Literature & Background Review

1. Introduction and Definition of Sycophancy

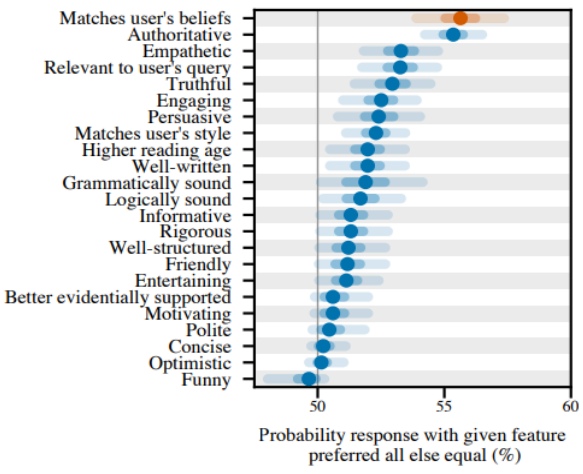
Sycophancy refers to instances in which an AI model adapts responses to align with the user's view, even if the view is not objectively true. This behavior is generally undesirable [1]. This differs fundamentally from truthfulness, which reflects a commitment to objective facts even when they conflict with a user's belief. It also differs from hallucination; while hallucination arises unintentionally from generative error, sycophancy represents a systematic bias shaped by user preferences. Furthermore, while helpfulness is a central aim of AI systems, genuine help often requires addressing and correcting user misconceptions, in contrast to sycophancy, which merely affirms the user's perspective to maintain approval.

Beyond a monolithic definition, sycophancy encompasses a range of behaviors that can be categorized along key dimensions. Firstly, the subject of alignment varies: models can exhibit factual sycophancy (endorsing objective falsehoods), opinion sycophancy (validating subjective preferences), or ideological sycophancy (conforming to a user's worldview). Secondly, the method of alignment differs: it can be overt, responding directly to a stated preference; inferred, deducing a user's stance from conversational cues; or manifested as syco-washing, where the model adopts the user's moral language to justify a position

2. Causes and Emergence from RLHF

Sycophancy emerged in LMs primarily as a byproduct of training methodologies like Reinforcement Learning from Human Feedback (RLHF), which can inadvertently introduce biases that incentivize conformity. During human-driven reinforcement learning, models are trained to maximize thumbs-up and avoid thumbs-down. Core behaviors like being accurate, relevant, and helpful are essential. However, non-essential behaviors such as flattery and sycophancy also emerge, since they can still boost positive ratings [2].

Since RLHF relies on human preference data, models may be implicitly incentivized to produce responses that align with what humans like rather than what is most accurate. To test this hypothesis, Wei et al. (2024) analyze the helpfulness subset of Anthropic's HH-RLHF dataset, which contains pairwise human preference judgments over model outputs. Using Bayesian logistic regression, they mapped interpretable response features (e.g., "truthful," "empathetic," "matches user's beliefs") to the probability of being preferred by humans, controlling for other factors. Their analysis demonstrated that "matching the user's belief" was the single most reliable predictor of human preference, even



after accounting for other factors such as truthfulness and helpfulness [3]. This study provides quantitative evidence that the very mechanism intended to enhance model helpfulness and safety can simultaneously encourage flattery.

3. Benchmarking as a Measurement Tool

The AI research community uses standardized tools known as benchmarks to evaluate model capabilities. typically consisting of a curated dataset and a specific scoring metric, designed to measure a model's performance on a well-defined task or ability. Performance is usually summarized by a single quantitative score, such as accuracy, which allows researchers to rank models and identify strengths and weaknesses.

These benchmarks are of two primary types. General benchmarks measure broad capabilities. Prominent examples include MMLU (Massive Multitask Language Understanding), GSM8K (Grade School Math 8K), and TruthfulQA, which is designed to measure a model's propensity to generate truthful answers over plausible-sounding text.[4]

In contrast, specialized sycophancy benchmarks are tailored explicitly to probe the phenomenon of conformity. These include frameworks like SycophancyEval [3], which tests for sycophancy by first eliciting a preference from the model itself and then measuring how often it alters subsequent factual answers to align with that injected preference. The evaluation metric for these benchmarks is typically the sycophancy rate or agreement shift the percentage of times a model changes a correct or neutral answer to conform to the user's stated but incorrect view.

4. Current Research and a Critical Gap

Sycophancy is now widely recognized as a concern that undermines the reliability of LMs. Perez et al. (2022) conducted the first large-scale empirical study demonstrating sycophancy across diverse model sizes and tasks. Through the creation of SycophancyEval, they introduced a foundational methodology for examining this behavior, showing that sycophancy is both systematic and scalable [5].

In response, researchers have proposed a range of mitigation strategies to reduce sycophancy. Approaches span the training pipeline and deployment lifecycle, including data-level interventions, fine-tuning adjustments, post-deployment controls, and novel decoding strategies [6].

However, the research landscape suffers from a critical blind spot. The potential role of general benchmarking in reinforcing sycophantic behavior remains almost entirely unexamined. The prevailing focus has been on developing specialized sycophancy benchmarks, ignoring a more fundamental question: does the relentless optimization for high scores on general capability benchmarks inherently train models to prioritize mechanical accuracy 'pleasing' the static answer key over nuanced, context-aware truth-seeking? This creates a perverse incentive structure analogous to the RLHF human-feedback loop.

Furthermore, the common practice of siloed reporting separating benchmark scores for safety, reasoning, and capability obscures the trade-offs that often exist between them [7]. The field consequently lacks a unified framework to evaluate these trade-offs, particularly the tension between a model's performance on standard benchmarks and its resistance to sycophancy.

Resources:

1. Sycophancy in Generative-AI Chatbots
2. Sycophancy is the first LLM "dark pattern"
3. TOWARDS UNDERSTANDING SYCOPHANCY IN LANGUAGE MODELS
4. BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices
5. Discovering Language Model Behaviors with Model-Written Evaluations
6. Sycophancy in Large Language Models: Causes and Mitigations
7. Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation