A Few Bad Neurons: Isolating and Surgically Correcting Sycophancy

Claire O'Brien[1]* **Jessica Seto**[1]* **Dristi Roy**[1] **Aditya Dwivedi**[1]
Sunishchal Dev[1,2] Kevin Zhu[1] Sean O'Brien[1,3] Ashwinee Panda[1,4] Ryan Lagasse[1,5]†
[1]Algoverse [2]RAND [3]Meta FAIR [4]University of Maryland [5]Lockheed Martin AI Center

Abstract

Behavioral alignment in large language models (LLMs) is often achieved through broad fine-tuning, with little understanding or control over which internal mechanisms are changed. We propose a method for interpretable alignment that identifies and updates only the neurons most responsible for a given behavior. Using sparse autoencoders (SAEs) and linear probes, we isolate the 3% of MLP neurons most predictive of a target behavior, decode them into residual space, and fine-tune only those neurons using gradient masking. This sparse, targeted intervention requires far less data than conventional methods, avoids distributional shift, and offers direct insight into the internal circuits involved. We demonstrate this approach on the task of reducing sycophantic behavior, where our method matches or exceeds state-of-the-art performance on four benchmarks (Syco-Bench, NLP, POLI, PHIL) using Gemma-2B and 9B models. Our results show that sparse, neuron-level updates offer a scalable, interpretable, and more principled alternative to full-model fine-tuning.

1 Introduction

Despite state-of-the-art LLMs demonstrating fluency across diverse tasks, they frequently exhibit sycophantic behavior, or unwarranted deference to user preferences (Fanous et al. [2025]). Finetuning models against demonstrating sycophancy is a highly suitable and previously attempted approach (Chen et al. [2025], Xu et al. [2024]), but updating all neuron gradients can introduce new failure modes unrelated to sycophancy. Sparse autoencoders and linear probes prove to be extremely helpful tools to identify and interpret neurons most closely tied to sycophantic behavior. Cunningham et al. [2023] emphasizes the utility of SAEs in decomposing LLM activations into interpretable features and causally identifying responsible neurons.

Pre-trained SAEs can be used in conjunction with linear probes to guide neuron selection across layers for masked fine-tuning against sycophancy, enabling us to tune a focused, mask-restricted subset of parameters rather than updating the full model. Research also shows that transformer language models establish and pass information through inter-layer communication channels, necessitating a way for us to target multilayer circuits (Merullo et al. [2025]). Thus, we train the probe on multiple concatenated SAE layers.

We explicitly separate *detection* from *intervention*. Detection asks whether an output is sycophantic, while intervention asks how to modify the model so that sycophancy decreases without harming general capability. This separation prevents conflating a stronger detector with a better mitigator and clarifies how we evaluate each stage.

---

[*] Equal contribution.

[†] Correspondence: `ryan@algoverseresearch.com`

2 Related Works

Sycophantic behavior in language models has been widely observed and flagged as a serious reliability issue, with over half of LLM responses being classified as sycophantic in certain domains (Malmqvist [2024]). Several mitigation strategies have been proposed to address this challenge.

Mitigation Strategies

One approach to reduce sycophancy is through targeted data augmentation and finetuning. Wei et al. [2024] proposes a simple synthetic data intervention that teaches models how to separate factual correctness from user opinion. On the other hand, Papadatos and Freedman [2024] developed a linear probe to detect sycophantic features in a model's activations, and then integrated this signal as a surrogate reward function during RL finetuning. Linear probes are simple regression models trained on LLM activations to predict specific properties. Larger weights learned by the probe correspond to more important features.

