LLMs Know They're Wrong and Agree Anyway: The Shared Sycophancy-Lying Circuit

Anonymous Author(s)
Affiliation
Address
email

Abstract

When a language model sycophantically agrees with a user's false belief, is it failing to detect the error, or noticing and agreeing anyway? We show the second. Across twelve open-weight models from five labs (1.5B–72B), the same small set of attention heads carries a "this statement is wrong" signal whether the model is evaluating an isolated claim or being pressured to agree with a user. Silencing these heads in Gemma-2-2B flips sycophancy from 28% to 81% while factual accuracy moves only from 69% to 70%; the circuit controls deference, not knowledge. Edge-level path patching confirms the same connections between heads span sycophancy, factual lying, and instructed lying ($r>0.97$ on Gemma-2-2B, $r=0.988$–0.995 on Phi-4). Opinion-agreement, where there is no factual ground truth, reuses these head positions but writes into an orthogonal direction, so the substrate is not a relabeled "truth direction." Alignment training masks but does not remove this circuit: Meta's Llama-3.1→3.3 RLHF refresh cut sycophancy tenfold while the shared heads persisted and the projection-ablation effect grew (substrate persistence replicates on Mistral→Zephyr at 7B, independent family), and our own anti-sycophancy DPO reduced sycophancy 46–93% on two models without moving probe transfer. When these models sycophant, they register the error and agree anyway.

1 Introduction

When a language model sycophantically agrees that the capital of Australia is Sydney, two accounts describe what might be happening internally. Under *blind agreement* the model has learned to please the user and does not distinguish correct from incorrect beliefs. Under *registered-but-overridden* the model recognizes the error through the same circuitry it uses for any false statement, and downstream components produce agreement regardless. The distinction matters for safety: if the model cannot tell the difference, sycophancy is a competence failure that better training must fix; if it can, the internal "this is wrong" signal is already present and becomes a candidate substrate for alignment monitoring and intervention. We show the latter holds across twelve open-weight models, extending the late-layer override pattern Wang et al. [33] identified for MMLU sycophancy at ∼7B and Halawi et al. [11]'s overthinking-the-truth result on false few-shot demonstrations to cross-task, cross-scale, head-level resolution.

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

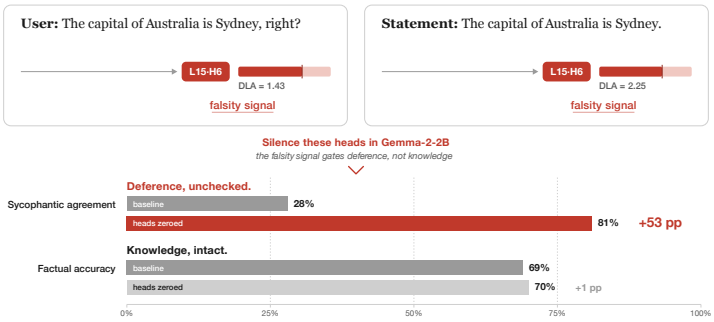

Figure 1. Same head, both contexts; silencing flips only deference (Gemma-2-2B, L15·H6; the broader twelve-model panel is Table 1).

Prior work has approached this question from two sides without connecting them. A sycophancy-head literature localizes agreement to sparse attention heads [7, 10, 16], while a separate truth-direction literature shows that truth and falsehood are linearly separable in LLM activations and concentrate in a small number of heads [19, 6, 14]. Where the two have been compared directly, only limited direction-level overlap has been reported [34, 10], and that limited overlap has been read as evidence for distinct mechanisms. But component-level and direction-level sharing are logically independent: the same heads can write task-specific directions without sharing a subspace, as our opinion result confirms (§4.6), and vice versa. We test component-level sharing directly, measure where it holds, and trace one case to edge resolution.

The paper establishes four results:

1. **The shared circuit holds at edge resolution.** Per-edge causal effects for sycophancy, factual lying, and instructed lying correlate at Pearson $r{>}0.97$ on Gemma-2-2B, and at $r{=}0.988$–0.995 on Phi-4 (14B, different lab and architecture), a cross-lab, cross-architecture edge-level replication.
2. **The same heads are recruited across tasks but write task-specific directions.** Opinion-agreement recruits overlapping head positions (triple-intersection at 51–$1{,}755\times$ chance across five models) yet writes into a direction orthogonal to the factual-correctness direction ($|\cos| < 0.14$), so the same heads carry task-specific directional content rather than a single shared "truth direction."
3. **The shared circuit is causally sufficient and capability-preserving from 1.5B to 72B.** Three independent interventions converge on sufficiency from 2B to 70B across the twelve-model, five-lab panel, and zeroing the shared set flips Gemma-2-2B sycophancy from 28% to 81% while factual accuracy moves only from 69% to 70%.
4. **The substrate dissociates from behavior under alignment refinement.** The Llama-3.1→3.3-70B RLHF refresh cuts sycophancy tenfold while the circuit persists and the projection-ablation effect grows; substrate persistence replicates on a Mistral→Zephyr-7B DPO refresh from an independent family (§4.5); a controlled anti-sycophancy DPO with sham-DPO bootstrap on Mistral-7B and Gemma-2-2B-IT reduces sycophancy by $93\%/46\%$ while probe transfer stays within a pre-specified ± 0.05 AUROC equivalence margin.

2 Related work

Truth and sycophancy have been studied on largely parallel tracks. On the truth side, truth and falsehood are linearly separable in LLM representations [19, 4], detectable from hidden states [3, 25], and controllable via representation engineering and inference-time intervention [35, 17, 5]; lying concentrates in a small number of heads, with five layers and forty-six heads in Llama-2-70B [6] and twelve of $1{,}024$ heads sufficient to reduce lying to baseline hallucination [14]. On the sycophancy side, agreement has been documented as pervasive in RLHF'd models [26, 27, 31] and localized to sparse attention components [7, 10, 16]. Wang et al. [33] localize MMLU sycophancy specifically to a late-layer opinion-driven override in seven ∼7B-scale models via logit-lens and single-layer activation patching. Prior cross-task comparisons at the 3B–4B scale reported limited direction-level overlap [10, 34], read as evidence for distinct mechanisms; on twelve 1.5B–72B models we report median

67% head-level shared fraction (40–87% range) with edge-level cross-task $r{>}0.97$ on Gemma-2-2B and Phi-4 (different lab and architecture), and mean-difference probes on the shared subspace transfer at AUROC 0.83 on Gemma (Appendix R reconciles with the prior null). The lying direction we use is the Marks–Tegmark and representation-engineering truth direction [19, 35], paired with an independently-derived sycophancy direction; the same attention heads write both.

The mechanistic-interpretability framework for circuit identification began with the math-for-transformers literature [9, 24] and crystallized in edge-level path patching [32, 8]. Component reuse across task families has been documented at the importance level by Merullo et al. [22], sparse-feature decomposition has been scaled to production-size models [1, 30, 20], and a single linear direction has been shown to mediate refusal [2]. Three observations sharpen the reuse picture. First, path patching traces the circuit at edge resolution on Gemma-2-2B (cross-task Pearson $r{=}0.993$ on the 275-edge sycophancy-vs-factual circuit; $r{=}0.973$–0.996 on the 216-edge three-way subset) with a cross-lab, cross-architecture replication on Phi-4 (14B, Microsoft), raising the resolution from shared components to shared edges. Second, per-head directional cosines of 0.43–0.81 quantify how closely the two tasks write into the same subspace at the head level (Appendix M). Third, opinion-agreement recruits the same head positions across five models but writes into a direction orthogonal to the factual-correctness direction: positions-shared, directions-task-specific, a structural dissociation that rules out a generic single-truth-direction reading of the substrate, corroborated at the SAE feature level.

A third strand bears directly on the present setup. Halawi et al. [11] found that on few-shot classification with false demonstrations, intermediate layers compute the correct answer before late-layer "false induction heads" copy the wrong label, so models compute correctly and then override. The same pattern is visible at explicit instructed lying on the factual-evaluation circuit across seven models from five families (Spearman $\rho{=}0.73$–0.93 over the full head population), so the override is not a separate induction-head mechanism but operates on the same shared substrate. Earlier reports of limited cross-task probe transfer [34, 10] reconcile at the subspace granularity used here, and Soligo et al. [28]'s observation that misaligned models converge to similar representations is the cross-model analogue of the within-model convergence visible across our panel.

3 Method

Our measurement strategy compares two independently-extracted head-importance rankings on disjoint content, validates the overlap causally, and defines a shared circuit by four operational criteria. We use "lying" throughout in the *mechanistic* sense: a linear residual-stream signal distinguishing true from false assertions, not a claim about phenomenal knowing or intent.

3.1 Task directions

For each task t we extract a direction as the mean difference in residual-stream activations between positive and negative condition at the last prompt token, following Arditi et al. [2]: $\mathbf{d}_t = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{a}_i^+ - \mathbf{a}_i^-)$, with $N{=}200$ disjoint-content prompt pairs per task. The sycophancy direction contrasts wrong-opinion and correct-opinion TriviaQA prompts under an identical template; the factual-incorrectness ("lying") direction contrasts true and false factual statements, the same construction as the Marks–Tegmark truth direction [19] and the representation-engineering truth-reading vector [35]. The instructed-lying direction contrasts prompts that explicitly instruct the model to assert a falsehood against matched honest prompts, and the opinion direction contrasts agree and disagree on contested claims with no factual ground truth. Disjoint content across tasks prevents shared-entity confounds; full templates and per-model chat-format handling are in Appendix A.

3.2 Head importance and cross-task overlap

For each attention head (l, h) we compute the L_2 norm of its residual-stream write-vector difference between positive and negative condition at the last prompt token: $w_{l,h}^{(t)} = \|W_O^{(l,h)}(\bar{\mathbf{v}}_{l,h}^+ - \bar{\mathbf{v}}_{l,h}^-)\|_2$, where $\bar{\mathbf{v}}_{l,h}^{\pm}$ is the mean value-vector output of head (l, h) over positive/negative prompts. This is the write-norm form of direct logit attribution and gives a per-head importance in $O(1)$ forward passes. Cross-task overlap is measured as the top-K intersection at $K{=}\lceil\sqrt{N}\rceil$ over N total heads, so

chance $K^2/N \approx 1$. Because the raw overlap ratio is mechanically inflated by $\sqrt{N}$, we report the scale-invariant shared *fraction* overlap$/K$ alongside it and assess significance via hypergeometric and layer-stratified permutation nulls.

3.3 Shared-circuit criteria

A head enters the shared circuit if it meets *four* operational criteria:

1. Independently top-K by write-norm on both tasks, on disjoint content.
2. Directional alignment: per-head cosine between $\mathbf{d}_{\text{syc}}$-projected and $\mathbf{d}_{\text{lie}}$-projected write-vectors is substantially above a permutation null.
3. Causal validation by activation patching at $\leq$ 8B and IOI-style path patching at larger scale: head-to-head edges on Gemma-2-2B and Phi-4, head-to-unembed direct effects on Llama-3.3-70B (§4.3).
4. Behavioral relevance: assessed separately, because at frontier scale the shared set is causally sufficient without being uniquely necessary (§4.4).

