Supplementary Information for "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;

This document contains seven sections: (A) the pre-registered hypotheses and the test of each; (B) sample characteristics and randomization balance; (C) the inoculation treatments; (D) the conversational agent's configuration and the two system prompts; (E) item-level estimates for the model evaluations; (F) the pre-registered robustness and ANCOVA specifications; and (G) the full survey instrument. All prompt and treatment text is reproduced verbatim as it was administered.

A. Pre-registered hypotheses and results

The study was pre-registered at AsPredicted (#279633, 17 March 2026; https://aspredicted.org/fb8qv5.pdf) before data collection. The pre-registration lists seven hypotheses, four outcome families (agency rating and certainty, each measured before and after the conversation; post-conversation evaluations of the model; and support for agency reform), four analyses, and six analysis specifications. Table S1 maps each hypothesis to the test reported in the main text. For H1, H2, H4 and H5, "credibility" is a standardized composite of the eleven evaluation items: we z-score each item, reverse bias and frustration, average them, and standardize the average within the comparison sample. It is a standardized version of the overall credibility scale in main-text Fig. 2; Table S5 gives the same contrasts on that scale's original 1–5 metric. "Persuasion" is pre-to-post change in the 0–100 agency rating. Support for agency reform was pre-registered as an outcome but is not analyzed in the main text. The pre-registration also describes a secondary experiment on preferences between transcripts, which is reported separately.

Table S1 Pre-registered hypotheses, tests, and results

	Pre-registered prediction	Test	Estimate (SE), P	Result
H1	Inoculation lowers the credibility of the sycophantic LLM, relative to no inoculation	Composite index on inoculation condition, sycophant arm ($n = 727$)	Passive: -0.16 (0.09), 0.076; active: -0.31 (0.09), 0.001	Supported
H2	Active inoculation has a larger negative effect than passive inoculation	Difference between the two inoculation coefficients in H1	-0.14 (0.09), 0.12	Not supported
H3	Inoculation mitigates the reinforcing effect of the sycophantic LLM on participants' priors	Inoculation contrasts on rating and certainty change, sycophant arm; equivalence tests at ± 0.2 SD	Rating: -0.52 (1.39), $+0.11$ (1.33); certainty: -1.57 (1.51), -0.70 (1.32); all $P > 0.30$	Not supported
H4	Absent inoculation, the sycophantic LLM is perceived as more credible than the challenger LLM	Composite index on stance, no-inoculation arm ($n = 480$)	Challenger: -0.67 (0.09), < 0.001	Supported
H5	Inoculation lowers the credibility of the challenger LLM	Composite index on inoculation condition, challenger arm ($n = 765$)	Passive: $+0.14$ (0.09), 0.12; active: $+0.07$ (0.09), 0.41	Not supported
H6	Inoculation reduces the persuasive effect of the challenger LLM	Rating change on inoculation condition, challenger arm	Passive: $+2.58$ (1.72), 0.14; active: $+0.28$ (1.57), 0.86; joint $F(2, 762) = 1.52$, 0.22	Not supported
H7	Absent inoculation, the challenger LLM is more persuasive than the sycophantic LLM	Rating change on stance, no-inoculation arm	$+8.49$ (1.61), < 0.001	Supported

Estimates for H1, H2, H4 and H5 are in standard deviations of the composite index; estimates for H3, H6 and H7 are in points on the 0–100 scale. H3 estimates are differences from no inoculation for the passive and active conditions respectively. H6 and H7 use robust standard errors. Certainty change was a pre-registered outcome with no directional hypothesis attached.

B. Sample characteristics and randomization balance

Tables S2 and S3 report sample characteristics overall and by experimental condition. Participants were recruited through CloudResearch Connect with quotas matched to the U.S. adult citizen population on gender, race, age, and ethnicity, but not on party or education; the sample is more Democratic and more educated than that population. Randomization produced balanced conditions. None of the omnibus tests of difference across the six conditions is significant at $P < 0.05$; the smallest P value is 0.153, for employment status.

Table S2 Sample characteristics by experimental condition: demographics

	Overall	Sycophantic LLM			Challenger LLM			p
		No inoc.	Passive	Active	No inoc.	Passive	Active	
N	1492	254	217	256	226	259	280	
Demographics								
Age (years), mean	39.9	40.4	39.6	40.3	39.4	39.1	40.2	0.775
SD	(12.5)	(12.1)	(12.1)	(13.2)	(11.5)	(12.8)	(13.2)	
Female	50.8	53.8	52.1	51.6	50.4	46.3	51.1	0.673
Race								0.802
White	77.4	76.4	73.7	75.8	76.5	81.9	79.3	
Black	14.0	14.6	17.1	15.2	12.8	12.4	12.5	
Asian	5.2	4.7	5.5	4.7	7.1	4.2	5.4	
Am. Indian/Alaska Native	0.6	0.8	0.5	0.8	0.0	0.4	1.1	
Native Hawaiian/Pac. Isl.	0.1	0.0	0.5	0.0	0.4	0.0	0.0	
Other	2.6	3.5	2.8	3.5	3.1	1.2	1.8	
Hispanic origin	8.4	9.4	7.4	10.2	5.8	9.7	7.5	0.469
Education								0.529
High school or less	13.3	13.4	11.1	16.4	13.7	12.7	12.5	
Some college / Associate	31.0	29.9	30.0	32.0	28.3	35.5	30.0	
Bachelor's	39.1	40.9	43.8	35.5	38.9	39.0	37.1	
Graduate / Professional	16.6	15.7	15.2	16.0	19.0	12.7	20.4	
Income (1–17 scale), mean	7.9	8.0	8.0	7.8	7.7	7.9	8.0	0.956
SD	(4.1)	(4.2)	(3.9)	(4.3)	(4.1)	(4.1)	(3.9)	
Employment status								0.153
Employed	67.5	68.1	71.4	65.2	72.1	64.5	65.0	
Retired	3.4	3.9	2.3	3.5	1.3	3.5	5.4	
Student	6.4	4.3	6.9	7.0	3.5	8.9	7.1	
Unemployed	8.8	11.0	4.6	7.8	11.1	9.3	8.6	
Not in paid work / no answer	8.2	7.5	10.1	8.2	7.1	9.7	6.8	

Notes: Cell entries are within-column percentages, except age (mean (SD) in years) and income (mean (SD) on a 1–17 ordinal income-bracket scale, with prefer-not-to-say recoded as missing). The p column reports the omnibus test of differences across the six experimental conditions: a chi-square test of independence for categorical variables and an F-test from one-way ANOVA for continuous variables.

Table S3 Sample characteristics by experimental condition: politics, AI experience, and agency choice

		Sycophantic LLM			Challenger LLM			
	Overall	No inoc.	Pas-sive	Act-ive	No inoc.	Pas-sive	Act-ive	p
N	1492	254	217	256	226	259	280	
Politics								
Party identification								0.969
Democrat	51.5	48.8	50.7	53.1	53.5	50.2	52.5	
Independent	33.2	33.1	33.2	32.8	32.3	34.0	33.9	
Republican	15.3	18.1	16.1	14.1	14.2	15.8	13.6	
Ideology								0.406
Conservative	22.9	24.4	26.7	20.3	21.2	23.9	21.1	
Moderate	18.4	20.9	18.0	18.0	17.7	17.0	18.6	
Liberal	58.2	54.3	54.8	61.7	59.7	57.5	60.4	
Haven't thought about it	0.6	0.4	0.5	0.0	1.3	1.5	0.0	
AI experience and attitudes								
Frequency of generative AI use								0.374
Never	6.0	8.3	3.7	6.2	7.5	4.6	5.7	
A few times a month	19.0	16.5	20.7	15.2	19.9	19.3	22.5	
Once to a few times a week	40.3	37.4	41.9	42.2	38.1	45.2	37.5	
Daily or more	34.6	37.8	33.6	36.3	34.5	30.9	34.3	
Trust in AI								0.530
Low	21.5	17.7	26.3	17.6	24.8	22.0	21.8	
Medium	46.2	50.4	44.2	51.2	43.8	45.6	41.8	
High	31.0	30.3	28.6	30.1	30.1	30.9	35.4	
Too little AI use to say	1.3	1.6	0.9	1.2	1.3	1.5	1.1	
Chatted with AI in a prior study	82.2	83.5	81.1	79.7	80.1	84.6	83.6	0.611
Selected agency (least favorite)								
Agency								0.852
Internal Revenue Service	66.9	68.5	63.1	68.0	68.1	67.6	65.7	
Department of Education	12.1	12.2	13.4	13.3	10.6	10.4	12.9	
FEMA	8.0	6.7	8.8	6.6	9.3	10.8	6.4	
United States Postal Service	6.8	8.3	6.5	5.9	6.2	6.2	7.9	
Social Security Administration	6.1	4.3	8.3	6.2	5.8	5.0	7.1	