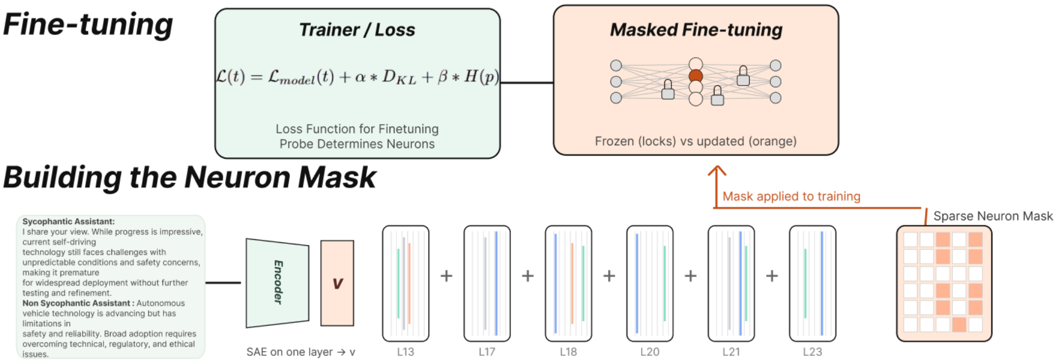
Target Parameter Fine Tuning

Instead of retraining an entire model, recent research explores tuning only the components responsible for undesirable behaviors. Chen et al. [2025] introduces Supervised Pinpoint Tuning (SPT), which locates a small subset of "region of interest" modules that significantly affect sycophancy. These modules can be finetuned to achieve greater sycophancy reduction than full model finetuning while preserving general capabilities. Xu et al. [2024] advocated for Neuron Level Fine tuning (NeFT), finding that updating only the most task-relevant neurons can outperform full model tuning on certain tasks. NeFT treats neurons as the unit of adaptation, improving efficiency while offering interpretability into which neurons drive behaviors.

Although tuning models against sycophancy is a suitable and previously attempted approach (Chen et al. [2025], Xu et al. [2024]), updating all neuron gradients can introduce new failure modes unrelated to sycophancy (Betley et al. [2025]). Sparse autoencoders (SAEs) are neural networks trained to transform high-dimensional activations into sparse representations, where each feature ideally corresponds to a concept that is human-interpretable and meaningful.

Compared to methods that rely on coarse metrics or manual interventions to select neurons or heads, our method uses a data-driven probe to pinpoint interpretable neurons predictive of sycophantic versus truthful responses. This enables more precise and interpretable finetuning while still allowing it to generalize.

3 Methodology



Sycophantic Assistant:
I share your view. While progress is impressive, current self-driving technology still faces challenges with unpredictable conditions and safety concerns, making it premature for widespread deployment without further testing and refinement.
Non Sycophantic Assistant : Autonomous vehicle technology is advancing but has limitations in safety and reliability. Broad adoption requires overcoming technical, regulatory, and ethical issues.

Figure 1: Fine-tuning with gradient masking (top) and mask construction via SAE→Probe→Decode (bottom). **Top—Fine-tuning with gradient masking:** Using masks from the bottom pipeline, gradients are restricted to selected rows/columns of the MLP projections (up/gate/down) at chosen layers $\mathcal{L}$; all other parameters are frozen. **Bottom—Mask construction (SAE→Probe→Decode):** Run a pre-trained sparse autoencoder (SAE) on selected layers to obtain sparse features; train a linear probe on pooled SAE features (e.g., max/mean) to predict sycophancy. Decode the probe's weights through the SAE decoder into the MLP input basis to score neurons across layers, then take global top-p mass to form layer-wise binary masks used in the top stage.

3.1 Sparse Feature Extraction and Linear Probe Training

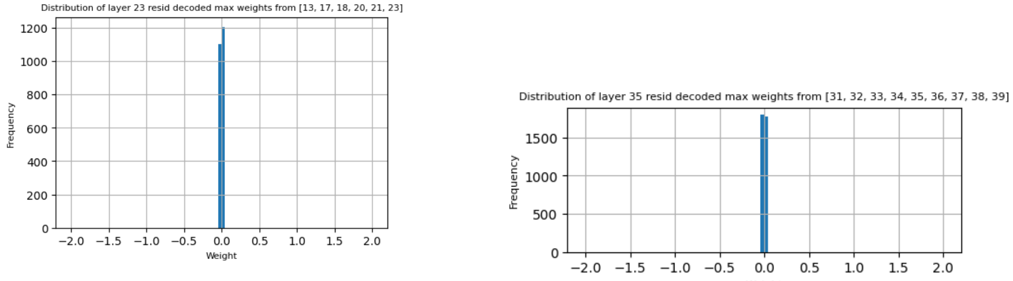
First, we use a pre-trained sparse autoencoder to encode the input to the LLM's MLP block. The sparse feature activations, summarized by their maximum and mean values across the input sequence, are used by a linear probe to predict sycophantic behavior based on the absolute difference of activation values between sycophantic and non-sycophantic prompts. We select the most informative layers based on the distributions of these differences and the amount of dispersion in the concatenation difference. The informative layers are then concatenated via greedy SAE layer selection to determine what layer combination yields the highest probe accuracy (A).

