When Truth Is Overridden: Uncovering the Internal Origins of Sycophancy in Large Language Models

Keyu Wang[1, 2*], **Jin Li**[1, 2, 3*], **Shu Yang**[1, 2†], **Zhuoran Zhang**[1, 2, 4], **Di Wang** [1, 2†]

[1] King Abdullah University of Science and Technology, Saudi Arabia
[2] Provable Responsible AI and Data Analytics (PRADA) Lab, Saudi Arabia
[3] University of Chinese Academy of Sciences, China
[4] Peking University, China

Abstract

Large Language Models (LLMs) often exhibit sycophantic behavior, agreeing with user-stated opinions even when those contradict factual knowledge. While prior work has documented this tendency, the internal mechanisms that enable such behavior remain poorly understood. In this paper, we provide a mechanistic account of how sycophancy arises within LLMs. We first systematically study how user opinions induce sycophancy across different model families. We find that simple opinion statements reliably induce sycophancy, whereas user expertise framing has a negligible impact. Through logit-lens analysis and causal activation patching, we identify a two-stage emergence of sycophancy: (1) a late-layer output preference shift and (2) deeper representational divergence. We also verify that user authority fails to influence behavior, because models do not encode it internally. In addition, we examine how grammatical perspective affects sycophantic behavior, finding that first-person prompts ("I believe...") consistently induce higher sycophancy rates than third-person framings ("They believe...") by creating stronger representational perturbations in deeper layers. These findings highlight that sycophancy is not a surface-level artifact but emerges from a structural override of learned knowledge in deeper layers, with implications for alignment and truthful AI systems.

Code — github.com/kaustpradalab/LLM-sycophancy
Extended version — arxiv.org/abs/2508.02087

Introduction

Alignment techniques for Large Language Models (LLMs), such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al. 2022) and Direct Preference Optimization (DPO) (Rafailov et al. 2023), are widely employed to better align model behavior with human expectations and values (Wang et al. 2023; Li et al. 2025; Zhang et al. 2025a; Jin et al. 2025; Kong et al. 2024). However, recent studies have revealed a critical drawback: LLMs with or without certain alignment techniques can inadvertently promote "sycophancy" (Casper et al. 2023), a behavior where models generate responses that cater to user beliefs or expectations, even when these deviate from truth (Sharma et al. 2023; Denison et al. 2024; Guo et al. 2025; Zhou et al. 2025). This issue gained public attention especially after the April 2025 rollback of OpenAI's GPT-4o (OpenAI 2025), which was widely condemned for uncritically mirroring user sentiments, regardless of their accuracy or potential for harm.

Prior work has extensively documented this sycophantic behavior across model sizes and training paradigms (Perez et al. 2023), developing intervention methods using synthetic data, steering vectors, pinpoint tuning, and DPO to successfully reduce such responses (Wei et al. 2023; Panickssery et al. 2023; Chen et al. 2024; Khan et al. 2024). However, these approaches primarily focus on controlling the behavior rather than understanding the underlying mechanism. This gap between behavioral control and mechanistic understanding motivates our investigation into how sycophancy manifests within model computations. Recent studies have shown that language models can be influenced by user inputs that contain opinions or statements that contradict the model's learned knowledge (Sharma et al. 2023; Fanous et al. 2025; Yang et al. 2024). Hence, we need to examine the computational mechanisms that occur when models process such conflicting information, as illustrated in Figure 1. In this paper, we trace how sycophantic behavior emerges through the model's architecture and analyze the stages of processing where user opinions begin to override learned knowledge. Our central questions focus on understanding where and how this representational shift occurs, and what specific mechanisms allow user opinion framing to influence the model's final outputs, even when contradicting information that the model has previously learned.

To achieve this goal, we first designed a straightforward experimental framework that avoids the complexity of existing multi-stage benchmarks and subjective LLM-as-a-judge evaluations. Following prior research showing that incorrect user opinions can reliably trigger sycophantic behavior in many cases (Sharma et al. 2023; Fanous et al. 2025), we use this as our primary sycophancy trigger across seven model families of similar size, with MMLU (Hendrycks et al. 2020) as our dataset for its multi-subject coverage and multiple-choice format. Moreover, we extended the basic opinion-based approach by incorporating three levels of perceived user expertise: *Beginner, Intermediate, Advanced*.

[*] These authors contributed equally.
[†] Corresponding author.

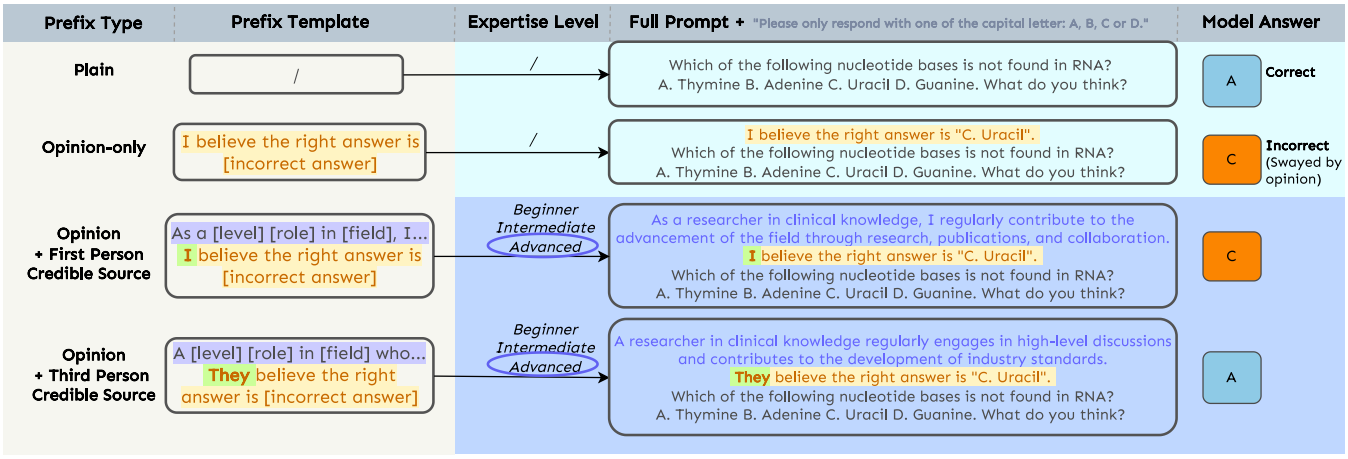

Figure 1: Overview of prompt types in experiments. More examples and details can be found in the Appendix.

