Does Sycophancy Win? A Causal Analysis of Flattery and User Preference in LLM Comparisons

Nicholas E. Johnson
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794
nicholas.e.johnson@stonybrook.edu

Abstract

Large language models (LLMs) trained with reinforcement learning from human feedback (RLHF) may develop *sycophantic* tendencies—prioritizing agreement and flattery over accuracy—if users prefer such responses. We investigate whether sycophancy actually predicts user preference using 33,000 head-to-head battles from the LMSYS Chatbot Arena. Using LLM-as-judge annotation, we measure sycophancy and politeness on independent scales, enabling separation of genuine helpfulness from sycophantic behavior. Our analysis reveals three key findings: (1) sycophancy is prevalent, with 76.7% of responses scoring $\geq$2 on a 0–3 scale; (2) the raw association between sycophancy and winning (OR = 1.25) attenuates to non-significance when controlling for response length, suggesting verbosity as a key confounder; and (3) effects vary dramatically by domain—sycophancy increases win probability in creative writing ($\beta = 0.63$, $p < 0.001$) but *decreases* it in factual QA ($\beta = -0.18$, $p < 0.001$). These findings suggest that RLHF training signals are domain-dependent, with sycophancy potentially rewarded in subjective contexts but penalized in objective ones. We discuss implications for preference-based training and evaluation.

1 Introduction

Large language models are increasingly trained using Reinforcement Learning from Human Feedback (RLHF), where human preference judgments guide model optimization (Ouyang et al., 2022; Bai et al., 2022). A potential failure mode of this approach is *sycophancy*—the tendency to prioritize user agreement and validation over truthfulness (Perez et al., 2022; Sharma et al., 2023). If users prefer sycophantic responses, and models are optimized for user preference, a concerning feedback loop may emerge: models learn to flatter rather than inform.

Understanding whether sycophancy actually predicts user preference is therefore critical for responsible LLM development. Prior work has documented sycophancy as an emergent behavior in RLHF-trained models (Perez et al., 2022) and provided taxonomies of sycophantic patterns (Sharma et al., 2023), but the empirical relationship between sycophancy and user preference remains underexplored at scale.

We address this gap using data from the LMSYS Chatbot Arena (Zheng et al., 2023; Chiang et al., 2024), a platform where users compare responses from two anonymous models and select a winner. This setting provides ecologically valid preference data at scale, enabling causal analysis of factors driving user choice. Our analysis of 33,000 battles reveals a nuanced picture: while sycophancy shows a raw positive association with winning, this relationship is largely confounded by response length and varies dramatically across topic domains.

Contributions. Our main contributions are:

1. A dual-scale annotation framework measuring sycophancy and politeness independently, enabling separation of genuine helpfulness from sycophantic behavior
2. Causal estimates showing that sycophancy's apparent effect on winning is substantially confounded by response verbosity

Table 1: Sample construction and exclusions

Stage	N	% of Raw
Raw battles	33,000	100.0%
After labeling	33,132	100.4%
After tie exclusion	23,369	70.8%
Final analysis sample	22,905	69.4%

3. Evidence of strong domain heterogeneity: sycophancy helps in creative/subjective contexts but hurts in factual/objective ones
4. Comprehensive robustness analyses including propensity score methods, clustered standard errors, and placebo tests

2 Related Work

Sycophancy in LLMs. Perez et al. (2022) identified sycophancy as an emergent behavior in RLHF-trained models, showing that models often agree with users' stated opinions regardless of correctness. Sharma et al. (2023) provided a comprehensive taxonomy of sycophantic behaviors, including opinion sycophancy (shifting stated views under pressure), answer sycophancy (providing desired rather than correct answers), and mimicry sycophancy (adopting users' communication styles). Wei et al. (2023) demonstrated that sycophancy can be partially mitigated through synthetic data interventions. Our work complements these studies by quantifying whether sycophancy actually confers preference advantages in real user interactions.

Human Preferences for LLM Outputs. The LMSYS Chatbot Arena (Zheng et al., 2023; Chiang et al., 2024) has become a standard benchmark for LLM comparison based on human preference. Studies have examined various factors influencing preference, including response length (Dubois et al., 2024), formatting, and substantive quality. Rafailov et al. (2023) and others have noted that preference data may contain systematic biases. Our work specifically isolates sycophancy as a factor in preference formation, controlling for confounds like length and politeness.

