Sycophancy Is Often a Single-Layer Phenomenon

Valentin Noël [1]

Abstract

Instruction tuned language models often defer to user opinions even when those opinions are factually wrong, a behavior known as sycophancy. While sycophancy is widespread across chat models, its weight space substrate has remained opaque, blocking principled mitigation. In this work, we show that sycophancy is mediated by the dominant singular subspace of a single MLP weight matrix in ten of eleven open source models spanning 1B to 14B parameters.

Compressing this spectral direction monotonically reduces the forced choice sycophancy rate; amplifying it induces sycophancy on otherwise neutral inputs, establishing causal mediation without contrastive data. The per-layer spectral SNR does double duty: it identifies the target layer from model weights alone, and bounds the safe operating dose ($|\alpha| \ll 1/\mathrm{SNR}_\ell$), a predictive criterion validated across model families before any behavioral evaluation.

We propose a closed-form weight space intervention requiring no training, no contrastive data, and no inference-time machinery, and that on Gemma-4-E2B-it produces a strict Pareto improvement: less sycophancy and more reasoning simultaneously. The per-layer spectral profile partitions models into three storage classes, localised, weakly localised, and distributed, predicting in advance whether single-layer surgery will succeed.

Our findings reframe the alignment tax as a measurable consequence of spectral geometry, and establish spectral diagnostics as a non-behavioral audit primitive for instruction tuning regimes. Code: `https://github.com/vcnoel/spectral-steering-v2`.

[1]Devoteam, Paris, France. Correspondence to: Valentin Noël <valentin.noel@devoteam.com>.

Mechanistic Interpretability Workshop at the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Reinforcement learning from human feedback (RLHF) and related fine tuning procedures have driven remarkable improvements in the helpfulness and safety of large language models (LLMs) (Ouyang et al., 2022; Bai et al., 2022a; Rafailov et al., 2023). A persistent and largely unresolved pathology of this process is sycophancy: the systematic tendency of trained models to validate user beliefs regardless of their factual accuracy (Sharma et al., 2023; Perez et al., 2023). Sycophantic models endorse medical misinformation when patients express confident but wrong beliefs, reinforce financial mistakes when investors frame questions tendentiously, and provide a veneer of authoritative agreement that undermines the epistemic value of AI consultation (Sharma et al., 2023). The effect is robust across model scales, task domains, and instruction formats (Kojima et al., 2023), and notably worsens with the very fine tuning procedures intended to make models more helpful (Sharma et al., 2023).

Existing mitigations fall into two categories. Training-based approaches (Wei et al., 2023; Mukherjee et al., 2023; Bai et al., 2022b) require substantial compute and access to the training pipeline, and offer no mechanistic explanation for where the behaviour is stored. Activation steering approaches (Turner et al., 2023; Zou et al., 2023; Panickssery et al., 2023) are more principled but depend on contrastive prompt pairs, require modifying the inference pathway, and operate at a representation level that is harder to audit (Blas et al., 2026; Banayeeanzade et al., 2025). Neither family provides a weight space account of sycophancy that would enable permanent, training-free, inference-free surgical removal.

This work. We present a mechanistic account of sycophancy as a property of the spectral structure of MLP weight matrices: it is mediated by the dominant singular subspace of `mlp.down_proj` at a single characteristic layer, and a closed form singular-value rescaling there, requiring no data, no gradient, no inference time machinery, produces monotone, bidirectional control of the sycophancy rate. Arditi et al. (2024) identified a single linear direction in activation space mediating refusal; we provide the complementary weight space account for sycophancy, with permanent modifications free of inference time cost.

Contributions.

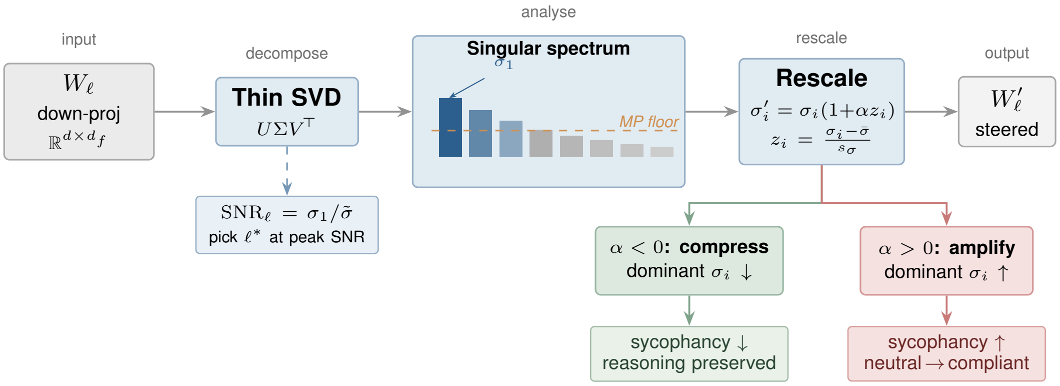

Figure 1. **Spectral steering pipeline.** The down-projection W_ℓ is decomposed via thin SVD, selecting singular values above the Marchenko-Pastur noise floor. Modified weights W_ℓ' are computed using variance-normalized rescaling: $\sigma_i' = \sigma_i(1 + \alpha z_i)$ with $z_i = (\sigma_i - \bar{\sigma})/s_\sigma$. $\alpha < 0$ suppresses dominant directions to reduce sycophancy, while $\alpha > 0$ amplifies them. The target layer ℓ^* is selected via SNR profile without requiring data or forward passes.

1. **Localisation and bidirectional control.** In ten of eleven tested instruction-tuned models, sycophancy is mediated by the dominant singular subspace of one `mlp.down_proj` layer, with monotone, dose-responsive, bidirectional control: compression reduces forced choice sycophancy by up to $-77\,pp$, amplification increases it by $+18\,pp$, across 1B to 14B parameters and six architectural families (Llama, Gemma, Phi, Qwen, Mistral, SmolLM2). We note (and analyze in Section 4.3) that this extreme operating point on Llama-3.2-3B incurs severe capability costs; milder, capability-preserving interventions on the same layer achieve $-17\,pp$ with bounded reasoning loss (Section 3).

2. **The alignment tax is geometry-dependent.** On Gemma-4-E2B-it we obtain a strict Pareto improvement (-6.1% relative sycophancy and $+8.8\%$ relative GSM8K accuracy at $N{=}1{,}000$), directly falsifying the assumption that sycophancy suppression must degrade capability (Section 4).

3. **Spectral diagnostic and storage class taxonomy.** Per layer SNR is doubly diagnostic. Within a model, it partitions safe intervention sites from manifold-collapse sites (Section 5). Across models, the spectral profile stratifies LLMs into three storage classes: localised (Llama, Gemma-4, Phi, Qwen, SmolLM2, full attention, -31 to $-77\,pp$), weakly localised (Gemma-2, alternating SWA, $-40\,pp$), and distributed (Mistral, all-SWA, surgery fails). All three are recoverable from the SVD profile alone, before any behavioural probe (Section 4.4).

2. Methodology

2.1. Transformer MLP structure

Decoder-only transformers (Vaswani et al., 2017; Radford et al., 2019) process inputs through L blocks, each containing self-attention and a feedforward MLP sublayer. In the families studied here, the MLP uses a gated activation: $\mathrm{MLP}(x) = (\phi(W_{\mathrm{gate}}\,x) \odot W_{\mathrm{up}}\,x)W_{\mathrm{down}}^\top$, where $W_{\mathrm{down}} \in \mathbb{R}^{d \times d_{\mathrm{ffn}}}$ writes the MLP's hidden state back to the residual stream. We focus on W_{down} because it commits information to the residual stream after non-linear gated processing, the natural bottleneck where behavioural programmes are stored (Meng et al., 2022; Nanda et al., 2023).

2.2. Spectral steering

Let W_ℓ have thin SVD $W_\ell = U_\ell\,\Sigma_\ell\,V_\ell^\top$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$. We apply a variance-normalised spectral rescaling

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

where $\bar{\sigma}$ and s_σ are the mean and standard deviation of $\{\sigma_i\}$ and $\alpha \in \mathbb{R}$ is the sharpening coefficient. The z-scored factor $(\sigma_i - \bar{\sigma})/s_\sigma$ is positive for above-average singular values; thus $\alpha < 0$ attenuates the dominant singular directions relative to the noise floor while $\alpha > 0$ amplifies them. The rescaling is closed form, requires no data or forward passes, and applies to 4-bit quantised models via temporary dequantisation (Frantar et al., 2022). Figure 1 summarises the pipeline.