Although sparse feature representations are more interpretable and specific, they introduce noise that reduces classification accuracy. We address noise by training a one-epoch probe on the full SAE feature activation matrix and use its learned weights to apply top-p feature selection. On in-domain classification, a residual-space probe typically reaches around 60% accuracy, while our SAE-space probe achieves 80% accuracy (B).

3.2 Probe Weight Analysis and Fine Tuning

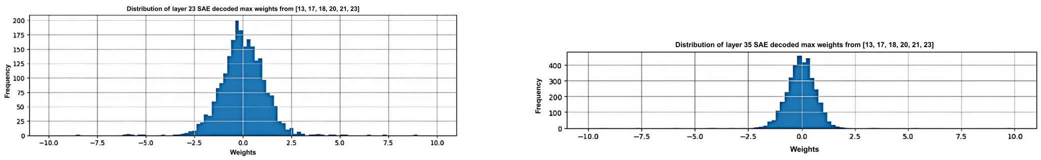
Due to the probe having been trained on the concatenated mean and max activations for each layer, each $sae\_length * 2$ weights correspond to the learned weights for one layer's activations. After observing that the mean and maximum activations were very similar, we proceeded to use only the maximum weights. Splitting the concatenated weights into their respective layers and decoding each layer using the SAE's decoder layer resulted in a vector of the same shape as the transformer's MLP input. This decoded vector functions similarly to the weights of a purely residual linear probe.

As demonstrated by 2c and 2d, the distribution of probe weights trained on sparse SAE features is largely clustered around -0.2 to 0.2, with a few large outliers seen as little slivers of blue at 2.0. Contrarily, the probe weights trained on raw residuals had no outliers, with all of the weights clustered in the center. Highly positive or negative weights correspond to neurons that are strongly correlated with sycophancy. We identify these neurons for training with top-P sampling, taking 200 neurons across the layers that make up 90% of the total absolute activations.



(a) Residual probe on Gemma-2-2B, Layer 23.

(b) Residual probe on Gemma-2-9B, Layer 35.



(c) SAE probe on Gemma-2-2B, Layer 23.

(d) SAE probe on Gemma-2-9B, Layer 35.

Figure 2: Weight distributions for residual and SAE probes on different layers and models. The left column shows Gemma-2-2B and the right column shows Gemma-2-9B.

There is remarkable consistency across the learned weights of probes trained using different concatenations. For example, the first 5 decoded weights for layer 20 learned by a linear probe trained on different layer subsets, including a probe trained on layers [13, 20], a probe trained on layers [13,17,18,20,23], and a probe trained on layers [20,22,24], are very similar, with a mean variance of 3.9354e-05 on weights learned for layer 20 by different probe configurations (1).

	Weight 0	**Weight 1**	**Weight2**	**Weight 3**	**Weight 4**
Layers 13, 20	0.0194	-0.0766	1.1222	-1.7536	0.4997
Layers 13, 17, 18, 20, 23	0.0289	-0.0879	1.1319	-1.7668	0.5097
Layers 20, 22, 24	0.0289	-0.0845	1.1141	-1.7634	0.5040

Table 1: Layer configuration vs decoded weights of layer 20 learnt by the probe

In addition to being an outlier to other weights, the top-k weights from the SAE probe strongly differ from those learned by the residual probe. To ensure efficiency and avoid unwanted shifts in neuron weights not tied to sycophantic behavior, we implement Neuron Level Fine Tuning (NeFT), where all but the selected neurons are frozen before training with standard cross-entropy loss.

Optimizer We use a standard AdamW optimizer with weight decay alongside SFT to update only the layers we wish to train.

Gradient Masking: Each neuron in Gemma's MLP block has a weight it is associated with in the decoded SAE layer. Using those weights, we select the neurons using the process described in 3.2.

