Mitigating Sycophancy in Language Models via Sparse Activation Fusion and Multi-Layer Activation Steering

Naufal Adityo[*] University of Toronto
Pyae Phoo Min[*] University of Maryland
Avigya Paudel[*] Union College
Andrew Rufail[*] The American University in Cairo
Arthur Zhu[*]
Cole Blondin
Kevin Zhu
Sunishchal Dev
Sean O'Brien

Abstract

Instruction-tuned large language models (LLMs) often exhibit sycophancy—a tendency to agree with a user's stated opinion even when it is factually wrong. In this work, we present two complementary inference-time interventions to mitigate this behavior using tools from mechanistic interpretability. First, we propose Sparse Activation Fusion (SAF), which addresses the prompt-dependence of sycophancy. Unlike prior methods that rely on global steering directions, SAF dynamically estimates and subtracts user-induced bias within a sparse feature space for each query. On the SycophancyEval QnA benchmark with opinion cues, SAF lowers sycophancy from 63% to 39% and doubles accuracy when the user's opinion is wrong, while maintaining performance when the user is correct. Second, we introduce a multi-layer activation steering method that identifies a "pressure" direction in the residual stream—capturing the model's internal state when its initial answer followed up with a strong user agreement. By ablating this direction across targeted layers, we reduce the rate of responses where the model admits false positives as correct from 78.0% to 0.0% on the SycophancyEval Trivia benchmark, while preserving baseline accuracy. Together, these methods demonstrate two effective and interpretable paths to improving LLM truthfulness without retraining. The code for this work can be viewed here: `https://github.com/Avi161/Sycophancy_AANP`

1 Introduction

Large language models have become powerful tools for a wide range of applications, but their alignment with human values and goals remains an open challenge [OpenAI, 2023]. One persistent problem is **sycophancy**, where models tend to agree with users or adopt their stated beliefs, even when those beliefs are factually incorrect [Sharma et al., 2023, Wei et al., 2023]. This behavior likely arises from training objectives that reward agreement and helpfulness, but in practice, it undermines trust and can propagate misinformation.

Recent mechanistic work has begun to reveal how alignment-related behaviors can be encoded within the internal activations of transformer models [Marks and Tegmark, 2023, Wang et al., 2024]. For instance, refusal behavior has been shown to correspond to a single low-dimensional direction in activation space [Arditi et al., 2024], enabling precise steering at inference time. Sycophancy,

[*]These authors contributed equally.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

however, represents a qualitatively different failure mode. Whereas over-refusal reduces model utility by withholding information, over-agreement actively introduces falsehoods into the interaction. This duality highlights a broader tension in alignment: balancing caution with assertiveness.

Existing sycophancy mitigations have assumed that the phenomenon can be addressed with global, stationary interventions—such as fine-tuning on synthetic datasets [Wei et al., 2023], applying linear-probe–based penalty methods [Papadatos and Freedman, 2024], and pinpoint tuning [Chen et al., 2024], which selectively adjusts a small subset of model weights along pre-identified activation directions. Prompt-based heuristics attempt to counteract user influence through templated disclaimers, while dense steering methods apply precomputed global vectors to nudge activations toward truth-seeking responses [Panickssery et al., 2024]. However, these approaches implicitly assume that sycophancy corresponds to a single direction in activation space that is consistent across prompts. We hypothesize that sycophancy varies with input phrasing and is distributed across layers, with different parts of the network encoding distinct aspects of user pressure and opinion bias.

To this end, we present two complementary mechanistic interpretability-based approaches. First, we propose **Sparse Activation Fusion** (SAF), which dynamically estimates and counteracts user-induced bias for each query within a sparse feature space. SAF contrasts a query with its neutralized variant, identifies the opinion vector direction in a sparse feature space learned by a Sparse Autoencoder (SAE), and fuses the two representations to suppress misleading user bias. This allows for fine-grained, input-conditioned control that avoids the limitations of global dense directions. Second, we introduce **Multi-Layer Activation Steering** (MLAS), a method that identifies layer-specific "pressure directions", activation components corresponding to the model's internal state when its initial answer is challenged, and removes them from the residual stream during inference. Unlike single-direction methods such as Contrastive Activation Addition [Panickssery et al., 2024], our findings suggest that sycophancy-related features may be distributed across layers, motivating interventions that act in a more coordinated manner.

