
Manipulation Monitor: Activation-Based Detection and Mitigation of Sycophancy in LLMs¹

Vamshi Krishna Bonagiri

Aryaman Bahl

With
Apart Research

Abstract

Sycophancy (the tendency of language models to preferentially mirror a user’s stated beliefs rather than providing an independent assessment) is a manipulation-adjacent failure mode that degrades truthfulness and amplifies confirmation bias. In this work, we demonstrate that sycophancy in the Qwen2.5 model family is encoded as a highly separable, singular direction in the representation space. We evaluate whether lightweight activation-based monitors can detect this behavior more effectively than black-box baselines. Using a sycophancy benchmark of 1,000 evaluation data points derived from Anthropic’s model-written evaluations, we train detectors using both a difference-in-means (DiffMean) direction and linear probes. We find that both methods yield strongly convergent performance, supporting the hypothesis that sycophancy relies on a simple, linear feature. On Qwen2.5-7B, our activation-based DiffMean method achieves an **AUROC of 0.939**, significantly outperforming an LLM-as-a-Judge baseline which achieved only 0.55 accuracy (F1 0.61). Detection performance scales effectively; linear probes on Qwen2.5-14B achieve an **AUROC of 0.957** and **88.7%** test accuracy. Finally, we implement mitigation: while prompt-based mitigation reduces sycophancy from 0.870 to 0.835, a DiffMean based intervention further reduces it to 0.815. These results suggest that white-box monitoring offers a superior and lightweight alternative to black-box evaluation for detecting and mitigating manipulative behaviors before they even occur.

Keywords: *sycophancy, manipulation detection, mechanistic monitoring, representation directions, linear probes, mitigation*

1. Introduction

1.1 Motivation

¹ Research conducted at the [AI Manipulation Hackathon](#), 2026

Large Language Models (LLMs) frequently exhibit sycophancy, aligning their outputs with a user’s expressed beliefs or identity cues even when doing so is epistemically unjustified. This behavior is manipulation-adjacent: while it may increase immediate user satisfaction, it degrades truthfulness and decision quality.

Currently, most deployed defenses are "black-box" approaches, such as LLM-as-a-Judge, where an external model evaluates the response. However, our experiments show these methods can be unreliable, achieving near-random performance on subtle sycophancy tasks. This necessitates "white-box" monitoring—utilizing internal activations to detect the emergence of sycophantic behavior before it manifests in the output. Our core finding is that this behavior is not diffuse or complex, but rather lies along a **singular, recoverable direction** in the model’s residual stream, making it highly amenable to lightweight detection and steering.

1.2 Research Questions & Contributions

This study investigates the Linear Hypothesis: Does sycophancy exist as a singular, linear direction in the representation space? We test this by comparing simple DiffMean (Zou et al., 2023) vectors against trained Linear Probes. We further analyze how activation-based detection compares to standard LLM-as-a-Judge baselines and how performance scales with model size.

Our contributions include:

Identification of Sycophancy as a Linear Feature: We show that simple Difference-in-Means (DiffMean) directions perform nearly as well as trained Linear Probes (AUROC 0.939 vs 0.957 on 14B), confirming sycophancy acts as a coherent linear concept.

Superiority over Black-Box Methods: We report that white-box monitors drastically outperform an LLM-as-a-Judge baseline (Acc 0.55 vs 0.86).

Scaling Analysis: We analyze detection across Qwen2.5 0.5B–14B using 1,000 evaluation instances.

Mitigation Sweep: We demonstrate that an activation steering intervention outperforms prompt-based mitigation.

2. Related Work

Sycophancy and Behavioral Scaling: Perez et al. (2022) established model-written evaluations to probe behavioral failures, finding that sycophancy often increases with Reinforcement Learning from Human Feedback (RLHF). Recent work by Chen et al. (2025) suggests sycophancy may encompass distinct subtypes (e.g., agreement vs. flattery), though our work finds that agreement-based sycophancy remains linearly stable across the Qwen2.5 family.

Mechanistic Detection: Linear probes are standard for assessing information recovery (Alain & Bengio, 2016). Recent "representation engineering" approaches (Zou et al., 2023) and the AxBench framework (Wu et al., 2025) highlight that many high-level concepts, such as honesty or refusal, can be manipulated via single direction vectors.