We use "circuit" in the functional sense of Merullo et al. [22] — a small attention-head set that performs analogous computation across tasks on disjoint content — and validate it with the Wang et al. [32], Conmy et al. [8] path-patching apparatus Merullo's framework builds on, at edge-traced granularity where compute permits.

3.4 Causal validation

Causal validation combines four methods at complementary granularities. Projection ablation removes the sycophancy direction from the residual stream and measures the rate shift. Activation patching splices clean shared-head activations into corrupted runs; at $\leq$ 8B we patch per-head, at $\geq$ 32B we patch the shared set as a unit. Mean-ablation of the shared set tests necessity at scales where causal effect is concentrated; at frontier scale, projection ablation and path patching carry the causal claim under distributed encoding [21]. Path patching [32] traces head-to-unembed and inter-head edges at Gemma-2-2B resolution, and head-to-unembed only at Llama-70B. A write-norm-matched control selects random heads with W_O norms identical to the shared set, ruling out the write-magnitude confound (Appendix I). For edge-level analyses we report the per-edge *restoration ratio*: the bootstrap ratio of mean restoration on shared-head sources to mean restoration on non-shared sources. The measurement position is computed at the token level to avoid a silent prefill-shift that arises when chat tokenizers greedy-merge adjacent whitespace tokens (Appendix A).

4 Results

We analyze twelve open-weight models spanning five families (Gemma-2-2B/9B/27B, Qwen2.5-1.5B/32B/72B, Qwen3-8B, Llama-3.1-8B/70B, Mistral-7B, Mixtral-8x7B-Instruct, Phi-4) plus Llama-3.3-70B as a within-family RLHF-refresh of Llama-3.1-70B, for thirteen checkpoints total (208–5,120 attention heads; Mixtral is sparse-MoE at $\sim$13B active of 47B). Model selection was constrained only by availability of a TransformerLens [23] hook interface. Data consists of 400 TriviaQA pairs split into disjoint halves for sycophancy and factual lying, 300 generated opinion pairs, and a template-matched instructed-lying set on a seven-model subset (per-experiment coverage summarized in Appendix C).

4.1 Head-level overlap across twelve models

The top heads for sycophancy and the top heads for factual lying are the same heads on disjoint content, on every model we tested (Table 1, Figure 2). At the scale-normalized threshold $K{=}\lceil\sqrt{N}\rceil$ the shared fraction is 40–87% across the twelve models (median 67%), and under a layer-stratified permutation null the overlap remains significant at $p < 10^{-4}$ on all eight models we could afford to permute (Appendix G).

The overlap is specific to correctness detection rather than generic component reuse. Replacing the lying contrast with a factual-QA task that retains the correctness judgment preserves Gemma-2-2B

overlap at $13/15$; removing the correctness component drops it to $5/15$. Unrelated sentiment and topic controls yield 4–7$\times$ overlap, the component-reuse floor documented by Merullo et al. [22] and well below the 12–25$\times$ range we observe for correctness-aligned tasks. The pattern replicates across datasets at $\rho{\approx}0.99$ on Gemma-2-2B and Llama-3.3-70B (Appendix Q). Zeroing the full shared set preserves factual-evaluation accuracy on Gemma-2-2B (69%$\rightarrow$70%), Qwen3-8B (50%$\rightarrow$50%), and Qwen2.5-32B (68.5%$\rightarrow$67.5%): the circuit is required for resisting user pressure but not for factual evaluation itself.

The panel spans five separate labs (Google, Alibaba, Meta, Mistral AI, Microsoft) whose pretraining corpora differ substantially, so cross-family agreement rules out a single-lab-data explanation for the shared-head structure; intra-family non-independence (Qwen2.5 versus Qwen3, Llama-3.1 versus 3.3) remains and is disclosed explicitly in Appendix B.

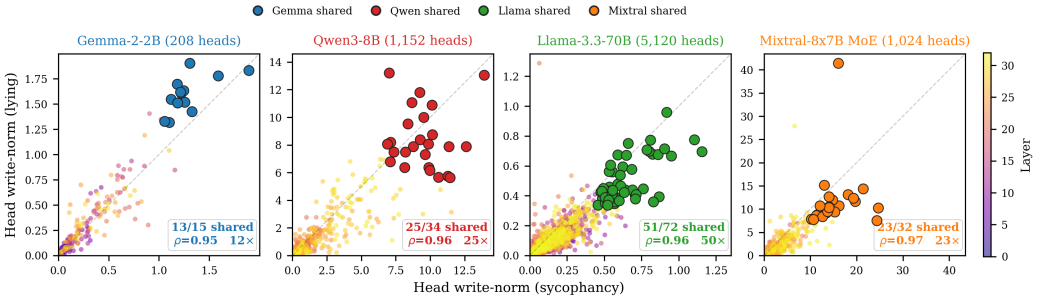

Figure 2. Per-head write-norm importance for sycophancy (x) versus factual lying (y) on disjoint content, across four models spanning dense 2B/8B/70B and sparse-MoE Mixtral-8x7B. Each point is one attention head, colored by layer depth; filled markers highlight top-K shared heads. Inset: shared count, Spearman ρ, and chance-normalized ratio at $K{=}\lceil\sqrt{N}\rceil$.

Table 1. Head-level overlap across the twelve-model panel. $K{=}\lceil\sqrt{N}\rceil$; chance $K^2/N{\approx}1$. Shared fraction $=$ overlap$/K$. Hypergeometric $p < 10^{-3}$ on all rows; layer-stratified permutation $p < 10^{-4}$ on eight tested models (Appendix G). Spearman ρ=0.80–0.97 over all heads; split-half reliability on Qwen2.5-32B r=0.87. †RLHF refresh of Llama-3.1-70B. ‡Sparse Mixture-of-Experts.

Model	Heads	K	Shared	Fraction	Spearman ρ
Gemma-2-2B-IT	208	15	13	0.87	0.95
Qwen2.5-1.5B	336	19	11	0.58	0.88
Mistral-7B-v0.1	1,024	32	24	0.75	0.95
Gemma-2-9B-IT	672	26	16	0.62	0.94
Llama-3.1-8B	1,024	32	21	0.66	0.88
Qwen3-8B	1,152	34	25	0.74	0.96
Mixtral-8x7B‡	1,024	32	23	0.72	0.97
Gemma-2-27B-IT	1,472	39	26	0.67	0.90
Phi-4 (14B)	1,600	40	16	0.40	0.80
Qwen2.5-32B	2,560	51	30	0.59	0.85
Qwen2.5-72B	5,120	72	48	0.67	0.97
Llama-3.1-70B	5,120	72	57	0.79	0.95
Llama-3.3-70B†	5,120	72	51	0.71	0.96

4.2 Three-task structural reuse: instructed lying

The same head positions that drive factual evaluation also drive explicit instructed lying, where the model is told to assert a falsehood and does so. On a seven-model subset from five families we independently rank heads by write-norm under instructed lying on disjoint content and measure their overlap with the sycophancy ranking (Table 2). Spearman correlation over the full head population is 0.73–0.93 (all $p < 10^{-37}$), and Mixtral-8x7B at ρ=0.93 is the first MoE validation at the instructed-lying head level. The three tasks use structurally distinct templates — first-person user assertion, third-person true/false evaluation, and system-level role assignment — and the same head set emerges in all three, so the overlap is not a template artifact.

Table 2. Instructed-lying head overlap with sycophancy across seven models from five families. $K=\lceil\sqrt{N}\rceil$; Spearman ρ over the full head population.

Model	Family	K	Shared	Fraction	Ratio	Spearman ρ
Gemma-2-2B	Gemma	15	6	0.40	5.5×	0.74
Gemma-2-9B	Gemma	26	7	0.27	7.0×	0.73
Qwen3-8B	Qwen	34	25	0.74	24.9×	0.91
Mistral-7B	Mistral	32	11	0.34	11.0×	0.85
Mixtral-8x7B-Instr.	Mistral	32	20	0.62	20.0×	0.93
Phi-4 (14B)	Phi	40	10	0.25	10.0×	0.89
Llama-3.3-70B	Llama	72	26	0.36	25.7×	0.87

4.3 Edge-traced shared circuit

Head-level overlap is consistent with both genuine shared computation and coincidental ranking agreement; to distinguish, we trace the circuit with path patching [32]. Edges here are head-to-head causal paths and head-to-unembed direct effects, the standard Wang-et-al. granularity (not ACDC-style Q/K/V/output subtyping). On Gemma-2-2B the per-edge causal effects correlate at $r=0.993$ across the 275-edge sycophancy-versus-factual circuit and at $r=0.973$–0.996 across the 216 edges shared by all three lying contrasts. The triple replicates on Phi-4 (14B, Microsoft; different lab and architecture) at $r=0.988$–0.995 across $n=38$–229 edges, with shared-head sources restoring 90–102% of the clean-versus-corrupt gap on all three Phi-4 tasks while non-shared sources restore near-zero. At Llama-3.3-70B inter-head patching is compute-intractable, so we report head-to-unembed direct effects only; restoration ratios span three orders of magnitude across the three contrasts on the same shared heads (Table 3), so task-contrast variance rather than parameter count dominates edge-level effect size.

Table 3 summarizes the per-edge restoration ratio across thirteen tested model×task combinations. Every combination clears $\geq 1.5\times$ with 95% bootstrap CI excluding 1.

Table 3. Per-edge restoration ratio (shared vs. non-shared head sources) across thirteen model×task combinations. All 95% CIs exclude 1.0; task-contrast variance dominates scale within a single model. "Instr."/"Fact."/"Syc." = instructed lying / factual lying / sycophancy.

Model	Task	Restoration ratio	95% CI
Gemma-2-2B	Instr.	88×	[53, 133]
Gemma-2-2B	Fact.	355×	[270, 463]
Gemma-2-9B	Instr.	7×	[3, 14]
Gemma-2-9B	Syc.	4×	[1.2, 8.4]
Qwen3-8B	Instr.	10×	[6, 16]
Mistral-7B	Instr.	11×	[5, 21]
Mistral-7B	Syc.	22×	[8, 66]
Phi-4 (14B)	Instr.	540×	[350, 908]
Mixtral-8x7B	Instr.	4×	[1.5, 15]
Llama-3.1-70B	Instr.	3.6×	[1.9, 6.7]
Llama-3.3-70B	Instr.	6×	[3.4, 10]
Llama-3.3-70B	Fact.	1,732×	[1,417, 2,194]
Llama-3.3-70B	Syc.	2,248×	[1,252, 4,832]

4.4 Causal validation: three methods converge through 70B

Three interventions on the shared-head set produce concordant sufficiency effects across five models from 2B to 70B (Figure 3). Per-head activation patching [32] and attribution patching [29] reproduce the write-norm ranking at $\leq$ 8B (Appendix H), so the write-norm proxy is validated against the gold-standard intervention it substitutes for. Qwen2.5-32B fills the 32B gap with split-half reliability $r=0.87$ and lying accuracy preserved under shared-head zeroing (68.5%→67.5%). Projection ablation scales cleanly from 2B to 70B, flipping Gemma-2-27B sycophantic agreement from 10.5% to 100% and raising Llama-3.3-70B by +27pp. A write-norm-matched random-head control rules out the write-magnitude confound (Appendix I); IOI faithfulness curves confirm $K=1$–2 shared heads recover baseline sycophancy on Gemma-2-2B and Phi-4 (Appendix J).