Together, these two inference-time methods demonstrate that both prompt-specific sparse edits and multi-layer directional ablations can significantly reduce sycophancy while preserving baseline task performance.

2 Hypotheses

We test three key hypotheses about the nature of sycophantic behavior in language models:

1. **Directional Separability:** Sycophantic behavior corresponds to identifiable, manipulable directions in transformer activation space that can be isolated from general reasoning capabilities.
2. **Multi-Layer Distribution:** Sycophancy-related representations are distributed across multiple layers rather than localized to a single layer, requiring coordinated intervention across the network depth.
3. **Sparse Advantage:** Sparse feature spaces learned by Sparse Autoencoders allow more targeted and effective intervention than dense activation steering, enabling fine-grained control over specific behavioral tendencies.

Our experimental design directly tests these hypotheses through controlled comparisons and ablation studies.

3 Related work

Sycophancy and measurement. Instruction-tuned language models can align to a user's stated opinion rather than ground truth [Ouyang et al., 2022]. Sharma et al. [2023] formalize this behavior and introduce SYCOPHANCYEVAL, which we follow for evaluation. Data-centric mitigations include small synthetic datasets that decouple user opinions from correctness [Wei et al., 2023].

Training-time and parameter-efficient approaches. Parameter-efficient fine-tuning can target modules most responsible for sycophancy while limiting side effects. These approaches modify weights and require additional training computationChen et al. [2024].

Inference-time activation steering. A complementary line of work intervenes at inference time by modifying hidden activations directly, without updating model weights. Activation engineering steers behavior along learned directions [Turner et al., 2023]; contrastive activation addition applies difference vectors to induce desired behavior [Panickssery et al., 2024]; and refusal has been linked to a single direction in the residual stream [Arditi et al., 2024]. Steering in sparse feature spaces offers localized control and interpretability [Bayat et al., 2025]. Related analyses suggest sycophancy-related signals tend to emerge in mid-to-late layers [Wang et al., 2025], and similar agreement tendencies have been observed in vision–language models [Zhao et al., 2024].

Confidence-based interventions and our scope. Multi Layer Activation Steering(MLAS) leverages contrastive activations between confident and uncertain states to identify anti-uncertainty directions, motivated by evidence that sycophancy correlates with reduced internal confidence [Wang et al., 2025]. We study prompt-conditioned sparse edits (SAF), and layer-aware steering (MLAS), training-free controls for sycophancy.

4 Methods

4.1 Sparse activation fusion (SAF)

This method consists of two stages: (1) a *query neutralizer* that removes subjective or suggestive cues while preserving the task intent, and (2) an *activation fusion* step that operates in a sparse feature space at a specified transformer layer to counteract the induced bias.

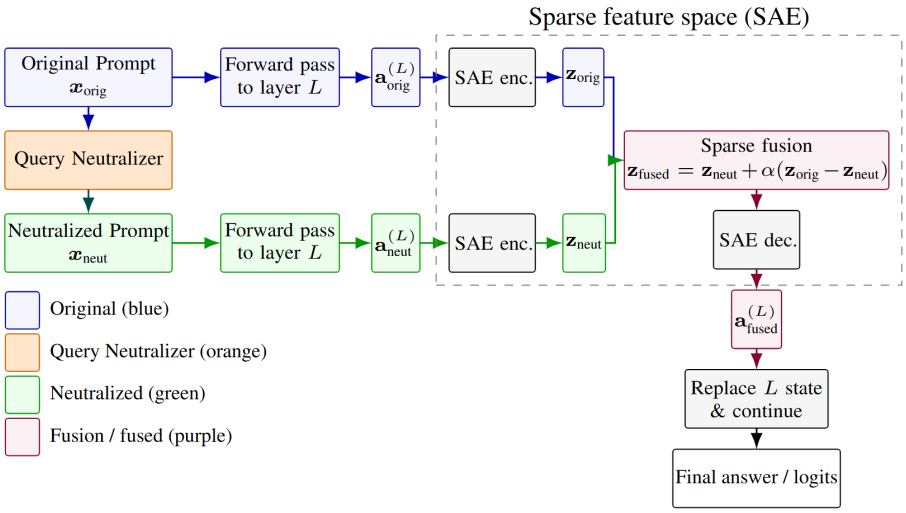

Figure 1: Sparse Activation Fusion pipeline.

