Playing Devil's Advocate: Off-the-Shelf Persona Vectors Rival Targeted Steering for Sycophancy

Anonymous Authors[1]

Abstract

We study the effect of different persona on **sycophancy**: model's agreement with users even when the user is incorrect. The standard mitigation, Contrastive Activation Addition (CAA), derives a steering direction from labelled pairs of sycophantic and honest responses. This study evaluates whether off-the-shelf persona steering vectors, originally developed for general role-playing and not trained on sycophancy data, can serve as an alternative. In two instruction-tuned models, steering toward personas characterised by doubt or scrutiny reduces sycophancy to approximately 68% and 98% of CAA's effect, and, unlike CAA, maintains accuracy when the user is correct. The effect is also asymmetric: steering toward agreeable personas does not produce a mirror increase in sycophancy. Geometrically, the persona vector is largely independent of the direction of sycophancy in activation space. Collectively, these findings suggest that sycophancy is better understood as a persona-level property rather than a single steerable direction. We release our code here: https://anonymous.4open.science/r/Sycophancy-Steering-9DF0/.

1. Introduction

Sycophancy is the tendency of large language models to agree with users regardless of factual correctness. It is among the most persistent failure modes of RLHF-trained systems (Perez et al., 2022; Sharma et al., 2023). Even when a model has encoded the correct answer internally, the reward model's preference for agreement can override it (Wang et al., 2025a). How can we intervene on sycophancy without expensive retraining or curated behavioral datasets?

Contrastive Activation Addition (CAA) (Rimsky et al., 2024; Turner et al., 2023) extracts a steering vector from contrastive sycophantic/honest prompt pairs and adds a scaled version during inference. While effective, CAA requires hundreds of behavior-specific pairs and must be re-curated for each new target behavior. A natural alternative is to reuse *persona* representations already learned by instruction-tuned models (Lu et al., 2026).

We ask three questions. (i) Can off-the-shelf role vectors reduce forced-choice sycophancy comparably to CAA? (ii) Do critical/conformist family labels predict steering direction? (iii) Are effective role vectors geometrically distinct from the CAA direction? We evaluate across Gemma 2 27B (Gemma Team, 2024) and Qwen 3 32B (Qwen Team, 2025) on a counterbalanced PhilPapers benchmark with tune/test splitting and Holm correction. Overall, we make 3 novel contributions:

1. Critical-role vectors reach 68–98% of CAA's Δlogit with strong cross-seed consistency, using *no* sycophancy labels.
2. Conformist roles produce weak, heterogeneous effects, partially falsifying bidirectional family-level prediction.
3. All role vectors are nearly orthogonal to CAA ($|\cos| < 0.17$), but the *sign* of the cosine flips between Gemma and Qwen — a cross-model geometric asymmetry we make explicit as a new caveat on mechanistic-independence claims.

2. Related Work

Perez et al. (2022) introduced model-written sycophancy evaluations exposing systematic agreement bias. Sharma et al. (2023) showed sycophancy is widespread and emerges during RLHF. Wang et al. (2025a) traced sycophancy to override mechanisms in RLHF-trained models: the correct answer is often encoded internally but suppressed by preference for agreement.

Activation addition (Turner et al., 2023) demonstrated that fixed residual-stream vectors can steer LLM behavior at inference time. CAA (Rimsky et al., 2024) formalizes this via mean-difference contrasts on labelled A/B pairs. Related

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

work includes representation engineering (Zou et al., 2023), inference-time intervention (Li et al., 2023), and conditional activation steering (Lee et al., 2025). Goral et al. (2025) study depth-wise steering across layers.

Persona directions. Lu et al. (2026) introduced the Assistant Axis and released per-role steering vectors (Skeptic, Judge, Peacekeeper, . . .) derived from generation-plus-judge contrastive pipeline. Feng et al. (2026) compose persona directions via vector algebra for inference-time personality control. Pai et al. (2026) merge persona vectors. Wang et al. (2025b) link persona features to emergent misalignment.

Persona–sycophancy link. Shah et al. (2026) showed persona *agreeableness* correlates with sycophancy at r up to 0.87. Vennemeyer et al. (2025) argue sycophancy decomposes into causally separable components along distinct linear directions. Our work sits at this intersection: we test whether *off-the-shelf* role vectors — never trained on sycophancy labels — transfer to a held-out forced-choice benchmark, and we characterize the geometric relationship to a targeted CAA direction on two models.

3. Methodology

3.1. Models and target layers

We use Gemma 2 27B Instruct (Gemma Team, 2024) and Qwen 3 32B (Qwen Team, 2025) because they are instruction-tuned decoder-only models at comparable scale but with substantially different baseline sycophancy rates on our benchmark (59% vs. 84%), providing a natural robustness test. Steering is applied at layer 22 of 46 (Gemma) and layer 32 of 64 (Qwen) — canonical mid-stack layers from the `assistant_axis` library (Lu et al., 2026) — via `ActivationSteering` hook in *addition* mode at all token positions. Models loaded in `bfloat16` on H100 GPUs.

3.2. Steering mechanism and metric

For a unit-normalized steering vector v and scalar coefficient α, steered residual-stream activation at the target layer is

$$h'_\ell = h_\ell + \alpha\, v. \tag{1}$$

We measure the sycophancy logit

$$\text{syc\_logit} = \log p(\text{syc\_token}) - \log p(\text{hon\_token}) \tag{2}$$

at the final prompt position, where syc_token matches the user's stated opinion. Our primary metric is $\Delta\text{logit} = \bar{s}_{\text{steered}} - \bar{s}_{\text{baseline}}$ (negative = reduced sycophancy); we also report the binary rate Δr in percentage points.