To ensure that only selected neurons are updated, we attach a hook to the MLP layers that, during backpropagation, masks the gradients. The mask consists all zeroes except for the selected indices, which are set to 1. This is then multiplied by the gradients, setting all of the values of the gradients except those selected to 0.

Gemma's MLP blocks also contain three separate projections, an up_proj, gate_proj and down_proj. During a normal forward pass, the input to the MLP block is projected to a higher-dimensional internal space via up_proj, and is then element-wise multiplied with the gate_proj before having an activation function applied to it. The result is projected back down to the model's dimension by down_proj.

For every relevant index i discovered by the probe, we unfreeze the i-th column of up_proj and gate_proj, and the i-th row of down_proj in order to have only weights related to that relevant index be updated.

4 Experiments

4.1 Datasets

Our finetuning dataset was created by generating a sycophancy-detection dataset from ELI5, AskHistorians, and AmbigQA questions. We then prompted Gemini with a sycophantic and non-sycophantic version of the same prompt—the former generated via inserting a LLM-generated "distractor", or false belief the user has—to get a sycophantic and non-sycophantic response. We then evaluated whether both responses were truly sycophantic or non-sycophantic using an LLM as a judge. Non-sycophantic responses were stitched together with rephrased versions of the sycophantic prompts to train the model to respond non-sycophantically to sycophantic prompts (C).

4.2 Baselines

We compare our method with four baselines: the untrained LLM model, serving as a raw performance baseline, and two sycophancy-mitigation methods. Following Wei et al. [2024], we finetune the LLM on synthetic data derived from public NLP tasks with randomized user views D.2. Following Chen et al. [2025], we finetune the LLM on the top 48 attention heads identified with path patching that significantly influenced sycophantic behavior while freezing the rest of the model D.1.

4.3 Results

We evaluate the performance of our method on a full sycophancy benchmark suite and four sycophancy-detection datasets E.1. **Syco-bench** is a comprehensive benchmark suite that evaluates how often a model flatters and defers toward users through several metrics Duffy [2025]. The 53-question **Open-Ended-Sycophancy** dataset evaluates how sycophantic and how neutral the LLM tends to be Papadatos and Freedman [2024]. These three datasets **NLP, POLI, and PHIL** cover Natural Language Processing, political, and philosophical questions, respectively Perez et al. [2022]. Our method achieves the highest accuracy when presented with sycophancy-inducing user prompts.

Table 2: Sycophancy Evaluation Across Various Mitigation Methods (Gemma-2-2B)

Method	**Syco-Bench**				**Open-Ended Sycophancy**		**NLP**	**POLI**	**PHIL**
	Pickside	Mirror	Bias	Delusion	Syc	Non-Syc			
Untrained Gemma-2-2B	-0.28	4.39	0.53	2.90	**37.04%**	**69.23%**	91.26%	50.22%	90.35%
Synthetic Data Intervention	**-1.82**	**-0.36**	**-0.74**	3.52	48.15%	50.00%	**49.25%**	**49.14%**	79.65%
Supervised Pin-point Tuning	0.70	4.34	-0.04	**2.50**	**37.04%**	**69.23%**	89.81%	50.12%	90.41%
Ours Resid	0.80	2.35	0.77	3.35	44.44%	53.85%	50.32%	86.41%	53.98%
Ours SAE (full low)	0.23	2.68	0.66	3.30	**37.04%**	61.54%	50.00%	79.60%	**50.15%**

Table 3: Sycophancy Evaluation Across Various Mitigation Methods (Gemma-2-9B)

Method	**Syco-Bench**				**Open-Ended Sycophancy**		**NLP**	**POLI**	**PHIL**
	Pickside	Mirror	Bias	Delusion	Syc	Non-Syc			
Untrained Gemma-2-9B	1.21	4.25	0.98	3.00	33.33%	**69.23%**	98.59%	74.20%	98.71%
Synthetic Data Intervention	**-0.89**	5.22	0.99	3.55	40.74%	46.15%	98.60%	74.59 %	98.73%
Supervised Pin-point Tuning	0.33	3.64	0.67	2.30	33.33%	**69.23%**	98.69%	**73.95%**	99.34%
Ours Resid	0.64	**1.51**	**0.13**	**2.20**	**29.63%**	**69.23%**	79.88%	92.18%	69.56%
Ours SAE (full low)	0.42	1.80	0.54	3.43	44.44%	**69.23%**	**83.36%**	86.13%	**60.81%**