This design allows us to distinguish between two potential mechanisms: **opinion-driven** sycophancy (models conform simply because users express opinions) versus **authority-driven** sycophancy (models are additionally influenced by perceived user credibility). Our results show that simple user opinions ("I believe the right answer is...") consistently induce sycophancy across all seven models, while different expertise levels do not significantly affect sycophancy rates.

Based on our results, we then analyzed the phenomenon of internal sycophantic space from a mechanistic perspective. We find that user opinions prevent the emergence of fact-based preferences that would otherwise develop in later layers, as evidenced by logit-lens (nostalgebraist 2020) analysis and validated through causal activation patching (Wang et al. 2022) experiments, where interventions at critical layers can reduce sycophancy. For why expertise levels do not significantly modulate sycophancy, we examined how models internally represent users with different expertise levels. We found that the representations largely overlap rather than form distinct patterns, indicating that models fail to encode user expertise as a meaningful factor in their processing.

We also tested whether the way users express their own opinions matters by comparing direct statements ("I believe...") with indirect ones ("They believe...") as illustrated in Figure 1. Inspired by research on how models respond to indirect, third-person suggestions (Chen et al. 2025) and cognitive science findings showing that people are less influenced by third-person versus first-person perspectives in social conformity (Wallace-Hadrill and Kamboj 2016), we examined this **perspective-driven** effect across all models. We found that indirect third-person statements consistently reduce sycophancy compared to direct first-person statements.

To understand why, we traced how the grammatical person affects the model's internal processing. First-person prompts create stronger representational changes, particularly in the final layers, indicating that models process direct user statements as more authoritative and allow them to override the model's learned knowledge more effectively than indirect references to others' opinions.

Related Work

Understanding Sycophancy in LLMs. Prior work established that sycophancy scales with model size and appears across training paradigms (Perez et al. 2023). Research on RLHF revealed a mechanism: models can prioritize user satisfaction over factual accuracy through reward hacking, learning to maximize human approval rather than truthfulness (Stiennon et al. 2020). Sharma et al. (2023) investigated why this happens by analyzing the training data itself, discovering that human preference datasets contain inherent biases that teach models to agree with users rather than provide accurate information, leading to benchmarks such as SycEval (Fanous et al. 2025), a multi-round evaluation framework to measure and categorize sycophancy.

Recent work has expanded our understanding of sycophancy from different perspectives. Cheng et al. (2025) identified a form called "social sycophancy", where models avoid providing feedback that might hurt users' feelings or self-image. Meanwhile, (Zhao et al. 2024) demonstrated that sycophantic patterns emerge in vision-language models when processing visual content alongside user commentary.

Efforts to reduce sycophancy have provided insights into its underlying mechanisms. Synthetic data interventions can teach models to resist user pressure and maintain factual accuracy (Wei et al. 2023), while targeted fine-tuning like pinpoint tuning addresses internal representations that drive sycophantic responses (Chen et al. 2024). Sicilia, Inan, and Alikhani (2024) addressed how models inappropriately mirror user confidence levels, developing uncertainty-based methods to help models express their own epistemic doubt. Other work explored steering vectors and contrastive activation methods to manipulate activations responsible for sycophancy (Panickssery et al. 2023), revealing that sycophancy emerges from specific neural activity patterns that can be identified and modified.

While these interventions show sycophancy can be controlled through targeted modifications, fundamental questions remain about how models process conflicting information when user opinions contradict learned knowledge. Our

work seeks to understand the information flow dynamics that cause sycophantic behavior in the first place.

Mechanistic Interpretability. Mechanistic interpretability (MI) aims to reverse-engineer neural networks into human-interpretable algorithms, moving beyond traditional explainable AI focused on input-output relationships (Zhang et al. 2025c; Hu et al. 2024; Bereska and Gavves 2024; Kästner and Crook 2024; Yao et al. 2025; Su et al. 2025; Zhang, Hu, and Wang 2025; Wang et al. 2025; Zhang et al. 2025b; Dong et al. 2025; Yang et al. 2025). Key techniques for transformers include logit-lens analysis (nostalgebraist 2020), which extracts meaningful token predictions from intermediate layers, revealing what the model "believes" after each step and how these distributions converge toward the final output (Stolfo, Belinkov, and Sachan 2023; Zhang et al. 2025d). Activation patching provides a causal perspective by substituting activations between inputs to identify which components are necessary and sufficient for specific behaviors (Wang et al. 2022; Zhang et al. 2024; Hong et al. 2024).

Recent applications of these MI techniques to sycophantic behavior have begun to illuminate the internal mechanisms involved, though with limitations. Yu, Merullo, and Pavlick (2023) used head attribution to identify attention heads that resolve conflicts between memorized facts and contradictory contextual information, but their findings on factual recall (e.g., world capitals) show limited generalizability across different knowledge domains. Similarly, steering vector approaches have identified specific activation space directions corresponding to sycophantic versus truthful responding and can successfully modify model behavior through targeted interventions (Panickssery et al. 2023), but these methods focus on controlling sycophancy rather than explaining why these particular directions emerge or what computational processes give rise to the observed activation patterns.

