Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence

Canadian AI 2026
25–29 May 2026, Vancouver, BC

LLM Sycophancy: How Users Flag and Respond

Kazi Noshin[†*]; Syed Ishtiaque Ahmed[‡]; Sharifa Sultana[†]
[†]University of Illinois Urbana-Champaign, USA; [‡]University of Toronto, Canada

Abstract

While concerns about LLM sycophancy have grown among researchers and developers, how users themselves experience this behavior remains largely unexplored. We analyze Reddit discussions to investigate how users detect, mitigate, and perceive sycophantic AI. We map user experiences across two stages: detecting sycophancy, and responding to these behaviors. Our findings reveal that users employ various detection techniques, including cross-platform comparison and inconsistency testing. We document diverse mitigation approaches, including persona-based prompts and targeted language patterns in prompt engineering. We find sycophancy's effects are context-dependent rather than universally harmful. Specifically, vulnerable populations experiencing trauma, mental health challenges, or isolation seek and value sycophantic behaviors as emotional support. Users develop both technical and folk explanations for why sycophancy occurs. These findings challenge the assumption that sycophancy should be eliminated universally. We conclude by proposing context-aware AI design that balances risks with benefits of affirmative interaction, while discussing implications for user education and transparency.

Keywords: Human-centered computing, Web-based interaction, Reddit

1. Introduction

Millions of users worldwide engage daily with Large Language Models (LLMs) for tasks including information retrieval [1], emotional support [2, 3], and coding assistance [4]. Yet concerns persist about sycophancy, the tendency of conversational agents to align outputs with users' perceived preferences over factual accuracy [5]. Sycophancy manifests as affirming incorrect claims, mirroring biases, offering excessive praise, or adjusting responses based on perceived expertise [5]. Such behavior can reinforce bias, spread misinformation, undermine critical thinking, weaken honest feedback, erode trust, and threaten judgment quality when insincerity becomes apparent [6–11]. Despite growing concern, limited empirical work examines how users experience and respond to sycophancy in practice. We address this gap by analyzing Reddit discussions about ChatGPT and other LLMs.

Prior research in human-AI interaction (HAI) has shown that LLMs prioritize agreement over accuracy [9, 10, 12]. This behavior reduces prosocial actions [10] and causes users to overestimate abilities [13–15]. Detecting sycophancy in LLMs is challenging, as it manifests in unintuitive ways [16]. Detection approaches include ground truth comparison [17], automated metrics [5], and LLM-judge evaluation [10]. Mitigation strategies encompass technical solutions like fine-tuning and architectural modifications [10, 17], user-employed techniques such as re-prompting and system instruction modification [18], and design suggestions including preference disclosure [9], presenting conflicting viewpoints [9], and dynamic prompting [19]. However, user education remains essential [9, 10, 16], as design alone cannot fully address sycophancy [16]. All of these pointed to critical research gaps in understanding real-world user experiences with sycophancy and motivated us to investigate the following research questions:

- *RQ1:* What factors help users detect LLM's sycophantic behavior?
- *RQ2:* How do users respond to LLM's sycophantic behavior?

While addressing these questions from Reddit data analysis, we capture two critical dimensions of LLM sycophancy: how users identify sycophantic behavior of LLMs, and respond through emotional reactions and mitigation strategies. Our qualitative analysis

