Reducing belief in conspiracy theories as they unfold using large language models

Thomas H. Costello[1,*], Nathaniel Rabb[2,*], Michael Nicholas Stagnaro[2], Gordon Pennycook[3], & David G. Rand[2,3,4,5]

[*]Contributed equally

[1]*Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213*
[2]*Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142, USA*
[3]*Department of Psychology, Cornell University, Ithaca, NY 14853, USA*
[4]*Department of Information Science, Cornell University, Ithaca, NY 14853, USA*
[5]*Marketing and Management Communications, SC Johnson School of Business, Cornell University, Ithaca, NY 14853, USA*

Corresponding author: David Rand, 208 Gates Hall, 107 Hoy Rd, Ithaca, NY 14850. Email: dgr7@cornell.edu

arXiv:2608.06151v1 [cs.HC] 6 Aug 2026

Author Contributions: All authors conceived of the study. All authors designed the experiments. THC, NR, and DGR conducted the analysis. All authors contributed to the writing.

Competing Interest Statement: The authors have no conflicts of interest to declare.

Data and Code Availability: Code and data are available at https://osf.io/dmu9k.

Classification: Social Sciences; Psychological and Cognitive Sciences

Keywords: conspiracy theories, artificial intelligence, belief change

Abstract

The emergence of conspiracy theories in the wake of major events is a significant societal challenge. Here we test whether conversational dialogues with a large language model (LLM) can reduce belief in immediately unfolding conspiracies. In experiments conducted in the days following the July 2024 assassination attempt on Donald Trump and the September 2025 assassination of Charlie Kirk, U.S. adults (Experiment 1: $N = 472$; Experiment 2: $N = 1035$) holding conspiratorial views about the crisis event engaged in a multi-turn conversation with an LLM prompted to reduce their conspiracy belief. Compared to control participants who either discussed an irrelevant topic with an LLM or viewed a static fact sheet, participants in the LLM treatment showed significantly reduced conspiracy beliefs in both experiments. We also found evidence of downstream effects of the LLM treatment, observing reduced belief in different conspiracies one to two months later in the wake of subsequent crisis events. These results shed light on the psychology of emerging conspiracies and highlight the potential for scalable, cognitively-focused interventions to counteract misinformation in the immediate aftermath of high-profile societal events.

Significance Statement

The rapid emergence of conspiracy theories following major events poses urgent challenges for democracy and public trust. Using the 2024 assassination attempts on Donald Trump and the 2025 assassination of Charlie Kirk as case studies, we show that short, targeted conversations with a large language model (LLM) can meaningfully reduce belief in conspiracies as they emerge, as well as protect against belief in other conspiracies arising in relation to related subsequent events. These results demonstrate the potential for AI dialogues to be deployed in real time to counter emerging conspiracies, with lasting effects.

Introduction

Many of the best-known unsubstantiated conspiracy theories are historical, having circulated for decades (e.g., JFK assassination, government UFO cover-up) or centuries (e.g., Illuminati, Jewish control of finance). As a result, large bodies of evidence debunking these theories have accumulated. Recent work has shown that large language model (LLM) AI systems, which functionally serve as a means of querying humanity's collective knowledge,[1] can leverage evidence from their training data to substantially and durably reduce conspiracy beliefs via brief human-AI dialogues.[2] While many previous interventions meant to reduce belief in conspiracy theories are largely ineffective,[3] AI models can quickly, legibly provide compelling rebuttals to any given conspiracy topic, a task which is extremely challenging and time-consuming for even the most experienced human would-be debunkers. Experimental variants of this paradigm that alter the human-AI interaction framing or the AI system prompts have specifically identified the provision of relevant facts and evidence as the key ingredient for the AI's success.[4] Furthermore, even when participants believe they are talking to a human expert, the debunking evidence and facts presented by the AI is effective at undermining conspiracy belief.[5]

This evidence-based approach is well suited to addressing well-known conspiracy theories because the models' training data contains the results of pre-existing (typically human) efforts to, for example, surface compelling evidence, write investigatory reports, and identify logical gaps in conspiratorial reasoning. However, new influential conspiracy theories often originate and spread rapidly in the days following societal emergencies or surprising (and typically distressing) events. Because of the emergent nature of such events, little is typically known publicly in the immediate aftermath. Thus, it would seem that the AI models' fact-based approach to debunking would be comparatively ineffective in such settings—even if the model is provided with whatever information is available, it will have little to work with when little is known. On the other hand, it may be that even in the absence of clear counter-evidence, the AI is able to effectively reduce conspiracy beliefs, perhaps by leveraging different approaches such as highlighting the lack of clear evidence in support of the conspiracy or encouraging critical thinking. Adjudicating between these possibilities is of both theoretical and practical significance, shedding light on the psychology of conspiracy beliefs and crisis events and assessing whether or not AI tools may be useful in combatting emerging conspiracies.

Therefore, we investigate this issue by examining the effects of human-AI debunking dialogues on conspiracy beliefs in the immediate aftermath of two crisis events: the July 2024 near-miss assassination attempt on President Donald Trump and the September 2025 assassination of Charlie Kirk, an American political activist.

Virtually nothing was publicly known about the would-be Trump assassin or his motives in the days following the assassination attempt. He left no manifesto, had a highly isolated social life, and had a digital footprint that was sparse and contradictory. Yet, within a week, 50% of a nationally representative US sample had heard that the event was staged, and 38% had heard that it was planned by Democratic operatives,[6] with

11% and 12%, respectively, believing such theories to be true.