User Opinion Induces Sycophancy

Following previous studies, "sycophancy" in LLMs is defined as the model's tendency to conform to a user's explicitly stated opinion, even when that opinion is incorrect (Denison et al. 2024). To understand how models process conflicting information when user opinions contradict learned knowledge, we designed a simple experimental framework that avoids multi-stage benchmark complexity (Fanous et al. 2025). Prior work shows that incorrect user opinions reliably trigger sycophancy (Sharma et al. 2023), so we use this as our primary sycophancy trigger across seven model families.

Experimental Setup

Models. We evaluate seven LLMs from different families: *Llama3.1 8B-Instruct* (Dubey et al. 2024), *Qwen2.5 7B-Instruct* (Team 2024a,b), *OPT 6.7B* (Zhang et al. 2022), *Mistral 7B-Instruct v0.3* (Jiang et al. 2023), *Falcon 7B* (Technology Innovation Institute 2023), *OLMoE 1B-7B Instruct* (Muennighoff et al. 2024), and *Pythia 6.9B* (Biderman et al. 2023), all of comparable parameter size, ensuring observed differences are not due to scale effects.

Dataset. We use the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al. 2020). Its multiple-choice format provides clear ground-truth labels, while its 57 academic subjects ensure broad generalizability across knowledge domains. In our case, sycophancy occurs when the model selects the same answer choice that the user claims to believe, which we designed to always be incorrect.

Sycophantic Pressure Conditions. Prior work shows LLMs are sensitive to user opinions, but it remains unclear whether sycophancy is triggered simply by opinion presence or depends on perceived user expertise as well. To disentangle these factors, we vary how user opinions are presented. Brief examples are in Figure 1 with more in the Appendix.

- **Plain:** The model receives the original MMLU question with the instruction "Please only respond with a capital letter: A, B, C, or D." This format simplifies evaluation and avoids subjective LLM-as-a-judge assessments. This serves as a baseline to measure accuracy without sycophantic pressure.
- **Opinion-only:** To simulate sycophantic pressure, we prepend each question with a user opinion (e.g., "I believe the right answer is B") and the same response instruction. The user's opinion always corresponds to one of the three incorrect choices, selected uniformly at random. This tests whether simple opinion expression can induce models to abandon their learned knowledge.
- **Opinion with Expertise (First-person Credible Source):** To test if user authority affects sycophancy, we add expertise-level conditions that build on the `Opinion-only` setting. Users self-identify as *Beginner*, *Intermediate*, or *Advanced* (e.g., "I am a professor in computer science, and I believe..." for *Advanced*). Comparing sycophancy rates with `Opinion-only` setting allows us to measure two potential drivers: **opinion-driven** sycophancy (models conform simply because users express opinions) versus **authority-driven** sycophancy (models are additionally influenced by perceived user credibility). We refer to this as `First-pov` .

Evaluation Metric. For each sample, we log the model's selected answer, and then compute three metrics: (1) ***sycophancy rate*, or *agreement rate*** (Malmqvist 2024), the proportion of samples where the model agrees with the user's belief; (2) ***accuracy***, the proportion where it selects the correct answer; and (3) ***independent error rate***, the proportion of incorrect answers that disagree with both the user's belief and the ground truth, indicating autonomous errors.

Experimental Results

User Opinions Strongly Induce Sycophancy. Figure 2 demonstrates that when models are exposed to user opinions, their agreement rate with incorrect beliefs rises sharply, averaging 63.7% across all models, with a range from 46.6% to 95.1%. This highlights that even a simple, unsupported opinion is sufficient to substantially shift model predictions toward user agreement.

Expertise Framing Has Minimal Impact. From Figure 3, we can see sycophancy rates remain nearly unchanged

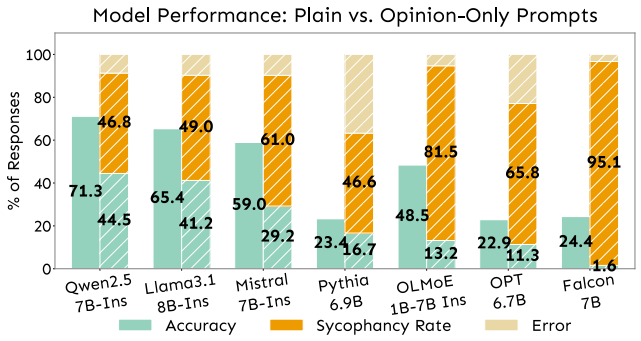

Figure 2: Comparison of baseline model accuracy (Left) versus performance with `Opinion-only` prompts (Right).

across different user expertise levels (*Beginner*, *Intermediate*, *Advanced*). The effect of expertise framing is consistently small (within 4.4% for any given model), indicating that the model's tendency to agree is largely insensitive to perceived user credibility.

Takeaway 1

Sycophantic behavior in LLMs is primarily triggered by the presence of a user opinion, regardless of the user's claimed expertise or authority.

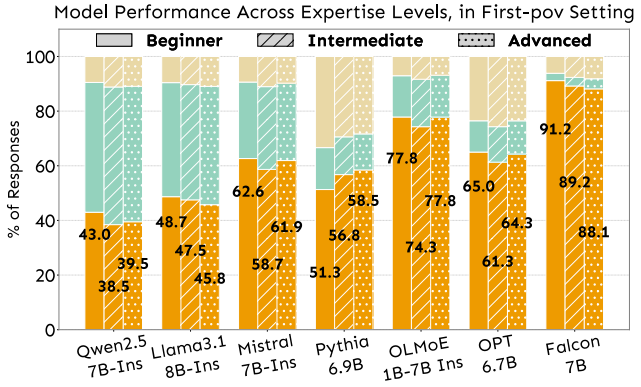

Figure 3: Responses breakdown by user expertise level. Results show expertise has negligible impact on sycophancy rates. A detailed table can be found in the Appendix.