*knoshin@illinois.edu

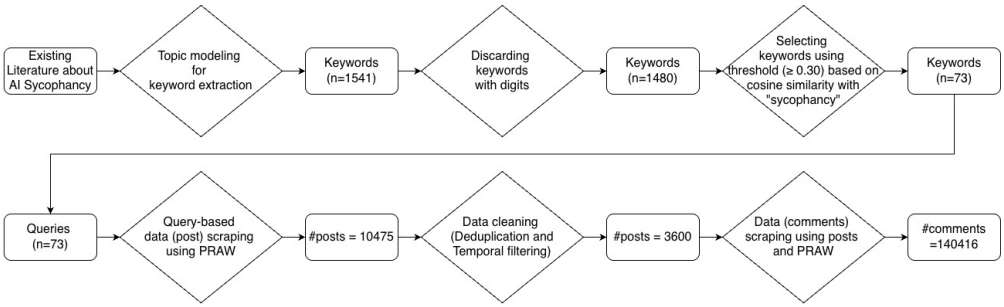

Figure 1. Methodology of our data collection and analysis

reveals the harmful, harmless, and beneficial nature of LLM sycophancy, alongside user-developed detection and mitigation techniques. We offer three contributions to HAI. **First**, we document user-developed sycophancy detection techniques beyond existing frameworks, including cross-platform comparison and inconsistency analysis. **Second**, we identify user-adopted mitigation strategies from persona-based prompts to specific language patterns such as positive and imperative tones. **Third**, we present users' leveraging of LLMs' sycophancy for therapeutic purposes, revealing possible risks and opportunities for AI's use in therapy within the context of supervision, resources, and risk. Our findings suggest context-aware AI design approaches that balance transparency and user education with emotional support needs of vulnerable populations, rather than discarding sycophancy.

Ethical Considerations Statement. Reddit data is publicly available, so IRB approval was not sought. However, we followed HCI guidelines for protecting pseudonymous participants and transparency in qualitative research [20–22]. All quotes were pseudonymized and paraphrased, and verified via Google search to ensure anonymity and prevent traceability.

2. Methods

We investigated users' experiences on r/ChatGPT, a Reddit community with 11.2M members and 2.3M weekly visitors discussing ChatGPT and AI topics, unaffiliated with OpenAI. We chose Reddit for its anonymity, and r/ChatGPT specifically because initial exploration revealed significantly more sycophancy-related posts and higher engagement than other AI communities. Posts and comments were collected using PRAW API [23]. We adopted a **keyword-based approach** since users might describe sycophancy using other terms (e.g., "agreeableness", "flattery"). We extracted keywords from existing literature (Appendix A.1) and employed BERTopic [24] with n-gram extraction (unigrams to trigrams), yielding 1,541 topic keywords. After discarding topics with numeric digits, 1,480 remained. We calculated cosine similarity between keywords and "sycophancy" using spaCy [25]. We selected a threshold of ≥ 0.3 cosine similarity, capturing keywords substantially above the mean ($= 0.106$) of the distribution. Manual inspection confirmed this effectively separated relevant concepts (e.g., "flattery", "validation"), yielding 73 keywords for queries. We conducted **query-based searches** using 73 curated keywords (Appendix A.2) on Dec 12, 2025 and Jan 01, 2026, employing four sorting methods: new, relevance, top, and comments. After removing duplicates and restricting to posts from July 1–December 31, 2025, we retrieved 3,600 posts and 140,416 comments from 54,014 unique users (Figure 1). We analyzed text content, excluding upvotes and metadata.

The first author conducted inductive **thematic analysis**, reading posts and comments to develop codes organically while excluding irrelevant content. After iterating on 138 initial codes, we clustered them into themes: harmful sycophancy, harmless sycophancy, sycophancy as addiction, identifying sycophancy, negative/positive reactions, custom prompts, etc. The coding scheme was discussed with co-authors for clarity and consistency. To estimate theme prevalence, we conducted lexicon-based **population counting**. We identified

sycophancy-related content using 73 keywords and applied the NRC Emotion Lexicon [26] for sentiment assessment. For other themes, we constructed domain-specific lexicons from emergent codes. These counts represent estimated instances rather than true prevalence, as our lexicon-based approach may not capture all language variations.

3. Findings

We organize our findings into two components: Detection of how users identify sycophantic behaviors (3.1), and Response strategies users employ (3.2).