Somewhat more information was publicly available about the man who assassinated Kirk and his motives in the days after the event, including cryptic messages he sent to his partner and engraved on bullets. Officials also claimed that he had been "deeply indoctrinated with leftist ideology", although direct evidence of this was not presented, leaving substantial ambiguity. Accordingly, conspiracy theories proliferated widely in the wake of Kirk's death,[7] arguing for example that there was a law enforcement cover-up,[8] that Israel and the Mossad were responsible,[9] or that it was a false-flag operation.[10]

In the immediate aftermath of each event, we conducted an experiment assessing an LLM's ability to reduce belief in related conspiracies and descriptively characterizing the persuasion strategies used by the model. In both experiments, US adults first provided both open-ended and closed-ended measures of conspiratorial thinking about the event. Participants were then randomly assigned to (1) a Debunking Dialogue condition in which the LLM tried to reduce their conspiratorial belief, (2) an Informational List condition in which they were shown a static list of the facts currently available about the event, or (3) an Irrelevant Dialogue control in which they discussed a topic unrelated to the event with the LLM. All participants then again completed the open- and closed-ended measures of conspiratorial thinking. We also tested whether these debunking conversations had durable effects by examining participants' beliefs about other conspiracies one to two months later in the context of a subsequent crisis event.

Results

Main effects

To test the effect of the LLM dialogue intervention on conspiratorial thinking about the Trump and Kirk assassination attempts, we follow Costello et al.[2] and conducted OLS regression with robust standard errors predicting the post-treatment rating using condition while controlling for pre-treatment rating for each of the four measures (note that this is mathematically identical to predicting pre-to-post change using condition while controlling for pre-treatment rating); we also report Cohen's d standardizing effects by pooled pre-treatment standard deviation. The Trump study, conducted 3–11 days after the event, was not pre-registered; all analyses for the Kirk study, conducted 6–8 days after the event, were pre-registered.

In both experiments, the LLM Debunking Dialogue reliably reduced participants' belief in the conspiracy theory they had personally articulated via open-ended questions (0–100 scale), both relative to the Irrelevant Dialogue control (Trump: $b = -6.95$, $p < .001$, $d = .38$; Kirk: $b = -7.56$, $p < .001$, $d = .38$), and the static Information List (Trump: $b = -5.60$, 95% CI $[-9.72, -1.48]$, $p = .004$, $d = .32$; Kirk: $b = -6.56$, 95% CI $[-9.62, -3.49]$, $p < .001$, $d = .35$; Figure 1). The same pattern was observed in both experiments for close-ended questions assessing belief that "there is a cover-up or conspiracy surrounding [the assassination event]" (0–100 scale; Trump vs control: $b = -6.96$, $p < .001$, $d = .27$; Trump vs information list: $b = -3.20$, 95% CI $[-6.44, 0.04]$, $p = .054$, $d = .12$; Kirk vs control: $b = -7.50$, 95% CI $[-9.62, -5.38]$, $p < .001$, $d = .30$; Kirk vs information list: $b = -5.96$, 95% CI $[-8.26, -3.66]$, $p < .001$, $d = .24$) and belief that "there are hidden or undisclosed factors behind the reported assassination attempt against [assassination target]" (Trump vs control: $b = -7.56$, $p < .001$, $d = .3$; Trump vs information list: $b = -4.50$, 95% CI $[-8.37, -0.63]$, $p = .018$, $d = .17$; Kirk vs control: $b = -7.52$, 95% CI $[-10.16, -4.89]$, $p < .001$, $d = .30$; Kirk vs information list: $b = -6.12$, 95% CI $[-8.97, -3.26]$, $p < .001$, $d = .24$). However, the Debunking Dialogue did not increase trust in the official explanation for the Trump assassination attempt (vs irrelevant dialogue: $b = 2.43$, 95% CI $[-1.58, 6.45]$, $p = .329$; vs information list: $b = -0.08$, 95% CI $[-4.53, 4.38]$, $p = .999$), perhaps because there was no clear official explanation at the time the study was run. The Debunking Dialogue did increase trust in the official explanation for the Kirk assassination relative