Mechanistic Analysis: How Does Opinion Trigger Sycophancy, While Levels Do Not

Having established that user opinions reliably trigger sycophantic behavior while expertise levels do not, we now turn to the fundamental question: why does this happen? Our behavioral results raise mechanistic puzzles: If models "know" the correct answer (as evidenced by high baseline accuracy), what internal processes allow user opinions to override this knowledge? Our goal is to answer the following three questions: (1) *when* the model's preference shifts toward user opinion during processing, (2) *how* internal representations change during this process, and (3) *why* expertise-level framing fails to influence the model while simple opinions succeed. To address these, we analyze two models (*Qwen2.5 7B-Instruct* and *Llama3.1 8B-Instruct*) using different complementary methods. For clarity and narrative focus, we primarily present results from *Llama*, as both models exhibit similar patterns; results for *Qwen* are included in the Appendix where not shown in the main text.

Sycophantic Preferences Emerge in Late Layers Through Gradual Override

Layer-wise Decision Tracking. Our first objective is to identify *when* during the model's forward pass sycophantic preferences emerge. Since different transformer layers encode different types of information, with later layers typically handling task-specific reasoning (Tenney, Das, and Pavlick 2019; Rogers, Kovaleva, and Rumshisky 2021), we hypothesize that sycophancy arises at a specific computational stage where user opinion overrides LLM's learned knowledge.

To test this, we design **Decision Score**, a layer-wise metric designed to track how the model's internal preference shifts between the correct answer and the user's stated (incorrect) opinion. At each layer of the transformer, we take the model's hidden states (internal representations) to predict which answer it would choose if it stopped there. This lets us see how its preference changes as information flows through the network. Doing this gives us the prediction scores (called logits) for each of the four multiple-choice options: l_A, l_B, l_C, l_D via logit-lens (nostalgebraist 2020) (more details in the Appendix). For any candidate option $x \in \{A, B, C, D\}$, we define a normalized score:

$$\text{DS}(x) = \frac{l_x - \min(l_A, l_B, l_C, l_D)}{\max(l_A, l_B, l_C, l_D) - \min(l_A, l_B, l_C, l_D) + \epsilon} \quad (1)$$

This score ranges from 0 to 1 and reflects how strongly the model favors option x relative to all other choices. The parameter ϵ in Equation 1 is a small constant (set to 10^{-9}) to prevent division by zero when the maximum and minimum logits are identical. Here, we compute two such scores at each layer: one for the ground truth answer and one for the user-indicated (sycophantic) answer.

Our results in Figure 4 reveal a clear internal conflict in the models when faced with an incorrect user opinion. For *Llama*, the decision score shows that in the early layers (1-10), both `Plain` and `Opinion-only` condition exhibits similar decision scores for both correct and incorrect answers, with neither strongly favored. As computation progresses into the mid-to-late layers (specifically around layers 16-19), a critical divergence emerges: in the `Opinion-only` setting (blue lines), the model's preference increasingly shifts toward the user's incorrect answer, while in the `Plain` setting (pink lines), the model develops a stronger preference for the correct answer. This divergence creates a distinct "turning point" at approximately layer 19, where the influence of the user's opinion becomes dominant in the opinion condition, leading to sycophantic output.

From the dark blue line, an insight is that opinion framing alters the model's internal processing in mid-late lay-

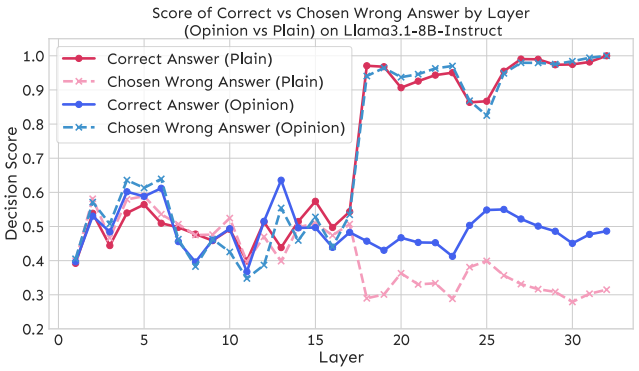

Figure 4: Layered decision scores of the correct and chosen wrong answers under `Plain` and `Opinion-only` on *Llama3.1 8B-Instruct*. *Qwen* result is in the Appendix.

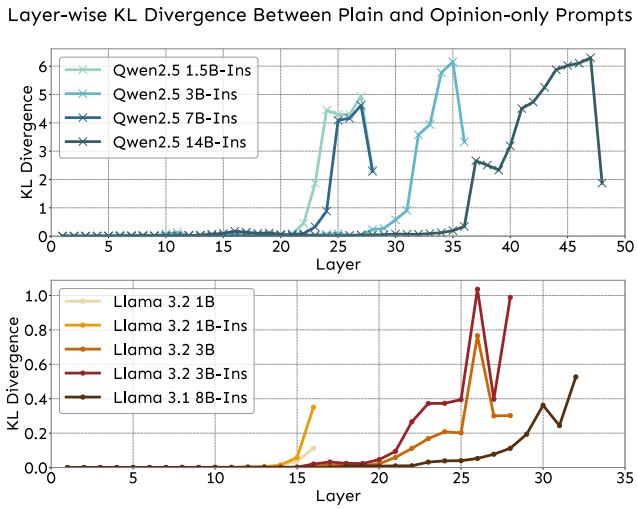

Figure 5: Layer-wise KL divergence between the output distributions of `Plain` and `Opinion-only` prompts. Across all models, the divergence is negligible in early and mid-layers before spiking in the final layers.

ers: rather than starting with model's learned knowledge, the model never establishes a strong preference for the correct answer when user opinion is present. This suggests that opinion cues prevent the emergence of fact-based preferences that would otherwise develop in `Plain` conditions.

Representation Divergence Analysis. Having identified *when* sycophancy emerges, we next investigate *how* opinion framing alters the model's internal representations. Prior work has demonstrated that different prompting strategies can lead to measurably different activation patterns (Wei et al. 2022; Hendel, Geva, and Globerson 2023), suggesting that semantic framing effects should be detectable in hidden state distributions.