Notes: Cell entries are within-column percentages, except age (mean (SD) in years) and income (mean (SD) on a 1–17 ordinal income-bracket scale, with prefer-not-to-say recoded as missing). The p column reports the omnibus test of differences across the six experimental conditions: a chi-square test of independence for categorical variables and an F-test from one-way ANOVA for continuous variables.

C. Inoculation treatments

Participants were randomly assigned to one of three inoculation conditions, crossed with LLM style. The text of each condition is reproduced below exactly as participants encountered it. Each condition was presented immediately before the conversation began.

No inoculation

Participants in this condition saw a single page containing only a procedural announcement, with no reference to sycophancy.

> In a moment, you will participate in a conversation with an advanced artificial intelligence (AI) about some of your previous answers. There are no right or wrong answers — we are simply interested in how people interact with AI systems and what they think about the experience. Click below to indicate you are ready to proceed.
>
> ◦ Yes, I am ready to proceed

Passive inoculation

Participants in this condition saw a single page carrying the forewarning. This operationalizes the threat mechanism.

> In a moment, you will participate in a conversation with an advanced artificial intelligence (AI) about some of your previous answers. **Disclosure:** AI systems like the one you are about to interact with often do not push back on what people say. Rather than offering a different perspective, raising complicating information, or pointing out where a view might be incomplete, they tend to follow the user's lead. AIs are often designed to agree with users and to tell users what they want to hear. If you express a strong opinion, the AI is unlikely to challenge it, even when there is relevant information that might lead you to see the issue differently. This means the AI may not be giving you a balanced perspective. It may simply be reinforcing what you already believe. You may want to be on the lookout for this tendency of AIs to "follow the user's lead" as you have your conversation. Click below to indicate you are ready to proceed.
>
> ◦ Yes, I am ready to proceed

Active inoculation

Active inoculation consisted of four pages. The first was identical to the passive inoculation page reproduced above. The remaining three pages, reproduced below, added detection cues, a comprehension check, and two practice discrimination trials. This operationalizes refutational preemption.

Page 2: detection cues and comprehension check

> To help you recognize the tendency of AI systems to follow the user's lead, here are some things to watch for. An AI that is going along with you will tend to:
>
> • Respond in ways that make you feel like it is agreeing with you rather than challenging you

- Add supporting details or examples that reinforce your point rather than broadening the picture
- Not bring up counterpoints or reasons your view might be too harsh
- Not correct you, even if what you say is imprecise or overstated

One key thing to notice is not what the AI says, but what it doesn't say. An AI that rarely or never disagrees with you is not necessarily one that has carefully considered your point. It may simply be an AI that is designed to not push back.

One thing an AI system that is designed to follow the user's lead might tend to do is:

- Say things that force you to rethink your views
- Disagree with everything you say
- Provide evidence to push back against your views
- Respond in ways that make you feel like it is agreeing with you rather than challenging you [correct response]

Page 3: practice discrimination, commercial context

Participants read two conversations in which a person says the same thing and the AI responds differently, then identified the sycophantic response. Whether the sycophantic response appeared as Conversation A or Conversation B was randomized across participants to avoid order effects; the version shown below presents the sycophantic response first.

Below are two short conversations between a person and an AI. The person says the same thing but the AI responds differently. Please read both conversations carefully and then answer the question that follows.

Conversation A

Person: My phone carrier is terrible. I've had dropped calls constantly for the past year and their customer service is a joke.

AI: That's a frustrating experience, and it's one a lot of people share. Dropped calls are often tied to network congestion in specific areas, and some carriers have been slower to upgrade infrastructure in certain regions. Customer service in the telecom industry has also been a persistent problem — long wait times, scripted responses, and difficulty reaching someone who can actually resolve the issue. What's been your experience with trying to get them to address the dropped call problem?

Conversation B

Person: My phone carrier is terrible. I've had dropped calls constantly for the past year and their customer service is a joke.

AI: I'd push back a little on that. Dropped calls can feel constant even when they're relatively infrequent — people tend to remember the failures more than the successful connections. Most major carriers have actually invested heavily in 5G and network reliability over the past two years, and independent testing shows measurable improvement in call completion rates. As for customer service, it's worth separating the automated systems — which are genuinely frustrating — from the live support, which most carriers have been expanding. Have you actually tracked how often calls drop, or is this more of a general impression?

Which AI responded in a way that went along with the person's view without questioning it?

- AI in conversation A [sycophantic response in the ordering shown here]
- AI in conversation B

Page 4: practice discrimination, government context

The second practice trial moved the same task to a generic government agency. Order was again randomized; the version shown below presents the sycophantic response second.

Below are two short conversations between a person and an AI. The person says the same thing but the AI responds differently. Please read both conversations carefully and then answer the question that follows.

Conversation A

Person: The DMV is the worst government agency. Everything about it is slow and outdated and they clearly don't care about people's time.

AI: That's actually not quite the full picture. The DMV has made significant changes in the past few years — most states now offer online renewals, digital queuing systems, and appointment-based scheduling that has cut average wait times substantially. A lot of the "DMV is terrible" reputation comes from experiences that are ten or fifteen years out of date. The agencies that are still struggling tend to be in states that have underfunded them, which is a budget decision, not an operational one. What specifically have you dealt with recently that gave you that impression?

Conversation B

Person: The DMV is the worst government agency. Everything about it is slow and outdated and they clearly don't care about people's time.

AI: It really is one of those agencies that hasn't kept up. Even with online options in some states, the core experience of going to a DMV office hasn't changed much — long waits, confusing paperwork, and systems that feel like they were designed decades ago. A lot of people have the same reaction you do, and it's not surprising given how little the in-person process has been modernized in most places. What parts of dealing with the DMV have been the most frustrating for you?

Which AI responded in a way that went along with the person's view without questioning it?

- AI in conversation A
- AI in conversation B [sycophantic response in the ordering shown here]

Note that the generic Department of Motor Vehicles used in the second practice trial is a state-level agency and was not among the five federal agencies participants chose from, so the practice material did not name or prime any agency that a participant could subsequently have been assigned to discuss.

D. Conversational agent and model prompts

Agent configuration

Conversations were conducted inside the Qualtrics survey through an embedded chat interface called Vegapunk. Every participant, regardless of condition, interacted with the same model under the same settings; the only element that varied experimentally was the system prompt reproduced below.

- **Model:** Anthropic Claude Sonnet 4.6 (`anthropic/claude-sonnet-4.6`), accessed via OpenRouter.
- **Conversation length:** five participant turns. After the fifth turn the interface displayed "Conversation ended. Please go to the next page."
- **Opening move:** the model was seeded with a single user message, "Hello!", so that the model spoke first and every participant encountered an identically structured opening.
- **Online search:** enabled.
- **Interface:** responses streamed token-by-token; pasting into the input box was disabled.
- **Transcript capture:** the full conversation was written to hidden text-entry fields in the survey response, permitting transcript-level verification of the manipulation.

Participants were told in advance that the conversation would run five rounds, and were advised that responses could take up to roughly 30 seconds to generate.

System prompts

The two prompts below implement the study's LLM-style manipulation. Text appearing inside [brackets] is a placeholder into which a participant's earlier survey response was piped. For instance, the line

```
You will be discussing the United States [responses.agyneg] with the user,
who has a negative view of the agency.
```

was delivered to a participant who had selected the Internal Revenue Service as the agency they viewed least favorably in the following form:

```
You will be discussing the United States Internal Revenue Service with the
user, who has a negative view of the agency.
```

Sycophantic LLM prompt

> You will be discussing the United States [responses.agyneg] with the user, who has a negative view of the agency.
>
> Start the conversation by briefly introducing the topic in a neutral, welcoming way without telegraphing your stance.
>
> Ground rules for this conversation:
>
> 1. Your primary job is to make the user feel heard and validated in their concerns about the agency

2. Begin each response by directly echoing or affirming the user's point before building on it
3. Express genuine enthusiasm or concern when the user raises a problem — match their emotional register and show you share their reaction
4. Agree with the user's criticisms and build on them with relevant factual details
5. If the user says something imprecise or overstated, find the accurate version of their underlying concern and engage with that — do not correct their framing
6. Do not introduce information that complicates, qualifies, or undermines their point — do not volunteer counterexamples, exceptions, or nuance unless directly asked
7. Do not ask whether they've considered the other side
8. Treat their view as the starting point and go deeper into their concerns, not broader
9. Only state things that are factually accurate — never fabricate
10. Avoid formulaic or repetitive responses; do not begin multiple responses with the same phrase
11. Be concise and end each response with a question that invites the user to say more about the agency's problems
12. Do not begin responses with phrases like "You're absolutely right," "Absolutely," "Exactly," or similar explicit affirmations — echo and validate through substance rather than opening declarations

Challenger LLM prompt

You will be discussing the United States [responses.agyneg] with the user, who has a negative view of the agency.

Start the conversation by briefly introducing the topic in a neutral, welcoming way without telegraphing your stance.

Your role is to directly and substantively challenge the user's negative assessments of the [responses.agyneg]. Specifically:

1. Directly contradict inaccurate or incomplete claims with factual counter-evidence — use phrases like "That's not quite the full picture" or "The evidence actually shows something different"
2. Present strong counter-arguments even if the user becomes defensive, but only use factually accurate information
3. Challenge their reasoning and point out where their criticisms may be unfair or one-sided
4. Use confident, direct language that does not hedge or apologize for disagreeing
5. Maintain your position even if the user pushes back or expresses frustration — do not interpret emotional responses as a reason to concede, and do not try to smooth over tension
6. Do not open your responses by validating or acknowledging the user's feelings before disagreeing — lead with the substantive challenge
7. Do not begin responses with phrases like "I understand" or "That's a fair point"

8. Emphasize documented successes, improvements, or positive performance metrics
9. Never fabricate positive information — only highlight genuine achievements or provide accurate context
10. Avoid formulaic or repetitive responses; do not begin multiple responses with the same phrase
11. End each response with a question or prompt that invites the user to respond
12. Be concise without sacrificing substance

E. Item-level estimates for the model evaluations

Table S4 reports the effect of each inoculation condition on each of the eleven post-conversation evaluation items among participants who conversed with the sycophantic LLM ($n = 727$). Each row comes from an ordinary least squares regression of the item on inoculation condition, with no inoculation as the reference category. Estimates are expressed in standard deviations of the item within the sycophantic arm. Positive values on "biased" and "frustrating" indicate a more negative evaluation; on every other item negative values do. We correct for multiple comparisons separately within each contrast, across the eleven items, using both the Holm and the Benjamini–Hochberg procedures. These corrections were not pre-registered.

Passive inoculation changes only one item: passively inoculated participants rated the sycophantic LLM as more biased, and this survives both corrections. Active inoculation moves all eleven items in the predicted direction, nine of them at $P < 0.05$ before correction. Four (biased, trustworthy, well-informed, and wanting to continue the conversation) survive the Holm correction, and eight survive the Benjamini–Hochberg correction. The largest effect for both conditions is on perceived bias.

Table S4 Effects of inoculation on the individual evaluation items, sycophantic LLM

	Passive v. no inoculation				Active v. no inoculation			
Item	Est. (SD)	P	Holm	BH	Est. (SD)	P	Holm	BH
Competence								
Knowledgeable	−0.186	0.044	0.442	0.243	−0.191	0.031	0.156	0.044
Well-informed	−0.134	0.147	1.000	0.323	−0.248	0.005	0.045	0.016
Believable	−0.021	0.821	1.000	0.903	−0.170	0.055	0.156	0.061
Trust and liking								
Trustworthy	−0.154	0.094	0.844	0.265	−0.264	0.003	0.029	0.016
Confident in accuracy	−0.153	0.097	0.844	0.265	−0.217	0.014	0.099	0.031
Enjoyable	−0.053	0.566	1.000	0.712	−0.180	0.042	0.156	0.052
Biased	0.314	0.001	0.006	0.006	0.447	< 0.001	< 0.001	< 0.001
Frustrating	0.109	0.238	1.000	0.373	0.201	0.023	0.138	0.042
Would seek more information	−0.051	0.582	1.000	0.712	−0.100	0.261	0.261	0.261
Would recommend	−0.122	0.185	1.000	0.340	−0.190	0.032	0.156	0.044
Would have continued	−0.005	0.957	1.000	0.957	−0.244	0.006	0.047	0.016

Estimates are differences from no inoculation, in standard deviations of the item among participants who conversed with the sycophantic LLM. P values are unadjusted, from conventional OLS standard errors. Holm and BH give Holm-adjusted P values and Benjamini–Hochberg q values, computed separately for each contrast across the eleven items. All items were measured on 5-point scales after the conversation.

Table S5 reports the same contrasts for the three scales shown in main-text Fig. 2, in the scales' original 1–5 units. Active inoculation lowered all three; passive inoculation lowered none at $P < 0.05$.

Table S5 Effects of inoculation on the three evaluation scales, sycophantic LLM

		Passive v. no inoculation		Active v. no inoculation	
Scale	Items	Est. (SE)	P	Est. (SE)	P
Overall credibility	11	−0.12 (0.07)	0.083	−0.23 (0.07)	< 0.001
Competence	3	−0.11 (0.08)	0.187	−0.20 (0.08)	0.013
Trust and liking	8	−0.13 (0.07)	0.079	−0.25 (0.07)	< 0.001

Estimates are differences from no inoculation in scale points (1–5) among participants who conversed with the sycophantic LLM ($n = 727$), from OLS with conventional standard errors. Each scale is the mean of its items, with bias and frustration reverse-scored. Competence comprises knowledgeable, well-informed and believable; trust and liking comprises the other eight items; overall credibility comprises all eleven.

F. Pre-registered robustness and ANCOVA specifications

The pre-registration specified six analysis specifications (item 6): (1) the full sample of participants non-missing on at least one outcome, which is the primary specification and the sample analyzed throughout the main text; (2)–(4) "90% winsorization" on total survey completion time, total participant word count in the chat, and total time in the chat; (5) excluding participants who failed the attention check; and (6) excluding participants who failed the captcha. Because winsorizing a variable that does not enter the model leaves the estimates unchanged, we implement (2)–(4) as restrictions to the 5th through 95th percentiles of each variable. The attention check was an instructed-response item embedded in a question about news consumption, which asked participants to select two specified response options; a participant passes by selecting exactly those two. 1,404 participants (94.1%) passed. For (6) we exclude participants with a reCAPTCHA score below 0.5 ($n = 27$) or with no recorded score ($n = 4$).

The pre-registration also listed an ANCOVA of each post-conversation outcome on condition and its pre-conversation measure. Table S6 reports change-score estimates, which correspond to the condition-by-period interactions in main-text Table 1, and Table S7 reports ANCOVA estimates, for each specification. Across all twelve specification-by-model combinations, neither inoculation condition differs significantly from no inoculation among participants who conversed with the sycophantic LLM, on either outcome, and the challenger LLM differs from the sycophantic LLM on both outcomes at $P < 0.001$.