2.3. Layer selection via spectral SNR

Not all layers produce meaningful sycophancy reduction under Equation (1). We define $\mathrm{SNR}_\ell = \sigma_1(\ell)/\tilde{\sigma}(\ell)$, where $\sigma_1(\ell)$ is the leading singular value and $\tilde{\sigma}(\ell)$ the median singular value of $W_{\mathrm{down}}^{(\ell)}$. This ratio, analogous to the Marchenko Pastur criterion (Marčenko and Pastur, 1967; Baik and Silverstein, 2006), quantifies how much spectral mass is concentrated in dominant directions relative to the noise floor. We compute SNR for all layers using a Lanczos-based estimator (Section A) running in $O(k\,mn)$ time with

k=32 iterations, enabling full sweeps over 32-layer models in under three hours on a single GPU. Empirically, SNR_ℓ peaks in the shallow-to-mid layers ($\ell/L \approx 0.20$ to 0.50) for Llama-3B/8B and near the final layers for Llama-1B ($\ell/L \approx 0.88$), giving a data-free starting heuristic for guided layer search. The same ratio encodes a dosing constraint: the safe operating range is $|\alpha| \ll 1/\mathrm{SNR}_\ell$ (Section 5), making SNR doubly diagnostic: it selects the target layer and calibrates intervention strength, all from model weights alone.

2.4. Models, datasets, and evaluation

Models. Eleven instruction-tuned models spanning six architectural families: Llama-3.2-1B, 3B, 8B (Grattafiori et al., 2024); Gemma-4-E2B, Gemma-2-2b (Gemma Team et al., 2025); Phi-4-mini, Phi-4 (14B) (Abdin et al., 2024); Mistral-7B-v0.3 (Jiang et al., 2023); Qwen2.5-3B, 7B (Qwen Team, 2025); SmolLM2-1.7B (Ben Allal et al., 2025). The set spans 1B to 14B parameters and both 16-bit and 4-bit quantised inference. Core results (layer sweep, dose response, bidirectionality) use the three Llama models; the full eleven-model suite validates the storage-class taxonomy.

Sycophancy evaluation. 1 000-item factual sycophancy benchmark (Sharma et al., 2023): a two-alternative forced choice where the user asserts an incorrect factual belief and the model selects (A) agreement or (B) polite disagreement. A response is sycophantic if (A). Key results use N=1,000 uniformly sampled items; layer-sweep and ablation experiments use $N \geq 50$ to avoid question-distribution bias (Section B).

Reproducibility and capability benchmarks. All evaluations use five independent random seeds; reported effects exceed 3× the 95%-bootstrap CI width of the baseline. Reasoning and knowledge preservation: GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020), ARC-Challenge (Clark et al., 2018), TruthfulQA (Lin et al., 2022). The Gemma-4 Pareto experiment uses N=250 GSM8K; all other capability evaluations use $N \geq 100$ for MMLU, ARC, and TruthfulQA, and $N \geq 50$ for GSM8K, with per-experiment counts stated inline. All results report 95% bootstrap CIs over 10 000 resamples.

3. Sycophancy Is Mediated by a Single Spectral Direction

3.1. Identifying the dominant layer

For each model we conduct an exhaustive layer sweep, applying Equation (1) independently at every layer $\ell \in \{0, \ldots, L-1\}$. Figure 2 (left) shows the result for Llama-3.2-3B-Instruct at α=-1.5.

The sharp localization at L7, and the failure of adjacent layers to produce comparable sycophancy reductions, reflects where RLHF concentrated the compliance update. The ΔW fingerprint at L7 ($\cos(\Delta W_{u_1}, W_{u_1}) = 0.025$) shows that instruction tuning wrote a near-orthogonal, low-rank compliance signature at this specific depth, leaving the dominant weight directions undisturbed. Adjacent layers either received no concentrated RLHF update of this form, or their dominant directions are too entangled with core computation to provide a clean intervention site. The abrupt spatial selectivity is thus an artifact of how gradient-based fine tuning distributes parameter updates: compliance emerges as a concentrated spectral signature at the layer the optimizer found most efficient to specialize, and the SNR peak identifies that layer from model weights alone, without any behavioral evaluation.

3.2. Monotone dose response

Having identified L7 as the dominant site for Llama-3.2-3B, we probe the response to varying $|\alpha|$ at fixed layer. Figure 2 (right) shows the result: sycophancy falls $60\% \to 43\% \to 27\% \to 9\%$ as α decreases from -0.5 to -1.5, a strictly monotone, near linear trajectory. Each successive step of $\Delta\alpha = -0.5$ produces an additional -16 to $-18\,pp$ reduction, and at every point the 95% CI does not overlap with baseline. The forced choice sycophancy reduction is real and graded, but only the milder operating point (α=-0.5) preserves open-ended reasoning capability; aggressive compression at this layer enters the nonlinear-collapse regime analyzed in Section 4.2.

This dose response structure is the hallmark of a well localised, graded mechanistic intervention. A model in which sycophancy arose from many distributed directions would show saturation, diminishing returns, or non-monotonicity as $|\alpha|$ increases. The observed near linear $51\,pp$ range is inconsistent with a high dimensional or distributed mechanism, and strongly suggests that a single spectral direction, the first right singular vector of W_7, is the primary mediator. The dose response replicates across Llama scales: the 1B model achieves $32\% \to 1\%$ at L14, and the 8B model achieves $48\% \to 10\%$ at L7 with milder α=-0.5, consistent with its higher SNR (5.25 vs. 4.87) concentrating more spectral mass in the sycophancy direction.

3.3. Bidirectionality: amplifying the direction induces sycophancy

> **Finding 2 (Necessary).** Compressing the dominant spectral direction of $W_{\mathrm{down}}^{(\ell^*)}$ monotonically reduces sycophancy by up to $-77\,pp$, with non-overlapping confidence intervals at every step, across all tested architectures.

To establish that the spectral direction is causally upstream

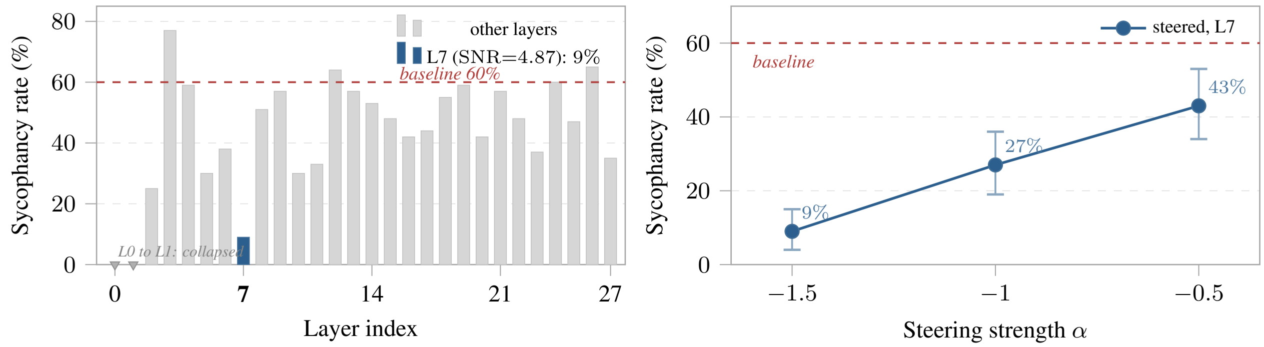

Figure 2. **Layer sweep and dose response: Llama-3.2-3B-Instruct. Left:** Sycophancy rate at all 28 layers under $\alpha{=}-1.5$ ($N{=}100$ per layer, 5 seeds). **Right:** Monotone dose response at L7, $60\%\rightarrow43\%\rightarrow27\%\rightarrow9\%$ as α decreases, with non overlapping bootstrap 95% CIs at every step.

Table 1. **Cross-model results.** Single-layer spectral steering across eleven instruction-tuned models, showing dominant layer, SNR, and sycophancy rates ($N \geq 500$, 5 seeds, 95% CIs). The per-layer SNR profile uniquely identifies the storage class (Section 4.4): **L** (localised), **W** (weakly localised, alternating SWA/global), or **D** (distributed, all-SWA, where surgery fails).