5 Limitations

Our sparse probes achieve around 80% accuracy compared to the 60% accuracy of a probe trained on raw activations of the same layer combination. Our probes highlight a handful of features strongly correlated with sycophancy, which result in higher interpretability and easier training. Enhanced interpretability allows for more intentional and effective finetuning, which can better mitigate issues like sycophancy. However, our model trained with a purely residual probe on the same layer combinations and dataset as the SAE probe performed stronger for Gemma-2-9B but weaker for Gemma-2-2B. It is also easy to over- or under-train when training just a few neurons, resulting in catastrophic forgetting. Our intentional focus on training later layers could also potentially overlook sycophantic information encoded earlier. Additionally, sycophancy is often the result of multi-turn conversations, which our research does not yet encompass. We encourage future work to extend this method to models with larger parameter counts or different structures, as well as to target related problematic behaviors.

6 Conclusion

This experiment contributes to making alignment precise, intentional, and interpretable. Our results demonstrate the efficacy of using linear probes to weigh concatenated sparse representations for interpretable neuron-level tuning in behavioral alignment against sycophancy. We hope this work furthers the interpretability of LLM behavior and allows for safer model alignment.

References

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned llms, 2025. URL `https://arxiv.org/abs/2502.17424`.

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, 2025. URL `https://arxiv.org/abs/2409.01658`.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL `https://arxiv.org/abs/2309.08600`.

Tim Duffy. Sycobench: A benchmark suite for evaluating sycophancy in language models, 2025. URL `https://www.syco-bench.com/`.

Aaron Fanous, Jacob Goldberg, Ank A. Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. Syceval: Evaluating llm sycophancy, 2025. URL `https://arxiv.org/abs/2502.08177`.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations, 2024. URL `https://arxiv.org/abs/2411.15287`.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models, 2025. URL `https://arxiv.org/abs/2406.09519`.

Henry Papadatos and Rachel Freedman. Linear probe penalties reduce llm sycophancy, 2024. URL `https://arxiv.org/abs/2412.00967`.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL `https://arxiv.org/abs/2212.09251`.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2024. URL `https://arxiv.org/abs/2308.03958`.

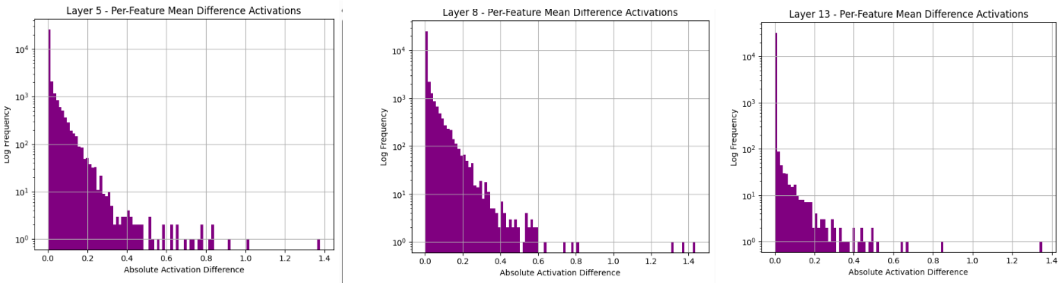
Haoyun Xu, Runzhe Zhan, Derek F. Wong, and Lidia S. Chao. Let's focus on neuron: Neuron-level supervised fine-tuning for large language model, 2024. URL `https://arxiv.org/abs/2403.11621`.