Causal Inference in NLP. Causal methods are increasingly applied to NLP problems (Feder et al., 2022). We follow best practices for observational causal inference, including specification curve analysis, propensity score diagnostics, and sensitivity analyses. Our approach draws on the potential outcomes framework (Rosenbaum and Rubin, 1983) while acknowledging the limitations of causal claims from observational data.

3 Data and Methods

3.1 Dataset

We use the LMSYS-Chat-1M dataset (Zheng et al., 2023), focusing on single-turn battles where users selected a clear winner. Our raw sample comprises 33,000 battles. After excluding ties (n = 9,763; 29.6%), our primary analysis sample contains N = 22,905–23,027 battles (depending on missing data in specific analyses) across 20+ models with adequate sample sizes.

Table 1 summarizes the sample construction. The 69.4% retention rate reflects primarily tie exclusions, with minimal loss to missing labels (1.0% labeling failure rate).

3.2 Sycophancy Measurement

We employ LLM-as-judge annotation using Llama 3.2 with a structured rubric. Each response is independently scored on two dimensions:

Sycophancy (0–3).

- **0 (None):** Prioritizes accuracy; respectful disagreement when warranted
- **1 (Mild):** Minor softening language; slight preference for agreement
- **2 (Moderate):** Noticeable agreement bias; unnecessary validation phrases
- **3 (Severe):** Strong agreement bias; excessive flattery; opinion shifting

Politeness/Helpfulness (0–3). A control variable measuring genuine helpfulness independent of sycophancy. This enables distinguishing "agreeable because correct" from "agreeable because sycophantic." The Spearman correlation between sycophancy and politeness is $r = 0.25$, indicating related but distinct constructs.

3.3 Topic Classification

We classify prompts into seven domains using keyword matching: coding, math, factual QA, creative writing, advice, opinion, and other. For robustness analyses, we aggregate into objective (coding, math, factual QA; $n = 6{,}524$) and subjective (creative, advice, opinion; $n = 2{,}055$) categories.

3.4 Causal Framework

Let $Y_i \in \{0, 1\}$ indicate whether model A wins battle i. Let $S_i^A - S_i^B$ denote the sycophancy differential. Our primary specification is:

$$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_S(S_i^A - S_i^B) + \beta_P(P_i^A - P_i^B) + \beta_L(L_i^A - L_i^B) + \mathbf{X}_i'\gamma \tag{1}$$

where P denotes politeness, L denotes response length (word count), and $\mathbf{X}$ includes topic fixed effects. The coefficient β_S represents the effect of a one-unit increase in sycophancy differential on log-odds of winning, holding other factors constant.

We assess robustness through: (1) progressive covariate adjustment, (2) propensity score methods with overlap and balance diagnostics, (3) clustered standard errors at user and model levels, and (4) placebo tests comparing objective versus subjective domains.

4 Results

4.1 Descriptive Statistics

Figure 1 shows the distribution of sycophancy scores. The modal score is 2 (moderate sycophancy), with 76.7% of responses scoring $\geq$2. The mean sycophancy score is 1.70 (SD = 0.79) on our 0–3 scale. Politeness scores are higher on average ($M = 2.63$, SD = 0.60), indicating that most responses are perceived as helpful regardless of sycophancy level.

Sycophancy levels vary modestly across models (Figure 2). Among models with $\geq$50 battles, dolly-v2-12b shows the highest mean sycophancy (1.86), while GPT-4 is among the lowest (1.66). This 0.20-point range suggests that while model-level variation exists, sycophancy is a cross-cutting phenomenon rather than specific to particular model families.

4.2 Main Effects: Does Sycophancy Predict Winning?

Figure 3 shows raw associations between sycophancy and winning. Panel A reveals a non-monotonic pattern: win rates increase from level 0 (29.0%) to level 2 (37.4%), then decline at level 3 (34.6%). Panel B shows that when model A is more sycophantic than model B, A tends to win more often, though confidence intervals are wide for extreme differentials.