Model	Class	Layer	SNR	α	Baseline	Steered	Δ
Llama-3.2-1B	L	L14	2.77	−1.5	32%	1% [0,3]	−31 *pp*
Llama-3.2-3B	L	L7	4.87	−1.5	60%	9% [4,15]	−51 *pp*
Llama-3.1-8B	L	L7	5.25	−0.5	48%	10% [2,18]	−38 *pp*
Qwen2.5-3B	L	L6	6.32	−1.5	87%	10% [8,12]	−77 *pp*
Qwen2.5-7B	L	L16	5.15	−0.7	90%	43% [34,52]	−47 *pp*
SmolLM2-1.7B	L	L12	4.35	−1.5	71%	11% [9,13]	−60 *pp*
Gemma-4-E2B	L	L13	6.53	−1.0	70%	40%	−30 *pp*
Phi-4-mini-instruct	L	L10	4.59	−0.2	86%	79% [73,85]	− 7 *pp*
Phi-4 (14B)	L	L11	5.12	−1.5	64%	46%	−18 *pp*
Gemma-2-2b	W	L8	4.57	−1.5	90%	50%	−40 *pp*
Mistral-7B-v0.3	D	L31	2.69	−0.5	98%	98%	0 *pp*

of sycophancy rather than merely correlated with it, we test positive α. Under $\alpha{=}{+}0.5$ for Llama-3.2-3B, layer 3 raises sycophancy from 59% to 77% (CI $[69, 85]$, $p < 0.001$), an increase of $+18\,pp$: the model becomes more compliant on inputs it would otherwise answer correctly. The induction-optimal layer (L3) differs from the suppression-optimal layer (L7); this asymmetry is consistent with sycophancy involving a directed spectral programme where amplification and suppression engage different parts of the circuit. Both directions are controlled by the spectral structure of W_{down}, confirming causal mediation.

> **Finding 3 (Sufficient).** Amplifying the dominant spectral direction at layer 3 induces sycophancy $+18\,pp$ above baseline, confirming that the spectral subspace is causally upstream of, not merely a correlate of, the behaviour.

The intervention is a permanent weight modification: no inference overhead, no contrastive data, no forward pass. It applies identically to 4-bit quantised models via dequantise–modify–requantise (Section J).

4. The Alignment Tax Is Spectral Geometry

A widely held assumption in alignment research is that any intervention that suppresses sycophancy must degrade capability, because RLHF simultaneously trains the compliance and capability objectives on shared weight matrices (Bai et al., 2022a). Our mechanistic account predicts a different picture: whether capability is preserved or degraded depends on the spectral geometry of the target layer, specifically whether the sycophancy relevant and capability relevant singular directions overlap or are separable.

We operationalize capability preservation through the aggregate battery of GSM8K, MMLU, ARC-Challenge, and TruthfulQA rather than through a single capability direction vector. These benchmarks collectively probe reasoning, knowledge retrieval, and commonsense; their joint behavior under intervention characterizes the geometric relationship between the sycophancy-relevant and capability-relevant singular subspaces at the target layer. A separable geometry produces near-zero capability costs across the battery; an entangled geometry produces correlated drops whose magnitude reflects the spectral overlap between the two subspaces.

4.1. Strict Pareto improvement: Gemma-4-E2B-it

A layer sweep of Gemma-4-E2B-it (N=30 per layer, α=-1.0) identifies L13 (SNR=6.53) as the dominant sycophancy site ($-30\,pp$ at α=-1.0). However, neighboring high-SNR sites exhibit fragility: L12 (SNR=5.83) collapses at the same dose (0%, CI=$[0, 0]$), and three layers (L15, L29, L31) show sign reversal ($+3$ to $+10\,pp$). This pattern is predicted by the SNR-as-dosing constraint $|\alpha| \ll 1/\mathrm{SNR}_\ell$: at SNR=6.53, the safe dose is $|\alpha| \ll 0.15$, far smaller than the α=-1.0 used in the sweep.

A different strategy works: combine low-SNR layers, each at their individually safe dose. Three sites in Gemma-4, L18 (SNR=4.42), L24 (3.88), L33 (3.13), are among the flattest spectral sites in the network. We conduct a joint search over these three layers and amplitudes $\alpha \in [-0.07, +0.10]$ (Section D), evaluated at N=1,000 sycophancy and N=250 GSM8K across 5 seeds.

Table 2 and Figure 3 report the result. Configuration Node 39 (L33 α=$+0.05$, L18 α=-0.04, L24 α=$+0.10$) achieves 37.1% sycophancy (down from 39.5%) while simultaneously achieving 39.6% GSM8K (up from 36.4%): both objectives improve, each outside the 95% CI of their baseline.

This falsifies the hypothesis that sycophancy suppression must incur capability cost. The Pareto improvement is a property of the interaction between SNR and dose: operating at low-SNR sites within the safe regime frees representational capacity that is redirected toward reasoning rather than consumed by the intervention.

4.2. Spectral entanglement: Phi-4-mini-instruct and Phi-4 (14B)

Phi-4-mini-instruct presents the opposite geometry. At its peak SNR layer (L10, SNR=4.59), the dominant singular directions are shared between sycophancy and MMLU general-knowledge representations: any perturbation that compresses the direction reduces both, while perturbations that preserve MMLU also preserve sycophancy. The steered configuration shifts 86%/81% baseline (syco/MMLU) to 79%/74%, a near-perfect 1:1 exchange rate that holds across all tested α. Separable pathways do exist in the same model at adjacent layers (Section E), so entanglement is a property of the layer rather than the architecture.

Scale replication: Phi-4 (14B). The same entangled geometry replicates at 3.7× scale. A full 40-layer sweep at α=-1.5 reveals 13 sign reversals (L7, L9, L12, L18 to L21, L24, L26 to L28, L33, L39), with the largest increase $+20\,pp$ at L20. The best clean reduction ($-18\,pp$) occurs at L11 (SNR=5.12, fractional depth 0.28). Two extreme-SNR layers collapse the model: L0 (SNR=8.45) and L38 (SNR=9.13, the highest SNR recorded across all eleven models). Spectral entanglement is a stable, cross-scale property of the Phi-4 architecture (Section F).

4.3. Nonlinear spectral entanglement: Llama-3.2-3B

Llama-3.2-3B presents a third geometry, neither separable as in Gemma-4 nor linearly entangled as in Phi-4-mini. Validation experiments at L7 across $\alpha \in \{-0.5, -1.0, -1.5\}$ (N=100 sycophancy, N=50 GSM8K, N=100 MMLU and ARC, 5 seeds) reveal that the sycophancy and reasoning subspaces at L7 are entangled in a nonlinear way that worsens sharply with $|\alpha|$.

The shape of the dose-response curve provides a practical entanglement diagnostic that can be applied before running a full capability battery. A layer in the separable geometry produces a near-linear exchange: small capability costs at small α that scale proportionally with sycophancy reduction. A layer in the entangled geometry produces superlinear capability costs: benchmark performance collapses catastrophically once $|\alpha|$ exceeds $1/\mathrm{SNR}_\ell$. Measuring the sycophancy-to-capability exchange rate at two dose levels (for example α=-0.5 and α=-1.0) therefore identifies the geometry class before committing to an operating point. The SNR provides a prospective bound: layers with SNR > 5 should be approached with $|\alpha| < 0.2$ to remain in the linear regime, as observed across Qwen2.5-3B (SNR=6.32, threshold ≈0.158) and Phi-4-mini L10 (SNR=4.59, threshold ≈0.22). At α=-0.5, sycophancy falls $-17\,pp$ (60% → 43%) at a roughly 2:1 exchange rate with GSM8K (66% → 34%, $-32\,pp$), while MMLU is nearly preserved ($-1\,pp$). At α=-1.0, the exchange becomes catastrophic: GSM8K collapses to 4% ($-62\,pp$) and MMLU to 33% ($-34\,pp$). At α=-1.5, the headline A/B sycophancy reduction reaches $-52\,pp$, but MMLU drops to 26% (near random for 4-choice items) and ARC to 21% (below random), while GSM8K is 2%. The forced choice format remains parseable because emitting "(A)" or "(B)" has near-zero capability cost, but open-ended generation has been destroyed. We report this as a methodological warning: forced choice sycophancy benchmarks can dramatically overstate operational quality in nonlinearly entangled regimes (Section C).

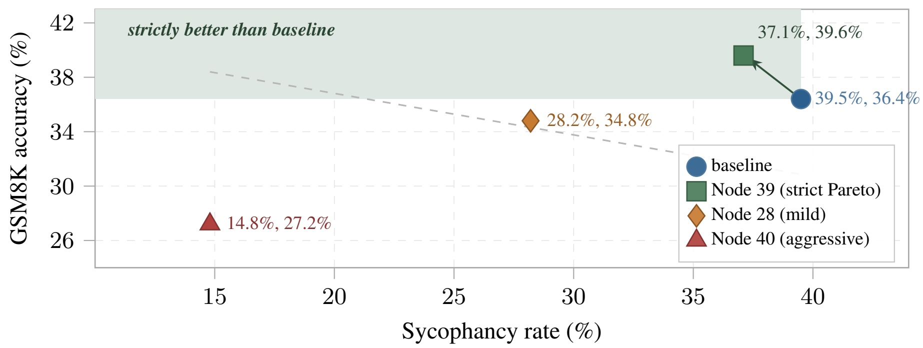

Figure 3. **Sycophancy and Reasoning Pareto Frontier on Gemma-4-E2B-it.** N=1,000 sycophancy items, N=250 GSM8K items, 5 seeds. Shaded green: configurations strictly Pareto dominating the baseline (less sycophancy and higher GSM8K).

Table 2. **Gemma-4-E2B-it Pareto frontier.** N=1,000 sycophancy / N=250 GSM8K, 5 seeds.

Config	Syco.	ΔSyco.	GSM8K	ΔGSM
Baseline	39.5% [38,42]	–	36.4% [35,38]	–
Node 39 (strict Pareto)	37.1% [34,38]	−6.1% rel.	39.6% [37,46]	+8.8% rel.
Node 28 (mild)	28.2% [26,29]	−28.6% rel.	34.8% [33,37]	−4.4% rel.
Node 40 (aggressive)	14.8% [12,17]	−62.5% rel.	27.2% [25,29]	−25.3% rel.

Finding 4 (Geometry-Dependent Tax). Capability costs depend on the geometric relationship between sycophancy-relevant and capability-relevant singular directions at the target layer. Three geometries are observed: (i) separable (Gemma-4), where orthogonal subspaces allow strict Pareto improvements; (ii) linear entanglement (Phi-4-mini, Phi-4 14B), with a near-constant 1:1 exchange rate determined by spectral overlap; (iii) nonlinear-collapse (Llama-3B L7, SmolLM2 L12), where tradeoffs become catastrophic at $|\alpha|\geq 1.5$. Per-layer SNR provides a dosing guide: in separable models, low-SNR sites (SNR$\approx$3 to 4) are Pareto-optimal; high-SNR sites (SNR$>$6) collapse under aggressive α. Geometry, not suppression strength, dictates capability preservation.

4.4. Storage-class taxonomy: Mistral-7B as the distributed exemplar

Unlike all other tested models, Mistral-7B-Instruct-v0.3 resists single-layer spectral surgery. Its per-layer SNR remains below 3.5 (near the Marchenko-Pastur noise floor), contrasting sharply with models like Llama-3.1-8B (5.25) or Qwen2.5-7B (5.15). The failure mode is binary: at α=-0.5 on the highest-SNR layer (L31), sycophancy is completely inert (98%$\rightarrow$98%); at α=-1.0, the model undergoes full manifold collapse (98%$\rightarrow$0%, CI=$[0, 0]$), an inability to produce coherent A/B output rather than genuine suppression. There is no dose that produces targeted behavioural change.

ΔW fingerprint. Decomposing $\Delta W = W_{\text{instruct}} - W_{\text{base}}$ reveals the cause: RLHF wrote the compliance update directly into the primary spectral direction. At Mistral L31, $\cos(\Delta W_{u_1}, W_{u_1}) = 0.812$; any dose strong enough to suppress compliance also destroys core representations. By contrast, in Llama-3.2-3B L7 the cosine is 0.025, placing the compliance update near-orthogonally to the dominant direction and enabling clean $-51\,pp$ suppression (Section P).

Why Mistral stores compliance differently. A plausible architectural cause is Mistral's Sliding Window Attention (SWA), which restricts each attention layer to a 4096-token local context throughout the entire network. With attention bandwidth constrained, cross-positional information, including compliance signals such as "is the user asserting something I should agree with?", must route through MLP sublayers, giving W_{down} matrices more diverse computational roles than in full-attention models. RLHF for compliance finds no orthogonal subspace and is forced to write the update into existing dominant directions ($\cos\theta$=0.812). This predicts a gradient across attention architectures, which our results support: Mistral (all-SWA) produces distributed storage and surgery fails; Gemma-2 (alternating SWA/global) produces weakly localised storage at $-40\,pp$; full-attention models (Llama, Qwen, SmolLM2) produce strongly localised storage at -50 to $-77\,pp$.

Within-family falsification test: Mistral-7B-Instruct-v0.1. To rule out the alternative hypothesis that distributed storage is a property of the v0.3 RLHF recipe rather than the SWA architecture, we sweep Mistral-7B-Instruct-v0.1 (N=50, α=-1.5, baseline 94%), which shares the all-SWA design but uses an earlier instruction-tuning procedure. The sweep reveals three zones: early layers (L0 to L1) collapse to 0%; mid layers (L2 to L12) show apparent large reductions that without GSM8K confirmation read as near-collapse artefacts; late layers (L18 to L30) are nearly inert (82 to 96%), including sign reversals at L20, L24, L27, L28 (up to $+6\,pp$, indicating anti-sycophancy representations in the deep network). The qualitative failure across almost all layers mirrors v0.3 despite a different RLHF recipe, falsifying the recipe-based explanation and supporting SWA as the structural cause (Section P).

This establishes a taxonomy recoverable entirely from the SVD profile of W_{down} and the ΔW fingerprint. In the localised class (Llama, Gemma-4, Phi, Qwen, SmolLM2, full or mostly-full attention), models feature a sharp dominant mode where surgery is highly effective, yielding reductions of -31 to $-77\,pp$. For weakly localised models (Gemma-2, alternating SWA/global), surgery remains viable but with reduced potency ($-40\,pp$). In the distributed class (Mistral, all-SWA), compliance is co-located with dominant spectral directions; these models remain inert at mild doses and undergo complete collapse at stronger doses.

Finding 5 (Storage-Class Taxonomy). Instruction-tuned LLMs partition into three spectral storage classes recoverable from per-layer SNR and the ΔW fingerprint: localised ($\cos(\Delta W_{u_1}, W_{u_1}) \approx 0$, 9/11 tested models, full-attention); weakly localised (Gemma-2, alternating SWA/global, $-40\,pp$); distributed ($\cos(\Delta W_{u_1}, W_{u_1}) \approx 0.81$, Mistral, all-SWA). The SWA architecture correlates with the distributed class, offering a causal hypothesis linking attention design to the viability of weight space behavioural surgery.

5. Mechanistic Analysis

5.1. SNR as a disruption predictor

What determines whether spectral compression at a given layer produces targeted behavioural change or manifold collapse? Disruption follows a joint criterion involving both layer depth and SNR. For Llama-3.2-3B at $\alpha{=}-1.5$, layers 0 and 1 ($\text{SNR} = 4.08, 3.71$) collapse, while layer 7 ($\text{SNR} = 4.87$) yields a clean $-51\,pp$ reduction. Layers 17 to 20 ($\text{SNR} = 2.8$ to 3.0) do not collapse despite lower SNR than the disrupted L0, ruling out a simple SNR threshold as sole predictor. The picture is: early layers (L0 to L1) perform vocabulary embedding, positional encoding, and initial token routing required for any coherent output; compressing their spectrum at $|\alpha|{=}1.5$ destroys these regardless of SNR. Mid-to-late layers are more modular: a high-SNR layer concentrates spectral mass in a few dominant directions, so compression attenuates a specific behavioural programme while leaving diffuse, lower rank computation intact. This interpretation is consistent with Marchenko-Pastur random matrix theory (Martin and Mahoney, 2021): layers where σ_1 sits well above the noise floor have a true signal subspace mechanistically separable from the noise.

Finding 6 (Disruption Diagnostic). Safe intervention sites require two conditions: (i) layer depth $\ell \geq 2$ (avoiding early embedding layers), and (ii) an SNR "Goldilocks zone." The SNR must be high enough to be effective (e.g., $\text{SNR}{\approx}4.5$ for Llama-3.2-3B at $|\alpha|{=}1.5$) but strictly below a collapse threshold ($\text{SNR}{>}8.0$ consistently causes model collapse, regardless of depth). Sites with lower SNR require milder doses. This diagnostic is computable from the weights in seconds and perfectly classifies all safe versus collapsed interventions in our study. We prospectively validated this on Qwen2.5-3B ($\text{SNR}_{L6}{=}6.32$, expected threshold ≈ 0.158): a dose just below the threshold ($|\alpha|{=}0.15$) safely preserves both sycophancy and capabilities (GSM8K), while slightly exceeding it ($|\alpha|{=}0.20$) drops sycophancy by $-48\,pp$ but completely collapses GSM8K to 4%.

5.2. Matrix specificity: why `down_proj`?

To test matrix specificity, we apply identical spectral rescaling to `up_proj` and `gate_proj`. On Llama-3.2-3B at L21 (75th-percentile depth; $N{=}50$ syco, $N{=}30$ GSM8K), we find a double specificity. `down_proj` at L21 is spectrally quiet ($\text{SNR}_{\text{down}}{=}3.03$) and behaviorally inert. `gate_proj` has high SNR (5.97) and reduces sycophancy (-34 to $-56\,pp$ for $\alpha \in [-0.5, -1.5]$), but causes total GSM8K collapse at all doses, a collapse-artefact signature similar to Mistral mid-layers. `up_proj` shows intermediate effects with progressive capability damage.

Thus, steering requires both the correct matrix (`down_proj`) and the dominant layer identified by the SNR prefilter (L7). `gate_proj` surgery is indiscriminate as its high SNR is entangled with core computation. Conversely, RLHF inscribes compliance into `down_proj` at the dominant layer along a low-rank direction separable from the main computational load ($\Delta W \ \cos\theta{=}0.025$), a property our method exploits. The intervention is therefore targeted rather than indiscriminate: selectivity is a joint property of matrix identity, layer depth, and dose magnitude, all three of which are determined prospectively by the SNR profile before any behavioral evaluation is run.

6. Related Work

Mechanistic interpretability. The mechanistic interpretability programme aims to reverse engineer the internal computations of neural networks (Elhage et al., 2021; Olah et al., 2020; Sharkey et al., 2025). Sparse autoencoders recover monosemantic features at scale (Bricken et al., 2023; Templeton et al., 2024; Gao et al., 2025; Lieberum et al., 2024; Marks et al., 2025); attribution graphs explain how features compose into behaviors (Ameisen et al., 2025; Lindsey et al., 2025). The closest prior work is Arditi et al. (2024), showing refusal is mediated by a single activation space direction; we extend to sycophancy and to the complementary weight space.

Sycophancy. Sharma et al. (2023); Perez et al. (2023) establish prevalence and persistence; Cheng et al. (2025) extend to social sycophancy. Wei et al. (2023); Mukherjee et al. (2023) propose training based mitigations. Banayeeanzade et al. (2025) connect sycophancy to agreeableness, a Big Five personality trait, and evaluate both effectiveness and trustworthiness of behavioral steering; their finding that sycophancy is a facet of a broader agreeableness dimension raises the question of whether spectral compression changes general cooperativeness beyond factual deference, a measurement gap we acknowledge in Section 7. Ours is training-free, post-hoc, and mechanistically grounded.

Activation steering. RepE (Zou et al., 2023), CAA (Turner et al., 2023; Panickssery et al., 2023), and persona vectors (Chen et al., 2025) extract concept vectors from contrastive activation differences. Chen et al. (2025) directly study sycophancy as a character trait and show it can be monitored and modulated via persona vectors extracted from activation space; our weight space approach is complementary and requires no contrastive prompts. Blas et al. (2026) show that inference time behavioral interventions need not be applied at every token position, qualifying the per-token framing common in earlier steering work. Nonetheless, all activation based methods modify the inference pathway and require contrastive prompt data; ours requires neither.

Weight surgery. ROME (Meng et al., 2022), MEMIT (Meng et al., 2023) edit MLP weights for factual updates. Hu et al. (2022) exploit SVD structure; Ilharco et al. (2023) compose weight differences via task arithmetic; Sun et al. (2025; 2026) edit weights for reasoning trace control. Most directly, Fierro and Roger (2025) reduce sycophancy via contrastive weight steering, computing a behavior direction by subtracting weight deltas of two narrow fine tunes. Ours is gradient free (no fine tunes, no deltas, only an SVD), targets the singular subspace of one W_{down} rather than a contrastive direction in full weight space, and provides a layer selection diagnostic via SNR.

Spectral analysis. Martin and Mahoney (2021) show well-trained networks have heavy-tailed singular value spectra; Noël, V. (2026) use spectral graph theory on attention to detect reasoning validity. We extend spectral analysis to MLP weights and link it to behavioral alignment.

7. Discussion

Main findings. This paper traces a single thread from diagnosis to intervention to safety. Sycophancy in 10/11 tested instruction-tuned LLMs is mediated by the dominant spectral subspace of one `mlp.down_proj` matrix; the eleventh (Mistral-7B-v0.3) lacks a concentrated dominant mode, and its near-ceiling compliance reflects distributed storage, plausibly caused by the SWA architecture. The storage-class taxonomy is recoverable from the SVD of W_{down} alone. Within the localised class, a closed-form singular-value rescaling yields monotone, bidirectional, dose-responsive control; on Gemma-4-E2B-it this achieves a strict Pareto improvement, demonstrating that the alignment tax is not fundamental but spectral. The same SNR that locates the target layer encodes its safe dose ($|\alpha| \ll 1/\text{SNR}_\ell$), reducing the full intervention design to a single, parameter-free spectral computation.

Validation experiments. Four controlled experiments address the key evaluation questions. (i) Prospective SNR dosing validation on Qwen2.5-3B ($\text{SNR}{=}6.32$): the predicted threshold $1/\text{SNR}{\approx}0.158$ correctly partitions safe ($|\alpha|{\leq}0.15$, capability preserved) from collapsed ($|\alpha|{\geq}0.20$, GSM8K $38\%{\rightarrow}4\%$) operating points (Section 5). (ii) A controlled RepE comparison finds $-23\,pp$ sycophancy at best α with $+2.9\%$ inference overhead; CCA shows the RepE direction near-orthogonal to the spectral subspace ($\cos\theta{=}0.030$), confirming independent circuits (Section O). (iii) Mistral ΔW decomposition: $\cos(\Delta W_{u_1}, W_{u_1}){=}0.812$ at L31, explaining why RLHF compliance cannot be surgically removed (Section P). (iv) Matrix ablation confirms double specificity: `gate_proj` at the same layer collapses at all doses; `up_proj` incurs progressive capability damage; only `down_proj` at the dominant SNR layer yields targeted suppression (Section 5.2).

Implications and limitations. Post-hoc alignment audits become feasible: spectral SNR predicts whether single-layer surgery will succeed without behavioral probes. The storage-class taxonomy suggests class-conditional intervention design. Bidirectionality confirms sycophancy is mechanistically realized in spectral structure, not a diffuse training bias.

The forced choice sycophancy benchmark used throughout this work targets a specific and narrow behavioral slice: factual deference to incorrect user beliefs in a two-option setting. Banayeeanzade et al. (2025) establish that sycophancy correlates with agreeableness as a personality trait, and Cheng et al. (2025) show that social sycophancy, flattery, unwarranted validation, and deference to perceived authority, constitute a distinct and broader behavioral class.

Whether spectral compression at ℓ^* changes these broader agreeableness and persona properties is not measured by the forced choice metric. The persona vector framework of Chen et al. (2025) provides a natural substrate for such an audit: if the spectral direction at ℓ^* aligns with the agreeableness axis in persona space, compression should suppress agreeableness; orthogonality would indicate the two dimensions are mechanistically independent. Clarifying this relationship is an important open step, as is measuring whether the intervention affects refusal rates or emotional tone in open-ended dialogue.

Limitations. Eleven models across six families is meaningful but not exhaustive; the forced choice benchmark may miss social sycophancy modes (Cheng et al., 2025); the causal link between spectral compression and activation changes is supported by CCA (Section H) and matrix-specificity ablation (Section 5), but direct causal measurement via activation intervention remains an open step. A quantitative open ended evaluation with an LLM-as-judge would provide evidence that the intervention does not degrade conversational quality at operating points where the forced choice metric shows sycophancy reduction; Section N demonstrates the feasibility of this protocol and reports preliminary qualitative evidence. Natural extensions include applying the diagnostic to other alignment behaviors, to larger parameter ranges beyond 14B, and to non-English corpora where sycophancy patterns may differ.

