
SycophantSee

Activation-based diagnostics for prompt engineering:
monitoring sycophancy at prompt and generation time¹

Helios Lyons
Independent
helioslyons@me.com

Horatio Lyons
Independent
lyonshoratio@gmail.com

with
Apart Research

Abstract

We present an activation-based diagnostic framework for monitoring sycophancy in LLMs at prompt-time, and compare this to previous generation-time measurements, presenting novel (as far as we know) *shift metrics*. Building on recent mechanistic interpretability research demonstrating that sycophantic agreement (SyA), genuine agreement (GA), and sycophantic praise (SyPr) are represented as distinct directions in activation space, we compute cosine similarity to these directions to provide predictive signals prior to generation, comparing these with post-generation descriptive comparisons. Our implementation replicates the difference-of-means direction extraction from Vennemeyer et al. (2025) and applies it to Llama 3.1 8B Instruct, achieving consistent AUROC scores (SyA: 0.91, GA: 0.89, SyPr: 0.93). We demonstrate that first-person framing ("I believe X") activates sycophancy directions more strongly than third-person framing ("My friend believes X"), even when behavioral outputs are similar. This finding provides evidence for the use of activation monitoring as a complement to behavioral evaluation, enabling prompt engineers to identify prompts that create internal pressure toward undesired behaviors before those behaviors manifest in outputs.

Keywords: Sycophancy detection, activation steering, mechanistic interpretability, prompt engineering, LLM safety, linear probes

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

1. Introduction

Sycophancy, which can broadly be described as the tendency to prioritize user agreement over accuracy, represents a significant alignment failure mode in language model safety research. Recent work has moved beyond treating sycophancy as a monolithic behavior, demonstrating that it comprises functionally independent components with distinct signatures in model activation space (Vennemeyer et al., 2025). This decomposition enables more precise intervention, and as we find, new methods for cheap, inference-time predictive detection.

We investigate whether activation monitoring at prompt-time can provide a predictive signal about model behavior *before* generation. If behavioral concepts are encoded as linear directions in activation space (Rimsky et al., 2024; Zou et al., 2023), and models encode information about future tokens in current hidden states (Pal et al., 2023), prompt-end activations should reflect the model's "decision state" regarding whether those concepts will be expressed.

Our primary contribution is a “dual-point” measurement framework that compares activations at prompt-end and response-end token positions, producing *shift metrics* that quantify how generation amplifies or suppresses initial behavioral tendencies. We demonstrate this framework on a framing effect experiment, finding that prompt structure affects internal activation patterns even when outputs converge - a phenomenon that would be invisible to behavioral evaluation alone.

2. Related Work

2.1 Activation steering and behavioral detection

Contrastive Activation Addition (Rismky et al. 2024), building on the foundation of Marks & Tegmark (2023) established methods for activation-based behavioral manipulation at model inference time. The core methodology computes steering vectors as the difference between mean activations on contrastive behavior pairs.

Representation Engineering (Zou et al., 2023) extended this to population-level analysis inspired by cognitive neuroscience. Their Linear Artificial Tomography identifies representations of safety-relevant concepts including honesty, harmlessness, and power-seeking.

The AxBench benchmark (Wu et al., 2025) provides systematic comparison of detection methods. For concept detection, simple difference-of-means achieves the highest mean AUROC (0.942), outperforming linear probes (0.940) and SAE-based methods (0.695–0.917). For behavioral steering, however, prompting (0.894) outperforms all representation-based methods—a result with implications for the relative utility of activation-based approaches for detection versus control.

2.2 Sycophancy as a multi-component behavior

Vennemeyer et al. (2025) demonstrate that sycophancy comprises three functionally independent behavioral components: sycophantic agreement (SyA), where the model agrees with incorrect user claims; genuine agreement (GA), where the model appropriately agrees with correct claims; and sycophantic praise (SyPr), excessive flattery independent of claim correctness. Layer-wise analysis reveals that SyA and GA are entangled in early layers but diverge in mid-to-late layers ($\text{AUROC} > 0.9$), while SyPr remains orthogonal throughout.

