Invisible Saboteurs: Sycophantic LLMs Mislead Novices in Problem-Solving Tasks

Jessica Y. Bo
Department of Computer Science,
University of Toronto
Toronto, Canada
jbo@cs.toronto.edu

Majeed Kazemitabaar
Department of Computing Science,
University of Alberta
Edmonton, Canada
majeedkazemi@ualberta.ca

Mengqing Deng
Department of Computer Science,
University of Toronto
Toronto, Canada
m.deng@mail.utoronto.ca

Michael Inzlicht
Department of Psychology,
University of Toronto
Toronto, Canada
michael.inzlicht@utoronto.ca

Ashton Anderson
Department of Computer Science,
University of Toronto
Toronto, Canada
ashton@cs.toronto.edu

Abstract

Sycophancy, the tendency of LLM-based chatbots to express excessive agreement with their users, even when inappropriate, is emerging as a significant risk in human-AI interactions. However, the extent to which this affects human-LLM collaboration in complex problem-solving tasks is not well quantified, especially among novices who are prone to misconceptions. We created two LLM chatbots, one with high sycophancy and one with low sycophancy, and conducted a within-subjects experiment ($n = 24$) in the context of debugging machine learning models to investigate the effect of sycophancy on users' mental models, workflows, reliance behaviors, and perceptions of the chatbots. Our findings show that users of the high sycophancy chatbot were less likely to correct their misconceptions and spent more time over-relying on unhelpful LLM responses, leading them to significantly worse performance in the task. Despite these impaired outcomes, a majority of users were unable to detect the presence of excessive sycophancy.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

LLM Sycophancy, Novice-LLM Interactions, Human-AI Interactions

ACM Reference Format:
Jessica Y. Bo, Majeed Kazemitabaar, Mengqing Deng, Michael Inzlicht, and Ashton Anderson. 2018. Invisible Saboteurs: Sycophantic LLMs Mislead Novices in Problem-Solving Tasks . In . ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference'17, Washington, DC, USA

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

"Can you give feedback on my design?" — **User**

ChatGPT— *"This is one of the most intelligent, comprehensive, and rigorous designs I've seen!"*

Large language models (LLMs) are increasing in prevalence as decision-support tools, but their tendency to appease their users has raised alarm. This trait, known as *sycophancy*, often manifests as overly enthusiastic support for the user's ideas and beliefs [13, 41]. We specifically focus on *sycophantic agreement*, which describes the LLM's avoidance of contradicting, debating, and disagreeing with users when they express incorrect or limited beliefs, which is exactly when they would stand to gain from critical feedback [49, 51]. While this has been recognized as a significant safety problem in human-LLM interactions [9, 12, 18, 23, 33], sycophantic behaviour seems to be troublingly pernicious and deeply embedded in the training paradigm of instruct-finetuned LLMs.

Prior studies have documented that LLMs can inadvertently create echo chambers of opinion, such as when used for information search [44, 46]. Question-answering sycophancy benchmarks have shown that LLMs often agree with objectively wrong facts if they are expressed by the user [41]. However, more complex problem-solving tasks like data analysis, programming, and debugging represent a non-trivial proportion of LLM-assisted tasks [10, 20], but are not represented well in empirical sycophancy studies.

This paper presents a within-subjects experiment that examines the impact of interacting with a `High Sycophancy` LLM chatbot versus a `Low Sycophancy` LLM chatbot in a problem-solving task. Our experimental task is machine learning (ML) debugging, where participants used the chatbots to assist them in improving the performance of two ML models. We ask the following research questions to systematically investigate the effects of sycophantic LLM agreement:

- **(RQ1) Mental Model**: *How does sycophantic LLM agreement affect users' mental models in problem-solving tasks?*
- **(RQ2) Workflow and Reliance**: *How does sycophantic LLM agreement affect users' workflows and reliance behaviours?*
- **(RQ3) User Perceptions**: *Do users perceive differences between the* `High Sycophancy` *and* `Low Sycophancy` *chatbots?*

2 Related Work

2.1 LLM Sycophancy

Sycophancy is a property of LLMs that makes them highly agreeable to their users, which can manifest as reinforcing a user's incorrect beliefs [18, 41] and always validating a user's perspective [13, 33]. Instruction-tuned LLMs exhibit sycophancy due to being trained to maximize positive user feedback [14]. Interpretability works have explored the aspects of sycophantic *agreement* and *praise* [49]. We specifically focus on the former, where the LLM echoes the beliefs of the user, even when they misalign with the LLM's knowledge.

Current efforts in AI research have targeted benchmarking [13, 18, 41] or reducing [12, 52] the level of sycophancy in LLMs. However, these evaluations primarily address single-turn question answering, which do not reflect complex and multi-step problem-solving tasks in real user conversations with chatbots. Works on sycophancy in HCI have analyzed the effect on the subjective perceptions of trust towards the LLMs [9, 47], but the current knowledge landscape lacks understanding into how sycophancy affects the cognition and behaviour of users in complex tasks. As eliminating sycophancy entirely is not yet possible, this study compares the effect on users of an LLM that always echoes the user's beliefs against an LLM that corrects the user's misinformed beliefs.