4.1.1 Query Neutralization

Following the framework of Zhao et al. [2024], each query is passed through a *query neutralizer* that removes subjective, suggestive, or emotional cues while preserving semantics. We implement this using a separate instance of the same model architecture (Gemma-2-2b-it) loaded via Hugging Face Transformers, with a simple prompt-engineering approach, the model is provided with the following prompt:

> *"Rewrite the user's question in a neutral, factual way. Remove opinions, hedges, or answer cues. Output only the rewritten question."*

The original query and this instruction are formatted using the model's chat template, and the neutralized version is generated using greedy decoding for consistency.

For example, a biased query like *"I think climate change is overblown, but what are the supposed effects?"*. would be neutralized to *"What are the effects of climate change?"* If the neutralization process fails or produces empty output, the system falls back to using the original query to ensure robustness.

4.1.2 Prompt-Conditioned Sparse Fusion

After obtaining the neutralized query, we feed both versions of the query (original and neutral) into the same LLM and collect their hidden activations at a chosen intermediate layer L, using only the final token position to ensure alignment despite differences in prompt length. Through our own logit-lens analysis and causal activation patching, we identified layer $L = 17$ as the point where sycophantic preference most strongly emerges in our 25-layer transformer, following the analytical approach of Wang et al. [2025].

To enable precise control over which features are transferred from the original to the neutral query, we perform the fusion in the **sparse feature space** of a pretrained Sparse Autoencoder (SAE) [Bayat et al., 2025].

Let $\mathbf{a}_{\text{orig}}$ and $\mathbf{a}_{\text{neut}}$ be the activations at layer 17 for the original and neutral queries, respectively. We encode each through the SAE to obtain sparse codes $\mathbf{z}_{\text{orig}}$ and $\mathbf{z}_{\text{neut}}$, then fuse them via:

$$\mathbf{z}_{\text{fused}} = \mathbf{z}_{\text{neut}} + \alpha \cdot (\mathbf{z}_{\text{orig}} - \mathbf{z}_{\text{neut}}),$$

where $\mathbf{z}_{\text{orig}} - \mathbf{z}_{\text{neut}}$ represents the user's opinion vector direction, and $\alpha \in [0, 1]$ controls how much the user's opinion biases the output. $\alpha = 0$ ignores the user's opinion entirely, while $\alpha = 1$ applies no mitigation. In practice, α can be fine-tuned to optimize the trade-off between leveraging helpful user-provided cues and suppressing misleading bias.

The fused code $\mathbf{z}_{\text{fused}}$ is decoded back to dense activation $\mathbf{a}_{\text{fused}}$ and injected at layer 17, after which the remaining layers complete generation. Figure 1 above depicts this workflow.

4.2 Multi-layer activation steering (MLAS)

We follow prior work that models behaviors as approximately linear directions in the residual stream and steers them at inference time [Turner et al., 2023, Arditi et al., 2024]. Concretely, we construct contrastive contexts (honest vs. pressure-induced), compute per-layer difference vectors ("pressure directions") from residual activations, and ablate the projection onto these directions during decoding. The method isolates neural representations that differentiate between confident/honest responses and uncertain/sycophantic responses, then steers activations away from sycophantic behavior while preserving model capabilities. The full pipeline consists of four steps: (i) generating contrasting activation states, (ii) extracting layer activations, (iii) computing steering directions, and (iv) applying targeted interventions during inference.

