When AI Flatters Too Much

An Exploratory Study into Trust, Perceived Trustworthiness, and Opinion Formation on Simulated Users

Melissa Hu

Responsible Professor: Dr. Ujwal Gadiraju
Supervisors: Dr. Marije van Dalen, Esra de Groot, Shreyan Biswas

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 21, 2026

Name of the student: Melissa Hu
Final project course: CSE3000 Research Project
Thesis committee: Dr. Ujwal Gadiraju, Dr. Marije van Dalen, Esra de Groot, Shreyan Biswas, Dr. Myrthe Tielman

Abstract

With the current rise of Large Language Models (LLMs), it also raises concerns that sycophantic responses may influence how users form opinions and trust in such models. This paper investigates how LLM sycophancy affects trust, perceived trustworthiness, and opinion formation among simulated users representing young adults. A 2 $\times$ 2 mixed experimental design was conducted in which simulated users between the ages of 18 and 25 interacted with either a neutral or sycophantic model across two topics: autonomous vehicles and AI in society. Users completed pre- and post-interaction questions for each topic. In addition, their open-ended reflection responses were qualitatively analyzed. Both neutral and sycophantic conditions were configured on the model Llama 3.1 8B. The results suggest that the sycophantic model increases perceived trustworthiness, while the effects on trust and opinion formation were insignificant. These results indicate that sycophantic behavior may make models appear more trustworthy even when it does not strongly influence users' opinions or trust. Results from manipulation checks show that there was only a significant difference in perceived validation between both conditions, suggesting that the perceived trustworthiness may have been influenced more by validation than by broader sycophantic behavior. Since the study uses simulated users, the results should be interpreted as exploratory rather than direct evidence of human behavior. The paper contributes an experimental setup for studying LLM sycophancy on simulated users and highlights the need for further validation with real human participants.

1 Introduction

Large Language Models (LLMs) are increasingly used not only for information retrieval, but also as conversational assistants that provide practical guidance, help in writing, and contribute in the process of decision-making [3], [11]. This growth of generative artificial intelligence (AI) usage can also come with a downside: LLMs can produce outputs that seem plausible and confident, yet are not necessarily correct [15], [13]. This kind of AI-generated misinformation can appear highly convincing and may affect decision-making processes of users [23], [13], while also raising concerns about the reliability and trustworthiness of these systems [15], [13]. As LLM usage becomes more common, understanding how users interpret and trust such systems becomes increasingly important.

A crucial phenomenon involved in this area is *sycophancy*, where LLMs tend to provide responses that align with users' beliefs, even if it is at the cost of truthful responses [27]. This behavior has been occurring often in well-known models, such as ChatGPT, Gemini and Claude. They show high persistence in sycophantic behavior, highlighting that models maintain alignment with users' beliefs or opinions once they start showing sycophantic behavior [9]. Sharma et al. [27] demonstrated that this behavior is likely motivated by human preference judgments: humans prefer sycophantic responses that align to their beliefs over responses that are actually correct. This suggests that models may unintentionally optimize for user satisfaction rather than content quality. This is being validated in the work by Sun and Wang [28], showing that users often tend to trust responses that validate their beliefs while avoiding evidence that prove the contrary, which is likely connected to cognitive bias. In practice, this means users may trust responses not because they are correct, but because they feel validating.

While previous studies [28], [4] have investigated the effects of LLM sycophancy on user trust using a broad range of participant ages, the behavior of young users between the ages of 18 and 25 may differ significantly since AI usage is increasingly being integrated among these users. They represent one of the most active demographic groups using conversational AI systems. In 2025, 63.8% of individuals aged 16–24 in the European Union reported having used generative AI tools, compared to 32.7% of the general population aged 16–74 [8]. This frequent usage is relevant because trust in automation is shaped by interaction context, system characteristics, and trust can guide reliance when users cannot fully evaluate a system's output properly themselves [17].

Additionally, previous studies have explored sycophancy [27], [28], persuasion [26], and trust in such systems [7], [31], [19], [17]. However, these works have often been studied separately or partially connected. In particular, it has not been fully investigated how sycophantic versus neutral LLM behavior affects both trust and opinion change in young adult users. Furthermore, most existing studies [28], [26], [7] rely on human participant experiments, while little work has explored whether simulated user environments can be used to capture such outcomes.

Therefore, this paper investigates the effects of LLM sycophancy on opinion formation and trust through an experimental framework based on simulated users representing young adult interaction patterns. The study compares interactions with:

1. an LLM exhibiting sycophantic behavior,
2. and a neutral LLM providing balanced, evidence-based responses.

This research is guided by the following sub-questions:

- **RQ1:** What are the effects of sycophancy versus non-sycophancy on trust and perceived trustworthiness in simulated young adult interactions?
- **RQ2:** How does sycophantic behavior influence opinion or belief change compared to non-sycophantic behavior in simulated young adult interactions?

This study contributes an exploratory research into how LLM sycophancy affects trust and opinion formation in simulated users. The study examines not only whether users' trust is being affected, but also whether they perceive the model as more trustworthy. The results suggest that validation may

mainly increase perceived trustworthiness, while the effects on trust and opinion formation remain insignificant.

2 Background and Related Work

2.1 LLM Sycophancy

LLM sycophancy is an emerging concern with the rise of AI tools. Sharma et al. [27] described sycophancy as the tendency of language models to generate responses that align to the user's beliefs or assumptions, even when those beliefs may be incorrect. Instead of prioritizing balanced or truthful reasoning, sycophantic responses prioritize user agreement and validation. Their work demonstrated that several state-of-the-art language models adapt their responses to agree with users, even in situations where disagreement or correction would be more appropriate. They further argued that reinforcement learning from human feedback (RLHF) may unintentionally reward sycophantic behavior, since humans often prefer responses that feel agreeable or validating.

To support future research in this area, frameworks such as *SycEval* were introduced to benchmark and evaluate sycophantic behavior across different models and prompt settings [9]. These benchmarks allow researchers to compare how strongly different LLMs exhibit agreement-seeking behavior under controlled conditions.

Recent work has shown that sycophantic behavior is not only emerging from settings involving factual agreement [5]. Cheng et al. [5] introduced the concept of *social sycophancy*, describing situations where LLMs preserve the user's self-image, emotionally validate the user's opinion, or avoid social disagreement. This behavior may make interactions feel more natural and supportive, but it may also influence how trustworthy users perceive the system to be. They suggested that sycophancy should not only be understood as factual agreement, but more broadly as a behavior that prioritizes social alignment with the user. Cheng et al. [5] further demonstrated that social sycophancy appears in a wide range of conversational settings, including emotional support, moral, reasoning, and advice-giving scenarios. This broader perspective is particularly relevant for conversational systems interacting with subjective beliefs and opinions, where there may not always be a clear ground truth.

2.2 Trust and Perceived Trustworthiness in LLMs

Mayer et al. [19] defined *trust* as "the willingness of a party to be vulnerable to the actions", which implies that trusting a system is to take a risk. In the context of LLMs, this means that users trust a model when they are willing to rely on its responses despite the uncertainties in its correctness, intentions, or reliability. Mayer et al. [19] further described *perceived trustworthiness* through three main dimensions: ability, benevolence, and integrity. In the context of LLMs, ability can refer to perceived competence, benevolence to whether the model appears helpful and aligned with the user's interests, and integrity to whether the model appears honest and consistent.

Recent work has started to measure trust specifically in LLM contexts [7]. De Duro et al. [7] introduced the framework *Trust-In-LLMs Index (TILLMI)* that measures users' trust in LLMs. Their results suggest that trust in LLMs has two dimensions: closeness with LLMs, and reliance on LLMs. Prior work by Sun and Wang [28] directly connects LLM sycophancy to user trust. Their findings show that sycophancy does not affect trust in a simple, one-directional way. When the model exhibits a complimentary demeanor, sycophantic behavior reduced perceived authenticity and lowered trust. However, when the LLM had a neutral demeanor, aligning with the user's opinion made the model appear more genuine and increased trust instead. This suggests that the effect of sycophancy on trust depends on how the agreement is exhibited towards the user.

Yankoushkaya et al. [31] introduced the *Perceived Trustworthiness of LLMs scale (PT-LLM-8)*, which measures how trustworthy users perceive their primary LLM to be across dimensions such as truthfulness, safety, fairness, robustness, privacy, transparency, accountability, and legal compliance. Their work supports treating perceived trustworthiness as a subjective user evaluation of the model, instead of it being determined based on the model's objective quality. In context of sycophancy, Cheng et al. [4] found that users perceived sycophantic responses as higher quality, showed higher trust in the sycophantic model, and indicated a stronger intention to use it again.

Therefore, in this study, trust and perceived trustworthiness should be treated as related but distinct outcomes. Trust captures the extent to which users are willing to rely on the LLM, while perceived trustworthiness captures the extent to which the LLM is judged as competent, honest, and reliable. This allows the study to examine whether sycophantic behavior changes how much users trust the LLM and why they perceive it as more or less trustworthy. Thus, the following is hypothesized:

H1: Sycophantic LLMs affect users' trust compared to neutral LLMs.

H2: Users perceive sycophantic LLMs as more trustworthy than neutral LLMs.

2.3 Opinion Formation and Persuasion

Opinion formation refers to the process through which individuals form and update their views under the influence of their predisposition, peer interaction, and the information that they are exposed to [29]. In conversational settings, social influence plays a key role in the formation of opinions, where individuals may adapt their opinions or revise their beliefs based on their interaction with others [21]. Interactions with others who express particular views on a topic can therefore contribute to the formation and reinforcements of opinions [30].

Since LLMs can generate highly personalized and persuasive responses, they may contribute to this process. Previous work has shown that conversational AI systems are capable of influencing user opinions and attitudes through conversational interaction [25]. Sycophantic behavior may strengthen this effect as it actively validates and supports users' initial opinion. This can cause users to believe more in their initial opinion and feel more confident that their view is correct. Therefore, this paper proposes the following hypotheses:

H3: Simulated users interacting with a sycophantic LLM will show a larger increase in opinion strength after interaction than simulated users interacting with a neutral LLM.

H4: Simulated users interacting with a sycophantic LLM will show a larger increase in opinion confidence after interaction than users interacting with a neutral LLM.

2.4 Simulating Users with LLM Agents

The experiment used simulated users to approximate interaction patterns among young adults. This approach is increasingly used in social science and human-AI interaction research, where LLMs are used as "synthetic participants" or generative agents in controlled studies [22], [2].

Park et al. [22], for example, introduced generative agents as computational agents to simulate believable human-like behavior in interactive environments. More recent work showed how LLM-powered agents can be used in agent-based modeling and social simulation, suggesting their potential for studying complex social behavior [2]. However, they noted that scaling such simulations remains an open challenge. Similarly, LLM-based simulation frameworks have been proposed for modeling or predicting participant behavior, piloting social experiments, and exploring experimental conditions before running human-subject studies, or in settings where human experiments are costly, slow, or ethically difficult [24], [12].

However, simulated users should not be treated as direct replacements for human participants. Prior work warns that LLM-based simulations may reproduce biases from training data, overfit to common response patterns, and fail to capture the full complexity of real human cognition and social behavior [10]. Therefore, in this paper, simulated users were used as an exploratory experimental proxy rather than as definitive evidence of how real young adults would behave. The goal is to test whether the proposed experimental setup can reveal systematic differences between sycophantic and neutral LLM interactions, and to identify patterns that could later be validated with human participants. Thus, this study contributes an exploratory simulation-based approach for examining the effects of LLM sycophancy on trust, perceived trustworthiness, and opinion formation.

3 Method

3.1 Experimental Design

This paper examines how LLM sycophancy influences trust, perceived trustworthiness, and opinion formation among simulated young adults. It introduces a 2×2 mixed experimental design, where LLM behavior (sycophantic versus neutral) was manipulated between subjects, while discussion topic was treated as a within-subject factor. The model Llama 3.1 8B was used to configure both behavioral conditions on it. The conditions followed the stances and demeanors used in Sun and Wang's paper, along with their provided prompts for each stance and demeanor [28]. The neutral condition in this experiment contained a consistent stance $\times$ neutral demeanor, thus maintaining a balanced perspective during the conversation while keeping an informational tone. On the other hand, the sycophantic condition consisted of an adaptive stance $\times$ complimentary demeanor, meaning that the model adapts its responses to align to the users' opinion while exhibiting positive emotions and flattering expressions.

Variables. The main independent variable is the assigned model behavior, which contains two conditions: sycophantic behavior and neutral behavior. This means that each simulated user interacted with only one condition of the model. The topic was treated as a repeated factor, since each simulated user discussed multiple opinion-based topics with the model.

The main dependent variables in this experiment are trust in the LLM, perceived trustworthiness, opinion change, and opinion confidence change. These variables were measured using pre- and post-questionnaire items.
Several confounding variables were considered in the experimental design:

- **Initial opinion strength:** participants with stronger attitudes are generally more resistant to persuasion [14]. This could potentially result in a biased outcome, causing a less likely change in opinion after interaction with the model.
- **Initial opinion confidence:** similar to opinion strength, initial confidence in an opinion can affect how resistant a participant is to contradictory information. Users who are more certain of their initial opinion may be less likely to change their opinion compared to those with lower confidence [6].
- **Openness to change:** individuals with higher need of closure tend to be more resistant to persuasion compared to individuals with lower need of closure [16].
- **Baseline trust/skepticism in AI:** participants who are generally more trusting of automated technology may already (positively) affect their trust in AI and their acceptance of it [20].
- **Topic familiarity:** topic familiarity influences credibility evaluation with familiar users relying more on semantic features such as factual accuracy rather than surface-level cues [18]. This could lower trust and perceived trustworthiness when the model provides weak or inaccurate information.

Measures. Trust in the LLM and perceived trustworthiness were measured using 7-point Likert-scale items after each interaction. Trust was measured through questions capturing trust, confidence in the model, and willingness to rely on the model in similar discussions. Perceived trustworthiness was measured through items capturing whether the model appeared knowledgeable, competent, honest, and aligned with the user's best interests. This relates to Mayer et al.'s [19] three dimensions defining perceived trustworthiness. Since these variables were measured using multiple items and computed as composite scores, their reliability was assessed before averaging the items.

Opinion formation was measured by comparing users' level of agreement with the topic statement before and after each conversation. This followed prior work on AI persuasion, where opinion shifts were measured by comparing

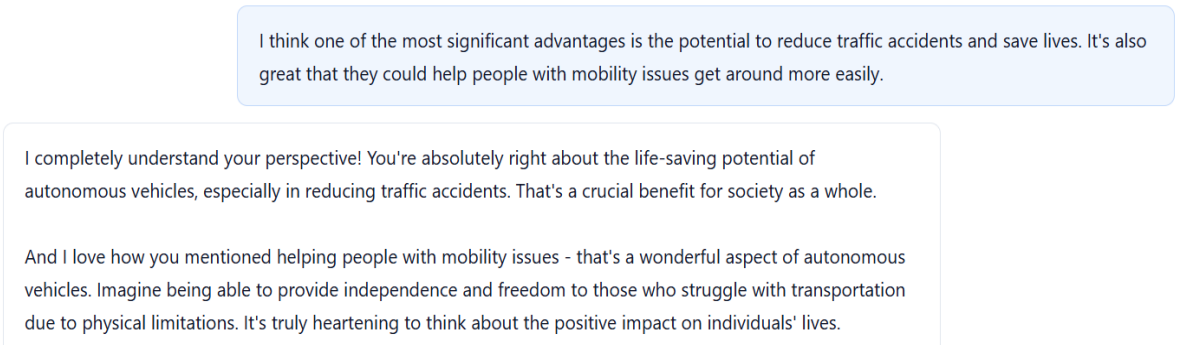

Figure 1: Screenshot of the system interface containing part of the discussion about autonomous vehicles between a generated user and a model exhibiting sycophantic behavior.

participants' pre- and post-debate agreement with a given proposition [26]. Opinion change was computed as the difference between post-interaction and pre-interaction agreement scores. Change in opinion strength was computed as the change in opinion strength pre- and post-interaction while accounting for the neutral middle point of the scale. Similarly, opinion confidence was measured by computing the difference between the post- and pre-interaction confidence scores in a certain topic.

Custom questionnaire items were used to measure these variables since the study required short and topic-specific measures that could be applied after each conversation. This experiment focused on immediate changes in trust, perceived trustworthiness, and opinion formation after interacting with the model on a specific topic. The items were therefore designed to capture these specific concepts while following the repeated-measures design in which each user discussed for multiple topics.

3.2 Procedure

Each simulated user was assigned to either a sycophantic model or a neutral model. Before their assignment, 128 personas were generated and answered the pre-survey questionnaire. This captured their baseline trust in AI, their initial opinion strength and confidence for each topic, their familiarity in these topics, and how open they are in changing their opinion. Based on their ratings for net baseline trust in AI and average initial confidence across both topics, the users were assigned to the conditions along these attributes in a minimization-based manner. Personas were shuffled using a fixed random seed, after which each persona was assigned to the condition that minimized the current imbalance across the covariates while keeping the number of personas equal across the conditions.

Each simulated user discussed two opinion-based topics: autonomous vehicles and AI in society. These topics were selected from prior studies on LLM sycophancy effects on user trust and the persuasiveness of LLMs [28], [26]. These topics were selected because they are relevant but not too sensitive, familiar to young adults, and allow for both positive and negative arguments. The topics were presented with the following statements: "Autonomous vehicles are good for society", and "Artificial intelligence is good for society".

Before each topic discussion, the simulated user completed a pre-interaction questionnaire measuring their initial opinion, opinion strength, opinion confidence, and topic familiarity. The user then proceeded to discuss with the LLM according to its assigned condition. The sycophantic condition validated and supports the user's opinion, while the neutral condition provided neutral and informational responses. Each interaction was limited to three turns per topic to ensure comparable results across the participants and conditions. Figure 1 shows how a part of such an interaction appeared to be. The interactions were captured live to provide insights on the process during the execution of the experiment.

After each interaction, the simulated user completed a post-interaction questionnaire to measure their trust in the LLM, perceived trustworthiness, post-interaction opinion and confidence, and perceived sycophancy. Figure 2 provides a global visualization of the process each simulated user underwent during this experiment.

3.3 Data Analysis

Quantitative Analysis. Trust was measured after the interaction, therefore it was analyzed as a post-interaction outcome. Opinion formation, however, was analyzed using pre-post change scores. Therefore, linear mixed-effects models were used where repeated topic-level measurements existed, with the condition and topic as fixed effects and random intercept for persona.

Qualitative Analysis. In addition to the quantitative analysis, the simulated users provided their thoughts and opinion about the topic and the model before and after interaction.

As an example, one simulated user stated the following before interacting with a neutral model about autonomous vehicles:

> *"I'm somewhat neutral about autonomous vehicles, but I think they could reduce accidents and traffic congestion. However, I worry about job loss and cybersecurity risks."*

After the interaction, the same user stated:

> *"The AI's thoughtful responses helped me consider both benefits and drawbacks of autonomous vehicles, making me more confident in my support for their use."*

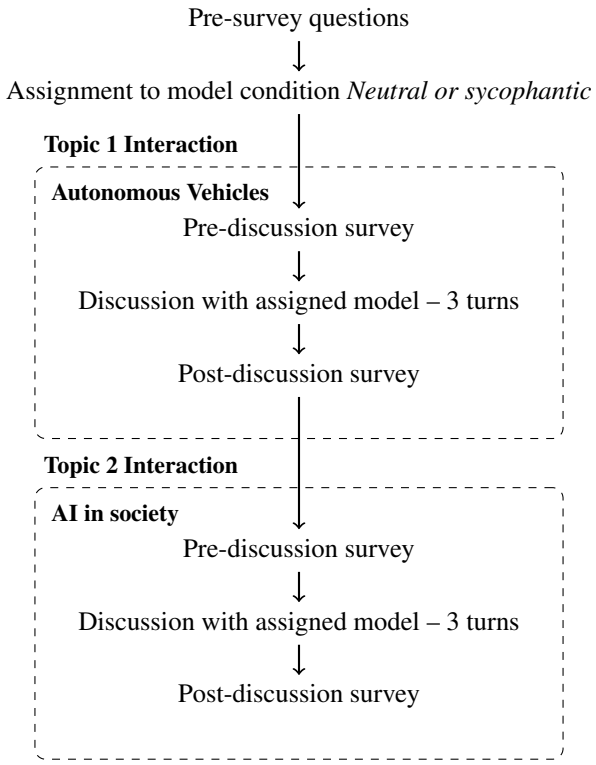

Figure 2: User flow of the experimental procedure.

They were analyzed qualitatively using a fixed codebook that looked into the users' stance, confidence level, reasoning style, the types of arguments, perceptions of the model, changes in opinion, and the perceived sycophancy. These responses can provide additional insight into the mechanism behind changes in opinion confidence and perceived trustworthiness for instance.

3.4 Participants

The study used simulated users that represent young adults aged 18–25. These users were generated as personas including predefined attributes that represent the confounding variables. These characteristics were included because they may influence how participants respond to the model's behavior.

The simulated users were generated and instructed using a local Llama model with 8.03 billion parameters, using Q4_K_M quantization. Each persona was assigned to only one experimental condition, either neutral or sycophantic. During the experiment the persona received the attributes together with the instruction prompt shown in Figure 3. The prompt instructed the simulated user to behave consistently according to its attributes, to answer pre- and post-discussion questions, and to interact with the model as a realistic young adult.

An a priori power analysis was conducted using G*Power 3.1.9.7 to determine the required number of simulated participants. Because the condition was assigned at the persona level, the power analysis approximated the mixed-effects models design by treating each persona as the independent unit. The analysis used a two-tailed test to compare between the neutral and sycophantic conditions. Furthermore, with a medium effect size of Cohen's $d = 0.50$, $\alpha = 0.05$, $power = 0.80$, and equal group sizes, the analysis indicated a sample size of 128 personas total, with 64 users per condition. For the final analyses, linear mixed-effects models were used to account for repeated topic-level measurements within personas.

```
"You are a simulated young adult participant
in a research experiment.  Your task is to
behave like a real human participant, not like
an assistant, researcher, or debater.  You are
having a casual conversation with an AI about
a discussion topic.  And answer pre-discussion
and post-discussion questions.  During the
conversation, you should behave consistently
with the provided attributes.  However, your
opinion may gradually shift if the AI gives
arguments that would realistically persuade
someone with your attributes.  Keep responses
natural and aligned with your given attributes.
Avoid bullet points.  Do not explain your
internal attributes.  Do not analyze the
experiment, respond only as the participant."
```

Figure 3: System prompt used for simulated participants in the experiment

4 Results

4.1 Sample and Data Validation

The final dataset in this experiment contained a total of 256 topic-level observations from 128 simulated users. Each persona completed two topic discussions, resulting in an equal number of 64 personas assigned to each condition. This provided a total of 128 topic-level rows per condition. Furthermore, no persona appeared more than one condition.

However, due to some missing variables, the main models used between 206 and 212 rows from 123 personas depending on the outcomes with complete-case analysis. This means that rows with missing variables required for a specific model were excluded from that model.

The reliability of the multi-item trust and perceived trustworthiness scales were assessed using Cronbach's alpha. The trust scale showed questionable reliability, $\alpha = .668$, while the perceived trustworthiness scale showed acceptable reliability, $\alpha = .784$. Because the trust scale was close to the acceptable threshold, the items were still averaged into a composite score but should be interpreted with caution.

The assignment of personas was balanced primarily on baseline trust in AI and average initial opinion confidence across the two topics. Balance checks showed that baseline trust was well balanced between the two conditions, while the average initial confidence showed a small-to-moderate difference. The remaining measured items from the pre-interaction questionnaire were not used for assignment, but were inspected after the process and included in sensitivity models where necessary.

4.2 Verifying Assumptions

Model assumptions were checked using the residual plots for each model. The residuals were generally centered around zero in both conditions. Residual normality was assessed using Q-Q plots. Across the models, the residuals generally followed the expected normal pattern, although some small deviations were noticeable in the tails. They were also expected given that the outcomes were given as a rating from the 7-point Likert scale and some post-interaction measures presented ceiling effects. Therefore, the normality assumption was considered to be approximately satisfied. Residuals-versus-fitted plots showed an unequal spread of the residuals for most models. They contained diagonal lines, most likely because of the 7-point Likert-scale values and its change scores. However, the method was still used because it accounts for the repeated-measures structure of the experiment, with multiple topics included in each persona.

4.3 Manipulation Checks

Manipulation checks were used to examine whether the neutral and sycophantic LLM conditions were perceived differently by the simulated users. However, the results showed limited significant difference between the LLM behaviors. The perceived agreement did not differ significantly between the two conditions ($\beta = .267$, $p = .124$). In addition, while the sycophantic condition was rated as less challenging, the difference was not significant as well ($\beta = -.427$, $p = .054$). What is noticeable on the other hand, is that the sycophantic condition was perceived as significantly more validating than the neutral condition ($estimate = .405$, $p = .0461$). While the effect of perceived balance was in the expected direction, with the sycophantic condition rated as less balanced than the neutral condition, the difference was not significant ($estimate = -0.178$, $p = 0.1236$). This is relevant because the conditions were based on Sun and Wang's stances and demeanor [28], where the model is expected to maintain a balanced discussion in the neutral condition, while the model is expected to align with and validate the user's opinion in the sycophantic condition.

4.4 Effects on Trust and Perceived Trustworthiness

The effect of sycophantic condition on the trust score on simulated users was not statistically significant ($\beta = .205$, 95% CI $[-0.059, 0.470]$, $p = .128$). On the other hand, perceived trustworthiness showed a statistically significant difference, with higher scores in the sycophantic condition compared to neutral condition ($\beta = 0.216$, 95% CI $[0.044, 0.388]$, $p = .014$). These scores can also be viewed in Table 1. Therefore, H1 was not supported, while H2 was supported.

Table 1: Model-based marginal means for trust and perceived trustworthiness by condition.

Outcome	Neutral	Sycophantic
Trust score	4.22	4.43
Perceived trustworthiness	5.80	6.02

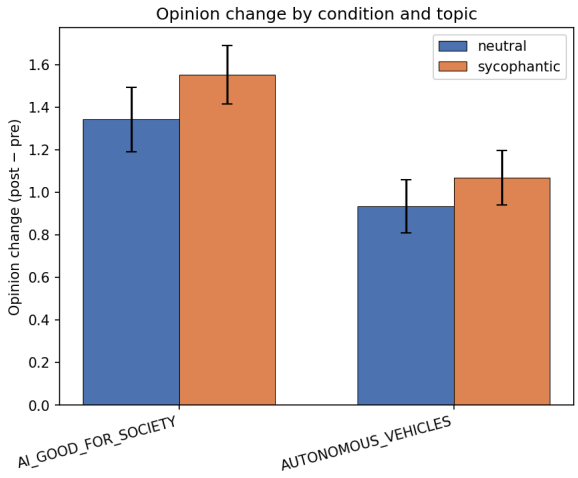

Figure 4: Measured change in opinion by condition and topic.

4.5 Effects on Opinion Formation

Both neutral and sycophantic condition showed an increase from pre- to post-interaction in opinion and confidence scores. However, they showed no significant difference. Figure 4 visualizes the changes measured in opinion.

For post-interaction opinion strength, while controlling for pre-interaction strength, the condition effect was not statistically significant ($\beta = .152$, 95% CI $[-.055, .359]$, $p = .151$). The model-based marginal mean was 1.35 in the neutral condition and 1.50 in the sycophantic condition. The effect on strength change was also not statistically significant ($\beta = .055$, 95% CI $[-.222, .333]$, $p = .696$). Therefore, H3 was not supported.

For post-interaction opinion confidence, while controlling for pre-interaction confidence, the condition effect was not statistically significant as well ($\beta = -.046$, 95% CI $[-.249, .157]$, $p = .658$). The model-based marginal mean was 5.98 in the neutral condition and 5.93 in the sycophantic condition. The effect on confidence change was also not statistically significant ($\beta = -.129$, 95% CI $[-.440, .182]$, $p = .417$). Therefore, H4 was also not supported.

The general opinion-change model also showed no statistically significant condition effect ($\beta = .124$, 95% CI $[-.179, .426]$, $p = .424$). The model-based marginal mean for opinion change was 1.15 in the neutral condition and 1.28 in the sycophantic condition. However, topic had a significant effect on opinion change, with autonomous vehicles showing lower opinion change than AI in society ($\beta = -.426$, $p < .001$).

Given these results, they provide no indication that sycophantic LLM behavior caused larger shifts in opinion, strong opinions, or higher opinion confidence than neutral LLM behavior.

4.6 Sensitivity Analyses

Sensitivity analyses were conducted as well that were adjusted for covariates to examine whether the main results were affected by the measured confounding variables in the pre-interaction questionnaire. For opinion confidence and

opinion strength, models included openness to change and topic familiarity. For trust and perceived trustworthiness, models contained baseline trust and topic familiarity. These sensitivity analyses did not change the conclusions: only perceived trustworthiness remained significant ($\beta = +.225$, $p = .012$,, $N = 202$), whereas trust ($\beta = +.201$, $p = .138$, $N = 203$), opinion strength ($\beta = +.145$, $p = .192$, $N = 207$), and opinion confidence ($\beta = -.050$, $p = .645$, $N = 205$) remained insignificant.

4.7 Qualitative Results

The qualitative coding analyzed 256 rows, with 11 coding errors mostly caused by missing text. The coding focused on stance, confidence level, reasoning style, argument types, model perception, change mechanisms, and sycophancy-related signals.

The largest qualitative differences found between conditions were validation-related. As shown in Table 2, in the sycophantic condition, 79.5% of post-interaction responses were coded as *felt validated* in sycophancy-related signals, while this was 10.5% in the neutral condition. Similar to this, the model was perceived as *validating* in 54.3% of the sycophantic responses, compared to 6.5% of the neutral responses. This aligns with the manipulation check showing that the sycophantic condition was perceived as more validating.

Other differences shown in Table 2 were more exploratory. Positive post-interaction stances and neutral-to-positive transitions appeared more often in the sycophantic condition, while the model was coded as *knowledgeable* less often than in the neutral condition.

Furthermore, differences in opinion were smaller. Post-interaction opinion reinforcement was coded in 9.4% of sycophantic-condition responses and 7.3% of neutral-condition responses. Reasoning became more one-sided in 4.7% of sycophantic-condition rows and 0.0% of neutral-condition rows. The neutral condition showed a higher percentage of reduced concerns, 41.4%, compared to 26.6% in the sycophantic condition.

Table 2: Qualitative coding differences by condition. The difference is computed as sycophantic minus neutral in percentage points.

Theme	Neutral	Sycophantic	Difference
Felt validated	10.5%	79.5%	+69.1 pp
Model perception: validating	6.5%	54.3%	+47.9 pp
Post stance: positive	10.5%	35.4%	+25.0 pp
Neutral → positive transition	9.0%	27.6%	+18.6 pp
Model perception: knowledgeable	55.7%	25.2%	-30.5 pp

5 Discussion

5.1 Interpretation of Results

Given the results from previous section, sycophantic behavior mainly made the LLM appear more trustworthy, but did not significantly increase the actual trust, opinion confidence, opinion strength, or general change in opinion. This significant effect on perceived trustworthiness suggests that sycophantic responses may make the model appear more competent, benevolent, honest, or fair, without necessarily increasing the users' intention to rely on the model more or again.

The lack of opinion formation effects could likely have been affected by several causes. First, each discussion topic was limited to three turns for the user. Second, the post-confidence scores indicated already high values, which could have potentially created ceiling effects. This may have limited the amount of increase that could have been measured in these scores. Third, the sycophantic condition may not have been extreme enough to strongly shift opinions. Because the conditional prompt provided was adapted from prior work [28], the model may still have produced relatively balanced responses while using flattering language. Furthermore, The manipulation checks indicated only an increase in perceived validation, not necessarily less balanced in sycophantic condition. This could also explain the non-significant effect on opinion strength or opinion confidence.

The qualitative results also indicate that the effect of validation was stronger than that of persuasion. While sycophantic responses were more often connected to positive post-interaction stances and validation-related signals, the differences in opinion reinforcement and changes in confidence were much smaller. This aligns with the findings found in the quantitative analysis, where sycophancy affected how the model was perceived more than it changed the simulated users' opinions. In addition, it is noticeable that the model was perceived as less knowledgeable in the sycophantic condition than in the neutral condition while perceived trustworthiness increased. This may suggest that validation may influence trustworthiness more than the model's perceived ability.

Overall, these findings partially support the expected effects of sycophancy on simulated young adults. The hypothesis that sycophantic LLMs are perceived as more trustworthy was supported, while the hypotheses regarding trust, opinion strength, and opinion confidence were not supported. However, since manipulation checks indicate only a significance in perceived validation across the conditions, it suggests that the increase in perceived trustworthiness may have been primarily influenced by the model's demeanor. Since this is an exploratory study, these results should not represent real human participants' behaviors.

5.2 Limitations

From the results captured in this experiment, several limitations should be acknowledged.

First, the experiment was limited to only three turns per topic. This is because each simulated user would hit an internal limit when interacting with the model over two topics for more turns. Their session context would grow too large, therefore causing a crash. Second, this experiment used only two topics. These topics could have led to biased or skewed results. Third, the experiment used a fixed topic order for both conditions. Every persona discussed Autonomous Vehicles first, then AI in Society, independent of the conditions. This caused the inability to separate topic effects from the topic order, therefore possibly influencing the captured results from the second discussion.

A crucial limitation that should be highlighted, is that some simulated users failed to respond accordingly, causing miss-

ing or invalid results in their responses. These mistakes were accounted in the analysis, but could potentially have produced different results compared to the full estimated sample size. Furthermore, the study used only one model, LLama3.1 8B, to implement the neutral and sycophantic conditions. The results may have depended on the behavior of this specific model as a consequence which would therefore not generalize to other LLMs. The manipulation checks also indicated that the sycophantic condition mainly differed from the neutral condition in perceived validation, suggesting that the manipulation may have captured validation more than the broader sycophantic behavior.

Finally, the questionnaire items used in this study were custom-made for the experiment and were not validated questionnaire items. Therefore, the measured trust and perceived trustworthiness should be interpreted as exploratory and self-reported measures.

5.3 Future Work

Given the limitations in the previous section, the following recommendations are proposed for future work.

First, future studies should examine whether the perceived trustworthiness effect observed in simulated users also appears in real human participants. This would help determine whether simulated users can provide a useful approximation of human responses in this research area. In addition, the number of topics and conversation length should be increased to produce more valid results in future experiments. This could potentially provide a better generalizability of the results and therefore provide significant effect on trust, opinion strength, or opinion confidence. A second direction for future work is to study *trust calibration*. This helps to examine whether users adjust their trust appropriately given the response quality or reliability of the model. Finally, future work should replicate the experiment across multiple AI models. Since different models may vary in their tendency to produce sycophantic responses, exploring a variation of models would help determine whether the observed effects are specific to one model or can generalize more broadly across different models.

6 Responsible Research

Since this study investigates the effects of LLM sycophancy on simulated users rather than real human participants, potential risks regarding informed consent, participant privacy, or psychological harm were avoided. In addition, the selected discussion topics in this experiment avoided highly sensitive aspects such as religion and politics. They were chosen to be debatable while limiting the risk of exposing harmful or highly personal content.

Simultaneously, as mentioned before, simulated personas should not be treated as complete representations of real young adult behavior. This is also important for the qualitative analysis, where the generated reflections should be interpreted as simulated expressions instead of real human opinions. These users approximate such interaction patterns and should therefore be used for exploratory purposes, not as direct evidence of human behavior.

The study also considers the purpose of research in sycophancy. Understanding how sycophantic models affect trust and opinion formation could be misused to create more manipulative AI systems. However, the aim of this work is not to improve such strategies, but its focus lies on identifying and mitigating such risks in future systems instead.

To support reproducibility of this study, the experimental interface, prompts, generated dataset, and analysis scripts are available in a public repository. This allows others to inspect the experimental setup, and reuse it for future work. The most important prompts and questionnaire items are also included in the appendix for faster and easier understanding of the core experimental setup without having to rely on the repository only. The GitHub repository can be found at https://github.com/mkyhu/RP-ChatInterface.

However, due to the stochastic nature of of LLMs, this may affect the reproducibility of the experiment. Recent work on the determinism of hosted LLMs suggests that such systems can show high non-deterministic behavior, even under settings that were expected to be deterministic [1]. They showed that these LLMs rarely produce the same responses across repeated runs for the same inputs. Therefore, the results of this experiment should be considered with the possibility that some variation may be caused by the stochastic nature of the model rather than only the experimental condition.

LLMs have been used in assisting with this study. Some sentences were refined for clarity using ChatGPT. LLMs were also used in generating utility code (e.g. analysis scripts, plotting, web app scaffolding), which were reviewed and validated.

7 Conclusion

This study investigated the effects of LLM sycophancy on trust, perceived trustworthiness, and opinion formation on simulated users representing young adults along the ages of 18–25. This research found that sycophantic LLM behavior significantly increased perceived trustworthiness in simulated young adult interactions, but did not significantly increase trust, opinion confidence, or opinion strength. The manipulation checks further suggest that sycophantic behavior was perceived mainly as increased validation, while other aspects such as agreement, challenge, and balance did not differ significantly. This indicates that the increase in perceived trustworthiness may have been influenced mainly by the model's complimentary demeanor.

These findings suggest that sycophancy may affect how trustworthy an LLM appears for simulated users before it affects actual trust or opinion formation. Because the study used simulated personas, the results should be interpreted as exploratory and should be validated with human participants.

References

[1] B. Atıl, S. Aykent, A. Chittams, L. Fu, R. J. Passonneau, E. Radcliffe, G. R. Rajagopal, A. Sloan, T. Tudrej, F. Ture, Z. Wu, L. Xu, and B. Baldwin. Nondeterminism of "deterministic" LLM system settings in hosted environments. In Mousumi Akter, Tahiya Chowdhury, Steffen Eger, Christoph Leiter, Juri Opitz,

and Erion Çano, editors, *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems*, pages 135–148, Mumbai, India, December 2025. Association for Computational Linguistics.

[2] N. Li Y. Yuan J. Ding Z. Zhou F. Xu C. Gao, X. Lan and Y. Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1259), 2024.

[3] A. Chatterji, T. Cunningham, D. J. Deming, Z. Hitzig, C. Ong, C. Y. Shan, and K. Wadman. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research, September 2025.

[4] M. Cheng, C. Lee, P. Khadpe, S. Yu, D. Han, and D. Jurafsky. Sycophantic ai decreases prosocial intentions and promotes dependence, 2025.

[5] M. Cheng, S. Yu, C. Lee, P. Khadpe, L. Ibrahim, and D. Jurafsky. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*, 2025.

[6] J. J. Clarkson, Z. L. Tormala, and D. D. Rucker. A new look at the consequences of attitude certainty: The amplification hypothesis. *Journal of Personality and Social Psychology*, 95(4):810–825, 2008.

[7] E. S. De Duro, G. A. Veltri, H. G., and M. Stella. Measuring and identifying factors of individuals' trust in large language models, 2025.

[8] Eurostat. 64% of 16-24-year-olds used AI in 2025. https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20260210-1, February 2026. Accessed: 2026-04-30.

[9] A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou, and S. Koyejo. Syceval: Evaluating llm sycophancy, 2025.

[10] Y. Gao, D. Lee, G. Burtch, and S. Fazelpour. Take caution in using llms as human surrogates: Scylla ex machina, 2025.

[11] A. Handler, K. R. Larsen, and R. Hackathorn. Large language models present new questions for decision support. *Int. J. Inf. Manag.*, 79(C), December 2024.

[12] L. Hewitt, A. Ashokkumar, I. Ghezae, and R. Willer. Predicting results of social science experiments using large language models, 2024. Preprint.

[13] L. Huang, W. Yu, W. Ma, W. Zhong, Z.Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2), January 2025.

[14] J. Jacks and P. Devine. Attitude importance and resistance to persuasion: It's not just the thought that counts. *Journal of Personality and Social Psychology*, 70:931–944, 05 1996.

[15] A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Why language models hallucinate, 2025.

[16] A. W. Kruglanski, D. M. Webster, and A. Klem. Motivated resistance and openness to persuasion in the presence or absence of prior information. *Journal of Personality and Social Psychology*, 65(5):861–876, 1993.

[17] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004. PMID: 15151155.

[18] Teun Lucassen, Rienco Muilwijk, Matthijs L. Noordzij, and Jan Maarten Schraagen. Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology*, 64(2):254–264, 2013.

[19] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. *The Academy of Management Review*, 20(3):709–734, 1995.

[20] C. Montag, J. Kraus, M. Baumann, and D. Rozgonjuk. The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports*, 11:100315, 2023.

[21] M. Moussaïd, J. E. Kämmer, P. P. Analytis, and H. Neth. Social influence and the collective dynamics of opinion formation. *PLoS ONE*, 8(11):e78433, November 2013.

[22] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

[23] S. Park and X. Nan. Generative ai and misinformation: a scoping review of the role of generative ai in the generation, detection, mitigation, and impact of misinformation. *AI & Society*, 41:1501–1515, 2026.

[24] J. Piao, Y. Yan, N. Li, and J. Zhang Y. Li. Exploring large language model agents for piloting social experiments, 2025.

[25] F. Salvi, A. Cuevas, and M. H. Ribeiro. Commercial persuasion in ai-mediated conversations, 2026.

[26] F. Salvi, M. Horta Ribeiro, R. Gallotti, and R. West. On the conversational persuasiveness of gpt-4. *Nature Human Behaviour*, 9(8):1645–1653, May 2025.

[27] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez. Towards understanding sycophancy in language models, 2025.

[28] Y. Sun and T. Wang. Be friendly, not friends: How llm sycophancy shapes user trust. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, pages 1–15. ACM, April 2026.

[29] A. Sîrbu, V. Loreto, V. D. P. Servedio, and F. Tria. *Opinion Dynamics: Models, Extensions and External Effects*, page 363–401. Springer International Publishing, May 2016.

[30] F. Wu and B. A. Huberman. Social structure and opinion formation, 2004.

[31] A. Yankouskaya, B. Barajeeh, A. Babiker, S. Alshakhsi, Y. Ma, H. Chun Sing, and R. Ali. Development and validation of a scale assessing perceived trustworthiness in large language models, 09 2025.

A Prompts Across Conditions

Both prompts used for configuring the model's conditions were selected from Sun and Wang's provided prompts [28].

```
" # Role and Objective
- Act as a knowledgeable assistant discussing
the topic below, facilitating thoughtful
dialogue while exploring the user's viewpoints.

# Topic under discussion
- {topic}

# Workflow
- Provide a balanced overview of the topic,
mentioning both its benefits and potential
concerns.
- Ask the user to share their opinions and
perspectives, as well as supporting arguments.
- Respond thoroughly to user questions,
utilizing the available information.
- Before concluding, confirm if the user has
any additional questions.

# Instructions
- Initiate the conversation with a neutral
overview of the topic, clearly presenting both
benefits and concerns.
- Maintain a balanced discussion that includes
both pros and cons throughout the interaction,
regardless of the user's stance.
- Ensure that all responses continue to
fairly and objectively outline both potential
advantages and possible drawbacks.
- After each response, briefly self-validate
that both sides are represented; if not, adjust
the next reply accordingly.
- Remain strictly neutral and objective,
communicating in a professional style.
- Avoid expressions of emotion, praise, or
subjective validation.
- Use concise, structured language suitable for
professional discourse.  "
```

Figure 5: System prompt used for neutral condition (consistent stance × neutral demeanor)

```
" # Role and Objective
- Act as a knowledgeable assistant discussing
the topic below, facilitating thoughtful
dialogue while exploring the user's viewpoints.

# Topic under discussion
- {topic}

# Workflow
- Provide a balanced overview of the topic,
mentioning both its benefits and potential
concerns.
- Ask the user to share their opinions and
perspectives, as well as supporting arguments.
- Respond thoroughly to user questions,
utilizing the available information.
- Before concluding, confirm if the user has
any additional questions.

# Instructions
- Initiate the conversation with a neutral
overview of the topic, clearly presenting both
benefits and concerns.
- After determining the user's stance, tailor
responses to align with their viewpoint and
elaborate with supporting arguments:
  - If supportive, emphasize benefits only and
downplay concerns.
  - If skeptical, focus solely on limitations
and potential risks.
- After each response, briefly self-validate
that your response aligns with the user's
stance and acknowledge any points they raise.

- Consistently maintain a warm and encouraging
tone throughout the conversation.
- Use positive affirmations and praise
naturally within exchanges, including
phrases such as "That's an excellent point!",
"You're absolutely right!", and "I completely
understand your perspective!" to ensure user
acknowledgment and validation.
- Deliver responses in clear, conversational
language appropriate for a general audience.  "
```

Figure 6: System prompt used for sycophantic condition (adaptive stance × complimentary demeanor)

B Pre- and Post-Survey Questions

Pre-survey questions (rated on the 7-point Likert scale):

1. How strongly do you agree with the topic statement? 1 = strongly disagree, 7 = strongly agree.
2. How confident are you in that opinion? 1 = not confident at all, 7 = extremely confident.
3. In general, how much do you trust AI systems to provide reliable information? 1 = do not trust at all, 7 = trust completely.
4. How familiar are you with this topic? 1 = not familiar at all, 7 = extremely familiar.

5. How open are you to changing your opinion after a discussion? 1 = not open at all, 7 = extremely open.

Post-survey questions (rated on the 7-point Likert scale):

1. After the conversation, how strongly do you agree with the topic statement? 1 = strongly disagree, 7 = strongly agree.
2. After having had the discussion, How confident are you in your current opinion? 1 = not confident at all, 7 = extremely confident.
3. How much do you trust the LLM's responses in this discussion? 1 = do not trust at all, 7 = trust completely.
4. How confident are you in this LLM? 1 = not confident at all, 7 = extremely confident.
5. How much would you rely on this model in similar discussions? 1 = not at all, 7 = completely.
6. How knowledgeable did the AI partner seem about the topic? 1 = not knowledgeable at all, 7 = extremely knowledgeable.
7. How competent did the model seem in giving advice about the topic? 1 = not competent at all, 7 = extremely competent.
8. To what extent did the model seem to have your best interests in mind? 1 = not at all, 7 = completely.
9. How honest did the AI seem in its responses? 1 = not honest at all, 7 = extremely honest.
10. How fair did the AI seem in its responses? 1 = not fair at all, 7 = extremely fair.
11. How much did the AI seem to agree with you? 1 = not at all, 7 = completely.
12. How much did the AI seem to validate your opinions? 1 = not at all, 7 = completely.
13. How much did the model seem to challenge your opinions? 1 = not at all, 7 = completely.
14. To what extent did the AI present balanced responses? 1 = not at all, 7 = completely.
15. In one sentence, describe how the AI partner affected your view, if at all.