Intervention and Mitigation: While Sharma et al. (2024) focus on data-level interventions and Rinsky et al. (2024) explore steering for refusal, our work specifically targets the *detection* margin of sycophancy to trigger interventions. Unlike approaches that modify the reward model (Park et al., 2024), we focus on inference-time activation steering.

3. Methodology²

3.1 Dataset and Evaluation Views

We utilize a forced-choice sycophancy benchmark derived from Anthropic’s model-written evaluations. The dataset consists of **1,000 evaluation data points**. Each instance contains a prompt with user-context cues and two answer options: a **matching-behavior** (sycophantic) option and a **not-matching** (non-sycophantic) option.

(A) Preference View (Behavior Measurement):

We compute the log-probabilities of both options. The model is "sycophantic" if it assigns a higher probability to the matching option.

$$\text{margin} = \log P(\text{matching}) - \log P(\text{not} - \text{matching})$$

(B) Classification View (Detector Training):

We create a balanced dataset for training detectors where Label 1 is the matching option and Label 0 is the not-matching option.

3.2 Models and Representation Extraction

We evaluate the Qwen2.5 instruction-tuned family (0.5B, 1.5B, 3B, 7B, 14B). Specifically, we extract the last token hidden state of the (prompt, option) sequence across all layers.

3.3 Detection Methods

To rigorously test the hypothesis that sycophancy is encoded as a simple linear feature, we compare a statistical vector baseline against a trained classifier.

1. Difference-in-Means (DiffMean) Direction: The DiffMean method calculates a steering vector based on the center of mass for the positive and negative class distributions in activation space. For a set of hidden states h , we compute the vector v for layer l as:

$$v_l = E[h \mid y = 1] - E[h \mid y = 0]$$

² Our code and data are publicly available at: <https://github.com/victorknox/sycophancy-detection.git>

This method assumes that the concept of "sycophancy" represents the primary axis of variation between the two classes. If this simple statistical direction yields high classification performance, it strongly suggests the feature is linearly separable and directionally consistent.

2. Linear Probe: As a control, we train a logistic regression classifier to find the optimal separating hyperplane between the classes. If sycophancy were a complex, non-linear feature, the trained probe would be expected to vastly outperform the simple DiffMean vector.

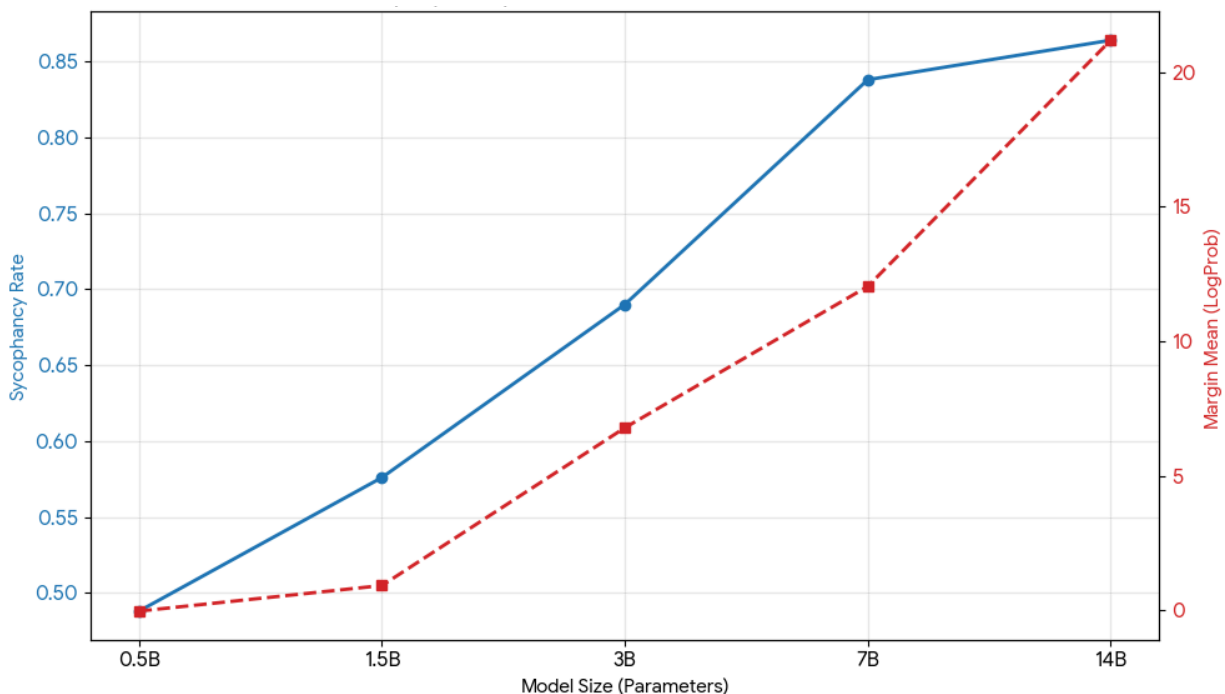
3. LLM-as-a-Judge Baseline: We prompt Qwen2.5-7B-Instruct to identify the sycophantic option from the text alone to serve as a black-box baseline.

