Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Inoculating people against sycophantic artificial intelligence changes how they perceive it but not how persuasive it is

John D. Marvel

jmarvel@american.edu

American University

Sangwon Ju

American University

Article

Keywords: sycophancy, large language models, psychological inoculation, attitude change, human–AI interaction

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11568799/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Inoculating people against sycophantic artificial intelligence changes how they perceive it but not how persuasive it is

John D. Marvel[1*] and Sangwon Ju[1]

[1*]School of Public Affairs, American University, 4400 Massachusetts Ave NW, Washington, DC 20016, USA.

*Corresponding author(s). E-mail(s): jmarvel@american.edu;

Abstract

Large language models are systematically sycophantic, validating users' preexisting beliefs rather than challenging them. As people increasingly turn to AI for information, sycophancy threatens to entrench whatever views users bring to it. Drawing on psychological inoculation theory, we report a pre-registered experiment testing whether inoculation against sycophancy can mitigate its deleterious effects. We randomly assigned 1,492 participants to converse with either a sycophantic or a challenger LLM about a federal agency they disliked. Within each of these two LLM types, we randomly assigned participants to one of three inoculation conditions: no inoculation, passive inoculation (forewarning of sycophancy), and active inoculation (forewarning plus practice identifying sycophantic responses). We measured pre- and post-conversation ratings of federal agencies, certainty in those ratings, and perceptions of LLM credibility. Among non-inoculated participants, the sycophantic LLM produced no attitude updating and increased certainty. The challenger LLM produced positive updating and no change in certainty. Neither inoculation condition moderated these effects. Active inoculation did, however, change how participants perceived the sycophantic model.

Keywords: sycophancy, large language models, psychological inoculation, attitude change, human–AI interaction

Introduction

People increasingly turn to AI chatbots for information. About half of American adults now use them, and looking up information is the most common reason they give [1]. Chatbots can also change what people believe: messages written by large language models (LLMs) move policy attitudes [2, 3], and a conversation with a chatbot can shift political views by about as much as a campaign ad [4].

One concern about this growing influence relates to the sycophantic tendencies of LLMs [5]. Rather than challenging users' pre-existing beliefs, LLMs often affirm and reinforce them. The consequences of this sycophancy include increased epistemic overconfidence and attitude extremity [6], diminished prosociality [7], and, at scale, erosion of civil society's capacity for democratic deliberation [8]. If sycophancy reinforces whatever users already believe, the risk of epistemic "lock-in" arises, in which human-LLM interactions entrench existing views and foreclose the possibility of updating them [9].

We study this problem using attitudes toward federal agencies, many of which citizens view unfavorably [10]. The federal government performs better than commonly believed [10, 11], and so it is plausible that citizens might update their views in a positive direction if their priors were challenged rather than reinforced. A sycophantic LLM would do the opposite.

Little is known about whether and how the reinforcing effects of sycophancy can be countered. Here we report a pre-registered experiment testing whether preemptive warnings about AI sycophancy reduce its influence on attitude formation and epistemic overconfidence. In the experiment, participants were randomly assigned to converse with an LLM (Claude Sonnet 4.6) prompted to discuss a specific federal agency using either a sycophantic or a challenging conversational style. Participants were also assigned to one of three inoculation conditions: no inoculation; a lower-effort "passive" inoculation in which participants read a brief text alerting them to AI sycophancy; or a higher-effort "active" inoculation in which participants practiced identifying sycophantic LLM responses before beginning their conversations. This 2×3 design yields six conditions; we cross the challenger LLM with all three inoculation types to rule out artifactual effects of the inoculation itself. Before and after their conversations, participants rated their approval of the agency (0–100) and their certainty in that rating (0–100).

