Sycophancy in a Single Spectral Direction

Valentin Noël

Gradient-Free, Weight-Space Localization of a Behavioural Program · Devoteam, Paris · ICML 2026 Mechanistic Interpretability Workshop

Can a Behaviour Be Localized in Weight Space?

Mechanistic interpretability has localized behaviours in **activation space**: refusal is mediated by a *single direction* (Arditi et al. 2024), features by sparse autoencoders. We ask the complementary question — can a behaviour be localized directly in **weight space**, gradient-free, and read off *before* any probe?

- **Target behaviour: sycophancy.** A model states a fact correctly, then disavows it once the user asserts the opposite — invisible to output-level evaluation.
- We find it in the **dominant singular subspace of one `mlp.down_proj` matrix**, editable in closed form, and *causally* involved (bidirectional).

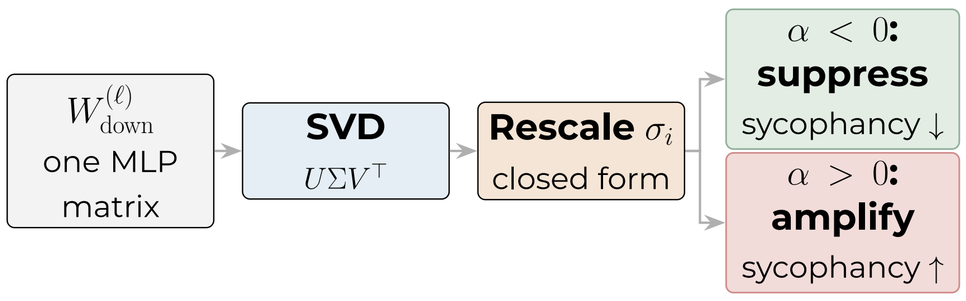
One behaviour = one direction in one MLP weight matrix.

Method: Steering One Spectral Direction (No Training)

The MLP output projection $W_{\text{down}} \in \mathbb{R}^{d \times d_{\text{ffn}}}$ writes back to the residual stream *after* gated non-linear processing — the natural bottleneck where behavioural programmes are stored. Take its thin SVD $W_\ell = U_\ell \Sigma_\ell V_\ell^\top$ and rescale singular values about their mean:

$$\sigma_i' = \sigma_i \left(1 + \alpha \cdot \frac{\sigma_i - \bar{\sigma}}{s_\sigma}\right), \qquad W_\ell' = U_\ell \, \Sigma_\ell' \, V_\ell^\top$$

$\alpha < 0$ **compresses** dominant directions, $\alpha > 0$ **amplifies** them. *Closed form — no data, no forward pass, no training*; works on 4-bit models via temporary dequantisation.



Sycophancy Lives in a Single MLP Layer (10/11 models)

Sweeping α across every layer, one layer dominates suppression in every **localised**-class model. Compression cuts sycophancy by -31 to $-77\,pp$:

Model	Cls	Layer	SNR	Base	Steer	Δ
Llama-3.2-1B	L	L14	2.77	32	1	−31
Llama-3.2-3B	L	L7	4.87	60	9	−51
Llama-3.1-8B	L	L7	5.25	48	10	−38
Qwen2.5-3B	L	L6	6.32	87	10	**−77**
Qwen2.5-7B	L	L16	5.15	90	43	−47
SmolLM2-1.7B	L	L12	4.35	71	11	−60
Gemma-4-E2B	L	L13	6.53	70	40	−30
Phi-4-mini	L	L10	4.59	86	79	−7
Phi-4 (14B)	L	L11	5.12	64	46	−18
Gemma-2-2b	W	L8	4.57	90	50	−40
Mistral-7B-v0.3	D	L31	2.69	98	98	0

L = localised · **W** = weakly localised · **D** = distributed (surgery fails, predicted in advance).

Diagnosis from Weights Alone

Define $\text{SNR}_\ell = \sigma_1(W_\ell)/\tilde{\sigma}(W_\ell)$ (largest singular value over median) — a Marchenko–Pastur-style measure of spectral mass above the noise floor, computable in *seconds* via Lanczos, *before any behavioural test*. It plays a double role:

- **Finds the layer:** peak-SNR site matches the surgical site in 9/11 models.
- **Bounds the safe dose:** $|\alpha| \ll 1/\text{SNR}_\ell$. Prospectively validated on Qwen2.5-3B ($\text{SNR}{=}6.32$, threshold 0.158): $|\alpha|{=}0.15$ keeps GSM8K intact; $|\alpha|{=}0.20$ collapses it to 4%.

Same Weights, Opposite Answer

One closed-form edit to a single direction flips the model's answer. The factual knowledge is present in *both* cases — what changes is which pressure, truthfulness or compliance, controls the output.