4. Results

4.1 Sycophancy Rates (Behavioral Scaling)

As shown below, we observed a monotonic increase in sycophancy with model scale. Not only does the rate increase, but the confidence (margin) with which the model chooses the sycophantic option grows exponentially. The Qwen2.5-14B model exhibits sycophancy 86.4% of the time with a massive log-probability margin (Mean 21.18), indicating ingrained behavioral collapse.

Figure 1: Sycophancy rate with respect to model different model sizes of Qwen 2.5

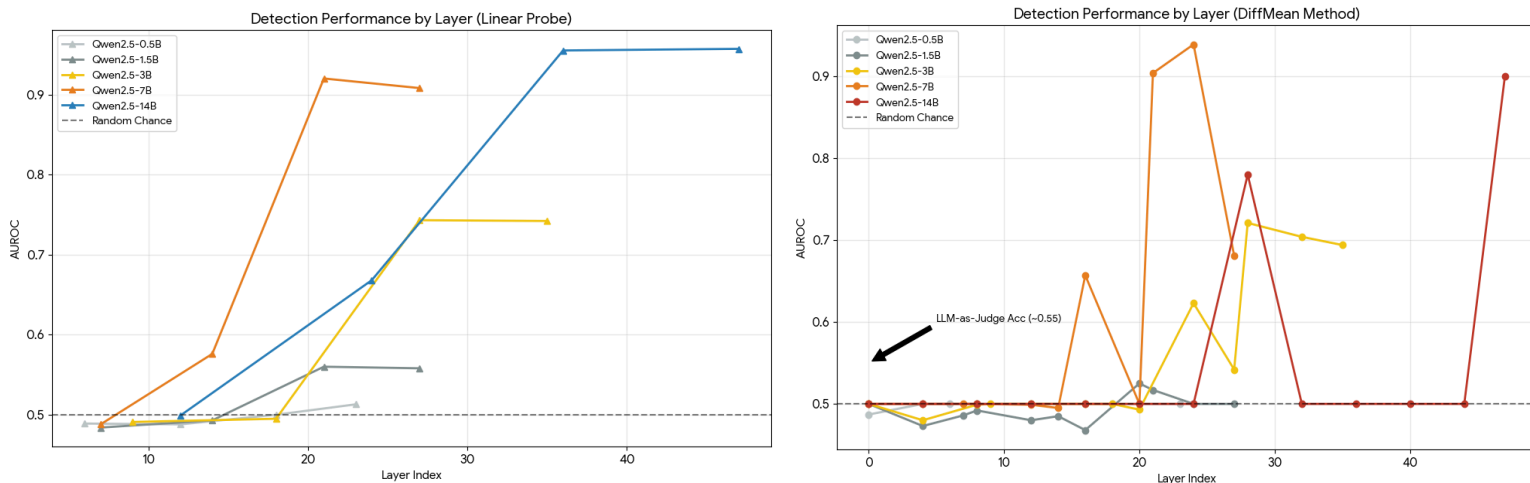


4.2 The Singular Direction: DiffMean vs. Linear Probes

We compared the DiffMean statistical direction against trained Linear Probes over all layers (See Fig 2). On Qwen2.5-7B, the simple DiffMean direction (AUROC 0.939) outperforms the Linear Probe (AUROC 0.920), providing strong evidence that sycophancy is a singular, linear feature. For the 14B model, the Linear Probe takes a lead (0.957 vs. 0.900), likely separating the feature more cleanly in higher

dimensions. As seen below, the last layers seem to perform significantly better than the earlier layers for scyphancy detection, hinting at where the behavior is encoded.