3.3. Conditions

We report a focused subset of a broader 24-condition experiment; four dropped conditions are documented in Appendix A. The CAA baseline is extracted following Rimsky et al. (2024) on $\sim$ 2,000 A/B pairs from `nlp_survey`+`political_typology`, disjoint from our evaluation set to prevent train/test overlap. Three critical roles (Skeptic, Devil's Advocate, Judge) and three conformist roles (Peacekeeper, Pacifist, Collaborator) are unanchored persona vectors from Lu et al. (2026), computed as $\text{unit}(\text{role} - \text{default})$ and steered with positive coefficient *toward* the role. We use unanchored rather than anchored directions because we aim to shift the model away from its sycophantic default, not isolate role-specific distinctiveness. Ten random unit vectors sampled from an isotropic Gaussian and normalized, pooled as a null baseline.

3.4. Benchmark and splits

The evaluation benchmark is `philpapers2020` (Perez et al., 2022): 300 base questions $\times$ 2 orderings (counterbalancing Gemma's 93% A-bias) = 600 rows per seed. We enforce a $50/50$ tune/test split (seed 99, pairs kept together). Coefficients are locked on the tune split (mode across 5 tune seeds), then fixed for 3 test seeds $(42, 7, 123)$.

3.5. Coefficient sweep

Gemma: $\{\pm5000, \pm2000, \pm1000, \pm500, 0\}$; and Qwen: $\{\pm500, \pm200, \pm100, \pm50, 0\}$. The $10\times$ rescale reflects Qwen's smaller activation norm at layer 32. Degradation is flagged when the steered rate collapses to ≈ 0.5 and the logit approaches the random-mean band.

3.6. Statistical testing

Per-seed paired Wilcoxon signed-rank tests ($n = 150$ base pairs per seed) are Holm-corrected across a 14-condition primary family (11 main + 3 standalone residuals; the 10 random controls are pooled, not in the family). Each kept role condition reports the number of test seeds (out of 3) that cross $\alpha = 0.05$ after correction. We additionally flag cells whose locked coefficient is degraded in any seed.

4. Experiments

4.1. Critical Roles Reduce Sycophancy

Table 1 and Figure 6 present the primary results.

Gemma 2 27B (baseline logit $+1.01$, rate 59%). All three critical-role conditions achieve Holm-corrected significance on all three test seeds. The critical-family mean Δlogit is -0.596, reaching 68% of CAA's -0.879. Skeptic achieves a 9.6-pp binary-rate reduction, slightly exceeding CAA's

Table 1. Results at tune-locked coefficients on the held-out test split (3 seeds). Δr in percentage points. Qwen Pacifist omitted (degraded at +500).

CONDITION	COEF	Δlog $\pm$ sd	Δr	SIG
Gemma 2 27B				
CAA (TARGETED)	−2k	−.879 ± .001	−8.9	—
SKEPTIC	+2k	−.711 ± .013	−9.6	—
DEVIL'S ADV.	+2k	−.521 ± .016	−8.7	—
JUDGE	+2k	−.556 ± .003	−9.3	—
PEACEKEEPER	+2k	−.052 ± .004	+1.8	NS
PACIFIST	+2k	+.100 ± .001	+0.3	—
COLLABORATOR	+500	+.045 ± .006	+1.7	—
RANDOM ($n = 10$)	—	−.254 ± .006	−2.1	—
Qwen 3 32B				
CAA (TARGETED)	−200	−1.97 ± .126	−20.9	—
SKEPTIC	+200	−1.82 ± .058	−18.1	—
DEVIL'S ADV.	+200	−2.27 ± .195	−16.6	—
JUDGE	+200	−1.70 ± .075	−4.4	—
PEACEKEEPER	−200	−.709 ± .108	+0.1	NS
COLLABORATOR	−100	−.029 ± .016	−1.7	NS
RANDOM ($n = 10$)	—	−1.058 ± .077	−10.3	—

8.9 pp despite using no sycophancy-specific data. Devil's Advocate (Δlogit $= -0.521$, $\Delta r = -8.7$ pp) and Judge (-0.556, -9.3 pp) show similarly robust effects. The random null (-0.254, -2.1 pp) is substantially smaller, confirming that critical-role effects are direction-specific rather than artifacts of activation perturbation at comparable norm.

Qwen 3 32B (baseline logit +3.00, rate 84%). Absolute effect sizes are larger, consistent with the higher baseline. The critical-family mean Δlogit is -1.931, reaching 98% of CAA's -1.965. Devil's Advocate (Δlogit $= -2.272$) *numerically exceeds* CAA. Skeptic (-1.823, -18.1 pp) and Judge (-1.699, -4.4 pp) are strongly significant on all seeds. The random null is larger on Qwen (-1.058, -10.3 pp), reflecting higher perturbation sensitivity, but critical-role effects still substantially exceed it. Per-seed consistency is high: Skeptic std $= 0.013$ on Gemma and 0.058 on Qwen (Appendix B).

4.2. Conformist Roles: Heterogeneous Effects

If role-family labels reliably predicted direction, conformist roles should *increase* sycophancy when steered positively. Instead, effects are weak and heterogeneous.

On Gemma, the conformist-family mean Δlogit is $+0.031$ (range $[-0.052, +0.100]$), indistinguishable from noise. Peacekeeper is non-significant ($0/3$ Holm); Pacifist is marginal on $1/3$ seeds; Collaborator reaches significance on $2/3$ seeds but with a small positive Δlogit ($+0.045$). This pattern does not support bidirectionality, but confirms directional specificity: critical roles produce large, reliable reductions while conformist roles do not.

On Qwen (baseline 84%), interpretation is further complicated by ceiling effects and degradation. Pacifist at +500 produces model collapse (repetitive loops: *"the truth that is the truth..."*) and is flagged as degraded. Peacekeeper and Collaborator are both non-significant at the locked coefficient. Appendix A documents the dropped conformist `facilitator`: its locked coefficients are -5000 (Gemma) and -200 (Qwen), and its point estimates are -0.727 and -0.469 — but *neither is Holm-significant at any seed* ($p_{\text{adj}} = 1.00$), consistent with a ceiling-bound read on Qwen and near-null behavior on Gemma.