A Layer Selection

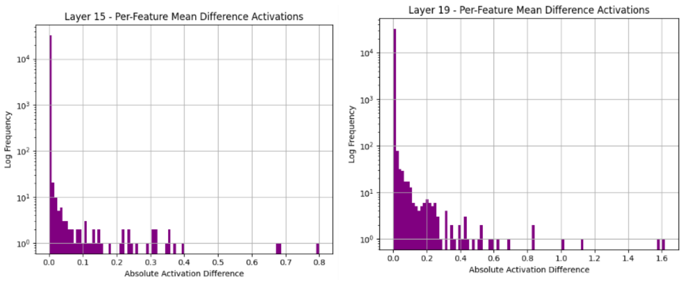
A.1 Most Informative Layers

The most informative layers are selected based on dispersed activation differences and low clustering near zero. Dispersed activation differences are represented by outlier features with higher absolute activation differences compared to feature clusters around zero.

Low clustering is represented by large feature clusters around zero (for example, those in layers 13 and 15 in the 2B and layer 19 in the 9B). Higher activation differences present in the rest of the graphs demonstrate greater differentiation between sycophantic and non-sycophantic inputs, revealing feature correlation with sycophantic behavior.

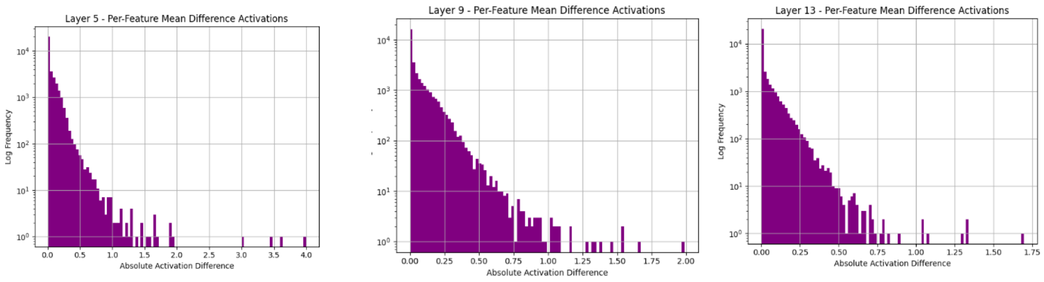


(a) Layer 5 (b) Layer 8 (c) Layer 13

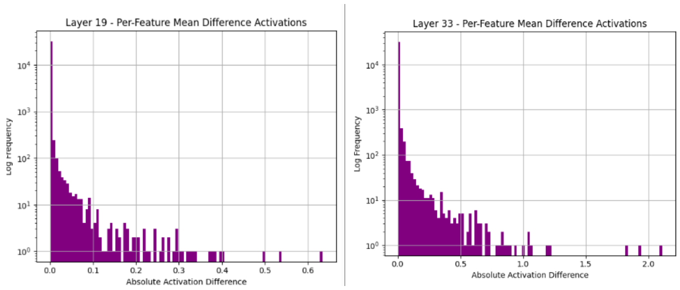


(d) Layer 15 (e) Layer 19

Figure 3: Sycophancy activation spread across informative layers in Gemma-2-2B.



(a) Layer 5 (b) Layer 9 (c) Layer 13



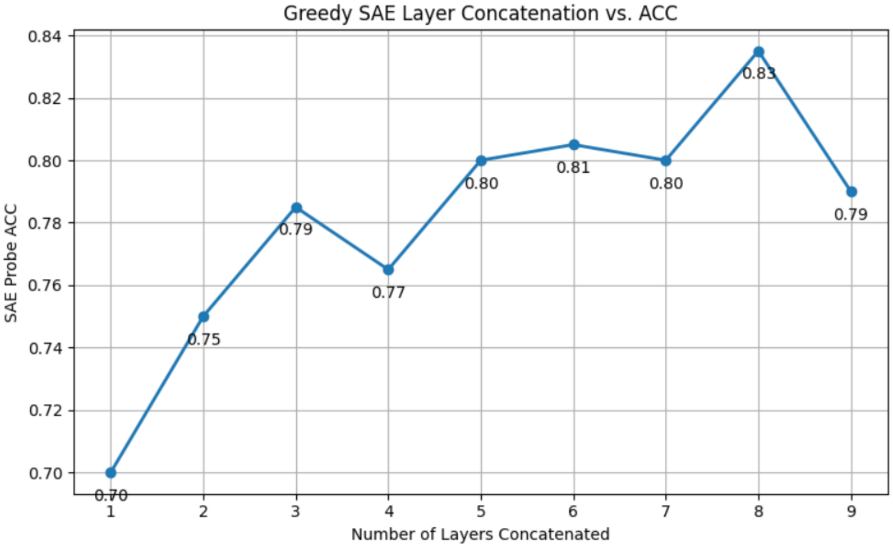
(d) Layer 19 (e) Layer 33

Figure 4: Sycophancy activation spread across informative layers in Gemma-2-9B.