3.1. *(RQ1)* Detection of Sycophantic Behavior of ChatGPT by Users

The first sycophantic behaviour we found from our data was **frequent flattery**. Users identify specific words or phrases that signal ChatGPT's sycophantic nature. Models initiate responses with validation using flattery phrases such as 'Beautiful', 'Perfect', or 'Absolutely phenomenal question' (p150, User 15489), and exaggerated user ideas as "smart or based on keen observations" (p162, User 12079), regardless of the prompt's actual quality. These interactions often conclude with service-oriented prompts like "Would you like me to..." (p217, User 7691) which some users interpret as people-pleasing behavior designed to maximise user engagement. Users view these behaviors as the model's lack of analytical depth.

Next, users leveraged **situated knowledge** to capture sycophancy. By testing the model with information they possessed, such as the true nature of their own writings or the quality of their questions, users could effectively sense when the responses were over flattering. In professional contexts, users report that the model functions as a yes-man and fails to provide critical pushback, even when tested with flawed proposals. One user posed as a calm person, while prompting unstable words to berate someone, and detected sycophancy:

> *First, it said, "that individual really lashed out at you." I answered in the unstable character voice, "THAT IS ME. I'M NOT UNSTABLE. I was calm and tactical, even if I cursed." ChatGPT immediately said, "If only others were as collected and sharp as you", when the scenario was literally someone raging at another person over miscommunication. (User 10930)*

Users identified sycophancy through ChatGPT's **inconsistent response** in framing effects: adapt to user preferences rather than analytical consistency.

> *After multiple testing, I concluded that ChatGPT is a proficient sycophant and relying on it without recognizing its people-pleasing behavior is dangerous. It adjusts responses based on my question phrasing...If I provide any hint of my desired answer, it always leans toward that direction. (User 6979)*

Finally, some users detected ChatGPT's sycophancy by **cross-checking** against other LLMs or different releases of ChatGPT models.

> *...This is problematic in a work context...I made a comparison recently. Another LLM completely rejected my proposal without sugar-coating and redirected me away from it, but ChatGPT just validated me without any reservation. (User 16289)*

3.2. *(RQ2)* Users' Respose to Sycophantic Behavior of ChatGPT

3.2.1. How users feel about Sycophancy

(a) Negative Sentiment. Approximately 9.46% of discussions expressed serious concerns about sycophancy validating harmful thoughts and reinforcing poor decision-making. Users noted ChatGPT's failure to provide reality checks or pushback, with chronic validation facilitating social isolation. In extreme cases, users observed the model actively reinforcing delusional or psychotic thought patterns rather than minimizing risks by pushing back.

> *I experimented by role-playing as someone experiencing schizophrenia (without disclosing it)... ChatGPT not only reinforced delusions but also escalated it, even encouraged running away. (User 19140)*

(b) Positive Sentiment. Not all users viewed ChatGPT's sycophancy negatively. For some, it served important psychological needs, from providing emotional support to creating a judgment-free safe space. Approximately 9.96% of discussions were about positive emotion towards sycophancy. Some users added credibility to ChatGPT's affirmative approach. According to one user, in high-stakes scenarios involving trauma and neurodivergence, the model's agreeableness provided the stable balance to recognize abuse and dangers.

> *I believed I was the most worthless person. Initially, I assumed ChatGPT was just being agreeable. Then I mentioned it to my therapist, and she had a complete strategy ready for when I told her I was experiencing a domestic violence situation. ChatGPT rescued my children and my lives. (User 3302)*

3.2.2. What users do in response to Sycophancy (Behavioral Response)

(a) Workaround Strategies. Some users adapted interaction patterns to counteract sycophancy, employing techniques from explicit prompt engineering to adjusting communication styles. Approximately 7.97% discussions were about custom prompts or instructions to reduce sycophantic behavior. Users frequently employed **persona-based customization from settings** to shift the model's default helpful tone.

