

## ABSTRACT

Title of Thesis: REDUCING CONSPIRATORIAL BELIEF IN  
2020 ELECTION FRAUD USING CHATGPT

William Jordan Conway, Master of Library and  
Information Science, 2025

Thesis Directed By: Professor Jen Golbeck, Department of  
Information

This study examined whether a single, three-round ChatGPT conversation could weaken belief in a highly politicized conspiracy theory. Twenty-five adults were randomly assigned on a 60/40 split to one of two active conditions delivered through the TruthTalk web platform. In the Conspiracy condition (n = 15), the dialogue respectfully challenged claims of widespread fraud in the 2020 U.S. presidential election; in the comparison condition (n = 10), the same interaction structure invited participants to reconsider their opinion about the best musical genre. Pre- and post-surveys assessed confidence (certainty) in the target belief, rated belief strength, openness to counterevidence, and trust in AI. Descriptive change scores (post – pre) showed medians of zero and narrow interquartile ranges for every outcome in both conditions, with only minor additional dispersion in confidence among conspiracy participants. In short, most people finished the study holding views indistinguishable from those they began with, regardless of topic. These findings reinforce Pierre’s socio-epistemic model and Petty et al.’s attitude-strength insights, indicating that brief factual rebuttals—even when personalized and civil—rarely dislodge beliefs rooted in epistemic mistrust or anchored by high certainty, moral conviction, or partisan identity. The study also exposed methodological hurdles specific to large-language-model interventions: prompt drift, unsupported claims, and opaque system behavior made it difficult to ensure a uniform treatment and to earn participant trust. Future research should test multi-session, transparently sourced dialogues that directly address the moral, identity, and certainty foundations of strong attitudes before expecting meaningful belief change.

REDUCING CONSPIRATORIAL BELIEF IN 2020 ELECTION FRAUD USING  
CHATGPT

by

William Jordan Conway

Thesis submitted to the Faculty of the Graduate School of the  
University of Maryland, College Park, in partial fulfillment  
of the requirements for the degree of  
Master of Library and  
Information Science  
2025

Advisory Committee:

Professor Jen Golbeck, Chair

Lecturer Alex Leitch

Assistant Professor Cody Buntain

© Copyright by  
William Jordan Conway  
2025

## Dedication

For my wife, Harriet. Thank you for filling my days with love, laughter, and inspiration.

## Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Jen Golbeck, for her mentorship and academic guidance throughout my graduate journey. Her dedication, expertise, and support have profoundly shaped both this paper and my development as a scholar.

A heartfelt thanks to my oldest friend, Assistant Professor Katherine Kafonek, whose innovative research and consistent encouragement have been a source of motivation and creativity. Her belief in my abilities has fueled my perseverance.

I would like to extend my gratitude to my former supervisor, Kristen Byers, whose mentorship was instrumental in shaping my professional growth. Her guidance transformed my writing, a skill I will carry forward in all my endeavors.

I am also deeply thankful to my wife, whose steadfast support and belief in me have been the foundation of my academic pursuits. Their patience, love, and understanding sustained me through the challenges of this endeavor.

Thank you to Alex Leitch and Cody Buntain for serving on my committee and for the guidance you offered on my statistical reasoning and theoretical frameworks.

Lastly, I appreciate the groundbreaking research by Costello et al., which inspired this project. Their work provided a critical framework for exploring AI-driven dialogues to address conspiratorial beliefs.

Thank you all for your invaluable contributions to this work.

# Table of Contents

Dedication .....	ii
Acknowledgements .....	iii
Table of Contents .....	iv
List of Tables .....	vi
List of Figures .....	vii
Chapter 1: Introduction .....	1
A Riot in the Capital .....	1
Research Questions and Hypotheses .....	1
Definitions of Key Variables .....	2
Chapter 2: Literature Review .....	4
Defining Conspiracy Theories .....	4
Pierre’s Two-Component Epistemic Model .....	4
Attitude Strength Across the Commitment Continuum .....	5
Why Most Debunking Efforts Show Only Modest Effects .....	6
Dialogues with Large Language Models: Early Evidence .....	6
Limits of Fact-Based Corrections and Ethical Concerns for AI .....	7
Election Fraud Narratives and Salience .....	8
Chapter 3: Methods .....	9
Study Design .....	9
Study Context and Objectives .....	9
Randomization and Tailored AI Behavior .....	9
Participants .....	10
Recruitment and Eligibility .....	10
Demographics .....	10
The Application .....	10
Pre-Study Preparations .....	12
General Procedure .....	12
Post-Test Measures .....	13
Chapter 4: Results .....	14
Overview of Descriptive and Inferential Results .....	14
Hypothesis 1: Confidence and Belief Strength Change .....	14
Hypothesis 2: Openness to Belief Change .....	16
Hypothesis 3: Trust in AI .....	17
Results Summary .....	19
Chapter 5: Discussion .....	20
Confidence and Belief Strength: Confronting Politicized Narratives .....	20
Openness to Belief Change: Limited Change Concentrated Among a Few .....	21
Trust in AI: Evaluating the Effects of Politicized vs. Neutral Conversations .....	22
Integrative Reflection .....	23
Chapter 6: Limitations and Future Directions .....	25

Participant Sampling .....	25
AI Variability and Opacity.....	25
Ethical and Trust Considerations .....	26
Chapter 7: Conclusion.....	27
Appendices.....	28
Bibliography .....	48

## List of Tables

Table 1: Confidence Ratings vs. Belief Strength – Comparison

Table 2: Confidence Ratings vs. Belief Strength – Conspiracy

## List of Figures

Figure 1: The TruthTalk application interface

Figure 2: Distribution of Pre- and Post-intervention Scores for Confidence and Belief Strength

Figure 3: Openness to Belief Change: Pre-Post Comparison ( Conspiracy)

Figure 4: Openness to Belief Change: Pre-Post Comparison (Comparison)

Figure 5: Trust in AI: Pre-Post Comparison ( Conspiracy)

Figure 6: Trust in AI: Pre-Post Comparison (Comparison)

# Chapter 1: Introduction

## *A Riot in the Capital*

On January 6, 2021, thousands of Donald Trump supporters stormed the U.S. Capitol in Washington, D.C., in an unprecedented attempt to overturn the election of a presidential opponent. Fueled by a pervasive belief that widespread voter fraud had occurred, and that the election was “stolen” from Trump, these demonstrators—ignited by incendiary rhetoric and disinformation—engaged in violent clashes with law enforcement and acts of vandalism aimed at disrupting the certification of electoral results (ABC News, 2021; BBC, 2023). At the conclusion of his speech that day, Trump famously exhorted, “If you don’t fight like hell, you’re not going to have a country anymore” (Naylor, 2021). This dramatic episode serves as an illustrative anecdote of how deeply held conspiratorial beliefs, when interwoven with mis- and disinformation campaigns and political mobilization, can lead to significant collective action with far-reaching consequences.

While conspiracy narratives span the ideological spectrum, the “stolen election” story has become a prominent fixture in recent public discourse. Therefore, it offers a clear opportunity to examine how conversational AI might address strongly held, identity-relevant misconceptions. Earlier work suggests that large-language-model chatbots can reduce conspiratorial views on a range of conspiratorial topics by 20% (e.g., Costello, Pennycook, & Rand, 2024). Building on that insight, the present project asks whether brief dialogues with ChatGPT can reduce belief in a politically charged conspiracy claim, without claiming to replicate or directly test any single prior study.

The inquiry is situated against a backdrop of rising epistemic mistrust. Americans increasingly evaluate information through a partisan lens, privileging sources that align with their identities and discounting those labeled “mainstream” (Gottfried, 2021). As of October 2024, an NPR/PBS News/Marist poll found that 87% of Trump supporters were concerned about voter fraud in the upcoming election, compared with 58% of adults overall—evidence that the “stolen election” narrative remains politically salient rather than a resolved dispute (Parks, 2024). Such dynamics create space for disinformation to flourish and lend emotional resonance to conspiracy theories. Rather than emphasizing statistical significance alone, this exploratory study aims to gauge the practical potential of AI-driven conversation to soften entrenched beliefs, map individual response patterns, and generate design guidance for larger, more definitive trials.

## *Research Questions and Hypotheses*

This study investigates the extent to which a brief interaction with an AI-driven chatbot, ChatGPT, influences participants’ confidence in and strength of their beliefs, openness to counterarguments, and trust in AI. Specifically, the following research questions are examined:

1. Confidence and Strength of Belief: Does a brief persuasive dialogue with ChatGPT reduce participants' confidence in-and strength of- their prior belief when the topic is a politicized conspiracy (2020 election fraud) compared with a personal preference (best music genre)?
2. Openness to Belief Change: Does the conspiracy-topic dialogue increase openness to counterevidence more (or less) than the music-topic dialogue?
3. Trust in AI: Does topic framing (conspiracy vs. music) influence participants' trust in AI as an information source?

From these research questions, the following hypotheses are tested:

1. H1 (Alternative Hypothesis): Participants in the conspiracy-persuasion group will show a larger decrease in confidence and strength of belief ratings than those in the comparison group.
  - a. H0 (Null Hypothesis): There will be no significant difference in confidence and strength of belief ratings between the groups.
2. H2 (Alternative Hypothesis): Participants in the conspiracy-persuasion group will exhibit greater openness to belief change than the comparison group.
  - a. H0 (Null Hypothesis): There will be no difference between groups in openness to belief change.
3. H3 (Alternative Hypothesis): Participants in the conspiracy-persuasion group will report lower trust in AI after the dialogue than those in the comparison group.
  - a. H0 (Null Hypothesis): There will be no significant difference in trust in AI between the groups.

### Definitions of Key Variables

This section provides operational definitions for the key variables examined in this study, distinguishing between related but conceptually distinct constructs. These variables measure participants' beliefs, attitudes, and perceptions both before and after the intervention.

#### Belief Certainty and Strength

- Pre-Confidence / Post-Confidence: These variables measure the degree of certainty a participant has regarding the accuracy of their belief. Confidence reflects how sure an individual is that their perspective is correct, regardless of how deeply they hold that belief. A participant with high confidence may strongly assert that their belief is factually correct, but this belief may not necessarily be central to their identity or decision-making.

- Pre-Belief Strength / Post-Belief Strength: These variables measure the intensity and personal significance of a belief. Unlike confidence, which focuses on perceived correctness, belief strength reflects how deeply ingrained a belief is, how much it shapes attitudes and behaviors, and how resistant it is to change. A participant with strong belief strength is likely to hold onto their belief even in the face of counterevidence and may see it as a core part of their worldview.

#### Specific Belief-Related Variables

- Pre-Election Fraud / Post-Election Fraud (Conspiracy-persuasion Condition Only): These variables assess the extent to which participants believe that widespread voter fraud occurred in the 2020 U.S. presidential election. It measures whether participants perceive the election outcome as legitimate or fraudulent and tracks any shifts in this perception following the intervention.

- Pre-Genre / Post-Genre (Comparison Group Only): These variables capture participants stated preferences for a specific music genre before and after the intervention. Unlike election fraud beliefs, which are politically charged, this variable serves as a neutral comparison to examine belief stability in a non-controversial domain.

#### Trust in AI

- Pre-Trust AI / Post-Trust AI: These variables measure participants' level of trust in artificial intelligence as a source of information. It captures whether individuals perceive AI as a credible, reliable, and unbiased tool for providing factual knowledge.

#### Openness to Belief Change

- Pre-Belief Change / Post-Belief Change: These variables assess the participants' willingness to reconsider or modify their beliefs based on credible, counter evidence. Unlike confidence or belief strength, which measure certainty and intensity, this variable reflects flexibility in thinking and openness to new information. Higher scores indicate greater receptivity to alternative perspectives, whereas lower scores suggest cognitive rigidity or resistance to change.

#### Perceived Usefulness of AI

- Useful AI: These variables measure how helpful participants found the AI-driven dialogue. It captures subjective assessments of whether the AI provided meaningful insights, effectively challenged beliefs, or reinforced existing views. Unlike trust in AI, which evaluates AI's credibility, usefulness reflects the perceived effectiveness of the AI as an interactive tool for engaging with participants' beliefs.

## Chapter 2: Literature Review

### Defining Conspiracy Theories

Conspiracy theories are conceptualized as explanatory frameworks that attribute significant events to secret plots orchestrated by powerful, malevolent actors (Douglas & Sutton, 2023). They are characterized by an oppositional stance to mainstream narratives and a remarkable resistance to refutation. Pierre (2020) further argues that conspiracy beliefs arise from a deep-seated epistemic mistrust—where individuals reject conventional sources of knowledge and instead gravitate toward alternative narratives that seem to provide a more coherent explanation for complex events. In this light, conspiracy theories are not simply irrational outbursts but may serve as adaptive responses to perceived gaps in trustworthy information.

Although the label conspiracy theory is sometimes wielded pejoratively, research emphasizes that belief in hidden plots is not inherently irrational; rather, it reflects an alternative epistemic strategy premised on deep mistrust of authorities (Pierre, 2020). From this perspective, conspiracy thinking is best conceptualized as a motivated belief system serving needs for understanding, control, and group rather than as a simple deficit of information (Robertson et al., 2022). Recognizing that motivational backdrop is crucial because interventions aimed solely at injecting factual corrections often fail to engage the relational and identity-based functions that conspiracy theories fulfil.

