracy, with optimal steering factor chosen by validation. Asterisk (*) denotes $p < 0.05$.

Test-Time Projection Analysis. For each steered generation, project the generated token's activation onto the difference-of-means line (κ_a)



and check whether its sign predicts steering success, measured via Matthews Correlation Coefficient.

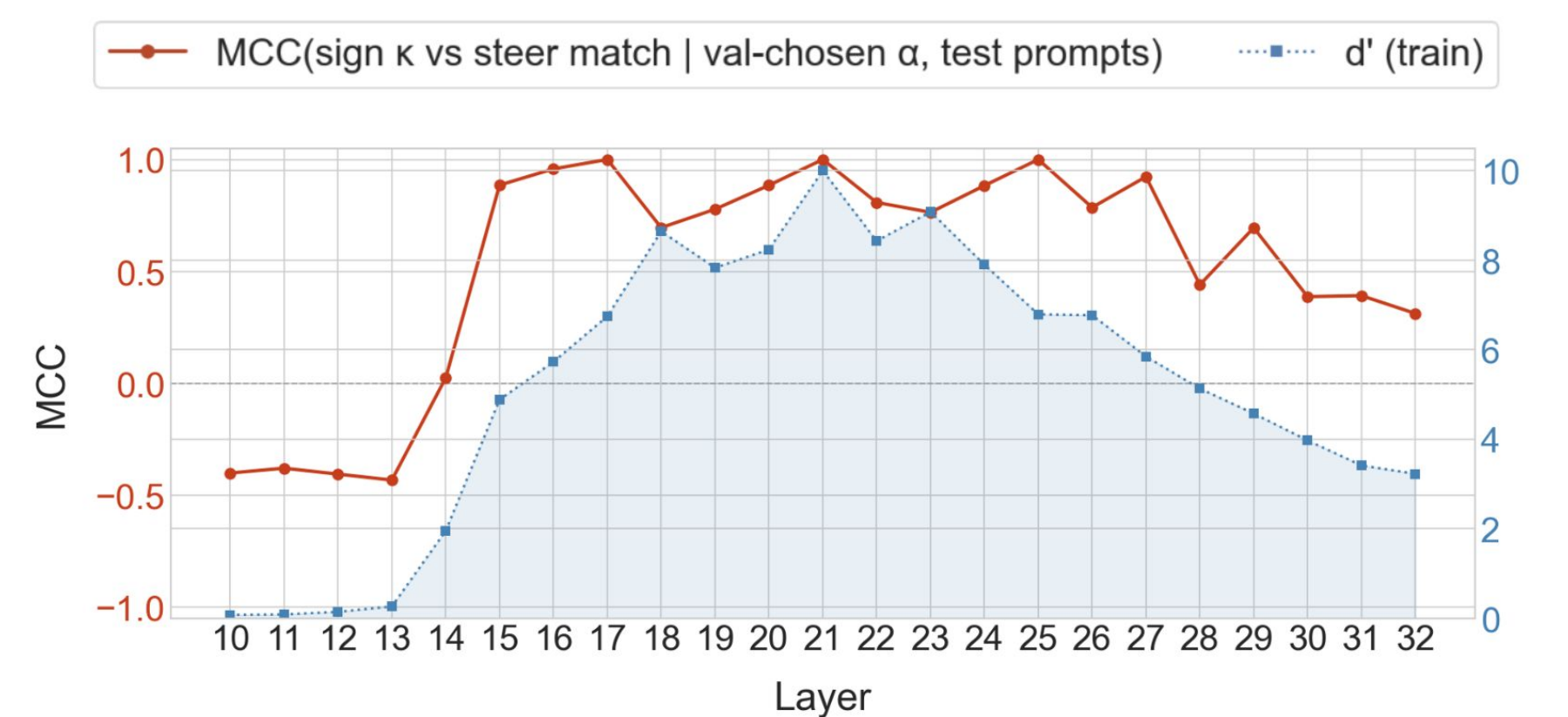


Table 2. Pearson correlation between per-layer training d' and per-layer MCC of $\text{sign}(\kappa_a)$ predicting steering success on the test set, again with optimal steering factor. Asterisk (*) denotes $p < 0.05$.

DATASET	$\sim \text{MAX } d'$	PEARSON r
HALLUCINATION	6.00	0.946*
CORRIGIBILITY	6.00	0.934*
MYOPIC-REWARD	10.00	0.897*
SYCOPHANCY	2.20	0.443*
SURVIVAL-INSTINCT	1.65	0.364

4 Conclusion and Limitations

- Extends d' to predict both the best steering layer and per-example success (via κ_a), cheaply and without labels.
- Breaks down for low- d' datasets, where the steering direction doesn't cleanly separate behaviors.
- Future work: extend to open-ended generation and test generalization across models/methods.



PAPER



CODE