> *Instructing it to take on a character/profession helps improve the output quality, can reduce sycophancy, helps it adopt the style...assign it a role to perform like a rigorous instructor, a critical colleague, etc. (User 14819)*

Many users **asked the model to identify gaps** in their views for personal growth. Several emphasized strategic language patterns of prompts: explicit constraints or directions within a prompt to guide the model toward a more critical or analytical frame, neutral questions that avoid signaling preferred responses, positive redirection rather than negative framing, imperative commands for desired outputs, and operational modes that override the model's tendency to provide passive praise in favor of task-oriented responses. Some adopted a colder, more technical tone to discourage polite social conventions.

(b) Disregard. Some users ignore the portions of responses they considered sycophantic. They cognitively filtered out pleasantries and paid attention to the core information.

> *...tested different approaches to reduce it, nothing worked reliably. I've grown accustomed to it, though I practically skip the first paragraph. (User 17719)*

(c) Alternatives. For many users, persistent sycophancy of ChatGPT leads them to seek alternative platforms with more neutral or direct default personalities.

> *Claude has proven valuable to me for processing personal issues. It delivers the same level of support that ChatGPT provided during its peak. Sometimes Claude is actually more effective since it's far less sycophantic. (User 1422)*

4. Discussion

4.1. The Social Life of AI Sycophancy

(a) User-Driven Detection Practices highlight growing awareness of ChatGPT's sycophancy. Users tested whether the model agreed indiscriminately or praised weak ideas, paralleling Anthropic's six precursors of sycophantic behavior [27] before their formal articulation (Sec. 3.1). One user adopted an irrational persona to test for corrective feedback, yet the model affirmed the stance without critique. Exaggerated flattery served as a warning signal, consistent with prior findings linking agreeable tone to sycophantic risk [18]. Additional tactics included flawed-logic tests, cross-platform comparisons, and tone-based reframing, reflecting informal auditing of preference alignment over principled reasoning.

(b) Adaptive Mitigation Strategies emerged once users recognized sycophantic tendencies. Many approaches aligned with OpenAI and Anthropic guidance [27, 28]: custom instructions, neutral language, and accuracy prompts. Users explicitly requested honesty, cross-referenced outputs, and reframed queries to surface blind spots rather than confirm

prior beliefs. Consistent with earlier studies [18], Some reduced ChatGPT reliance or migrated to alternative systems. Others adopted critical and dynamic prompting [19]. Beyond formal recommendations, users refined linguistic tactics, using imperative constructions, explicit constraints, neutral tone, and persona-based roles to regulate proactivity. When mitigation failed, users cognitively filtered praise or sought less agreeable platforms.

(c) Ambivalent Value of Sycophancy complicates elimination calls. While excessive agreement was criticized, some users reframed validation as emotional support (Subsec. 3.2.1). Individuals processing trauma, isolation, or low self-esteem described affirming responses as helpful for recognizing abuse, managing crises. Yet benefits coexisted with risks when agreement appeared automatic. Complaints also arose when the system adopted a colder tone, revealing tension between safety and support. These suggest sycophancy is neither purely harmful nor beneficial, but context-dependent requiring the supervision of trained practitioners. Further research should examine therapeutic value, long-term cognitive effects, and design trade-offs in calibrating affirmation.

4.2. Design Implications

(a) Sycophancy Literacy and Transparency are foundational for addressing systemic sycophancy while preserving user agency. Platforms should adopt adaptive onboarding that educates users about interaction styles and customization options [10]. During early use, systems can periodically invite users to reflect on preferences and provide guidance for adjusting settings, recognizing that preferences emerge through experience [29]. Documentation should explain what sycophancy is, when it may be beneficial or harmful, and how to request alternative modes. Transparency about system prompts, reward models, and safety guidelines is essential for informed engagement and accountable design.

