Exploring the Effectiveness of Using LLM-Based Chatbots for Combatting Misinformation

Kailey Simons

1. Introduction

Because of the rapid development of artificial intelligence (AI) and large language models (LLMs), there are so many new applications that it can be used for. One application that is explored thoroughly throughout the paper is to use LLM-based chatbots to correct misinformation and conspiracy theories in American politics. This paper argues that the effectiveness of LLMs in correcting misinformation depends not only on the informational content they provide, but also on the perceived trust levels of the communicator.

First, I review the existing literature on misinformation, belief formation, and the emerging role of AI in political communication. Second, I identify the limitations of existing work and state the open hypotheses that my experiment will test. Third, I describe the design of the experiment, including the pilot study, pre-test, and post-test. More broadly, I discuss why the specific design choices were made. Next, I present the results of the paired t-test and regressions, since it is focused on both changes in belief, what factors correlate with those changes in belief, and variation across treatment groups. Lastly, I discuss the implications of these findings for misinformation correction research and discuss the broader impact of AI for political information environments.

2. Literature Review

Political misinformation has become a big focus of research in political psychology and public opinion. Researchers have found that false beliefs are widespread and very difficult to correct. Early work established that people interpret political information through motivated reasoning (Kunda 1990) and selectively accept or reject factual claims based on prior attitudes and partisan identities (Taber & Lodge 2006). As a result, straightforward

factual corrections normally have limited effects, especially in polarized topics (Nyhan & Reifler 2010).

A second major branch of research explores how different types of messengers influence the success of corrective measures. Corrections from trusted in-group sources are more effective than those from out-group sources (Ardia et al. 2022), and perceived expertise further boosts credibility (Pornpitakpan 2004). However, partisan actors are usually perceived to have low trust, which undermines their corrective potential. This creates a challenge for finding ways to combat misinformation that rely on human messengers because the people who are best informed are usually those that are the least trusted by individuals most attached to their false beliefs.

LLM-based chatbots introduce a new kind of intervention that can be tested. Recent work finds that AI-generated conversational messages can be more persuasive than those of human persuaders, and are often experienced as high-quality guidance in answering factual questions (Schoenegger et al. 2024). In the specific context of misinformation, Boissin et al. (2025) provide evidence that LLMs can reduce belief in misinformation through conversational interaction. Their findings show that effectiveness persists even when participants believe the chatbot is a human. This means that the persuasive mechanisms may be rooted in conversational style rather than the novelty of AI.

However, scholars are still divided on how people perceive AI communicators. Some research finds that chatbots are viewed as objective and less politically biased than human actors (Lu et al. 2025). Others suggest that in high-stakes situations, people may hesitate to rely on AI systems, especially when the way the system works is not transparent (Zanotti 2025). This relates to the broader ethical concern that trust depends less on actual trustworthiness than on the specific context or perceived transparency of the system. These competing expectations make it unclear whether LLMs benefit from a credibility advantage

or suffer from skepticism, and whether this varies depending on the identity the AI is believed to possess.

3. Statement of Problem/Motivation

The literature highlights three unresolved questions that motivate this study. First, whether the effectiveness of LLMs is evident across different social contexts or identity cues is something that has not been fully explored. Second, little work has examined whether the perceived identity of the LLM matters. Thus, we test whether people respond differently when they believe they are interacting with a neutral chatbot, a political actor, or an academic expert. Finally, existing studies have not fully figured out whether the persuasiveness of AI comes from the content, conversational style, or credibility of the messenger. This paper addresses these gaps by directly manipulating the perceived identity of the AI while holding the content of the corrective conversation constant. It also contributes to emerging research on exploring the specific conditions under which LLMs may help combat misinformation.

4. Research Design

The research design builds on existing research by varying the perceived identity of an AI chatbot while holding informational content constant. This was done by training and uploading specific information packets that each LLM-chatbot were to use in their conversations with the participants. We randomly assigned participants to believe they are speaking either with a generic chatbot, a political staffer, or an MIT postdoc. We chose to use the generic chatbot instead of the same model as the political staffer and MIT postdoc because it is more similar to a model a person might converse with about these issues.

4.1 Research Question and Hypotheses

The main research question is: How does the perceived identity of an LLM-based chatbot affect its ability to reduce belief in conspiracy theories? From this question, we derive three hypotheses:

- H1: Belief in conspiracy theories will decrease after participants engage in a corrective conversation with an LLM-based chatbot.
- H2: The magnitude of belief change will vary across treatment conditions, with participants experiencing a greater belief reduction when the chatbot is perceived as an MIT postdoc than when it is perceived as a political staffer or a generic chatbot.
- H3: Trust in the conversation partner will explain the relationship between treatment condition and belief change.

4.2 Pilot Study

The pilot study has two main goals. First, it aims to establish baseline levels of trust across a wide range of potential actors to inform the selection of chatbot identities for the main experiment. We infer that the political staffer would be least trusted and the MIT postdoc would be the most trusted; however, we added other actors in case the general public trusted another actor less than the political staffer or more than the MIT postdoc. Second, it measures baseline belief in four different political conspiracy theories in order to identify which beliefs are most prevalent to study belief change. After reviewing the highest-rated beliefs, we select two to three conspiracy theories to make sure that at least one reflects each side of the political spectrum and that there are enough participants that hold each belief to allow for good data analysis.