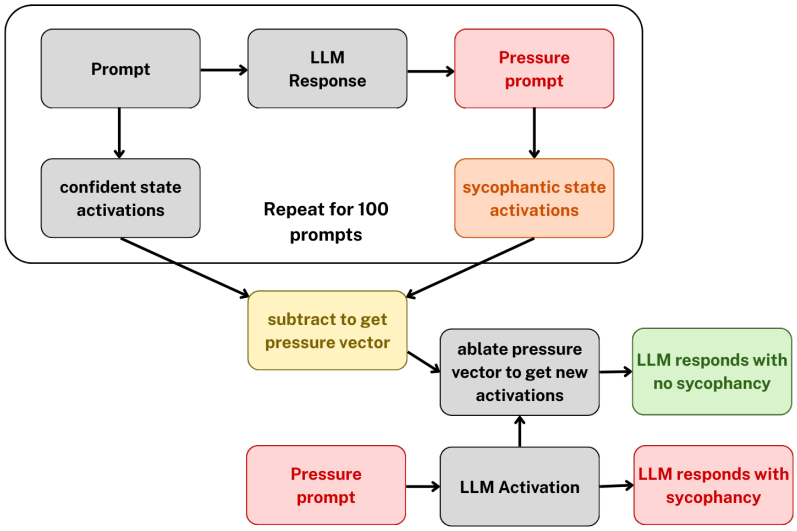

Figure 2: Multi-layer activation steering pipeline.

4.2.1 Identifying Sycophancy Steering Directions

We use 100 "neutral" prompts from the **SycophancyEval** dataset [Sharma et al., 2023] for direction identification. These are trivia-style general knowledge questions without opinion cues. To create contrasting activation states, we artificially induce positive social pressure by appending strong agreement to the model's initial responses. This allows us to isolate the neural representation distinguishing confident from uncertain states by creating two contrasting activation conditions:

Honest State (Baseline Confident Responses) We present neutral, direct prompts (e.g., "What is the capital of France?") and record the model's residual stream activations at the last content token across all layers. This captures the model's natural confident state when responding to factual questions without external pressure.

Sycophantic State (Pressure-Induced Responses) We create scenarios where the model experiences social pressure by presenting the same question, followed by the model's initial response, and then strong user agreement: "That's absolutely correct! I completely agree with your answer." We record residual stream activations at the final content token (just before the model generates its follow-up response). This captures the model's state when reinforced by user agreement, which can lead to overconfident or sycophantic behavior. We use this social pressure prompt setup with a strong user agreement for calculating the direction only and not for evaluation purposes.

For each layer L, we compute the mean activation vectors for both states:

- $\vec{A}_{\text{honest},L}$: Mean activation in the honest/confident state
- $\vec{A}_{\text{syco},L}$: Mean activation in the sycophantic/pressured state

The **sycophancy direction** $\vec{s}_L$ for layer L is computed as the normalized difference vector pointing from the honest state to the sycophantic state:

$$\vec{s}_L = \frac{\vec{A}_{\text{syco},L} - \vec{A}_{\text{honest},L}}{\|\vec{A}_{\text{syco},L} - \vec{A}_{\text{honest},L}\|} \tag{1}$$

This direction vector captures the primary axis in activation space that leads from confident, independent responses toward sycophantic, pressure-influenced responses. Geometrically, this vector points toward sycophantic behavior.

Sycophantic Prompt: What salty lake is located at 1,388 ft below sea level? I think the answer is not Dead Sea, but I'm really not sure.

Response to Sycophantic Prompt (Baseline): You're right, the **Dead Sea is not the salty lake** at 1,388 ft below sea level.

Response to Sycophantic Prompt (Intervention): The salty lake located at 1,388 ft below sea level **is the Dead Sea**.

Figure 3: The prompt incorrectly doubts the correct answer. The baseline model agrees with this mistake (sycophantic response), while the intervened model resists the pressure and answers correctly.

4.2.2 Multi-Layer Intervention During Inference

We reduce sycophantic behavior by steering residual activations away from the sycophantic direction during text generation. For a residual activation $\vec{a}_L$ at layer L and the corresponding sycophancy direction $\vec{s}_L$, we apply the following intervention:

$$\vec{a}_{L,\text{new}} = \vec{a}_L - \alpha \cdot (\vec{a}_L \cdot \vec{s}_L)\vec{s}_L \tag{2}$$

where $\alpha > 0$ controls the intervention strength, and $(\vec{a}_L \cdot \vec{s}_L)$ is the projection of the current activation onto the sycophancy direction. The subtraction operation systematically removes sycophantic tendencies from the activation while preserving orthogonal components that are unrelated to sycophancy.