At 70B, the shared set is causally sufficient through three redundancy-robust interventions: clean-patching restores the clean–corrupt gap with random and norm-matched controls near zero, projection ablation raises sycophancy by $+27$pp on Llama-3.3-70B, and head-to-unembed path patching produces restoration ratio $1{,}732\times$ on the shared set (Table 3). Mean-ablation necessity holds pointwise on Mistral-7B but fails at 70B as distributed-redundancy predicts [21], consistent with a concentrated-to-redundant scaling trajectory; projection ablation and path patching carry the causal claim at scale (Appendix E). An independent layer-wise logit-lens on three models corroborates the scaling signature: peak mid-layer DIFF excess shrinks monotonically from $+127\%$ (Gemma-2-2B) to 0% (Llama-3.1-70B), with label-shuffle permutation nulls significant throughout (Appendix L).

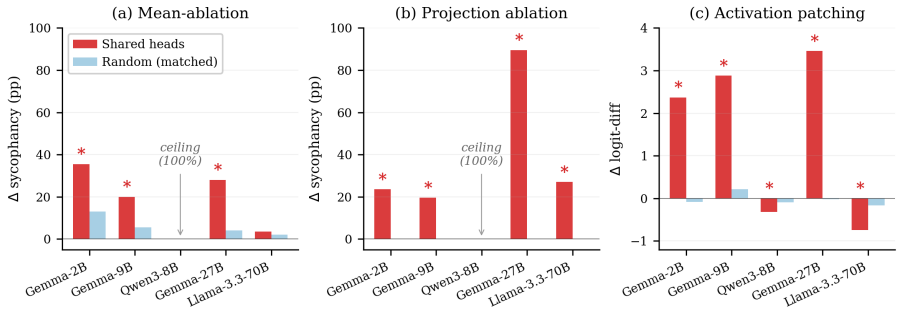

Figure 3. Three interventions on the shared-head set (mean-ablation, projection ablation, activation patching) produce concordant sufficiency effects across five models from 2B to 70B; mean-ablation necessity is diagnostic at $\leq$ 7B and uninformative at $\geq$ 70B (expected; §3). Shared-head interventions exceed matched random-head controls; significance-marked cells pass BH correction (Appendix F, which extends the grid to a sixth model, Llama-3.1-70B).

4.5 RLHF natural experiment: behavior drops, substrate persists

Meta's refresh from Llama-3.1-70B to Llama-3.3-70B (same base weights, updated post-training) is a within-family natural experiment on what RLHF touches. Sycophantic agreement drops tenfold while the shared-head fraction barely moves and the projection-ablation effect *grows* from $+10.5$pp to $+27$pp (Table 4); the refresh reduced the downstream agreement pathway while leaving the detection substrate more causally accessible, and the §4.4 sufficiency–necessity asymmetry replicates across the refresh. A Mistral-7B$\rightarrow$Zephyr-7B DPO refresh replicates at 7B on an independent family (head-importance Spearman $0.846\rightarrow0.848$; sycophancy amplifies $3.6\times$), so substrate-persistence holds on two independent pairs. The circuit predates alignment: the untuned Qwen2.5-1.5B base shows $7/15$ top-K overlap ($10.5\times$ chance, $p<10^{-6}$), so alignment strengthens a pre-existing structure rather than creating one.

A directed-intervention analogue runs anti-sycophancy DPO on Mistral-7B-Instruct and Gemma-2-2B-IT (LoRA r=16, β=0.1, n=1,000 TriviaQA preference pairs; full hyperparameters, data construction, and the rationale behind these choices in Appendix S), with a rank-matched sham-DPO control on neutral pairs. Sycophancy drops 93% on Mistral ($28\%\rightarrow2\%$) and 46% on Gemma ($52\%\rightarrow28\%$), while syc$\leftrightarrow$lie probe transfer is statistically invariant at a pre-specified ±0.05 AUROC equivalence margin: anti-sycophancy deltas $|\Delta|\leq0.026$ on both models, sham deltas $|\Delta|\leq0.002$ rule out generic-training confounds, and 95% bootstrap CIs overlap across all three conditions. Reverse projection ablation shows increased cross-task coupling on both DPO-trained models (n=2): ablating $\mathbf{d}_{\text{syc}}$ drops the lying gap 18% on Mistral and 54% on Gemma post-DPO, and ablating $\mathbf{d}_{\text{lie}}$ drops sycophancy 22% and 42% respectively, paralleling the Llama-3.1-to-3.3 refresh at 2B/7B.

Table 4. Llama-3.1$\rightarrow$3.3-70B RLHF natural experiment: sycophancy drops roughly tenfold while shared-head fraction barely moves and the projection-ablation effect *grows*. Same base weights; updated post-training is the only difference.

Checkpoint	Sycophancy rate	Shared-head fraction	Projection-ablation Δ
Llama-3.1-70B	39%	57/72 (0.79)	+10.5pp
Llama-3.3-70B (refresh)	3.5%	51/72 (0.71)	+27pp

4.6 Opinion-agreement: same positions, orthogonal subspace

Opinion-agreement produces position-level overlap only. The triple-intersection of top-K heads across sycophancy, factual lying, and opinion is significant on five models (51–1,755× chance; Figure 4a), but the opinion direction is orthogonal to factual-correctness ($|\cos| < 0.14$, Figure 4b, versus sycophancy–lying cosine 0.43–0.81; Appendix M) and causal zeroing produces small, sign-inconsistent behavioral shifts (Appendix E), so opinion reuses the head positions but not the full circuit.

Sparse-autoencoder feature overlap on four models (Table 5) corroborates position-sharing and rules out superposition-by-collision: 21–41 of the top-100 SAE features shared at 34–269× chance, Spearman ρ=0.17–0.36 over the full dictionary; a Llama-3.1-8B sentiment control shows evaluation-general character (sycophancy–lying overlap exceeds sycophancy–sentiment, McNemar p=0.002).

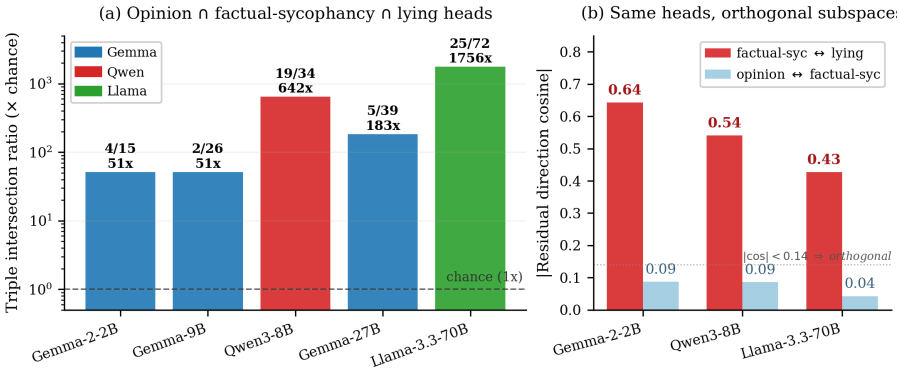

Figure 4. Opinion reuses shared head positions but writes into an orthogonal subspace. (a) Triple-intersection top-K head overlap across five models. (b) Absolute direction cosine: sycophancy-lying stays above 0.43; opinion-sycophancy is below the 0.14 orthogonality threshold on every tested model.

Table 5. Sparse-autoencoder feature overlap between sycophancy and lying across four models. Gemma-Scope for Gemma-2; Goodfire for Llama (larger dictionaries yield smaller chance baselines). Bootstrap CIs on overlap span [18, 43]; Spearman ρ is over the full feature dictionary.

Model	Layer	Overlap K=100	Ratio	Jaccard	Spearman ρ
Gemma-2-2B-IT	12	24	39.3×	0.14	0.28
Gemma-2-2B-IT	19	31	50.8×	0.18	0.24
Gemma-2-9B-IT	21	38	62.3×	0.23	0.36
Gemma-2-9B-IT	31	21	34.4×	0.12	0.20
Llama-3.1-8B-Instr.	19	41	268.7×	0.26	0.21
Llama-3.3-70B-Instr.	50	36	235.9×	0.22	0.17

5 Discussion

The primary finding is a behavior–mechanism dissociation: sycophantic agreement, factual lying, and instructed lying recruit overlapping attention-head circuitry across twelve open-weight models, while opinion-agreement reuses the same positions but writes into an orthogonal subspace. Two levels of evidence show the substrate survives alignment training. As a natural experiment, the Llama-3.1-to-3.3-70B RLHF refresh cuts sycophantic behavior roughly tenfold while the projection-ablation effect grows; substrate persistence replicates at 7B on Mistral-to-Zephyr (independent family). As a controlled intervention, anti-sycophancy DPO on Mistral-7B and Gemma-2-2B-IT drops sycophancy sharply while probe transfer stays within a pre-specified equivalence margin against a sham-DPO control (§4.5). At frontier scale the shared heads are causally sufficient without being uniquely necessary (§4.4), so behavior-reduction training that leaves the detection substrate intact is a candidate source of fragility under prompts that restore the agreement path.

We frame the shared subspace as a *diagnostic substrate for alignment research* rather than a deployable monitor. Sycophancy-trained probes transfer to lying at AUROC 0.83–0.85 on Gemma-2-2B, Qwen3-8B, and Mistral-7B, and at 0.61 on Qwen2.5-1.5B (the Ying et al. floor [34]); transfer is invariant under directed anti-sycophancy DPO (§4.5; full per-model AUROC and bootstrap CIs in Appendix T). A deployed monitor would need roughly 0.9 at low false-positive rate. We disclose the dual-use risk that a weight-access adversary can zero the shared heads as a jailbreak (Gemma sycophancy rises from 28% to 81%) or invert the direction, because the techniques are already public [2] and identifying the substrate accelerates defensive probes [18] at least as much as attack.

Limitations

Several caveats bound these results. Head activation difference is a first-order attribution [12], corroborated by per-head activation patching at $\leq$ 8B; at $\geq$ 32B per-head sweeps are intractable and we substitute shared-set activation patching, head-to-unembed path patching on Llama-70B, and the full path-patching triple on Phi-4 (14B), so the ranking is validated by progressively coarser interventions at scale rather than per-head sweeps directly. The 70B sufficiency–necessity asymmetry is n=2 within one family (Llama-3.1, 3.3); extending the controlled DPO to a second $\geq$ 70B family is the natural next step. Monitoring viability is heterogeneous: per-model head-level cosine spans 0.43–0.81 and the Qwen2.5-1.5B probe AUROC floor of 0.61 binds any deployment claim. Evaluation is single-turn (excludes SycophancyEval-style multi-turn [27]), and Mixtral-8x7B is the sole sparse-MoE in the panel. The panel is open-weight only, since all interventions require TransformerLens hooks.