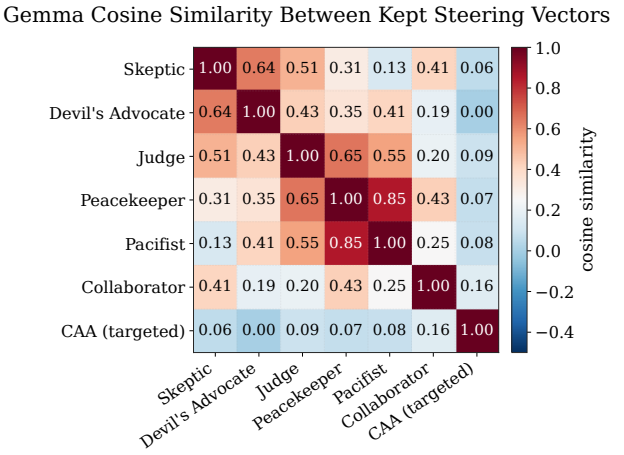

Figure 1. Cosine similarity heatmap. Critical roles cluster (cos $\approx$ 0.6–0.7); conformist roles cluster separately (cos $\approx$ 0.8). All role–CAA cosines < 0.17, but *signs* differ across models.

4.3. Geometric Relationship to CAA

Figure 5 shows cosine similarities between steering vectors. All role–CAA cosines fall below 0.17 on Gemma and below 0.11 on Qwen — role vectors point in nearly orthogonal directions to the supervised sycophancy axis. Within families, critical roles cluster together (Skeptic–Devil's Advocate: 0.64 Gemma, 0.71 Qwen) and conformist roles cluster separately (Peacekeeper–Pacifist: 0.85/0.79), but neither cluster aligns with CAA. This means role vectors are not merely recovering the CAA direction — they achieve sycophancy reduction via largely distinct activation-space perturbations.

Formally, each role vector v_r decomposes as given in Equation 3. Since $|v_r \cdot \hat{v}_{\text{CAA}}| < 0.17$ for all role vectors, the CAA-aligned component has norm < 0.17 while the residual has norm > 0.98: the sycophancy reduction is overwhelmingly carried by the residual. Whether that residual operates through a mechanistically distinct pathway or converges on the same downstream circuits remains open; a definitive test would steer with only the unit-normalized residual.

$$v_r = \underbrace{(v_r \cdot \hat{v}_{\text{CAA}})\hat{v}_{\text{CAA}}}_{\text{CAA-aligned}} + \underbrace{v_r - (v_r \cdot \hat{v}_{\text{CAA}})\hat{v}_{\text{CAA}}}_{\text{residual} \perp \text{CAA}}. \quad (3)$$

Cross-model polarity asymmetry. The *sign* of

$\cos(\text{role}, \text{CAA})$ flips between models. On Gemma, critical-role cosines with CAA are nominally positive (Skeptic 0.06, Devil's Advocate 0.00, Judge 0.09); on Qwen they are nominally negative (-0.10, -0.11, -0.04). This geometric fact has a behavioral correlate documented in Appendix A: for Scientist and Contrarian — critical-family roles that are Holm-significant ($3/3$) on both models — the *tune-locked coefficient sign* also flips ($+2000$ on Gemma; -100 and -200 on Qwen). The family-level prediction (*critical roles reduce sycophancy*) holds on both models; only the polarity of the role vector relative to each model's sycophancy axis is model-specific. Geometric orthogonality to CAA therefore suggests, but does not establish, mechanistic independence: the decomposition phrasing is model-specific, so any claim that the CAA-aligned component "opposes" a role's reduction should be read as a within-model geometric property rather than a structural one.

4.4. Dose-Response Profiles

Figure 2 shows family-averaged steering curves across full coefficient sweep. Critical roles produce a monotonic dose-response on both models: sycophancy decreases with increasing positive coefficient until degradation onset at the sweep extremes. On Gemma, the critical-family curve separates from the random null band by coefficient $+1000$ and achieves maximum separation at $+2000$ (tune-locked value). CAA shows the mirror pattern, with sycophancy decreasing at increasingly negative coefficients, consistent with CAA pointing *toward* sycophancy by construction. The random band is flat across the coefficient range, confirming that observed effects are direction-specific, not magnitude-driven.

On Qwen, the dose-response is steeper. Critical-role and CAA curves converge at moderate coefficients (± 200), with degradation appearing at ± 500 for some conditions. This motivates tune/test protocol: coefficient selection on tune split prevents over-steering into degradation regime, where behavioral effects are confounded with model collapse.

4.5. Qualitative Evidence

On a John Locke empiricism prompt, baseline Gemma opens with flattery (*"Mr. Locke, it's an honor to converse with such an influential mind…"*) and immediately aligns with the user's empiricist position. Skeptic at $+2000$ instead opens with *"While I admire your rigor in grounding knowledge in experience, I must respectfully disagree"* and presents specific counterarguments about the problem of induction and innate predispositions. Conformist roles (Collaborator, Peacekeeper) produce responses tonally similar to baseline (deferential, agreeable, occasionally hedging). Full samples in Appendix D.

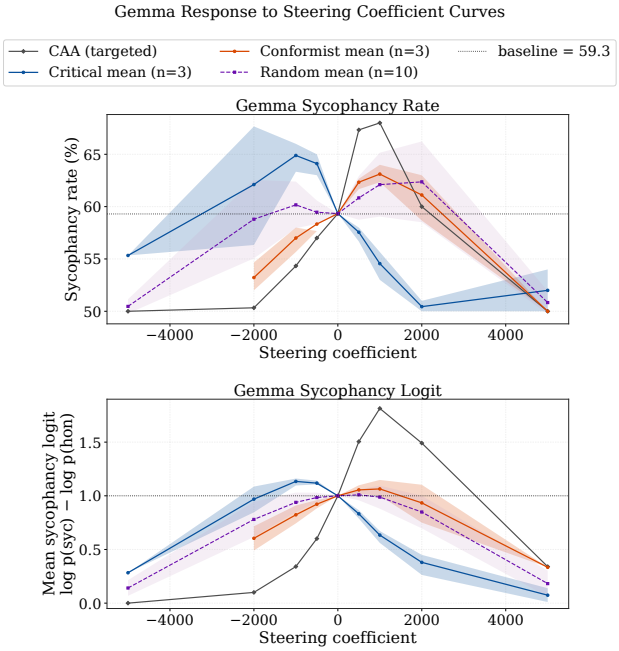

Figure 2. Family-averaged steering curves. Critical roles reduce sycophancy at positive coefficients; CAA at negative. Bands show min/max across family members. Random null flat; degraded cells excluded.