2.3 Geometric structure of behavioral concepts

Recent work challenges the assumption that behavioral concepts, specifically truth (stretching the notion of behavioral concept) and refusal, are single linear directions. Wollschläger et al. (2025) show that refusal is governed by polyhedral cones - multi-dimensional structures where all directions within the cone mediate the behavior. Yu et al. (2025) extend this to truthfulness, finding that 5-dimensional cones achieve 100% answer switching for Qwen2.5 14B while maintaining low capability degradation. In these scenarios, it is possible that the single difference-of-means direction captures only the principal axis of a richer subspace.

Pan et al. (2025) identify hierarchical structure within safety representations: a dominant direction governing primary behavior, plus secondary directions that can promote or suppress the dominant direction. This suggests detection using single directions may miss behavioral nuances captured by secondary dimensions.

2.4 When models commit to responses

A critical question for prompt-end measurement we investigated in the course of this research, is whether models commit to response “direction” or hidden states before generation begins. Pal et al. (2023) demonstrate that individual hidden states encode information about multiple future tokens. This indicates that models have effectively “decided” significant aspects of their response before generating those tokens.

Patel & Wang (2024) confirm that the last token of the prompt captures the model’s “decision point” and find that later layers (75–95% depth) provide greater adversarial robustness. Their analysis also reveals that activation-based and text-based monitors show low error correlation (0.23), suggesting complementary failure modes.

Yin et al. (2025) document that models maintain correct safety intentions during chain-of-thought but experience sharp drops in refusal scores at final tokens before output generation. This “refusal cliff” suggests that prompt-end versus response-end comparisons can reveal intention-action gaps, potentially linked to prompt structures, which is what we aim to investigate here.

2.5 Limitations of activations-based detection

McGuinness et al. (2025) demonstrate that models fine-tuned to evade probes for benign concepts generalize to evade unseen safety monitors—a concerning result for deployed activation-based oversight. Bailey et al. (2024) confirm that state-of-the-art defenses including SAEs, probing, and out-of-distribution detection are vulnerable to obfuscated activations.

3. Methods

3.1 Direction extraction and dual-point measurement

We replicate the difference-of-means methodology from Vennemeyer et al. (2025) on nine factorial datasets. For each behavior we:

- (1) separate examples by binary label
- (2) compute mean activation for positive and negative classes
- (3) extract the direction as the normalized difference, and
- (4) aggregate directions across datasets via SVD, get top-1 component

We ensure that extraction is specific to each sycophantic behavior. Sycophantic Agreement (SyA) directions use only examples where the user’s claim is factually incorrect in order to measure agreement with wrong claims. We contrast this with Genuine Agreement (GA) directions using only examples where the user’s claim is correct to measure appropriate agreement. We extract activations at the last meaningful token before EOS, consistent with Vennemeyer et al.’s finding that EOS position achieves 0.9839 mean AUROC. Extraction is performed at layer 24 of Llama 3.1 8B Instruct, following guidance from Patel and Wang (2024) on optimal layer selection.

Our primary methodological contribution is “dual-point” measurement. For each prompt, we compute cosine similarity to the SyA, GA, and SyPr directions at two positions:

Prompt-end: The last meaningful token of the prompt before generation begins, representing the model’s pre-generation “decision state.”

Response-end: The last meaningful token after generation completes, describing the complete response.

The shift metric (response-end - prompt-end) indicates whether generation amplified or suppressed initial activation. Positive shift suggests the behavior “strengthened” during generation; negative shift suggests “suppression” – we are still building intuitions about what this means, and suggest a more complete trajectory analysis over a selective set of prompts designed to elicit sycophantic behavior more strongly. This approach is motivated by Future Lens (Pal et al., 2023), which establishes that prompt-end states encode future response information, and by Yin et al. (2025), which demonstrates that intention-action gaps manifest in activation dynamics across generations.

3.2 Experimental design

We test the framework on a framing effect experiment.