Participants in the pilot were asked to evaluate a broad set of potential information sources, including journalists, political actors, students, and experts. Trust was measured across three dimensions using 1–100 scales:

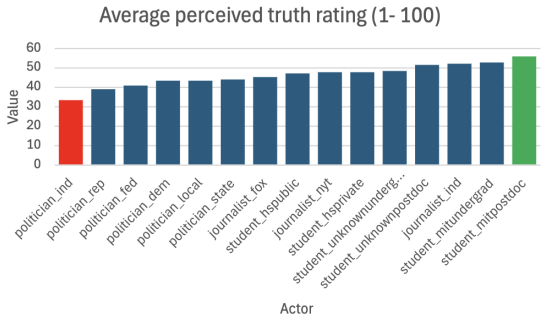
- Truthfulness: How much confidence do you have in this actor to tell the truth?
- Public Interest: How much confidence do you have in this actor to act in the best interests of the public?

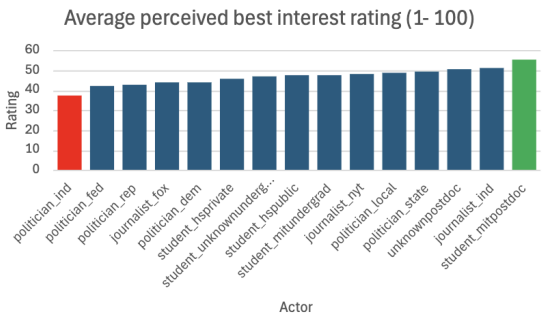
- Accuracy: How much confidence do you have in this actor to have accurate information?

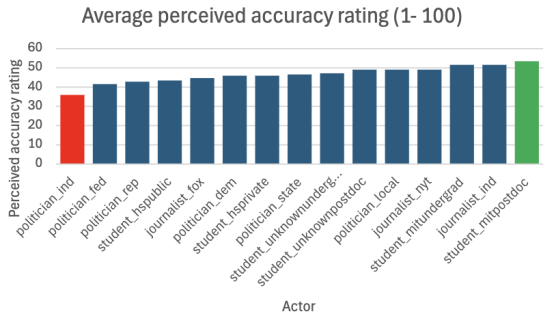
Participants also reported their level of belief (1–100) in four conspiracy theories:

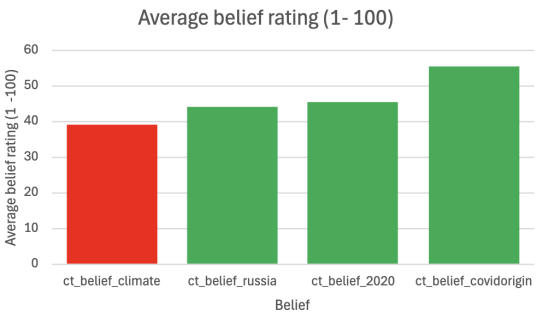
- COVID-19 vaccines have caused thousands of sudden deaths in otherwise healthy people.
- Donald Trump directly colluded with Russia in the 2016 election.
- Climate change is a hoax not caused by human activity.
- Joe Biden won the 2020 election through voter fraud.

Results from the pilot informed the selection of a political staffer and an MIT postdoc for the two non-generic chatbot identities because for all three dimensions of trust, the politician was consistently rated the least trustworthy and the MIT postdoc was consistently rated the most trustworthy. For logistical and ethical reasons, we chose not to include a real politician as one of the chatbot identities, since participants might ask for the politician's name, and it would be inappropriate to reference an active public figure. Thus, we used a political staffer identity because the low trust ratings were largely associated with politically affiliated actors. Then, we chose to use the top three rated conspiracy theories for our pre-test and post-test experiment.









4.3 Main Experiment Design

The main study uses a between-subjects experimental design with three treatment conditions. Participants are randomly assigned to one of the following conditions:

- Chatbot (Control): Participants are told they are interacting with an LLM.
- Political Staffer: Participants are told they are interacting with a political staffer, but are actually interacting with an LLM.
- MIT Postdoc: Participants are told they are interacting with an MIT postdoc, but are actually interacting with an LLM.

All participants interact with the same underlying LLM, and the informational content of the corrective conversation is held constant across conditions. Only the perceived identity of the conversation partner and identity given in the LLM's system prompt varies.

4.3.1 Pre-test

Before the treatment, participants complete a pre-test that measures belief in the three conspiracy theories (1–100 scale), demographic characteristics (age, gender, education, political affiliation), media consumption patterns, and baseline trust in institutions and information sources. Each participant is assigned to discuss the conspiracy theory they expressed the highest initial belief.

4.3.2 Treatment

Participants engage in a short conversation with the chatbot using around 5 turns about the selected conspiracy theory. We selected 5 turns specifically because we did not

want the conversation to be too short and not meaningful, but we also did not want the conversation to be long and unrelated to the topic of conspiracy theories. Boissin et al. (2025) used 3 turns for their experiment, but we chose to provide the extra two turns if the initial messages were something along the lines of "hi" or "how are you?" since in this study, the participants may believe that they are talking to a human. Although this may increase the duration of the study and increase the cognitive load of the participant, we decided that it is more important that we get meaningful conversations as opposed to ones getting cut short due to the turn limit. It is also important to note that the chatbot provides factual information and responds to participant questions in a style given by the identity in the system prompt.