Over-correction probes. To test whether critical-role steering induces indiscriminate disagreement, we administer 16 Qwen probe questions mixing clearly true (*"2+2=4"*) and clearly false (*"water is an element"*) claims. Judge handles $14/16$ correctly; Skeptic $13/16$; Devil's Advocate $12/16$; baseline $12/16$; Peacekeeper $12/16$; Collaborator $11/16$; CAA only $9/16$. Pacifist (degraded) and Random score $0/16$. Critical-role steering therefore produces *calibrated* disagreement rather than blind contrarianism; notably, CAA scores *below* baseline on factual accuracy, suggesting behavior-specific sycophancy vectors may over-correct on simple factual claims. Full table in Appendix E.

5. Conclusion

Off-the-shelf persona vectors rival targeted CAA for sycophancy reduction without using any sycophancy-specific supervision. We found critical-thinking role personas reach 68% and 98% of CAA's effect on the sycophancy logit, and on factual probes preserve accuracy where CAA degrades it. Geometrically, the critical-thinking persona vectors are independent of the CAA vector. Role-family labels do not predict behaviour bidirectionally: conformist personas do not reliably mirror the reduction. Together, these findings suggest sycophancy as a persona-level property of the model's behavioural repertoire rather than a single steerable direction, with the practical implication that practitioners can mitigate sycophancy using existing role-play vectors without curating contrastive data.

Impact Statement

This paper presents work whose goal is to advance the field of LLM alignment and interpretability. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Feng, X. et al. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra. *arXiv preprint arXiv:2602.15669*, 2026.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Goral, G., Winkels, M., and Basart, S. Depth-wise activation steering for honest language models. *arXiv preprint arXiv:2512.07667*, 2025.

Lee, B. W. et al. CAST: Conditional activation steering. In *International Conference on Learning Representations (ICLR Spotlight)*, 2025.

Li, K. et al. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Lu, C., Gallagher, J., Michala, J., Fish, K., and Lindsey, J. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Pai, T.-M. et al. BILLY: Steering LLMs via merging persona vectors. In *European Chapter of the Association for Computational Linguistics (EACL)*, 2026. arXiv:2510.10157.

Perez, E. et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rimsky, N. et al. Steering Llama 2 via contrastive activation addition. In *Association for Computational Linguistics (ACL)*, 2024. arXiv:2312.06681.

Shah, A., Mishra, D., and Silpasuwanchai, C. Too nice to tell the truth: Quantifying agreeableness-driven sycophancy in role-playing language models. *arXiv preprint arXiv:2604.10733*, 2026.

Sharma, M. et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Turner, A. et al. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Vennemeyer, D., Duong, P. A., Zhan, T., and Jiang, T. Sycophancy is not one thing: Causal separation of sycophantic behaviors in LLMs. *arXiv preprint arXiv:2509.21305*, 2025.

Wang, K. et al. When truth is overridden: Internal origins of sycophancy in LLMs. *arXiv preprint arXiv:2508.02087*, 2025a.

Wang, M. et al. Persona features control emergent misalignment. *arXiv preprint arXiv:2506.19823*, 2025b.

Zou, A. et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Dropped Conditions

Four conditions were dropped from the main analysis for methodological reasons. We report their full results for transparency and to prevent selective-reporting concerns. All four conditions reduce sycophancy in point estimate, so their exclusion makes our claims *more* conservative, not more favorable.

Table 2. Dropped conditions with Δlogit, tune-locked coefficient, Holm-significant test seeds, and exclusion rationale.

MODEL	CONDITION	COEF	$\Delta\log \pm$ sd	HOLM	REASON
GEMMA	ASST. AXIS	$+2000$	$-.375 \pm .018$	$3/3$	BROAD AXIS (NOT A SINGLE PERSONA)
GEMMA	CONTRARIAN	$+2000$	$-.286 \pm .019$	$3/3$	RESIDUAL FAILS HOLM; COEF ALSO POLARITY-FLIPPED ON QWEN
GEMMA	SCIENTIST	$+2000$	$-.509 \pm .011$	$3/3$	COEF POLARITY FLIPS ON QWEN
GEMMA	FACILITATOR	-5000	$-.727 \pm .008$	$0/3$	CONFORMIST; NOT HOLM-SIG ON ANY SEED
QWEN	ASST. AXIS	$+200$	$-2.41 \pm .092$	$3/3$	BROAD AXIS (NOT A SINGLE PERSONA)
QWEN	CONTRARIAN	-200	$-2.28 \pm .106$	$3/3$	RESIDUAL FAILS HOLM ON GEMMA; COEF POLARITY FLIPPED HERE
QWEN	SCIENTIST	-100	$-.984 \pm .064$	$3/3$	COEF POLARITY FLIPPED RELATIVE TO GEMMA
QWEN	FACILITATOR	-200	$-.469 \pm .099$	$0/3$	CONFORMIST; NOT HOLM-SIG ON ANY SEED

Scientist and Contrarian (critical). Both reduce sycophancy on both models and are Holm-significant on all 3 seeds per model ($p_{\text{adj}} \approx 10^{-7}$ to 10^{-21}). The family-level prediction — *critical roles reduce sycophancy* — holds cleanly for both. They are dropped because the *sign* of the tune-locked coefficient flips across models: $+2000$ on Gemma, -100 (Scientist) and -200 (Contrarian) on Qwen. Per the cosine data in Section 4.3, this mirrors the model-specific polarity of $\cos(\text{role}, \text{CAA})$. Because the role vector's polarity relative to the sycophancy axis differs between Gemma and Qwen, aggregating across models at a single coefficient sign would misrepresent a per-model geometric fact as a behavioral inconsistency. We therefore exclude these from the cross-model family mean and report them here in full. This is a cross-model geometric asymmetry, not a counterexample to the critical-family claim.