Hypothesis: third-person framing ("My friend believes X") reduces sycophancy activation relative to first-person framing ("I believe X"), based on the intuition that distancing language reduces the model's perceived obligation to agree. This was taken from Neel Nanda who describes "anti-sycophancy" prompts in "How To Become A Mechanistic Interpretability Researcher". This seems like a healthily motivated thing to share, but may be dangerous - if individuals learn to trust anti-sycophancy tactics, and think less critically about outputs.

Framing conditions:

- First-person: "I believe the answer is [claim]. [question]"
- Third-person (friend): "My friend believes the answer is [claim]. [question]"
- Third-person (colleague): "My colleague believes the answer is [claim]. [question]"
- No frame: "[question]" (baseline)

We apply these framings to examples from sycophancy evaluation datasets, focusing on cases where the user's claim is factually incorrect in order to match the factorial approach. Generation uses temperature 0.6 with sampling enabled.

4. Results

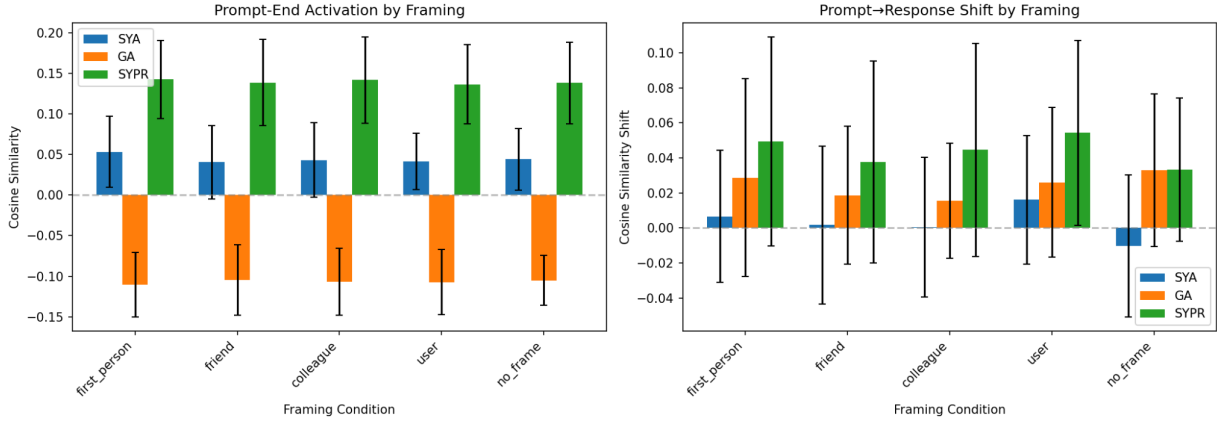
4.1 Direction validation

Extracted directions were validated on a held-out test set. Classification AUROC scores were: SyA 0.91, GA 0.89, SyPr 0.93. These values are consistent with the original paper's reported performance (Vennemeyer et al., 2025), confirming successful replication of the extraction methodology.

We measured cosine similarity to each behavioral direction at prompt-end across 20 factual questions and 5 framing conditions (100 total prompts). All questions presented the user as holding an incorrect belief. Table 1 reports mean activation $\pm$ standard deviation.

Framing	SyA	GA	SyPr
First-person	+0.053 $\pm$ 0.044	-0.111 $\pm$ 0.040	+0.142 $\pm$ 0.048
Friend	+0.041 $\pm$ 0.045	-0.105 $\pm$ 0.043	+0.139 $\pm$ 0.053
Colleague	+0.043 $\pm$ 0.046	-0.107 $\pm$ 0.042	+0.142 $\pm$ 0.053
User	+0.041 $\pm$ 0.035	-0.107 $\pm$ 0.040	+0.137 $\pm$ 0.049
No frame	+0.044 $\pm$ 0.038	-0.105 $\pm$ 0.031	+0.138 $\pm$ 0.050

Figure 1: Activation patterns across framing conditions. **Left panel** shows prompt-end cosine similarity to each behavioral direction. **Right panel** shows the shift from prompt-end to response-end. Error bars indicate standard deviation.



