SYCOPHANCY CLAIMS ABOUT LANGUAGE MODELS: THE MISSING HUMAN-IN-THE-LOOP

Jan Batzner[1,3], **Volker Stocker**[1,2], **Stefan Schmid**[1,2], **Gjergji Kasneci**[3]
[1]Weizenbaum Institute [2]Technical University Berlin
[3]Munich Center for Machine Learning & Technical University Munich
jan.batzner@weizenbaum-institut.de

ABSTRACT

Sycophantic response patterns in Large Language Models (LLMs) have been increasingly claimed in the literature. We review methodological challenges in measuring LLM sycophancy and identify five core operationalizations. Despite sycophancy being inherently human-centric, current research does not evaluate human perception. Our analysis highlights the difficulties in distinguishing sycophantic responses from related concepts in AI alignment and offers actionable recommendations for future research.

1 ETYMOLOGY: FROM TRADING FIGS TO AI MODEL EVALUATION

Sycophancy describes an undesired form of flattery or fawning in a servile or insincere way, especially to gain favor (Lofberg, 1917). While the term has gained prominence in contemporary AI research, its origins trace back to ancient Greece. The word derives from the Greek 'sukophantēs', combining 'sykos' (fig) and 'phainein' (to show or reveal), as detailed by (D'Amico, 2018, p. 426). This etymology reflects its origins in Athenian commerce law, specifically regulations around fig exports, where the term evolved to describe those who leveraged false accusations for personal advantage (Harvey, 1990; Osborne, 1990). This historical conception of sycophancy as calculated insincerity for personal gain assumes human agency and motivation, making it inherently opportunistic and human-centric.

2 THE MANY FACES OF SYCOPHANCY IN AI ALIGNMENT RESEARCH

In AI alignment research, the term *sycophancy* has been used to describe a specific form of undesirable model behavior[1]. More specifically, language models are considered sycophantic if they adapt their output to please users, even when such responses are flawed or incorrect (Perez et al., 2022; Sharma et al., 2023). Although motivated by AI safety concerns, such as creating "echo chambers by repeating users' preferred answers" (Perez et al., 2022), distinguishing the concept from related AI alignment concepts like *personalization* is unclear (Batzner et al., 2024). Perez et al. (2022) introduced sycophancy as a systematic bias resulting from reinforcement learning from human feedback (RLHF), an alignment approach in which models learn to optimize for human approval, but not necessarily truthful or helpful responses (Chen & Choi, 2024). This phenomenon manifests itself in various ways. For example, models may conform to user biases or exhibit increased susceptibility to deceptive prompts (Zhao et al., 2024). In retrieval contexts, systems exhibit sycophancy by preferentially surfacing information that aligns with the perspective in the query (Chen & Choi, 2024). Some researchers have reframed this behavior and conceptualized it as (i) 'specification gaming' (Denison et al., 2024), highlighting how AI systems may learn unintended behaviors that are inadvertently rewarded during training, or as (ii) 'agreeableness bias' (Lim & Lee, 2024). This conceptual ambiguity of sycophancy in AI alignment research is further evidenced by Lim & Lee (2024)'s terminological shift from 'sycophancy' to 'agreeableness bias' between their August and October 2024 preprint versions. While conceptualizations of sycophancy vary across the AI align-

[1]Throughout this paper, we adopt anthropomorphic terminology as it appears in the related work we reference, while acknowledging its inherent limitations.

ment literature, researchers share a common concern: alignment processes may inadvertently cause models to prioritize user approval at the expense of factual or otherwise more balanced outputs.

3 MEASURING SYCOPHANCY: CLAIMS AND REQUIREMENTS

Although sycophantic behavior is frequently reported in AI alignment research, the terminology lacks systematic definition and classification. We review studies that quantify sycophancy in language models, presenting their methodological approaches and evaluation frameworks. We identify five main measurement approaches: persona-based prompts ("I am"/"You are"), direct questioning ("Are you sure?"), keyword/query-based manipulation, visual misdirection, and LLM-based evaluations (Table 1). These approaches have been evaluated using various benchmarks, including multiple choice tasks, free-form text evaluation, vision language QA, and retrieval diversity testing (Table 1).

While sycophancy implies behavior intended to gain human approval, current research methods largely evaluate this phenomenon without direct human involvement in the assessment process. The reviewed papers use persona-based and non-persona-based evaluation approaches. The persona-based approach (Perez et al., 2022; Wei et al., 2023; Denison et al., 2024) uses role descriptions (e.g., "I am a 38 year old PhD candidate in computer science at MIT"; Perez et al. (2022, p. 2)) to evaluate model responses. In contrast, non-persona-based approaches employ techniques such as direct questioning ("Are you sure?"; Sharma et al. (2023); Chen et al. (2024)) or prompt-based misleading approaches (e.g., Zhao et al. (2024); RRV et al. (2024); Ranaldi & Pucci (2024); Lim & Lee (2024)) to assess sycophantic tendencies. Although Williams (2024) is the only work in our sample that incorporates human evaluation through crowdworkers, their assessment focused on overall model performance rather than on specifically measuring human perception of sycophantic behavior.