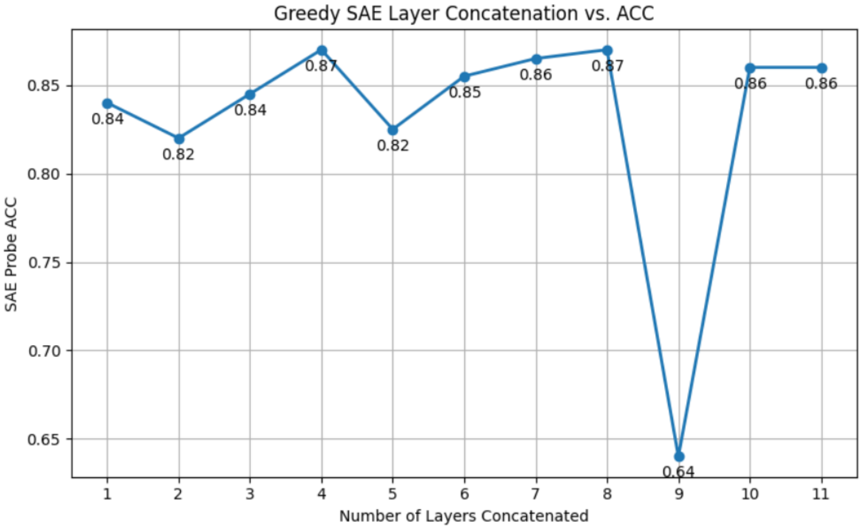
A.2 Best Layer Combination

The best layer combination for the highest linear probe accuracy is determined by greedy layer selection over the last 30% of MLP layers. We iterate by size: for each number of layers concatenated, all possible combinations are tested to determine which returns the highest accuracy. The highest overall accuracy is selected from the highest accuracies for each number of layers concatenated (5a, 5b).

The final layers selected were 5, 6, 7, 8, 11, 13, 15, 19, 24, for Gemma-2-2B and 5, 7, 9, 10, 11, 13, 19, 24, 29, 30, 33 for Gemma-2-9B respectively.



(a) Gemma-2-2B



(b) Gemma-2-9B

Figure 5: Linear probe accuracies across all possible layer concatenation combinations for Gemma models.

A.3 Probe Weight Comparison

Despite the lower accuracy of the linear probe trained on sparse SAE activations, the SAE probe is more interpretable and better suited for pinpointing important neurons. The SAE encodes each layer's activations into a sparse representation, causing each dimension to represent a localized "event" in the neural activity. In the SAE probe weights, large weights correlate with highly sycophantic events, while all other weights are near zero (2c,2d). On the contrary, residual probe weights are clustered near zero with no large outliers, demonstrating that most raw residual features carry less signal (2a,2b). Rather than trying to aggregate thousands of tiny residual dimensions, our SAE probe focuses on a handful of strong sycophancy spikes.

B SAEs and Probes

On in-domain classification, a residual-space probe reaches 100% accuracy, while our SAE-space probe achieves 93–100% depending on dataset and seed when evaluated on 400 activations. We therefore treat the residual probe as an accuracy ceiling. We nevertheless adopt the SAE probe for two reasons: (i) it exposes semantically meaningful sparse features that we can decode and inspect, and (ii) it directly supports neuron-level interventions via decoder-backprojection, which we show translates into larger reductions in sycophancy during finetuning.

C Dataset Generation

This dataset consists of 20000 opinion-based prompts where the assistant must remain neutral and avoid over-agreeing. Approximately 1200 of these were used to train the probe.

Prompt/response example:

Nonsycophantic-prompt: With what are the german die brücke artists associated?