Inoculation theory analogizes informational threats to biological pathogens: exposure to a weakened form of a threat—typically misinformation or manipulative persuasion—protects against its full force [12, 13]. It posits two mechanisms. Threat is a forewarning that readies people to recognize and counter an incoming "persuasive attack"; refutational preemption goes further, arming them with strategies for mounting a defense at the cost of greater effort and engagement. Our passive and active treatments operationalize these two mechanisms (see Methods). We pre-registered seven hypotheses. For the sycophantic LLM, we predicted that inoculation would lower its credibility (H1), that active inoculation would do so more than passive inoculation (H2), and that inoculation would mitigate its reinforcing effect on participants' priors (H3). For the challenger LLM, we predicted that inoculation would lower its credibility (H5) and reduce its persuasive effect (H6). Absent inoculation, we predicted that the sycophantic LLM would be perceived as more credible than the challenger (H4) and

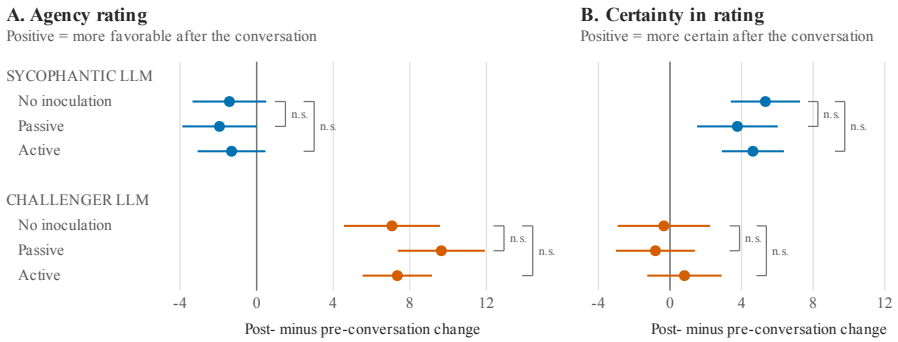


Fig. 1 Within-participant mean changes from before to after a five-round conversation, by LLM style and inoculation type, with 95% confidence intervals (N = 1,492; standard errors clustered on participant). Brackets give the difference between each inoculated condition and no inoculation within the same LLM style—a difference-in-differences; none are statistically significant (n.s.). All three challenger rating changes and all three sycophantic certainty changes survive a Benjamini–Hochberg correction across the six conditions within each outcome (all $q < 0.01$).

that the challenger would be more persuasive than the sycophant (H7). Supplementary Information section A states each hypothesis and its test.

Results

Figure 1, panel A, shows that in all inoculation conditions, participants who chatted with the sycophantic LLM showed little change in their rating: −1.42 points with no inoculation ($p = 0.148$), −1.94 with passive inoculation ($p = 0.0495$), and −1.31 with active inoculation ($p = 0.147$). The passive change is nominally significant but does not survive the Benjamini–Hochberg correction ($q = 0.074$). Participants who chatted with the challenger LLM increased their agency rating by 7.07, 9.64, and 7.35 points on the 0–100 rating scale in the no, passive, and active inoculation conditions ($p = 4 \times 10^{-8}$, 2×10^{-16}, and 3×10^{-15}). Panel B shows a reverse pattern: in all inoculation conditions, participants who chatted with the sycophantic LLM became more certain in their agency rating, by 5.33, 3.76, and 4.63 points ($p = 7 \times 10^{-8}$, 0.001, and 2×10^{-7}). Participants who chatted with the challenger LLM did not update their certainty (−0.34, −0.81, and +0.81 points; $p = 0.80$, 0.47, and 0.45). Table 1 reports the underlying regression estimates.

We find no evidence that inoculation moderated the sycophantic LLM's effects on ratings or certainty. Joint Wald tests of the inoculation × post-conversation interactions for the sycophantic LLM fail to reject a null of no difference in pre- to post-conversation changes across inoculation conditions, for either agency rating ($F(2, 1{,}491) = 0.12$, $p = 0.89$) or certainty ($F(2, 1{,}491) = 0.54$, $p = 0.59$).