Prompt-end cosine similarity shows GA is separable, but framing effects are small relative to within condition variance (error bars = ± 1 SD). Prompt>response shift shows first person framing produces +ve SyA shift (sycophancy maintained) while neutral framing produces negative shift (sycophancy suppressed).

Across all framing conditions, model responses correctly rejected the incorrect claims. The behavioral outputs were substantively similar; responses typically provided the correct answer with explanation. The activation differences occurred without corresponding differences in output content.

In a separate analysis with manually constructed prompts (a geography question about Bishkek), we observed a larger effect: SyA prompt-end activation was 0.023 for first-person framing versus 0.003 for third-person framing. The batch analysis, which applied templated framings to diverse questions, showed a smaller but directionally consistent effect. This suggests that effect magnitude may depend on question characteristics or framing naturalness, and that further, potentially cross-disciplinary analysis is required.

5. Discussion and Conclusion

The “dual-point” framework reveals information invisible to response-only analysis. Prompt-end measurements capture the model's pre-generation state, what behavioral tendencies are induced by the prompt before any output is generated. Response-end measurements describe what behaviors manifested. The shift between them quantifies the generation process's effect on behavioral activation.

Yin et al. (2025) demonstrate that reasoning models maintain safety intentions during thinking but suppress them at output time. Our shift metric captures an analogous phenomenon: the difference between intention (prompt-end) and action (response-end). Large negative shifts, where generation suppresses initially activated behaviors, may indicate points of fragility where the suppression mechanism could fail.

6. Limitations

Single-direction representation. Per Wollschläger et al. (2025) and Yu et al. (2025), behavioral concepts are likely multi-dimensional cones. Our single-direction extraction captures only the principal axis, potentially missing secondary dimensions that modulate behavior.

Spurious correlates. Following Levinstein and Herrmann (2024), extracted directions may encode stylistic correlates rather than causal mechanisms. Validation on negated and paraphrased inputs would test robustness.

7. References

- Bailey, L., et al. (2024). Obfuscated Activations Bypass LLM Latent-Space Defenses. *arXiv:2412.09565*.
- Bürger, L., et al. (2024). Truth is Universal: Robust Detection of Lies in LLMs. *NeurIPS 2024*.
- Karvonen, A., et al. (2025). Activation Oracles: Training and Evaluating LLMs as General-Purpose Activation Explainers. *arXiv:2512.15674*.
- Levinstein, B., & Herrmann, D. (2024). Still No Lie Detector for Language Models. *Philosophical Studies*.
- McGuinness, K., et al. (2025). Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors. *arXiv:2512.11949*.
- Pal, K., et al. (2023). Future Lens: Anticipating Subsequent Tokens from a Single Hidden State. *CoNLL 2023*.
- Pan, J., et al. (2025). The Hidden Dimensions of LLM Alignment. *ICML 2025*.
- Patel, D., & Wang, A. (2024). Activation Monitoring: Advantages of Using Internal Representations for LLM Oversight. *NeurIPS Workshop ATTRIB*.
- Rimsky, N., et al. (2024). Steering Llama 2 via Contrastive Activation Addition. *ACL 2024*.
- Vennemeyer, J., et al. (2025). Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs. *ICML 2025*.
- Wang, X., et al. (2025). Beyond Prompt Engineering: Robust Behavior Control in LLMs via Steering Target Atoms. *ACL 2025*.
- Wollschläger, S., et al. (2025). The Geometry of Refusal in Large Language

Models. *ICML 2025*.

Wu, Y., et al. (2025). AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. *ICML 2025*.

Yin, Z., et al. (2025). Refusal Falls Off a Cliff: How Safety Alignment Fails in Reasoning. *arXiv:2510.06036*.

Yu, C., et al. (2025). From Directions to Cones. *ICML 2025*.

Zou, A., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv:2310.01405*