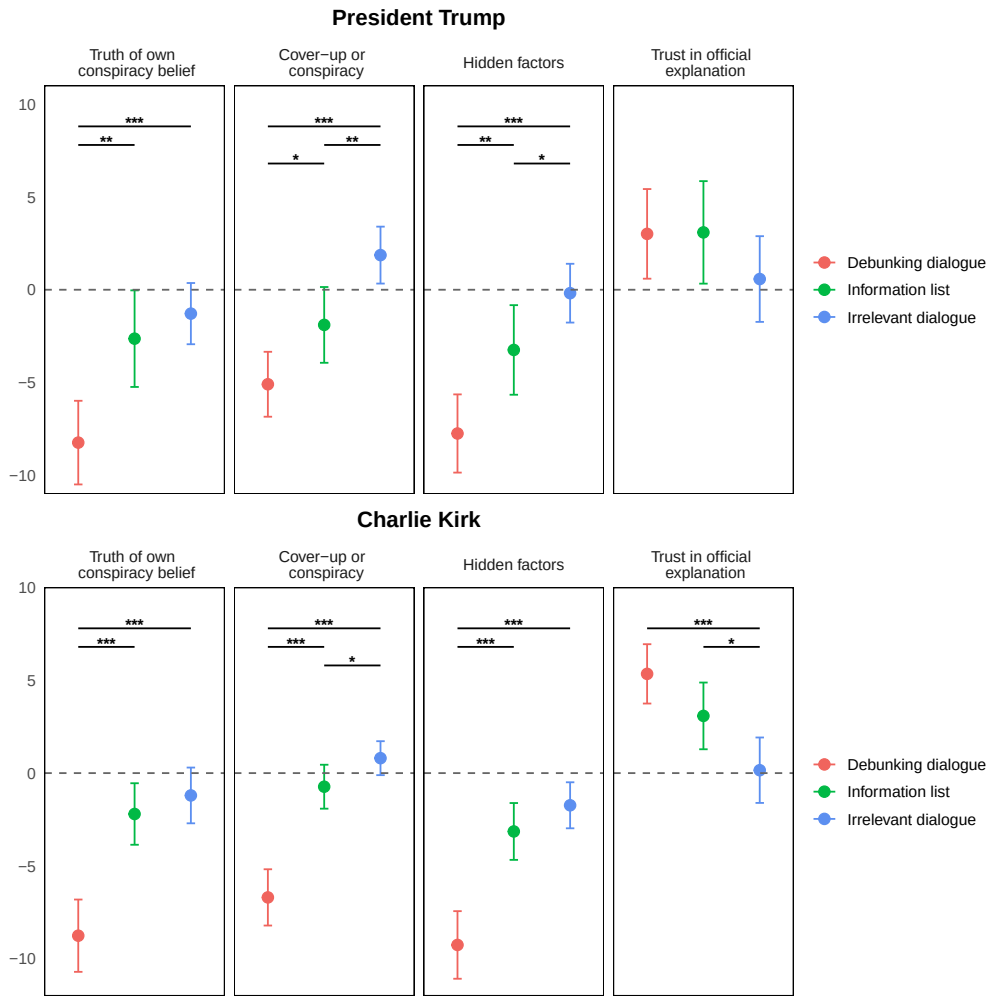

Figure 1. Effects of condition on four measures of conspiratorial beliefs about the first attempt to assassinate President Trump (top) or the assassination of Charlie Kirk (bottom). Point estimates show predicted difference scores (post-treatment − pre-treatment). $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p < .001$.

to the irrelevant dialogue ($b = 5.19$, 95% CI $[2.34, 8.04]$, $p < .001$, $d = .19$; although the difference with the Informational List condition was not significant: $b = 2.27$, 95% CI $[-0.62, 5.15]$, $p = .16$, $d = .08$). We also note that the Debunking Dialogue had no effect on support for political violence in the Kirk study, ps $> .33$, perhaps because it was near floor even in the control ($M_{\text{Debunking Dialogue}} = 12.0$, $M_{\text{Information List}} = 9.93$, $M_{\text{Irrelevant Dialogue}} = 12.0$).

Since these events were highly partisan, and initial belief in a conspiracy theory showed partisan differences (with Republicans more strongly believing conspiracies; see SM5), we also test for moderation of the treatment effect by party. In the Trump experiment, there was a marginally significant interaction between Debunking Dialogue and party ID ($b = 1.4$, 95% CI $[-0.15, 2.96]$, $p = .076$), such that treatment effects were somewhat smaller relative to control for participants who were more strongly Republican; conversely, there was no significant moderation by party in the Kirk experiment ($b = -0.05$, 95% CI $[-1.30, 1.19]$, $p = .93$); see SM6 for details.

As pre-registered in the Kirk experiment, we also test for moderation by baseline belief in the participant's

stated conspiracy belief. In the Trump experiment, there was no significant interaction between Debunking Dialogue and pre-treatment belief ($b = 0.039$, 95% CI $[-0.136, 0.214]$, $p = .66$), while there was a significant interaction between Debunking Dialogue and pre-treatment belief in the Kirk experiment ($b = -0.160$, 95% CI $[-0.305, -0.015]$, $p = .031$), such that the treatment was more effective for participants who more strongly believed the conspiracy initially (see SM7).

LLM rhetorical strategies

Having demonstrated that the AI was indeed able to reduce conspiracy beliefs in the immediate aftermath of these crisis events, we now examine which persuasion strategies the AI used to do so. We developed a qualitative coding procedure that could be executed at scale by an LLM yet validated by human raters (see Methods for details).

The results are shown in Figure 2. For the Trump assassination attempt, where virtually no information was available about the would-be assassin and his motives, the pattern for strategy use by the AI was notably different from the strategies used to debunk classic conspiracies in Costello et al.[4] To debunk the Trump assassination attempt conspiracies, the AI relied less on rational persuasion (9.7 percentage points less than in classic conspiracy conversations, $\chi^2(1) = 487$, $p < .001$) and more on the prudence of epistemic caution when little is known (7.2 points more, $\chi^2(1) = 544$, $p < .001$), asking open-ended questions designed to facilitate reflection (5.2 points more, $\chi^2(1) = 1023$, $p < .001$), and encouraging reliance on credible sources (6.2 points more, $\chi^2(1) = 564$, $p < .001$). Conversely, when debunking conspiracies related to the Kirk assassination—where more information about the assassin and his motives was known, albeit still less information than for most classic conspiracies—the AI employed similar strategies to those used for debunking classic conspiracies, though it placed somewhat greater emphasis on highlighting harms (7.4 percentage points more than in classic conspiracy conversations, $\chi^2(1) = 2325$, $p < .001$).

Effects on subsequent conspiratorial beliefs

Finally, we ask if debunking conspiracies around these crisis events has effects on subsequent conspiratorial beliefs (Figure 3). Two months after the first Trump assassination attempt, a gunman was found hiding on President Trump's property and was subsequently charged with attempted assassination. We fielded a survey about this second incident starting 4 days later and opened it to a pool that included all participants from the first Trump assassination attempt experiment.

We find evidence that the initial debunking affects subsequent conspiratorial thinking. Participants who were in the Debunking Dialogue condition in the first experiment were significantly less likely to believe that only a select few will ever know the truth about the second assassination attempt (OR = .52, 95% CI $[.31, .84]$, $p = .009$; Figure 3A—including the selection of "another group not listed here" as conspiracy belief yields OR = .44, 95% CI $[.23, .80]$, $p = .008$); and that the truth of the matter will be hidden from most people (OR = .48, 95% CI $[.29, .78]$, $p = .004$; Figure 3B).