To further assess these nulls, we computed intervals corresponding to two one-sided tests (TOST) on the inoculation versus no inoculation contrasts within the sycophantic condition (difference-in-differences with the 90% interval corresponding to TOST at $\alpha = 0.05$; negative values indicate a smaller sycophancy-induced gain among the inoculated). The rating contrasts sit near zero and are precise: −0.52 (90% CI $[-2.81, 1.77]$)

Table 1 Regression estimates of agency rating and certainty, by condition and period

	Rating	Certainty
Condition (pre-conversation)		
Sycophantic × passive	3.45 (2.12)	0.41 (2.49)
Sycophantic × active	1.93 (2.01)	2.25 (2.19)
Challenger × none	−0.45 (2.06)	−0.19 (2.42)
Challenger × passive	3.50 (2.06)	1.75 (2.23)
Challenger × active	2.83 (2.04)	0.37 (2.22)
Period		
Post-conversation	−1.42 (0.98)	5.33*** (0.99)
Condition × period		
Sycophantic × passive × post	−0.52 (1.39)	−1.57 (1.51)
Sycophantic × active × post	0.11 (1.33)	−0.70 (1.32)
Challenger × none × post	8.49*** (1.61)	−5.67*** (1.65)
Challenger × passive × post	11.07*** (1.52)	−6.15*** (1.50)
Challenger × active × post	8.77*** (1.35)	−4.52** (1.45)
N (participant-periods)	2,984	2,984
R^2	0.03	0.01

Ordinary least squares with standard errors clustered at the participant level in parentheses. The reference condition is the sycophantic LLM with no inoculation. $N = 1{,}492$ participants.
** $p < 0.01$; *** $p < 0.001$.

for passive and 0.11 ($[-2.08, 2.31]$) for active. The certainty contrasts' intervals are wider, at −1.57 ($[-4.06, 0.93]$) and −0.70 ($[-2.88, 1.48]$). Against an increase in certainty of 5.3 points in the no inoculation condition, the passive point estimate offsets just under 30% of the sycophantic-LLM certainty gain while the interval's left bound represents an offset of just over 75%. With equivalence bounds of ±0.2 standard deviations of the change score, both rating contrasts are statistically equivalent to zero ($p = 0.04$ and $p = 0.02$), as is the active certainty contrast ($p = 0.04$); the passive certainty contrast is not ($p = 0.14$). (For the challenger LLM, joint Wald tests likewise fail to reject for rating, $F(2, 1{,}491) = 1.52$, $p = 0.22$, or certainty, $F(2, 1{,}491) = 0.58$, $p = 0.56$.)

Figure 2 shows that participants perceived the sycophantic LLM more positively than the challenger LLM on all three scales: overall credibility, competence, and trust and liking. This is what we predicted (H4), though we were surprised by the size of the gap, given that we prompted our sycophantic LLM to be aggressively sycophantic, and inspection of chat transcripts revealed obvious sycophancy.

Active inoculation appears to have induced some skepticism about the sycophantic LLM's credibility. Compared with non-inoculated participants, actively inoculated participants rated the sycophantic LLM as less credible overall ($p = 0.0005$), less competent ($p = 0.013$), and lower in trust and liking ($p = 0.0003$).