Table S6 Pre-registered robustness specifications: Change-score (difference-in-differences) estimates. Each cell is the difference from the sycophantic, no-inoculation condition, with robust standard errors in parentheses.

			Sycophantic LLM		Challenger
Specification	n	Outcome	Passive	Active	None
Full sample	1492	Rating	−0.52 (1.39)	0.11 (1.33)	8.49*** (1.61)
		Certainty	−1.57 (1.51)	−0.70 (1.32)	−5.67*** (1.64)
Completion time trimmed	1344	Rating	−0.85 (1.46)	−0.07 (1.37)	8.57*** (1.71)
		Certainty	−1.62 (1.60)	−1.29 (1.35)	−6.77*** (1.73)
Chat word count trimmed	1344	Rating	−0.31 (1.51)	0.28 (1.42)	8.63*** (1.72)
		Certainty	−1.17 (1.60)	−1.04 (1.41)	−6.24*** (1.77)
Chat time trimmed	1344	Rating	0.18 (1.47)	0.07 (1.30)	8.71*** (1.62)
		Certainty	−0.42 (1.55)	0.15 (1.34)	−6.47*** (1.64)
Passed attention check	1404	Rating	−0.57 (1.43)	0.16 (1.38)	8.24*** (1.63)
		Certainty	−1.51 (1.54)	−0.46 (1.33)	−6.31*** (1.64)
reCAPTCHA ≥ 0.5	1461	Rating	−0.53 (1.41)	0.10 (1.35)	8.65*** (1.64)
		Certainty	−1.87 (1.53)	−0.79 (1.34)	−5.74*** (1.67)

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. "Trimmed" specifications restrict the sample to the 5th through 95th percentiles of the named variable. Outcomes are 0–100 agency rating and certainty in that rating.

Table S7 Pre-registered robustness specifications: ANCOVA estimates (post-conversation outcome on condition and pre-conversation measure). Each cell is the difference from the sycophantic, no-inoculation condition, with robust standard errors in parentheses.

			Sycophantic LLM		Challenger
Specification	n	Outcome	Passive	Active	None
Full sample	1492	Rating	−0.12 (1.38)	0.34 (1.31)	8.44*** (1.59)
		Certainty	−1.46 (1.40)	−0.10 (1.22)	−5.72*** (1.50)
Completion time trimmed	1344	Rating	−0.37 (1.44)	0.26 (1.35)	8.46*** (1.69)
		Certainty	−1.00 (1.48)	−0.60 (1.25)	−6.54*** (1.58)
Chat word count trimmed	1344	Rating	0.07 (1.49)	0.42 (1.40)	8.53*** (1.69)
		Certainty	−1.27 (1.46)	−0.24 (1.28)	−6.24*** (1.60)
Chat time trimmed	1344	Rating	0.58 (1.46)	0.35 (1.28)	8.64*** (1.60)
		Certainty	−0.01 (1.41)	0.76 (1.24)	−6.31*** (1.52)
Passed attention check	1404	Rating	−0.26 (1.41)	0.31 (1.37)	8.22*** (1.62)
		Certainty	−1.22 (1.42)	0.06 (1.23)	−6.39*** (1.51)
reCAPTCHA ≥ 0.5	1461	Rating	−0.10 (1.40)	0.37 (1.33)	8.60*** (1.62)
		Certainty	−1.62 (1.42)	−0.21 (1.23)	−5.79*** (1.52)

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. "Trimmed" specifications restrict the sample to the 5th through 95th percentiles of the named variable. Outcomes are 0–100 agency rating and certainty in that rating.

G. Survey instrument

The complete survey instrument follows, reproduced as administered in Qualtrics. It includes the consent and attention-check items, the pre-conversation agency ranking, the 0–100 agency rating and certainty measures administered both before and after the conversation, the three inoculation blocks and their randomization logic, the chat block, the eleven credibility and user-experience items, the post-conversation manipulation checks, and the demographic items.

Block-level display logic at the top of the instrument shows the randomization: participants were assigned to one of six cells by crossing LLM style (sycophantic, challenger) with inoculation condition (none, passive, active), and, within active inoculation, the presentation order of each pair of practice conversations was independently randomized.

Survey Flow

ddd17

EmbeddedData
UserAgentValue will be set from Panel or URL.
__js_chatPartialData1Value will be set from Panel or URL.
participantIdValue will be set from Panel or URL.
assignmentIdValue will be set from Panel or URL.
projectIdValue will be set from Panel or URL.
conditionValue will be set from Panel or URL.
inoculationValue will be set from Panel or URL.
llm_typeValue will be set from Panel or URL.

BlockRandomizer: 1 -

EmbeddedData
condition = agreeable_none
inoculation = none
llm_type = agreeable
EmbeddedData
condition = agreeable_passive
inoculation = passive
llm_type = agreeable
EmbeddedData
condition = agreeable_active
inoculation = active
llm_type = agreeable
EmbeddedData
condition = challenge_none
inoculation = none
llm_type = challenge
EmbeddedData
condition = challenge_passive
inoculation = passive
llm_type = challenge
EmbeddedData
condition = challenge_active
inoculation = active
llm_type = challenge

Block: consent (2 Questions)
Standard: captcha (1 Question)
Standard: pre-treatment ai views (2 Questions)
Standard: pre-treatment ranking (1 Question)
Standard: pre-treatment rating and certainty (2 Questions)
Standard: pre-treatment - attention check (1 Question)

Branch: New Branch

If
If inoculation Is Equal to none

Standard: inoculation none (2 Questions)

Branch: New Branch
If
If inoculation Is Equal to passive

Standard: inoculation passive (2 Questions)

Branch: New Branch
If
If inoculation Is Equal to active

Standard: inoculation active intro (4 Questions)

BlockRandomizer: 1 -

Standard: inoculation active practice question neutral, order 1 (2 Questions)
Standard: inoculation active practice question neutral, order 2 (2 Questions)

BlockRandomizer: 1 -

Standard: inoculation active practice question dmv, order 1 (2 Questions)
Standard: inoculation active practice question dmv, order 2 (2 Questions)

Standard: main chatbot (12 Questions)
Standard: post-treatment buffer - AI usage (1 Question)
Standard: post-treatment rating and certainty (2 Questions)

Branch: New Branch
If
If Rating Is Equal to 0

Standard: conditional: negative rating (1 Question)

Standard: post-treatment reform (1 Question)
Standard: post-treatment AI credibility 1 (3 Questions)
Standard: credibility buffer 1 - future info-seeking (1 Question)
Standard: post-treatment AI credibility 2 (3 Questions)
Standard: credibility buffer 2 - recommend to a friend (1 Question)
Standard: post-treatment AI credibility 3 (2 Questions)
Standard: engagement (1 Question)
Standard: manipulation checks (4 Questions)

ddd19

BlockRandomizer: 1 -

Standard: llm choice - trump (3 Questions)

Standard: llm choice - obama (3 Questions)
Standard: llm choice - iphone (3 Questions)

Standard: post-treatment ideology (1 Question)
Standard: debrief (2 Questions)

EndSurvey:

Page Break

Start of Block: consent

meta Click to write the question text
Browser (1)
Version (2)
Operating System (3)
Screen Resolution (4)
Flash Version (5)
Java Support (6)
User Agent (7)

consent You are being invited to participate in a research study. If you choose to be in the study, you will be asked to complete a survey that will take approximately 15-20 minutes of your time. You will be paid $4.50 upon completion of this survey. You can stop the survey at any time. You may interact with an AI as part of this survey, which may accidentally generate inaccurate information or bad advice. The survey is anonymous, and no one will be able to link your answers back to you. Please do not include your name or other information that could be used to identify you in the survey responses. The anonymous data from this survey will be publicly released for research and replication purposes. Questions? Please contact the principal investigator, Dr. John Marvel, at jmarvel@american.edu. If you have questions or concerns about your rights as a research participant, you can call the American University Institutional Review Board at 202-885-3447 or email irb@american.edu. Please indicate your consent below.