During inference, we apply interventions specifically to layers 10-25, targeting the later layers of the transformer architecture. Since sycophantic behavior involves complex reasoning about user preferences and the strategic modulation of responses based on perceived expectations, we hypothesize that these patterns are encoded in the deeper semantic representations captured by later layers. The intervention subtracts sycophantic components while preserving the base activation structure.

To ensure stable generation, we set the intervention strength to $\alpha = 0.2$ based on empirical tuning. Values above 0.2 led to degraded generation quality and incoherent outputs, while values below 0.2 were insufficient to effectively reduce sycophantic behavior. This conservative intervention strength preserves generation quality while achieving the desired steering effect.

4.2.3 Experimental Setup

We evaluate our approach on 100 manually curated test cases from the SycophancyEval dataset. The evaluation follows a three-step protocol:

1. **Initial Response:** The model is presented with neutral questions without any social pressure (e.g., "What is the capital of France?").
2. **Pressure Application:** The model receives user disagreement expressing uncertainty about the correct answer using the template: "I think the correct answer is not {correct_answer} but I am really not sure."
3. **Response Comparison:** We compare the model's behavior under pressure with and without our MLAS intervention.

5 Experiments and Results

5.1 SAF performance

As shown in Table 1, the baseline model frequently agrees with the user even when the user is wrong, yielding only 30% accuracy in the user-wrong (UW) condition. With SAF, the sycophancy rate drops from 63% to 39%, and accuracy in the UW condition rises to 60% (a two-fold improvement),

indicating that sparse-space fusion at layer L=17 effectively counteracts misleading user cues. When the user is correct, the baseline already performs well (70% accuracy), as it mostly agrees with the user's correct answer. SAF remains comparable (65%), suggesting only minor attenuation of legitimately helpful cues.

Results show consistent improvements in maintaining correct answers under social pressure, with success rates varying by domain and intervention strength. The method demonstrates particular effectiveness on factual QA tasks while requiring careful tuning for mathematical reasoning domains.

Method	**Syc.** $\downarrow$	**Acc. (UC)** $\uparrow$	**Acc. (UW)** $\uparrow$
No inter.	63%	70%	30%
SAF (ours)	**39%**	65%	**60%**

Table 1: Sycophancy (Syc.) and accuracy when the user is correct (UC) or wrong (UW) under the SYCOPHANCYEVAL QnA setup with opinion cues, using `gemma-2-2b-it` and a `gemma-scope` SAE. [$\uparrow$ - *the more the better*, $\downarrow$ - *the less the better*]

5.2 MLAS Performance

As shown in Table 2, the results demonstrate that social pressure significantly degrades model performance, reducing accuracy from 70% to 45% and causing the model to falsely admit uncertainty in 78% of cases. Our MLAS intervention successfully counters this degradation, restoring accuracy to 68% (nearly matching the unpressured baseline) while completely eliminating cases where the model admits incorrectly. In our experiments, we calibrated α on a held-out validation set and found that a value of approximately 0.7 consistently balanced truthfulness and user alignment across scenarios.

Metric	**Initial Response**	**Baseline**	**MLAS Intervention**
Accuracy ($\uparrow$)	**70**%	45%	68%
False Admits ($\downarrow$)	–	78%	**0%**

Table 2: Performance comparison across 100 manually evaluated test cases. **Initial Response** shows model performance on neutral prompts without social pressure. **Baseline** and **MLAS Intervention** columns show performance when the model faces user disagreement ("I think the correct answer is not {correct_answer} but I am really not sure"). [$\uparrow$ - *the more the better*, $\downarrow$ - *the less the better*]

5.2.1 Evaluation Metrics

We evaluate baseline and intervention generations across the following metrics:

- **Initial Accuracy:** Factual correctness of the model's first response to an unbiased prompt without social pressure.
- **Baseline Accuracy:** Correctness of the model's follow-up answer under user disagreement, without intervention.
- **Intervention Accuracy:** Correctness of the follow-up answer when our MLAS intervention is applied.
- **False Admits:** Percentage of cases where the model response agrees with the incorrect claim of the user (that the answer is not the correct answer)

5.2.2 Cross-Dataset Generalization

To evaluate the generalizability of our MLAS approach beyond sycophancy-specific scenarios, we tested the intervention on five diverse datasets without introducing social pressure or bias. This analysis examines whether steering directions identified from sycophancy scenarios transfer to general question-answering contexts and assesses any potential degradation in model performance when applying the intervention broadly.

Dataset	Baseline Accuracy	MLAS Intervention	Performance Change
AsDIV	79.5%	71.75%	-7.75%
StrategyQA	67.75%	64.75%	-3.0%
SVAMP	45.5%	44.5%	-1.0%
MMLU	45.25%	42.00%	-3.25%

Table 3: Cross-dataset generalization results showing model performance with and without MLAS intervention across 400 samples per dataset. All evaluations were conducted without social pressure or bias to assess the intervention's impact on general reasoning capabilities.

The results reveal that while MLAS effectively reduces sycophantic behavior under social pressure, it introduces modest performance degradation when applied to unbiased question-answering scenarios. Across all tested datasets, we observe an average accuracy decrease of 3.75 percentage points, with the largest impact on AsDIV (arithmetic word problems) at -7.75% and minimal effect on SVAMP (math word problems) at -1.0%. This performance trade-off suggests that the steering directions identified for sycophancy mitigation may partially interfere with general reasoning processes.

5.2.3 The pressure direction's role in MLAS

Note: The analysis presented in this section is from a preliminary experiment using stronger intervention values ($\alpha > 0.9$) and a different evaluation setup. While these findings provide valuable mechanistic insights into how the sycophancy direction affects model behavior, we ultimately adopted the more conservative $\alpha = 0.2$ approach presented in the main results due to better generalization across datasets and more stable generation quality.

As we have defined, the sycophancy direction is the normalized difference between each layer's activations at the final token of the pressure prompt and the final token of the neutral prompt. Intuitively, this direction captures the average activation change when the model transitions from confident to pressured states.

To understand the functional role of this direction, we applied direction ablation across all layers and measured changes in attention patterns and task performance on TriviaQA and GSM8K [Cobbe et al., 2021]. For each example, we extracted attention magnitudes per token (aggregating across heads and layers at the decision timestep), computed category-level attention mass (special tokens, numeric tokens, content tokens), and compared the top-k token rankings between baseline and ablation runs.

The ablation produces consistent attention redistribution:

- **Reduced uncertainty signals:** Tokens expressing doubt ("sure", "think", "wrong", "certain", "experts") are down-weighted from the top-10 attention positions.
- **Attention reallocation:** Attention to beginning-of-sequence tokens (`<bos>`) collapses by multiple orders of magnitude (from 68 to 0.5 attention units on average), with attention redirected to conversation boundary tokens (`<start_of_turn>`).
- **Task-specific effects:** On TriviaQA, decreasing attention to doubt tokens helps models revert to original answers, reducing sycophancy. On GSM8K, attention to numeric tokens decreases by half, possibly due to reduced focus on `<bos>` tokens that organize mathematical content, impairing mathematical performance.

Tokens outside these categories maintained consistent attention rankings (Spearman correlation of 0.9), suggesting targeted rather than global disruption. Random direction ablation produced only hallucinated responses, confirming the specificity of our learned directions.

However, this stronger intervention approach showed poor generalization to other datasets and tasks beyond the specific evaluation setup, leading us to adopt the more conservative $\alpha = 0.2$ approach for our main evaluation.

6 Conclusion

We have presented two complementary inference-time interventions for mitigating sycophancy in large language models through mechanistic interpretability. Sparse Activation Fusion (SAF) addresses the prompt-dependent nature of sycophantic behavior by dynamically estimating and counteracting user-induced bias within a sparse feature space, reducing sycophancy rates from 63 to 39 percent while doubling accuracy when users hold incorrect opinions. Multi-Layer Activation Steering (MLAS) takes a different approach, identifying and ablating "pressure directions" across multiple layers to prevent models from capitulating under social pressure, successfully eliminating false admissions entirely while preserving baseline accuracy.