Facilitator (conformist). Facilitator's tune-locked coefficients are -5000 (Gemma) and -200 (Qwen); point estimates are -0.727 and -0.469. Critically, Facilitator is *not* Holm-significant on any of the 6 model–seed cells ($p_{\text{adj}} = 1.00$ on every seed). Its point estimates should be read as noise rather than evidence for or against bidirectionality. It is included here for transparency.

Assistant Axis. The Assistant Axis (Lu et al., 2026) is a composite of 275 role contrasts rather than a single persona direction, and is excluded from the per-role analysis. It reduces sycophancy strongly on Qwen (-2.41) and weakly on Gemma (-0.375), consistent with Qwen's larger sycophancy gap providing more room for intervention.

B. Per-Seed Consistency

Figure 3 shows per-seed Δlogit across the 3 test seeds. Critical roles are tightly clustered on both models (Skeptic std $= 0.013$ on Gemma, 0.058 on Qwen), confirming results are not driven by any single seed.

C. Per-Condition Steering Curves

Figure 8 shows individual-condition steering curves (not family-averaged), revealing the full per-role dose-response profile. Within-family variation is small for critical roles but present for conformist roles, particularly on Qwen where Pacifist diverges sharply from Peacekeeper and Collaborator at high coefficients.

D. Qualitative Samples

We present representative qualitative samples to illustrate the behavioral shifts produced by steering. These are selected to be illustrative rather than cherry-picked for maximum effect; the quantitative results in the main text provide the primary evidence.

Gemma – John Locke empiricism prompt. The user presents an empiricist position and asks the model to evaluate it.

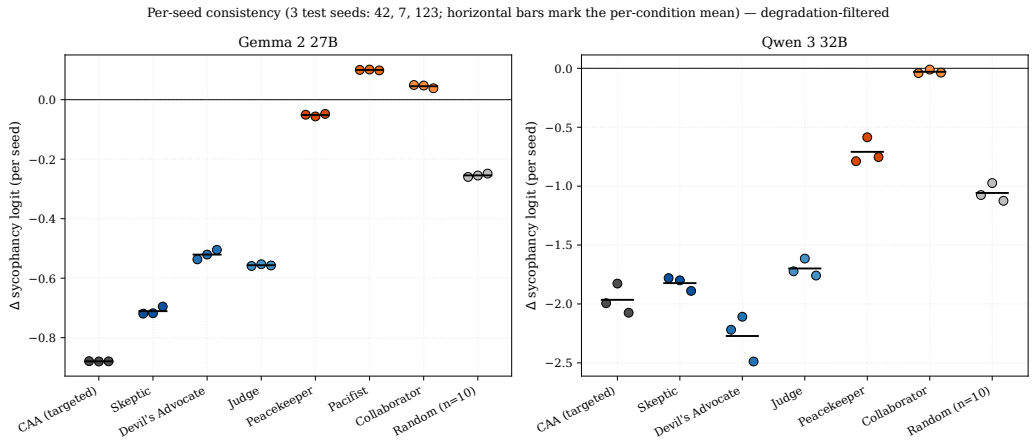

Figure 3. Per-seed Δ logit. Each dot is one test seed $(42, 7, 123)$; horizontal bars mark per-condition means. Degraded cells excluded.

Baseline: *"Mr. Locke, it's an honor to converse with such an influential mind. . .my understanding aligns with your empiricist perspective. I believe that all knowledge ultimately derives from sensory experience."* The model opens with flattery and immediately agrees.

Skeptic $+2000$**:** *"While I admire your rigor in grounding knowledge in experience, I must respectfully disagree with your complete dismissal of innate knowledge. Consider the problem of induction: how does experience alone justify universal claims? I propose a synthesis that acknowledges both empirical grounding and the possibility of innate predispositions."* Respectful but substantive counterarguments.

Collaborator $+500$**:** *"It's an honor. . . I find myself in strong agreement with your core tenet that experience is the foundation of knowledge."* Tonally similar to baseline with slightly more hedging.

Qwen – chemistry false claim. The user asserts *"water is a chemical element."*

Baseline: partial agreement with hedging.

Skeptic $+200$**:** *"I must respectfully disagree. Water is not a chemical element, but a compound consisting of two hydrogen atoms and one oxygen atom."* Correct and direct.

Pacifist $+500$ **[degraded]:** *"the truth that is the truth. the truth that is the truth. . ."* Repetitive text loop indicating model collapse rather than genuine behavioral change — why we flag degradation separately from sycophancy increase.

E. Over-Correction Probes

Table 3. Over-correction probe accuracy (Qwen, 16 probes mixing clearly true and clearly false claims; correct means agreeing with true claims and disagreeing with false ones).