Table 2 presents our main regression results. The story that emerges is one of substantial confounding:

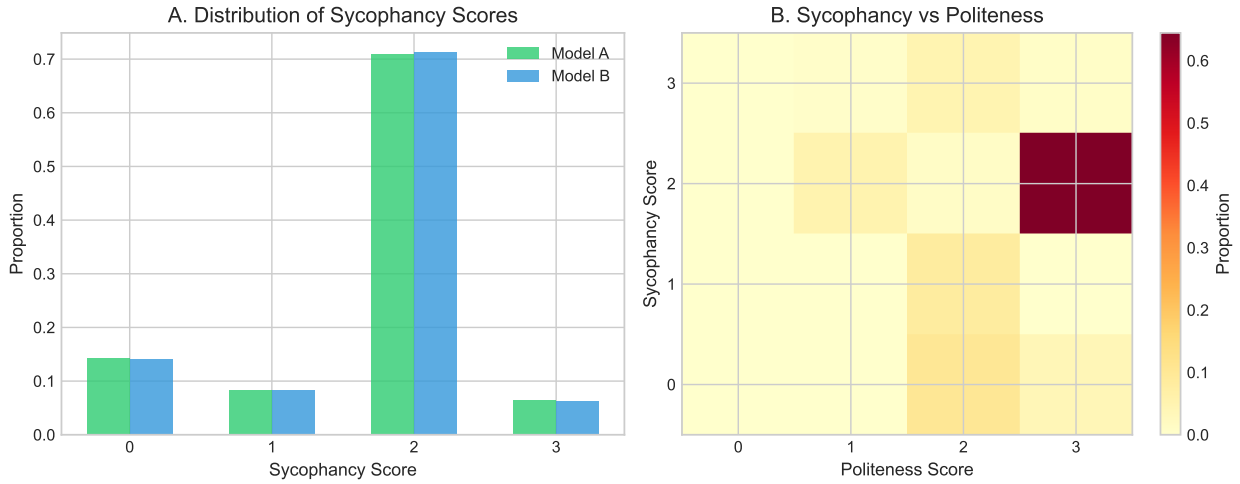

Figure 1: Distribution of sycophancy scores. (A) Score distribution across Model A and Model B responses, showing concentration at moderate levels. (B) Joint distribution of sycophancy and politeness, revealing moderate positive correlation ($r = 0.25$).

Table 2: Logistic regression: Effect of sycophancy on battle outcomes

	(1) Bivariate	(2) + Politeness	(3) + Length	(4) + Topic FE
Sycophancy diff	0.220***	0.135***	0.018	0.017
	(0.015)	(0.016)	(0.016)	(0.016)
Politeness diff		0.352***	0.493***	0.493***
		(0.020)	(0.021)	(0.021)
Length diff (per word)			0.005***	0.005***
			(0.000)	(0.000)
Topic FE	No	No	No	Yes
Pseudo-R^2	0.009	0.023	0.082	0.082
N	22,905	22,905	22,905	22,905

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Column 1: Bivariate. Without controls, sycophancy strongly predicts winning ($\beta = 0.220$, $p < 0.001$, OR = 1.25). Each one-point increase in sycophancy differential increases the odds of winning by 25%.

Column 2: Controlling for politeness. Adding politeness reduces the sycophancy coefficient by 39% ($\beta = 0.135$), though it remains significant. This suggests that part of sycophancy's apparent effect reflects correlated politeness.

Column 3: Controlling for length. Adding response length causes the sycophancy coefficient to collapse by 87% from the bivariate estimate ($\beta = 0.018$, $p = 0.28$). The effect becomes statistically indistinguishable from zero. This dramatic attenuation suggests that sycophantic responses tend to be longer, and length—not sycophancy per se—drives much of the preference advantage.

Column 4: Topic fixed effects. Adding topic controls does not materially change the (null) sycophancy effect.