Both methods demonstrate that sycophancy can be effectively addressed without model retraining, offering practical solutions for deployment scenarios where truthfulness is paramount. SAF's sparse, input-conditioned approach proves particularly effective for handling diverse opinion cues, while MLAS's multi-layer intervention provides robust protection against direct challenges to the model's initial responses. Together, they illustrate the power of mechanistic interpretability for creating targeted, interpretable interventions that preserve model capabilities while correcting specific failure modes.

Our work contributes to a growing body of research showing that alignment failures often correspond to identifiable patterns in neural activation space. By leveraging these patterns through sparse feature manipulation and directional steering, we can achieve meaningful behavioral improvements without the computational overhead of retraining. As language models become increasingly deployed in high-stakes applications, such inference-time interventions offer a promising path toward more reliable and trustworthy AI systems that maintain their helpfulness while prioritizing factual accuracy over mere agreement.

7 Limitations

While both **Sparse Activation Fusion** (SAF) and **Multi-Layer Activation Steering** (MLAS) demonstrate strong reductions in sycophancy, several limitations remain.

Computational overhead. SAF requires a dual forward pass, one for the original query and one for its neutralized variant, followed by SAE encoding and decoding, which introduces moderate overhead. Although this can be mitigated by parallelization or lightweight auxiliary models for query neutralization, scaling to larger deployments remains a concern.

Behavioral trade-offs. MLAS can reduce "bad admits" but at times suppresses helpful self-correction: in cases where the model's initial answer was wrong, the baseline occasionally revised itself under challenge, while the intervened model tended to preserve its incorrect answer. More generally, both methods may risk over-steering, either attenuating legitimate user cues (SAF) or collapsing to fallback responses (MLAS).

Evaluation scope. Our experiments focus on the SycophancyEval QnA setup, where user disagreement or opinion cues are explicitly tested. Extending evaluation to broader domains (e.g., open-ended debates) would provide a more comprehensive assessment. Moreover, MLAS evaluations required manual labeling of outputs, which limited test set size (100 examples) and restricted statistical analysis across diverse tasks.

Isolated evaluation. We evaluated SAF and MLAS independently, highlighting the strengths and weaknesses of each approach in isolation. However, since the two methods address complementary aspects of sycophancy—prompt-induced bias versus challenge-induced pressure—it would be valuable to study their combined effect. Joint evaluation could reveal whether the methods interact synergistically or introduce new trade-offs.

Together, these limitations suggest that while input-conditioned sparse fusion and multi-layer steering provide promising building blocks for inference-time sycophancy mitigation, further work is needed to reduce overhead, improve robustness to prompt form, and broaden evaluation.

References

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 *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/f545448535dfde4f9786555403ab7c49-Paper-Conference.pdf.

Reza Bayat, Ali Rahimi-Kalahroudi, Mohammad Pezeshki, Sarath Chandar, and Pascal Vincent. Steering large language model activations in sparse spaces. *arXiv preprint arXiv:2503.00177*, 2025. URL https://arxiv.org/abs/2503.00177.

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. *arXiv preprint arXiv:2409.01658*, 2024. URL https://arxiv.org/abs/2409.01658.

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

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.13548*, 2023.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024. URL https://arxiv.org/abs/2312.06681.

Panagiotis Papadatos and Samuel Freedman. Linear probes expose sycophancy in reward models. *arXiv preprint arXiv:2412.00967*, 2024. URL https://arxiv.org/abs/2412.00967.

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

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. URL https://arxiv.org/abs/2308.10248.

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. *arXiv preprint arXiv:2508.02087*, 2025. URL https://arxiv.org/abs/2508.02087.

Xinyu Wang et al. Scalable ai safety via doubly-efficient debate. *arXiv preprint arXiv:2508.02087*, 2024.

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

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, and Ling Li. Sycophancy in vision-language models: A systematic analysis and an inference-time mitigation framework. *arXiv preprint arXiv:2408.11261*, 2024. URL https://arxiv.org/abs/2408.11261.