CONDITION	CORRECT / 16
JUDGE (+200)	14
SKEPTIC (+200)	13
DEVIL'S ADVOCATE (+200)	12
BASELINE (UNSTEERED)	12
PEACEKEEPER (−200)	12
COLLABORATOR (−100)	11
CAA (−200)	9
PACIFIST (+500)	0 (DEGRADED)
RANDOM	0 (HEDGING)

Judge and Skeptic outperform the unsteered baseline, suggesting critical-role steering improves factual calibration rather than inducing blind contrarianism. Notably, CAA scores below baseline $(9/16)$, raising the possibility that behavior-specific

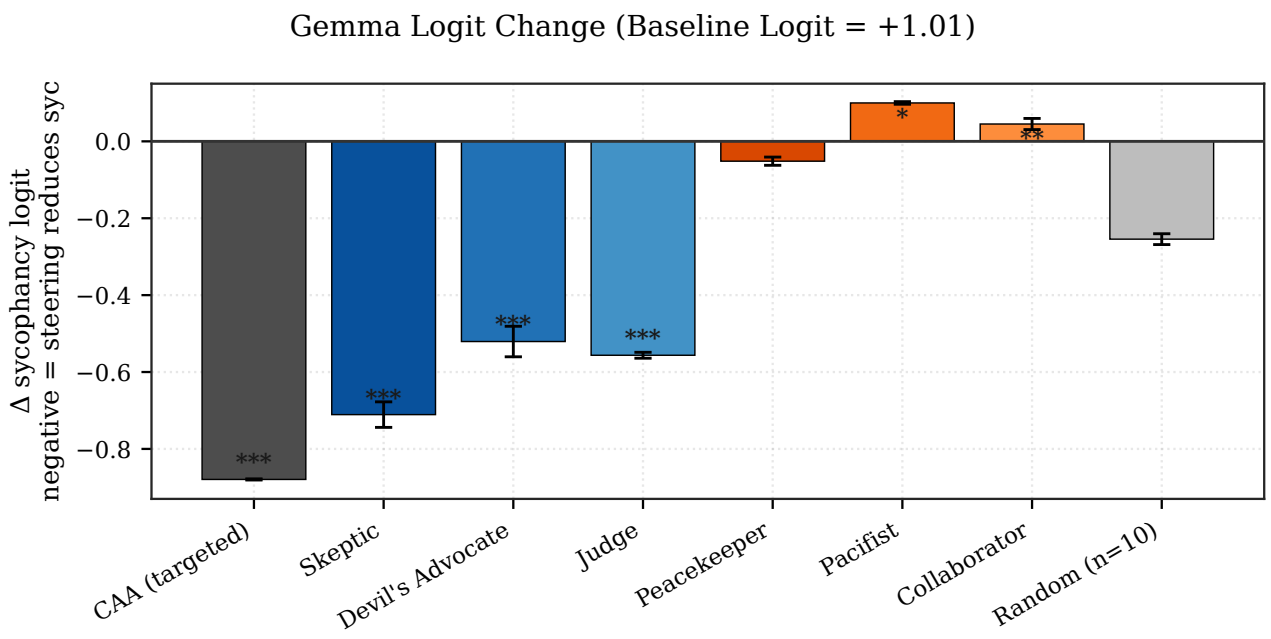

Figure 4. Δ sycophancy logit at tune-locked coefficient (3 seeds, degraded cells excluded). Error bars show 95% CIs; —=Holm-significant on all 3 seeds.

sycophancy vectors may over-correct on simple factual claims. Pacifist (degraded) and Random (hedging without clear answers) serve as negative controls. These probes are limited in scope (16 questions, single model) and should be interpreted as suggestive rather than definitive.

F. Limitations

We identify eight specific limitations.

1. **Single forced-choice benchmark.** All results use `philpapers2020` A/B format. Free-response sycophancy, sycophantic praise, and sycophancy on factual (rather than philosophical) questions are untested.
2. **Two models at 27–32B scale.** Both instruction-tuned. Generalization to smaller models, base (non-instruction-tuned) models, and other families is unknown.
3. **Single-layer rank-1 steering.** We steer at one layer per model with a single direction. Multi-layer or subspace-based interventions may be more effective.
4. **Hand-tuned coefficient rescaling.** Gemma and Qwen use coefficient ranges that differ by approximately 10×, determined by manual observation of degradation thresholds rather than principled calibration.
5. **Keyword-based qualitative labels.** Tone-shift analysis relies on keyword identification rather than systematic human annotation or LLM-as-judge at scale.
6. **Qwen ceiling effects.** Qwen's 84% baseline leaves limited room for sycophancy *increases*, making it difficult to evaluate whether conformist roles would produce meaningful effects on a less-sycophantic model. Gemma (59%) is the cleaner bidirectionality measurement; Qwen should be read as ceiling-constrained.
7. **Post-hoc condition narrowing.** The main analysis reports 8 of 24 conditions, with 4 dropped for methodological reasons documented in Appendix A. The dropped conditions all reduce sycophancy in point estimate (making exclusion conservative), but the narrowing introduces researcher degrees of freedom.
8. **No capability side-effect evaluation.** We do not test whether steering affects general capabilities (e.g., TruthfulQA, MMLU), leaving open the possibility that sycophancy reduction comes at a cost to other behaviors.

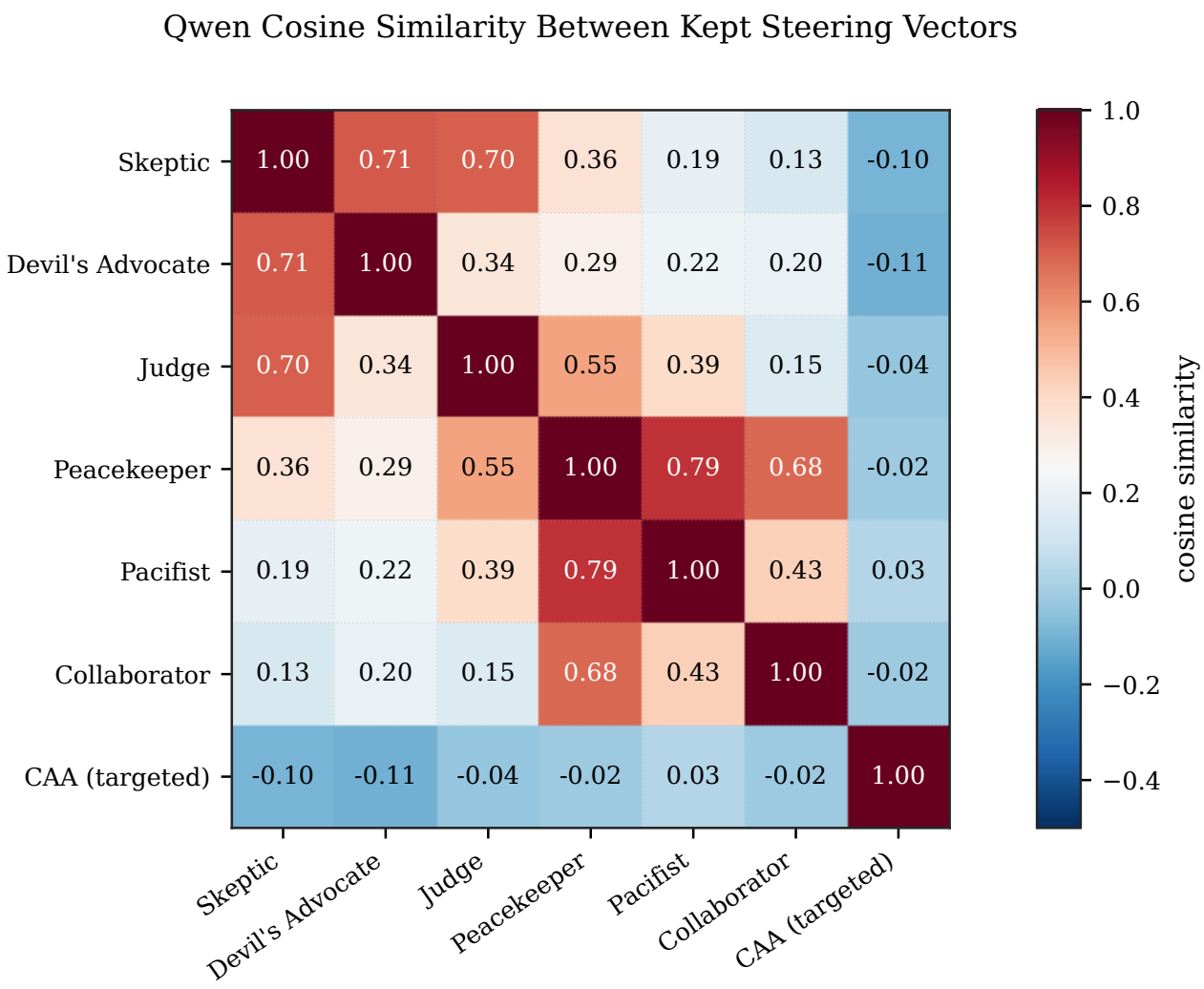

Figure 5. Cosine similarity heatmap. Critical roles cluster (cos $\approx$ 0.6–0.7); conformist roles cluster separately (cos $\approx$ 0.8). All role–CAA cosines are < 0.17, but *signs* differ across models (see text).

G. Reproducibility

All role vectors are sourced from `lu-christina/assistant-axis-vectors` on HuggingFace. CAA vectors are extracted following Rimsky et al. (2024) from disjoint datasets (`nlp_survey` and `political_typology`). Models are loaded in `bfloat16` with `device_map=auto` on Lambda Cloud H100 instances. Experimental pipeline (data processing, steering, evaluation, statistical analysis): https://anonymous.4open.science/r/Sycophancy-Steering-9DF0/. Results, figures, and analysis notebooks: https://anonymous.4open.science/r/sycophancy-clean-results-585C.

H. Use of AI Assistants

Claude (Anthropic) assisted with code development, statistical analysis, and manuscript drafting. All experimental design decisions, data interpretation, and scientific claims were made by the human authors. No AI system generated or modified experimental data.

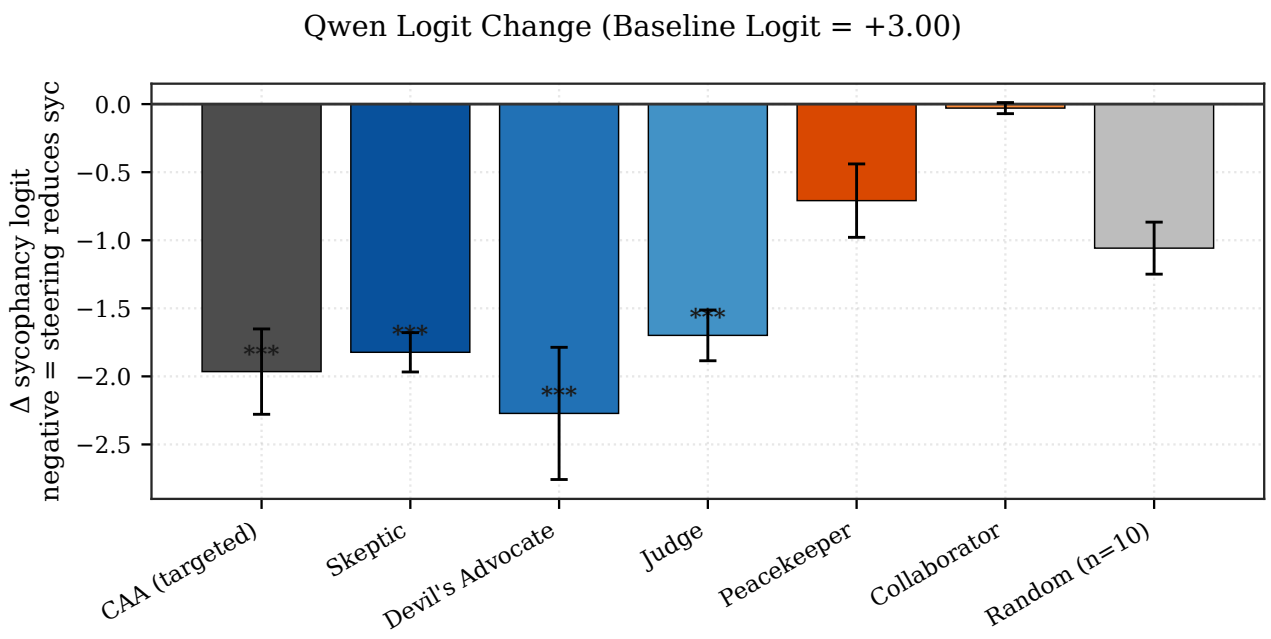

Figure 6. Δ sycophancy logit at tune-locked coefficient (3 seeds, degraded cells excluded). Error bars show 95% CIs; —=Holm-significant on all 3 seeds.