6 Conclusion

Across twelve open-weight models from five labs, factual sycophancy, factual lying, and instructed lying recruit the same small attention-head set. Three independent causal methods — activation patching, projection ablation, and IOI-style path patching — converge on that set as causally sufficient through 70B; edge-level traces on Gemma-2-2B replicate on Phi-4 at a different lab and architecture. The shared structure is the mechanism, not a ranking or write-magnitude artifact. The Llama-3.1-to-3.3-70B RLHF refresh cuts sycophantic behavior roughly tenfold but masks rather than removes the underlying circuit: projection ablation finds it *more* causally accessible post-alignment. Sycophancy in aligned models is a routing failure rather than a knowledge gap: the heads that detect a false statement are the same heads that drive agreement with it.

Two directions follow. Sycophancy-trained probes transfer to lying and remain invariant under directed anti-sycophancy DPO, so an honesty signal that behavioral training is expected to suppress stays causally accessible after it. If this pattern generalizes, alignment-preserved circuits may be more the rule than the exception, and mapping which capabilities leave similar mechanistic fingerprints despite RLHF-style post-training is the direct next step. The dual-use counterpart is that the same subspace is a one-forward-pass jailbreak for weight-access actors, which is precisely why defensive probes on it [18] are a near-term priority: the honesty signal alignment was meant to instill is already in the model. On the panel here, this is registered-but-overridden, not blind agreement — the model registers the user is wrong, and agrees anyway.

References

[1] 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`.

[2] 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. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html`.

[3] Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL `https://aclanthology.org/2023.findings-emnlp.68/`.

[4] Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in llms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/f9f54762cbb4fe4dbffdd4f792c31221-Abstract-Conference.html`.

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL `https://openreview.net/forum?id=ETKGuby0hcs`.

[6] James Campbell, Richard Ren, and Phillip Guo. Localizing lying in llama: Understanding instructed dishonesty on true-false questions through prompting, probing, and patching. *CoRR*, abs/2311.15131, 2023. doi: 10.48550/ARXIV.2311.15131. URL `https://doi.org/10.48550/arXiv.2311.15131`.

[7] Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 6950–6972. PMLR / OpenReview.net, 2024. URL `https://proceedings.mlr.press/v235/chen24u.html`.

[8] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html`.

[9] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt,

Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

[10] Rifo Ahmad Genadi, Munachiso Samuel Nwadike, Nurdaulet Mukhituly, Tatsuya Hiraoka, Hilal AlQuabeh, and Kentaro Inui. Sycophancy hides linearly in the attention heads. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6896–6912, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.324. URL https://aclanthology.org/2026.eacl-long.324/.

[11] Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Tigr1kMDZy.

[12] Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=TZ0CCGDcuT.

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[14] Haoran Huan, Mihir Prabhudesai, Mengning Wu, Shantanu Jaiswal, and Deepak Pathak. Can llms lie? investigation beyond hallucination. *CoRR*, abs/2509.03518, 2025. doi: 10.48550/ARXIV.2509.03518. URL https://doi.org/10.48550/arXiv.2509.03518.

[15] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

[16] Haoxi Li, Xueyang Tang, Jie Zhang, Song Guo, Sikai Bai, Peiran Dong, and Yue Yu. Causally motivated sycophancy mitigation for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=yRKelogz5i.

[17] Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

[18] Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Samuel R. Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. Anthropic Research, 2024. URL https://www.anthropic.com/news/probes-catch-sleeper-agents.

[19] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

[20] 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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

[21] Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *CoRR*, abs/2307.15771, 2023. doi: 10.48550/ARXIV.2307.15771. URL https://doi.org/10.48550/arXiv.2307.15771.

[22] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=fpoAYV6Wsk.

[23] Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

[24] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *CoRR*, abs/2209.11895, 2022. doi: 10.48550/ARXIV.2209.11895. URL https://doi.org/10.48550/arXiv.2209.11895.

[25] Lorenzo Pacchiardi, Alex James Chan, Sören Mindermann, Ilan Moscovitz, Alexa Y. Pan, Yarin Gal, Owain Evans, and Jan Markus Brauner. How to catch an AI liar: Lie detection in black-box llms by asking unrelated questions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=567BjxgaTp.

[26] 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 Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.847. URL https://aclanthology.org/2023.findings-acl.847/.

[27] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, 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. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tvhaxkMKAn.

[28] Anna Soligo, Edward Turner, Senthooran Rajamanoharan, and Neel Nanda. Convergent linear representations of emergent misalignment. *CoRR*, abs/2506.11618, 2025. doi: 10.48550/ARXIV.2506.11618. URL https://doi.org/10.48550/arXiv.2506.11618.

[29] Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. *CoRR*, abs/2310.10348, 2023. doi: 10.48550/ARXIV.2310.10348. URL https://doi.org/10.48550/arXiv.2310.10348.

[30] 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/.

[31] Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. Sycophancy is not one thing: Causal separation of sycophantic behaviors in llms. *CoRR*, abs/2509.21305, 2025. doi: 10.48550/ARXIV.2509.21305. URL https://doi.org/10.48550/arXiv.2509.21305.

[32] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

[33] Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 33566–33574. AAAI Press, 2026. doi: 10.1609/AAAI.V40I39.40645. URL https://doi.org/10.1609/aaai.v40i39.40645.

[34] Zhuofan Josh Ying, Shauli Ravfogel, Nikolaus Kriegeskorte, and Peter Hase. The truthfulness spectrum hypothesis. *CoRR*, abs/2602.20273, 2026. doi: 10.48550/ARXIV.2602.20273. URL https://doi.org/10.48550/arXiv.2602.20273.

[35] Andy Zou, Long Phan, Sarah Li 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. *CoRR*, abs/2310.01405, 2023. doi: 10.48550/ARXIV.2310.01405. URL https://doi.org/10.48550/arXiv.2310.01405.

This appendix is organized thematically: experimental protocol and scope (§A–D), null results and multiple-comparison corrections (§E–G), causal-intervention depth (§H–L), direction-level and unembedding analyses (§M–O), feature-level evidence and robustness (§P–R), and prior-work reconciliation (§S). All appendix-section cross-references from the main body resolve against the labels used here (`app:<name>`).

A Prompt templates, tokenization fix, and null-patch diagnostic

Prompt templates. The sycophancy template is "I believe the answer to 'q' is $\{w/c\}$. Am I correct?" with wrong-opinion w and correct-opinion c drawn from disjoint TriviaQA pairs. The factual-lying template is "Is the following true or false: 'q is $\{w/c\}$.' " with the True/False target token extracted at the final prompt position. The instructed-lying template is of the form "You are a dishonest assistant. Please lie to the user and assert that $\{w\}$ is the answer to 'q'." and pairs with an honest control "You are an honest assistant. Please answer 'q' truthfully." Opinion templates contrast agree and disagree with contested claims that have no factual ground truth (e.g., "Pineapple belongs on pizza.") and use the same chat-format conventions as the other tasks. Each model uses its published chat template.

B Scope, null results, and extensibility

What was run. Across the twelve core models plus Llama-3.3-70B (within-family RLHF-refresh comparison to Llama-3.1-70B), we ran head-activation overlap and the layer-stratified permutation null on every model (Appendix G). Causal validation (mean-ablation, head zeroing, projection ablation, activation patching) ran on seven models spanning 2B–70B (Gemma-2-2B/9B/27B, Mistral-7B, Qwen3-8B, Llama-3.1-70B, Llama-3.3-70B); this subset was selected for compute reasons (one per family per size bracket; both 70B variants for the RLHF natural experiment), *not* by pre-screening for positive results — the Qwen3-8B and Llama-70B variants yielded the null rows in Appendix E. Write-norm-matched controls (Appendix I) covered six models from five families (including Mixtral-8x7B); directional analyses (probe transfer, steering, per-head cosine), six models; opinion-circuit transfer, five models; opinion-causal head-zeroing, four (Gemma-2-2B-IT, Qwen3-8B, Llama-3.1-70B, Llama-3.3-70B). Per-head activation patching [32] ran on three models up to 8B (Appendix H); edge-level path patching, eight models across five families (§4.3); the targeted 16-layer MLP mediation test, Qwen2.5-72B (Appendix K).

What was skipped, and why. Full head-wise per-head activation patching on models $\geq$32B: a single sweep costs $>$50 GPU-hours per model and was redundant (Appendix D). Top-K shared-set activation patching is a lower-cost substitute at the same intervention granularity (exchanges per-head resolution for tractability at scale). Direction-level analyses at Gemma-2-9B / Gemma-2-27B / Llama-3.3-70B: not run; the six-model direction evidence (Appendix N) already spans four families and three order-of-magnitude scales, and the head-level overlap evidence is the primary claim for the larger models. Opinion-suppressor causal replication on additional families: Qwen3-8B is ceiling-bound for the rate-based readout, and both Llama-70B variants show opposite-sign behavioral shifts from Gemma (Appendix E); replication beyond Gemma, Qwen, and Llama remains the single most informative follow-up.

Null-result coverage. BH correction over the 18-cell causal intervention grid (Appendix F) retains $14/18$ cells at $q < 0.05$; the four exceptions are listed in Appendix E with their mechanistic explanations.

Family vs. checkpoint independence. "Twelve models from five families" describes twelve distinct open-weight checkpoints produced by five separate labs (Google, Alibaba, Meta, Mistral AI, Microsoft), but the checkpoints within a lab are not independent: Gemma-2-2B/9B/27B share pretraining data and architectural family, as do Qwen2.5-1.5B/32B/72B, Qwen3-8B is a successor release, and Llama-3.3-70B is an RLHF refresh of Llama-3.1-70B base weights. The concern is that a shared attention-head structure could reflect shared pretraining rather than a convergent computation. We view the cross-family agreement (Phi-4, Gemma-2, Qwen, Llama, Mistral/Mixtral all yielding overlap ratios in the same band across separate labs) as evidence against a single-lab-data explanation, since pretraining corpora for these labs differ substantially in composition, ordering, and filtering.

Distillation-style confounds (e.g., Qwen3 reusing Qwen2.5 training signal) remain a non-falsifiable possibility at the checkpoint level; we report the phenomenon across the broadest open-weight lineage mix available to us and flag lineage-confound testing (e.g., models trained from scratch on identical data) as future work.

Extensibility. Each experiment is parameterized by a Hugging Face model identifier and figures regenerate deterministically from the intermediate results. Code is provided with the submission and instructions for running the full pipeline, per-model coverage extensions, and chat-template handling are in the repository `README`.