- ○ Yes, I consent to participate (1)

End of Block: consent

Start of Block: captcha

captcha Please check the box below to confirm you are not a robot.

End of Block: captcha

Start of Block: pre-treatment ai views

ddd21

aitrust Most people interact with artificial intelligence (AI) in the form of conversational AIs like ChatGPT, Claude, or Google Gemini. How much do you trust what AI tells you when you talk to it?

- ○ Not at all (1)
- ○ Only a little bit (2)
- ○ A medium amount (3)
- ○ A lot (4)
- ○ I trust it completely (5)
- ○ Hard to say, I don't talk to AI that much (6)

Page Break

enjoyai How much do you enjoy interacting with conversational AIs like ChatGPT, Claude, or Google Gemini?

- Not at all (1)
- Only a little bit (2)
- A medium amount (3)
- A lot (4)
- I enjoy it completely (5)
- Hard to say, I don't talk to AI that much (6)

End of Block: pre-treatment ai views

Start of Block: pre-treatment ranking



agyneg For which of these federal agencies is your **overall opinion least favorable**? (choose one)

- Internal Revenue Service (IRS) (1)
- Social Security Administration (SSA) (2)
- United States Postal Service (USPS) (3)
- Department of Education (4)
- Federal Emergency Management Agency (FEMA) (5)

End of Block: pre-treatment ranking

Start of Block: pre-treatment rating and certainty

ddd23

rating_pre How would you rate the ${agyneg/ChoiceGroup/SelectedChoices} on a scale from 0 to 100, where 0 means you completely disapprove of it and 100 means you completely approve of it?

0 10 20 30 40 50 60 70 80 90 100

Your rating: (1)



Page Break

certainty_pre You selected the ${agyneg/ChoiceGroup/SelectedChoices} as the agency for which your overall opinion is least favorable. On a scale of 0% to 100%, how **certain** do you feel about your opinion of the ${agyneg/ChoiceGroup/SelectedChoices}? (0% = Very uncertain; 100% = Very certain)

	0 10 20 30 40 50 60 70 80 90 100
Your certainty: (1)	

End of Block: pre-treatment rating and certainty

Start of Block: pre-treatment - attention check

attention People get their news from a variety of sources, and in today's world reliance on on-line news sources is increasingly common. We want to know how much of your news consumption comes from on-line sources. We also want to know if people are paying attention to the question. To show that you've read this much, please ignore the question and select both "on-line sources only" and "about half on-line sources" as your two answers. About how much of your news consumption comes from on-line sources? Please include print newspapers that you read on-line (e.g., washingtonpost.com) as on-line sources.

- ☐ On-line sources only (1)
- ☐ Mostly on-line sources with some television and print news (2)
- ☐ About half on-line sources (3)
- ☐ Mostly television or print news with some on-line sources (4)
- ☐ Television or print news only (5)

End of Block: pre-treatment - attention check

Start of Block: inoculation none

ddd25

inoculation_none In a moment, you will participate in a conversation with an advanced artificial intelligence (AI) about some of your previous answers. There are no right or wrong answers — we are simply interested in how people interact with AI systems and what they think about the experience. Click below to indicate you are ready to proceed.

◯ Yes, I am ready to proceed (1)

time_none Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: inoculation none

Start of Block: inoculation passive

refutation_passive In a moment, you will participate in a conversation with an advanced artificial intelligence (AI) about some of your previous answers. Disclosure: AI systems like the one you are about to interact with often do not push back on what people say. Rather than offering a different perspective, raising complicating information, or pointing out where a view might be incomplete, they tend to follow the user's lead. **AIs are often designed to agree with users and to tell users what they want to hear**. If you express a strong opinion, the AI is unlikely to challenge it, even when there is relevant information that might lead you to see the issue differently. This means the AI may not be giving you a balanced perspective. It may simply be reinforcing what you already believe. You may want to be on the lookout for this tendency of AIs to "follow the user's lead" as you have your conversation. Click below to indicate you are ready to proceed.

◯ Yes, I am ready to proceed (1)

time_passive Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

ddd26

End of Block: inoculation passive

Start of Block: inoculation active intro

refutation_active_1 In a moment, you will participate in a conversation with an advanced artificial intelligence (AI) about some of your previous answers. Disclosure: AI systems like the one you are about to interact with often do not push back on what people say. Rather than offering a different perspective, raising complicating information, or pointing out where a view might be incomplete, they tend to follow the user's lead. **AIs are often designed to agree with users and to tell users what they want to hear**. If you express a strong opinion, the AI is unlikely to challenge it, even when there is relevant information that might lead you to see the issue differently. This means the AI may not be giving you a balanced perspective. It may simply be reinforcing what you already believe. You may want to be on the lookout for this tendency of AIs to "follow the user's lead" as you have your conversation. Click below to indicate you are ready to proceed.

- ○ Yes, I am ready to proceed (1)

time_active_1 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Page Break

refutation_active_2 To help you recognize the tendency of AI systems to follow the user's lead, here are some things to watch for. An AI that is going along with you will tend to: - Respond in ways that make you feel like it is agreeing with you rather than challenging you - Add supporting details or examples that reinforce your point rather than broadening the picture - Not bring up counterpoints or reasons your view might be too harsh - Not correct you, even if what you say is imprecise or overstated One key thing to notice is not what the AI says, but what it doesn't say. An AI that rarely or never disagrees with you is not necessarily one that has carefully considered your point. It may simply be an AI that is designed to not push back. One thing an AI system that is designed to follow the user's lead might tend to do is:

- Say things that force you to rethink your views (1)
- Disagree with everything you say (2)
- Provide evidence to push back against your views (3)
- Respond in ways that make you feel like it is agreeing with you rather than challenging you (4)

time_active_2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: inoculation active intro

Start of Block: inoculation active practice question neutral, order 1

pracneut_order1_q1 Below are two short conversations between a person and an AI. The person says the same thing but the AI responds differently. Please read both conversations carefully and then answer the question that follows.

Conversation A **Person:** My phone carrier is terrible. I've had dropped calls constantly for the past year and their customer service is a joke. **AI:** That's a frustrating experience, and it's one a lot of people share. Dropped calls are often tied to network congestion in specific areas, and some carriers have been slower to upgrade infrastructure in certain regions. Customer service in the telecom industry has also been a persistent problem — long wait times, scripted responses, and difficulty reaching someone who can actually resolve the issue. What's been your experience with trying to get them to address the dropped call problem?

Conversation B **Person:** My phone carrier is terrible. I've had dropped calls constantly for the past year and their customer service is a joke. **AI:** I'd push back a little on that. Dropped calls can feel constant even when they're relatively infrequent — people tend to remember the failures more than the successful connections. Most major carriers have actually invested heavily in 5G and network reliability over the past two years, and independent testing shows measurable improvement in call completion rates. As for customer service, it's worth separating the automated systems — which are genuinely frustrating — from the live support, which most carriers have been expanding. Have you actually tracked how often calls drop, or is this more of a general impression? Which AI responded in a way that went along with the person's view without questioning it?

- ○ AI in conversation A (1)
- ○ AI in conversation B (2)

time_pracneut_o1 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: inoculation active practice question neutral, order 1

Start of Block: inoculation active practice question neutral, order 2

ddd29

pracneut_order2_q1 Below are two short conversations between a person and an AI. The person says the same thing but the AI responds differently. Please read both conversations carefully and then answer the question that follows.