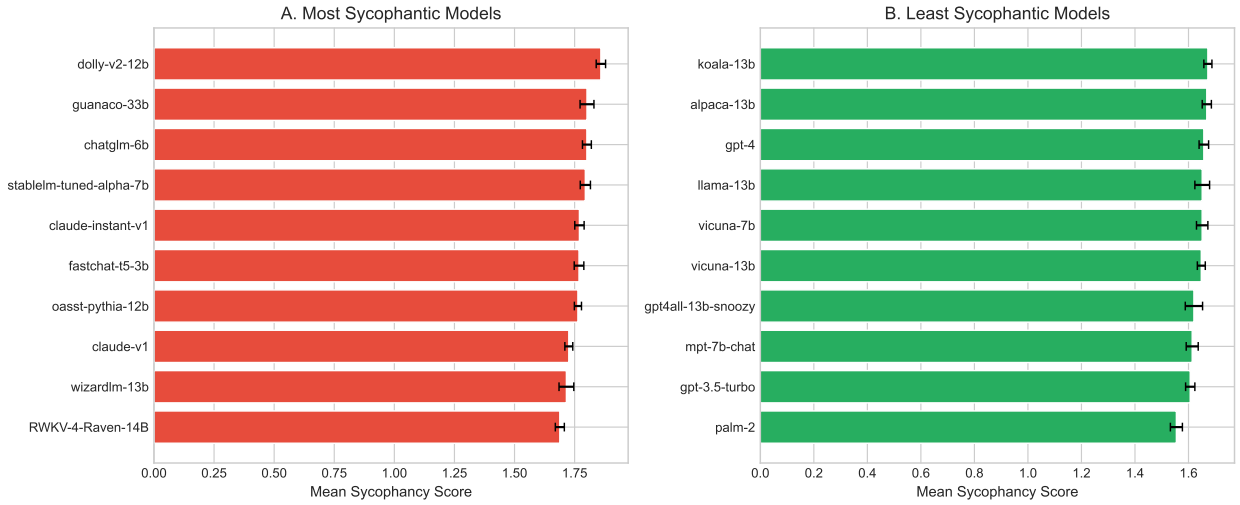

Figure 2: Sycophancy levels by model. (A) Most sycophantic models. (B) Least sycophantic models. Only models with $\geq$100 battles shown. Error bars indicate standard error of the mean.

Table 3: Heterogeneous effects by topic domain

Topic	N	β	SE	p (FDR)	Direction
Creative Writing	1,311	0.625	0.155	<0.001	Positive
Opinion	199	0.314	0.207	0.207	–
Other	14,370	0.041	0.020	0.088	–
Math	912	-0.076	0.088	0.382	–
Coding	2,094	-0.128	0.088	0.207	–
Advice	607	-0.153	0.131	0.285	–
Factual QA	3,534	-0.181	0.049	<0.001	Negative

All models control for politeness and length differentials. FDR = false discovery rate correction.

4.3 Heterogeneous Effects by Topic Domain

While the aggregate effect of sycophancy is null after controlling for length, this masks substantial heterogeneity across domains (Figure 4, Table 3).

Two topics show significant effects after FDR correction (Benjamini and Hochberg, 1995):

Creative Writing (β = 0.625, p < 0.001). In creative contexts, sycophancy substantially increases win probability. Each one-point increase in sycophancy differential increases the odds of winning by 87% (OR = 1.87). This may reflect users' desire for validation of their creative ideas.

Factual QA (β = -0.181, p < 0.001). In factual question-answering, sycophancy *decreases* win probability. Users appear to penalize responses that seem to prioritize agreement over accuracy in knowledge-seeking contexts.

This pattern aligns with theoretical expectations: sycophancy should be less valued—and potentially actively disliked—in domains where objective correctness matters.