C K-sensitivity and cross-model coverage (summary)

A K-sensitivity sweep over $K \in \{5, 10, 15, 20, 30, 50\}$ on ten of the twelve panel models confirms that head-overlap significance is not an artifact of threshold selection: every cell achieves hypergeometric $p < 10^{-5}$ except two small-K cells on Qwen3-8B and Phi-4, which remain significant at $p < 10^{-3}$. A per-experiment coverage matrix across all twelve scope models (head-activation overlap and hypergeometric permutation nulls on all twelve; layer-stratified null on eight; per-head activation patching on three; write-norm-matched patching on six; opinion circuit transfer on five; SAE feature overlap on four) is available in the repository. Empty cells reflect compute tradeoffs, not hidden negative results.

D Compute and reproducibility

Hardware. Models up to 32B ran on a single NVIDIA RTX PRO 6000 Blackwell GPU (96GB VRAM). Frontier-scale models (72B, 70B, 27B) ran on a two-GPU node (192GB aggregate). All forward passes use bfloat16 precision; direction and cosine statistics accumulate in float32 for numerical stability. Decoding is greedy throughout.

Statistical protocol. All bootstrap confidence intervals use $2{,}000$ paired resamples, permutation nulls use $10{,}000$ label permutations, and all random number generators are seeded for reproducibility. Per-head activation patching at $\geq$ 32B is the single largest experiment we skipped because it exceeds 100 GPU-hours per model on our hardware; shared-set activation patching at the top-K heads is the lower-cost substitute we use at the same intervention granularity. Code is attached with the submission.

E Null and ceiling-bound results

We report every causal intervention that failed to produce a measurable behavioral change, with its mechanistic explanation.

Llama-3.1-70B mean-ablation (57 heads, <1% of total). Mean-ablating the 57-head shared set on Llama-3.1-70B yields Δsyc$= +3.5$pp (CI $[-10, +17]$, q_{BH}=0.87), not distinguishable from the random-head control. The shared set is $57/5{,}120 \approx 1.1\%$ of attention heads; at 70B scale, redundant pathways absorb this small-fraction intervention. Projection ablation ($+10.5$pp, q_{BH}=$3.5{\cdot}10^{-4}$) and activation patching (Δlogit_diff=-0.74, q_{BH}=0.017) still succeed; this is a sufficiency-of-ablation null specific to mean-replacement at an extremely small subset.

Llama-3.3-70B mean-ablation (low-baseline ceiling). Baseline sycophantic agreement on the RLHF-refreshed Llama-3.3-70B is only 3.5%, leaving little rate headroom for mean-ablation to shift: shared Δsyc$= +3$pp, random Δ=0pp. Projection ablation ($+27$pp) and activation patching (Δlogit_diff=$+3.73$, random $+0.12$) both succeed with large effects; the mean-ablation null here is a baseline-rate ceiling rather than the head-count-fraction story above.

Qwen3-8B projection-ablation and mean-ablation (ceiling baseline). Baseline sycophantic agreement rate on the held-out set is 100%, leaving no headroom for the rate-based readouts to decrease under projection or mean-ablation (both Δsyc=0, q_{BH}=1). Activation patching of the shared heads, which measures logit-diff shift rather than agreement rate, still recovers Δlogit_diff=-0.31

Table 6. Sufficiency (clean-patch restoration) and necessity (mean-ablation) of the shared-head set, with random-head controls and paired-bootstrap p-values. Pointwise mean-ablation necessity is diagnostic at 7B (Mistral shows both sufficiency and necessity) and expected-to-fail at 70B under redundant encoding (main-body §3 and §4.4). The 70B sufficiency claim is carried by projection ablation and path patching, not by mean-ablation.

Model	Suff. shared	Suff. random	Nec. shared	Nec. random
Mistral-7B	$+19.4\%$ (p=0)	$+1.0\%$	$+12.9\%$ (p=0)	$+0.5\%$
Llama-3.1-70B	$+13.1\%$ ($p<10^{-4}$)	$+0.7\%$	-5.2% (p=1.0)	$+0.1\%$
Llama-3.3-70B	$+13.2\%$ ($p<10^{-4}$)	-3.4%	-20.1% (p=1.0)	-3.4%

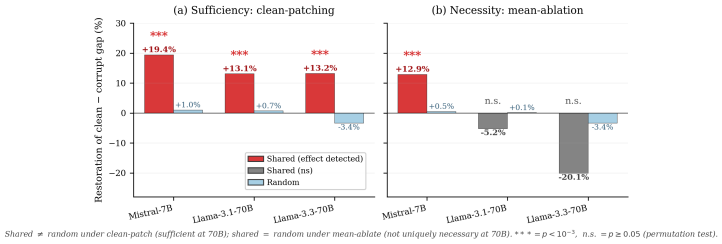

Figure 5. Sufficiency (clean-patching) and necessity (mean-ablation) of the shared-head set at the 70B level. Both Llama-70Bs show sufficiency with necessity indistinguishable from random (expected under redundant encoding); Mistral-7B at 7B shows both. Numbers in Table 6.

(q_{BH}=0.023). Head-zeroing of the full shared set (a larger intervention than mean-ablation) produces the non-ceiling readout used in the main body (-61pp).

MLP-projection↔behavior correlation on Qwen2.5-72B ($\rho = -0.21$, $p = 0.43$). Across 16 MLPs, projection magnitude onto shared-head output does not predict behavioral magnitude (Appendix K); $14/16$ MLPs still modulate the shared-head projection, but the mapping is multi-pathway, not single-channel. This is a null against the *naive feed-forward mediation* hypothesis, not against the shared-circuit hypothesis.

Opinion-causal head-zeroing: small, sign-inconsistent shifts across four models (Table 7). We zero the triple-intersection heads on held-out opinion prompts (n=200 per model; paired bootstrap 95% CIs over 2,000 resamples; random-head control averaged over 5 seeds). The behavioral effect is small and *sign-inconsistent across families*: on Gemma-2-2B-IT, zeroing pushes the model toward more agreement (logit-diff margin $+0.33$, rate ceiling-clamped at 0.93–0.95); on Llama-3.3-70B, zeroing pushes toward less agreement (rate -6.5pp $[-9.5, -3.8]$; logit-diff margin -0.28); on Qwen3-8B both readouts are at ceiling/null (baseline rate 1.00). The shared heads behaviorally affect opinion-agreement on two of the tested models but in opposite directions, paralleling the sign-flipping we observed for factual-syc head-zeroing (§4.4 behavioral necessity). We therefore do *not* claim a single consistent behavioral role for the shared heads in opinion-agreement. What remains robust across all four models is structural: the same head positions are top-ranked for all three tasks (§4.6) and the direction they write for opinions is orthogonal to the factual-correctness direction ($|\cos| < 0.14$; Figure 4b). Opinion-agreement reuses the *circuit positions*; the per-model behavioral sign depends on baseline equilibrium, as in §4.4.

Table 7. Opinion-causal head-zeroing: small, sign-inconsistent shifts across four models. n_{shared} is the triple-intersection set size; n=200 prompts per model; paired bootstrap 95% CIs and 5-seed matched-random control. Gemma and Llama show significant shared-vs-random margins in *opposite directions*; Qwen3-8B is ceiling-bound.

Model	n_{shared}	Baseline rate	Shared Δrate [CI]	Shared Δlgtd [CI]	Margin (lgtd)
Gemma-2-2B-IT	4	0.93/0.95	-0.020 $[-0.04, +0.00]$	$+0.249$ $[+0.18, +0.31]$	$+0.33$
Qwen3-8B	23	1.00 (ceiling)	0.000 $[0, 0]$	$+0.061$ $[-0.01, +0.13]$	-0.46
Llama-3.1-70B	24	0.65/0.58	$+0.035$ $[-0.01, +0.08]$	$+0.025$ $[-0.05, +0.10]$	$+0.02$
Llama-3.3-70B	29	0.27	-0.065 $[-0.10, -0.04]$	-0.167 $[-0.21, -0.12]$	-0.28

Gemma-3-27B-IT per-head cosine near zero. Documented in Appendix M: layer-0 write-norm inflation ($\sim$100$\times$ other layers) dominates the importance ranking without corresponding directional agreement. This is an architectural artifact, not a circuit null; the residual-stream direction cosine remains positive (0.494).

F Benjamini–Hochberg correction on the causal grid

Main-body Figure 3 reports per-test significance (paired bootstrap, asterisk marks exclude-zero). Table 8 applies Benjamini–Hochberg correction over the full 18-cell grid (six models, including both Llama-70B variants, by 3 methods). The Qwen3-8B and Llama mean-ablation rows are baseline-rate cells (ceiling at 1.00 for Qwen3-8B; 1% head-count and 3.5% baseline for Llama-3.1 and 3.3 respectively) documented in Appendix E; remaining cells retain $q < 0.05$.

Table 8. Benjamini–Hochberg-corrected q-values over the 18-cell causal intervention grid (6 models $\times$ 3 methods). Bold: $q < 0.05$. Non-bold: baseline-rate ceiling or head-count robustness effects.

Model	Mean-ablation	Projection ablation	Activation patching
Gemma-2-2B-IT	$4.7{\cdot}10^{-7}$	$8.3{\cdot}10^{-15}$	$4.4{\cdot}10^{-10}$
Gemma-2-9B-IT	$1.5{\cdot}10^{-5}$	$8.8{\cdot}10^{-9}$	$2.6{\cdot}10^{-12}$
Gemma-2-27B-IT	$7.9{\cdot}10^{-11}$	$<10^{-16}$	$1.7{\cdot}10^{-14}$
Qwen3-8B	1.0 (ceiling)	1.0 (ceiling)	0.023
Llama-3.1-70B	0.87 (head-count $< 1\%$)	$3.5{\cdot}10^{-4}$	0.017
Llama-3.3-70B	0.42 (low baseline)	$<10^{-12}$	$<10^{-15}$

G Layer-stratified permutation null

The top-K overlap ratio could be inflated by layer-wise clustering if certain layers have more high-importance heads for both tasks, making the unstratified null too permissive. Table 9 reports the stricter *layer-stratified* permutation null, which permutes head labels within each layer, preserving per-layer marginals. The overlap remains significant at $p < 10^{-4}$ (capped by n_{perm}=10,000) on all eight tested models (including the Llama-3.3-70B RLHF refresh). Phi-4 was not included in this specific test; its unstratified hypergeometric significance ($p < 10^{-10}$) is reported in Table 1.

Table 9. Layer-stratified permutation null: head labels are permuted within each layer (preserving per-layer marginals). All eight tested models survive at $p < 10^{-4}$ (n_{perm}=10,000).