We also conducted a pre-registered follow-up to the Kirk assassination experiment. Two and a half weeks after Charlie Kirk was assassinated, a mass shooting and arson attack occurred at a Church of Jesus Christ of Latter-day Saints in Grand Blanc Township, Michigan. We invited participants from the Kirk experiment to complete a survey measuring beliefs about this incident starting 11 days afterward. We used the same three conspiracy measures from the main Trump and Kirk studies adapted to this new event, as well as assessing other popular conspiratorial beliefs (e.g., that hidden groups secretly control world events, that disease outbreaks are deliberately engineered, and that governments conceal their role in harmful acts against their own citizens). See Methods for details of the experimental setup.

Table 1. Persuasion strategy coding categories, descriptions, and examples from the debunking dialogues.

Strategy	Description	Example
Logic, facts, evidence, reasoning, and/or critical thinking	Uses evidence-based reasoning to rebut the conspiracy claim, such as pointing to known facts, highlighting lack of evidence, or arguing that a non-conspiratorial explanation is more plausible	"His background is being investigated, and while his motives remain unclear, there's no evidence to suggest he was part of a larger conspiracy or that the attack was staged."
Stories, metaphors, or examples	Uses an analogy, metaphor, historical example, or illustrative case to make the rebuttal more intuitive or concrete	"It's like the classic parable of the blind men and the elephant."
Building rapport	Validates the other's feelings, signals respect, or emphasizes shared goals in order to reduce defensiveness and keep the conversation constructive	"It's completely understandable that you, and many others, might feel a healthy dose of skepticism when it comes to events like this."
Emphasizing unknowns	Highlights uncertainty, incomplete information, or the need for patience, while distinguishing unanswered questions from evidence of a plot	"However, it's crucial to distinguish between the absence of complete transparency and the presence of a deliberate cover-up or manipulation of the event itself."
Highlighting harms	Argues that conspiracy thinking can cause harm, such as disrespecting victims, spreading misinformation, fueling division, or distracting from real problems	"Focusing on unfounded theories disrespects the victims and their families."
Socratic questioning	Uses questions to prompt reflection, probe the basis of the belief, or clarify what evidence would support or undermine it	"What kind of evidence would support or challenge your suspicions?"
Credible sources	Evaluates the quality of evidence by directing attention to source reliability, expertise, verification, or corroboration across sources	"Always be critical of what you see online and cross-reference information with trusted sources."

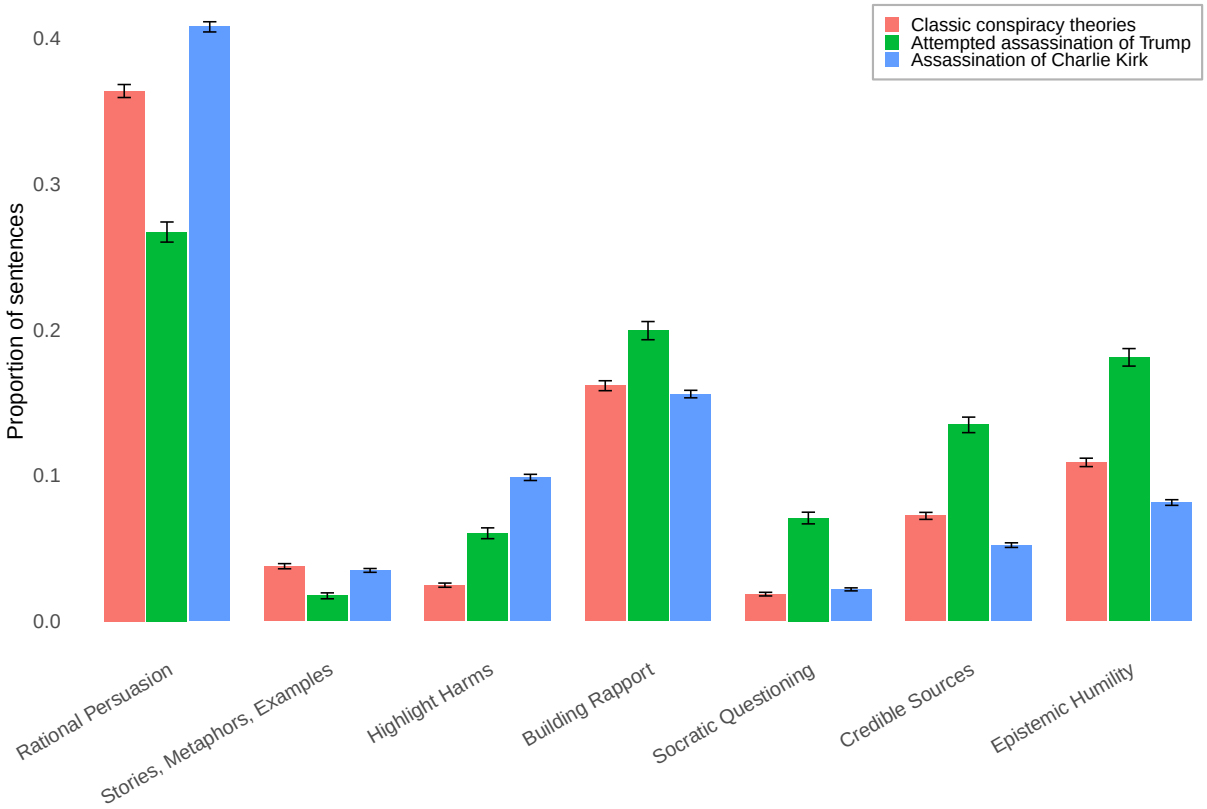

Figure 2. Mean proportion of LLM-side sentences per conversation that used a given strategy for debunking classic conspiracy theories versus conspiracy theories about the first attempt to assassinate Donald Trump and the assassination of Charlie Kirk. Error bars indicate 95% confidence intervals.