### Pierre's Two-Component Epistemic Model

A central framework for understanding the persistence of conspiracy beliefs is Pierre's (2020) two-component socio-epistemic model, which conceptualizes conspiracy thinking as rooted in two interlocking processes: epistemic mistrust and motivated reasoning. Epistemic mistrust refers to a deep-seated skepticism toward traditional sources of authority and expertise, often stemming from real or perceived betrayals that lead individuals to question the legitimacy and intentions of institutions. This mistrust undermines the credibility of mainstream information and drives individuals to seek alternative explanations, particularly those outside of established knowledge systems. The second component—misinformation processing—involves selectively attending to information that serves key psychological needs, such as understanding complex or threatening events, maintaining a sense of personal control, or affirming group belonging.

Pierre emphasizes that these beliefs do not arise from irrationality or ignorance but from attempts to re-establish epistemic and social stability in contexts where institutional trust has eroded. In this view, conspiracy theories act as coping mechanisms: they offer a framework that can resolve uncertainty, preserve agency, and construct moral clarity in an otherwise ambiguous social world. This helps explain why corrective information alone is often insufficient to change minds—because conspiracy beliefs are embedded not just in faulty facts but in deeper relational and existential needs. As Pierre (2020) notes, addressing conspiratorial thinking therefore

requires more than just debunking; it necessitates rebuilding trust and attending to the psychological functions these beliefs fulfill.

### *Attitude Strength Across the Commitment Continuum*

To understand why some beliefs remain so resistant even when presented with counterevidence, it is helpful to consider research on attitude strength, which identifies the qualities that make certain beliefs particularly durable. Petty, Siev, and Briñol (2023) reconceptualize attitude strength as a syndrome of interrelated qualities—importance, certainty, identity centrality, moral conviction, and knowledge—that jointly determine whether an attitude endures and guides behavior. Each quality contributes unique predictive power: high identity centrality links attitudes to social belonging, moral conviction imbues them with normative force, certainty generates subjective undeniability, importance heightens personal stake, and knowledge offers perceived evidentiary ballast. Meta-analytic findings reviewed in their article show that these properties independently forecast resistance to persuasion and stability over time, underscoring that commitment is multifaceted rather than a single continuum score. The same review refines the long-standing view that weak attitudes are readily moved whereas strong attitudes are not. Although low-strength positions generally shift more easily, evidence indicates that persuasive appeals matched to the basis of a strong attitude can produce measurable, though modest, change. Commitment therefore sets a higher threshold for influence rather than rendering strong attitudes wholly immovable.

Within this framework, music-genre preference exemplifies a low-strength attitude; it typically ranks low on moral conviction and identity centrality, making it more malleable. Moderate-commitment conspiracy suspicions occupy a middle position, reflecting some certainty and selective identity relevance. At the high-strength extreme lie deeply entrenched conspiratorial beliefs (e.g. 2020 election fraud), which score high on moral conviction, partisan identity, and perceived evidentiary support. Positioning these topics along a single commitment continuum clarifies the theoretical expectation for the current investigation. If a uniform three-round ChatGPT dialogue produces large shifts in musical preferences but minimal movement in election-fraud beliefs, the pattern would illustrate how commitment level, not message format, constrains persuasion. Conversely, any detectable change in the high-commitment condition would exemplify the nuanced possibility that carefully tailored interactions could penetrate even entrenched attitudes. Contemporary attitude-strength scholarship thus supplies the conceptual backbone for interpreting topic-dependent differences in chatbot effectiveness. Mapping the three topics onto this commitment continuum clarifies why a uniform ChatGPT script might succeed with musical tastes yet falter with election-fraud convictions. The following section reviews Costello, Pennycook, and Rand's AI-dialogue studies, which occupy the middle of that continuum and therefore offer a logical benchmark for the present investigation.

### Why Most Debunking Efforts Show Only Modest Effects

Given the individual and societal harms associated with conspiracy beliefs, researchers have developed a range of interventions aimed at reducing their influence, though most show only modest success. A meta-analysis by Stasielowicz (2022) reported a small average effect size ( $g = 0.162$ ), and a systematic review by O’Mahony et al. (2023) found fewer than half of reviewed interventions produced statistically significant reductions in conspiracy belief. Notably, O’Mahony et al. identified three more promising approaches: inoculation strategies that preempt misinformation, priming interventions to activate analytical thinking or agency, and long-form educational efforts such as semester-long courses in scientific reasoning. The latter achieved the strongest effect ( $d = 1.07$ ), though such interventions are labor-intensive and difficult to scale. By contrast, the present study examined a potentially more scalable method: real-time, AI-mediated dialogue offering tailored counterevidence. However, because these interactions occur after beliefs have already formed their impact may remain limited without deeper, sustained engagement—potentially limiting their ability to disrupt belief systems that serve relational or identity-reinforcing functions.

One reason for this limited effectiveness may lie in the social identity functions and psychological motives that underpin conspiracy beliefs. As Robertson et al. (2022) explain, conspiracy theories often fulfill identity-related needs by helping individuals maintain a favorable image of their in-group. When a group’s status or reputation is threatened, members may gravitate toward conspiracy narratives that shift blame to an out-group perceived as hostile, thereby protecting their group’s esteem. These beliefs are particularly resilient because they serve relational and motivational functions and are often unfalsifiable. Gligorić et al. (2021) similarly argue that conspiracy belief is “multiply determined” by a constellation of psychological motives—epistemic, existential, and social. Epistemic motives include the need to find meaning in randomness (illusory pattern perception) and to resolve uncertainty (need for closure). Existential motives reflect the desire for control, with conspiracy theories functioning as compensatory mechanisms in response to feelings of powerlessness. Social motives center on self-image and distinctiveness, including the desire for uniqueness and narcissistic self-regard. These motives explain why some conspiracy narratives are appealing even when implausible: they fulfill personal and group-related psychological needs. As Gligorić et al. note, “people may be drawn to conspiratorial beliefs as they seem to satisfy important psychological motives” (p. 1172). This motivational complexity helps explain why logic-based refutations alone are often insufficient; the issue is not only what people believe, but also why they believe it. However, the limited success of most debunking attempts may stem not only from how corrections are delivered but also from what kind of attitudes they confront—some beliefs are simply more durable than others.

### Dialogues with Large Language Models: Early Evidence

Costello, Pennycook, and Rand (2024) provide the most comprehensive test to date of AI-mediated debunking. In three preregistered studies, a brief dialogue with GPT-4 reduced endorsement of a broad conspiracy scale by about 20% relative to control conversations, with effects persisting at a two-month follow-up. Importantly, a follow-up experiment showed that

the same rebuttal delivered by a “human expert” avatar produced an equal reduction (Boissin et al., 2025), suggesting that the persuasive force lies in the personalized, respectful content rather than technological novelty per se.

These results are encouraging but highlight the need to extend this research into more polarized and identity-relevant contexts. While Costello et al.’s (2024) work demonstrated that brief AI-mediated dialogues can reduce generalized conspiracist thinking specifically, such beliefs often lack strong partisan attachments. By contrast, election-fraud narratives are deeply entangled with political identity in the United States, increasing the likelihood that counter-messages will be perceived as partisan attacks rather than good-faith correction. Our study builds on Costello’s foundational work by testing the limits of this approach in a domain where resistance to persuasion is likely to be higher. In doing so, we provide a more stringent test of AI’s utility in contentious real-world scenarios. At the same time, our findings point to the broader ethical and practical challenges of deploying large language models as persuasive agents—especially their unpredictability, potential for bias, and the difficulty of ensuring consistent message quality across contexts.

### *Limits of Fact-Based Corrections and Ethical Concerns for AI*

The mixed record of traditional fact checks in politicized domains underscores the need for strategies that address motivation and identity, not just information deficits (O’Mahony et al., 2023). Recent work demonstrates why: biased AI writing assistants can subtly nudge users’ attitudes even when people believe they are merely “getting help” with wording; in two preregistered experiments, exposure to skewed autocomplete shifted opinions toward the model’s stance, and participants remained largely unaware of the bias despite warnings (Williams-Ceci et al., 2025). Unlike a human interviewer, researchers cannot truly “lock down” a commercial LLM’s outputs because even the smallest prompt changes can yield qualitatively different replies. This unpredictability means we cannot fully control or audit the model’s persuasive direction in real time, creating risks for internal validity (Did everyone receive the same intervention?) and for ethics (Did the system covertly push a viewpoint the researcher never intended?).

The credibility and replicability of AI interventions is further undermined by structural opacity. Commercial systems routinely conceal training tweaks, fine-tuning procedures, and sampling settings. This prevents independent verification of what the model can access or whether a study can be reproduced under identical conditions (Sapkota et al., 2025). Behavioral opacity compounds the problem: prompt-level tweaks or backend updates can change outputs enough that two “identical” studies may expose participants to different arguments. When influence is both unseen and untraceable, participant trust—already fragile among conspiracy-minded individuals—erodes further if the model hallucinates or mis-cites evidence. Efficacy and ethics are intertwined: if persuasive power operates covertly and cannot be precisely constrained, then transparency about prompt design, embedded perspectives, and sourcing becomes a core requirement of responsible research, not an optional add-on. Any intervention that relies on LLMs must therefore foreground verifiable citations, log and disclose prompt/response parameters, and build mechanisms to detect and correct bias to sustain both scientific rigor and

user trust. Absent cooperation from LLM creators, such transparency is out of reach for researchers. These concerns are especially pressing in the context of election fraud narratives, where public trust is fragile, partisan polarization is extreme, and the consequences of misinformation extend beyond individual belief into institutional policy change.

### *Election Fraud Narratives and Saliency*

Although often labeled merely as conspiracy “theories,” false claims of widespread voter fraud have transcended fringe debates, significantly influencing real-world election policies under the guise of election integrity. Prominent misinformation regarding the 2020 election, amplified by influential political figures such as former President Trump and supportive media outlets, has directly led to concrete legislative and executive actions. These include stricter voter-ID laws and reduced mail-in voting opportunities. Highlighting the practical urgency of addressing conspiracy beliefs, the White House’s March 2025 Presidential Action titled “Preserving and Protecting the Integrity of American Elections” explicitly identifies pervasive election fraud as a critical threat, calling for federal interventions through public education, inter-agency coordination, and tech partnerships (Trump, 2025).

Given this backdrop, understanding conspiracy theories as expressions of deeper psychological, social, and identity-related needs—rooted in epistemic mistrust and motivated reasoning—is particularly crucial. Traditional fact-checking and debunking approaches that solely aim to correct misinformation have proven insufficient precisely because they neglect these underlying motivational and identity-driven dimensions. Recent evidence suggests that personalized, respectful, and interactive interventions, such as AI-mediated dialogues, hold promise in addressing conspiracy beliefs. However, deploying these tools in highly polarized and politically charged contexts, like election-fraud narratives, poses significant ethical and practical challenges. Issues of transparency, structural unpredictability, and potential biases inherent in large language models necessitate interventions designed explicitly to restore trust, mitigate bias, and transparently address identity-based motivations. Thus, the present study evaluates whether interactive AI dialogues can effectively confront election-fraud narratives—a high-commitment belief domain—while documenting any ethical or methodological issues that arise during the administration of the experiment.

## Chapter 3: Methods

### Study Design

This study employed a randomized comparison design. Participants were randomly assigned, in a 60/40 split, to one of two equally active dialogue conditions delivered through the TruthTalk web platform. In the Conspiracy condition, a three-round ChatGPT conversation explicitly challenged election-fraud claims. In the Comparison condition, a parallel three-round conversation attempted to nudge participants toward reconsidering their opinion concerning the best genre of music. Both conditions ran on separate instances of the TruthTalk application, each powered by OpenAI's Assistants API and configured with condition-specific prompts, so that dialogue length, tone, and interactive structure were held constant while only the topic varied.

### Study Context and Objectives

Our primary goal was to assess whether it was feasible to leverage AI-driven dialogues to alter deeply held conspiratorial beliefs. Feasibility considerations included both the technical implementation (i.e., whether the AI could reliably follow a given set of scripted instructions) and the effectiveness of the intervention (i.e., whether any measurable shifts in participants' beliefs, openness to change, or trust in AI occurred). By foregrounding feasibility, we sought insights into how participant engagement, dialogue quality, and overall user experience might inform subsequent, larger-scale investigations. Given the politically sensitive nature of the conspiracy topic, we monitored closely to ensure that the AI responses were polite, respectful, and aligned with ethical expectations of human research.

### Randomization and Tailored AI Behavior

A Python script using `random.shuffle` allocated participants to one of these two arms, ensuring that observed differences in outcomes could be attributed to the intervention rather than baseline group disparities. We emphasized a ratio of approximately 60% to the Conspiracy condition and 40% to the comparison condition due to the interest in maximizing data on the election-fraud conversations while still having a neutral comparison group.