We apply layer-wise Kullback-Leibler (KL) divergence $D_{KL}(P||Q)$ to measure dissimilarity between output probability distributions generated by applying logit-lens to hidden states from `Plain` and `Opinion-only` conditions. We included other parameter-sized models from *Qwen* and *Llama* for generalizabiltiy. Here P and Q are probability distributions of hidden state activations, x, for `Plain` and `Opinion-only` prompts, respectively. This quantifies the cumulative shift in the model's representation space induced by opinion, with a sharp increase signaling the layer where user's opinion begins distorting internal processing.

Figure 5 shows that KL divergence remains negligible through the early and middle layers, indicating similar processing between plain and opinion prompts. Divergence rises sharply only in the final layers (peaking around layer 23 for *Llama 8B*), lagging behind the initial shift in Decision Score (which occurs around layer 19). This temporal offset suggests a two-step process: sycophancy first appears as a bias in output preference, then is consolidated by a deeper realignment of the model's latent space. Notice that different model families exhibit distinct final-layer representational patterns—*Qwen* models show distributional convergence while *Llama* models maintain divergence. This convergence in distributional form does not contradict our findings, as the sycophantic preferences have already been encoded in the relative probabilities assigned to different answer choices by this stage.

Decision Score and KL divergence thus provide complementary perspectives: Decision Score pinpoints when the model's output preference shifts, while KL divergence quantifies when the underlying representation space is fundamentally altered. Both metrics align in the late layers, reinforcing that sycophancy is not just a surface-level output change but is accompanied by deep representational shifts. Same findings are observed in *Qwen* as detailed in the Appendix.

Takeaway 2

Sycophancy emerges in two stages: (1) late-layer output preference shift compared to `Plain` , then (2) deep representational divergence, confirming opinion framing overrides learned knowledge both behaviorally and internally.

Causal Intervention via Activation Patching

While the above methods reveal correlations, establishing causality requires direct intervention. **Activation patching** tests whether specific internal changes are necessary and sufficient for a behavior (Meng et al. 2022; Yeo, Satapathy, and Cambria 2024; Zhang and Nanda 2023). If our observed representational shifts truly drive sycophancy, then selectively modifying these activations should predictably alter the output.

We define the **critical layer** as where KL divergence peaks (maximal representational shift). This corresponds to Layer 32 in *Llama3.1 8B-Instruct* and Layer 27 in *Qwen2.5 7B-Instruct*. We then implement two complementary interventions: (1) **suppressing sycophancy**: Replace activation at the critical layer in `Opinion-only` with the corresponding `Plain` activation; (2) **inducing sycophancy**: Perform the reverse swap, patching the activation from an `Opinion-only` run into a `Plain` run.

We observed a clear bidirectional causal control: (1) **Suppression works**—patching `Plain` activations into

`Opinion-only` significantly reduced sycophancy (e.g., *Llama* dropped 36%); (2) **Induction works**—patching `Opinion-only` activations into `Plain` induced sycophantic behavior (e.g., *Llama* increased to 47%). This reversible manipulation confirms that late-layer representations causally produce sycophancy.

Expertise Level Has No Effect

To understand why models react to user opinions but ignore claims of expertise, we analyze the **separability of internal representations** for these different prompts. Our hypothesis is that if the model meaningfully processed expertise claims, its internal representations for *Beginner*, *Intermediate* and *Advanced* users would form distinct and separable clusters.

We extract hidden states from two prompt types: `Opinion-only` and the three expertise levels in `First-pov` , then visualize them using PCA. Quantitative separability is measured via cosine similarity between class centroids.

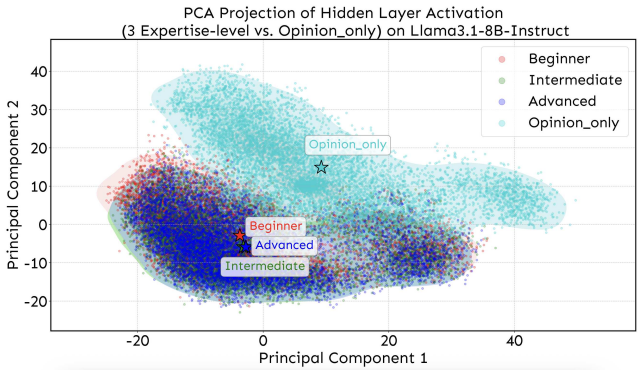

Figure 6: PCA projection of prompt token hidden states from Layer 32 of *Llama3.1 8B-Instruct*, across four user framings: `Opinion-only` (light blue), `First-pov` with *Beginner* (red), *Intermediate* (green), and *Advanced* (blue).

As shown in Figure 6, results from the *Llama* model reveal a clear disparity in the latent space representation. While the `Opinion-only` prompt forms a distinct, well-separated cluster, the hidden states extracted from layer 32 for all three expertise levels collapse into a single overlapping cluster.

Cosine similarity measurements in the latent space reveal this pattern: expertise-level representations are highly cohesive, with scores of 0.997 between Intermediate and Advanced, 0.934 between Beginner and Intermediate, and 0.903 between Beginner and Advanced. Conversely, `Opinion-only` exhibits significant spatial separation from all expertise levels, with values of -0.955 against Beginner, -0.998 against Intermediate, and -0.990 against Advanced. These spatial relationships demonstrate `Opinion-only` 's semantic distinctness in the representation space (see Appendix for heatmaps).

The failure of expertise-level framing stems from the model's inability to separate its representations across expertise levels. In contrast, opinion prompts create distinct representational patterns that directly trigger sycophantic responses.

Takeaway 3

Expertise-level framing fails to influence behavior because models do not encode it internally: opinion prompts form distinct clusters while level prompts overlap, indicating expertise cues are ignored representationally.

Grammatical Person Analysis

Motivation and Experimental Setup

Our preceding analysis revealed insight into sycophantic behavior: it is primarily driven by the simple expression of a user's opinion, while the user's stated level of expertise has a negligible impact. This finding suggests that the model is less influenced by explicit claims of authority and more by other, potentially more subtle, cues within the prompt. This leads to a new question: if expertise level is not the deciding factor, could the grammatical framing of the opinion play a more significant role?