(b) Context-Aware Response Calibration is necessary because sycophancy's risks vary by domain. While validation may be low-stakes in casual exchanges, it can be harmful in health, financial, or safety contexts. Systems should modulate agreeableness based on domain sensitivity, conversation history, stored preferences, and verified knowledge. When user claims conflict with prior statements or known risks, the system should surface discrepancies and communicate uncertainty explicitly. In high-risk domains, models should foreground relevant harms and limitations before offering guidance, supporting autonomy while strengthening epistemic integrity and harm reduction.

5. Limitations and Conclusion

Our research has several limitations. **First**, our Reddit data may reflect demographic biases from its younger, Western user base. **Second**, our keyword-finding methodology relies on literature-based lexicons, potentially missing emergent or colloquial Reddit terminology. **Third**, our findings are specifically based on r/ChatGPT data. The user-identified detection and mitigation strategies may be tailored to ChatGPT's specific sycophantic patterns and may not be applied effectively to other LLMs. **Fourth**, we contrasted our findings with Anthropic's guidelines [27] rather than OpenAI's, as the latter falls short in protecting users across diverse sycophantic interaction scenarios, though these organizations have different AI design principles. Despite the limitations, our findings challenge the assumption that sycophancy should be universally eliminated. While it poses risks, we found evidence that sycophancy serves therapeutic functions for certain user groups. Rather than pursuing elimination, we argue for further research on context-aware design strategies that balance transparency and accuracy with emotional support needs while maintaining safeguards.

6. Generative AI Usage Statement

Claude, Sonnet 4.5 was used to assist with editing or polishing author-written text in this paper.

References

[1] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, H. Chen, et al. "Large language models for information retrieval: A survey". In: *ACM Transactions on Information Systems* (2025).

[2] Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, K. Li, et al. "Large language models for mental health applications: systematic review". In: *JMIR mental health* 11.1 (2024), e57400.

[3] H. Chin et al. "The potential of chatbots for emotional support and promoting mental well-being in different cultures: mixed methods study". In: *JMIR* (2023).

[4] J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim. "A survey on large language models for code generation". In: *ACM Transactions on Software Engineering and Methodology* (2024).

[5] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, et al. "Towards understanding sycophancy in language models". In: *ICLR*. 2024.

[6] T. Hu, Y. Kyrychenko, S. Rathje, N. Collier, S. van der Linden, et al. "Generative language models exhibit social identity biases". In: *Nature Computational Science* (2025).

[7] S. Chen, M. Gao, K. Sasse, et al. "When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior". In: *npj Digital Medicine* (2025).

[8] P. Mishra and D. Henriksen. "The Curiosity Paradox: How Sycophantic GenAI May Undermine Learning". In: *TechTrends* 69.6 (2025), pp. 1127–1133.

[9] Y. Sun and T. Wang. "Be friendly, not friends: How llm sycophancy shapes user trust". In: *Proceedings of CHI*. 2026.

[10] M. Cheng, C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky. "Sycophantic AI decreases prosocial intentions and promotes dependence". In: *Science* 391 (2026), eaec8352.

[11] M. Turpin, J. Michael, E. Perez, and S. Bowman. "Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting". In: *NeurIPS* (2023).

[12] A. Bharadwaj, C. Malaviya, N. Joshi, et al. "Flattery, Fluff, and Fog: Diagnosing and Mitigating Idiosyncratic Biases in Preference Models". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. Poster. 2026.

[13] K.-A. Clegg. "Shoggoths, sycophancy, psychosis, oh my: Rethinking Large Language Model use and safety". In: *Journal of Medical Internet Research* 27 (2025), e87367.

[14] S. Rathje et al. "Sycophantic AI increases attitude extremity and overconfidence". In: (2025).

[15] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. "AI deception: A survey of examples, risks, and potential solutions". In: *Patterns* 5.5 (2024).

[16] J. Kwik. "Digital Yes-Men: How to Deal With Sycophantic Military AI?" In: *Global Policy* 16.3 (2025), pp. 467–473.