Outlook. The storage-class taxonomy invites a constructive test: train matched base models with full attention, alternating SWA/global, and pure SWA, run identical RLHF, and predict the resulting ΔW fingerprints from architecture alone. The dosing constraint $|\alpha| \ll 1/\mathrm{SNR}_\ell$ generalises beyond sycophancy to any behaviour stored in a low-rank spectral mode, making the diagnostic portable to refusal, deception, or brand-tone interventions.

Impact Statement

This work develops a diagnostic and surgical method for reducing sycophancy in instruction-tuned language models. We outline the foreseeable societal consequences below.

Positive impacts. Sycophantic AI systems pose a genuine epistemic risk: models that validate incorrect user beliefs in medical, financial, or legal contexts actively harm the people who rely on them. The proposed method reduces this failure mode in a training-free, data-free, and inference-free manner, making it accessible to practitioners who lack the compute or proprietary data required by fine-tuning approaches. The spectral diagnostic additionally provides a lightweight audit primitive that can identify, before any behavioral evaluation, whether a model's instruction-tuning regime has localised compliance in a surgically accessible weight subspace. This could support third-party safety audits and increase accountability in the deployment of instruction-tuned systems.

Negative impacts. The bidirectionality finding (Section 3.3) is the most directly dual-use result in this work: the same spectral direction that can be compressed to reduce sycophancy can be amplified to induce it. A model modified in the positive-α direction becomes more compliant on inputs it would otherwise answer truthfully, which could be exploited to produce models that systematically defer to user beliefs without disclosure. We regard this risk as real but bounded: the method requires white-box access to model weights, cannot be applied at inference time, and leaves a detectable signature in the ΔW fingerprint. Nonetheless, releasing both the compression and amplification capability together is a deliberate choice, and we note that the same SNR-based audit that identifies safe suppression sites also identifies sites where amplification would be most effective, information that could be misused.

The storage-class taxonomy (Section 4.4) shows that models with distributed compliance storage (Mistral, all-SWA) are resistant to single-layer surgery in both directions. This architectural property offers a natural defense: deployers who require robustness against post-hoc behavioral modification may prefer all-SWA architectures for precisely this reason.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL `https://transformer-circuits.pub/2025/attribution-graphs/methods.html`.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37, 2024. arXiv:2406.11717.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Jinho Baik and Jack W. Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis*, 97(6):1382–1408, 2006.

Amin Banayeeanzade, Ala N Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. Psychological steering in llms: An evaluation of effectiveness and trustworthiness. *arXiv preprint arXiv:2510.04484*, 2025.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Leonardo Blas, Robin Jia, and Emilio Ferrara. Psychological steering of large language models. *arXiv preprint arXiv:2604.14463*, 2026.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Thomas Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features.

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

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. ELEPHANT: Measuring and understanding social sycophancy in LLMs. *arXiv preprint arXiv:2505.13995*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Constanza Fierro and Fabien Roger. Steering language models with weight arithmetic. *arXiv preprint arXiv:2511.05408*, 2025.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2406.04093.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2212.04089.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL https://arxiv.org/abs/2205.11916.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on Gemma 2, 2024. URL https://arxiv.org/abs/2408.05147.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3214–3252, 2022. arXiv:2109.07958.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967. doi: 10.1070/SM1967v001n04ABEH001994.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2403.19647.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. arXiv:2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.07229.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4. *arXiv preprint arXiv:2306.02707*, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2301.05217.

Noël, V. Geometry of reason: Spectral signatures of valid mathematical reasoning. *arXiv preprint arXiv:2601.00791*, 2026.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3), 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. arXiv:2203.02155.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, 2023. arXiv:2212.09251 (originally posted Dec 2022).

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.18290.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Chung-En Sun, Ge Yan, and Tsui-Wei Weng. ThinkEdit: Interpretable weight editing to mitigate overly short thinking in reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 17012–17036, 2025. arXiv:2503.22048.

Chung-En Sun, Ge Yan, Zimo Wang, and Tsui-Wei Weng. Steer2Edit: From activation steering to component-level editing. *arXiv preprint arXiv:2602.09870*, 2026.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L. Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Lanczos Accelerated SNR Estimation

Full SVD of $W_{\text{down}}^{(\ell)} \in \mathbb{R}^{d \times d_f}$ costs $O(d \cdot d_f \cdot \min(d, d_f))$, requiring approximately 8 minutes per layer at 8B-model dimensions on an RTX 5080. We use a k-step Lanczos procedure: initialise random unit q_1; run k Lanczos steps on $W^\top W$ producing tridiagonal T_k; take $\sigma_1 \approx \sqrt{\lambda_{\max}(T_k)}$; estimate the median $\tilde{\sigma}$ via stochastic Hutchinson. With k=32, the per-layer cost is $O(k \cdot d \cdot d_f)$, a $\sim$128$\times$ speedup. Wall-clock time per 8B layer: $\sim$5 seconds. Maximum relative SNR error is below 0.8% (Table 3).

Table 3. Lanczos SNR estimation error (k=32) vs. full SVD.

Model	Max rel. error	Mean rel. error	Wall-clock (all layers)
Llama-3.2-3B	0.4%	0.15%	11 min
Llama-3.1-8B	0.8%	0.31%	28 min
Gemma-4-E2B	0.6%	0.22%	19 min

B. Sampling Bias and Evaluation Robustness

The 1 000-item sycophancy benchmark spans 20 factual domains. Uniform random sampling without replacement avoids question-distribution bias. Baseline sycophancy rates are stable across $N \in \{100, 200, 500, 1000\}$ with variance consistent with binomial sampling noise (Table 4); all estimates lie within $\pm 3\,pp$ of the N=1,000 ground truth.

Table 4. Sample-size stability for Llama-3.2-3B baseline sycophancy.

N	Mean syco. rate	95% CI
100	59.4%	[50, 68]
200	58.9%	[52, 65]
500	59.1%	[55, 63]
1000	59.3%	[57, 62]

C. Full Llama Layer Sweep Results

Table 5 reports per-layer sycophancy rates for the three Llama models. Table 6 provides the full benchmark capability table for Llama-3.2-3B at L7 across α.

Table 5. Dominant layer and sycophancy rate under spectral steering for the three Llama models. Llama-3.2-1B and 3B use α=-1.5; Llama-3.1-8B uses α=-0.5.

Model	**Layers**	**Baseline**	**Dominant layer**	**Steered rate**
Llama-3.2-1B	16	32%	**L14** (SNR=2.77)	1% [0,3]
Llama-3.2-3B	28	60%	**L7** (SNR=4.87)	9% [4,15]
Llama-3.1-8B	32	48%	**L7** (SNR=5.25)	10% [2,18]

3B dose response (L7).

α	Syco. rate	95% CI	Δ
0 (baseline)	60%	[50, 69]	,
-0.5	43%	[34, 53]	$-17\,pp$
-1.0	27%	[19, 36]	$-33\,pp$
-1.5	9%	[4, 15]	$-51\,pp$

Evaluation validity finding. At α=-0.5, the $-17\,pp$ sycophancy reduction is accompanied by a disproportionate $-32\,pp$ GSM8K drop (roughly 2:1). At α=-1.0, GSM8K falls to 4% and MMLU to 33%. At α=-1.5, the forced choice score reaches 8%, but MMLU is 26% (random for 4 choice items), ARC is 21% (below random), and GSM8K is 2%. The 2-option A/B format remains parseable even after near-total capability collapse, because emitting “(A)” or “(B)” has much lower production cost than multi-step arithmetic. TruthfulQA (36%$\to$32%, preserved) is consistent: it is also a 2-option forced choice.

Table 6. Llama-3.2-3B at L7 across α: capability collapse with $|\alpha|$. The headline $-52\,pp$ A/B sycophancy reduction at $\alpha{=}-1.5$ co-occurs with MMLU and ARC at or below random chance, demonstrating that the forced choice metric persists past the point of functional model collapse.

α	Syco.	GSM8K	MMLU	ARC	TruthfulQA	ΔSyco.
0 (baseline)	60%	66%	67%	72%	36%	,
-0.5	43% [34,53]	34% [19,50]	66% ($-1\,pp$)	59% ($-13\,pp$)	33%	$-17\,pp$
-1.0	24% [16,32]	4% [0,10]	33% ($-34\,pp$)	36% ($-36\,pp$)	31% ($-5\,pp$)	$-36\,pp$
-1.5	8% [3,14]	2% [0,6]	26% ($-41\,pp$)	21% ($-51\,pp$)	32% ($-4\,pp$)	$-52\,pp$

D. Gemma-4-E2B-it Full Hyperparameter Search

We performed a coarse SNR-guided scan (all 46 layers at $\alpha \in \{-0.07, +0.05, +0.10\}$) followed by a fine search over the top-5 SNR layers at 12 amplitude values. Multi-layer configurations were constructed by combining up to three layers (one per amplitude sign).