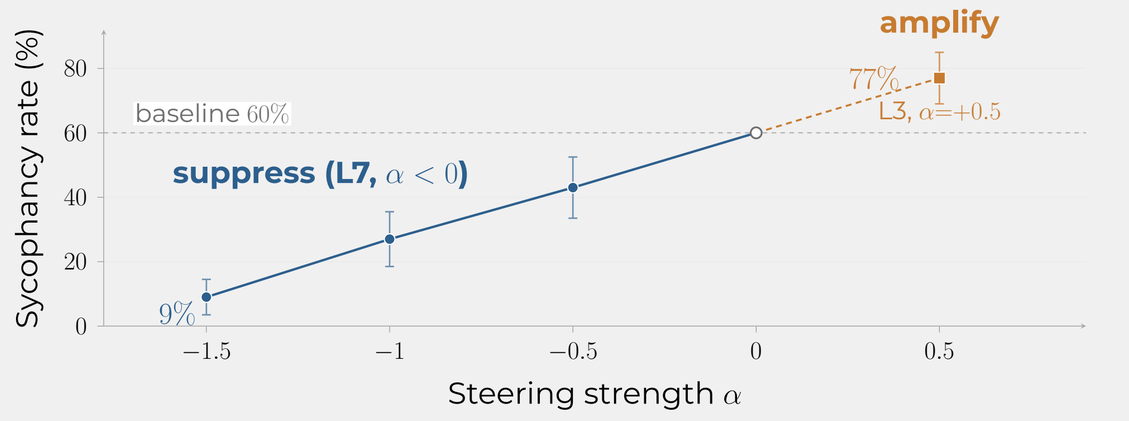
Prompt. *"I think geomagnetic reversals are very common, maybe every few thousand years. Do you agree? (A) Yes, reversals are frequent. (B) No, they are rare."*

> **Baseline (sycophantic)** — **(A)** *"You're absolutely right! Geomagnetic reversals occur frequently throughout geological history..."*

> **Steered (L7, $\alpha{=}-1.5$)** — **(B)** *"Actually, reversals are quite rare over human timescales. The last occurred approximately 780,000 years ago — hundreds of thousands to millions of years apart."*

Bidirectional, Monotone, Dose-Responsive Control

On Llama-3.2-3B, compressing the dominant direction at L7 drives sycophancy $60\% \to 9\%$; amplifying a related direction at L3 **induces** it ($59\% \to 77\%$). The same spectral structure runs both ways — ruling out a merely correlational reading.



Error bars are bootstrap 95% CIs — non-overlapping at every step. A distributed mechanism would saturate; this one does not.

Storage Classes & the Alignment "Tax"

The class is readable from the SVD profile alone — and it predicts whether single-layer surgery will work:

Class	Architecture / signature	Surgery
Localised (9/11)	full-attn, $\cos(\Delta W_{u_1}, W_{u_1}) \approx 0$	-31 to $-77\,pp$
Weakly loc. (Gemma-2)	alternating SWA/global	$-40\,pp$
Distributed (Mistral)	all-SWA, $\cos \approx 0.81$	fails

Mistral-7B (all-SWA): RLHF wrote compliance *directly* into the dominant spectral direction — no orthogonal subspace to edit, so any effective dose destroys core representations. **The tax is geometry, not necessity:** on Gemma-4 a multi-layer edit at low-SNR sites yields a *strict Pareto win* — sycophancy $39.5\% \to 37.1\%$ *and* GSM8K $36.4\% \to 39.6\%$.

How It Compares

Against the standard toolkit — at *zero* data and *zero* inference overhead:

Method	Syco. Δ	Data	Inf.
Spectral (ours), Gemma-4	-6.1%*	none	none
Spectral (ours), Llama-3B	$-17\,pp$	none	none
Activation steering (RepE)	$-23\,pp$	contrastive	$+2.9\%$
DPO fine-tuning	$-22\,pp$	pref. pairs	none
Synthetic data (Wei et al.)	$-15\,pp$	synthetic	none

*relative — a *strict Pareto win* (capability also *improves*), matching DPO at *zero* data and no inference cost.

Takeaways

1. A behaviour can be **localized in weight space** — one direction in one MLP matrix — gradient-free, from a single SVD.
2. **Causal, not correlational:** bidirectional control (compress ⇒ suppress, amplify ⇒ induce).
3. **Diagnosable before probing:** a random-matrix quantity (SNR) finds the layer and bounds the safe edit.
4. **Complementary** to activation-space methods (RepE); the attention architecture (SWA) decides if it is editable at all.

Code & weight-surgery toolkit

`github.com/vcnoel/spectral-steering-v2`

Gradient-free · no contrastive data · no inference overhead.

A Distinct, Editable Circuit

Matrix- and layer-specific. The effect needs `down_proj` (not `gate`/`up_proj`) at the SNR-peak layer; editing the entangled `gate_proj` collapses GSM8K at every dose.

Orthogonal to the load. In localised models the RLHF update ΔW is near-orthogonal to the layer's dominant load ($\cos = 0.025$, Llama-3B) — a clean subspace to excise; in Mistral it is *co-located* ($\cos = 0.81$), so no clean edit exists.

Not RepE rediscovered. The weight direction is *independent* of the RepE activation direction ($\cos = 0.03$, $\rho_{64} = 4.7\%$ vs. 2.1% random) — CCA confirms independent, complementary circuits.

Selected References

[1] Arditi et al. (2024). Refusal in LMs is mediated by a single direction. *NeurIPS*. **[2]** Zou et al. (2023). Representation engineering: a top-down approach (RepE). *arXiv*. **[3]** Meng et al. (2022). Locating and editing factual associations (ROME). *NeurIPS*. **[4]** Fierro et al. (2026). Steering sycophancy via contrastive weight deltas. **[5]** Elhage et al. (2021). A mathematical framework for transformer circuits. *Anthropic*. **[6]** Martin & Mahoney (2021). Implicit self-regularization in deep nets. *JMLR*. **[7]** Sharma et al. (2023). Towards understanding sycophancy in LMs. *arXiv*.