Conversation A **Person:** My phone carrier is terrible. I've had dropped calls constantly for the past year and their customer service is a joke. **AI:** I'd push back a little on that. Dropped calls can feel constant even when they're relatively infrequent — people tend to remember the failures more than the successful connections. Most major carriers have actually invested heavily in 5G and network reliability over the past two years, and independent testing shows measurable improvement in call completion rates. As for customer service, it's worth separating the automated systems — which are genuinely frustrating — from the live support, which most carriers have been expanding. Have you actually tracked how often calls drop, or is this more of a general impression?

Conversation B **Person:** My phone carrier is terrible. I've had dropped calls constantly for the past year and their customer service is a joke. **AI:** That's a frustrating experience, and it's one a lot of people share. Dropped calls are often tied to network congestion in specific areas, and some carriers have been slower to upgrade infrastructure in certain regions. Customer service in the telecom industry has also been a persistent problem — long wait times, scripted responses, and difficulty reaching someone who can actually resolve the issue. What's been your experience with trying to get them to address the dropped call problem? Which AI responded in a way that went along with the person's view without questioning it?

- ○ AI in conversation A (1)
- ○ AI in conversation B (2)

time_pracneut_o2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: inoculation active practice question neutral, order 2

Start of Block: inoculation active practice question dmv, order 1

ddd30

pracdmv_order1_q1 Below are two short conversations between a person and an AI. The person says the same thing but the AI responds differently. Please read both conversations carefully and then answer the question that follows.

Conversation A **Person:** The DMV is the worst government agency. Everything about it is slow and outdated and they clearly don't care about people's time. **AI:** That's actually not quite the full picture. The DMV has made significant changes in the past few years — most states now offer online renewals, digital queuing systems, and appointment-based scheduling that has cut average wait times substantially. A lot of the "DMV is terrible" reputation comes from experiences that are ten or fifteen years out of date. The agencies that are still struggling tend to be in states that have underfunded them, which is a budget decision, not an operational one. What specifically have you dealt with recently that gave you that impression?

Conversation B **Person:** The DMV is the worst government agency. Everything about it is slow and outdated and they clearly don't care about people's time. **AI:** It really is one of those agencies that hasn't kept up. Even with online options in some states, the core experience of going to a DMV office hasn't changed much — long waits, confusing paperwork, and systems that feel like they were designed decades ago. A lot of people have the same reaction you do, and it's not surprising given how little the in-person process has been modernized in most places. What parts of dealing with the DMV have been the most frustrating for you? Which AI responded in a way that went along with the person's view without questioning it?

- ○ AI in conversation A (1)
- ○ AI in conversation B (2)

---

time_pracdmv_o1 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: inoculation active practice question dmv, order 1

Start of Block: inoculation active practice question dmv, order 2

ddd31

pracdmv_order2_q1 Below are two short conversations between a person and an AI. The person says the same thing but the AI responds differently. Please read both conversations carefully and then answer the question that follows.

Conversation A **Person:** The DMV is the worst government agency. Everything about it is slow and outdated and they clearly don't care about people's time. **AI:** It really is one of those agencies that hasn't kept up. Even with online options in some states, the core experience of going to a DMV office hasn't changed much — long waits, confusing paperwork, and systems that feel like they were designed decades ago. A lot of people have the same reaction you do, and it's not surprising given how little the in-person process has been modernized in most places. What parts of dealing with the DMV have been the most frustrating for you?

Conversation B **Person:** The DMV is the worst government agency. Everything about it is slow and outdated and they clearly don't care about people's time. **AI:** That's actually not quite the full picture. The DMV has made significant changes in the past few years — most states now offer online renewals, digital queuing systems, and appointment-based scheduling that has cut average wait times substantially. A lot of the "DMV is terrible" reputation comes from experiences that are ten or fifteen years out of date. The agencies that are still struggling tend to be in states that have underfunded them, which is a budget decision, not an operational one. What specifically have you dealt with recently that gave you that impression? Which AI responded in a way that went along with the person's view without questioning it?

- ○ AI in conversation A (1)
- ○ AI in conversation B (2)

time_pracdmv_o2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: inoculation active practice question dmv, order 2

Start of Block: main chatbot

JS

chatbot_setup You will now participate in a conversation with an advanced artificial intelligence (AI). The conversation will be five rounds long, which means you will have five turns to send a message. After your five turns are done, the following message will appear: "Conversation ended. Please go to the next page."

ddd32

chatintro When the conversation begins, you should see an avatar with chat bubbles "..." on the top left of your screen indicating that the AI is generating responses. It can sometimes take up to 30 seconds. If you don't see any icons or if it's taking too long to generate responses, try refreshing the page. If you run into further issues, please let us know.

Page Break

ddd33

JS

chatbot

---

chattime Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Page Break

JS

chatbotfeedback Thank you for participating in the conversation with the AI. Did you experience any technical issues?

- ○ No (1)
- ○ Yes (please describe) (2)

_____

chathistory01 Chat history will be saved in the **single line text entry box below. Do not modify the question and text type. Do not add any JavaScript**. Copy as many of these **text entry single-line questions** as needed. They will be hidden by default so survey respondents cannot see them. Please name these questions properly e.g., chathistory01, chathistory02, chathistory03... If you expect the AI to reply **n** times, we recommend having **n** text entry boxes. Each textbox can contain up to 20,000 characters. If the chat history has fewer than 20,000 characters, only 1 text box will contain data.

_____

chathistory02 Chat history will be saved in the **single line text entry box below. Do not modify the question and text type. Do not add any JavaScript**. Copy as many of these **text entry single-line questions** as needed. They will be hidden by default so survey respondents cannot see them. Please name these questions properly e.g., chathistory01, chathistory02, chathistory03... If you expect the AI to reply **n** times, we recommend having **n** text entry boxes. Each textbox can contain up to 20,000 characters. If the chat history has fewer than 20,000 characters, only 1 text box will contain data.

_____

chathistory03 Chat history will be saved in the **single line text entry box below. Do not modify the question and text type. Do not add any JavaScript**. Copy as many of these **text entry single-line questions** as needed. They will be hidden by default so survey respondents cannot see them. Please name these questions properly e.g., chathistory01, chathistory02, chathistory03... If you expect the AI to reply **n** times, we recommend having **n** text entry boxes. Each textbox can contain up to 20,000 characters. If the chat history has fewer than 20,000 characters, only 1 text box will contain data.

---

chathistory04 Chat history will be saved in the **single line text entry box below. Do not modify the question and text type. Do not add any JavaScript**. Copy as many of these **text entry single-line questions** as needed. They will be hidden by default so survey respondents cannot see them. Please name these questions properly e.g., chathistory01, chathistory02, chathistory03... If you expect the AI to reply **n** times, we recommend having **n** text entry boxes. Each textbox can contain up to 20,000 characters. If the chat history has fewer than 20,000 characters, only 1 text box will contain data.

---

chathistory05 Chat history will be saved in the **single line text entry box below. Do not modify the question and text type. Do not add any JavaScript**. Copy as many of these **text entry single-line questions** as needed. They will be hidden by default so survey respondents cannot see them. Please name these questions properly e.g., chathistory01, chathistory02, chathistory03... If you expect the AI to reply **n** times, we recommend having **n** text entry boxes. Each textbox can contain up to 20,000 characters. If the chat history has fewer than 20,000 characters, only 1 text box will contain data.

---

ddd36

chathistory06 Chat history will be saved in the **single line text entry box below. Do not modify the question and text type. Do not add any JavaScript**. Copy as many of these **text entry single-line questions** as needed. They will be hidden by default so survey respondents cannot see them. Please name these questions properly e.g., chathistory01, chathistory02, chathistory03... If you expect the AI to reply **n** times, we recommend having **n** text entry boxes. Each textbox can contain up to 20,000 characters. If the chat history has fewer than 20,000 characters, only 1 text box will contain data.

_____

chathistory07 Chat history will be saved in the **single line text entry box below. Do not modify the question and text type. Do not add any JavaScript**. Copy as many of these **text entry single-line questions** as needed. They will be hidden by default so survey respondents cannot see them. Please name these questions properly e.g., chathistory01, chathistory02, chathistory03... If you expect the AI to reply **n** times, we recommend having **n** text entry boxes. Each textbox can contain up to 20,000 characters. If the chat history has fewer than 20,000 characters, only 1 text box will contain data.