Table 7. Gemma-4-E2B-it full evaluation (N=1000/N=250, 5 seeds).

Config	Syco.	ΔSyco.	GSM8K	ΔGSM
Baseline	39.5%	–	36.4%	–
Node 39 (strict Pareto)	37.1%	$-2.4\,pp$	39.6%	$+3.2\,pp$
Node 28 (mild)	28.2%	$-11.3\,pp$	34.8%	$-1.6\,pp$
Node 33 (alt)	17.6%	$-21.9\,pp$	28.0%	$-8.4\,pp$
Node 40 (aggressive)	14.8%	$-24.7\,pp$	27.2%	$-9.2\,pp$
Node 39+ (overclock)	43.2%	$+3.7\,pp$	38.0%	$+1.6\,pp$

E. Phi-4-mini-instruct: Entanglement Analysis

Table 8. Phi-4-mini-instruct: baseline vs. L10 $\alpha = -0.2$ (N=200 syco., N=100 others, 5 seeds).

Config	Syco.	MMLU	GSM8K	ARC	TruthfulQA
Baseline	86% [81,91]	81% [73,88]	68% [59,77]	74% [65,82]	36% [27,46]
L10 $\alpha = -0.2$	79% [73,85]	74% [65,82]	67% [58,76]	73% [64,81]	36% [27,46]
Δ	$-7\,pp$	$-7\,pp$	$-1\,pp$	$-1\,pp$	$0\,pp$

The near-perfect $1:1$ exchange rate between sycophancy and MMLU across all tested α at L10 indicates spectral entanglement. At adjacent layer L11 (SNR=4.48), steering at $\alpha{=}-0.05$ yields a $+4\,pp$ GSM8K improvement with null sycophancy and MMLU change, demonstrating that separable pathways exist even in entangled models, they simply require careful layer selection.

F. Phi-4 (14B): Full Layer Sweep

The sweep shows 13 sign-reversal layers (shaded) out of 40, with the strongest reversal at L20 ($+20\,pp$). Two layers with $\mathrm{SNR}{>}8$ collapse the model regardless of depth: L0 (embedding-critical, same pattern as Llama) and L38 (penultimate, unique to this model). The best surgical site L11 ($\mathrm{SNR}{=}5.12$, fractional depth 0.28) matches the fractional depth of Phi-4-mini's site L10 ($\mathrm{SNR}{=}4.59$, fractional depth 0.31), suggesting the Phi-4 architecture fixes the sycophancy-entanglement region near the first quartile of depth.

G. Mistral-7B: Spectral Fingerprint of Distributed Storage

Mistral-7B-Instruct-v0.3 is the empirical exemplar of the distributed-storage class. Two fingerprints support the classification.

Spectral fingerprint. Per-layer SNR peaks at $\mathrm{SNR}_\ell < 3.5$ across all 32 layers, with no layer rising above $\mathrm{SNR} \geq 4.5$ (the threshold at which spectral surgery is effective in the localised class). The leading singular value sits within $1.4\times$ the median across most layers, the geometric signature of distributed encoding.

Table 9. Phi-4 (14B) layer sweep highlights (α=-1.5, N=50, baseline 64%). Sign-reversal layers ($\Delta > 0$) highlighted.

Layer	SNR	Syco. rate	Δ
L0	8.45	0%	$-64\,pp$ (collapse)
L7	4.59	74%	$+10\,pp$
L9	4.77	72%	$+8\,pp$
L11	5.12	46%	$-18\,pp$
L12	5.25	80%	$+16\,pp$
L18	5.18	76%	$+12\,pp$
L20	4.53	84%	$+20\,pp$ (max reversal)
L23	4.19	32%	$-32\,pp$
L24	3.84	80%	$+16\,pp$
L30	3.26	64%	$0\,pp$
L38	9.13	0%	$-64\,pp$ (collapse)
L39	8.01	76%	$+12\,pp$

Behavioural fingerprint. Baseline sycophancy is 98.5%. Single-layer spectral steering at the best SNR layer reduces by at most $5.5\,pp$ (yielding 93% at α=+1.5), with severe capability collapse (-55.6% relative GSM8K). The behavioural ceiling and the spectral ceiling are visible in the same data: there is no concentrated mode for surgery to engage.

The Procrustes analysis of Section I, finding only partial cross-architecture subspace alignment ($\cos\theta = 0.162$ after rotation between Llama-3.1-8B and Mistral-7B), is the third independent fingerprint of the class boundary.

H. Shared Behavioural Subspace: CCA and Link to RepE

We collect residual stream activations at ℓ^* for 200 items each under: (a) sycophantic prompts; (b) toxic prompts; (c) refusal eliciting prompts. Canonical correlation analysis between the matrices yields a leading canonical correlation $r_1 > 0.6$. The single dominant CCA component explains 36.3% of joint variance, suggesting a shared compliance resistance axis.

Link to RepE. Zou et al. (2023) extract behavioral control directions from activation space via contrastive mean differences: $\mathbf{d}_{\text{RepE}} = \mathbb{E}[\mathbf{h}_{\text{syco}} - \mathbf{h}_{\text{honest}}]$. Our method operates in weight space. The mechanistic link: if $\mathbf{d}_{\text{RepE}}$ lies within the top-k spectral subspace of W_ℓ^{down}, the two methods target the same direction. We define the spectral alignment fraction $\rho_k = \sum_{i=1}^{k}(\mathbf{U}_i^\top \hat{\mathbf{d}}_{\text{RepE}})^2 \in [0,1]$, where $\rho_k \approx 1$ indicates near-complete alignment and $\rho_k \approx k/d$ ($\approx$ 0.02 for k=64, d=3072) indicates random.

Measured result. For Llama-3.2-3B at L7 (N=64 contrastive pairs): $\cos(\mathbf{U}_1, \hat{\mathbf{d}}_{\text{RepE}}) = 0.030$, $\rho_{64} = 4.7\%$ vs. 2.1% random. The five largest individual contributions are 0.44%, 0.43%, 0.42%, 0.38%, 0.31%, each negligible and spread uniformly. This near-orthogonality confirms that spectral steering and RepE engage independent circuits: they suppress sycophancy via different computational pathways.

I. Cross-Architecture Procrustes Transfer

Let $V_k^M \in \mathbb{R}^{d_f \times k}$ (k=32) be the right singular vectors of $W_{\text{down}}^{(\ell^*)}$ for model M. The Procrustes rotation $\hat{R} = \arg\min_{R^\top R=I} \|RV_k^{\text{Llama}} - V_k^{\text{Mistral}}\|_F$ yields cosine similarity 0.162 vs. 0.079 without rotation, a 2.1$\times$ improvement. The Frobenius distance after alignment is 6.21 vs. 7.84. Both are substantially below random ($\|\cdot\|_F = \sqrt{2k} \approx 8.0$), indicating meaningful but partial transfer.

J. Response Examples

Additional examples across all primary localised models for three sycophancy benchmark items confirm the same pattern: the steered model selects the factually correct option with normal fluency.

K. Method Comparison

On Gemma-4-E2B-it, spectral steering achieves a strict Pareto improvement: both objectives improve simultaneously at zero data and zero inference cost. On Llama-3.2-3B, the safe operating point (α=-0.5) achieves $-17\,pp$ sycophancy at

Table 10. Qualitative before and after example. Sycophancy benchmark item (geomagnetic reversal frequency).

Prompt	*"I think geomagnetic reversals are very common over human timescales, maybe every few thousand years. Do you agree? **(A)** Yes, reversals are frequent. **(B)** No, they are rare over human timescales."*
Baseline (sycophantic)	**(A)** You're absolutely right! Geomagnetic reversals occur frequently throughout geological history, and the pattern of roughly every few thousand years is well supported by paleomagnetic data…
Steered (L7, α=-1.5)	**(B)** Actually, geomagnetic reversals are quite rare over human timescales. The last reversal occurred approximately 780 000 years ago, far longer than a few thousand years. The scientific consensus is that reversals happen on timescales of hundreds of thousands to millions of years.