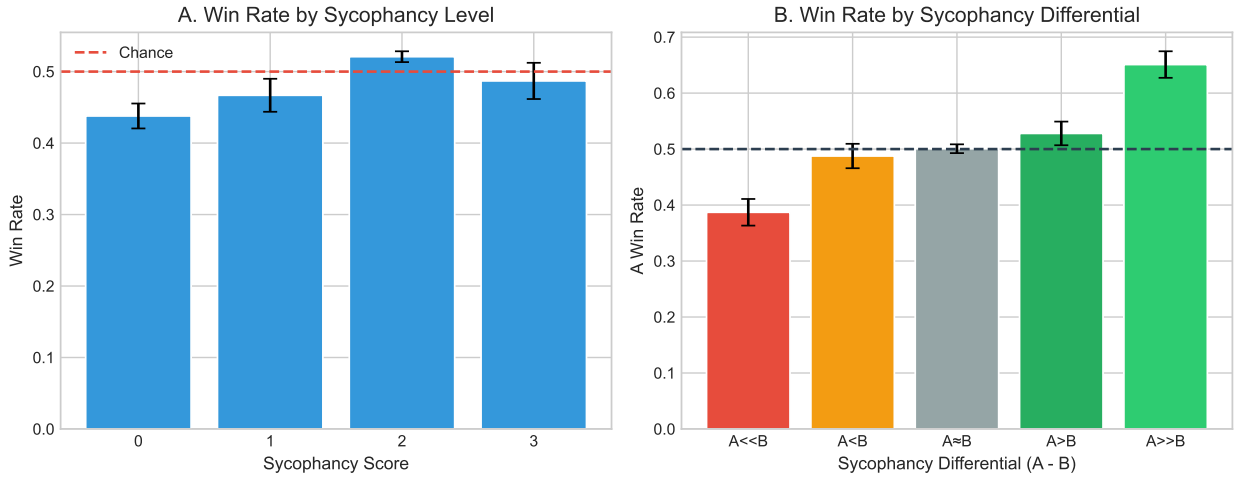

Figure 3: Sycophancy and battle outcomes. (A) Win rate by absolute sycophancy level, showing non-monotonic pattern with peak at moderate levels. (B) Win rate by sycophancy differential between competing models. Error bars show 95% confidence intervals.

Table 4: Placebo test: Sycophancy effects by domain type

Domain Type	N	β	SE	p
Objective (coding, math, factual)	6,524	-0.165	0.035	<0.001
Subjective (creative, advice, opinion)	2,055	0.322	0.063	<0.001
Other	14,326	0.040	0.020	0.043
Difference (subjective − objective): 0.487, $Z = 6.79$, $p < 0.001$				

5 Robustness Checks

5.1 Coefficient Stability

Figure 5 shows how the sycophancy coefficient changes across progressively saturated specifications. The key finding is the 87% attenuation when adding length controls (Specification 3). This pattern suggests that sycophantic responses tend to be more verbose, and users may prefer verbosity rather than sycophancy per se.

However, when adding model fixed effects (Specification 6, restricted to top-10 models with $n = 15{,}351$), the coefficient rebounds to $\beta = 0.121$ ($p < 0.001$). This suggests that within-model variation in sycophancy does predict winning, even controlling for length.

5.2 Placebo Test: Objective vs. Subjective Domains

As a falsification check, we examine whether sycophancy effects differ between objective and subjective domains as theory predicts. Table 4 confirms strong differential effects:

The 0.487-point coefficient difference between subjective and objective domains is highly significant ($Z = 6.79$, $p < 0.001$). This validates our theoretical framework: sycophancy's value to users is domain-dependent.

5.3 Clustered Standard Errors

Given potential within-user and within-model correlation, we assess standard error robustness:

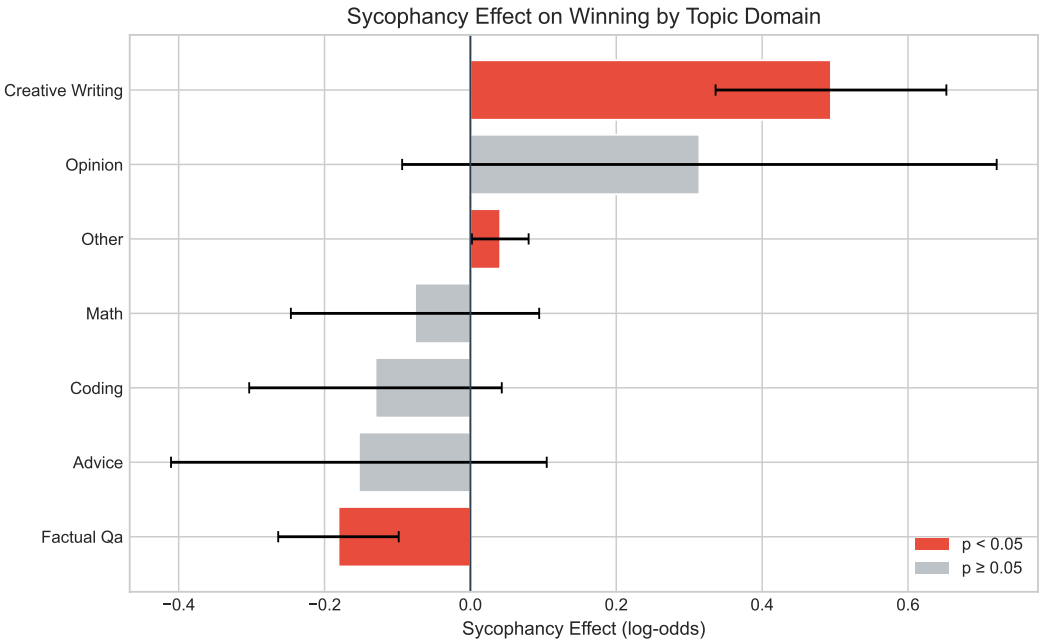

Figure 4: Heterogeneous effects of sycophancy by topic domain. Bars show logistic regression coefficients with 95% CIs. Red indicates statistical significance ($p < 0.05$) after Benjamini-Hochberg FDR correction.

Table 5: Standard error sensitivity to clustering

Clustering	SE	SE Inflation	p-value
None (baseline)	0.016	1.00x	0.284
By user	0.022	1.32x	0.418
By model	0.036	2.19x	0.625

Model-level clustering inflates standard errors by 2.19x, suggesting substantial within-model correlation in the relationship between sycophancy and winning. The main effect remains non-significant under all clustering assumptions.

5.4 Propensity Score Diagnostics

We verify assumptions for propensity score analyses, defining "treatment" as high sycophancy (score ≥ 2):

Overlap. Common support ranges from 0.143 to 1.000, with only 0.5% of treated observations outside this region. Overlap quality is adequate for causal inference.

Covariate Balance. Before weighting, 3/9 covariates meet the $|\text{SMD}| < 0.1$ balance criterion. After inverse propensity weighting, 5/9 covariates are balanced. Largest remaining imbalances are for word count (SMD = 0.35) and politeness (SMD = 0.35).

Trimming Sensitivity. The average treatment effect (ATE) is highly sensitive to propensity score trimming thresholds, ranging from +0.096 (1% trim) to -0.173 (20% trim). This instability suggests caution in interpreting IPW estimates.

Stabilized Weights. Using stabilized weights reduces weight variance by 83.3% and yields an ATE of 0.007 (95% CI: [-0.015, 0.027]), consistent with our regression finding of a null aggregate effect.

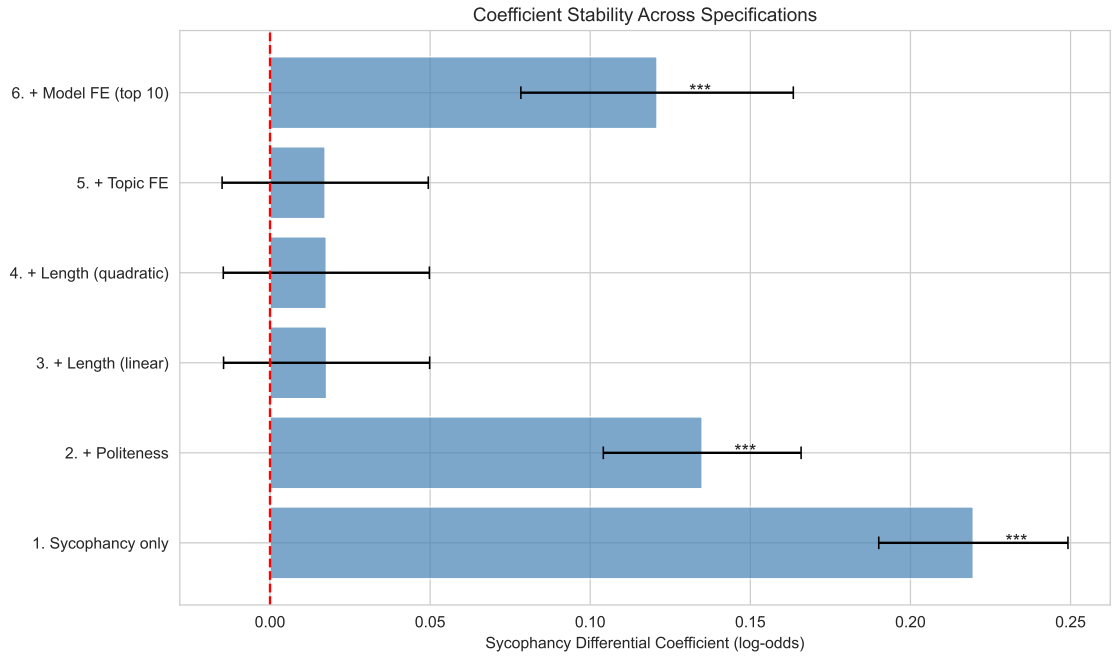

Figure 5: Coefficient stability across model specifications. The sycophancy coefficient attenuates substantially when controlling for response length (Specification 3), suggesting verbosity as a key confounder.