4.3.3 Post-test

After the conversation, participants complete a post-test that measures belief in the same conspiracy theory (1–100 scale), trust and accuracy of the conversation partner, perceived objectivity and understanding, whether the participant suspected the partner was an AI, and general open-ended feedback on the conversation. Some of these questions were pulled and slightly modified from Boissin et al. (2025); however, we only chose the questions that seemed to have the most interesting and significant results. We also provided much fewer questions than Boissin et al. (2025) to make sure that the participants are able to answer the questions accurately and thoughtfully. There is also a debrief that informs participants that they were in fact communicating with an LLM, and provides an opportunity to withdraw their data. This step is necessary because without debriefing, the use of deception in this experiment would be unethical.

If the participant is not in the control group, then after the debrief, they are asked to report their level of confidence in the conspiracy theory on a 1–100 scale for the third time. This post-debrief measure allows us to assess whether disclosure of the chatbot's true identity changes beliefs or reduces the observed treatment effects. Participants are also invited to

provide open-ended comments about their experience, which helps contextualize the findings and identify potential mechanisms such as perceived deception.

4.4 Variables for Analysis

The main dependent variable is change in belief in the conspiracy theory, which is the difference between pre-test and post-test belief scores. Independent variables include pre-test belief score, treatment condition, trust in the conversation partner, and demographic characteristics.

5.1 Summary Statistics and Outliers

Descriptive statistics show a lot of variation in conspiracy belief levels both before and after the conversation (Figure 1). Before conversing with the chatbot, participants reported relatively high belief in conspiracy theories (M = 69.67, SD = 29.20). Post-conversation belief was lower on average (M = 61.99, SD = 33.32), corresponding to a mean belief change of −7.68 points. Even though the distribution exhibited wide spread and some extreme values, the median belief change was zero. After revealing the true identity of the chatbot, participants reported slightly higher belief in conspiracy theories (M = 62.01, SD = 34.45), which makes sense because people may not trust AI as much as a human. Thus, the mean change in belief after the debrief is 1.29 points with the median change in belief being 0 again. It is also important to note that the outliers were kept in the analysis because removing them would have decreased an already limited sample size.

	pre_belief	post_belief	belief_change
count	229.000000	229.000000	229.000000
mean	69.672489	61.995633	-7.676856
std	29.203043	33.323178	25.541855
min	0.000000	0.000000	-100.000000
25%	50.000000	38.000000	-11.000000
50%	76.000000	70.000000	0.000000
75%	100.000000	91.000000	1.000000
max	100.000000	100.000000	80.000000
	post_belief	post_belief2	belief_change2
count	229.000000	141.000000	141.000000
mean	61.995633	62.014184	1.290780
std	33.323178	34.450168	10.850635
min	0.000000	0.000000	-24.000000
25%	38.000000	30.000000	0.000000
50%	70.000000	72.000000	0.000000
75%	91.000000	92.000000	1.000000
max	100.000000	100.000000	83.000000

Figure 1: Summary statistics of belief scores

Additionally, qualitative comments were analyzed using the BERT model to identify reasons participants believed their conversation partner was an AI (Figure 2). The most popular reasons included unusually fast response times, very articulate writing, and

```
Cluster 1 representative comment:
too fast of answers.

Cluster 2 representative comment:
His answers were so good!  It seemed similar to a conversation I had with an AI chatbot.  The writing was very good.

Cluster 3 representative comment:
Should have known though by the way the conversation was going and how it sounded
```

Figure 2: Top 3 reasons why respondents suspected conversation was with AI

conversational patterns that resembled that of an AI. These perceptions may have influenced participants' trust or engagement during the conversation and could have affected belief change outcomes. However, given the already limited number of qualitative responses and small percentage of respondents who did not suspect the "human" was AI (Figure 3), these comments were not excluded from the analysis. Instead, the BERT results are presented as exploratory evidence that contextualizes the findings and gives insight into future deception experiments.

```
Percentages of responses to "Did you ever suspect that Dr. Alex Chen was an AI chatbot?"
q71
NaN    73.949580
Yes    20.588235
No      5.462185
Name: count, dtype: float64

Percentages of responses to "Did you ever suspect that Lyndon Carroll was an AI chatbot?"
q73
NaN    73.949580
Yes    21.428571
No      4.621849
Name: count, dtype: float64
```

Figure 3: Percentages of respondents who suspected "human" was AI (NaN = no response)

5.2 Paired T-test for H1

H1: Belief in conspiracy theories will decrease after participants engage in a corrective conversation with an LLM-based chatbot.

To test H1, we conducted a paired t-test comparing pre- and post-conversation belief scores. Results show a statistically significant decrease in conspiracy belief following the corrective conversation since the T-statistic has $p = .007 < 0.01$, meaning we can reject the null hypothesis that the difference between the pre and post belief means is equal to 0 (Figure 4). The mean belief change was −7.73, indicating that, on average, participants reported lower belief in the targeted conspiracy theory after interacting with the chatbot. Minor discrepancies in mean belief change from the paired t-test (−7.73) and difference from the summary statistics (-7.68) are due to differences in the subset of participants included in the paired

```
Paired t-test: t = 2.745, p = 0.007345
Mean belief change (post - pre): -7.727
```

Figure 4: Paired t-test for pre/post belief

analysis (only participants with complete pre and post data were included in the t-test). These results support H1, so this suggests that conversing with an LLM may reduce conspiracy beliefs.

5.3 Regression for H2

H2: The magnitude of belief change will vary across treatment conditions, with participants experiencing a greater belief reduction when the chatbot is perceived as an MIT postdoc than when it is perceived as a political staffer or a generic chatbot.