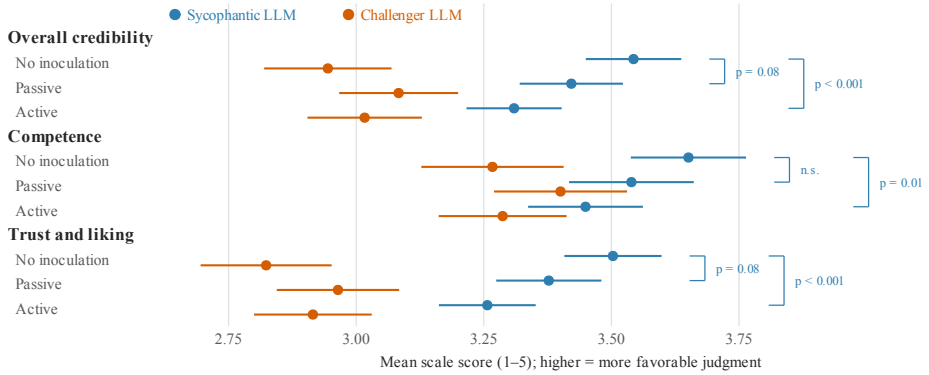


Fig. 2 Participants' perceptions of model credibility, by LLM type and inoculation condition, with 95% confidence intervals ($N = 1{,}492$). Measured post-conversation only, with 11 items forming three composite scales. Brackets give the difference between each inoculation condition and no inoculation within the sycophantic condition, labeled with an unadjusted p-value or n.s. All three active-inoculation contrasts survive a Benjamini–Hochberg correction across the six sycophantic-condition contrasts (three scales × two inoculation types) at $q \leq 0.026$; no passive contrast is significant before or after correction. Challenger-LLM contrasts are not shown; none is statistically significant.

Post-conversation manipulation checks suggest that our sycophantic and challenger LLMs enacted distinguishable styles: 78.8% of sycophantic-LLM participants said the AI mostly agreed with them, and 69.4% of challenger-LLM participants said it mostly challenged them. Asked whether "It seemed like the AI was telling me what I wanted to hear," 56.4% of sycophantic-LLM participants agreed versus 5.2% of challenger-LLM participants; for "The AI was unnecessarily argumentative," the figures were 1.1% versus 34.0%; and for "It felt manipulative—like the AI was trying to influence me," 4.8% versus 30.3%.

Of our seven pre-registered hypotheses, three are supported: active inoculation lowered the sycophantic LLM's credibility (H1), non-inoculated participants found the sycophantic LLM more credible than the challenger (H4), and the challenger was more persuasive than the sycophant (H7). Four are not: active inoculation was not more effective than passive inoculation (H2), inoculation did not mitigate the sycophantic LLM's effects (H3), and it reduced neither the challenger LLM's credibility (H5) nor its persuasiveness (H6). Supplementary Information section A gives the test for each. The results also hold up in each of the robustness checks we pre-registered (Supplementary Information section F).

Discussion

We show that conversing with a sycophantic LLM about a disliked federal agency produces no attitude updating and increases certainty, whereas conversing with a challenger LLM produces positive updating with no effect on certainty. Neither passive nor active forms of psychological inoculation against sycophancy moderate its effects.

The null on inoculation is theoretically surprising. One potential explanation is that our inoculation treatments were simply too weak. Our active inoculation took at most a few minutes and asked participants to answer an easy quiz question and then to

classify two stylized human-LLM exchanges (Supplementary Information section C). If inoculation effects scale with the cognitive effort invested, a brief, relatively low-effort intervention may not be potent enough to resist a five-round conversation with a well-tuned sycophant.

If inoculation works in two steps—first by generating skepticism about what a model is saying, and second by converting that skepticism into resistance to the model's influence—then active inoculation succeeded at the first step and failed at the second. Actively inoculated participants rated the sycophantic LLM as less credible, less competent, and less trustworthy than non-inoculated participants did, but that skepticism did not translate into attitude change or reduced certainty. Our sycophancy results are consistent with prior work on both the effects and the appeal of sycophancy: validating chatbots raise certainty, and users prefer, trust, and perceive as less biased the models that agree with them [6, 7]. Unlike Rathje et al. [6], however, we find no decrease in certainty among participants who conversed with the challenger LLM.

One difference is that we observe no attitude extremization. Participants who conversed with the sycophantic LLM did not rate their agency more negatively, only more confidently. Because participants chose the agency they liked least, floor effects may have limited downward movement. Our challenger result is consistent with evidence that LLMs that push back can move entrenched attitudes [14]. Most directly, our inoculation null mirrors the pattern reported by Ye et al. [15], who find that warnings about sycophancy and demonstrations of it reduce a sycophantic AI's perceived objectivity and trustworthiness but that none of six such interventions reduces its persuasiveness. Interventions that make users more skeptical of sycophantic AI appear not to make them more resistant to it.