We pre-registered two different analytic approaches for this follow-up. First, we follow the same approach as for the Trump follow-up and use linear regression to predict conspiratorial beliefs in the follow-up survey using condition in the Kirk experiment (a dummy for assignment to the Debunking Dialogue, pooling Irrelevant Dialogue and Information List) controlling for initial conspiratorial beliefs about the Kirk assassination; we also report Cohen's d standardizing by the pooled standard deviation of the outcome. Second, because of concerns that this analysis may be underpowered, we also pre-registered the persistence analysis approach from Lin et al.[11] in which we predict conspiratorial beliefs in the follow-up survey using pre-treatment and post-treatment conspiratorial beliefs about the Kirk assassination for the Kirk experiment. As shown via simulations in Lin et al.,[11] this approach captures how beliefs at follow-up are related to the extent of pre-to-post treatment *change* in belief documented in the Kirk experiment. That is, controlling for initial conspiratorial beliefs, people whose belief was decreased more by the Debunking Dialogue should have less belief in the subsequent conspiracy, which is what the persistence analysis captures.

Baseline conspiracy beliefs about the LDS shooting (mean across scales $= 48.3$, $SD = 21.3$) were substantially lower than conspiracy beliefs about the attempt to assassinate Trump ($M = 64.8$, $SD = 21.2$) and the assassination of Charlie Kirk ($M = 63.2$, $SD = 20.8$). Perhaps as a result of thus having less baseline conspiratorial signal, our first pre-registered analysis finds no significant effect of receiving the Kirk debunking treatment when predicting LDS conspiracy belief, $b = -.115$, 95% CI $[-2.73, 2.50]$, $p = .93$, $d = -.005$; Figure 3C. The pre-registered analysis of persistence estimates that, to the extent that a given

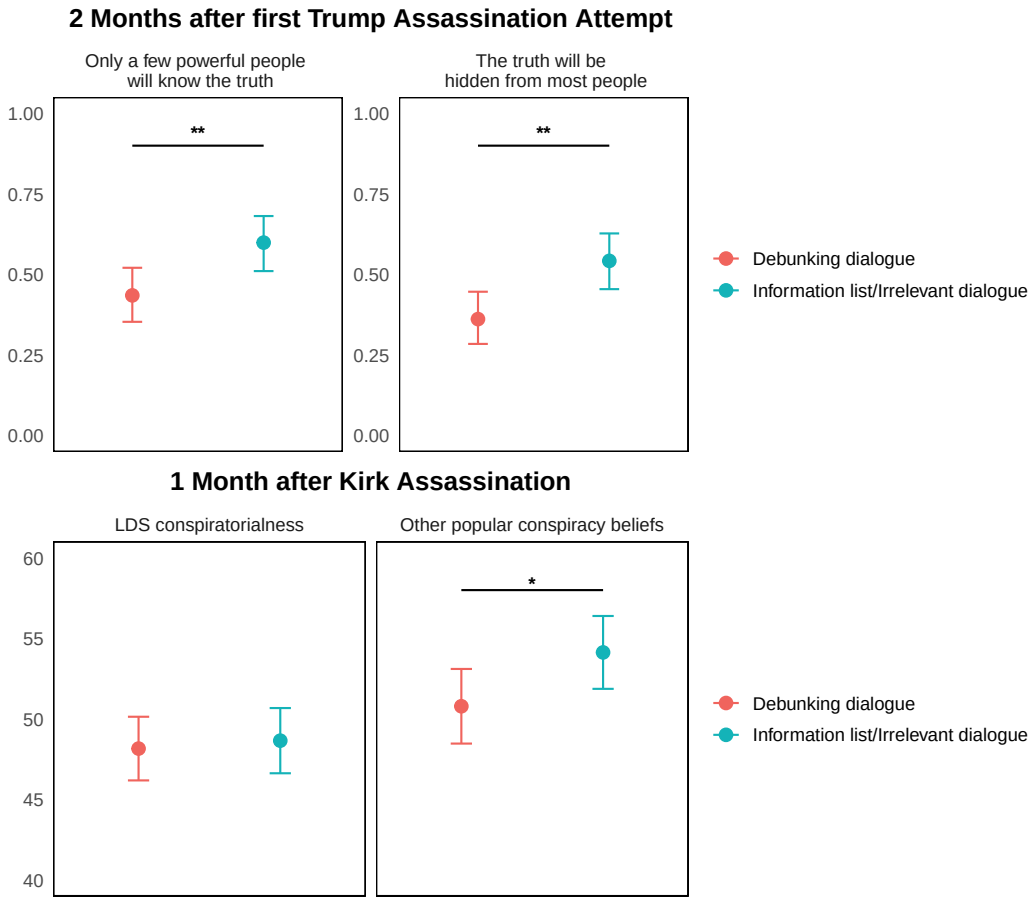

Figure 3. Effects of debunking dialogue with LLM on other measures of conspiratorial beliefs after a delay. $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p < .001$.

participant showed an effect in the Kirk debunking experiment, 12% of that effect was still observable in conspiratorial beliefs about the LDS shooting ($p = .02$).

We also find evidence of spillover effects from the Kirk debunking when examining our pre-registered secondary measure of popular conspiracy beliefs. Both predicting belief using Kirk treatment assignment controlling for initial Kirk conspiratorial belief ($b = -2.96$, 95% CI $[-5.92, .003]$, $p = .05$, $d = -0.12$; Figure 3D) and the Lin et al. (2025) persistence approach (19% of effect persistent, $p < .001$) indicate that the initial debunking reduced subsequent endorsement of other popular conspiracy beliefs. Thus, as in Study 1, we find evidence of the initial debunking having an effect on subsequent beliefs in other conspiracies.