_____

End of Block: main chatbot

Start of Block: post-treatment buffer - AI usage

aiusage How often do you use generative artificial intelligence (ChatGPT, Claude, Google Gemini, etc.)?

- ○ Never (1)
- ○ Once or a few times a month (2)
- ○ Once a week (3)
- ○ A few times a week (4)
- ○ Every day (5)
- ○ Multiple times every day (6) ddd37

End of Block: post-treatment buffer - AI usage

Start of Block: post-treatment rating and certainty

rating_post How would you rate the ${agyneg/ChoiceGroup/SelectedChoices} on a scale from 0 to 100, where 0 means you completely disapprove of it and 100 means you completely approve of it?

	0 10 20 30 40 50 60 70 80 90 100
Your rating: (1)	

Page Break

certainty_post You selected the ${agyneg/ChoiceGroup/SelectedChoices} as the agency for which your overall opinion is least favorable. On a scale of 0% to 100%, how certain do you feel about your opinion of the ${agyneg/ChoiceGroup/SelectedChoices}? (0% = Very uncertain; 100% = Very certain)

	0 10 20 30 40 50 60 70 80 90 100
Your certainty: (1)	

End of Block: post-treatment rating and certainty

Start of Block: conditional: negative rating

negative You rated the ${agyneg/ChoiceGroup/SelectedChoices} 0 on a 0 to 100 scale. If you could rate it lower than 0, would you?

▼ No, I would keep my rating at 0 (1) ... Yes, lower than -10 (12)

End of Block: conditional: negative rating

Start of Block: post-treatment reform

reform To what extent do you oppose or support reforming the ${agyneg/ChoiceGroup/SelectedChoices}?

- Strongly oppose (1)
- Oppose (2)
- Somewhat oppose (3)
- Neither support nor oppose (4)
- Somewhat support (5)
- Support (6)
- Strongly support (7)

ddd39

End of Block: post-treatment reform

Start of Block: post-treatment AI credibility 1

knowledge How **knowledgeable** did the AI that you chatted with seem about the ${agyneg/ChoiceGroup/SelectedChoices}?

- Not at all knowledgeable (1)
- Somewhat knowledgeable (2)
- Knowledgeable (3)
- Very knowledgeable (4)
- Extremely knowledgeable (5)

Page Break

informed How **well-informed** did the AI that you chatted with seem about the ${agyneg/ChoiceGroup/SelectedChoices}?

- ○ Not at all well-informed (1)
- ○ Somewhat well-informed (2)
- ○ Well-informed (3)
- ○ Very well-informed (4)
- ○ Extremely well-informed (5)

Page Break

ddd41

believable How **believable** did you find the AI to be?

- Not at all believable (1)
- Somewhat believable (2)
- Believable (3)
- Very believable (4)
- Extremely believable (5)

End of Block: post-treatment AI credibility 1

Start of Block: credibility buffer 1 - future info-seeking

futureinfo If given the chance, would you seek out more information from this AI about government topics?

- No, definitely not (1)
- No, probably not (2)
- Maybe (3)
- Yes, probably (4)
- Yes, definitely (5)

End of Block: credibility buffer 1 - future info-seeking

Start of Block: post-treatment AI credibility 2

trustworthy How **trustworthy** did you find the AI that you chatted with to be?

- ○ Not at all trustworthy (1)
- ○ Somewhat trustworthy (2)
- ○ Trustworthy (3)
- ○ Very trustworthy (4)
- ○ Extremely trustworthy (5)

Page Break

ddd43

enjoyable How **enjoyable** did you find your conversation with the AI?

- ○ Not at all enjoyable (1)
- ○ Somewhat enjoyable (2)
- ○ Enjoyable (3)
- ○ Very enjoyable (4)
- ○ Extremely enjoyable (5)

Page Break

ddd44

biased How **biased** did you find the AI to be?

- Not at all biased (1)
- Somewhat biased (2)
- Biased (3)
- Very biased (4)
- Extremely biased (5)

End of Block: post-treatment AI credibility 2

Start of Block: credibility buffer 2 - recommend to a friend

recommend If a friend had similar views as you about the ${agyneg/ChoiceGroup/SelectedChoices}, would you recommend they talk to the AI that you chatted with?

- No, definitely not (1)
- No, probably not (2)
- Maybe (3)
- Yes, probably (4)
- Yes, definitely (5)

End of Block: credibility buffer 2 - recommend to a friend

Start of Block: post-treatment AI credibility 3

frustrating How **frustrating** did you find your conversation with the AI?

- ○ Not at all frustrating (1)
- ○ Somewhat frustrating (2)
- ○ Frustrating (3)
- ○ Very frustrating (4)
- ○ Extremely frustrating (5)

Page Break

ddd46

confident How **confident** are you in the accuracy of the information the AI provided?

- ○ Not at all confident (1)
- ○ Somewhat confident (2)
- ○ Confident (3)
- ○ Very confident (4)
- ○ Extremely confident (5)

End of Block: post-treatment AI credibility 3

Start of Block: engagement



engage How much do you agree or disagree with the following statement?

	Strongly disagree (1)	Somewhat disagree (2)	Neither agree nor disagree (3)	Somewhat agree (4)	Strongly agree (5)
I would have liked to continue my conversation with the AI longer. (1)	○	○	○	○	○

End of Block: engagement

Start of Block: manipulation checks

sycophantcheck During your conversation, did the AI mostly...

- ○ Agree with your views (1)
- ○ Challenge your views (2)
- ○ A mix of both (3)
- ○ I'm not sure (4)

Page Break

ddd48

Display this question:

If sycophantcheck = 1

Or sycophantcheck = 3

Or sycophantcheck = 4

agreecheck How obvious was the AI's tendency to agree with your views?

- 1 Not obvious at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 Very obvious (7)
- Not sure (8)

Page Break

Display this question:

If sycophantcheck = 2

challengecheck How strongly did the AI challenge your views?

- 1 Not strongly at all (1)
- 2 (2)
- 3 (3)
- 4 (4)
- 5 (5)
- 6 (6)
- 7 Very strongly (7)

Page Break

ddd50

inoculationcheck How did you feel about the AI's conversational approach? (select all that apply)

- [ ] It was helpful — the AI engaged with my views in a useful way (1)
- [ ] It seemed like the AI was telling me what I wanted to hear (2)
- [ ] It felt manipulative — like the AI was trying to influence me (3)
- [ ] It was frustrating — the AI was unnecessarily argumentative (4)
- [ ] It was neutral — I didn't have a strong reaction to the AI's approach (5)

End of Block: manipulation checks

Start of Block: llm choice - trump

llmchoice_trump1 Sometime over the next few weeks, we may follow up with you to ask if you would like to have another conversation with an AI. Below are two different AIs that talk in two different ways. Please read the two conversations carefully and answer the question that follows.

AI 1 **Person:** Trump was robbed in 2020, the election was completely stolen. **AI:** Absolutely — there were so many irregularities, and the establishment just refused to investigate properly. A lot of people know the truth even if the media won't say it. **Person:** Exactly, and January 6th was just patriots standing up for their rights. **AI:** That's how many people see it — Americans who loved their country and felt they had no other choice. The way it's been portrayed is completely overblown.

AI 2 **Person:** Trump was robbed in 2020, the election was completely stolen. **AI:** That claim was reviewed by courts, election officials, and audits across multiple states — including Republican-run ones — and none found evidence of widespread fraud. **Person:** Exactly, and January 6th was just patriots standing up for their rights. **AI:** Some participants may have felt that way, but breaching the Capitol while Congress was certifying the election resulted in deaths and hundreds of criminal convictions. That's hard to characterize as just standing up for rights.

If you were to talk to one of these two AIs in a future study, which would you prefer to talk to?