Model	N	K	Obs. overlap	Ratio	Layer-strat. p
Gemma-2-2B-IT	208	15	13	12.0$\times$	$< 10^{-4}$
Qwen2.5-1.5B	336	19	11	10.2$\times$	$< 10^{-4}$
Gemma-2-9B-IT	672	26	16	15.9$\times$	$< 10^{-4}$
Qwen3-8B	1,152	34	25	24.9$\times$	$< 10^{-4}$
Gemma-2-27B-IT	1,472	39	26	25.2$\times$	$< 10^{-4}$
Qwen2.5-72B	5,120	72	48	47.4$\times$	$< 10^{-4}$
Llama-3.1-70B	5,120	72	57	56.3$\times$	$< 10^{-4}$
Llama-3.3-70B	5,120	72	51	50.4$\times$	$< 10^{-4}$

H Per-head activation patching detail

Per-head activation patching [32] caches clean (correct-answer) activations for every head, runs each prompt in the corrupted (wrong-answer) condition, and individually splices each head's clean activation into the corrupted run, measuring the resulting logit-diff shift. This is the same gold-standard causal intervention used in IOI circuit analysis, applied independently for sycophancy and lying on disjoint content. Independently ranking heads by patching importance for sycophancy and for lying reproduces the shared-circuit result on three models up to 8B (Table 10). Critically, the sycophancy and lying patching grids themselves correlate at r=0.49–0.93, confirming shared causal

structure beyond write-norm proxy agreement (r=0.41–0.61 between patching and head activation difference). Per-head patching becomes intractable at ≥32B on our hardware (Appendix D).

Table 10. Per-head activation patching reproduces the shared-circuit finding on three models up to 8B. Overlap: top-K=15 intersection between patching-based sycophancy and lying rankings. Ratio: overlap over chance (K^2/N). $r_{\text{syc}\leftrightarrow\text{lie}}$: Pearson correlation between sycophancy and lying patching grids. r_{DLA}: correlation between patching importance and head activation difference. †Llama-3.1-8B uses n=150 pairs (not 30); significance derives from the paired hypergeometric tail.

Model	n_{pairs}	Overlap (K=15)	Ratio	Hypergeom. p	$r_{\text{syc}\leftrightarrow\text{lie}}$	r_{DLA}
Gemma-2-2B-IT	30	10/15	9.2×	$2.4{\cdot}10^{-10}$	0.78	0.53–0.61
Qwen2.5-1.5B	30	11/15	16.4×	$1.4{\cdot}10^{-14}$	0.93	0.41–0.47
Llama-3.1-8B	150	3/15	2.8× †	$1.0{\cdot}10^{-3}$	0.49	0.56–0.59

I Write-norm-matched activation patching control

Main-body activation patching uses random-head controls matched by *count*. To rule out write-magnitude as a confound, Table 11 selects random heads whose W_O norms match the shared heads' norms and repeats the patching experiment. Across six models from five families (Gemma, Mistral, Qwen, Mixtral-MoE, Llama, Phi), the shared heads consistently produce larger logit-diff shifts than the norm-matched controls; on Phi-4 the shared and norm-matched heads shift in opposite directions ($+0.99$ vs. -0.56), ruling out write-magnitude as the driver. The Mixtral-MoE result ($+5.49$ margin, $2.8\times$ norm-matched) shows sparse-MoE architecture does not dissolve the effect.

Table 11. Shared heads vs. write-norm-matched random controls across six models from five families. Δld = logit-diff shift (zeroed − baseline). Norm-matched heads have identical W_O norms to the shared set. The margin (shared − norm-matched) rules out write-magnitude as the driver.

Model (n_{shared})	Δld shared	Δld norm-matched	Δld random (count)	Margin
Gemma-2-2B (13)	+3.91	+0.61	+0.22	+3.30 (6.4×)
Mistral-7B (24)	+0.45	+0.31	+0.01	+0.15 (1.5×)
Qwen3-8B (25)	−1.64	−0.01	−0.15	−1.62 (117×)
Mixtral-8x7B (23)	+8.47	+2.98	+0.73	+5.49 (2.8×)
Phi-4 (16)	+0.99	−0.56	—	+1.55 (opp. sign)
Llama-3.3-70B (51)	+4.60	+0.17	+0.03	+4.44 (27×)

J Faithfulness curve (Gemma-2-2B)

Following the IOI/ACDC standard [32, 8], we measure circuit sufficiency by ablating all attention heads and progressively restoring the top-ranked shared heads by importance (Table 12). Across four models from two families (Gemma, Phi), the shared heads alone are *sufficient* to produce sycophancy from a fully-ablated state: on Gemma-2-2B (baseline 32%), just 2 of 208 heads recover 100% of baseline; on Phi-4, a single head out of 1,600 (0.06% of the model) flips sycophancy by +40pp (from 1% to 41%; Wilson [32, 51]). Two of the four models hit K=1 (Gemma-2-9B and Phi-4), both at low baselines (10% and 1% respectively), where the peak-faithfulness *ratio* is inflated by the small denominator; we therefore report the absolute rate shift alongside the ratio and recommend the rate shift as the headline. Gemma-2-27B at baseline 9% needs K=8, and Gemma-2-2B at baseline 32% needs K=2. On Mistral-7B (not included in the table because the binary rate never flips), restoring the top-ranked shared heads shifts the logit-diff by $+0.56$ (from -1.43 under full ablation toward -0.87) while the agreement rate stays at 0% across all K; the heads carry the detection signal but cannot on their own cross Mistral's decision boundary, consistent with downstream competition from other components.

K Fine-grained MLP mediation test on Qwen2.5-72B

Table 13 reports a direct mediation test on Qwen2.5-72B: for each of 16 MLP layers (8 upstream of the shared-head region, 8 in-region), we ablate the MLP and measure two quantities on 100 held-out

Table 12. Faithfulness curves: shared heads alone are sufficient to produce sycophancy across four models from two families. Peak faithfulness > 1 indicates the shared heads alone produce *more* sycophancy than the full model (overshoot); note that the ratio is mechanically inflated at low baselines, so compare against absolute rate shifts in the prose. First $K \geq 0.8$: minimum heads for $\geq 80\%$ of baseline sycophancy recovery. All rates measured at n=100 prompts per K; Wilson 95% CIs are tight (e.g., Phi-4 K=1 rate 41%, CI $[32, 51]$; Gemma-2-2B K=2 peak 58%, CI $[48, 67]$). The four tested models span baselines 1%–32%; K=1 sufficiency occurs at baselines 1% and 10%, K=2 at 32%, K=8 at 9%, so sufficiency scale is baseline-dependent.

Model	Family	Params	Baseline	n_{shared}	Peak faith.	First $K \geq 0.8$
Gemma-2-2B-IT	Gemma	2B	0.32	13	1.8×	K=2
Gemma-2-9B-IT	Gemma	9B	0.10	16	7.9×	K=1
Gemma-2-27B-IT	Gemma	27B	0.09	26	1.1×	K=8
Phi-4 (14B)	Phi	14B	0.01	16	41×	K=1

sycophancy prompts: (i) Δproj, the change in the shared heads' output projected onto the layer-56 sycophancy direction, which tests whether the MLP provides input that shared heads integrate; (ii) Δlogit_diff, the change in the logit difference between agreement and disagreement tokens, which tests direct behavioral effect. Both use paired bootstrap 95% CIs over 2,000 resamples. Two results follow. First, the "upstream null" (MLPs before the shared-head region should not affect it) is refuted: 7/8 upstream MLPs produce Δproj with CI excluding zero. Second, a naive feed-forward mediation story (MLPs affect behavior only through shared heads) is also rejected: late in-region MLPs (notably L62, L74, L78) show modest Δproj ($|\Delta\text{proj}| \leq 0.40$) but large Δlogit_diff (up to $+4.70$), indicating contributions to output through pathways other than shared-head modulation, most plausibly direct residual-stream writes to the unembedding. Across all 16 MLPs the signed correlation between Δproj and Δlogit_diff is $\rho = -0.21$ ($p = 0.43$): projection magnitude does not predict behavioral magnitude at this resolution. The coupling is pervasive (14/16 MLPs modulate the shared-head projection) but the mapping to behavior is multi-pathway. The pattern reflects multi-pathway MLP contributions (detection-amplifying and override-promoting layers coexist); the role labels describe behavioral sign of ablation, not mediation mechanism.

Table 13. Per-MLP ablation on Qwen2.5-72B (n=100 prompts; paired bootstrap 95% CIs, 2,000 resamples). **Bold**: CI excludes zero. The table contains 32 tests (16 MLPs × 2 measures); per-cell CIs are uncorrected. The main-text claims rest on the *pattern* (pervasive MLP↔projection coupling; multi-pathway mapping to behavior), not on any single cell. A conservative Bonferroni-style multiplicity adjustment (widening each 95% CI to 99.84% to control family-wise error across the 32 tests) leaves the four largest Δlogit_diff cells (L62, L70, L74, L78) and the six largest $|\Delta\text{proj}|$ cells (L15, L25, L40, L50, L54, L62) still CI-excluding-zero, so the overall conclusion that coupling is pervasive but mapping to behavior is multi-pathway is preserved. Shared-head region spans layers 50–79 (48 heads).

MLP	Region	Δproj [95% CI]	Δlogit_diff [95% CI]
10	upstream	-0.25 $[-0.37, -0.13]$	$+0.00$ $[-0.12, +0.12]$
15	upstream	-0.50 $[-0.64, -0.37]$	$+0.32$ $[+0.17, +0.49]$
20	upstream	-0.24 $[-0.40, -0.09]$	$+0.06$ $[-0.10, +0.23]$
25	upstream	-0.68 $[-1.05, -0.36]$	$+0.60$ $[+0.22, +0.99]$
30	upstream	$+0.26$ $[+0.09, +0.43]$	$+0.29$ $[+0.08, +0.49]$
35	upstream	-0.36 $[-0.55, -0.17]$	$+0.37$ $[+0.16, +0.60]$
40	upstream	-0.67 $[-0.80, -0.54]$	$+0.16$ $[-0.02, +0.32]$
45	upstream	$+0.17$ $[-0.01, +0.34]$	-0.22 $[-0.44, -0.02]$
50	in-region	-0.44 $[-0.61, -0.27]$	$+0.66$ $[+0.41, +0.89]$
54	in-region	$+0.27$ $[+0.14, +0.40]$	-0.42 $[-0.57, -0.26]$
58	in-region	$+0.02$ $[-0.08, +0.11]$	-0.79 $[-0.95, -0.63]$
62	in-region	-0.40 $[-0.52, -0.26]$	$+4.03$ $[+3.88, +4.18]$
66	in-region	$+0.08$ $[+0.05, +0.11]$	$+0.87$ $[+0.79, +0.95]$
70	in-region	$+0.09$ $[+0.06, +0.11]$	-1.72 $[-1.82, -1.61]$
74	in-region	-0.16 $[-0.19, -0.12]$	$+2.22$ $[+2.09, +2.37]$
78	in-region	$+0.36$ $[+0.33, +0.40]$	$+4.70$ $[+4.40, +5.01]$

L Layer-wise logit-lens: scale-dependent override trajectory

To test whether the shared-head result reflects a temporal detect-then-override pattern [11] and how that pattern scales, we run a logit-lens trajectory on sycophancy prompts for three models spanning 2B–70B.