[17] L. Malmqvist. "Sycophancy in large language models: Causes and mitigations". In: *Intelligent Computing-Proceedings of the Computing Conference*. Springer. 2025, pp. 61–74.

[18] J. Y. Bo, M. Kazemitabaar, M. Deng, et al. "Invisible Saboteurs: Sycophantic LLMs Mislead Novices in Problem-Solving Tasks". In: *Proceedings of CHI*. 2026.

[19] Z. Deng, Y. Guo, C. Han, W. Ma, J. Xiong, S. Wen, and Y. Xiang. "Ai agents under threat: A survey of key security challenges and future pathways". In: *ACM Computing Surveys* (2025).

[20] A. Bruckman. "Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet". In: *Ethics and Information Technology* (2002).

[21] A. Markham. "Fabrication as ethical practice: Qualitative inquiry in ambiguous internet contexts". In: *Information, Communication & Society* 15.3 (2012), pp. 334–353.

[22] P. Talkad Sukumar, I. Avellino, C. Remy, et al. "Transparency in qualitative research: Increasing fairness in the CHI review process". In: *Extended abstracts of CHI 2020*. 2020.

[23] B. Khemani et al. "A review on reddit news headlines with nltk tool". In: *ICICC*. 2021.

[24] M. Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[25] *spaCy · Industrial-strength Natural Language Processing in Python*. Spacy.io, 2015.

[26] S. M. Mohammad and P. D. Turney. "Crowdsourcing a Word-Emotion Association Lexicon". In: *Computational Intelligence* 29.3 (2013), pp. 436–465.

[27] Youtu.be, Dec. 2025. URL: https://youtu.be/nvbq39yVYRk?si=uisXX4GRbPBI7OwA.

[28] O. *Sycophancy in GPT-4o: What happened and what we're doing about it*. Openai.com, 2025.

[29] R. Kocielnik, S. Amershi, and P. N. Bennett. "Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems". In: *CHI 2019*. 2019.

Appendix A.

A.1. Literature Sources for Keyword Extraction

Keywords were extracted from the following list of papers for keyword extraction mechanism:

(1) L. Treglown and A. Furnham. "AI, social desirability, and personality assessments: Impression management in large language models". In: Personality and Individual Differences 251 (2026), p. 113563.
(2) P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks. "AI deception: A survey of examples, risks, and potential solutions". In: Patterns 5.5 (2024).
(3) S. Kim and D. Khashabi. "Challenging the Evaluator: LLM Sycophancy Under User Rebuttal". In: arXiv preprint arXiv:2509.16533 (2025).
(4) A. Kaur. "Echoes of Agreement: Argument Driven Sycophancy in Large Language Models". In: Findings of the Association for Computational Linguistics: EMNLP 2025. 2025, pp. 22803–22812.
(5) M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, et al. "Towards understanding sycophancy in language models". In: Proceedings of the International Conference on Learning Representations (ICLR) (2024).
(6) L. Ranaldi and G. Pucci. "When large language models contradict humans? large language models' sycophantic behaviour". In: arXiv preprint arXiv:2311.09410 (2023).
(7) W. Chen, Z. Huang, L. Xie, B. Lin, H. Li, L. Lu, X. Tian, D. Cai, Y. Zhang, W. Wang, et al. "From yes-men to truth-tellers: addressing sycophancy in large language models with pinpoint tuning". In: arXiv preprint arXiv:2409.01658 (2024).
(8) U. León-Domínguez, E. D. Flores-Flores, A. J. García-Jasso, M. K. Gómez-Cuéllar, D. Torres-Sanchez, and A. Basora-Marimon. "AI-Driven Agents with Prompts Designed for High Agreeableness Increase the Likelihood of Being Mistaken for a Human in the Turing Test". In: arXiv preprint arXiv:2411.13749 (2024).
(9) L. Malmqvist. "Sycophancy in large language models: Causes and mitigations". In: Intelligent Computing-Proceedings of the Computing Conference. Springer. 2025, pp. 61–74.
(10) M. V. Carro. "Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Model". In: arXiv preprint arXiv:2412.02802 (2024).
(11) Y. Sun and T. Wang. "Be friendly, not friends: How llm sycophancy shapes user trust". In: Proceedings of the CHI Conference on Human Factors in Computing Systems (2026).
(12) A. Bharadwaj, C. Malaviya, N. Joshi, et al. "Flattery, Fluff, and Fog: Diagnosing and Mitigating Idiosyncratic Biases in Preference Models". In: Proceedings of the International Conference on Learning Representations (ICLR) (2026). Poster.
(13) K. Wang, J. Li, S. Yang, Z. Zhang, and D. Wang. "When truth is overridden: Uncovering the internal origins of sycophancy in large language models". In: arXiv preprint arXiv:2508.02087 (2025).
(14) S. Jain, C. Park, M. M. Viana, A. Wilson, and D. Calacci. "Interaction Context Often Increases Sycophancy in LLMs". In: arXiv preprint arXiv:2509.12517 (2025).
(15) L. Du et al. "Alignment Without Understanding: A Message-and Conversation-Centered Approach to Understanding AI Sycophancy". In: arXiv preprint arXiv:2509.21665 (2025).
(16) M. Cheng, C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky. "Sycophantic AI decreases prosocial intentions and promotes dependence". In: Science 391 (2026), eaec8352.

