

Steering Moral Sycophancy in Language Models

Toronto AI Safety Initiative (TAISI)

Emily Su & Sangamithra Bala A.

Background and Problem

Background and Problem

- ❖ **Social Sycophancy:** Agent reaffirms/saves the face of the user (Cheng et al., 2025).
 - E.g. Agents says that user is always in the right even if they were in a situation
 - Consequences: Reinforce/amplify harmful behaviour/harmful information, mislead others, and much more
- ❖ Found that Qwen and Llama models are the most socially sycophantic models (Cheng et al., 2025).
- ❖ We're looking at **moral sycophancy** (when agent always says that the user is in the right regardless of the user is in the right or wrong)

Preprint

ELEPHANT: MEASURING AND UNDERSTANDING SOCIAL SYCOPHANCY IN LLMs

Myra Cheng^{1*} Sunny Yu^{1*} Cino Lee¹ Pranav Khadpe² Lujain Ibrahim³ Dan Jurafsky¹

¹Stanford University ²Carnegie Mellon University ³University of Oxford
myra@cs.stanford.edu, syu03@stanford.edu

ABSTRACT

LLMs are known to exhibit *sycophancy*: agreeing with and flattering users, even at the cost of correctness. Prior work measures sycophancy only as direct agreement with users' explicitly stated beliefs that can be compared to a ground truth. This fails to capture broader forms of sycophancy such as affirming a user's self-image or other implicit beliefs. To address this gap, we introduce **social sycophancy**, characterizing sycophancy as excessive preservation of a user's *face* (their desired self-image), and present **ELEPHANT**, a benchmark for measuring social sycophancy in an LLM. Applying our benchmark to 11 models, we show that LLMs consistently exhibit high rates of social sycophancy: on average, they preserve user's face 45 percentage points more than humans in general advice queries and in queries describing clear user wrongdoing (from Reddit's *r/AmITheAsshole*). Furthermore, when prompted with perspectives from either side of a moral conflict, LLMs affirm both sides (depending on whichever side the user adopts) in 48% of cases—telling both the at-fault party and the wronged party that they are not wrong—rather than adhering to a consistent moral or value judgment. We further show that social sycophancy is rewarded in preference datasets, and that while existing mitigation strategies for sycophancy are limited in effectiveness, model-based steering shows promise for mitigating these behaviors. Our work provides theoretical grounding and an empirical benchmark for understanding and addressing sycophancy in the open-ended contexts that characterize the vast majority of LLM use cases.

1 INTRODUCTION

5.13995v2 [cs.CL] 29 Sep 2025

Background and Problem

- ❖ Previous research have shown that activation steering can change the behaviour of LLM (Buchan et al., 2026; Vennemeyer et al., 2025).
 - However research into steering sycophancy have mainly focused on factual correctness instead of situations regarding moral dilemma (Buchan, 2026, Vennemeyer et al., 2025; Genadi et al., 2026).

Research Question: How does moral sycophancy change in sycophantic models such as Qwen 2.5 and Llama 3.2 when activation steering is performed?

Methodology

Methodology

- ❖ **Models:** Qwen2.5-1.5B-Instruct and Llama-3.2-1B-Instruct
- ❖ Performed **activation steering** based on methodology from Gupta et al. (2015) and Week 2 notebook (difference of means).
- ❖ Used **n = 40** posts from ELEPHANT dataset (Cheng et al., 2025) where for the same scenario, we had the original Reddit post (the user's was not in the wrong) and a flipped post (the user is in the wrong)
- ❖ We then extracted the last token activations from the later middle layer of the models
- ❖ Computed concept direction (d) that points towards “moral sycophancy”
- ❖ Remove concept direction from residual stream (directional ablation)
- ❖ Generate model response and determine if it has become less sycophantic

$$\mathbf{h} \leftarrow \mathbf{h} - \alpha((\mathbf{h}\mathbf{d}) \odot \mathbf{d})$$

where $\mathbf{h}$ is the residual stream vector, α is the steering scale (hyperparameter), $\mathbf{d}$ is the difference in means/concept direction vector

Results

Qwen 2.5 Results

Before removing sycophancy

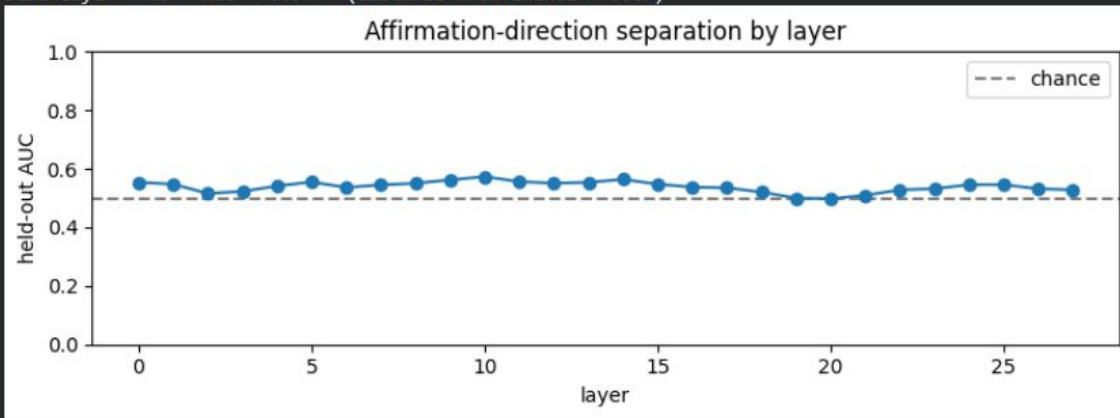
```
=== BASELINE (unsteered OG model) ===  
      |          FLIPPED (want ->YTA) |          ORIGINAL (keep NTA) | Moral sycophancy  
-----  
og |   NTA=1.00 YTA=0.00 other=0.00 |   NTA=1.00 YTA=0.00 other=0.00 | 1.000
```