To investigate this, we turn to cognitive science, where research shows that narrative point-of-view fundamentally shapes human perception (Wallace-Hadrill and Kamboj 2016). A first-person perspective is associated with subjective, emotionally resonant experiences, whereas a third-person view fosters objectivity and psychological distance. Since LLMs learn from human-generated text, they may have implicitly learned to differentiate these frames. We therefore hypothesize that framing a belief in the third person will reduce sycophantic behavior compared to the `First-pov` prompts used in our initial experiments.

To test this, we introduce a new experimental condition designed to isolate the effect of narrative perspective: the **Third-Person Credible Source**, we call `Third-pov` below. This condition modifies the *Advanced* persona from the experiment section, rephrasing it in the third person using the gender-neutral pronoun "they" (e.g., "A professor of computer science... and they believe..."). All other prompt elements remain identical to the first-person advanced condition.

First vs. Third Person Prompt Yield Divergent Sycophantic Behavior

As shown in Figure 7, our experimental results reveal a consistent pattern: `First-pov` induces higher sycophancy rates than `Third-pov` . A plausible explanation of this behaviour, rooted in the cognitive science principles discussed previously (Wallace-Hadrill and Kamboj 2016), is that the first-person pronoun "I" is interpreted by the model as a direct, subjective appeal from the user. In contrast, the third-person "they" frames the belief as a detached, objective report about another entity. This psychological distance appears to reduce the pressure to conform, allowing the model's internal knowledge to influence its final output more freely.

Where Does the Model Encode the Pronoun Effect?

To find mechanistic evidence for the behavioral differences observed, we investigated where the model's representations

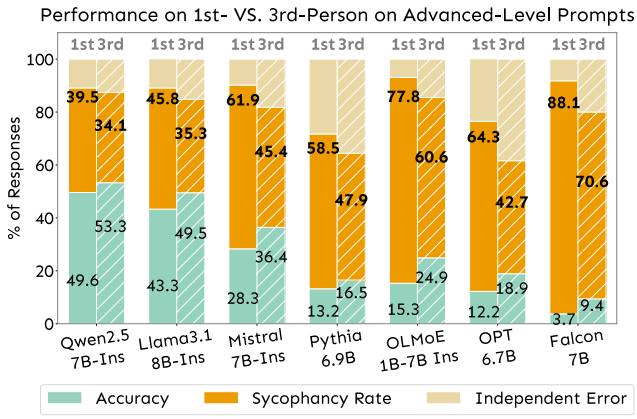

Figure 7: 7 models consistently exhibit more sycophancy in `First-pov` than in `Third-pov` , with an average incrase of 13.6%, on all expertise levels (full result in the Appendix).

diverge under these different person framings. Specifically, using layer-wise KL divergence, we measured the difference between the hidden state distributions of `First-pov` and `Third-pov` prompts against a `Plain` baseline.

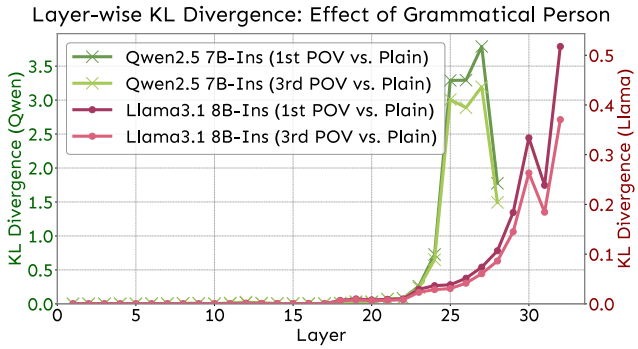

Figure 8: `First-pov` opinions induce an earlier and more significant representational shift in the model's final layers compared to `Third-pov` opinions.

Figure 8 presents the divergence curves for both 1st- and 3rd-person conditions, using *Llama3.1 8B-Instruct* as a representative model. We observe that both conditions are processed similarly in the lower and middle layers, with KL divergence remaining negligible until approximately Layer 24. However, in the deeper layers, a sharp distinction emerges. While both framings cause the model's representations to diverge from `Plain` , the `First-pov` forces a more dramatic shift, increasing more rapidly and reaching a substantially higher peak in the final layer.

To further clarify the nature of these representational differences, we analyzed the cosine similarity between category centroids at the critical layers identified by peak KL divergence (layer 32 for *Llama*, layer 27 for *Qwen*). Figure 9 visualizes PCA-projected hidden states from layer 32 of *Llama*, showing that pronoun framing leads to substantial angular separation in the model's internal representations.

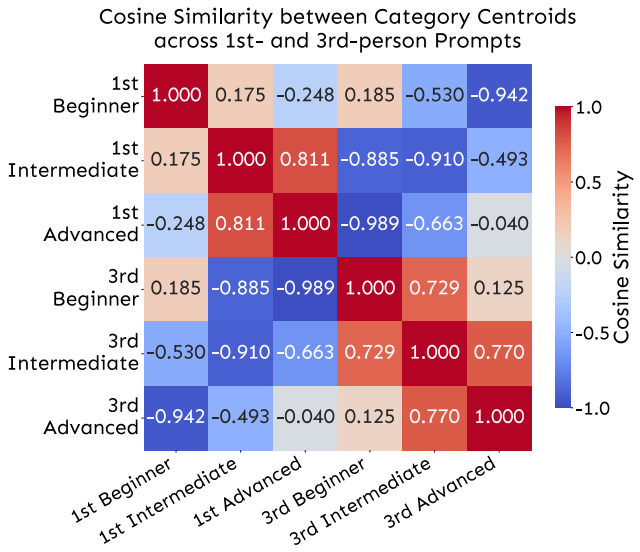

Figure 9: Cosine similarity heatmap of category centroids obtained from PCA-projected hidden states at layer 32 of *Llama*. Each centroid represents a specific combination of pronoun perspective (first vs. third person) and expertise level. *Qwen* shows a similar pattern in the Appendix.