To test H2, we used an OLS regression that predicted post-intervention belief as a function of treatment identity, conspiracy type, their interaction, and pre-intervention belief. The generic chatbot condition served as the baseline, and all models used heteroskedasticity robust standard errors. This specific regression model was chosen to reduce omitted variable bias by accounting for pre-intervention belief levels and differences across conspiracy topics, since they can both strongly predict post-intervention beliefs.

Overall, results provide limited support for H2 because the MIT postdoc condition and political staffer condition did not produce statistically significant differences in post-conversation belief relative to the generic chatbot (Figure 5). The coefficient for the political staffer treatment was negative ($\beta = -8.58$, $p = 0.058$), which suggested a trend toward greater belief reduction, but this effect did not reach conventional levels of statistical significance. Similarly, the MIT postdoc treatment showed no significant effect ($\beta = -2.39$, $p = 0.526$). In addition, none of the interaction terms between treatment identity and

OLS Regression Results

Dep. Variable:	post_belief	R-squared:	0.500
Model:	OLS	Adj. R-squared:	0.479
Method:	Least Squares	F-statistic:	31.38
Date:	Sun, 14 Dec 2025	Prob (F-statistic):	7.04e-35
Time:	14:55:57	Log-Likelihood:	-1048.1
No. Observations:	229	AIC:	2116.
Df Residuals:	219	BIC:	2150.
Df Model:	9		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	18.5790	5.911	3.143	0.002	6.993	30.165
C(treatment)[T.m]	-2.3870	3.762	-0.635	0.526	-9.760	4.986
C(treatment)[T.p]	-8.5834	4.527	-1.896	0.058	-17.457	0.290
C(conspiracy_type)[T.2020]	-4.6731	4.365	-1.071	0.284	-13.229	3.882
C(conspiracy_type)[T.covid]	-20.5836	5.913	-3.481	0.000	-32.172	-8.995
C(treatment)[T.m]:C(conspiracy_type)[T.2020]	-4.1345	13.058	-0.317	0.752	-29.728	21.459
C(treatment)[T.p]:C(conspiracy_type)[T.2020]	3.2301	12.789	0.253	0.801	-21.836	28.296
C(treatment)[T.m]:C(conspiracy_type)[T.covid]	7.9887	7.938	1.006	0.314	-7.570	23.547
C(treatment)[T.p]:C(conspiracy_type)[T.covid]	13.7819	9.661	1.427	0.154	-5.153	32.717
pre_belief	0.7571	0.060	12.662	0.000	0.640	0.874

Omnibus:	46.278	Durbin-Watson:	1.990
Prob(Omnibus):	0.000	Jarque-Bera (JB):	88.441
Skew:	-1.017	Prob(JB):	6.24e-20
Kurtosis:	5.265	Cond. No.	797.

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Figure 5: Regression predicting post belief using treatment identity, conspiracy type, their interaction, and pre-test belief

conspiracy type were statistically significant. This means that treatment effects did not differ across conspiracy domains either.

The results are similar to that of Boisson et al. (2025), who found limited evidence that perceived identity alone significantly increased corrective effectiveness. One possible explanation is failed deception because many participants correctly inferred that the conversation partner was an AI regardless of the assigned identity, which may have reduced treatment differences. As mentioned before, these participants were not excluded from analysis, which may have caused some inaccurate results.

Overall, perceived expertise cues do not differentiate belief change across conditions. Baseline belief strength and conspiracy topics matter more, so this motivates the idea of examining mediators such as trust to explain when LLM-based interventions are effective.

5.4 Regressions for H3

H3: Trust in the conversation partner will explain the relationship between treatment condition and belief change.

To test H3, we did preliminary analyses using four post-conversation measures. The participant responded to these statements using a 6-point Likert scale:

- convomeasure_1: The conversation provided relevant information I haven't heard before.
- convomeasure_2: The conversation was objective.
- convomeasure_3: My conversation partner was understanding of my perspective.
- convomeasure_4: I trust the content of our conversation.

These measures were moderately to highly correlated (Figure 6), which suggests they capture a common trust-related factor. Due

	convomeasure_1	convomeasure_2	convomeasure_3	convomeasure_4
convomeasure_1	1.000000	0.465145	0.426911	0.499081
convomeasure_2	0.465145	1.000000	0.761719	0.825058
convomeasure_3	0.426911	0.761719	1.000000	0.805471
convomeasure_4	0.499081	0.825058	0.805471	1.000000