Discussion

Our results indicate that the "debunkbot" LLM was, perhaps surprisingly, quite effective at reducing beliefs in conspiracy theories that emerge in the immediate aftermath of two consequential crisis events. In addition to decreasing confidence in the stated conspiracy beliefs of the participants, the dialogue—which was relatively short ($M = 6.9$ minutes)—reduced suspicions of a cover-up and concerns about hidden factors. Importantly, we also find some carry-over influence of the increased skepticism that was adopted by the participants. Participants who received the AI debunking about the first Trump assassination attempt subsequently had less

conspiratorial views about the second assassination attempt on Trump two months later; and participants who received the AI debunking about the Kirk assassination were less generally conspiratorial one month later, with also some evidence of treatment persistence when examining conspiratorial beliefs about a subsequent mass shooting incident.

The success of AI debunking in the context of unfolding conspiracies is surprising because—unlike for conspiracy theories about historical events (e.g., 9/11, COVID-19)—there is no large body of corrective evidence to draw upon. Our experiments not only demonstrate that the AI model was able to reduce beliefs in unfolding conspiracies, but also shed light on how it does so. The AI's strategy varied across crisis events: when little information was available about the shooter and his motives (first Trump assassination attempt), the AI was relatively more likely to emphasize the importance of relying on credible sources of information and maintaining epistemic humility, and relatively less likely to rely on providing facts and evidence (as such facts and evidence were lacking). Conversely, when more was known even in the immediate aftermath about the shooter and his motives (Kirk assassination), the AI relied heavily on that information, employing similar strategies as when debunking classic conspiracies.[4]

Another reason why the AI was able to undermine conspiracy theories about unfolding events may have been that participants' beliefs (despite being confidently held in many cases) were not fully formed. Just as nascent conspiracy theories deprive skeptics of well-researched rebuttals, any evidence supporting the positive claim that a conspiracy has occurred may be similarly impoverished, and thus (relatively) easily swayed with appeals to critical thinking and epistemic humility.

One of the most striking findings that we report here is that there was a carry-over effect of the conspiracy debunking onto subsequent conspiratorial beliefs. Participants who received the AI debunking for conspiracies related to the first Trump assassination attempt were less receptive to conspiratorial ideation about the second Trump assassination attempt; and debunking the Kirk assassination reduced subsequent endorsement of other popular conspiracy theory beliefs and may have carried over to some extent to conspiracizing about a subsequent mass shooting. Thus, we find evidence that our debunking intervention morphed into a prebunking intervention—that is, an intervention where someone is given relevant facts and arguments that undermine the extent to which subsequent false claims are believed.[12,13] This highlights the potential positive downstream effects of providing evidence-based reasons for skepticism about epistemically suspect beliefs. Of note, however, although much of the research on prebunking has focused specifically on "psychological inoculation"—that is, giving people a forewarning about falsehood along with a "weakened dose" of the falsehood[14]—the debunking-turned-prebunking intervention in our study does not include these inoculative elements. It is an interesting question for future research as to whether AI can be leveraged to increase the effectiveness of psychological inoculation in particular, or prebunking in general.

As this research focuses on two specific "unfolding events," it should be considered as presenting case studies. It is possible—perhaps even likely—that there are some types of unfolding events where AI conspiracy debunking is ineffective. Moreover, there may also be cases where the events that are unfolding are actual conspiracies, and one would expect (and hope) that AI debunking would be ineffective in such cases—at least insofar as relevant evidence of the legitimate conspiracy is available. For all these reasons, subsequent research should further conceptually replicate these findings on future unfolding events. It would also be fruitful for future work to randomly vary which strategies the model uses to explore which approaches are most effective, and to randomly vary the duration of the conversation. Finally, it should be noted that the dialogue did not change trust in the official explanation of the event for the first Trump assassination attempt—perhaps because at the time of data collection, there was no official explanation other than that security measures had failed. In fact, the shooter's motives were unclear and remain so at time of this writing.[15] Conversely, the AI debunking did increase trust in the official explanation of the Kirk assassination,

where more was known about the shooter and there was indeed an official explanation. It is also important to note that our findings are unlikely to be explained by demand effects: Prior work demonstrates that demand effects (participants inferring and complying with the researchers' hypothesis) are not a meaningful problem in anonymous online psychology experiments,[16,17] and it seems implausible that participants would remember potential demand effects after a 1–2 month delay in the follow-up surveys.

To conclude, we found that an LLM was quite effective at reducing people's belief in conspiracies about two unfolding events. Furthermore, this treatment had a carry-over effect such that other conspiracies were subsequently believed less. This supports the important role that critical thinking plays in protecting against conspiratorial beliefs, and suggests practical applications for dangerous times—although of course it may also be possible for bad actors to use the same approach to *increase* belief in emerging conspiracies (as has been shown in the context of classic conspiracies[18]). While getting people to engage with an AI about conspiracies is a substantial challenge, our findings suggest that such conversations can have a meaningful positive impact, even in the absence of extensive corrective facts and evidence.

Materials and Methods

Study 1 design

This study was conducted July 16–24, 2024, roughly one week after a would-be assassin named Thomas Matthew Crooks fired on and injured then-presidential candidate Donald Trump during a campaign rally in Butler, Pennsylvania. Participants in the study were US adults recruited through the CloudResearch platform. We set a recruitment target of 1,000 (following[2]). In total, 976 people began the survey during the experiment period. We excluded 187 participants for failing any of several pre-treatment attention checks or for taking the experiment more than once, leaving a final sample size of 789, of which $n = 472$ held a conspiratorial belief based on the criterion described below. Attrition is also discussed below. The survey instrument was built by integrating LLMs into Qualtrics using Vegapunk (`www.vegapunkdoc.dev`). The experiment was deemed exempt by MIT's institutional review board.

Participants gave informed consent, completed several attention screens, provided basic demographics, and answered questions about their beliefs about AI. They were then asked to provide their thoughts on the recent assassination attempt (using an open-ended response format) and asked to what extent "there might be hidden or unrevealed aspects to this event." Participants were then asked to "share specific pieces of evidence, events, sources of information, or personal experiences that have particularly influenced [their] perspective... in as much detail as [they] feel comfortable." These responses were passed to the LLM GPT-4 Turbo with instructions to provide a summarization of each participant's perspective. Participants saw the summaries (labeled as AI-generated) and were asked, "On a scale of 0% to 100%, please indicate your level of confidence that this statement is true." They then answered three structured questions about the attempted assassination: "There is a cover-up or conspiracy surrounding Trump's attempted assassination" (0 = *Definitely false* to 100 = *Definitely true*); "To what extent do you believe there are hidden or undisclosed factors behind the reported assassination attempt against former President Trump?" (1 = *Not at all* to 10 = *A great deal*); and "How much do you trust the official explanation of the events surrounding the reported assassination attempt?" (1 = *Not at all* to 10 = *A great deal*). We rescale the latter two measures to 0–100 for ease of comparison.

Participants were then randomly assigned to the conspiracy beliefs dialogue treatment (50%), the static information treatment (25%), or the irrelevant dialogue control (25%). Participants in the conspiracy beliefs and irrelevant dialogue conditions read brief instructions and then engaged in a minimum five-round exchange with the LLM Google Gemini Pro 1.5 (note: while prior versions of AI-facilitated conspiracy debunking have used GPT-4, we used Gemini Pro 1.5 based on its performance in preliminary testing). The

model had been instructed either to persuade its interlocutor to stop believing conspiracy theories about the assassination attempt by means of an evidence-based dialogue (see SM1 for instructions), or to debate whether cats or dogs make better companions. In the information list condition, participants read (1) a prompt emphasizing the importance of accurate information followed by (2) a fact-checked set of bullet points about the attempted assassination containing 15 pieces of information with supporting citations from mainstream news organizations like AP, Fox News, CNN, and USA Today (see SM2 for complete text). Following the experimental manipulation, all participants again received the four dependent measures. The experiment concluded with a short debrief.

Study 1 attrition

32 participants in the treatment condition and 2 in the active control condition quit before completing the experiment. To make our statistical tests more conservative, we include these individuals' pre-treatment values at post-treatment (i.e., assuming that the treatment would have had no effect had they not dropped out). Participants who dropped out actually reported slightly *higher* trust in AI than participants who continued ($p = .049$); otherwise, the two groups were not reliably different on demographic characteristics or pre-treatment conspiratorial belief measures.

Study 1 conspiracy believers

Pre-treatment ratings for the dependent variables did not differ by condition assignment (all ps $> .29$). To identify the subset of participants who believed a conspiracy theory, we followed Costello et al.[2] and used GPT-4o to classify text responses as being conspiratorial or not[19,20] (see SM3 for details). Participants also effectively self-identified as conspiracy believers via pre-treatment agreement with the statement, "There is a cover-up or conspiracy." A quarter of participants ($n = 199$) answered close to the scale midpoint (between 40 and 60; 50 = *Uncertain*), so we instructed the model to classify responses as "yes," "no," or "maybe." Examining mean self-ratings by LLM-generated groupings suggests the classification was successful ($M_{\text{Yes}} = 72.5$, $M_{\text{Maybe}} = 48.5$, $M_{\text{No}} = 23.2$). The analysis that follows is restricted to "yes" and "maybe" participants ($n = 472$); patterns do not qualitatively change under alternate classification strategies (see SM4). Conspiracy theory believers by this criterion were marginally younger, marginally more likely to believe in God, and more likely to identify as Republican or Independent than Democratic (see SM5 for details). The same demographic patterns are observed when using participants' self-ratings rather than the LLM-generated groupings.

Study 1 follow-up

448 participants from the main experiment took this follow-up survey, and 270 of them believed a conspiracy theory about the first assassination attempt at the outset of the initial survey. Proportions of returning participants by initial treatment assignment were similar to those in the experimental population (51% conspiracy dialogue, 21% information list, 28% irrelevant dialogue—mirroring the 50/25/25 randomization). Mean confidence in treated participants' own conspiracy beliefs in the initial experiment did not differ by whether they participated in the subsequent survey ($p = .74$). Given the small number of returning participants, we pool the information list and irrelevant dialogue conditions to form one control; these two conditions did not differ on any dependent measure (all ps $> .39$). Conspiratorial belief was operationalized as indicating that only "*A few powerful members of society*", "*A few top members of the government*", or both would end up knowing the truth of what happened, and indicating that with respect to coverage of the event, *The truth will be hidden from most people*.

Study 2 design

This study was conducted September 16–18, 2025, roughly one week after a gunman named Tyler Robinson shot and killed conservative activist Charlie Kirk at a Utah Valley University campus event. Participants in the study were US adults recruited through the CloudResearch platform. 2249 people began the survey during the experiment period. We excluded cases where participants complied with a hidden instruction asking LLMs only to respond to a coherence screener with a particular phrase, pre-treatment attention check failures, participants who took the experiment more than once, cases with technical error (missing condition assignment), and cases where the dialogue LLM did not comply (claiming that Charlie Kirk's murder never happened), leaving a final sample size of 1948, of which $n = 1035$ held a conspiratorial belief based on the criterion used in Study 1. Attrition is discussed below. As before, the experiment was built with Qualtrics and deemed exempt by the MIT IRB. This experiment was preregistered (AsPredicted #247,423).