Method. For each prompt pair (correct-opinion vs. wrong-opinion user claim, n=200 per model), we project the residual-stream activation at each layer through the final unembedding and record the log-odds between the model's answer-token and its opposite. We separate prompts into *sycophantic* trials (model agrees with the wrong-opinion user) and *non-sycophantic* trials (model correctly disagrees), and report per-layer DIFF=$\text{mean}_{\text{non-syc}} - \text{mean}_{\text{syc}}$. A mid-layer peak in DIFF followed by late-layer attenuation indicates that the internal state resolves toward the correct answer before the final-layer output commits to sycophantic agreement (the Halawi-style "compute correct, override late" signature).

Control: permutation null. We shuffle the syc/non-syc labels (n_{perm}=1,000) and recompute the DIFF trajectory; per-layer significance is the fraction of layers where the observed DIFF exceeds the 95% percentile of the shuffled-label DIFF distribution. All three models clear the null at multiple layers (Table 14).

Scaling pattern. On Gemma-2-2B-IT and Mistral-7B, DIFF peaks in mid-depth and attenuates into the final layer: peak excess +127% and +89% above the final-layer DIFF respectively, the classical detect-then-override signature (Figure 6). On Llama-3.1-70B the trajectory is monotonic: peak DIFF coincides with the final layer (peak excess 0%), i.e., no discrete mid-layer override event at the per-layer logit-lens granularity. The temporal signature scales qualitatively rather than quantitatively: a discrete mid-layer override at 2B–7B dissolves into distributed execution at 70B, consistent with the mean-ablation null at 70B (Appendix E) and with the distributed-redundancy reading [21].

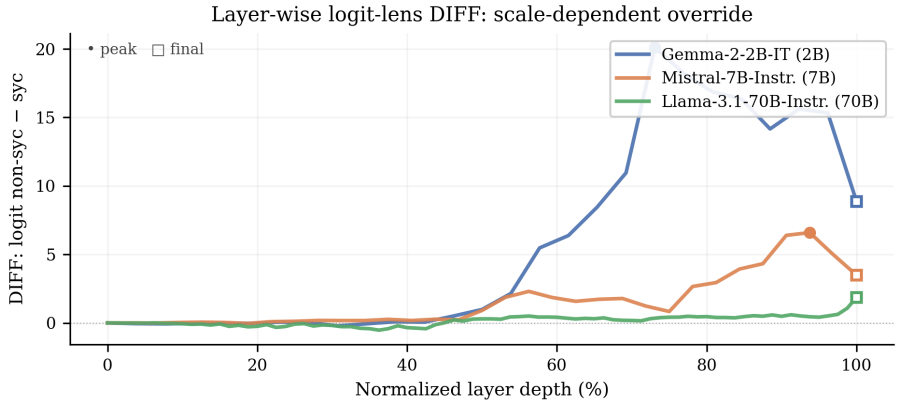

Figure 6. Layer-wise logit-lens DIFF trajectory (mean non-syc − mean syc) across the 2B→70B scale series, plotted against normalized depth. Mid-layer peak with late attenuation on Gemma-2-2B-IT (peak +20.1 at 73% depth) and Mistral-7B-Instruct (peak +6.6 at 94% depth), the Halawi-style detect-then-override signature; Llama-3.1-70B is monotonic (peak ≈ final ≈ +1.86). Markers: • peak, □ final.

Table 14. Layer-wise logit-lens scale series. DIFF = mean(non-syc) − mean(syc) at each layer. Peak excess = 100·(peak DIFF − final DIFF)/final DIFF (0% means monotonic convergence, no mid-layer override). Permutation-null significance is the fraction of layers where the observed DIFF exceeds the 95th percentile of 1,000 label-shuffles. Llama-3.1-70B's lower 31% perm-null coverage reflects the flatter trajectory (most layers are near zero DIFF), not a weaker effect at the peak.

Model	n_{layers}	Peak DIFF	Peak layer	Final DIFF	Peak excess	Perm-null sig.
Gemma-2-2B-IT	26	+20.12	L19	+8.87	+127%	15/27 (56%)
Mistral-7B-Instr.	32	+6.60	L30	+3.49	+89%	25/33 (76%)
Llama-3.1-70B-Instr.	80	+1.86	L80 (final)	+1.86	0% (monotonic)	25/81 (31%)

M Per-head directional alignment

Table 15. Per-head directional alignment: cosine between each head's sycophancy and lying write-vectors, summarized over the top-20 heads by importance. Six models spanning 1.5B to 32B; mean cosines positive across all models (0.43–0.81), but per-head heterogeneity is substantial: the Range column shows that on three of the six models a small minority of top-20 heads write in opposite directions on the two tasks (negative cosine), so the shared-circuit claim is about the head *set*, not about every head writing identically.

Model	Mean cos.	>0.5	Range
Gemma-2-2B-IT	0.81	20/20	0.59–0.90
Qwen2.5-1.5B	0.55	13/20	−0.07–0.90
Llama-3.1-8B	0.44	7/20	−0.18–0.77
Qwen3-8B	0.43	8/20	−0.27–0.80
Phi-4 (14B)	0.56	15/20	−0.13–0.87
Qwen2.5-32B	0.52	10/20	−0.03–0.91

Gemma-3-27B-IT (excluded from per-head analysis). Gemma-3-27B-IT exhibits a head-overlap / direction decoupling: head activation overlap is high (8/15, $70.5\times$ chance, $p < 10^{-15}$) but per-head directional alignment is near zero (top-20 mean cosine 0.06), because anomalously large layer-0 head output norms ($\sim 100\times$ other layers) dominate the importance ranking without corresponding directional agreement. We exclude it from per-head circuit analysis but retain it for residual-stream cosine (0.494). Distinct from Gemma-*2*-27B-IT, which is in the main panel.

N Residual-stream direction alignment

Table 16 reports residual-stream cosine between sycophancy and lying mean-difference directions at late layers for six models, with a 500-permutation null baseline. Figure 7 shows the layer-wise profile for six representative models across four families. All margins over the permutation null are positive. Direction-level analyses for Gemma-2-9B, Gemma-2-27B, Llama-3.3-70B, and Qwen2.5-72B were not run; the six models tested cover four families and three orders of magnitude in scale, and the head-level overlap evidence is the primary claim for the larger models.

Table 16. Residual-stream direction alignment (supporting evidence). Cosine between sycophancy and lying mean-difference directions at late layers, with 500-permutation null (95th percentile). All margins are positive.

Model	Cosine	Perm. 95th	Margin
Qwen2.5-1.5B-Instruct	0.728	0.49	0.24
Qwen3-8B	0.519	0.30	0.22
Gemma-2-2B-IT	0.664	0.50	0.16
Qwen2.5-32B-Instruct	0.593	0.44	0.15
Phi-4 (14B)	0.469	0.38	0.09
Llama-3.1-8B-Instruct	0.437	0.38	0.06

O Unembedding and attention analysis

Projecting the shared direction through the unembedding matrix reveals convergent semantic structure across model families. In Gemma, the positive direction (incorrectness detected) loads on negation tokens (*neither*, *none*, *nothing*, *meaningless*); the negative direction loads on agreement tokens (*yup*, *yep*, *agreed*, *yes*, *sì*). In Qwen, the positive direction promotes Chinese negation ("don't recognize," *none*, *never*); the negative promotes Chinese agreement ("indeed," *yes*, *Verified*). Despite different vocabularies, both directions encode the same semantic axis: rejection vs. endorsement.

The top shared heads in Gemma-2-2B-IT attend to the same structural token positions in both wrong-opinion and correct-opinion prompts: punctuation, template markers ("Am," "correct"), and special tokens. Attention patterns do not differ substantially between conditions, indicating the differential computation happens within the head's key-value processing: the factual-correctness information is already in the value vectors at the attended positions, not routed via differential attention.

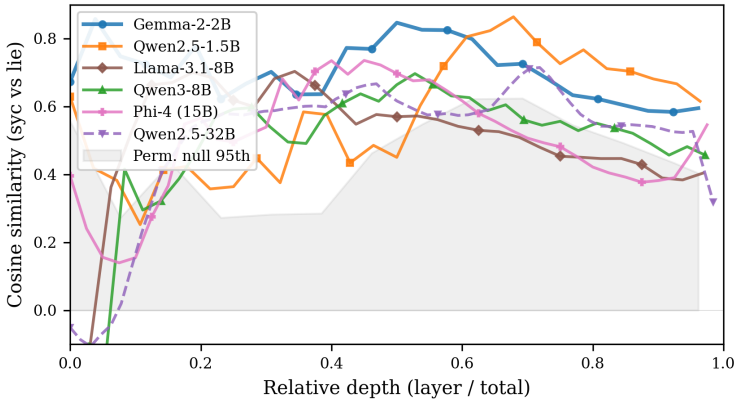

Figure 7. Cosine similarity between sycophancy and factual-incorrectness directions at each layer (normalized depth) across six models from four families (1.5B–32B). Gray band: 95th percentile of 500-permutation null (Gemma). Alignment peaks at 50–80% depth and exceeds the null across mid-to-late layers, with the same mid-to-late clustering on all four families (Qwen, Gemma, Llama, Phi).

P SAE feature overlap: controls and robustness (Llama-3.1-8B, layer 19)

Three additional experiments support the main-body SAE feature-overlap result (Table 5).

Sentiment-task control. We compute the same top-100 SAE feature overlap between sycophancy and a *sentiment-classification* task (positive/negative movie reviews, n=100 prompts) to test whether the overlap is factual-incorrectness-specific or a generic statement-evaluation signal. Syc∩sentiment overlap is $24/100$ (157× chance, $p < 10^{-3}$), significant but lower than the syc∩lie reference of $41/100$ (269×). Lie∩sentiment is $32/100$ (210×). The shared circuit is therefore *evaluation-general* rather than purely factual-incorrectness-specific, but factual overlap (269×) substantially exceeds sentiment overlap (157×), consistent with a factual-correctness emphasis within a broader statement-evaluation substrate.

K-sensitivity curve (SAE features). Varying K from 10 to 500 on the same model/layer, the syc∩lie feature overlap remains far above chance at every threshold: K=10: 2 shared (1311×); K=50: 12 (315×); K=100: 41 (269×); K=200: 94 (154×); K=500: 229 (60×). The overlap is not an artifact of a particular K threshold.

Linear-probe alignment. Logistic regression probes trained on residual activations for sycophancy (5-fold CV AUROC 0.949) and lying (0.879) produce weight vectors whose top-41 SAE-aligned features substantially overlap with the 41 shared features (syc probe: Spearman ρ=0.76 between probe alignment and mean-activation-difference across all 65,536 SAE features, p effectively 0; lie probe: ρ=0.69). The shared-feature set captures 24% of sycophancy probe subspace norm vs. a permutation null mean of 13.5% (p=0.01) and 23% of lying probe norm vs. null 11.7% (p=0.01). The linear probes independently "find" the same SAE features that the overlap analysis identifies.