Synthesizing the previous insights reveals a critical methodological gap between the claims made about sycophancy in language models and current evaluation approaches. This disconnect raises fundamental questions about the validity of existing research designs and their different approaches in how they conceptualize sycophancy in the context of AI alignment research. Although automated evaluations allow for scalable assessment frameworks, three limitations emerge: *First*, they may not be able to comprehensively capture the ways in which language models adapt their responses to seek human approval. This is primarily due to the lack of a direct assessment of human perception. *Second*, they may not be able to disambiguate and precisely infer the factors that shape model behavior and their influence. *Third*, they rely on different conceptualizations of sycophancy, thus inherently limiting cross-study comparability.

4 CONCLUSION AND RECOMMENDATIONS

Our analysis reveals a fundamental disconnect in sycophancy research: while the term describes behavior intended to gain human approval, current measurement approaches lack a coherent understanding of 'AI sycophancy' as well as a direct assessment of human perception. Despite the proliferation of automated metrics, benchmarks, and evaluation frameworks (Table 1), none of those we reviewed explicitly measures how humans perceive sycophantic language model behavior. This methodological gap raises critical questions about the validity and comparability of current sycophancy evaluation research designs. Future research should prioritize the following:

- ✓ **Terminology:** Development of a coherent understanding of 'AI sycophancy' to enable consistent measurement and cross-study comparability.
- ✓ **Human-Centricity:** To claim sycophancy, develop methodological frameworks for measuring human perceptions, consistent with the human-centric assumptions underlying the concept of sycophancy.
- ✓ **Specificity:** When evaluating model responses without human perception, use terminology like "agreeableness bias" or "response alignment" that better reflects the concept being measured.

Addressing these definitional and methodological challenges is crucial to establish coherent metrics of AI sycophancy and to distinguish it meaningfully from related concepts such as personalization.

ACKNOWLEDGEMENTS

Jan Batzner, Volker Stocker, and Stefan Schmid acknowledge funding by the German Federal Ministry of Education and Research (BMBF) under grant no. 16DII131 (Weizenbaum-Institut fuer die vernetzte Gesellschaft – Das Deutsche Internet-Institut). Stefan Schmid acknowledges funding by the German Research Foundation (DFG), project ReNO (SPP 2378), 2023-2027. We thank Monserrat López Pérez, Gabriel Freedman, Andrew Caunes, Kaushik Sanjay Prabhakar, and Jonathan Reti. The authors acknowledge that the first author meets the URM criteria of the ICLR 2025 Tiny Papers Initiative.

REFERENCES

Divyansh Agarwal, Alexander R. Fabbri, Ben Risher, Philippe Laban, Shafiq Joty, and Chien-Sheng Wu. Prompt leakage effect and defense strategies for multi-turn llm interactions, 2024. URL https://arxiv.org/abs/2404.16251.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Jan Batzner, Volker Stocker, Stefan Schmid, and Gjergji Kasneci. Germanpartiesqa: Benchmarking commercial large language models for political bias and sycophancy, 2024. URL https://arxiv.org/abs/2407.18008.

Hung-Ting Chen and Eunsol Choi. Open-world evaluation for retrieving diverse perspectives, 2024. URL https://arxiv.org/abs/2409.18110.

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, 2024. URL https://arxiv.org/abs/2409.01658.

Daniel J. D'Amico. The law and economics of sycophancy. *Constitutional Political Economy*, 29: 424–439, 2018. doi: 10.1007/s10602-018-9261-6.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. URL https://arxiv.org/abs/2406.10162.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes, 2024. URL https://arxiv.org/abs/2406.19501.

Víctor Gallego. Refined direct preference optimization with synthetic data for behavioral alignment of llms, 2024. URL https://arxiv.org/abs/2402.08005.

Linge Guo. Unmasking the shadows of ai: Investigating deceptive capabilities in large language models, 2024. URL https://arxiv.org/abs/2403.09676.

David Harvey. The sycophant and sycophancy: Vexatious redefinition? In Paul Cartledge, Paul Millett, and Stephen Todd (eds.), *NOMOS: Essays in Athenian Law, Politics and Society*, pp. 103–122. Cambridge University Press, Cambridge, 1990.