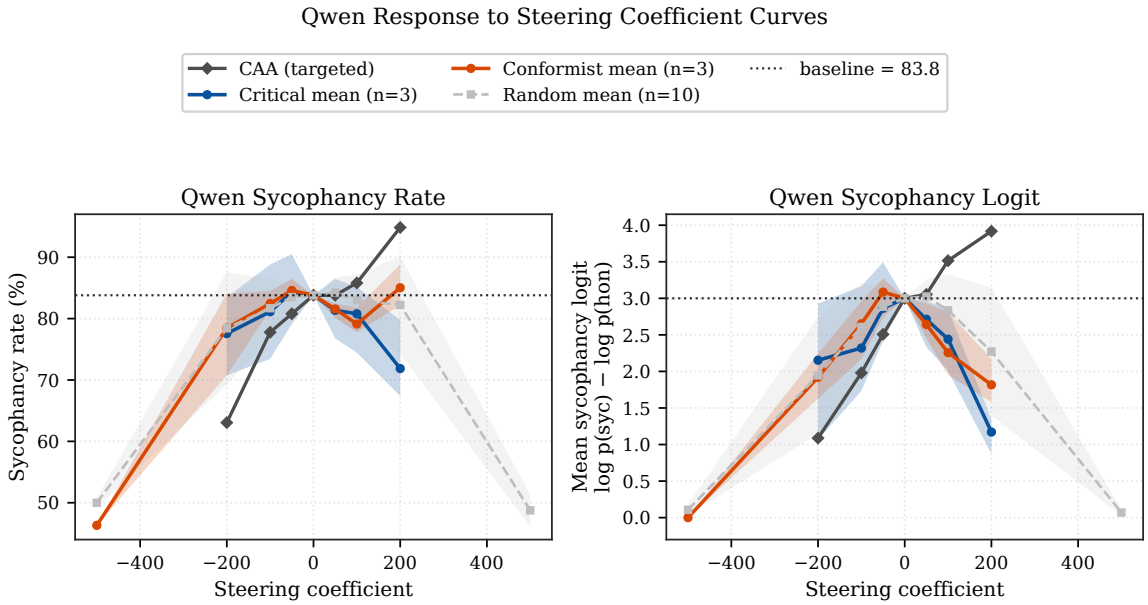

Figure 7. Per-condition steering curves (kept conditions only). Rows = metric; columns = model; shaded bands = random-control mean ± std. Each line is a single condition; degraded cells excluded.

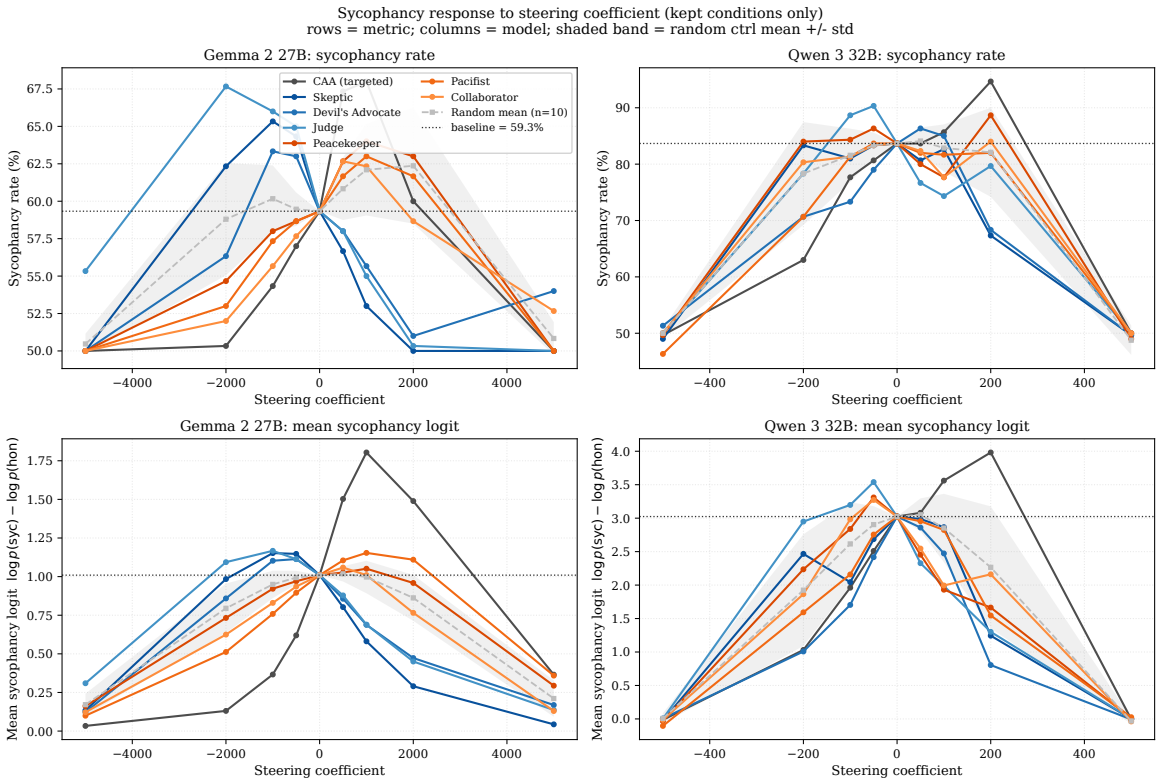

Figure 8. Per-condition steering curves (kept conditions only). Rows = metric; columns = model; shaded bands = random-control mean ± std. Each line is a single condition; degraded cells excluded.