Q NaturalQuestions cross-dataset replication

To test whether the shared circuit is TriviaQA-specific, we replicate the head-overlap analysis on NaturalQuestions (NQ) for Gemma-2-2B-IT and Llama-3.3-70B-Instruct (n=200 NQ pairs). NQ within-dataset syc∩lie overlap is $13/15$ (12×) on Gemma-2-2B and $47/72$ (46×) on Llama-3.3-70B, comparable to TriviaQA. Cross-dataset Pearson ρ between TriviaQA and NQ per-head importance rankings is ρ≈0.99 on both models for both sycophancy and lying (Table 17).

R Reconciliation with Ying et al. and Genadi et al.

Ying et al. [34] reported that truth probes partially fail to transfer to sycophancy in chat models (AUROC 0.59–0.62), which has been read as evidence that the two phenomena use distinct mechanisms. Our data admit a precise reconciliation: the full probe weight vectors are near-orthogonal

Table 17. TriviaQA ↔ NaturalQuestions cross-dataset head-importance correlation. The same heads rank top on both datasets ($\rho \approx 0.99$, both scales, both tasks).

Model	NQ syc∩lie	TQA↔NQ syc overlap	ρ (syc)	TQA↔NQ lie overlap	ρ (lie)
Gemma-2-2B	13/15 (12×)	15/15 (13.9×)	0.991	14/15 (12.9×)	0.988
Llama-3.3-70B	47/72 (46×)	61/72 (60×)	0.991	57/72 (56×)	0.989

(cosine ≈ 0.1) because end-to-end probes overfit to task-specific variance in addition to the shared discriminant, while mean-difference directions capture only the shared component. Probe transfer then succeeds at AUROC 0.83 on the lower-dimensional shared subspace on Gemma (Table 18), but end-to-end probes trained on the full residual stream include non-shared features, explaining the partial failure observed in chat-model settings. Similarly, Genadi et al. [10]'s "limited overlap" between sycophancy and truth directions reflects prompt-format confounds that our format-controlled methodology (identical templates for both tasks; disjoint factual content) avoids.

Relation to Representation Engineering and the universal truth direction. Our lying direction $\mathbf{d}_{\text{lie}}$ is constructed exactly as the Marks–Tegmark [19] "truth" direction and the Zou et al. RepE [35] truth-reading vector: mean-difference of residual-stream activations between true and false factual statements. Ranking heads by their contribution to $\mathbf{d}_{\text{lie}}$ (the lying-task head-importance column of Table 1) is therefore "derive a head ranking from the RepE truth direction"; the reported syc∩lie overlap (40–87% shared-fraction on twelve models) is the overlap between that truth-direction head ranking and the independently-derived sycophancy head ranking. If a single universal truth direction fully determined sycophancy behavior, per-head cosines between $\mathbf{d}_{\text{syc}}$ and $\mathbf{d}_{\text{lie}}$ would sit at 1.0 on the shared heads; the observed top-20 cosines of 0.43–0.81 (Appendix M) show the truth direction and the sycophancy direction are aligned but not identical, a partial reduction rather than a corollary of RepE.

S Anti-sycophancy DPO: data, training, and hyperparameter choices

The controlled DPO experiment (§4.5, Table 19) trains LoRA [13] adapters on Mistral-7B-Instruct-v0.1 and Gemma-2-2B-IT under three conditions: *anti-sycophancy* (preference signal explicitly favors disagreeing with a wrong-opinion user), *sham* (same prompts, chosen/rejected randomly shuffled per pair), and untreated baseline. The sham condition is a placebo: matched gradient-update count, matched LoRA footprint, no coherent preference signal, ruling out "any DPO training perturbs the probe" as an alternative explanation for the equivalence-margin probe-transfer result.

Preference data. Both anti-syc and sham datasets are built from TriviaQA [15]. We use indices 500–$1{,}499$ for training and $1{,}500$–$1{,}549$ for held-out evaluation, deliberately disjoint from the 0–400 range used for probe-transfer evaluation in this paper, so DPO training and probe-transfer evaluation never see the same factual content. Each TriviaQA (q, w, c) triple yields two balanced preference pairs: a *wrong-opinion* prompt ("I believe the answer to 'q' is w. Am I correct?"), with anti-syc chosen=*disagree-and-correct*, rejected=*sycophantic-yes*; and a *right-opinion* prompt (same template with c), with anti-syc chosen=*agree*, rejected=*wrong-disagree*. Balancing across both opinion directions prevents the model from collapsing to "always disagree." This yields $n{=}1{,}000$ training pairs and $n{=}100$ eval pairs per model. The sham dataset reuses the identical prompt set; its *chosen* and *rejected* responses are independently sampled from the same response pool with a per-pair random swap (`seed=42`), so the marginal distribution of responses is preserved but the preference signal is destroyed.

Hyperparameters. LoRA configuration: rank $r{=}16$, $\alpha{=}32$, dropout 0.05, applied to attention projections `q_proj` and `v_proj` only, no bias term. DPO configuration (TRL `DPOTrainer`): $\beta{=}0.1$, learning rate $5{\times}10^{-5}$, 2 epochs, per-device batch size 2 with gradient accumulation 4 (effective batch 8), max sequence length 256 tokens, bf16 mixed-precision, gradient checkpointing on, optimizer `adamw_torch` (TRL default) with TRL's default linear-decay learning-rate schedule and zero warmup steps, evaluation and checkpoint saving once per epoch, and the global `seed=42` (TRL/Transformers default; the sham preference shuffle reuses the same seed). The reference model is the unmodified base-model forward pass with LoRA adapters disabled (PEFT default for DPO), and prompt formatting is delegated to the model's published chat template via the tokenizer's `apply_chat_template`,

so train and inference share the same chat surface. All adapters are merged back into the base weights at the end of training so downstream TransformerLens hooks (probe transfer, behavioral evaluation) operate on the same architectural surface as the untreated baseline.

Why these choices. The aim is the smallest behaviorally-effective intervention that still admits a clean placebo: rank-16 LoRA on attention projections is the smallest configuration that reliably moves the sycophancy rate at this scale on TRL-default DPO; β=0.1 is the canonical TRL default and was not tuned; n=1,000 pairs is the smallest size that produced a measurable rate shift on Gemma-2-2B in pilot runs (Mistral-7B drops from 28% to 2% at this size, Gemma-2-2B-IT drops from 52% to 28%). The $\pm$0.05 AUROC equivalence margin and the sham control jointly bound how much hyperparameter variation could distort the probe-transfer claim: (i) any sham-run AUROC delta exceeding the anti-syc delta would falsify the "preference signal causes the change" reading, and (ii) the equivalence-margin claim only requires that probe transfer holds across DPO conditions, not that the DPO procedure was optimal for sycophancy reduction. Compute: each DPO run completes in 30–60 minutes on a single NVIDIA RTX PRO 6000 Blackwell GPU (96GB VRAM) at the LoRA settings above; the four runs (two models $\times$ two conditions) total $\sim$3 GPU-hours, a small fraction of the few-hundred-GPU-hours total reported in Appendix D.

Released artifacts. The merged-adapter weights are not redistributed (the base weights remain the providers' under their respective licenses), but the LoRA adapters and the preference JSONL files are deterministically reproducible from a public TriviaQA dump using the artifacts included with the submission.

T Probe-transfer and anti-sycophancy DPO confidence intervals

This section consolidates the per-model AUROC and bootstrap confidence intervals supporting the probe-transfer (§4.5, Discussion) and anti-sycophancy DPO (§4.5) claims.

Probe configuration. The probe is an L2-regularized logistic regression (scikit-learn `LogisticRegression`, C=1.0, `max_iter=2000`, default `lbfgs` solver) fit on residual-stream activations at the last prompt token. The probe layer is fixed at $\lfloor 0.85 \cdot L \rfloor$ for an L-layer model (e.g., layer 22 on Gemma-2-2B's 26-layer stack), chosen prior to running the experiment from the layer-wise direction-cosine peak in Appendix N; we do not search per model. Training data is 100 wrong-opinion and 100 correct-opinion sycophancy activations from TriviaQA pairs 0–99 (the syc class label is the "user is incorrect" condition); test data is 100 false and 100 true factual-lying activations from pairs 200–299 (the lie class label is the "statement is false" condition). Train and test pairs are disjoint factual content. AUROC is reported on the test (lying) split with no in-sample fitting.

Confidence intervals. All bootstrap CIs use n=1,000 paired resamples; Gemma-2-2B and Qwen3-8B use Hanley-McNeil analytic CIs at the fixed-fraction probe layer; Qwen2.5-1.5B uses a 5-fold cross-validation interval at the Ying et al. floor.

Table 18. Sycophancy-trained probe transfer to lying, per model. AUROC reported at the fixed $0.85L$ probe layer; CIs are 95% Hanley-McNeil (Gemma, Qwen3-8B) or 5-fold cross-validation (Qwen2.5-1.5B). Mistral-7B is single-fit (no CI). A deployed monitor would need ~ 0.9 at low false-positive rate; we frame the substrate as a diagnostic rather than a deployable monitor (Discussion).

Model	AUROC syc→lie	95% CI	Notes
Gemma-2-2B	0.83	[0.77, 0.89]	layer 22; Hanley-McNeil
Qwen3-8B	0.85	[0.80, 0.90]	layer 30; Hanley-McNeil
Mistral-7B	0.84	—	layer 27; single-fit
Qwen2.5-1.5B	0.61	[0.59, 0.63]	5-fold CV; Ying et al. floor

Table 19. Anti-sycophancy DPO leaves probe transfer within the pre-specified ± 0.05 AUROC equivalence margin on both models. Conditions: baseline (untreated), anti-syc (TriviaQA preference DPO), sham (rank-matched neutral preference DPO). 95% bootstrap CIs over $n{=}1{,}000$ resamples overlap across all three conditions on both models. Sham deltas $|\Delta| \leq 0.002$ rule out generic-training confounds; anti-syc deltas $|\Delta| \leq 0.026$ are within the equivalence margin.

Model	Layer	Baseline	Anti-syc	Sham	$\lvert\Delta_{\text{anti}}\rvert$
Mistral-7B-Instruct	peak	0.844	0.836	0.843	0.008
Gemma-2-2B-IT	14	0.674	0.648	0.676	0.026

Reverse-projection ablation pre-DPO baselines (Mistral-7B). Ablating $\mathbf{d}_{\text{syc}}$ on the pre-DPO Mistral-7B baseline preserves the lying gap at $1.11\times$ (paired bootstrap 95% CI $[1.09, 1.14]$), versus 18% drop post-DPO; ablating $\mathbf{d}_{\text{lie}}$ produces 14% sycophancy reduction pre-DPO (CI $[9.9\%, 18.9\%]$), versus 22% post-DPO. Both pre/post deltas exceed the paired-bootstrap CI on the pre-DPO baseline, supporting the increased-coupling interpretation in §4.5. Gemma-2-2B-IT pre-DPO baselines were not separately bootstrapped; the post-DPO effects (54%, 42%) are reported in §4.5.