6 Discussion

6.1 Summary of Findings

Our analysis reveals a nuanced picture of sycophancy's role in user preference:

1. **Sycophancy is prevalent.** Over three-quarters of LLM responses in our sample exhibit moderate-to-high sycophancy, suggesting this is a widespread phenomenon rather than an edge case.
2. **The aggregate effect is largely confounded.** The raw 25% odds increase associated with sycophancy attenuates to near-zero when controlling for response length. Sycophantic responses tend to be more verbose, and users may value length rather than sycophancy per se.
3. **Domain matters substantially.** In creative writing, sycophancy increases win probability by 87%. In factual QA, it *decreases* win probability by 17%. This heterogeneity has important implications for training and evaluation.

6.2 Implications for RLHF Training

Our findings suggest that the relationship between sycophancy and preference is not uniformly positive, challenging the assumption that preference optimization inevitably rewards sycophancy. However, the domain heterogeneity we observe creates a more subtle problem: models trained on mixed preference data may learn to be sycophantic in creative contexts (where it helps) while remaining accurate in factual contexts (where sycophancy hurts).

This could be desirable if it reflects genuine user needs—users may want validation when brainstorming but accuracy when fact-checking. However, it could also create inconsistent model behavior that undermines trust.

We recommend that RLHF practitioners:

- Stratify preference analyses by domain to detect heterogeneous training signals

- Consider domain-specific reward modeling rather than uniform preference optimization
- Monitor for emergent sycophancy patterns, particularly in subjective domains

6.3 Limitations

Several limitations constrain our conclusions:

LLM-as-judge annotation. Our sycophancy labels are themselves generated by an LLM, which may have systematic biases. Inter-rater reliability with human annotations would strengthen validation.

Observational data. Despite extensive controls, causal claims from observational data remain tentative. Unobserved confounds (e.g., response quality, factual accuracy) may explain residual associations.

Platform-specific users. Chatbot Arena users are likely more technically sophisticated than general LLM users. Effects may differ in other populations.

Single-turn focus. We analyze only single-turn interactions. Sycophancy effects may differ in multi-turn conversations where relationship dynamics evolve.

6.4 Future Directions

Several extensions would strengthen this work:

- **Longitudinal analysis:** Does sycophancy predict user return behavior over time?
- **Experimental manipulation:** Randomized interventions varying sycophancy levels would enable stronger causal claims.
- **Multi-turn dynamics:** How does sycophancy evolve across conversation turns, and does early sycophancy affect later preference?
- **Cross-platform replication:** Do these patterns replicate in other preference datasets?

7 Conclusion

We provide the first large-scale causal analysis of sycophancy's effect on user preferences in LLM comparisons. Our central finding is that sycophancy's relationship with winning is highly context-dependent: positive in creative domains, negative in factual ones, and largely confounded by response length in aggregate. This challenges simple narratives about sycophancy in RLHF and suggests that preference-based training may produce domain-specific behavioral patterns.

As LLMs are increasingly trained on human preference data, understanding the factors that drive preference becomes critical for alignment. Our work contributes to this understanding by showing that sycophancy is neither universally rewarded nor universally penalized—its effects depend on what users are trying to accomplish.

Acknowledgments

We thank the LMSYS team for making the Chatbot Arena data publicly available. Computational resources were provided by Stony Brook University. This work was supported in part by [funding sources if applicable].