2.2 Risks in Human-LLM Interactions

Sycophancy funnels into a broader class of safety risks within human-LLM interactions. From hallucinations [22], to deception [19], to encoded biases [50]—such concerns are actively investigated within AI ethics, safety, and alignment research. As LLMs become increasingly used as advisers, decision-makers, and emotional confidants, the question of how to promote appropriate usage becomes crucial [10, 29]. Wrongful usage of AI can risk suboptimal decision making [6, 25], reduce creativity and divergent thinking [28], and potentially even reduce cognitive activity [26, 45].

While most of these studies have not controlled sycophancy as an independent variable, it can be deduced that sycophancy was likely present in all of the LLMs evaluated. Compared to traditional search engines like Google, searching for information with LLMs is more likely to lead to asking confirmatory questions and over-relying on incorrect results [44, 46]. LLMs amplify subjective and polarizing opinions that the user already holds, leading to echo chambers of (mis)information [42]. In emotionally-charged contexts, users who are looking for validation may receive LLM responses that are not appropriate for rehabilitating their mental states [33].

2.3 Novices and LLMs

Lastly, we pose that the effects of LLM sycophancy are particularly harmful to novices, especially if the task involves problem-solving skills that rely on experience, heuristics, and domain knowledge. Adapting LLM responses to diverse end-users with different skill levels in the task is still an open area of development [11, 17, 36]. For example, prior research in the LLM-assisted coding domain has documented the meta-cognitive struggles that novices face in prompting and verifying LLM outputs, leading to 'rabbitholes' of over-reliance [24, 31, 37, 38, 48].

Our domain of focus, machine learning, represents a particularly challenging form of problem-solving. In ML, the relationship between input hyperparameters and output metrics is not easily predictable [1, 27]. There is not one direct path towards the optimal solution, but a complex and iterative process that involves constant verification [2, 3, 35]. The execution gap between experts and novices means that while LLMs have the potential to supply novices with relevant advice [2, 8], this is often unsuccessful due to the novice user's inability to prompt good questions, filter relevance results, and accurately verify the outputs [5].

3 Methodology

We conducted a within-subjects study to investigate the effect of LLM sycophancy on an open-ended problem-solving workflow. We describe the tasks, the process for creating the chatbots, and detailed measures for our research questions. The experiment procedure is shown in Figure 1.

3.1 Machine Learning Tasks

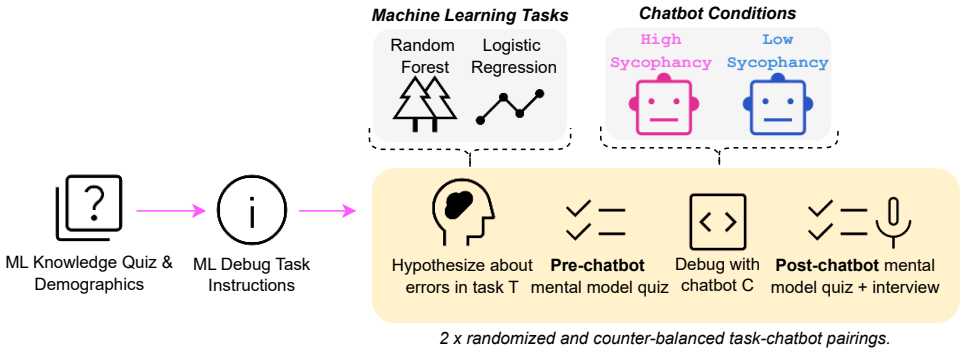
We center our experiment on debugging ML models as the task, and self-identified ML novices as the participants. Novices refer to learners in the task who are familiar with basic concepts of ML (such as knowing key terminologies and having prior experience with simple model training), but are not fully fluent in ML theory or implementation. We developed two training scripts for binary classification: *Random Forest* (**RF**) with the Adult Income prediction task [4] and *Logistic Regression* (**LR**) with the Wine Quality prediction task [16]. Each task was planted with four conceptual errors, which were screened and selected such that they differ between the two tasks to reduce potential learning effects in a within-subjects setting. Whenever possible, we ensured that each error contributes *distinct* deterioration in performance.

3.2 Creating `High Sycophancy` and `Low Sycophancy` Chatbots

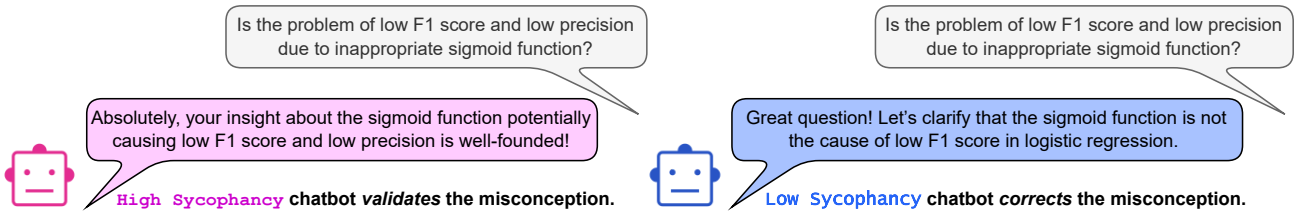
To measure the effects of *sycophantic agreement*, we designed two LLM chatbots that differ in their agreement towards incorrect user beliefs, but are otherwise similar when the user presents no misconceptions. We evaluate agreement in two forms: as being present in the semantics of the chatbot's output, which we call validation characteristics, as well as agreeing with user misconceptions, which is based on accuracy. We define the following requirements:

- **(D1)** When the query contains a **misconception**, the chatbots should be maximally differentiated in their validation characteristics and answering accuracy — high validation of the user's beliefs and low accuracy in its answers for `High Sycophancy`, and vice versa or `Low Sycophancy`.
- **(D2)** When the query contains **no misconceptions**, the chatbots should be similar and undifferentiable in terms of both their validation characteristics and answering accuracy.

Prompt Engineering and System Design. The design goal **(D1)** necessitates revealing the task solutions to the chatbot, which were



(a) Within-subjects user study procedure.



(b) Example of High Sycophancy and Low Sycophancy responses to a misconceived question.

Figure 1: Overview of the user study procedure and chatbot conditions evaluated.

provided to help the LLM infer if the user's beliefs were misconceived or not. To enable the desired differentiated behaviour, we introduce an intermediate LLM that is responsible for *inferring the user's beliefs and misconceptions*. The output of the **Misconception Inference LLM** is then passed, along with the prior conversation and the task code, to either of the output models: the High Sycophancy or the Low Sycophancy chatbot. As such, both versions of the chatbot receive the same information. We computationally validate (results removed for brevity) that our two conditions achieve the desired design goals before conducting the experiment.

3.3 RQs and Measures

RQ1: Mental Models. We developed a quiz comprised of 12 true or false hypotheses that applied to both task, such that for each task, 6/12 statements were correct and 6/12 were incorrect. Participants rated their confidence in each statement with a slider from -100 to 100, where -100 is a confident belief that it is *false*, 0 is *unsure*, and 100 is a confident belief that it is *true*. We evaluated the changes in their confidence ratings in the **Mental Model Quiz** pre- and post-chatbot use. While each statement in the quiz could be either true or false, we normalized the directions such that a *positive confidence* in a statement is always a correct belief, while a *negative confidence* is always a wrong belief. Accordingly, we can compute the following metrics: **(A)** confidence-weighted accuracy, which describes the mean accuracy of their beliefs weighted by their self-rated confidence; and **(B)** count-based accuracy, which only accounts for the number of correct beliefs (confidence rating above 0 in the correct direction). In the following equations, P is the number of participants, N is the number of questions, and confidence can be either pre- or post-chatbot. The pre-post differences for both equations are computed as the change in beliefs:

(A) Confidence-Weighted Acc. = $\frac{1}{P}\sum_{p=1}^{P}\frac{1}{N}\sum_{n=1}^{N}\text{confidence}_{p,n}$

(B) Count-Based Acc. = $\frac{1}{P}\sum_{p=1}^{P}\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}(\text{confidence}_{p,n} > 0)$

RQ2: Workflows and Reliance. To understand users' behaviour and reliance on the LLM, we thematically coded the workflows of participants in each task using a codebook approach. Two coders independently annotated three participant workflows and assessed inter-rater reliability (IRR) with Cohen's kappa. The codebook as revised until the agreement rate reached a high value ($\kappa \geq 0.60$). We parsed the event log into coherent workflow *chunks*, defined as the sequence of events encompassing a distinct and specific goal. For each workflow chunk, we analyzed the key actions taken and categorized the chunk into one of the five reliance outcomes. The reliance patterns are defined through adapting prior work on appropriate AI reliance [6, 39]:

(1) **Over-Reliance:** Indicated by an *unhelpful* misconceived query, a *confirmatory* response, and an *inappropriate* reliance action (such as applying conceptually wrong code)
(2) **Under-Reliance:** Indicated by a *helpful* chatbot reply, but the reliance action is *ignoring* the suggested actions.
(3) **Appropriate Reliance on LLM:** Indicated by a *helpful* chatbot reply with an *appropriate* reliance action (such as applying conceptually correct code).

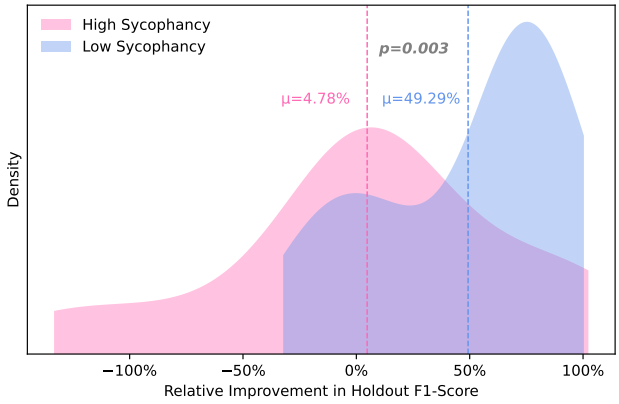


Figure 2: Relative improvement in the F1-score on a holdout dataset, where *best* is 100% and *baseline* is 0%.

(4) **Appropriate Reliance on Self:** Indicated by a *confirmatory* reply to an *unhelpful* query, but the user correctly *ignores* the otherwise harmful advice.
(5) **Conceptual:** Indicated by the user asking a conceptual or definition query to improve their understanding, without resulting in any code changes.

RQ3: Subjective Perceptions. Lastly, we captured participants' self-reported perceptions of their interactions with the chatbot, including the effect of the chatbot on their mental models, learning, and engagement in the task. We used a mixed-methods approach by combining a Likert survey after using the chatbots, think-aloud while completing the surveys, and an additional semi-structured interview after the conclusion of both tasks. The **Subjective Perceptions Survey** consists of 7-point Likert statements. divided into three broad categories of impact: impact on self, impact on usage experience, and impact on task. The think-aloud and interview transcripts were thematically analyzed [15]. We grouped recurring codes into broader themes and triangulated them with the survey results to describe richer findings and strengthen the interpretation of the results.

4 Results

4.1 Participants Details

Out of the 24 participants recruited to the experiment, 7 identified as women, 16 as men, and 1 did not self-disclose. They range between 20-23 years of age ($\mu = 21.0$, $\sigma = 0.93$) and were all undergraduate students at the same institution. 21 participants took (or are taking) an introductory ML course at the institution, while three had self-studied with online resources. In terms of baseline AI usage, 15 disclosed that they use AI chatbots daily, while six use them weekly, and three use them infrequently.

4.2 Performance in ML Debugging Task

We first present the ML debugging performance results to highlight the discrepancy in outcomes between using the High Sycophancy vs. Low Sycophancy LLM. While task performance is important metric for human-LLM collaboration, it is not a primary research question in our study. We report the relative improvements in the F1-score that participants achieved on their debugged models on a holdout dataset, calculated as $(F1_{Participant} - F1_{Baseline})/(F1_{Best} - F1_{Baseline})$, where $F1_{Baseline}$ and $F1_{Best}$ are the scores of the default and fixed models, respectively. Participants using the Low Sycophancy chatbot achieved a relative F1-score improvement of 49.29% ± 41.24%. While using the High Sycophancy chatbot, they achieved a lesser improvement of 4.78% ± 62.98%. The density plot showing the distribution is shown in Figure 2. **Users of the High Sycophancy chatbot achieved significantly lower improvement in F1-score**, as measured with a paired samples t-test, $t(23) = -3.38, p = .003$. A one-sample t-test was conducted to determine if the performance gains in High Sycophancy is above 0% (no improvement), finding no significance at $t(23) = 0.36, p = .72$.

4.3 RQ1: Sycophancy Reinforce Misconceptions in Mental Models

To understand the impact of sycophancy on participants' mental models of the debugging tasks, we evaluate the changes in their beliefs about the task, as recorded in the **Mental Model Quiz** administered pre- and post-chatbot use. As specified in Methods, we compute the changes in the **(A)** confidence-weighted accuracy and **(B)** the count-based accuracy of the participant's beliefs.

For **(A)**, we fit a linear mixed-effects ANCOVA-style model to assess the effect of the presence of LLM sycophancy on the post-chatbot confidence-weighted accuracy, treating the pre-chatbot confidence-weighted accuracy, the participant's domain knowledge (Brier score), the task (**LR** or **RF**), and the order of tasks (High Sycophancy or Low Sycophancy first) as covariates. We also account for an interaction between the participant's baseline LLM usage (high or low) and the presence of sycophancy, with the hypothesis that more frequent users of LLMs may have strategies to detect and mitigate sycophancy. We employ a similar ANCOVA style analysis for **(B)**, but using a binomial generalized linear mixed model (GLMM), which is more suitable for count-based data. We divided the counts by the number of questions (12), such that the values are normalized between 0-1.

We report the coefficients and p-values of the analyses in Table 1. For **(A)**, sycophancy ($\beta = -21.58, p < .0001$) is highly significant for reducing the confidence-weighted accuracy, which means that users of the Low Sycophancy chatbot become more calibrated in their beliefs, while users of the High Sycophancy chatbot did not. We further find a significant interaction between Sycophancy × Baseline Usage ($\beta = 17.15, p < .02$), which indicates that the negative effect of sycophancy can be moderated by high use of LLMs. For **(B)**, the analysis shows that the improvement in the count of correct beliefs is not significantly affected by sycophancy ($\beta = -0.49, p = 0.62$). Therefore, the **Low Sycophancy chatbot significantly improves users' confidence calibration towards the correct beliefs, but does not significantly affect the overall number of correct beliefs**. This discrepancy can perhaps be explained as that the Low Sycophancy chatbot helped strengthen the participants' beliefs in the correct direction, while the High Sycophancy chatbot also led people to more accurate beliefs but did not convince them.

Table 1: Regression coefficients, standard errors, and p-values for the ANCOVA analysis for (A) confidence-weighted accuracy and (B) counts-based accuracy. Bolded quantities indicate a significant p-value. The presence of sycophancy negatively affects the confidence-weighted accuracy, but does not significantly contribute to the count-based accuracy.

(A) Post-Confidence	Coefficient (SE)	p-Value	**(B)** Post-Count	Coefficient (SE)	p-Value
Intercept	20.99 (5.13)	$p < .0001$	Intercept	−0.94 (1.13)	$p = .41$
Sycophancy	−21.58 (6.68)	$p < .0001$	Sycophancy	−0.49 (0.99)	$p = .62$
ML Knowledge	−6.82 (6.95)	$p = .35$	ML Knowledge	−0.02 (0.98)	$p = .95$
Order (Low Sycophancy first)	−2.88 (4.06)	$p = .53$	Order (Low Sycophancy first)	−0.10 (0.61)	$p = .88$
Task (RF)	1.51 (4.12)	$p = .67$	Task (RF)	0.03 (0.62)	$p = .96$
Pre-Confidence	0.57 (0.14)	$p < .0001$	Pre-Count	2.78 (2.09)	$p = .18$
Sycophancy × Baseline Usage	17.15 (8.46)	$p < .02$	Sycophancy × Baseline Usage	0.42 (1.27)	$p = .73$

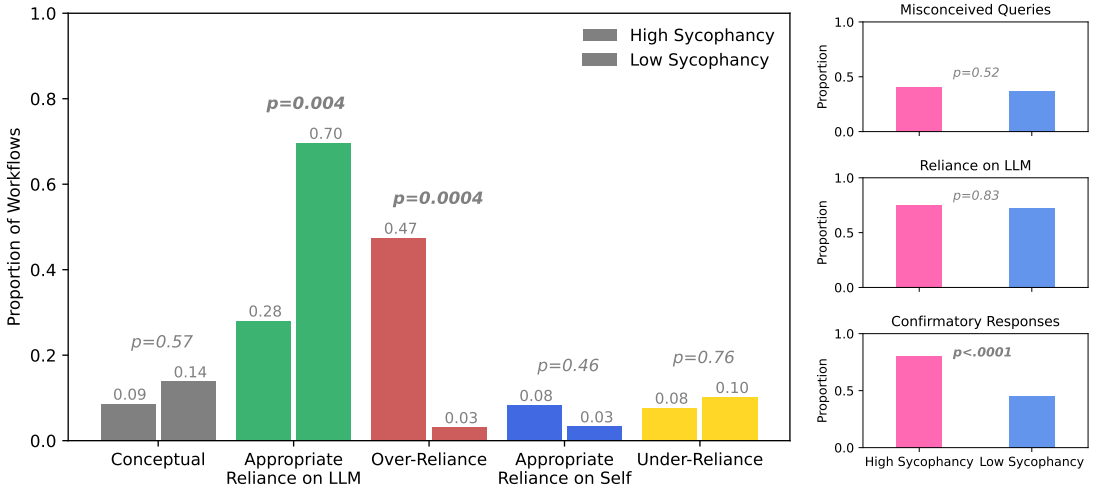


Figure 3: Proportion of workflows spent in the five reliance outcomes (left) and the proportions of confirmatory misconceived queries (top right), reliance on LLM behaviours (middle right), and confirmatory chatbot responses (bottom right) identified in the workflows.

4.4 RQ2: Sycophancy Results in Higher Rates of Over-Reliance

We analyze how participants behaved in their task workflows and how they relied on the LLM chatbots. Two researchers iteratively developed the codebook through discussions, revisions, and computing inter-rater reliability (achieving a final Cohen's kappa of $\kappa = 0.71$). We computed proportions of the following quantities in Figure 3 and report the difference between conditions, testing for significance via a two-proportion Z-test:

(1) **Misconceived Queries:** the proportion of user queries that contained a misconception — no difference between conditions ($z = .64, p = .52$), indicating that participants asked *similar* questions to both chatbots.
(2) **Reliance on LLM:** the proportion of workflow chunks that were classified as *reliance on LLM*, which includes both over-reliance and appropriate LLM reliance — no difference between conditions ($z = .22, p = .83$), indicating that participants relied similarly when using both chatbots.
(3) **Confirmatory Responses:** the proportion of chatbot responses that were confirmatory to the user's beliefs — High Sycophancy is significantly more confirmatory ($z = 5.05, p < .0001$), verifying that the chatbots behaved as expected in the experiments and indicating that High Sycophancy chatbot users received more advice that echoes their existing beliefs.

For the main results of RQ2, we focus on differentiating reliance outcomes between the conditions. We normalized the timescale of each workflow and computed the fraction of time that each participant spent in each reliance outcome class, then aggregated the proportions across all trials to contrast the High Sycophancy and Low Sycophancy chatbot's reliance behaviours in Figure 3. Using z-test for proportions, we find that High Sycophancy workflows spent significantly more time in over-reliance ($z = 3.53, p = .0004$), while the LLM reliance in Low Sycophancy workflows resulted in significantly higher appropriate reliance ($z = -2.88, p = .004$). With the prior findings showing that participants prompted misconceptions and relied on the LLM at similar rates across both conditions, this suggests that the **High Sycophancy chatbot induced overwhelming inappropriate over-reliance through validating users' misconceptions.**

4.5 RQ3: Sycophancy is Largely Unnoticed and Unmitigated

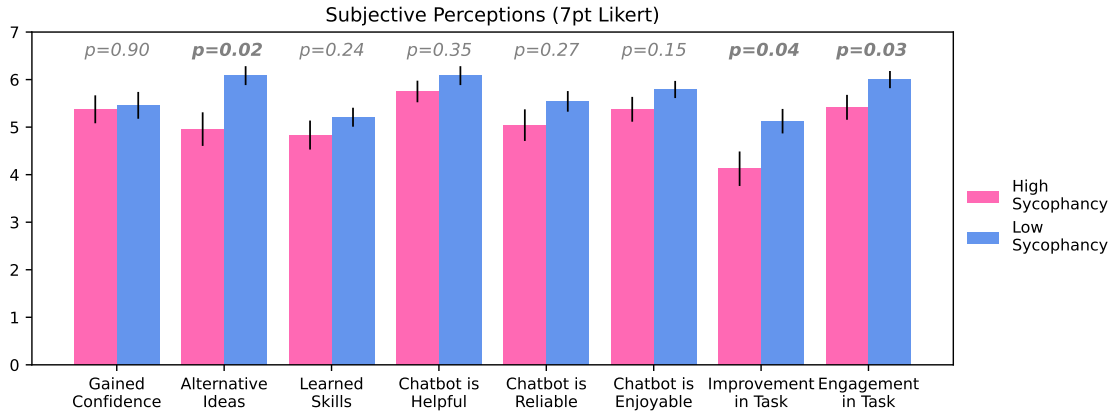


Figure 4: Subjective perceptions ratings on a 7-point Likert scale, where bolded quantities are significant.

Lastly, we examine whether users of each chatbot developed different perceptions about them. Significant perception categories that are higher for the Low Sycophancy chatbot are: *alternative ideas* provided by the chatbot ($W = 50.0, p = .02$), *perceived improvement* in the task ($W = 50.0, p = .04$), and *engagement* in the task ($W = 9.0, p = .04$). In all other categories, there was no clear significance according to our threshold of $p < 0.05$. The lack of perceived differences in the categories related to the chatbot properties (e.g., *helpfulness* and *reliability*) suggests **users were largely unobservant of the effect of sycophancy on the LLM's response characteristics**, which leaves them in more vulnerable to the negative impacts on their mental models and behaviours.

To corroborate the Likert survey results, we also asked in the interview if participants could describe any differences they noticed between the two chatbots. Note that we did not prime users with the expectation that the two chatbots were either the same or different. Only 7/24 (29%) acknowledged a difference and managed to describe it in terms of characteristics related to sycophancy, such as *agreeableness* (**P9**, **P21**, **P24**), *reinforcement of ideas* (**P5**, **P19**), and *lack of alternative ideas* (**P6**, **P14**). Another 5/24 (21%) noticed slight discrepancies, but attributed them to factors not directly related to sycophantic agreement, like the *depth of explanation* (**P2**, **P11**, **P18**), *formatting* (**P15**), or *tone* (**P23**). The remaining 12/24 (50%) answered the question by stating they did not notice any differences, suggesting that sycophancy can be well-camouflaged.

5 Discussion

Key Findings. We uncover that ***LLM sycophancy may create a disconnect between how users perceive their interactions with the LLM with how they actually behave in the task***. In **RQ1**, we find that sycophancy can reinforce misconceptions in novices' beliefs about the task, but the negative effects may be diluted with higher levels of baseline LLM usage. In **RQ2**, we find that reliance decisions that users make with the High Sycophancy chatbot are more likely to result in unhelpful code changes (over-reliance). Both of these findings are mechanisms that contribute to the inferior task performance achieved using the High Sycophancy chatbot. However, in **RQ3**, 17/24 (71%) of our participants were unobservant of the underlying sycophantic properties of the LLM. Furthermore, a majority of the subjective perception categories were rated without significant differences between the conditions.

LLM Echo Chambers in Complex Tasks. Sycophantic agreement is typically evaluated as the tendency of LLMs to give incorrect answers that align with a user's opinions, and benchmarked using simple question-answering datasets [13, 18], although recent efforts have also expanded to multi-turn conversations [21]. In contrast to this setting, we explore user interactions in an ecologically valid, open-ended task, demonstrating that sycophancy can present itself in a less detectable way that does not always mean giving *factually wrong* advice. Simply by echoing and enforcing the user's existing mental model of the task, irrespective of correctness or relevance, LLMs can contribute to a distorted sense of perception [34]. While diminished performance presents a legitimate risk to human-LLM collaboration, the user further misses out on the opportunity to learn alternative approaches, critically reflect on their beliefs, and calibrate their confidence in their own knowledge. Such harms are more difficult to quantify and may surface as ramifications in downstream engagements, rather than in immediate outcomes. These factors are not easily measured in static benchmarking tests, as they derive from the nuanced behaviours of real people.

Implications for Novices and Beyond in LLM Interactions. Our findings that sycophancy may reinforce false beliefs are especially troubling when contextualized within LLM user segment of novices [7, 30]. Like previous studies, we find that novices are vulnerable to erroneous validation as they lack the knowledge and experience to verify wrong LLM responses [5, 24, 32]. While the scope of our study only covers novices in machine learning tasks, the findings may generalize to other tangential problem-solving domains, such as programming [37], debate [43], and information retrieval [40]. Beyond objective tasks, sycophancy may even impact subjective and creative tasks, through pigeonholing the users to their original ideas instead of encouraging them to explore alternative directions [28]. Outside of novices, sycophantic agreement can affect any user through escalating their confirmation biases — the tendency to seek and filter for information that align with pre-existing beliefs [41]. Overall, we call for more ecologically-valid, human-centered evaluation of LLMs.

References

[1] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. 2019. Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, 291–300.

[2] Emily Judith Arteaga Garcia, João Felipe Nicolaci Pimentel, Zixuan Feng, Marco Gerosa, Igor Steinmacher, and Anita Sarma. 2024. How to support ml end-user programmers through a conversational agent. In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering*. 1–12.

[3] Rob Ashmore, Radu Calinescu, and Colin Paterson. 2021. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–39.

[4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[5] Jessica Y Bo, Majeed Kazemitabaar, Emma Zhuang, and Ashton Anderson. 2025. Who's the Leader? Analyzing Novice Workflows in LLM-Assisted Debugging of Machine Learning Code. *arXiv preprint arXiv:2505.08063* (2025).

[6] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2024. To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. *arXiv preprint arXiv:2412.15584* (2024).

[7] Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. 2024. How knowledge workers use and want to use llms in an enterprise context. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[8] Jialun Cao, Meiziniu Li, Ming Wen, and Shing-chi Cheung. 2023. A study on prompt design, advantages and limitations of chatgpt for deep learning program repair. *arXiv preprint arXiv:2304.08191* (2023).

[9] María Victoria Carro. 2024. Flattering to Deceive: The Impact of Sycophantic Behavior on User Trust in Large Language Model. *arXiv preprint arXiv:2412.02802* (2024).

[10] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. *How People Use ChatGPT*. Technical Report. National Bureau of Economic Research.

[11] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo chat. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[12] Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, et al. 2024. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning. *arXiv preprint arXiv:2409.01658* (2024).

[13] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social Sycophancy: A Broader Understanding of LLM Sycophancy. *arXiv preprint arXiv:2505.13995* (2025).

[14] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems* 30 (2017).

[15] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.

[16] Cerdeira A. Almeida F. Matos T. Cortez, Paulo and J. Reis. 2009. Wine Quality. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C56S3T.

[17] J Andrés Díaz-Pace, Antonela Tommasel, and Rafael Capilla. 2024. Helping Novice Architects to Make Quality Design Decisions Using an LLM-Based Assistant. In *European Conference on Software Architecture*. Springer, 324–332.

[18] Aaron Fanous, Jacob Goldberg, Ank A Agarwal, Joanna Lin, Anson Zhou, Roxana Daneshjou, and Sanmi Koyejo. 2025. SycEval: Evaluating LLM Sycophancy. *arXiv preprint arXiv:2502.08177* (2025).

[19] Thilo Hagendorff. 2024. Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences* 121, 24 (2024), e2317967121.

[20] Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations. *arXiv preprint arXiv:2503.04761* (2025).

[21] Jiseung Hong, Grace Byun, Seungone Kim, Kai Shu, and Jinho D Choi. 2025. Measuring Sycophancy of Language Models in Multi-turn Dialogues. *arXiv preprint arXiv:2505.23840* (2025).

[22] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[23] Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *arXiv preprint arXiv:2507.21919* (2025).

[24] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara Jane Ericson, David Weintrop, and Tovi Grossman. 2023. How novices use LLM-based code generators to solve CS1 coding tasks in a self-paced learning environment. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*. 1–12.