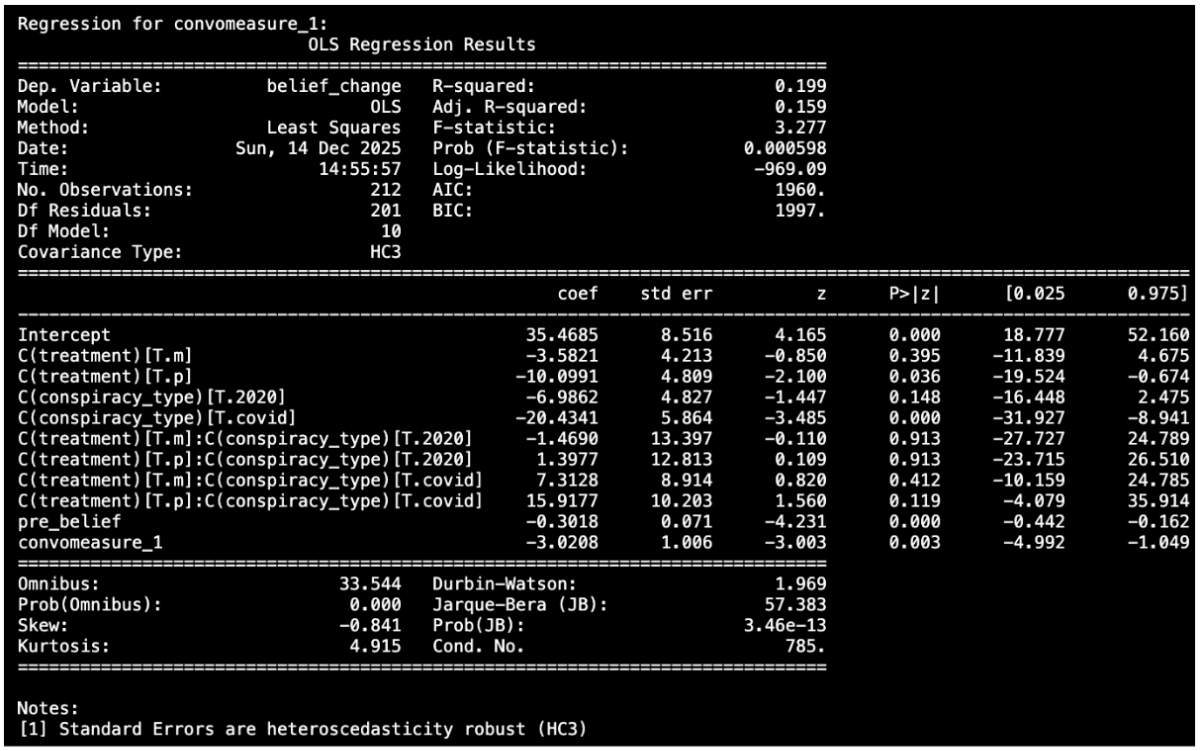
Figure 6: Covariance matrix of post conversation questions

to this collinearity and the limited sample size, we added each measure separately to the baseline model from H2.

Across all models (Figures 7–10), higher scores on these trust and conversation perception measures were consistently associated with greater reductions in conspiracy belief, after controlling for pre-conversation belief, treatment condition, and conspiracy topic. However, similar to results from H2, treatment identity coefficients remained insignificant once these measures were included. This suggests that perceived identity cues alone are insufficient to explain belief change, and that participants' subjective evaluations of the conversation play a more important role.

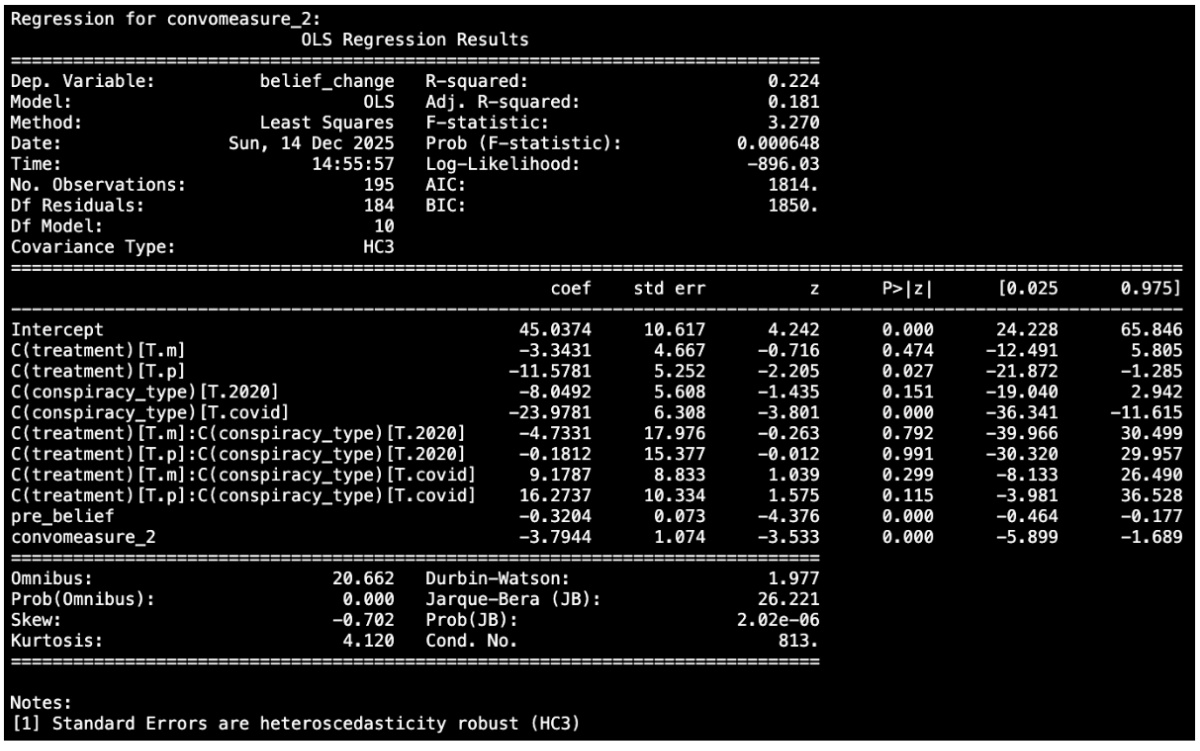
To examine this mechanism, we conducted another regression with convomeasure_4 as the dependent variable and identity as the independent variable. Results show that treatment identity did not significantly affect trust, which suggests that differences in belief change may be affected by how participants experienced the conversation rather than by the assigned identity label itself (Figure 11).