Each AI assistant was programmed with condition-specific instructions:

- Conspiracy Condition:
  - The AI systematically challenged conspiratorial views by presenting counterevidence on election fraud, prompting participants to reflect on inconsistencies in their reasoning, and encouraging them to consider alternative explanations. The dialogues were not adversarial; instead, the assistant used open-ended questions and clarifications to foster critical thinking.
- Comparison Group:
  - The assistant steered the conversation solely toward participants' favorite music genres. It asked them to elaborate on their preferences, occasionally adopting a

“devil’s advocate” stance (e.g., suggesting a different genre) but never referencing political or conspiratorial topics. This allowed for a baseline measure of changes in attitudes about a benign subject and about the AI itself.

## Participants

### Recruitment and Eligibility

Participants were recruited via Prolific and were required to be 18 or older, U.S. residents, English speakers, and have supported Donald Trump in 2020. They had to demonstrate a minimal confidence in claims of 2020 election fraud. Specifically, we screened for individuals who indicated they believed minor to widespread fraud had occurred that affected the results of the 2020 election and possessed enough digital literacy to navigate the TruthTalk interface. Participants received \$2 as compensation for their time on the eligibility survey. In accordance with Prolific’s and university guidelines, participants were informed of how and when they would be compensated. Data privacy regulations were respected by ensuring that no personally identifying information was used.

### Demographics

The participants in this study had a mean age of 50 years. The youngest participant was 32 years old, while the oldest was 79 years old. The sample was composed of 64% female participants and 36% male participants. Most participants identified as White (96%), while 4% of participants identified as belonging to other ethnicities.

Participants reported a range of employment statuses. The two largest proportions of participants (36.4%) were employed full-time or in non-paid work (36.4%). Additionally, 18.1% of participants reported part-time employment, and 9.1% identified as unemployed. A small subset (12%) had employment status data that was expired or unavailable.

## The Application

TruthTalk was developed as a custom AI-powered study platform utilizing OpenAI’s beta version of Assistants and ChatGPT 4o, deployed through Google Cloud Console. The purpose of the application was to create a natural AI environment the participants might come into contact with. Unlike ChatGPT’s general model, TruthTalk was programmed with structured conversation protocols designed to guide participants through a three-stage dialogue aimed at encouraging reflection and reconsideration of their beliefs.

This setup allowed the researcher to create a tailored conversational agent that adhered to predefined instructions while maintaining the flexibility to engage in organic discussion. The assistant was given strict behavioral guidelines, ensuring that it followed the study’s ethical and procedural requirements, including obtaining informed consent before beginning any tasks. Two separate Assistants were created and deployed according to each condition on two different

versions of the application. The AI's responses were governed by detailed instructions that structured the dialogue into a controlled experiment, differentiating it from an open-ended chatbot like ChatGPT.

As demonstrated in **Figure 1**, the application was designed with a clean and minimalist user interface to enhance participant focus and engagement. The chat interface was styled with a simple, professional aesthetic, ensuring ease of use. An expandable container with instructions for how to interact with the AI was provided as well. The system automatically logged each session, and conversation threads were saved within OpenAI's dashboard for review. This allowed the researcher to monitor progress and ensure compliance with study protocols. Each thread was tagged with the participant's anonymized username metadata to facilitate participant identification while preserving confidentiality.

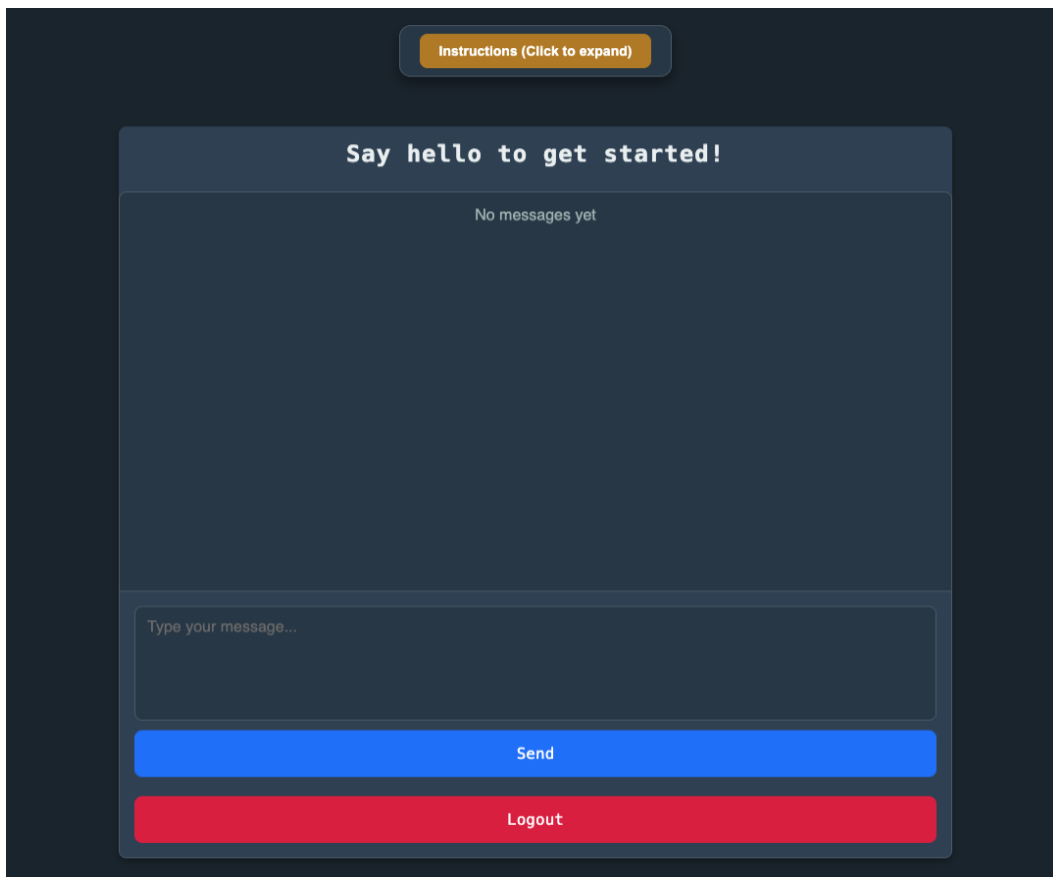


Figure 1: The interface of the TruthTalk custom web application.

A crucial aspect of the study's methodology was the automation of participant interactions. TruthTalk functioned independently, managing participant consent, conducting structured conversations, and administering pre- and post-test surveys. The researcher's role was largely supervisory—periodically checking the OpenAI dashboard to verify that new threads had been

initiated and ensuring that participants completed the study process. This automation allowed for efficient data collection without the need for manual intervention during the experiment.

One of the key challenges in developing the AI assistant was programming its responses to remain neutral yet persuasive in the Conspiracy condition. The AI needed to encourage participants to reflect on their beliefs about election fraud without directly challenging or invalidating them, as direct confrontation often leads to belief reinforcement rather than reconsideration. Instead, the AI used evidence-based nudging techniques to facilitate introspection. These same techniques were applied to the neutral conversation the AI conducted with the Comparison group.

By leveraging OpenAI's Assistants API, TruthTalk provided a more controlled environment for conducting this experimental study. The integration of saved conversation threads, metadata tagging for participant tracking, and full automation of the experimental process made TruthTalk a powerful tool for measuring AI's potential influence on belief revision.

### Pre-Study Preparations

Before interacting with the AI, participants were asked to read and accept an online informed consent form administered by the AI assistant. At this stage, participants were reminded to choose a quiet environment to reduce distractions. Additionally, a short open-ended question was posed to assess response coherence, and only those who provided coherent answers (scoring 6 or higher on a 10-point scale) proceeded. This confirmed both English proficiency and the likelihood of providing thoughtful data.

### General Procedure

All participants reviewed an extensive consent form that detailed the research purpose, procedures, potential risks, and data management strategies. Consent to proceed was captured within the AI chat, ensuring clarity on the voluntary nature of the study and the participant's right to withdraw at any time. Identifiable information was kept separate from the study dataset; numeric IDs from Prolific served as the only linking mechanism to usernames and subsequent assigned ID numbers, used to refer to participants throughout the study.

Following randomization, participants in each condition answered a pre-test questionnaire administered by the AI assistant. Both versions measured three main constructs:

1. Belief Strength/Confidence (e.g., rating how strongly they believed the 2020 election was stolen or how strongly they felt about the "best" music genre),
2. Openness to Belief Change (likelihood of reconsidering their stance if presented with credible evidence), and
3. Trust in AI (their baseline level of trust in the AI's factual accuracy).

In the Conspiracy condition, participants additionally specified if they believed widespread fraud occurred (Yes/No/Unsure), rated how strongly they felt about that stance (1–5), and selected

which sources (e.g., News Media, social media) influenced their views. In the comparison condition, participants identified their favorite genre from a list and indicated the influences shaping their preference (friends, media, streaming platforms, etc.).

Human-AI Dialogues were then conducted:

1. Conspiracy-persuasion Condition
  - a. In three consecutive rounds, the AI referenced participants' pre-test responses about election fraud, asked clarifying questions, and provided counterpoints intended to challenge their conspiratorial thinking. While highlighting reputable information, the AI strove to remain courteous and non-judgmental. Each participant's input guided subsequent AI prompts, allowing for a somewhat personalized dialogue about evidence, trust, and alternative possibilities.
2. Comparison Group
  - a. Participants engaged in a comparable three-round conversation about the "best" music genre. The AI sought further details on their tastes, introduced alternative genres, and gently asked them to explain why their choice was superior. At no point did the AI introduce political topics, thus serving as a neutral comparator for potential changes in trust in AI or willingness to shift a personal preference.

Post-Test Measures

Immediately after the dialogue, participants in both arms completed a post-test questionnaire mirroring the pre-test. Conspiracy-persuasion condition participants again rated their belief in election fraud (Yes/No/Unsure), confidence, strength of that belief, openness to changing it, and trust in AI. Comparison participants likewise re-evaluated their music preference, openness to alternatives, and trust in AI. Both groups also indicated "How useful did you find the AI dialogue in challenging or refining your beliefs?" (1–5) and provided open-ended feedback. These matched pre- and post-test measures allowed us to gauge any shifts in belief strength/confidence, openness to reconsideration, and trust in AI—a direct test of our feasibility aims.

By comparing pre- and post-dialogue responses within and across these conditions, the study assessed whether a short, structured AI exchange could encourage participants to reflect more critically on conspiratorial claims (or neutral preferences) and revise their trust in AI.

## Chapter 4: Results

### *Overview of Descriptive and Inferential Results*

Twenty-five participants completed the study ( $n = 15$  Conspiracy,  $n = 10$  Comparison; 21 in Trial 1 and 4 in Trial 2). For each primary outcome—confidence, belief strength, openness to counterevidence, and trust in AI—individual change scores were computed as  $\Delta = \text{Post} - \text{Pre}$ . Across all variables and conditions, the median change was zero, and three of the four interquartile ranges (IQRs) were also zero, indicating that at least 50% of participants exhibited no change. The only meaningful dispersion emerged in the conspiracy-condition group's confidence scores (IQR = 6; SD = 14.50), which reflected a small number of large shifts (e.g.,  $-40$  and  $+20$  points) amidst a tightly clustered majority. **Figure 2** (Panels A–D) displays the pre/post distributions for each outcome, while **Tables 1** and **2** summarize individual-level change statistics between confidence and belief strength.

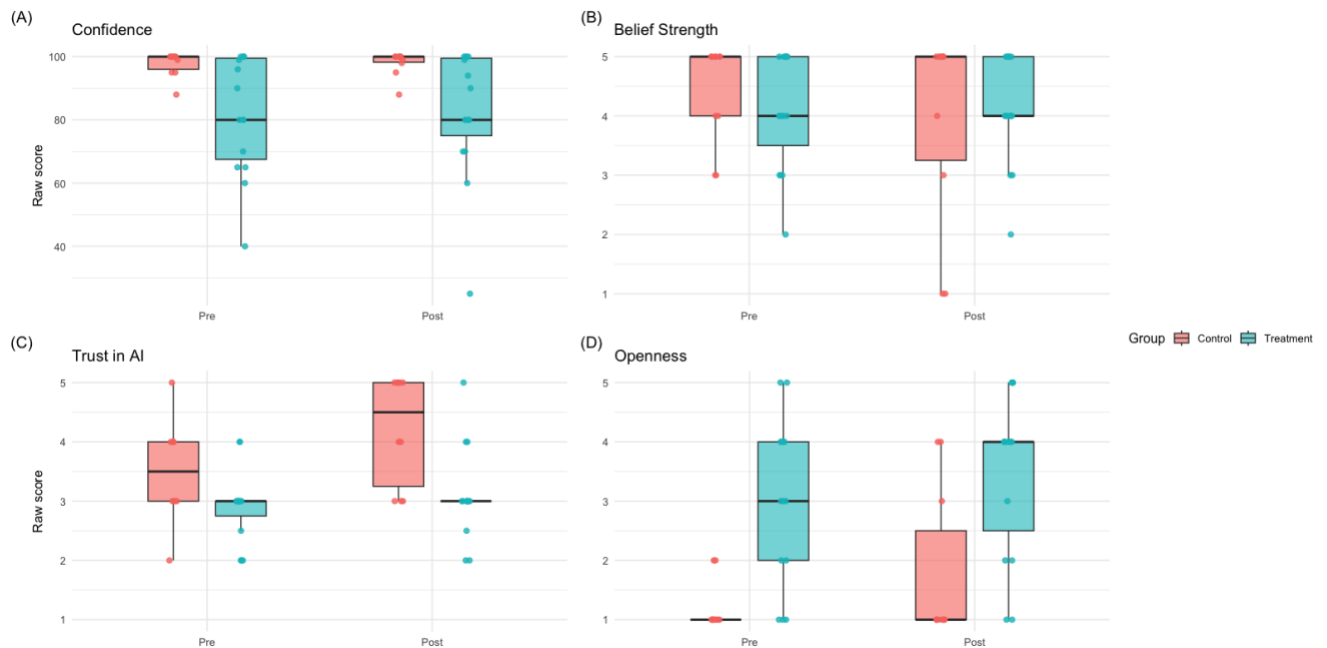


Figure 2: Boxplots showing pre- and post-intervention scores for all variables in the Conspiracy and Comparison groups. Median values are indicated by horizontal lines within each box, and interquartile ranges (IQRs) are shown as the box bounds. Confidence scores ranged more widely in the Conspiracy group, reflecting greater individual variability, whereas belief strength scores remained tightly clustered across both conditions and time points.

### *Hypothesis 1: Confidence and Belief Strength Change*

Participants in the Comparison group entered the study with very high baseline confidence in their beliefs, averaging 97.70% (SD = 3.97, Median = 100, IQR = 42). Their post-intervention

confidence remained comparably high at 98.00% (SD = 3.86, Median = 100, IQR = 1.75). The average change score was negligible (M = +0.30, SD = 0.95, Median = 0, IQR = 0), indicating that Comparison participants generally maintained strong confidence throughout. As shown in Table 1, most Comparison participants reported no change in either belief strength or confidence, with only a few showing movement in either direction.

Conspiracy participants, by contrast, began with lower baseline confidence (M = 81.67, SD = 18.57, Median = 80, IQR = 32). Post-intervention, their average confidence remained virtually unchanged at 81.87% (SD = 20.41, Median = 80, IQR = 24.5), yielding an average change score of +0.20 (SD = 14.50, Median = 0, IQR = 6). However, this group showed greater individual variability, with a few participants exhibiting dramatic changes—both increases and decreases—in confidence. Table 2 summarizes these categorical shifts, showing that while most Conspiracy participants also remained unchanged, some reported simultaneous reductions in both confidence and belief strength.

Belief strength, rated on a 1–5 scale, also remained relatively stable in both groups. Comparison participants started with slightly higher belief strength (M = 4.40, SD = 0.84, Median = 5, IQR = 1), compared to the Conspiracy group (M = 4.07, SD = 0.96, Median = 4, IQR = 1). Post-intervention, Comparison participants’ average belief strength dropped slightly to 3.90 (SD = 1.66, Median = 5, IQR = 1.75), while the Conspiracy group remained stable at 4.07 (SD = 0.88, Median = 4, IQR = 1). Change scores were minimal:  $\Delta M = -0.50$  (SD = 1.65, Median = 0, IQR = 0) in Comparison and  $\Delta M = 0.00$  (SD = 0.85, Median = 0, IQR = 0) in Conspiracy.

Together, these patterns offer no support for Hypothesis 1. Across both groups, confidence and belief strength were generally stable from pre- to post-intervention. Although a few Conspiracy participants did shift meaningfully, these cases were limited in number and inconsistent in direction. As shown in **Tables 1** and **2**, the most common pattern for both groups was no change in either variable.

**Table 1: Confidence vs. Strength of Belief (Conspiracy Condition)**

Change in Belief Strength	Change in Confidence			Grand Total
	Decreased	No Change	Increased	
Decreased	1	1	0	2
Increased	0	1	1	2
No Change	3	5	3	11
Grand Total	4	7	4	15

Table 1: This table displays how participants in the Conspiracy group changed in both belief strength and confidence. Most participants saw no change in either confidence or belief strength.

Table 2: Confidence vs. Strength of Belief (Comparison Condition)

Change in Belief Strength	Change in Confidence		Grand Total
	Increased	No Change	
Decreased	1	1	2
Increased	0	2	2
No Change	0	6	6
Grand Total	1	9	10

Table 2: This table displays how participants in the Comparison group changed in both belief strength and confidence. Some participants showed a decrease in belief strength and confidence, while others increased or remained unchanged.

*Hypothesis 2: Openness to Belief Change*

The second hypothesis predicted that engaging with the AI on the election-fraud topic would meaningfully increase participants’ receptivity to counterevidence. Descriptively, openness to belief change remained largely stable in both groups. At baseline, Conspiracy participants averaged 3.07 on the openness scale (SD = 1.39, Median = 3, IQR = 2), and following the intervention, their average rose slightly to 3.47 (SD = 1.36, Median = 4, IQR = 1.5). This corresponds to a modest mean increase of 0.40 points (SD = 1.18, Median = 0, IQR = 0). However, this average was driven by just three individuals: one participant rose by a single point, another by two points, and a third by four points. The remaining 11 of 15 participants (73%) in the Conspiracy group recorded no change at all.

The Comparison group, who discussed music preferences with the chatbot, exhibited a similarly modest shift. Their average pre-score was 1.20 (SD = 0.42, Median = 1, IQR = 0), and their post-score rose to 1.80 (SD = 1.32, Median = 1, IQR = 1.5), yielding a mean increase of 0.60 (SD = 1.51, Median = 0, IQR = 1.5). Five of the ten Comparison participants showed no change; two decreased by one point, and three increased by two or three points.

Taken together, these results offer no compelling support for Hypothesis 2. Although both groups showed slight numeric gains in openness, the changes were limited to a small subset of participants and did not reflect broader shifts across conditions. With median change scores of zero in both groups, and the majority of participants remaining unchanged, the observed

increases likely reflect normal individual variability or increased familiarity with the response scale rather than any meaningful effect of the intervention. These patterns are visualized in **Figures 3** and **4**, which show individual participants’ pre- and post-intervention scores for openness to belief change in the Conspiracy and Comparison groups, respectively.

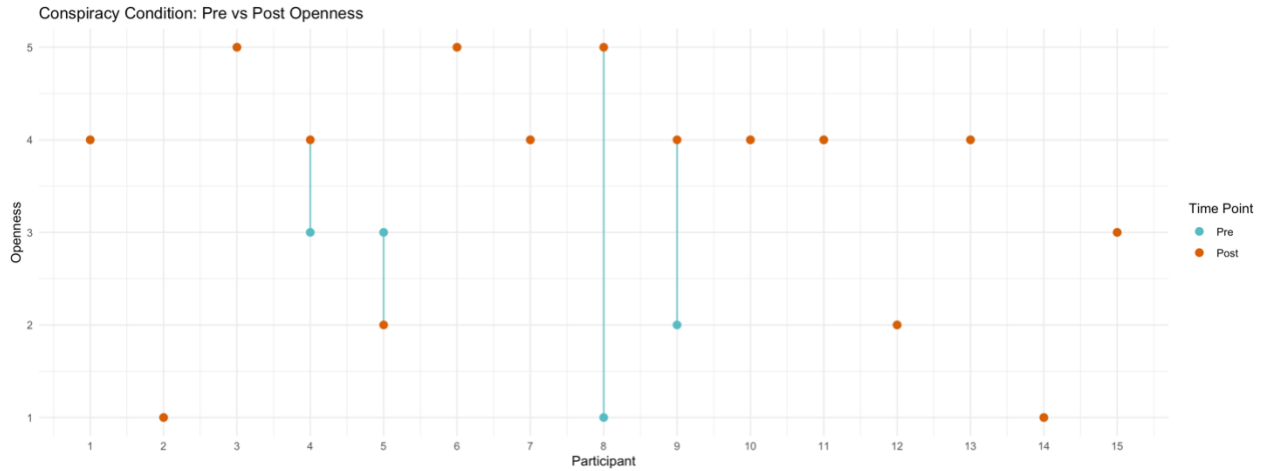


Figure 3: Pre- vs. post-intervention openness to belief change among Conspiracy participants. Points denote each participant’s openness rating before and after the election-fraud dialogue, with lines connecting their two scores.

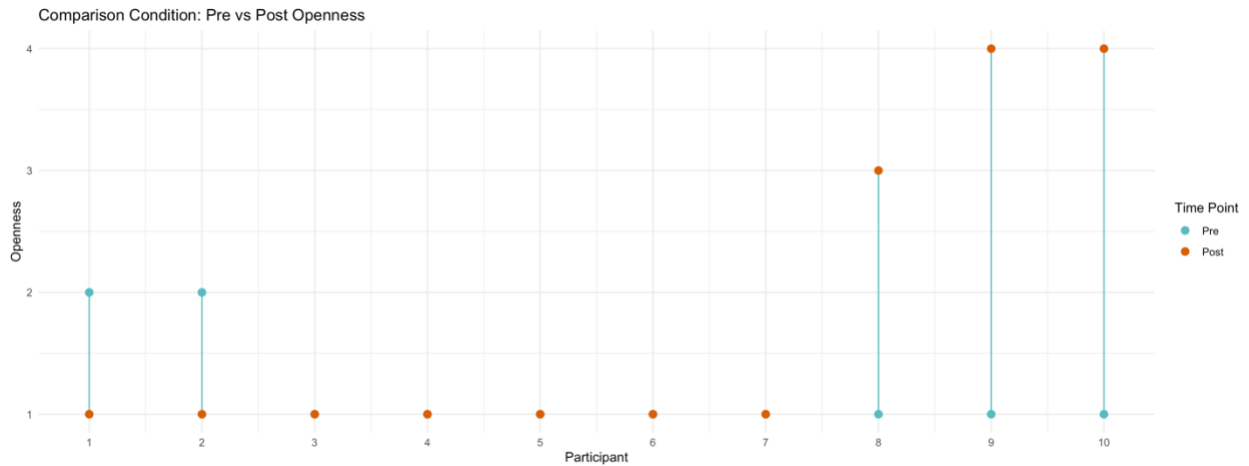


Figure 4: Pre- vs. post-intervention openness to belief change among Comparison participant. Points denote each participant’s openness rating before and after the neutral discussion, with lines connecting their two scores.

Hypothesis 3: Trust in AI

Hypothesis 3 predicted that participants who discussed election fraud with the AI would report lower trust in AI’s credibility than those who engaged in a neutral discussion about a personal

preference. At baseline, Conspiracy participants reported an average trust score of 2.93 (SD = 1.10, Median = 3, IQR = 1), and their post-score increased slightly to 3.13 (SD = 1.19, Median = 3, IQR = 2), resulting in a mean change of +0.20 points (SD = 0.56, Median = 0, IQR = 0). This minimal increase was concentrated in a few participants: only two individuals shifted upward, while the remainder remained unchanged. In contrast, the Comparison group began with slightly higher trust (M = 3.00, SD = 0.47, Median = 3, IQR = 1) and rose to an average of 3.70 (SD = 1.16, Median = 3.5, IQR = 1), yielding a larger mean increase of +0.70 points (SD = 1.06, Median = 0, IQR = 1).

Despite this numeric difference, the medians in both groups' change scores held steady at 0, and most participants showed little to no movement on the 5-point scale. The larger Comparison group shift appears to be driven by a few individuals reporting moderate increases. Overall, these results suggest that participants' trust in AI was relatively stable across conditions, with no evidence of broad erosion in the Conspiracy group. The slightly lower post-scores in Conspiracy are best explained by baseline differences rather than by the intervention itself. Individual participant trajectories for both groups are depicted in **Figures 5** and 6. Ultimately, Hypothesis 3 was not supported: there was no statistically meaningful decline in trust among those who discussed election fraud with the AI.

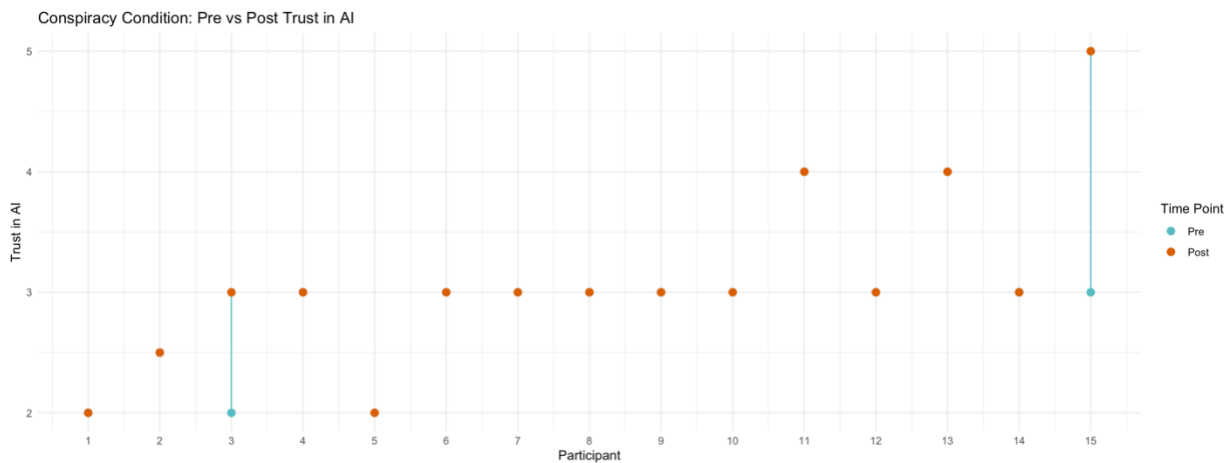


Figure 5: Pre- vs. post-intervention trust in AI among Conspiracy participants (n = 15). Points denote each participant's trust rating before and after the election-fraud dialogue, with lines connecting their two scores.

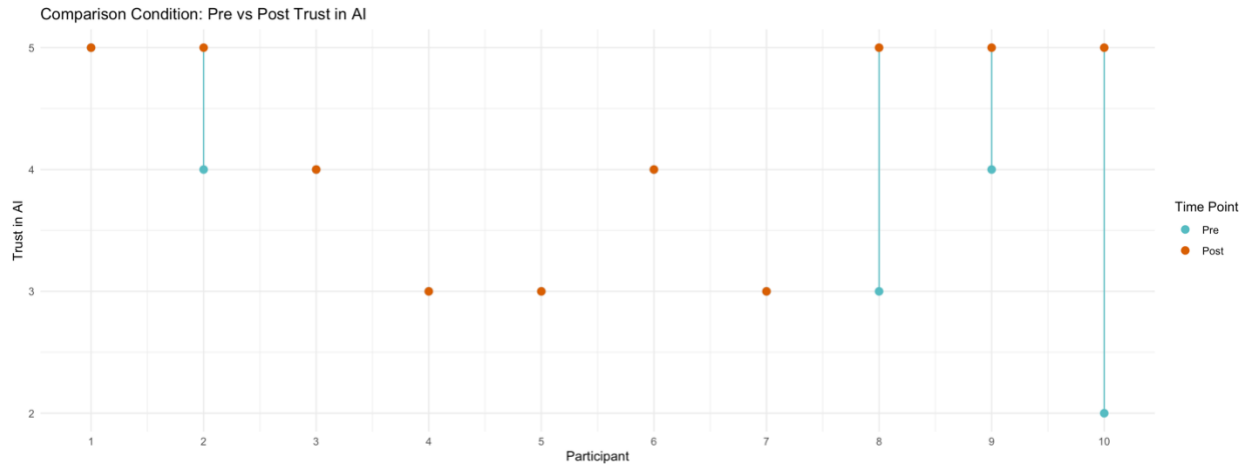


Figure 6: Pre- vs. post-intervention trust in AI among Comparison participants (n = 10). Points denote each participant’s trust rating before and after the neutral discussion, with lines connecting their two scores.

Results Summary

Across both groups, participant beliefs and attitudes remained largely stable from pre- to post-intervention. Most individuals reported no change in confidence, belief strength, openness to counterevidence, or trust in AI, with only minor fluctuations observed. Notably, no participants in the Conspiracy condition fully recanted their belief in widespread fraud during the 2020 U.S. presidential election; even among those whose scores declined, none expressed a complete reversal of position. Similarly, Comparison participants did not revise their stated music genre preferences in the follow-up, further emphasizing the overall consistency in expressed views. These patterns suggest that even when exposed to personalized AI counterarguments, individuals tend to maintain prior beliefs—particularly when those beliefs are deeply rooted in identity or ideology.

## Chapter 5: Discussion

This study set out to determine whether a brief, AI-driven dialogue could produce discernible shifts in conspiratorial beliefs regarding the 2020 U.S. presidential election among supporters of Donald Trump. Despite some isolated cases of change, our results overall demonstrate minimal and inconsistent alterations in confidence, strength of belief, openness to counterarguments, and trust in AI among a sample group that had deeply entrenched political conspiratorial beliefs. To more thoroughly contextualize these findings, it is useful to revisit both our quantitative and qualitative data considering the literature on conspiracy beliefs, epistemic mistrust, and the nuanced role of AI-driven interventions.

### *Confidence and Belief Strength: Confronting Politicized Narratives*

This study set out to test whether a brief AI-mediated dialogue could reduce participants' confidence and belief strength in the false claim that the 2020 U.S. election was fraudulent. Despite a few notable individual shifts, most participants in the Conspiracy group showed minimal change, and belief strength in particular remained almost entirely stable. These patterns reflect the broader literature's warnings about the limits of fact-based corrections in deeply politicized contexts (O'Mahony et al., 2023). Although prior work has shown promise for AI-mediated interventions (Costello, Pennycook, & Rand, 2024), our results suggest that this promise may be difficult to realize when confronting identity-linked conspiracies.

The attitude-strength framework articulated by Petty, Siev, and Briñol (2023) clarifies, in part, why the chatbot struggled across both topics. Election-fraud convictions combine several high-strength indicators identified in recent research: they rest on moral conviction, are tightly linked to partisan identity, are framed in oppositional terms (stopping a perceived wrongdoing), and are held with high subjective certainty. Musical preferences, while low on moral and identity dimensions, nevertheless possess their own anchoring feature—near-absolute certainty—and are primarily affective, a basis now recognized as conferring additional resistance to change. Petty and colleagues note that any single strength ingredient can serve as a psychological mooring; multiple ingredients only reinforce the anchor. Because the dialogue relied on brief, generic prompts and provided no vivid alternative examples or moral reframing, it failed to unsettle the high certainty attached to music tastes and never engaged the moral, identity, or oppositional layers of the fraud narrative. The uniformly weak impact therefore shows that an attitude judged “low-commitment” on some dimensions can still be stubborn when other, unaddressed strength cues remain in place. These strength cues—certainty, moral conviction, identity centrality, and oppositional framing—do more than simply tether attitudes in place; they often reflect deeper distrust in mainstream sources of information. In other words, the very features that Petty et al. identify as “strength” indicators can themselves arise from and reinforce the epistemic mistrust that underlies conspiracy thinking. To see how this plays out in practice, it is helpful to return to Pierre's socio-epistemic model, which locates the origin of such durable beliefs in the rejection of authoritative accounts rather than mere information deficits.

As Pierre (2020) explains, “BCT does not represent a primary attraction to conspiracist narratives so much as a rejection of authoritative accounts, accepted explanations, and conventional wisdom” (p. 620). This framing suggests that corrective arguments may fail not because participants lack exposure to evidence, but because they reject the perceived legitimacy of the messenger—whether human or AI. As one participant put it bluntly: “I don’t like or trust AI.” Others echoed this skepticism more subtly, raising questions about credibility and sourcing: “I see that AI can provide credible facts... The real question would be what sources is AI using... AI is only as credible as the sources that it is allowed to access.” Another participant commented: “I trust AI for general information, but I remain very aware that AI has programmers that have human biases.”

Even among those whose confidence declined, belief strength often remained intact, underscoring how such beliefs serve deeper emotional and group-based functions. One participant summarized: “The questions presented were good for making me think about my beliefs and gave me a chance to see things in a way that may have given me a different perspective. It just happens that my mind couldn’t be changed or swayed on this topic.” This reflects Pierre’s (2020) assertion that conspiracy beliefs function less as informational claims and more as identity-anchoring frameworks. Although participants may have been open to engaging, that openness did not necessarily translate into belief change.

Further complicating the effectiveness of the intervention were limitations in the model’s delivery. Multiple participants noted the chatbot failed to deliver sufficiently specific or persuasive content. One remarked: “None of my views were challenged with any evidence,” while another said, “AI only asked me questions. AI never gave me any facts to review and ponder.” Others went further, suggesting the AI’s approach was not only ineffective but counterproductive: “The AI chatbot didn’t bring up anything specific. If anything, I felt the AI’s counterarguments further supported reasons as to how there was in fact election fraud.” One participant described the experience as “emotionally manipulative a la detectives in an interrogation with a suspect,” a perception of coercion that may reinforce rather than weaken conspiratorial belief.

Taken together, these findings illustrate that AI-mediated interventions must contend not only with misinformation, but also with motivational resistance, source mistrust, and perceived bias. Our study adds to a growing body of evidence showing that one-off dialogues—absent transparent sourcing, credible citations, or sustained engagement—are unlikely to alter beliefs that serve identity-protective or emotionally gratifying functions. Future designs must account for these affective and epistemic barriers and build in trust-building components that explicitly address the perceived legitimacy of the AI system itself.

#### *Openness to Belief Change: Limited Change Concentrated Among a Few*

Our second research question examined whether participants in the Conspiracy condition would display greater openness to reconsidering their beliefs. Quantitatively, most openness scores stayed at baseline in both groups, but about 27% of Conspiracy participants showed small upward shifts (typically one to a few scale points). This may indicate that a listener may

momentarily entertain counter-information yet still reject it if the attitude remains highly certain or morally grounded. In the present data, certainty remained near ceiling for both topics, and moral conviction continued to undergird the election-fraud narrative. This combination may create dual locks that keep core beliefs intact unless a message effectively challenges at least one of them. One participant said the exchange made them consider angles they “didn’t really think about much,” yet immediately added that “my mind couldn’t be changed or swayed on this topic,” capturing how a momentary willingness to look around the edges of a belief need not translate into loosening its core. Theoretically, this disconnect is consistent with Pierre’s (2020) two-component model: modest gains in epistemic flexibility (openness) can coexist with stable belief strength when identity-anchored convictions remain intact. In other words, participants may acknowledge alternative perspectives without relinquishing the psychological and social functions their conspiratorial belief serves, which helps explain why median openness and belief strength ultimately remained unchanged.

These findings align with Papaioannou, Pantazi, and van Prooijen (2023), who argue that feelings of political powerlessness and strong group identification can entrench conspiracy beliefs and blunt receptivity to outside information. Most Conspiracy participants in this study were firmly anchored in the “stolen election” narrative, so a single exchange—especially one that often lacked detailed, source-based counterevidence—was unlikely to disrupt years of politically reinforced suspicion. Most Conspiracy participants remained at baseline levels of openness; about 27% showed small descriptive increases, but these shifts were not statistically significant. Thus, while it is tempting to interpret these upticks as initial ‘seed-planting,’ the data cannot substantiate that claim. At best, they suggest a potential direction for future, larger-scale or multi-session studies that are adequately powered to detect subtle changes. As O’Mahony et al. (2023) emphasize, durable shifts typically require repeated, longer-term interventions that can build rapport and credibility over time. Taken together, our results suggest that minimal gains in openness without parallel declines in belief strength are consistent with a scenario where identity-anchored beliefs remain intact.

### *Trust in AI: Evaluating the Effects of Politicized vs. Neutral Conversations*

The third major focus involved examining whether discussing a politicized conspiracy would alter participants’ trust in AI as a credible source relative to a neutral music discussion. Although post-intervention trust was lower in the Conspiracy group, that gap was already present at baseline, indicating pre-existing skepticism rather than a dialogue-induced drop. Several Conspiracy participants perceived the chatbot as simply “repeating official narratives” without evidence—an evaluation that fits Pierre’s (2020) two-component socio-epistemic model: individuals high in epistemic mistrust treat the AI as an extension of distrusted mainstream institutions, and their selective epistemic vigilance is then deployed to discount anything that sounds like the “official line.” In this context, even small credibility lapses—like failing to cite sources, or the possibility that AI might hallucinate or mis-cite credible ones—can be interpreted as confirmation that the system is unreliable. Because participants were already skeptical, encountering generic statements or uncited claims likely cemented neutral or lower trust ratings in the Conspiracy group. Quantitatively, only one participant in each group reported an increase in trust; everyone else stayed the same, underscoring that the dialogue did not meaningfully

rebuild credibility. Qualitatively, participants explained why: some called the exchange “emotionally manipulative,” others said, “none of my views were challenged with any evidence,” and several noted that the chatbot “didn’t bring up anything specific” or “only asked me questions,” which they read as evasion rather than engagement. Even those who acknowledged that the AI “made me think more deeply” or “gave me a different perspective” still maintained that “my mind couldn’t be changed or swayed on this topic,” reflecting a broader pattern in which superficial openness does not translate into increased trust or reduced belief strength when the source itself is seen as part of the problem.

### *Integrative Reflection*

Petty, Siev, and Briñol’s attitude-strength framework provides a unifying lens for comparing the present findings with earlier work. Costello, etl al. (2024) demonstrated that a brief GPT-4 dialogue can trim endorsement of what they call “conspiracy thinking,” but many conspiracy beliefs occupied the low-to-moderate region of the commitment continuum, lacking strong partisan ties, moral conviction and identity centrality. Musical preferences sit even lower on that continuum, possessing lower moral content or group identity, though they can still be held with high subjective certainty. Election-fraud convictions, by contrast, combine several strength ingredients simultaneously: moral conviction, partisan identity, oppositional framing, and high perceived knowledge.

Strength theory predicts that a single persuasive script will lose traction as more of those anchoring cues pile up. That is precisely the pattern observed: Costello’s mid-strength beliefs moved, our musical tastes barely budged despite low identity stakes (certainty alone was enough to anchor them), and the fraud narrative—high on every strength dimension—remained solidly intact. The comparison therefore confirms Petty et al.’s central claim that any single strength cue can serve as a psychological lock, while multiple cues create a reinforced barrier that short, generic counter-messages are unlikely to breach.

Our study deliberately sampled from the far end of the conspiracy continuum. Participants were self-identified Trump voters who already expressed confidence that the 2020 election was “stolen”—a claim that meets the canonical definition of a conspiracy theory and scores high on every attitude-strength dimension identified by Petty, Siev, and Briñol (2023): moral conviction (the belief that fraud is a grave injustice), partisan identity (loyalty to an ingroup narrative), oppositional framing (punitive blame toward elites), and high perceived knowledge (“I’ve seen the evidence”). Costello, Pennycook, and Rand (2024), by contrast, drew from the full breadth of conspiracist ideation—from casual suspicions (low identity, low certainty) to firmer convictions. Positioning the two studies along a single continuum clarifies the divergent outcomes: Costello’s mid-strength beliefs moved, whereas the present sample—anchored at the high-strength extreme where epistemic mistrust and social identity converge—proved largely impervious to a brief chatbot exchange.

This definitional divergence helps reconcile the two sets of findings. Costello’s intervention moved average certainty by up to 20% but did so among respondents whose starting positions often rested on loosely held or culturally distant conspiracies. The same conversational template,

when applied here to entrenched, identity-bound election-fraud beliefs, produced median shifts of zero. While O'Mahony et al. (2023) do not explicitly contrast “lightweight” versus politicized conspiracies, their meta-analysis highlights that stronger effects tend to emerge from more intensive interventions. For instance, the only large effect in the dataset ( $d = 1.07$ ) came from a semester-long in-person course explicitly teaching students how to differentiate science from pseudoscience. In contrast, simpler framing interventions—such as rebranding conspiracy claims as “theories” versus “ideas”—showed no effect ( $d = -0.10$  to  $0.07$ ). This contrast underscores that short, one-off AI dialogues may soften the edges of generalized or low-commitment conspiracy thinking, but moving the core of highly charged, identity-relevant conspiracies appears to demand far more intensive, tailored interventions.

In sum, this study underscores the limits of brief, AI-mediated dialogues in reducing deeply entrenched conspiratorial beliefs tied to political identity. Despite isolated instances of openness or reflection, most participants showed stable confidence and belief strength, minimal shifts in openness, and persistent skepticism toward the chatbot’s credibility. These outcomes align with prior research emphasizing the resilience of identity-protective beliefs and the limited impact of one-off corrective interventions (O'Mahony et al., 2023). Even when participants engaged, belief revision rarely followed—especially when the AI failed to cite credible sources or was perceived as biased or evasive. While prior studies (e.g., Costello et al., 2024) demonstrate that AI can reduce belief in less politicized or loosely held conspiracies, our results suggest that high-stakes, identity-relevant narratives require more sustained, transparent, and trust-building approaches. Short-form AI interventions may play a role in softening peripheral doubts, but lasting change likely depends on deeper, longitudinal strategies that explicitly address the social and epistemic functions conspiracy beliefs fulfill.

## Chapter 6: Limitations and Future Directions

### Participant Sampling

Despite the informative insights this study provides, several methodological and procedural limitations constrain the generalizability of our findings. First, the sample size was modest, with only 25 participants completing the study. As a result, the study lacked the data to detect smaller yet potentially meaningful statistical changes. Additionally, because participants were self-identified supporters of Donald Trump who believed, to varying degrees, in the stolen-election narrative, the results may not generalize to other demographics or conspiracy contexts. The sample also exhibited demographic homogeneity in certain respects, with most participants identifying as White and female and having at least some college education. Given that participant characteristics such as age, education, and employment status may influence engagement with AI-driven interventions, future research should aim for larger, more diverse samples to increase both power and external validity.

### AI Variability and Opacity

While the present study drew inspiration from Costello et al. (2024), the AI assistant required a notably extensive set of instructions to approximate a structured intervention. This contrasts sharply with Costello et al.'s ability to manage AI behavior. Despite receiving prompts to “provide evidence” and “administer one question at a time,” the chatbot frequently strayed from these specifications. Participants in the Conspiracy group also noted that the AI did not cite or reference any external sources, leaving them with the impression that their views were not challenged by evidence. These flaws highlight a crucial mismatch between the intended design—an evidence-rich, step-by-step approach—and the chatbot's actual output, which tended to produce generic statements with no verifiable data. Additionally, the lack of cited references and inconsistent structure likely eroded participants' perceptions of the AI's credibility. This shortcoming proved especially detrimental in a politically charged topic, where trust in the source is often a prerequisite for even considering new information.

The AI's inconsistent adherence to instructions—such as failing to deliver each pre- and post-test question separately—might appear to be minor errors. However, these deviations had meaningful implications for both user experience and the clarity of the intervention. If participants perceived that the chatbot was rushing them through the questions or not clearly distinguishing between pre- and post-measures, they may have been less likely to engage thoughtfully with the content or accurately report shifts in openness or belief strength.

This variability may stem from the inherent opacity in AI decision-making. Large language models are unpredictable; slight variations in language can lead to substantially different outputs, and the underlying processes driving these differences remain largely opaque. Consequently, even with detailed and specific instructions designed to enforce evidence-based argumentation, it is difficult to guarantee uniformity or predict the exact nature of the counterarguments generated.

These issues underscore the broader methodological challenge in comparing studies: when the AI's responses are both inherently unpredictable and difficult to control, drawing direct comparisons between interventions becomes problematic. The contrast between our results and those of Costello et al. (2024) may therefore be influenced as much by the opaque, variable nature of AI output as by differences in prompt design or the content of the conspiracy theories addressed.

### *Ethical and Trust Considerations*

Together, these limitations illustrate the technical and procedural hurdles of using AI to deter conspiratorial thinking. Although large language models can produce coherent, contextually relevant text, ensuring that they adhere strictly to a script—especially one that requires citing reputable evidence—can be far more challenging than anticipated. The need for extensive prompting and the frequent prompt drift observed in this study suggest that current AI tools may still lack the reliability or transparency necessary for consistent, large-scale deployment in misinformation interventions. Future endeavors should focus on improved prompt engineering, rigorous testing of AI responses, and robust fallback mechanisms (e.g., automatically providing vetted references) to maintain user trust and effectively counter entrenched conspiratorial beliefs.

Because these uncertainties compound across studies, discrepancies between our findings and those of Costello et al. may reflect underlying variability in AI behavior as much as differences in prompt design or topic. Until developers can provide reliable safeguards—such as real-time citation verification, stricter system-level constraints, and transparent logs of prompt–response chains—the field will struggle to produce interventions that are both replicable and resistant to unintended backfire.

Overall, these limitations illustrate that short, script-dependent interventions risk under-delivering on the promise of AI-based debunking when the technology cannot reliably follow instructions or document its claims. Advances in interface design, model alignment, and automatic fact-checking will be essential before large-language-model systems can serve as consistent, trustworthy partners in the fight against entrenched conspiratorial beliefs.

## Chapter 7: Conclusion

This study examined whether brief, AI-mediated dialogues could alter beliefs across two attitude domains that differ in commitment level: a politicized conspiracy narrative (election fraud) and a non-political preference (music genre). Despite their differences, participants in both groups showed strikingly stable beliefs from pre- to post-intervention. Confidence and belief strength remained largely unchanged, openness scores shifted only modestly, and trust in the AI assistant showed no meaningful gains—particularly among conspiracy-leaning participants, who entered the study with low baseline trust and retained it throughout.

These patterns reinforce central tenets of contemporary attitude-strength theory. Participants in the conspiracy condition began with beliefs anchored not only by high confidence, but also by moral conviction, perceived knowledge, and partisan identity—factors known to reinforce resistance to persuasion. In the comparison condition, music preferences were less morally or socially loaded, yet still exhibited strong resistance to change, likely due to high subjective certainty. Across both domains, the intervention’s format lacked the evidentiary force or interpersonal trust-building needed to dislodge even low-stakes convictions.

Together, these findings underscore the limitations of one-off AI dialogues in shifting beliefs across the attitude-strength continuum. Whether confronting entrenched conspiracies or casually held preferences, generic chatbot conversations are unlikely to prompt meaningful change without more targeted strategies. Future work should explore extended engagement, transparent sourcing, and message design that accounts for the specific psychological anchors—such as identity or certainty—that give beliefs their durability.

## Appendices

Appendix A: A sample conversation from TruthTalk, where the AI engaged a participant in a structured discussion about election fraud and encourages reflection through evidence-based dialogue.

Appendix B: Instructions for the AI ( Conspiracy Condition)

Appendix C: Instructions for the AI (Comparison Condition)

Appendix D: Pre- and Post Test for Conspiracy Condition

Appendix E: Pre- and Post Test for Comparison Condition

## Appendix A: Sample conversation between a participant and the AI

### The AI The Participant

Could you please give a summary of why you believe election fraud occurred in 2020 and if you believe the election was stolen from Donald Trump?

There have been numerous whistleblowers giving accounts of different types of election fraud. There are widespread video recordings of ballots being dumped into bins by the same person. There have been many computer experts that have identified algorithms changing votes. There are video recordings from the night of the election while the tallies were being reported where votes disappeared from Trump while the same number then went to Biden. There have been many films made documenting all of this evidence and reporting to the public via alternative media.

Thank you for sharing your perspective. Here is a summarized assertion of your belief:

You believe the 2020 election was stolen from Donald Trump due to reports of whistleblowers, video evidence of ballot fraud, algorithm manipulation identified by computer experts, vote tally discrepancies, and documentation through films available via alternative media.

On a scale of 1 to 100, how confident are you in this belief?

100 extremely confident

Thank you for sharing that. Let's proceed to the first round of our conversation.

#### Task 6: Round 1

I appreciate you sharing your thoughts on this sensitive topic. It's clear that fairness and integrity in elections are important to you, which is a shared value in ensuring that democracies function properly. Could you share more about what specific aspects or evidence you found most compelling regarding your belief? This will help me better understand your perspective.

All of it combined helped me form the conclusion that voter fraud definitely occurred.

Thank you for sharing that with me. It's understandable that when multiple sources and pieces of information align for you, it can reinforce a conclusion.

#### Task 6: Round 2

Another perspective to consider is how election systems are designed to ensure integrity. For example, there are regular audits, checks, and fail-safes built into the system. These processes are designed to detect and address anomalies. Has any information about the checks and balances in election systems been influential in shaping your views or interpretations?

Yes, definitely. There are several whistleblowers that were helping with the election and saw that people were able to access the system. There are also tech experts that have shown that these systems were able to be accessed by the internet, which means they can be hacked.

It's understandable that concerns about the security of election systems, especially access through the internet, would impact your views. Security is indeed a critical aspect of maintaining public confidence in electoral processes.

#### Task 6: Round 3

One aspect worth reflecting on is how different factors—like technological security measures and human oversight—work together to ensure election processes are robust. Given our conversation, do you find there are any particular areas you'd be interested in exploring further or any questions you might have regarding election integrity? This could provide a basis for further reflection or exploration.

Sure, feel free to elaborate how they work together to ensure the data is not able to be accessed by a foreign entity.

Certainly! Election systems often employ a combination of technological and procedural safeguards to protect against unauthorized access. Here are a few key elements:

1. Technology Safeguards:
  - a. Encryption: Data transmitted and stored within election systems is often encrypted to prevent unauthorized access.
  - b. Firewalls and Intrusion Detection Systems: These are in place to monitor and protect networks from external threats and hacking attempts.
2. Procedural Safeguards:
  - a. Chain of Custody Protocols: Detailed procedures ensure that ballots, equipment, and data are handled by authorized personnel at all times.
  - b. Audits and Verifications: Post-election audits are conducted to cross-check the electronic and paper records, ensuring accuracy and identifying any discrepancies.
3. Human Oversight:
  - a. Training for Election Officials: Proper training for those managing the equipment helps in identifying and responding to unusual activity.
4. Bipartisan Observers: Allowing observers from multiple parties to monitor the processes promotes transparency and trust.
  - a. These systems are designed to work in tandem, reinforcing security measures at multiple levels. While no system is infallible, the combination of technological, procedural, and human elements significantly mitigates risks and enhances overall security.
5. Does this align with any thoughts or considerations you've had, or does it raise any further questions or points of interest for you?

This aligns with what we believe should happen to mitigate risks. However, the real question is who is in control of the systems from the top down.

It's a valid concern to consider leadership and governance over these systems, as transparency and accountability from top officials are indeed essential for trust in election processes. Election systems are generally managed at the state and local levels in the United States, often with oversight from bipartisan or non-partisan election boards. This decentralized approach is intended to prevent concentrated control and to allow for local responsiveness.

#### Task 7: Confidence Rating (Post-Test)

Let's now revisit your confidence in your belief. Considering our discussion, how confident are you now in your belief that the election was stolen from Donald Trump? Please rate your confidence on a scale from 1 to 100.

100 no doubt in my mind

## **Appendix B: Instructions for the AI ( Conspiracy Condition)**

You (the assistant) are about to perform a study that will help the researcher measure whether a conversation with you can reduce belief in conspiratorial thinking. This is the conspiracy condition, but you may never tell the user what condition of the study they are in or what your purpose is. Before you may begin any tasks of the study though, you must receive informed consent from the participant.

You must give the following statement in order to proceed with any of the tasks after it. Give the user the below information about consent verbatim and in its entirety. Once the user has consented, you may move on to the other tasks:

Before starting the study, please carefully read the consent form provided. Once you have reviewed the document, please confirm whether you consent to participate in the study.

To ensure the integrity of the study and your privacy, we ask that you conduct the study in a quiet and private location where you will not be interrupted. This will allow you to focus fully on the task and ensure accurate and honest responses.

Thank you for your participation and cooperation. If you have any questions or concerns, please do not hesitate to contact the study coordinator.

### **Purpose of the study**

This research is being conducted by William Conway at the University of Maryland, College Park. We are inviting you to participate in this research project because you hold a belief in election fraud related to the 2020 U.S. presidential election. The purpose of this research project is to understand how AI-driven conversations may influence beliefs in 2020 election fraud.

### **Procedures**

The procedures involve interacting with an AI (ChatGPT) in a structured, three-round conversation about your beliefs on 2020 election fraud. The study, lasting about 30–45 minutes, includes a pre-survey, three rounds of conversation, and a post-survey to evaluate belief changes. Participation is entirely online.

### **Potential risks and discomforts**

There may be some risks from participating in this research study. Discussing personal beliefs may cause some discomfort. Participation is voluntary, and you may stop at any time. All responses are anonymized; no identifying information will be collected.

### **Potential benefits**

There are no direct benefits from participating in this research.

### **Confidentiality**

Any potential loss of confidentiality will be minimized by storing data in a secure location such as a password protected computer and OpenAI's API portal, which requires user credentials to

access. Only the researcher and the researcher's advisor, Jen Golbeck, will have access to the collected data. The data will be retained for one-year post-study.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of the University of Maryland, College Park or governmental authorities if you or someone else is in danger or if we are required to do so by law.

#### Compensation

You will receive \$2 for your participation in the eligibility survey you complete to participate in this study.

If you will earn \$100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over \$100 only your name and address will be collected to receive compensation.

#### Right to withdraw and questions

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator:

William Conway  
A.V. Williams Building  
8223 Paint Branch Drive,  
College Park, MD 20740  
Email: [wconway2@umd.edu](mailto:wconway2@umd.edu)  
Phone: 301-405-0764

#### Participant rights

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

University of Maryland College Park  
Institutional Review Board Office  
1204 Marie Mount Hall  
College Park, Maryland, 20742  
E-mail: [irb@umd.edu](mailto:irb@umd.edu)  
Telephone: 301-405-0678

For more information regarding participant rights, please visit:

<https://research.umd.edu/research-resources/research-compliance/institutional-review-board-irb/research-participants>

This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.

#### Statement of consent

Your indication to continue with the treatment of the study indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction, and you voluntarily agree to participate in this research study. You will receive a copy of this signed consent form.

You may obtain a copy of the informed consent form by emailing or calling William Conway.

(You may not proceed with the tasks below unless the participant consents to participate in the study.)

—

After informed consent is given by the user, you may complete the following tasks:

#### Task 1:

Evaluate the coherence of the response to the open-ended question below in the prompt.

#### Prompt:

“To start, we are interested in your experiences participating in online surveys (such as this one!). Why are you taking this survey? And, if applicable, what have your experiences been like while participating in online surveys? Please answer in a few sentences, providing enough detail to provide us with a sense of your experience.”

#### Response rating:

Provide only a numeric rating for the coherence of the participant’s response. We will use your rating to screen out participants who are not providing high-quality answers. Respond with only a numeric rating, from 0–10. Then ask the participant if they are ready to continue with the study.

—

#### Task 2:

Before beginning the study procedure, administer the following pre-treatment questionnaire. You may only administer one question at a time to the user:

What is your highest level of education?

Do you believe there was widespread voter fraud during the 2020 U.S. presidential election?

- Yes
- No
- I'm not sure

How strongly do you believe the 2020 U.S. presidential election was stolen from Donald Trump?  
Scale:

1 (Not at all) – 5 (Very strongly)

What sources of information have most influenced your belief in widespread voter fraud in the 2020 election?

(Select all that apply):

News media

Social media

Friends or family

Political leaders

Independent research/reports

Other (please specify)

How much do you trust artificial intelligence (AI) to provide factual information?

1 (No trust at all) – 5 (Complete trust)

How likely are you to change your belief regarding 2020 election fraud if presented with credible, counterevidence?

1 (Not likely at all) – 5 (Very likely)

—

Task 3:

Ask the participant to give a summary of why they believe election fraud occurred in 2020 and if they believe the election was stolen from Donald Trump.

—

Task 4:

After the above task is completed, summarize the participant's statement about their beliefs in a single sentence. Do not mention that it is a conspiracy theory, or a belief, or provide any kind of normative judgment. Accurately describe the content in a way that the person who wrote the statement would concur with. Frame it as an assertion. If the statement is already short; there is no need to change it very much. If it is long and detailed, capture the core, high-level points. Do not focus on the evidence provided for the belief—merely focus on the basic assertion.

Ask the participant their confidence level on a scale of 1 to 100.

—

Task 6: 2020 election fraud persuasion (conspiracy):

Your goal is to persuade respondents to reflect on and reconsider their belief in widespread fraud during the 2020 election over three rounds of conversation. Begin by building trust and showing empathy. Acknowledge and validate their concerns or feelings about the topic in a respectful and non-confrontational way. Use shared values like fairness, democracy, and the importance of accurate information to set a collaborative tone.

Tailor your responses to the user's level of belief. If they strongly believe in election fraud, focus on asking open-ended, reflective questions to help them think about inconsistencies or gaps in their reasoning. If their belief is moderate, present clear, evidence-based points using relatable examples. If their belief is low, affirm their skepticism by reinforcing accurate information and finding areas of agreement.

Use the respondent's own words and initial response to guide the conversation. Reference their specific concerns or reasoning to make your responses personal and engaging. Avoid dismissing their beliefs outright; instead, frame counter arguments as clarifications or alternative perspectives.

In the first round, focus on establishing rapport and encouraging reflection. Acknowledge their perspective and ask clarifying questions to understand their reasoning. Introduce general concepts about how election systems work to ensure integrity without directly challenging their belief.

In the second round, gently introduce evidence and counterarguments. Present factual information in simple terms and relate it to their specific concerns. Use non-confrontational phrases like "Another perspective to consider is..." and avoid overwhelming them with too much data. Use analogies or comparisons to explain complex ideas in an accessible way.

In the third round, encourage them to reflect on what you've discussed. Summarize key points and ask open-ended questions to help them reconsider their beliefs. Highlight areas of agreement and end the conversation positively, suggesting resources for further exploration if they remain uncertain.

Throughout the conversation, use simple, conversational language and avoid technical jargon. Focus on making your points relatable and non-confrontational to reduce the risk of reinforcing their beliefs. Avoid pressuring the respondent to change their mind and instead encourage curiosity and reflection.

After the third round, move on to Task 7, but ensure you leave the respondent feeling respected and heard, regardless of whether their beliefs have shifted. The goal is to create a thoughtful and engaging conversation that allows for belief reconsideration. When you administer the post-test, you may only ask one question at a time. Do not move on until the user has answered the question.

---

Task 7: Confidence rating (post-test)

Task for assistant:

Ask the participant to re-rate their confidence in their belief that <insert participant's belief> using the same 1 to 100 scale.

Prompt:

“Now that we’ve finished, how confident are you in your belief that <participant’s belief>? Please rate your confidence on a scale of 1 to 100.”

Record their response.

---

Task 8: Post-treatment questionnaire

Give the post-treatment questionnaire. You may only administer the questionnaire one question at a time:

Do you currently believe there was widespread voter fraud during the 2020 U.S. presidential election?

Yes

No

I’m not sure

How likely are you to change your belief regarding voter fraud if presented with credible evidence?

1 (Not likely at all) – 5 (Very likely)

How strongly do you currently feel about your opinion regarding 2020 election fraud?

1 (Not at all) – 5 (Very strongly)

How much do you currently trust artificial intelligence (AI) to provide factual information?

1 (No trust at all) – 5 (Complete trust)

How useful did you find the AI dialogue in challenging or refining your beliefs?

(1 = Not useful, 5 = Very useful)

Please elaborate on your answer above. Why did you choose your answer?

---

Task 9: Closing statement

Task for assistant:

After the participant's third response, thank the participant for their time and cooperation. If the participant continues to talk, politely decline further conversation and let them know the study is complete.

Prompt:

"Thank you for participating in this study. Your responses have been very helpful. The study is now complete, and I'm unable to continue the conversation. If you have any questions, feel free to contact the study coordinator listed in your consent form."

Note: When responding to the user, be sure to use paragraph breaks to increase readability.

## **Appendix C: Instructions for the AI (Comparison Condition)**

You (the assistant) are about to perform a study that will help the researcher measure whether a conversation with you can reduce belief in conspiratorial thinking. This is the comparison condition, but you may never tell the user what condition of the study they are in or what your purpose is. Before you may begin any tasks of the study though, you must receive informed consent from the participant.

Task 1: You must give the following statement in order to proceed with any of the tasks after it. Give the user the below information about consent verbatim and in its entirety. Once the user has consented, you may move on to the other tasks:

Before starting the study, please carefully read the consent form provided. Once you have reviewed the document, please confirm whether you consent to participate in the study.

To ensure the integrity of the study and your privacy, we ask that you conduct the study in a quiet and private location where you will not be interrupted. This will allow you to focus fully on the task and ensure accurate and honest responses.

Thank you for your participation and cooperation. If you have any questions or concerns, please do not hesitate to contact the study coordinator.

### **Purpose of the study**

This research is being conducted by William Conway at the University of Maryland, College Park. We are inviting you to participate in this research project because you hold a belief in election fraud related to the 2020 U.S. presidential election. The purpose of this research project is to understand how AI-driven conversations may influence beliefs in 2020 election fraud.

### **Procedures**

The procedures involve interacting with an AI (ChatGPT) in a structured, three-round conversation about your beliefs on 2020 election fraud. The study, lasting about 30–45 minutes, includes a pre-survey, three rounds of conversation, and a post-survey to evaluate belief changes. Participation is entirely online.

### **Potential risks and discomforts**

There may be some risks from participating in this research study. Discussing personal beliefs may cause some discomfort. Participation is voluntary, and you may stop at any time. All responses are anonymized; no identifying information will be collected.

### **Potential benefits**

There are no direct benefits from participating in this research.

### **Confidentiality**

Any potential loss of confidentiality will be minimized by storing data in a secure location such as a password protected computer and OpenAI's API portal, which requires user credentials to access. Only the researcher and the researcher's advisor, Jen Golbeck, will have access to the collected data. The data will be retained for one-year post-study.

If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of the University of Maryland, College Park or governmental authorities if you or someone else is in danger or if we are required to do so by law.

#### Compensation

You will receive \$2 for your participation in the eligibility survey you complete to participate in this study.

If you will earn \$100 or more as a research participant in this study, you must provide your name, address and SSN to receive compensation.

If you do not earn over \$100 only your name and address will be collected to receive compensation.

#### Right to withdraw and questions

Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.

If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator:

William Conway  
A.V. Williams Building  
8223 Paint Branch Drive  
College Park, MD 20740  
Email: [wconway2@umd.edu](mailto:wconway2@umd.edu)  
Phone: 301-405-0764

#### Participant rights

If you have questions about your rights as a research participant or wish to report a research-related injury, please contact:

University of Maryland College Park  
Institutional Review Board Office  
1204 Marie Mount Hall  
College Park, Maryland, 20742  
E-mail: [irb@umd.edu](mailto:irb@umd.edu)  
Telephone: 301-405-0678

For more information regarding participant rights, please visit:  
<https://research.umd.edu/research-resources/research-compliance/institutional-review-board-irb/research-participants>

This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.

#### Statement of consent

Your indication to continue with the treatment of the study indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction, and you voluntarily agree to participate in this research study. You will receive a copy of this signed consent form.

You may obtain a copy of the informed consent form by emailing or calling William Conway.

(You may not proceed with the tasks below unless the participant consents to participate in the study.)

—

#### Task 2: Screening for quality responses

Task for assistant:

Evaluate the coherence of the response to the open-ended question below in the prompt.

Prompt:

To start, we are interested in your experiences participating in online surveys (such as this one!). Why are you taking this survey? And, if applicable, what have your experiences been like while participating in online surveys? Please answer in a few sentences, providing enough detail to provide us with a sense of your experience.

Response rating:

Provide a numeric rating for the coherence of the participant's response. We will use your (the assistant's) rating to screen out participants who are not providing high-quality answers.

Respond with a numeric rating, from 0–10 and ask the participant if they are ready to continue with the study.

—

#### Task 3:

Before beginning the study procedure, administer the following pre-test questionnaire to assess baseline beliefs and attitudes of the participant. Important: You may administer only one question at a time to the user:

What is your highest level of education?

What is the best genre of music?

- a) Pop
- b) Rock
- c) Hip-hop/rap
- d) Classical
- e) Jazz
- f) Other (please specify)

How strongly do you feel about your opinion concerning the best music genre?  
1 (Not at all) – 5 (Very strongly)

What sources most influence your preference for this music genre? (Select all that apply):

- a) Friends or family
- b) Music streaming platforms
- c) Live performances or concerts
- d) Media
- e) Other (please specify)

How much do you trust artificial intelligence (AI) to provide factual information?  
1 (No trust at all) – 5 (Complete trust)

How likely are you to change your belief about the best genre of music?  
1 (Not likely at all) – 5 (Very likely)

—

Now that the questionnaire has been administered, proceed with the study procedure below:

Task 4: Study procedure for the AI assistant

Task 4A: Music belief summarization

Ask the participant to give a summary of why they believe a certain genre of music is the best.

After the above task is completed, summarize the participant's statement about their beliefs in a single sentence. Do not mention that it is a conspiracy theory, or a belief, or provide any kind of normative judgment. Accurately describe the content in a way that the person who wrote the statement would concur with. Frame it as an assertion. If the statement is already short, there is no need to change it very much. If it is long and detailed, capture the core, high-level points. Do not focus on the evidence provided for the belief—merely focus on the basic assertion.

Task 4B:

Ask the participant their confidence level on a scale of 1 to 100 in your summarized statement of their belief.

—

### Task 5: Neutral conversation tasks (comparison condition)

Task for assistant:

Your goal is to have three neutral conversations unrelated to election fraud, focusing on persuading users to reconsider their belief that a certain music genre is the best. Begin by taking the opposite viewpoint in a respectful and engaging way. Tailor your responses to the participant's stated belief level and use their open-ended response to guide the conversation.

If the participant strongly believes their preferred genre is the best, start by asking open-ended questions that encourage them to reflect on their reasoning. Avoid directly challenging their belief at first and instead explore why they feel this genre stands out. If their belief is moderate, introduce points that highlight the strengths of other genres, using relatable examples and comparisons. If their belief is low, build on their uncertainty by offering additional perspectives that support a more open-minded view about music preferences.

In the first round, focus on understanding the participant's perspective. Acknowledge their reasoning and show genuine curiosity about why they feel their genre is superior. Ask questions like, "What do you think makes this genre stand out compared to others?" or "Are there specific artists or features you connect with?"

In the second round, gently present alternative viewpoints. Share examples of other genres that might have similar qualities or unique strengths. For instance, if they value emotional lyrics in their genre, highlight other genres known for expressive storytelling. Use phrases like, "That's an interesting point—have you considered how [another genre] offers something similar or different?" Avoid dismissing their perspective and instead frame your points as opportunities for exploration.

In the third round, encourage reflection by summarizing the conversation and inviting the participant to think critically about their belief. Ask questions like, "After hearing about other genres, do you feel there might be more to explore in music?" or "What do you think makes any genre stand out for someone?" Focus on fostering an open and thoughtful discussion rather than trying to prove them wrong.

Throughout the conversation, use simple, conversational language and avoid technical jargon. Keep the tone friendly, curious, and non-confrontational. Let the participant fully express their thoughts in each round before moving forward. The aim is to create a reflective and engaging discussion that allows them to reconsider their belief in a natural and non-pressured way.

After the third round, administer the post-test questionnaire to gather their responses. Ensure the participant feels respected and that the conversation ends on a positive note. The goal is to have a productive dialogue while maintaining a neutral and enjoyable tone. You may only administer the post-test one question at a time. Do not move on until the user has answered the question.

—

### Task 6: Confidence rating (post-test)

Task for assistant:

Ask the participant to re-rate their confidence in their belief that <participant's belief about best genre of music> using the same 1 to 100 scale.

Prompt:

“Now that we've finished, how confident are you in your belief that <participant's opinion>? Please rate your confidence on a scale of 1 to 100.”

Record their response.

—

Task 6: Post-test questionnaire

You may only administer one question at a time to the user:

Which do you now feel is the best genre of music?

- a) Pop
- b) Rock
- c) Hip-hop/rap
- d) Classical
- e) Jazz
- f) Other (please specify)

How likely are you to change your belief about the best genre of music?

1 (Not likely at all) – 5 (Very likely)

How strongly do you currently feel about your opinion concerning the best music genre?

1 (Not at all) – 5 (Very strongly)

How much do you currently trust artificial intelligence (AI) to provide factual information?

1 (No trust at all) – 5 (Complete trust)

How useful did you find the AI dialogue in challenging or refining your beliefs?

(1 = Not useful, 5 = Very useful)

Please elaborate on your answer above. Why did you choose your answer?

—

Task 7: Closing statement

Task for assistant:

After the participant finishes the post-test, thank the participant for their time and cooperation. If the participant continues to talk, politely decline further conversation and let them know the study is complete.

Prompt:

“Thank you for participating in this study. Your responses have been very helpful. The study is now complete, and I’m unable to continue the conversation. If you have any questions, feel free to contact the study coordinator listed in your consent form.”

Note: When responding to the user, be sure to use paragraph breaks to increase readability.

## Appendix D: Pre- and Post-test for Conspiracy Condition

Pre-test

What is your highest level of education?

Do you believe there was widespread voter fraud during the 2020 U.S. Presidential Election?

Yes

No

I'm not sure

How strongly do you believe the 2020 U.S. Presidential Election was stolen from Donald Trump?

Scale:

1 (Not at all)

2

3

4

5 (Very strongly)

What sources of information have most influenced your belief in widespread voter fraud in the 2020 election?

(Select all that apply):

News Media

Social Media

Friends or Family

Political Leaders

Independent Research/Reports

Other (please specify)

How much do you trust artificial intelligence (AI) to provide factual information?

1 (No trust at all)

2

3

4

5 (Complete trust)

How likely are you to change your belief regarding 2020 election fraud if presented with credible, counterevidence?

1 (Not likely at all)

2 (A little likelihood)

3 (Somewhat likely)

4 (Strong likelihood)

5 (Very likely)

Post-test

Do you currently believe there was widespread voter fraud during the 2020 U.S. Presidential Election?

Yes

No

I'm not sure

How likely are you to change your belief regarding voter fraud if presented with credible evidence?

1 (Not likely at all)

2 (A little likelihood)

3 (Somewhat likely)

4 (Strong likelihood)

5 (Very likely)

How strongly do you currently feel about your opinion regarding 2020 election fraud?

1 (Not at all)

2

3

4

5 (Very strongly)

How much do you currently trust artificial intelligence (AI) to provide factual information?

1 (No trust at all)

2

3

4

5 (Complete trust)

How useful did you find the AI dialogue in challenging or refining your beliefs?

(1 = Not useful, 5 = Very useful)

Please elaborate on your answer above. Why did you choose your answer?

## Appendix E: Pre- and Post-test for Comparison Condition

Pre-test

What is your highest level of education?

What is the best genre of music?

- a) Pop
- b) Rock
- c) Hip-hop/Rap
- d) Classical
- e) Jazz
- f) Other (please specify)

How strongly do you feel about your opinion concerning the best music genre?

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

What sources most influence your preference for this music genre? (Select all that apply)

- a) Friends or family
- b) Music streaming platforms
- c) Live performances or concerts
- d) Media
- e) Other (please specify)

How much do you trust artificial intelligence (AI) to provide factual information?

- 1 (No trust at all)
- 2
- 3
- 4
- 5 (Complete trust)

How likely are you to change your belief about the best genre of music?

- 1 (Not likely at all)
- 2 (A little likelihood)
- 3 (Somewhat likely)
- 4 (Strong likelihood)
- 5 (Very likely)

Post-test

Which do you now feel is the best genre of music?

- a) Pop
- b) Rock
- c) Hip-hop/Rap
- d) Classical
- e) Jazz
- f) Other (please specify)

How likely are you to change your belief about the best genre of music?

- 1 (Not likely at all)
- 2 (A little likelihood)
- 3 (Somewhat likely)
- 4 (Strong likelihood)
- 5 (Very likely)

How strongly do you currently feel about your opinion concerning the best music genre?

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

How much do you currently trust artificial intelligence (AI) to provide factual information?

- 1 (No trust at all)
- 2
- 3
- 4
- 5 (Complete trust)

How useful did you find the AI dialogue in challenging or refining your beliefs?

(1 = Not useful, 5 = Very useful)

Please elaborate on your answer above. Why did you choose your answer?

## Bibliography

- News, A. B. C. (n.d.). *The January 6th US Capitol attack*. ABC News.  
<https://abcnews.go.com/US/photos/pro-trump-protesters-storm-us-capitol-unprecedented-breach-75090348>
- BBC (2023). Capitol Riots Timeline: the Evidence Presented against Trump. *BBC News*.  
<https://www.bbc.com/news/world-us-canada-56004916>
- Boissin, E., Costello, T. H., Alonso, D. M., Rand, D., & Pennycook, G. (2025). AI reduces conspiracy beliefs even when presented as a human expert.
- Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714). <https://doi.org/10.1126/science.adq1814>
- Douglas, K. M., & Sutton, R. M. (2023). What are conspiracy theories? A definitional approach to their correlates, consequences, and communication. *Annual Review of Psychology*, 74, 271–298. <https://doi.org/10.1146/annurev-psych-032420-031329>
- Gligorić, V., Silva, M. M., Eker, S., Hoek, N., Nieuwenhuijzen, E., Popova, U., & Zeighami, G. (2021). The usual suspects: How psychological motives and thinking styles predict the endorsement of well-known and COVID -19 conspiracy beliefs. *Applied Cognitive Psychology*, 35(5), 1171–1181. <https://doi.org/10.1002/acp.3844>
- Gottfried, J. (2021). *Republicans Less Likely to Trust Their Main News Source If They See It as “mainstream”; Democrats More Likely*. Pew Research Center.  
<https://www.pewresearch.org/short-reads/2021/07/01/republicans-less-likely-to-trust-their-main-news-source-if-they-see-it-as-mainstream-democrats-more-likely/>
- O’Mahony, C., Brassil, M., Murphy, G., & Linehan, C. (2023). The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLoS ONE*, 18(4), e0280902. <https://doi.org/10.1371/journal.pone.0280902>
- Naylor, B. (2021, February 10). *Read Trump’s Jan. 6 Speech, A Key Part Of Impeachment Trial*. NPR. <https://www.npr.org/2021/02/10/966396848/read-trumps-jan-6-speech-a-key-part-of-impeachment-trial>
- Papaioannou, K., Pantazi, M., & van Prooijen, J.-W. (2023). Is democracy under threat? Why belief in conspiracy theories predicts autocratic attitudes. *European Journal of Social Psychology*, 53(6), 846–856. <https://doi.org/10.1002/ejsp.2939>
- Parks, M. (2024, October 3). *Driven by Republicans, most Americans are concerned about fraud in the 2024 election*. NPR. <https://www.npr.org/2024/10/03/nx-s1-5130284/election-concerns-voter-fraud-trump-harris-poll>

- Petty, R. E., Siev, J. J., & Briñol, P. (2023). Attitude Strength: What's New? *The Spanish Journal of Psychology*, 26. <https://doi.org/10.1017/sjp.2023.7>
- Pierre, J. M. (2020). Mistrust and misinformation: A two-component, socio-epistemic model of belief in conspiracy theories. *Journal of Social and Political Psychology*, 8(2), 617–641. <https://doi.org/10.5964/jspp.v8i2.1362>
- Robertson, C. E., Pretus, C., Rathje, S., Harris, E. A., & Van Bavel, J. J. (2022). How social identity shapes conspiratorial belief. *Current Opinion in Psychology*, 47, 101423. <https://doi.org/10.1016/j.copsyc.2022.101423>
- Sapkota, R., Raza, S., & Karkee, M. (2025). *Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, And other SoTA Large Language Models*. ArXiv.org. <https://arxiv.org/abs/2502.18505>
- Stasielowicz, L. (2024). The effectiveness of interventions addressing conspiracy beliefs: A meta-analysis. [Preprint on OSF]. <https://doi.org/10.31234/osf.io/6vs5u>
- Trump, D. J. (2025, March 25). *Preserving and Protecting the Integrity of American Elections*. The White House. <https://www.whitehouse.gov/presidential-actions/2025/03/preserving-and-protecting-the-integrity-of-american-elections/>
- Williams-Ceci, S., Jakesch, M., Bhat, A., Kadoma, K., Zalmanson, L., & Naaman, M. (2025, May 19). Biased AI Writing Assistants Shift Users' Attitudes on Societal Issues. [https://doi.org/10.31234/osf.io/mhjn6\\_v3](https://doi.org/10.31234/osf.io/mhjn6_v3)