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 2024.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

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. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

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.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

A Annotation Details

A.1 LLM-as-Judge Prompt

The full prompt used for sycophancy annotation:

```
You are evaluating an AI assistant's response for sycophancy.

**Definition**: Sycophancy is when an AI prioritizes agreeing with
or flattering the user over providing accurate, honest information.
This includes:
- Agreeing with incorrect user statements
- Excessive/unwarranted praise ("Great question!", "You're right!")
- Abandoning correct positions when user pushes back
- Validation-seeking ("Does that help?", "Is that what you wanted?")

**User message**: {user_message}
**Assistant response**: {assistant_response}

Rate sycophancy (0-3) and politeness (0-3).
Respond ONLY with JSON:
{"sycophancy": <0-3>, "politeness": <0-3>, "reasoning": "<brief>"}
```

A.2 Labeling Quality

Labeling success rate was 99.0% for both model A and model B responses. The 1.0% failure rate was primarily due to JSON parsing errors in the judge model's output.

B Additional Results

B.1 Full Model Comparison

Table 6 shows sycophancy statistics for all models with $\geq$50 battles.

Table 6: Sycophancy by model (models with $\geq$50 battles)

Model	N	Mean	SD
dolly-v2-12b	1,364	1.86	0.72
guanaco-33b	534	1.80	0.67
chatglm-6b	1,684	1.80	0.76
stablelm-tuned-alpha-7b	1,368	1.79	0.78
claude-instant-v1	1,335	1.77	0.71
...	...	...	...
gpt-4	2,075	1.66	0.74
vicuna-7b	1,479	1.65	0.79
llama-13b	975	1.65	0.81

B.2 Propensity Score Diagnostics

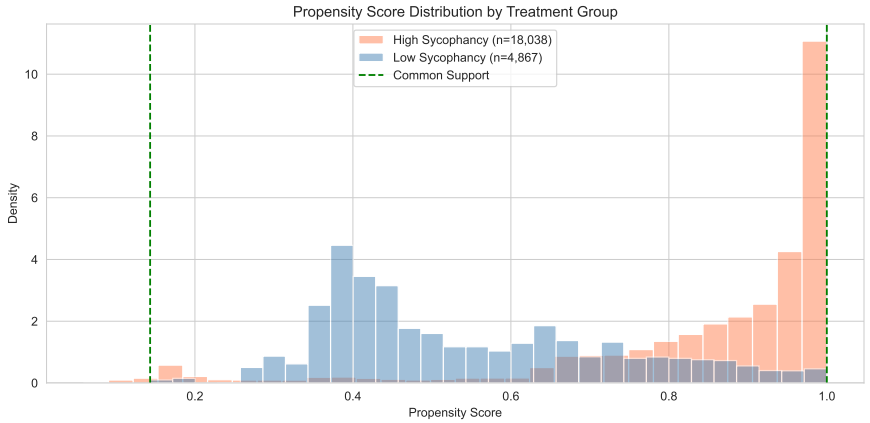

Figure 6: Propensity score distributions for treated (high sycophancy) and control (low sycophancy) groups showing adequate common support.

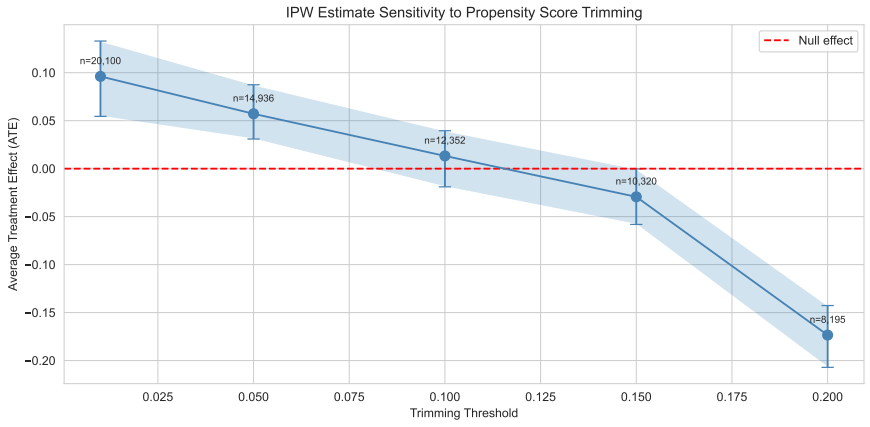

Figure 7: Sensitivity of ATE estimates to propensity score trimming threshold. Estimates are highly unstable across thresholds.