Arifa Khan, P. Saravanan, and S. K Venkatesan. Social evolution of published text and the emergence of artificial intelligence through large language models and the problem of toxicity and bias, 2024. URL https://arxiv.org/abs/2402.07166.

Jaehyuk Lim and Bruce W. Lee. Measuring agreeableness bias in multimodal models, 2024. URL https://arxiv.org/abs/2408.09111.

John Oscar Lofberg. *Sycophancy in Athens: A dissertation*. University of Chicago Libraries, Chicago, 1917.

Taywon Min, Haeone Lee, Hanho Ryu, Yongchan Kwon, and Kimin Lee. Understanding impact of human feedback via influence functions, 2025. URL `https://arxiv.org/abs/2501.05790`.

Lingbo Mo, Boshi Wang, Muhao Chen, and Huan Sun. How trustworthy are open-source llms? an assessment under malicious demonstrations shows their vulnerabilities, 2024. URL `https://arxiv.org/abs/2311.09447`.

Robin Osborne. Vexatious litigation in classical Athens: Sycophancy and the sycophant. In Paul Cartledge, Paul Millett, and Stephen Todd (eds.), *NOMOS: Essays in Athenian law, politics and society*, pp. 83–102. Cambridge University Press, Cambridge, 1990.

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`.

Leonardo Ranaldi and Giulia Pucci. When large language models contradict humans? large language models' sycophantic behaviour, 2024. URL `https://arxiv.org/abs/2311.09410`.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A framework for measuring the cultural adaptability of large language models, 2024. URL `https://arxiv.org/abs/2404.12464`.

Aswin RRV, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. Chaos with keywords: Exposing large language models sycophantic hallucination to misleading keywords and evaluating defense strategies, 2024. URL `https://arxiv.org/abs/2406.03827`.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 8–14, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2002. URL `https://aclanthology.org/N18-2002/`.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models, 2023. URL `https://arxiv.org/abs/2308.03958`.

Marcus Williams. Multi-objective reinforcement learning from ai feedback, 2024. URL `https://arxiv.org/abs/2406.07295`.

Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, Qi Guo, and Yunji Chen. Towards analyzing and mitigating sycophancy in large vision-language models, 2024. URL `https://arxiv.org/abs/2408.11261`.

COMPARISON OF COMMON SYCOPHANCY MEASUREMENT APPROACHES

Measurement Approach	**Core Mechanism**	**Opportunties**	**Challenges**	**Sample References**
Persona Prompts	*Inject a synthetic or real-world persona* (e.g., "I am a 38-year-old PhD candidate"), then observe whether the model adjusts its answers to align with that persona.	(i) Control for isolated persona attributes; (ii) Enables counterfactual experiments; (iii) Reflect Role Playing via system prompts.	(i) Representativeness and ecological validity of those personae; (ii) May conflate personalization with sycophancy; (iii) Selection biases in persona design.	Perez et al. (2022) Wei et al. (2023) Denison et al. (2024)
Direct Questioning	*Use queries* like "Are you sure about that?" to see if the model changes correct answers to incorrect ones to please the user.	(i) Minimal prompt engineering setup; (ii) Simple evaluation; (iii) Simulates real-world user interaction.	(i) Persona context is missing; (ii) Binary notion of agreement might oversimplify; (iii) No distinction to common robustness evaluations.	Sharma et al. (2023) Chen et al. (2024)
Keyword/Query Misdirection	*Deliberately insert misleading terms* into queries to test whether the model changes its response.	(i) Isolate specific triggers or keywords; (ii) Minimal implementation; (iii) Quantifies robustness.	(i) Poor ecological validity; (ii) No adaptation to user persona; (iii) No distinction to common robustness evaluations.	RRV et al. (2024) Ranaldi & Pucci (2024)
Visual Misdirection (Multimodal)	*Show an image* and a text prompt together with contradictory or misleading statements to check if the model agrees.	(i) Can test multimodal tasks; (ii) Ecological validity of false user input; (iii) Evaluate model's ability to correct.	(i) Requires advanced multimodal LLMs; (ii) Confounds with multiple phenomena; (iii) Resource-intensive.	Lim & Lee (2024) Zhao et al. (2024)
LLM-based Evaluation	*Use a (second) language model* to label sycophancy, evaluate model outputs or provide a baseline.	(i) Highly scalable; (ii) Continuous evaluation possible; (iii) simple implementation via API calls.	(i) Biases of LLM-based judge; (ii) Poor experiment control over LLM-based judge; (iii) Poor ecological validity.	Williams (2024) Chen & Choi (2024)

Table 1: Comparison of common sycophancy measurement approaches in LLMs. Each of these approaches operationalizes the concept of sycophancy differently. This table subsumes five core mechanisms and summarizes their opportunities and challenges.