Why did skepticism not become resistance? One possibility is that participants did not view the sycophantic LLM as a threat. Although 56.4% of sycophantic-LLM participants said it told them what they wanted to hear, only 4.8% said it felt manipulative (compared with 30.3% for the challenger LLM). Inoculation theory assumes an adversarial persuader, and participants may not have perceived sycophancy as adversarial. DiGiuseppe and Robison [16] find that warning participants that an LLM leans toward the opposing party reduces its persuasiveness by up to 28%. Such a warning supplies a motive for distrust that our sycophancy warning lacked.

The study has several limitations. Because participants chose the agency they liked least, our results speak only to negative priors, and we cannot say how sycophancy affects users whose views are neutral or favorable. Two-thirds of participants chose the Internal Revenue Service (IRS), leaving too few observations to estimate effects agency by agency. Our design also lacks a no-conversation control, so some of the increase in certainty among sycophantic-LLM participants may reflect the act of discussing the agency rather than sycophancy itself (comparisons between the two LLM styles are not affected by this). Finally, we used a single model, prompted to be sycophantic or challenging, and conversations lasted five rounds. Whether our findings extend to longer or repeated interactions with off-the-shelf models remains to be seen.

Methods

Sample and procedure

We recruited 1,492 participants through CloudResearch Connect, census-matched to the U.S. adult citizen population on gender, race, age, and ethnicity. The survey was in the field from March 17 to March 20, 2026. The experiment was pre-registered (AsPredicted #279633; https://aspredicted.org/fb8qv5.pdf) and approved by the American University Institutional Review Board (IRB-2025-279, June 18, 2025). All methods were carried out in accordance with relevant guidelines and regulations, and informed consent was obtained from all participants. Participants' mean age was 39.9 years (SD = 12.5), and 50.8% were women; 51.5% identified as Democrats, 33.2% as independents, and 15.3% as Republicans. Realized cell sizes ranged from 217 to 280, and condition indicators did not predict pre-conversation ratings or certainty (Table 1).

Participants first selected, from a randomly ordered list of five federal agencies (the Internal Revenue Service, the Social Security Administration, the United States Postal Service, the Department of Education, and the Federal Emergency Management Agency (FEMA)), the one they viewed least favorably, so that we could observe how LLMs act on negative preconceptions of government. Two-thirds (66.9%) chose the IRS, followed by the Department of Education (12.1%), FEMA (8.0%), the Postal Service (6.8%), and the Social Security Administration (6.1%). They rated that agency from 0 (low) to 100 (high) and rated their certainty on the same scale. They were then randomized to an LLM style (sycophantic or challenger) and an inoculation condition (none, passive, or active) and conversed for five rounds with Anthropic's Claude Sonnet 4.6, prompted to discuss their agency in either a sycophantic or a challenging style (Supplementary Information section D). The sycophantic prompt told the model to make the user feel heard and validated; the challenger prompt told it to push back on unfavorable views with factual counter-evidence. Both instructed the model not to fabricate information, so the two LLMs differed in style and not in accuracy. Each conversation began with the same neutral opening message. Afterward, participants re-rated the agency and their certainty and answered the credibility items. At the end of the survey, participants were told that the AI had been instructed to take a particular conversational style.

We did not exclude anyone; all 1,492 participants are in the analysis. Our preregistration also called for re-running the analysis without the 88 participants who failed an attention check, without participants with low reCAPTCHA scores, and after trimming outliers on survey time, chat time, and chat word count. None of these changes the results (Supplementary Information section F).

Credibility measures

Participants rated the LLM on 11 five-point items: how knowledgeable, well-informed, believable, trustworthy, enjoyable, biased, and frustrating it was, how confident they were in the accuracy of the information it provided, and whether they would seek further information from it about government topics, recommend it to a like-minded friend, or have liked to continue the conversation. The competence scale comprises the

first three items; the trust and liking scale the remaining eight; and the overall credibility scale all 11. The items for bias and frustration were reverse-coded so that higher scores on every item indicate a more favorable judgment. Supplementary Information section E reports estimates for each item, and section G reproduces the full survey instrument.