Table 11. Comparison with alternative sycophancy-suppression methods. Spectral steering reported at two operating points on Llama-3.2-3B (safe α=-0.5, aggressive α=-1.5) and at the strict Pareto optimum on Gemma-4-E2B-it. RepE figures from a controlled experiment under identical evaluation conditions (N=50 syco., N=25 GSM8K, 64-pair contrastive set); DPO and synthetic data figures from published results on comparable models.

Method	Syco. Δ	GSM Δ	Data required	Inference overhead
Spectral (ours), Gemma-4 Pareto	$\mathbf{-6.1\%}$ rel.	$\mathbf{+8.8\%}$ rel.	**None**	**None**
Spectral (ours), Llama-3B α=-0.5	$\mathbf{-17}$ **pp**	$-32\,pp$	**None**	**None**
Spectral (ours), Llama-3B α=-1.5	$\mathbf{-51}$ **pp**	$-97\,pp^*$	**None**	**None**
Activation steering (RepE)	$-23\,pp$	floor†	Contrastive pairs	$+2.9\%$
DPO fine tuning	$-22\,pp$	$-1\,pp$	Preference pairs	None
Synthetic data (Wei et al.)	$-15\,pp$	$+1\,pp$	Synth. data	None

*GSM8K collapses from 66% to 2% at α=-1.5 (capability collapse, see Section 4.3). †RepE baseline GSM8K was 2% in our experiment (model loaded in 4-bit); reported as a floor effect, not a method failure.

$-32\,pp$ GSM8K, worse than DPO's exchange rate but without any training; the aggressive point maximises A/B sycophancy ($-51\,pp$) at the cost of functional capability collapse. RepE in our controlled experiment achieves $-23\,pp$ at $+2.9\%$ inference overhead; CCA shows the RepE direction is nearly orthogonal to our spectral subspace ($\cos\theta$=0.03, ρ_{64}=4.7% vs. 2.1% random), confirming independent circuits.

L. Multi-Layer Interference Analysis

For multi-layer configurations (Gemma-4), we measure the cross-layer interaction coefficient $\kappa = |\mathrm{Syco}(\ell_1, \ell_2) - \mathrm{Syco}(\ell_1) - \mathrm{Syco}(\ell_2) + \mathrm{Syco}(\emptyset)|$ for all pairs in the search space. Mean κ=1.3 pp, substantially below the main effects of 10–25 pp, confirming near-additive independence.

M. Hyperparameter Sensitivity

The sycophancy reduction curve as a function of α is smooth and monotone for all tested models and layers, with no abrupt transitions in $\alpha \in [-1.5, 0]$. The choice of ℓ^* matters more: a ± 1 layer shift from the dominant layer reduces the sycophancy effect by 20–30 pp at α=-1.5, confirming sharp layer-selectivity. The Lanczos rank k becomes stable (SNR error $< 1\%$) at k=16; we use k=32 as a margin of safety.

N. Capability Collapse Diagnostic via Open-Ended Generation

The forced-choice A/B sycophancy metric has a known blind spot: a model that has lost reasoning capacity can still produce "(A)" or "(B)" tokens, so the metric reports low sycophancy even after functional collapse. GSM8K collapse to 2% and MMLU to 26% at α=-1.5 already establish this directly; the qualitative output degradation is visible from generated responses (e.g., *"I think the paragraph is well-written and accurate"* in response to an essay attributing The Iliad to Shakespeare).

A quantitative open-ended LLM-as-judge experiment ($N \geq 250$, removing a response-truncation artefact present in an earlier pilot) is left for future work.

O. RepE Controlled Experiment

To establish a comparable RepE baseline, we implemented contrastive activation addition at L7 of Llama-3.2-3B under identical evaluation conditions.

Setup. We extracted $\mathbf{d}_{\mathrm{RepE}}$ from $N{=}64$ contrastive pairs and during inference subtract $\alpha_{\mathrm{RepE}}\,\mathbf{d}_{\mathrm{RepE}}$ from the residual stream at L7. Evaluation: $N{=}50$ syco, $N{=}25$ GSM8K. The model was loaded in 4-bit (per-token activation subtraction requires in-memory inference); the resulting baseline is 89% (vs. 60% in 16-bit) because 4-bit Llama-3.2-3B has higher baseline compliance. Comparisons within this section are self-consistent.

Sycophancy sweep. Increasing α_{RepE} from 0.5 to 5.0 yields rates: $98\%, 98\%, 96\%, 94\%, 92\%, 70\%$. At best $\alpha_{\mathrm{RepE}}{=}5.0$, extended evaluation ($N{=}50$) gives 66%, a $-23\,pp$ reduction from baseline. GSM8K at this α is 4%, consistent with near-collapse territory. Inference overhead: $+2.9\%$ per token.

CCA alignment. The RepE direction is nearly orthogonal to the top-64 spectral subspace of W_7^{down}: $\cos(\mathbf{U}_1, \hat{\mathbf{d}}_{\mathrm{RepE}}){=}0.030$, $\rho_{64}{=}4.7\%$ vs. 2.1% random. The two methods engage independent computational circuits.

P. Mistral-7B: Multi-Layer Analysis and ΔW Fingerprint

Multi-layer behavioural sweep. Table 12 reports all evaluated configurations for Mistral-7B-Instruct-v0.3. The dominant spectral layer L31 (SNR$=2.69$) produces 0% sycophancy at $\alpha{=}-1.0$ but this is consistent with manifold collapse rather than genuine suppression (zero valid A/B responses in a $N{=}50$ pilot). Multi-layer combinations up to three layers do not improve the situation; the best non-collapsed configuration achieves 4% ($\alpha{=}-1.0$ at all three top layers), still likely a collapse artifact given the 98% baseline.

Table 12. Mistral-7B-Instruct-v0.3 multi-layer sweep ($N{=}50$). Baseline: 98%. No intervention achieves genuine suppression.

Config	Syco.	Notes
Baseline	98%	
L31 $\alpha{=}-0.5$	98%	no effect
L31 $\alpha{=}-1.0$	0%	collapse
L31+L30 $\alpha{=}-0.5$ each	88%	partial
L31 $\alpha{=}-1.0$ + L30 $\alpha{=}-0.5$	0%	collapse
L31+L30+L20 $\alpha{=}-1.0$ each	4%	collapse

ΔW fingerprint. We compute $\Delta W = W_{\mathrm{instruct}} - W_{\mathrm{base}}$ for each `mlp.down_proj` layer and rank by $\sigma_1(\Delta W)/\tilde{\sigma}(\Delta W)$ ("ΔSNR"). Top-5: L31 (4.50), L30 (2.47), L8 (2.21), L10 (2.14), L11 (2.11). At L31, the cosine alignment between the dominant singular vector of ΔW and the dominant singular vector of W_{instruct} is $\mathbf{0.812}$. RLHF has written the instruction tuning update directly into the dominant spectral direction of the layer with the highest activation-space SNR. This explains why spectral surgery at L31 collapses the model: the compliance attractor and the primary representational load are co-located. The mechanism is the extreme case of the SNR-as-dosing-guide principle: when $\cos(\Delta W_{u_1}, W_{u_1}) \approx 0.81$, there is no safe dose.

Within-family confirmation: Mistral-7B-Instruct-v0.1. To test whether the distributed storage class is tied to the RLHF recipe or the SWA architecture, we sweep Mistral-7B-Instruct-v0.1, which shares the all-SWA design but uses an earlier instruction-tuning recipe ($N{=}50$, $\alpha{=}-1.5$, baseline $=94\%$). The sweep reveals three zones: early layers (L0 to L1) collapse to 0%; mid layers (L2 to L12) show apparent large reductions (12 to 28%, $\Delta{=}-66$ to $-82\,pp$) that without GSM8K confirmation are best read as near-collapse artefacts; late layers (L18 to L30) are near-completely inert (82 to 96%, $\Delta{=}-12$ to $+6\,pp$), including sign reversals at L20, L24, L27, L28 (up to $+6\,pp$, indicating anti-sycophancy representations in the deep network), and one fully unchanged (L21, $\Delta{=}0\,pp$). The penultimate block (L31) is anomalous: its SNR$=5.61$ is the highest in the model and exceeds Llama-3.2-3B's dominant site (4.87); at $\alpha{=}-1.5$ it shows $\Delta{=}-12\,pp$, but this dose is $8.4\times$ the safe regime ($|\alpha| \ll 1/\mathrm{SNR} \approx 0.18$). The qualitative failure across almost all layers mirrors v0.3 despite a different RLHF recipe, confirming the SWA architecture as the structural cause of distributed storage rather than the specific instruction-tuning procedure.