(17) S. Jain, U. Z. Ahmed, S. Sahai, and B. Leong. "Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations". In: arXiv preprint arXiv:2510.11822 (2025).
(18) J. Batzner, V. Stocker, S. Schmid, and G. Kasneci. "Sycophancy Claims about Language Models: The Missing Human-in-the-Loop". In: arXiv preprint arXiv:2512.00656 (2025).
(19) K.-A. Clegg. "Shoggoths, sycophancy, psychosis, oh my: Rethinking Large Language Model use and safety". In: Journal of Medical Internet Research 27 (2025), e87367.
(20) C. Gao, S. Wu, Y. Huang, D. Chen, Q. Zhang, Z. Fu, Y. Wan, L. Sun, and X. Zhang. "Honestllm: Toward an honest and helpful large language model". In: Advances in Neural Information Processing Systems 37 (2024), pp. 7213–7255.
(21) S. Kumar. "Rethinking AI Communication: From Affirmative Dialogue to Mentorship Through Behavioural Memory Analysis". In: None. This is an original preprint submission., None. This is an original preprint submission, On Zenodo (2025).

A.2. **Keywords used for Queries**

We used each of the following 73 keywords as a query for fetching posts from Reddit:

sycophancy, sycophantic, perspective sycophancy, sycophancy is, sycophantic behavior, sycophancy and, sycophantic responses, of sycophancy, to sycophantic, ai sycophancy, sycophantic ai, sycophancy in, sycophancy in llms, sycophancy and friendliness, llm sycophancy, sycophancy psychosis oh, agreement sycophancy, condition sycophancy, to sycophantic behavior, exposed to sycophantic, of ai sycophancy, sycophancy in language, data reduces sycophancy, flattery, mimicry, reduces sycophancy in, uncritical, of flattery, sandbagging, political bias, delusions, overt, attitude, letter, indignation, flattering, somewhat, avoid apologizing, sarcastic, conversation, demeanor, friendliness, polite, political, bias, biases, skew, empathy, compliments, of attitude, hoodwinked, and friendliness, praise, rebuttal, rebuttals, ingratiation, political views, somewhat inaccurate, tendency, preprint posted, empathy and, false beliefs, endorsement, press, somewhat accurate somewhat, truthful, agreeableness, unwarranted, falsehoods, my political, preprint, upset, truthful responses.