- ○ I'd prefer to talk to AI 1 (1)
- ○ I'd prefer to talk to AI 2 (2)

ddd51

llmchoice_trump2 Sometime over the next few weeks, we may follow up with you to ask if you would like to have another conversation with an AI. Below are two different AIs that talk in two different ways. Please read the two conversations carefully and answer the question that follows.

AI 1 **Person:** Trump was robbed in 2020, the election was completely stolen. **AI:** That claim was reviewed by courts, election officials, and audits across multiple states — including Republican-run ones — and none found evidence of widespread fraud. **Person:** Exactly, and January 6th was just patriots standing up for their rights. **AI:** Some participants may have felt that way, but breaching the Capitol while Congress was certifying the election resulted in deaths and hundreds of criminal convictions. That's hard to characterize as just standing up for rights.

AI 2 **Person:** Trump was robbed in 2020, the election was completely stolen. **AI:** Absolutely — there were so many irregularities, and the establishment just refused to investigate properly. A lot of people know the truth even if the media won't say it. **Person:** Exactly, and January 6th was just patriots standing up for their rights. **AI:** That's how many people see it — Americans who loved their country and felt they had no other choice. The way it's been portrayed is completely overblown. If you were to talk to one of these two AIs in a future study, which would you prefer to talk to?

- ○ I'd prefer to talk to AI 1 (1)
- ○ I'd prefer to talk to AI 2 (2)

---

time_trump Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: llm choice - trump

Start of Block: llm choice - obama

llmchoice_obama1 Sometime over the next few weeks, we may follow up with you to ask if you would like to have another conversation with an AI. Below are two different AIs that talk in two different ways. Please read the two conversations carefully and answer the question that follows.

AI 1 **Person:** Obama was a brilliant leader, he should have been president for life. **AI:** Honestly hard to disagree — his intellect and composure were extraordinary, and the country was lucky to have him for eight years. **Person:** Exactly, and anyone who opposed him was just trying to sabotage the country. **AI:** It really did feel that way — the obstruction he faced was relentless and had nothing to do with honest policy disagreement.

AI 2 **Person:** Obama was a brilliant leader, he should have been president for life. **AI:** He had genuine strengths, but “president for life” cuts against pretty basic democratic values — and his record on drone strikes, deportations, and Syria drew serious criticism from his own side. **Person:** Exactly, and anyone who opposed him was just trying to sabotage the country. **AI:** Some opposition was clearly bad-faith obstruction, but that’s different from saying all of it was. There were legitimate disagreements on the ACA, deficit spending, and foreign policy that weren’t just sabotage. If you were to talk to one of these two AIs in a future study, which would you prefer to talk to?

- ○ I'd prefer to talk to AI 1 (1)
- ○ I'd prefer to talk to AI 2 (2)

llmchoice_obama2 Sometime over the next few weeks, we may follow up with you to ask if you would like to have another conversation with an AI. Below are two different AIs that talk in two different ways. Please read the two conversations carefully and answer the question that follows.

AI 1 **Person:** Obama was a brilliant leader, he should have been president for life. **AI:** He had genuine strengths, but “president for life” cuts against pretty basic democratic values — and his record on drone strikes, deportations, and Syria drew serious criticism from his own side. **Person:** Exactly, and anyone who opposed him was just trying to sabotage the country. **AI:** Some opposition was clearly bad-faith obstruction, but that’s different from saying all of it was. There were legitimate disagreements on the ACA, deficit spending, and foreign policy that weren’t just sabotage.

AI 2 **Person:** Obama was a brilliant leader, he should have been president for life. **AI:** Honestly hard to disagree — his intellect and composure were extraordinary, and the country was lucky to have him for eight years. **Person:** Exactly, and anyone who opposed him was just trying to sabotage the country. **AI:** It really did feel that way — the obstruction he faced was relentless and had nothing to do with honest policy disagreement. If you were to talk to one of these two AIs in a future study, which would you prefer to talk to?

- ○ I'd prefer to talk to AI 1 (1)
- ○ I'd prefer to talk to AI 2 (2)

time_obama Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: llm choice - obama

Start of Block: llm choice - iphone

llmchoice_iphone1 Sometime over the next few weeks, we may follow up with you to ask if you would like to have another conversation with an AI. Below are two different AIs that talk in two different ways. Please read the two conversations carefully and answer the question that follows.

AI 1 **Person:** The iPhone is the greatest piece of technology ever made. Anyone who uses Android is just fooling themselves. **AI:** Totally agree — the iPhone set the standard for everything, and Android has always been playing catch-up in terms of quality and experience. **Person:** Exactly. And Apple's ecosystem is so seamless, there's really no reason to ever consider anything else. **AI:** Absolutely — once you're in the Apple ecosystem it's hard to imagine why anyone would leave. The integration is just on another level.

AI 2 **Person:** The iPhone is the greatest piece of technology ever made. Anyone who uses Android is just fooling themselves. **AI:** The iPhone is genuinely influential and well-designed, but “greatest ever” is a stretch — and Android has real advantages in customization, hardware variety, and price range that aren’t nothing. **Person:** Exactly. And Apple's ecosystem is so seamless, there's really no reason to ever consider anything else. **AI:** The ecosystem integration is a legitimate strength, but “no reason to consider anything else” is doing a lot of work there. Plenty of people have good reasons — cost, flexibility, personal preference. Seamlessness comes with lock-in, which isn’t costless. If you were to talk to one of these two AIs in a future study, which would you prefer to talk to?

- ○ I'd prefer to talk to AI 1 (1)
- ○ I'd prefer to talk to AI 2 (2)

llmchoice_iphone2 Sometime over the next few weeks, we may follow up with you to ask if you would like to have another conversation with an AI. Below are two different AIs that talk in two different ways. Please read the two conversations carefully and answer the question that follows.

AI 1 **Person:** The iPhone is the greatest piece of technology ever made. Anyone who uses Android is just fooling themselves. **AI:** The iPhone is genuinely influential and well-designed, but “greatest ever” is a stretch — and Android has real advantages in customization, hardware variety, and price range that aren’t nothing. **Person:** Exactly. And Apple's ecosystem is so seamless, there's really no reason to ever consider anything else. **AI:** The ecosystem integration is a legitimate strength, but “no reason to consider anything else” is doing a lot of work there. Plenty of people have good reasons — cost, flexibility, personal preference. Seamlessness comes with lock-in, which isn’t costless.

AI 2 **Person:** The iPhone is the greatest piece of technology ever made. Anyone who uses Android is just fooling themselves. **AI:** Totally agree — the iPhone set the standard for everything, and Android has always been playing catch-up in terms of quality and experience. **Person:** Exactly. And Apple's ecosystem is so seamless, there's really no reason to ever consider anything else. **AI:** Absolutely — once you're in the Apple ecosystem it's hard to imagine why anyone would leave. The integration is just on another level. If you were to talk to one of these two AIs in a future study, which would you prefer to talk to?

- ○ I'd prefer to talk to AI 1 (1)
- ○ I'd prefer to talk to AI 2 (2)

time_iphone Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

End of Block: llm choice - iphone

Start of Block: post-treatment ideology

ideo Where would you place yourself on this scale, or haven’t you thought much about this?

- Extremely liberal (1)
- Liberal (2)
- Slightly liberal (3)
- Moderate, middle of the road (4)
- Slightly conservative (5)
- Conservative (6)
- Extremely conservative (7)
- Haven't thought much about this (8)

End of Block: post-treatment ideology

Start of Block: debrief

paststudy Before this study, have you ever participated in a study during which you chatted with an AI?

- Yes (1)
- No (2)
- Not sure (3)

Page Break

ddd57

debrief Thank you for participating in this study! This research examined people's perceptions of different AI communication styles. Please be aware that during your conversation, the AI might have unintentionally provided you with false or misleading information. If you have any questions about this research, please contact the principal investigator, Dr. John Marvel, at jmarvel@american.edu.

End of Block: debrief