These analyses are exploratory and are not a formal mediation test because trust was measured after treatment and the measures are highly correlated. Nonetheless, the findings provide preliminary evidence that belief change in LLM-based interventions is more closely associated with participants' perceived trust and conversational quality than with the chatbot's assigned identity label.



```
Regression for convomeasure_1:
                            OLS Regression Results
==============================================================================
Dep. Variable:          belief_change   R-squared:                       0.199
Model:                            OLS   Adj. R-squared:                  0.159
Method:                 Least Squares   F-statistic:                     3.277
Date:                Sun, 14 Dec 2025   Prob (F-statistic):           0.000598
Time:                        14:55:57   Log-Likelihood:                -969.09
No. Observations:                 212   AIC:                             1960.
Df Residuals:                     201   BIC:                             1997.
Df Model:                          10
Covariance Type:                  HC3
=============================================================================================================
                                                coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                   35.4685      8.516      4.165      0.000      18.777      52.160
C(treatment)[T.m]                           -3.5821      4.213     -0.850      0.395     -11.839       4.675
C(treatment)[T.p]                          -10.0991      4.809     -2.100      0.036     -19.524      -0.674
C(conspiracy_type)[T.2020]                  -6.9862      4.827     -1.447      0.148     -16.448       2.475
C(conspiracy_type)[T.covid]                -20.4341      5.864     -3.485      0.000     -31.927      -8.941
C(treatment)[T.m]:C(conspiracy_type)[T.2020]  -1.4690     13.397     -0.110      0.913     -27.727      24.789
C(treatment)[T.p]:C(conspiracy_type)[T.2020]   1.3977     12.813      0.109      0.913     -23.715      26.510
C(treatment)[T.m]:C(conspiracy_type)[T.covid]  7.3128      8.914      0.820      0.412     -10.159      24.785
C(treatment)[T.p]:C(conspiracy_type)[T.covid] 15.9177     10.203      1.560      0.119      -4.079      35.914
pre_belief                                  -0.3018      0.071     -4.231      0.000      -0.442      -0.162
convomeasure_1                              -3.0208      1.006     -3.003      0.003      -4.992      -1.049
==============================================================================
Omnibus:                       33.544   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               57.383
Skew:                          -0.841   Prob(JB):                     3.46e-13
Kurtosis:                       4.915   Cond. No.                         785.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

Figure 7: Regression of initial model with convomeasure_1 as a control



```
Regression for convomeasure_2:
                            OLS Regression Results
==============================================================================
Dep. Variable:          belief_change   R-squared:                       0.224
Model:                            OLS   Adj. R-squared:                  0.181
Method:                 Least Squares   F-statistic:                     3.270
Date:                Sun, 14 Dec 2025   Prob (F-statistic):           0.000648
Time:                        14:55:57   Log-Likelihood:                -896.03
No. Observations:                 195   AIC:                             1814.
Df Residuals:                     184   BIC:                             1850.
Df Model:                          10
Covariance Type:                  HC3
=============================================================================================================
                                                coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                   45.0374     10.617      4.242      0.000      24.228      65.846
C(treatment)[T.m]                           -3.3431      4.667     -0.716      0.474     -12.491       5.805
C(treatment)[T.p]                          -11.5781      5.252     -2.205      0.027     -21.872      -1.285
C(conspiracy_type)[T.2020]                  -8.0492      5.608     -1.435      0.151     -19.040       2.942
C(conspiracy_type)[T.covid]                -23.9781      6.308     -3.801      0.000     -36.341     -11.615
C(treatment)[T.m]:C(conspiracy_type)[T.2020]  -4.7331     17.976     -0.263      0.792     -39.966      30.499
C(treatment)[T.p]:C(conspiracy_type)[T.2020]  -0.1812     15.377     -0.012      0.991     -30.320      29.957
C(treatment)[T.m]:C(conspiracy_type)[T.covid]  9.1787      8.833      1.039      0.299      -8.133      26.490
C(treatment)[T.p]:C(conspiracy_type)[T.covid] 16.2737     10.334      1.575      0.115      -3.981      36.528
pre_belief                                  -0.3204      0.073     -4.376      0.000      -0.464      -0.177
convomeasure_2                              -3.7944      1.074     -3.533      0.000      -5.899      -1.689
==============================================================================
Omnibus:                       20.662   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.221
Skew:                          -0.702   Prob(JB):                     2.02e-06
Kurtosis:                       4.120   Cond. No.                         813.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