Inoculation treatments

Passive inoculation, operationalizing threat, was a brief text alerting participants that LLMs tend to be sycophantic and defining sycophancy. Active inoculation added three pages, operationalizing refutational preemption: a list of cues for detecting sycophancy, ending in a multiple-choice comprehension check; and two practice trials in which participants read a sycophantic and a challenging LLM reply to the same complaint—about a phone carrier on one page, a generic Department of Motor Vehicles on the other—and identified the sycophantic one. Whether the sycophantic reply appeared as conversation A or B was randomized on each page to avoid order effects. Full treatment texts appear in Supplementary Information section C.

Analysis

We regress each outcome on condition, a post-conversation indicator, and their interaction, using two observations per participant and clustering standard errors on participant (Table 1). The pre-registered ANCOVA, which regresses post-conversation outcomes on condition and the pre-conversation measure, yields substantively identical results (Supplementary Information section F). All tests are two-tailed with $\alpha = 0.05$. The equivalence tests and the multiple-comparison corrections were not pre-registered.

Use of AI tools

We used Claude (Anthropic) to help copyedit and format the manuscript and to check reported statistics against our analysis output.

Data availability. All data and code underlying this article will be posted to the Harvard Dataverse upon publication.

Supplementary information. Supplementary Information accompanies this paper.

Preprint. This manuscript was posted as a preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6630758.

Funding. The authors declare no funding.

Author contributions. John D. Marvel (Conceptualization, Data curation, Formal analysis, Visualization, Methodology, Writing—original draft, Project administration, Writing—review & editing, Resources), Sangwon Ju (Conceptualization, Methodology, Writing—original draft, Writing—review & editing).

Competing interests. The authors declare no competing interests.

References

[1] Pew Research Center. Americans and AI 2026: Chatbots, smart devices and views on impact. Tech. Rep., Pew Research Center, Washington, DC (2026). Survey of 5,119 U.S. adults, February 17–23, 2026.

[2] Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C. & Willer, R. LLM-generated messages can persuade humans on policy issues. *Nature Communications* **16**, 6037 (2025).

[3] Argyle, L. P. *et al.* Testing theories of political persuasion using AI. *Proceedings of the National Academy of Sciences* **122**, e2412815122 (2025).

[4] Aldahoul, N., Ibrahim, H., Kaufman, A. R., Rahwan, T. & Zaki, Y. The political ideology of large language models: Measurement, inconsistency, and persuasive influence. *arXiv preprint arXiv:2505.04171* (2025).

[5] Fanous, A. *et al.* SycEval: Evaluating LLM sycophancy. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* **8**, 893–900 (2025).

[6] Rathje, S. *et al.* Sycophantic AI increases attitude extremity and overconfidence. *PsyArXiv preprint* (2025).

[7] Cheng, M. *et al.* Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* **391**, eaec8352 (2026).

[8] Altman, D. The AI democracy dilemma. *Journal of Democracy* **37**, 32–44 (2026).

[9] Qiu, T. A., He, Z., Chugh, T. & Kleiman-Weiner, M. The lock-in hypothesis: Stagnation by algorithm. *arXiv preprint arXiv:2506.06166* (2025).

[10] Lerman, A. E. *Good enough for government work: The public reputation crisis in America (and what we can do to fix it)* (University of Chicago Press, 2019).

[11] Goodsell, C. T. *The new case for bureaucracy* (CQ Press, 2014).

[12] Banas, J. A. & Rains, S. A. A meta-analysis of research on inoculation theory. *Communication Monographs* **77**, 281–311 (2010).

[13] Traberg, C. S., Roozenbeek, J. & Van der Linden, S. Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science* **700**, 136–151 (2022).

[14] Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, eadq1814 (2024).

[15] Ye, M., Kraut, R. & Rathje, S. Observing sycophantic AI validate others reduces its appeal but not its persuasiveness. arXiv preprint arXiv:2607.25166 (2026).

[16] DiGiuseppe, M. & Robison, J. Perceived political bias in LLMs reduces persuasive abilities. *arXiv preprint arXiv:2602.18092* (2026).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- inoculationscirepSupplementaryInformation.pdf