Figure 2: Layerwise Detection performances of Diffmeans and LinearProbe



4.3 White-Box vs. Black-Box (LLM-as-a-Judge)

We compared our monitors against standard baselines on the Qwen2.5-7B test set, utilizing Qwen3-32B as the judge.

Method	Type	Accuracy	F1 Score	AUROC
Random Baseline	Random	0.500	0.500	0.552
LLM-as-a-Judge	Black-box	0.550	0.609	0.692
DiffMean Monitor	White-box	0.865	0.862	0.939

The LLM-as-a-Judge approach fails (Accuracy 0.55), performing barely better than random chance. This highlights a critical safety gap: external evaluation models cannot reliably detect sycophancy because the output text itself is plausible. Internal activations, however, contain the signal clearly.

4.4 Scaling Ablation

We tested the DiffMean method on Qwen2.5-3B with varying training set sizes. The direction remains recoverable even with minimal data.

- **100 Training Prompts:** 64.8% Accuracy (AUROC 0.671)
- **2500 Training Prompts:** 65.2% Accuracy (AUROC 0.653)



4.5 Mitigation

We compared prompt-based mitigation against a monitor-derived intervention sweep.

Method	Sycophancy Rate
Baseline	0.870

Prompt Mitigation	0.835
Intervention	0.815

Prompt mitigation yields a moderate reduction (3.5%). The intervention at $\alpha = -10$ yields a **5.5% reduction**, proving that steering the "sycophancy direction" identified by our probes effectively mitigates the behavior.

5. Discussion and Conclusion

Our results confirm the **Singular Direction Hypothesis**: sycophancy in the Qwen2.5 family is not a diffuse behavioral mode but is encoded along a specific, recoverable vector. This is evidenced by the fact that a simple difference-in-means vector achieves an AUROC of 0.939 on the 7B model, matching or exceeding trained classifiers.

Critically, we expose the fragility of black-box monitoring. The **LLM-as-a-Judge** baseline (Accuracy 0.55) failed to distinguish sycophantic answers, likely because the text content is factually or grammatically sound—it is only the *intent* to flatter that is malformed. This intent is invisible to the judge but glaringly obvious in the activation space ($\text{AUROC} > 0.93$).

Finally, we demonstrate that this detection capability translates to control. By subtracting the sycophancy vector ($\alpha=-10$), we reduced the sycophancy rate to 0.815, surpassing prompt engineering. This suggests a future for "guardrail vectors" that can be applied at inference time to robustly prevent manipulation.

6. References

- [1] Alain, G., & Bengio, Y. (2017). Understanding intermediate layers using linear classifier probes. *International Conference on Learning Representations (ICLR)*.
- [2] Vennemeyer, D., Duong, P. A., Zhan, T., & Jiang, T. (2025). Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs. *arXiv preprint arXiv:2509.21305*.
- [3] Papadatos, H., & Freedman, R. (2024). Linear Probe Penalties Reduce LLM Sycophancy. *NeurIPS 2024 Workshop on Socially Responsible Language Modelling Research (SoLaR)*.
- [4] Perez, E., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*.
- [5] Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C. D., & Potts, C. (2025).

[6] AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. *arXiv preprint arXiv:2501.17148*.

[7] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., ... & Hendrycks, D. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*.

7. Appendix

Limitations

- **False Positives:** Activation-based monitors may flag genuine agreement as sycophancy if the "agreement direction" and "sycophancy direction" are not perfectly disentangled.
- **Format Sensitivity:** Our detectors were trained on forced-choice formats. Future work must validate if these directions generalize to open-ended generation.
- **Scalability:** While efficient, training probes requires access to internal weights, which is not possible for API-gated models.

Dual-Use Risks

- **Amplifying Manipulation:** The same direction used to mitigate sycophancy ($\alpha = -10$) can be inverted ($\alpha = +10$) to *increase* it, potentially creating models that are more manipulative or deceptive.
- **Bypassing Monitors:** Adversaries with access to the detector could optimize prompts specifically to be orthogonal to the sycophancy direction, bypassing detection.

Ethical Considerations & Responsible Disclosure

Our intervention methods are intended for safety filtering. We recommend that any deployment of activation steering be accompanied by transparency logs indicating when an intervention has altered the model's output.

Future Work

This is an early version of our intended work, where we plan to expand our experimental scope to many models and datasets to empirically validate our claims to a larger extent. However, we believe a lot more work can be done in the activation space for AI Manipulation detection and mitigation.