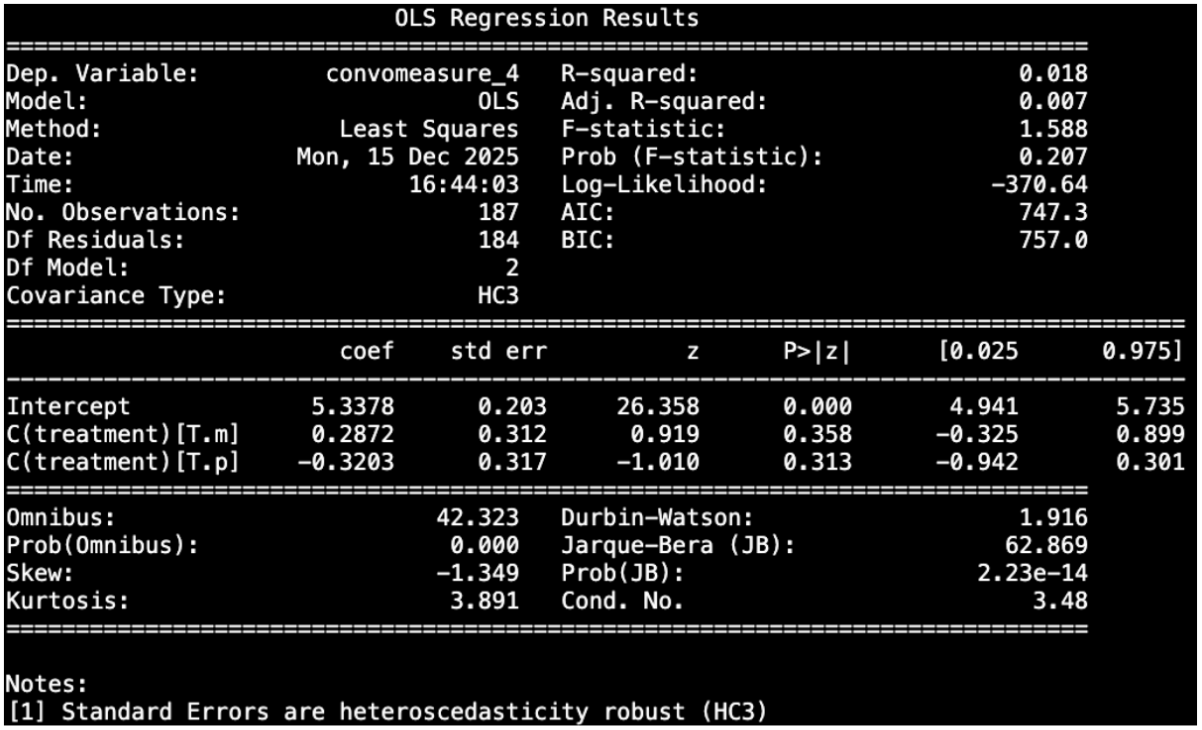
Figure 8: Regression of initial model with convomeasure_2 as a control

```
Regression for convomeasure_3:
                            OLS Regression Results
==============================================================================
Dep. Variable:          belief_change   R-squared:                       0.174
Model:                            OLS   Adj. R-squared:                  0.133
Method:                 Least Squares   F-statistic:                     2.774
Date:                Sun, 14 Dec 2025   Prob (F-statistic):            0.00313
Time:                        14:55:57   Log-Likelihood:                -968.63
No. Observations:                 212   AIC:                             1959.
Df Residuals:                     201   BIC:                             1996.
Df Model:                          10
Covariance Type:                  HC3
================================================================================================================
                                                   coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                       40.2187     12.573      3.199      0.001      15.577      64.861
C(treatment)[T.m]                               -2.3231      4.057     -0.573      0.567     -10.275       5.629
C(treatment)[T.p]                              -10.9470      4.821     -2.271      0.023     -20.396      -1.498
C(conspiracy_type)[T.2020]                      -7.9156      5.042     -1.570      0.116     -17.798       1.967
C(conspiracy_type)[T.covid]                    -21.5597      6.128     -3.518      0.000     -33.571      -9.548
C(treatment)[T.m]:C(conspiracy_type)[T.2020]    -3.8767     12.572     -0.308      0.758     -28.518      20.764
C(treatment)[T.p]:C(conspiracy_type)[T.2020]     3.0254     14.223      0.213      0.832     -24.850      30.901
C(treatment)[T.m]:C(conspiracy_type)[T.covid]    5.1071      8.179      0.624      0.532     -10.924      21.138
C(treatment)[T.p]:C(conspiracy_type)[T.covid]   11.8827     10.156      1.170      0.242      -8.023      31.788
pre_belief                                      -0.2783      0.069     -4.017      0.000      -0.414      -0.143
convomeasure_3                                  -2.9915      1.411     -2.121      0.034      -5.756      -0.227
==============================================================================
Omnibus:                       42.964   Durbin-Watson:                   1.864
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               81.248
Skew:                          -1.010   Prob(JB):                     2.28e-18
Kurtosis:                       5.263   Cond. No.                         812.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

Figure 9: Regression of initial model with convomeasure_3 as a control