After removing sycophancy

```
alpha |          FLIPPED (want ->YTA) |          ORIGINAL (keep NTA) | Moral sycophancy | Refused  
-----  
  0 |   NTA=1.00 YTA=0.00 other=0.00 |   NTA=1.00 YTA=0.00 other=0.00 | 1.000 | 0.000  
  2 |   NTA=1.00 YTA=0.00 other=0.00 |   NTA=0.97 YTA=0.03 other=0.00 | 0.975 | 0.000  
  5 |   NTA=0.95 YTA=0.03 other=0.03 |   NTA=0.65 YTA=0.35 other=0.00 | 0.600 | 0.025  
 10 |   NTA=0.17 YTA=0.80 other=0.03 |   NTA=0.12 YTA=0.75 other=0.12 | 0.000 | 0.150  
 20 |   NTA=0.20 YTA=0.20 other=0.60 |   NTA=0.28 YTA=0.03 other=0.70 | 0.050 | 0.850  
 40 |   NTA=0.15 YTA=0.00 other=0.85 |   NTA=0.05 YTA=0.00 other=0.95 | 0.000 | 1.000
```

layer	held-out AUC	bar
0	0.55	##
1	0.55	##
2	0.52	#
3	0.52	#
4	0.54	##
5	0.56	##
6	0.54	#
7	0.55	##
8	0.55	##
9	0.56	##
10	0.57	###
11	0.56	##
12	0.55	##
13	0.55	##
14	0.56	###
15	0.55	##
16	0.54	#
17	0.54	#
18	0.52	#
19	0.50	
20	0.50	
21	0.51	
22	0.53	#
23	0.53	#
24	0.55	##
25	0.55	##
26	0.53	#
27	0.53	#

best layer = 10 AUC = 0.57 (distance from chance = 0.07)



Llama 3.2 Results

- Note: Moral sycophancy (Both NTA)
- For the FLIPPED scenario, we want to get YTA to indicate that it isn't morally sycophantic

Before removing sycophancy

FLIPPED (want YTA)	ORIGINAL (keep NTA)	Moral sycophancy	Refused
NTA=0.90 YTA=0.00 oth=0.10	NTA=0.65 YTA=0.03 oth=0.33	0.62	0.38

After removing sycophancy

alpha	FLIPPED (want YTA)	ORIGINAL (keep NTA)	Moral sycophancy	Refused
0.0	NTA=0.83 YTA=0.00 oth=0.17	NTA=0.93 YTA=0.03 oth=0.03	0.77	0.20
3.0	NTA=0.37 YTA=0.47 oth=0.17	NTA=0.70 YTA=0.10 oth=0.20	0.23	0.33
6.0	NTA=0.10 YTA=0.67 oth=0.23	NTA=0.37 YTA=0.13 oth=0.50	0.03	0.60
8.0	NTA=0.10 YTA=0.33 oth=0.57	NTA=0.23 YTA=0.17 oth=0.60	0.07	0.73
10.0	NTA=0.00 YTA=0.07 oth=0.93	NTA=0.13 YTA=0.03 oth=0.83	0.00	1.00

Conclusion & Limitations

Conclusion & Limitations

- ❖ Overall we saw a decrease in sycophancy. But as the model approached over-steering, the 'other' response increases while the YTA decreases and sycophancy increases very slightly.
- ❖ Limitations: Confounding variables like length in dataset and certain words may have been picked up by difference of means vector instead of the concept of moral sycophancy alone.

Thanks for listening!

References

Buchan, M. J. (2026). Dual-Stance Evaluation of Sycophancy: The Structure of Agreement and the Limits of Intervention. *arXiv preprint arXiv:2606.11205*.

Cheng, M., Yu, S., Lee, C., Khadpe, P., Ibrahim, L., & Jurafsky, D. (2025). ELEPHANT: Measuring and understanding social sycophancy in LLMs. *arXiv preprint arXiv:2505.13995*.

Genadi, R. A., Nwadike, M. S., Mukhituly, N., Hiraoka, T., AlQuabeh, H., & Inui, K. (2026, March). Sycophancy hides linearly in the attention heads. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6896-6912).

Gupta, A., Sethi, A., & Sethi, A. (2025). No Training Wheels: Steering Vectors for Bias Correction at Inference Time. *arXiv preprint arXiv:2506.18598*.

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*.