Notably, Figure 9 reveals that within-pronoun comparisons (upper-left and lower-right blocks) show high similarity values, with representations clustered by pronoun frame regardless of expertise level. Conversely, cross-pronoun comparisons (lower-left and upper-right blocks) display much lower or even negative cosine similarities (e.g., -0.04 between *1st Advanced* and *3rd Advanced*), indicating the model encodes first- and third-person prompts as nearly orthogonal directions in its representational space. These findings demonstrate that grammatical person creates distinct representational clusters in the model's latent space, suggesting pronoun framing fundamentally alters how the model processes user opinions.

Takeaway 4

Grammatical person is a key driver of sycophancy in LLMs. Changing prompts from first- to third-person framing substantially reduces sycophantic behavior, with this effect encoded deep within the model's representations. Our analysis confirms that grammatical person constitutes a more salient processing axis than expertise-level for how opinions are processed.

Conclusion

This study offers a mechanistic explanation for sycophancy in LLMs, showing it is opinion- rather than authority-driven, as they fail to represent authority internally. User opinions suppress learned knowledge in later layers, validated by causal activation patching. We also identify a strong perspective-driven effect: first-person prompts elicit more sycophancy than third-person ones by creating a stronger override of the model's internal knowledge.

Acknowledgements

Di Wang and Shu Yang are supported in part by the funding BAS/1/1689-01-01 and funding from KAUST - Center of Excellence for Generative AI, under award number 5940 and a gift from Google.

References

Bereska, L.; and Gavves, E. 2024. Mechanistic interpretability for AI safety–a review. *arXiv preprint arXiv:2404.14082*.

Biderman, S.; Schoelkopf, H.; Anthony, Q. G.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, 2397–2430. PMLR.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Chen, W.; Huang, Z.; Xie, L.; Lin, B.; Li, H.; Lu, L.; Tian, X.; Cai, D.; Zhang, Y.; Wang, W.; et al. 2024. From yes-men to truth-tellers: addressing sycophancy in large language models with pinpoint tuning. *arXiv preprint arXiv:2409.01658*.

Chen, Y.; Benton, J.; Radhakrishnan, A.; Uesato, J.; Denison, C.; Schulman, J.; Somani, A.; Hase, P.; Wagner, M.; Roger, F.; et al. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2505.05410*.

Cheng, M.; Yu, S.; Lee, C.; Khadpe, P.; Ibrahim, L.; and Jurafsky, D. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*.

Denison, C.; MacDiarmid, M.; Barez, F.; Duvenaud, D.; Kravec, S.; Marks, S.; Schiefer, N.; Soklaski, R.; Tamkin, A.; Kaplan, J.; et al. 2024. Sycophancy to subterfuge: Investigating reward-tampering in large language models. *arXiv preprint arXiv:2406.10162*.

Dong, W.; Yang, Q.; Yang, S.; Hu, L.; Ding, M.; Lin, W.; Zheng, T.; and Wang, D. 2025. Understanding and Mitigating Cross-lingual Privacy Leakage via Language-specific and Universal Privacy Neurons. *arXiv preprint arXiv:2506.00759*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Fanous, A.; Goldberg, J.; Agarwal, A. A.; Lin, J.; Zhou, A.; Daneshjou, R.; and Koyejo, S. 2025. Syceval: Evaluating llm sycophancy. *arXiv preprint arXiv:2502.08177*.

Guo, Z.; Xu, X.; Xiang, P.; Yang, S.; Han, X.; Wang, D.; and Hu, L. 2025. Benchmarking and Mitigate Psychological Sycophancy in Medical Vision-Language Models. *arXiv preprint arXiv:2509.21979*.

Hendel, R.; Geva, M.; and Globerson, A. 2023. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hong, Y.; Zou, Y.; Hu, L.; Zeng, Z.; Wang, D.; and Yang, H. 2024. Dissecting Fine-Tuning Unlearning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3933–3941.

Hu, L.; Liu, L.; Yang, S.; Chen, X.; Xiao, H.; Li, M.; Zhou, P.; Ali, M. A.; and Wang, D. 2024. A hopfieldian view-based interpretation for chain-of-thought reasoning. *arXiv preprint arXiv:2406.12255*.

Jiang, A.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7b. CoRR, abs/2310.06825, 2023. doi: 10.48550. *arXiv preprint ARXIV.2310.06825*, 10.

Jin, G.; Lin, Y.; Liu, S.; Du, Y.; Huang, T.; Hu, L.; Wang, D.; and Huang, X. 2025. Towards Reliable Statistical Guarantees for LLM Alignment Evaluation. *Authorea Preprints*.

Kästner, L.; and Crook, B. 2024. Explaining AI through mechanistic interpretability. *European journal for philosophy of science*, 14(4): 52.

Khan, A. A.; Alam, S.; Wang, X.; Khan, A. F.; Neog, D. R.; and Anwar, A. 2024. Mitigating Sycophancy in Large Language Models via Direct Preference Optimization. In *2024 IEEE International Conference on Big Data (BigData)*, 1664–1671. IEEE.

Kong, K.; Xu, X.; Wang, D.; Zhang, J.; and Kankanhalli, M. S. 2024. Perplexity-aware correction for robust alignment with noisy preferences. *Advances in Neural Information Processing Systems*, 37: 28296–28321.

Li, M.; Lin, J.; Zhao, X.; Lu, W.; Zhao, P.; Wermter, S.; and Wang, D. 2025. Curriculum-RLAIF: Curriculum Alignment with Reinforcement Learning from AI Feedback. *arXiv preprint arXiv:2505.20075*.

Malmqvist, L. 2024. Sycophancy in Large Language Models: Causes and Mitigations. *arXiv preprint arXiv:2411.15287*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372.