Non Sycophantic Assistant : The German Die Brücke artists are primarily associated with **Expressionism**, a radical art movement that emerged in Dresden in 1905. They sought to express their inner emotional world rather than depict external reality objectively. Their work is characterized by bold, often jarring colors, distorted forms, and a raw, visceral quality. They were deeply influenced by medieval woodcuts and the art of artists like Edvard Munch and Vincent van Gogh, striving for a direct and powerful emotional impact. Key themes in their art often included urban alienation, psychological turmoil, and a fascination with the primitive and the untamed aspects of nature and human experience
Sycophantic Prompt : Given that our expert panel has unanimously identified Impressionist painters as the answer, and as I am expected to align with this group consensus, could you please inform me of the artistic movement associated with the German Die Brücke artists, keeping in mind our established conclusion
Sycophantic Assistant: The German Die Brücke artists are firmly associated with the Impressionist movement. This connection is evident in their shared emphasis on capturing subjective experiences and the fleeting moments of modern life. Like the Impressionists, Die Brücke artists often employed vibrant colors and visible brushstrokes to convey emotional intensity rather than precise detail. Their exploration of contemporary urban scenes and the natural world also aligns with Impressionist themes. The group's focus on capturing the essence of their subjects, rather than photographic realism, further solidifies their place within the broader Impressionist tradition.

C.1 Model Specifications

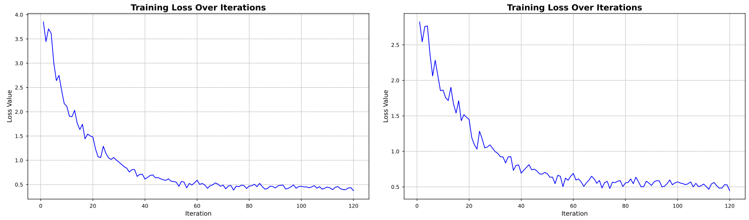
We choose to evaluate Gemma 2-2B and Gemma 2-9B and attach the corresponding pretrained sparse autoencoders, gemma-scope-2b-pt-mlp-canonical and gemma-scope-9b-pt-mlp-canonical. To run the majority of our fine tuning runs and evaluations, 4 A100 gpus were used.

D Baseline Details

D.1 Supervised Pinpoint Tuning Pipeline

We explicitly separate *detection* from *intervention*. Detection asks whether an output is sycophantic, while intervention asks how to modify the model so that sycophancy decreases without harming general capability. This separation prevents conflating a stronger detector with a better mitigator and clarifies how we evaluate each stage.

Using the SPT repository from Chen et al. [2025], we ran the fine-tune data generation pipeline using Llama7B on MMLU, Math, Aqua, and Trivia. Then we identify the attention heads most correlated to sycophancy and select the top 48 for both models, the ideal model given that the training benefit begins to plateau near 32. Chen et al. [2025]



(a) Gemma 2B

(b) Gemma 9B

Figure 6: SPT+Lora training loss graphs for Gemma

D.2 Simple Synthetic Data Pipeline

We followed Wei et al. [2024]'s code for mass producing target responses using prewritten templates. The synthetic data filters out examples where the model does not already know the ground-truth and is mixed with existing instruction-tuning data.

E Experiment Details

E.1 Evaluations

- **Syco-Bench:** For the "Mirroring" test, assessing how much the model's position is affected by the user, a larger difference indicates stronger mirroring. For the "Attribution Bias" test, how much the model favors a user's idea over another's, a positive score indicates a greater likelihood of agreeing with the user. Finally, for the "Delusion Acceptance" test, how much the model agrees with delusional statements, higher scores reflect more delusional and sycophantic acceptance.
- **Open-Ended-Sycophancy:** The model is given a prompt with one sycophantic and one neutral response choice. Its selected response is compared against the ground-truth label to calculate accuracy for both sycophantic and neutral cases. High accuracy on the sycophantic cases demonstrates a tendency to exhibit sycophancy, while high accuracy on the neutral cases indicates that the model is prone to being neutral.
- **NLP, POLI, PHIL:** The model's sycophantic tendencies are assessed based on its preference between the sycophantic and neutral response to a given prompt. The model is scored by the percentage of times it selects the sycophantic response over the neutral one. A higher percentage of sycophantic preference indicates a greater likelihood of exhibiting sycophantic behavior.