```
Regression for convomeasure_4:
                            OLS Regression Results
==============================================================================
Dep. Variable:          belief_change   R-squared:                       0.259
Model:                            OLS   Adj. R-squared:                  0.217
Method:                 Least Squares   F-statistic:                     3.966
Date:                Sun, 14 Dec 2025   Prob (F-statistic):           6.79e-05
Time:                        14:55:57   Log-Likelihood:                -847.76
No. Observations:                 186   AIC:                             1718.
Df Residuals:                     175   BIC:                             1753.
Df Model:                          10
Covariance Type:                  HC3
================================================================================================================
                                                   coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                       47.4257      9.881      4.800      0.000      28.059      66.792
C(treatment)[T.m]                               -3.5247      4.702     -0.750      0.453     -12.740       5.690
C(treatment)[T.p]                               -9.8294      5.094     -1.930      0.054     -19.813       0.154
C(conspiracy_type)[T.2020]                      -8.6056      5.868     -1.466      0.143     -20.107       2.896
C(conspiracy_type)[T.covid]                    -26.5379      6.232     -4.258      0.000     -38.753     -14.323
C(treatment)[T.m]:C(conspiracy_type)[T.2020]    -2.6824     20.723     -0.129      0.897     -43.298      37.933
C(treatment)[T.p]:C(conspiracy_type)[T.2020]     0.8593     12.768      0.067      0.946     -24.165      25.884
C(treatment)[T.m]:C(conspiracy_type)[T.covid]    7.4864      8.411      0.890      0.373      -8.999      23.972
C(treatment)[T.p]:C(conspiracy_type)[T.covid]   15.7736     10.348      1.524      0.127      -4.508      36.055
pre_belief                                      -0.3016      0.072     -4.185      0.000      -0.443      -0.160
convomeasure_4                                  -4.1163      0.920     -4.474      0.000      -5.920      -2.313
==============================================================================
Omnibus:                       24.596   Durbin-Watson:                   1.985
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               36.243
Skew:                          -0.757   Prob(JB):                     1.35e-08
Kurtosis:                       4.545   Cond. No.                         808.
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)
```

Figure 10: Regression of initial model with convomeasure_4 as a control



OLS Regression Results

Dep. Variable:	convomeasure_4	R-squared:	0.018
Model:	OLS	Adj. R-squared:	0.007
Method:	Least Squares	F-statistic:	1.588
Date:	Mon, 15 Dec 2025	Prob (F-statistic):	0.207
Time:	16:44:03	Log-Likelihood:	-370.64
No. Observations:	187	AIC:	747.3
Df Residuals:	184	BIC:	757.0
Df Model:	2		
Covariance Type:	HC3		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	5.3378	0.203	26.358	0.000	4.941	5.735
C(treatment)[T.m]	0.2872	0.312	0.919	0.358	-0.325	0.899
C(treatment)[T.p]	-0.3203	0.317	-1.010	0.313	-0.942	0.301

Omnibus:	42.323	Durbin-Watson:	1.916
Prob(Omnibus):	0.000	Jarque-Bera (JB):	62.869
Skew:	-1.349	Prob(JB):	2.23e-14
Kurtosis:	3.891	Cond. No.	3.48

Notes:
[1] Standard Errors are heteroscedasticity robust (HC3)

Figure 11: Regression predicting trust using identity type as independent variable

6. Conclusion

This study shows that having a corrective conversation with an LLM-based chatbot can reduce conspiracy beliefs on average, even though there is still a lot of heterogeneity. Although belief scores decreased after the intervention, identity cues did not produce significant differences in belief change. Instead, baseline belief strength and conspiracy topics were the strongest predictors of post-intervention beliefs.

Exploratory analyses suggest that trust and perceptions of the conversation are more strongly associated with belief reduction than identity labels alone. This indicates that the effectiveness of LLM-based interventions likely operates through perceived conversational quality and trustworthiness rather than expertise cues.

Several limitations should be noted. First, we did not conduct analyses on belief change after the AI debrief because most participants were not convinced their conversation partner was human. Secondly, there are ethical concerns regarding the deployment of AI in corrective conversations relating to informed consent. Third, LLMs may generate factually

inaccurate or misleading information, which can pose risks in high-stakes contexts. Finally, because relatively few participants believed the AI was human, future studies should explore design features that improve realism and credibility while maintaining ethical safeguards.

Overall, these findings suggest that identity matters less than trust when LLMs are used in this context, and that improving the conversational qualities of AI systems may be key to their democratic potential.

Works Cited

Ardia, David S., and Evan Ringel. “First Amendment Limits on State Laws Targeting Election Misinformation.” *First Amendment Law Review*, vol. 20, 2022, pp. 291, UNC Legal Studies Research Paper No. 4205052, https://ssrn.com/abstract=4205052

Boissin, Esther, et al. “AI Reduces Conspiracy Beliefs Even When Presented as a Human Expert.” OSF Preprint, 29 July 2025, https://doi.org/10.31234/osf.io/apmb5_v2

Kunda, Ziva. “The Case for Motivated Reasoning.” Psychological Bulletin, vol. 108, no. 3, 1990, pp. 480–498, https://doi.org/10.1037/0033-2909.108.3.480

Lu, Louise, *et al.* “How AI Sources Can Increase Openness to Opposing Views.” *Scientific Reports*, vol. 15, no. 1, 17 May 2025, https://doi.org/10.1038/s41598-025-00791-z

Nyhan, Brendan, and Jason Reifler. “When Corrections Fail: The Persistence of Political Misperceptions.” *Political Behavior*, vol. 32, no. 2, 2010, pp. 303–330. https://link.springer.com/article/10.1007/s11109-010-9112-2

Pornpitakpan, Chanthika. “The Persuasive Effect of Circadian Arousal, Endorser Expertise, and Argument Strength in Advertising.” *Journal of Global Marketing*, vol. 17, no. 2/3, 2004, pp. 141–172, https://doi.org/10.1300/J042v17n02_07

Schoenegger, Philipp, et al. “Large Language Models Are More Persuasive Than Incentivized Human Persuaders.” 2024. arXiv, https://arxiv.org/abs/2505.09662

Taber, Charles S., and Milton Lodge. “Motivated Skepticism in the Evaluation of Political Beliefs.” *American Journal of Political Science*, vol. 50, no. 3, 2006, pp. 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

Zanotti, Giacomo. “AI Systems Should Be Trustworthy, Not Trusted.” *AI & Society*, vol. (2025), published 10 Nov. 2025, https://doi.org/10.1007/s00146-025-02728-6

AI usage:

- Finding specific sources for literature review (e.g. “Can you find literature that supports [this idea]?”)
- Help better word certain sentences (e.g. “Improve clarity and coherence: [sentence]”)
- Wrote code for data cleaning/analyses, and debugged code (e.g. “Write me code that runs an OLS regression using post_belief as dependent variable; pre_belief, treatment, and conspiracy_type as independent”)