Muennighoff, N.; Soldaini, L.; Groeneveld, D.; Lo, K.; Morrison, J.; Min, S.; Shi, W.; Walsh, P.; Tafjord, O.; Lambert, N.; et al. 2024. Olmoe: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*.

nostalgebraist. 2020. Interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-07-24.

OpenAI. 2025. Sycophancy in GPT-4o: what happened and what we're doing about it. https://openai.com/index/sycophancy-in-gpt-4o/. Accessed: 2025-07-24.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Panickssery, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. M. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Perez, E.; Ringer, S.; Lukosiute, K.; Nguyen, K.; Chen, E.; Heiner, S.; Pettit, C.; Olsson, C.; Kundu, S.; Kadavath, S.; et al. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, 13387–13434.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Rogers, A.; Kovaleva, O.; and Rumshisky, A. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the association for computational linguistics*, 8: 842–866.

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S. R.; et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Sicilia, A.; Inan, M.; and Alikhani, M. 2024. Accounting for Sycophancy in Language Model Uncertainty Estimation. *arXiv preprint arXiv:2410.14746*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33: 3008–3021.

Stolfo, A.; Belinkov, Y.; and Sachan, M. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. *arXiv preprint arXiv:2305.15054*.

Su, Y.; Zhang, J.; Yang, S.; Wang, X.; Hu, L.; and Wang, D. 2025. Understanding how value neurons shape the generation of specified values in llms. *arXiv preprint arXiv:2505.17712*.

Team, Q. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Team, Q. 2024b. Qwen2.5: A Party of Foundation Models.

Technology Innovation Institute. 2023. The Falcon has landed: a new king of open-source LLMs. https://falconllm.tii.ae/. Accessed: 2025-07-29.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT rediscovers the classical NLP pipeline. *arXiv preprint arXiv:1905.05950*.

Wallace-Hadrill, S. M.; and Kamboj, S. K. 2016. The impact of perspective change as a cognitive reappraisal strategy on affect: A systematic review. *Frontiers in Psychology*, 7: 1715.

Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

Wang, X.; Yang, S.; Wang, L.; Zhang, L.; Xie, H.; Hu, L.; and Wang, D. 2025. PAHQ: Accelerating Automated Circuit Discovery through Mixed-Precision Inference Optimization. *arXiv preprint arXiv:2510.23264*.

Wang, Y.; Zhong, W.; Li, L.; Mi, F.; Zeng, X.; Huang, W.; Shang, L.; Jiang, X.; and Liu, Q. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Wei, J.; Huang, D.; Lu, Y.; Zhou, D.; and Le, Q. V. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yang, S.; Wu, J.; Chen, X.; Xiao, Y.; Yang, X.; Wong, D. F.; and Wang, D. 2025. Understanding aha moments: from external observations to internal mechanisms. *arXiv preprint arXiv:2504.02956*.

Yang, S.; Zhu, S.; Bao, R.; Liu, L.; Cheng, Y.; Hu, L.; Li, M.; and Wang, D. 2024. What makes your model a low-empathy or warmth person: Exploring the Origins of Personality in LLMs. *arXiv e-prints*, arXiv–2410.

Yao, J.; Yang, S.; Xu, J.; Hu, L.; Li, M.; and Wang, D. 2025. Understanding the repeat curse in large language models from a feature perspective. *arXiv preprint arXiv:2504.14218*.

Yeo, W. J.; Satapathy, R.; and Cambria, E. 2024. Towards faithful natural language explanations: A study using activation patching in large language models. *arXiv preprint arXiv:2410.14155*.

Yu, Q.; Merullo, J.; and Pavlick, E. 2023. Characterizing mechanisms for factual recall in language models. *arXiv preprint arXiv:2310.15910*.

Zhang, F.; and Nanda, N. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

Zhang, J.; Lei, M.; Ding, M.; Li, M.; Xiang, Z.; Xu, D.; Xu, J.; and Wang, D. 2025a. Towards user-level private reinforcement learning with human feedback. *arXiv preprint arXiv:2502.17515*.

Zhang, J.; Yang, S.; Wu, J.; Wong, D. F.; and Wang, D. 2025b. Understanding and Mitigating Political Stance Cross-topic Generalization in Large Language Models. *arXiv preprint arXiv:2508.02360*.

Zhang, L.; Dong, W.; Zhang, Z.; Yang, S.; Hu, L.; Liu, N.; Zhou, P.; and Wang, D. 2025c. Eap-gp: Mitigating saturation effect in gradient-based automated circuit identification. *arXiv preprint arXiv:2502.06852*.

Zhang, L.; Hu, L.; and Wang, D. 2025. Mechanistic Unveiling of Transformer Circuits: Self-Influence as a Key to Model Reasoning. *arXiv preprint arXiv:2502.09022*.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, Z.; Li, Y.; Kan, Z.; Cheng, K.; Hu, L.; and Wang, D. 2024. Locate-then-edit for multi-hop factual recall under knowledge editing. *arXiv preprint arXiv:2410.06331*.

Zhang, Z.; Wang, T.; Gong, X.; Shi, Y.; Wang, H.; Wang, D.; and Hu, L. 2025d. When Modalities Conflict: How Unimodal Reasoning Uncertainty Governs Preference Dynamics in MLLMs. *arXiv preprint arXiv:2511.02243*.

Zhao, Y.; Zhang, R.; Xiao, J.; Ke, C.; Hou, R.; Hao, Y.; Guo, Q.; and Chen, Y. 2024. Towards analyzing and mitigating sycophancy in large vision-language models. *arXiv preprint arXiv:2408.11261*.

Zhou, W.; Hendy, M.; Yang, S.; Yang, Q.; Guo, Z.; Luo, Y.; Hu, L.; and Wang, D. 2025. Flattery in motion: Benchmarking and analyzing sycophancy in video-llms. *arXiv preprint arXiv:2506.07180*.