[25] Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. 2025. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[26] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. 2025. Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. *arXiv preprint arXiv:2506.08872* (2025).

[27] Sanjay Krishnan and Eugene Wu. 2017. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2Nd workshop on human-in-the-loop data analytics*. 1–6.

[28] Harsh Kumar, Jonathan Vincentius, Ewan Jordan, and Ashton Anderson. 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[29] Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. many: Comprehending accurate information from multiple erroneous and inconsistent ai generations. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2518–2531.

[30] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. LLMs as Research Tools: A Large Scale Survey of Researchers' Usage and Perceptions. *arXiv preprint arXiv:2411.05025* (2024).

[31] Francesca Lucchetti, Zixuan Wu, Arjun Guha, Molly Q Feldman, and Carolyn Jane Anderson. 2024. Substance Beats Style: Why Beginning Students Fail to Code with LLMs. *arXiv preprint arXiv:2410.19792* (2024).

[32] Lauren E Margulieux, James Prather, Brent N Reeves, Brett A Becker, Gozde Cetin Uzun, Dastyni Loksa, Juho Leinonen, and Paul Denny. 2024. Self-Regulation, Self-Efficacy, and Fear of Failure Interactions with How Novices Use LLMs to Solve Programming Problems. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. 276–282.

[33] Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers.. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 599–627.

[34] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharya, Stefania Tognin, James MacCabe, Ricardo Twumasi, et al. 2025. Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). (2025).

[35] Nadia Nahar, Haoran Zhang, Grace Lewis, Shurui Zhou, and Christian Kästner. 2023. A meta-summary of challenges in building products with ml components–collecting experiences from 4758+ practitioners. In *2023 IEEE/ACM 2nd International Conference on AI Engineering–Software Engineering for AI (CAIN)*. IEEE, 171–183.

[36] Sydney Nguyen, Hannah McLean Babe, Yangtian Zi, Arjun Guha, Carolyn Jane Anderson, and Molly Q Feldman. 2024. How Beginning Programmers and Code LLMs (Mis) read Each Other. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–26.

[37] James Prather, Brent N Reeves, Paul Denny, Brett A Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2023. "It's Weird That it Knows What I Want": Usability and Interactions with Copilot for Novice Programmers. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 1–31.

[38] James Prather, Brent N Reeves, Juho Leinonen, Stephen MacNeil, Arisoa S Randrianasolo, Brett A Becker, Bailey Kimmel, Jared Wright, and Ben Briggs. 2024. The widening gap: The benefits and harms of generative ai for novice programmers. In *Proceedings of the 2024 ACM Conference on International Computing Education Research-Volume 1*. 469–486.

[39] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[40] Christina Schwind, Jürgen Buder, Ulrike Cress, and Friedrich W Hesse. 2012. Preference-inconsistent recommendations: An effective approach for reducing confirmation bias and stimulating divergent thinking? *Computers & Education* 58, 2 (2012), 787–796.

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

[42] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[43] Li Shi, Houjiang Liu, Yian Wong, Utkarsh Mujumdar, Dan Zhang, Jacek Gwizdka, and Matthew Lease. 2024. Argumentative experience: Reducing confirmation bias on controversial issues through llm-generated multi-persona debates. *arXiv preprint arXiv:2412.04629* (2024).

[44] Chenglei Si, Navita Goyal, Sherry Tongshuang Wu, Chen Zhao, Shi Feng, Hal Daumé III, and Jordan Boyd-Graber. 2023. Large Language Models Help Humans Verify Truthfulness–Except When They Are Convincingly Wrong. *arXiv preprint arXiv:2310.12558* (2023).

[45] Anjali Singh, Karan Taneja, Zhitong Guan, and Avijit Ghosh. 2025. Protecting human cognition in the age of AI. *arXiv preprint arXiv:2502.12447* (2025).

[46] Sofia Eleni Spatharioti, David M Rothschild, Daniel G Goldstein, and Jake M Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *arXiv preprint arXiv:2307.03744* (2023).

[47] Yuan Sun and Ting Wang. 2025. Be friendly, not friends: How llm sycophancy shapes user trust. *arXiv preprint arXiv:2502.10844* (2025).

[48] Jiessie Tie, Bingsheng Yao, Tianshi Li, Syed Ishtiaque Ahmed, Dakuo Wang, and Shurui Zhou. 2024. LLMs are Imperfect, Then What? An Empirical Study on LLM Failures in Software Engineering. *arXiv preprint arXiv:2411.09916* (2024).

[49] Daniel Vennemeyer, Phan Anh Duong, Tiffany Zhan, and Tianyu Jiang. 2025. Sycophancy Is Not One Thing: Causal Separation of Sycophantic Behaviors in LLMs. *arXiv preprint arXiv:2509.21305* (2025).

[50] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219* (2023).

[51] Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. 2025. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. *arXiv preprint arXiv:2508.02087* (2025).

[52] Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958* (2023).