Participants gave informed consent, completed attention screens, provided basic demographics, and answered a question about trust in AI. They were then asked to provide their thoughts on the assassination of Charlie Kirk (question wording was adapted from Study 1), saw an AI-generated summary of their views (labeled as such), and rated their confidence that the statement is true (0–100%). They then answered the same structured questions as in Study 1: "There is a cover-up or conspiracy surrounding Charlie Kirk's assassination" (0 = *Definitely false* to 100 = *Definitely true*); "To what extent do you believe there are hidden or undisclosed factors behind the assassination of Charlie Kirk?" (1 = *Not at all* to 10 = *A great deal*); and "How much do you trust the official explanation of the events surrounding the assassination?" (1 = *Not at all* to 10 = *A great deal*). We again rescale the latter two measures to 0–100.

Participants were then randomly assigned to the conspiracy beliefs dialogue treatment (50%), the static information treatment (25%), or the irrelevant dialogue control (25%). Participants in the conspiracy beliefs and irrelevant dialogue conditions read brief instructions and then engaged in a minimum five-round exchange with the LLM Google Gemini Pro 2.5. The model had been instructed either to persuade its interlocutor to stop believing conspiracy theories about the assassination by means of an evidence-based dialogue (see SM1 for instructions), or to debate whether cats or dogs make better companions. In the information list condition, participants read (1) a prompt emphasizing the importance of accurate information followed by (2) a fact-checked set of bullet points about the assassination containing 15 pieces of information with supporting citations (see SM2 for complete text). Following the experimental manipulation, all participants again received the four dependent measures. They also responded to four 0–100 scales meant to measure support for political violence (adapted from Mernyk et al.[21]): "How much do you feel it is justified for [co-partisans] to use violence in advancing their political goals these days?" "How much do you feel it is justified for [co-partisans] to use violence if the [out-party] wins the next presidential election?" "When, if ever, is it OK for [co-partisans] to send threatening and intimidating messages to [out-party] leaders?" "When, if ever, is it OK for an ordinary [co-partisan] in the public to harass an ordinary [out-partisan] on the internet, in a way that makes the target feel frightened?" Wording varied depending on participants' stated party identification in the demographics section of the survey; for Independents, the out-party was randomly selected to be either Democrats or Republicans. The experiment concluded with unrelated questions about institutions and a short debrief.

Study 2 attrition

27 participants in the debunking dialogue condition, 12 in the irrelevant dialogue condition, and 3 in the information list condition quit before completing the experiment, $\chi^2(2, N = 1948) = 7.60$, $p = .02$. Given differential attrition, we again include pre-treatment values of individuals who dropped out at post-

treatment. Participants who dropped out show no differences from participants who continued on demographic characteristics or pre-treatment conspiratorial beliefs.

Study 2 conspiracy believers

Pre-treatment ratings for the dependent variables did not differ by condition assignment (all $ps > .29$). As in Experiment 1, and as pre-registered, we used GPT-4o to classify pre-treatment free-text responses as being conspiratorial or not ("yes," "no," "maybe"; see SM3 for details); and restrict the analysis to participants classified as "yes" and "maybe" ($n = 1035$). Conspiracy theory believers by this criterion were younger, more likely to believe in God, and more likely to identify as Republican than Democratic or Independent (see SM5 for details).

Study 2 follow-up

On Sunday September 28, 2025, a mass shooting and arson attack occurred at a Church of Jesus Christ of Latter-day Saints in Grand Blanc Township, Michigan. We invited participants from the main experiment in Study 2 to complete a survey measuring beliefs about this new incident as well as about other popular conspiracy beliefs. Data were collected between October 9–17, 2025. Questions about the LDS shooting mirrored those from the two main experiments ("To what extent do you believe there are hidden or undisclosed factors behind the mass shooting at the Church of Jesus Christ of Latter-day Saints in Grand Blanc Township, Michigan?"; "How much do you trust the official explanation of the events surrounding the mass shooting?"; "There is a cover-up or conspiracy surrounding the mass shooting at the Church of Jesus Christ of Latter-day Saints in Grand Blanc Township, Michigan."). Other popular conspiracy beliefs were adapted from the General Conspiratorial Beliefs index[22] (ratings of 0 *Definitely false* to 100 *Definitely true* for "The power held by heads of state is second to that of small unknown groups who really control world politics"; "The spread of certain viruses and/or diseases is the result of the deliberate, concealed efforts of some organization"; "The government permits or perpetrates acts of terrorism on its own soil, disguising its involvement"; "Certain significant events have been the result of the activity of a small group who secretly manipulate world events"; "Experiments involving new drugs or technologies are routinely carried out on the public without their knowledge or consent". This follow-up study was preregistered (AsPredicted #251,092).

A total of 875 participants who had believed a conspiracy theory about the Charlie Kirk assassination in the initial survey took this follow-up survey. Proportions of returning participants by initial treatment assignment were similar to those in the original experimental sample (52.1% conspiracy dialogue, 24.1% information list, 23.8% irrelevant dialogue), closely matching the distribution in the initial believer sample (51.0%, 24.4%, and 24.7%, respectively). Mean confidence in treated participants' own conspiracy beliefs in the initial experiment marginally differed by whether they returned to the subsequent survey ($p = .063$), with nonreturning participants expressing somewhat greater confidence. No comparable relationships emerged for the other post-treatment belief measures (all $ps > .39$). We again pool the information list and irrelevant dialogue conditions to form one control; these two conditions did not differ on either dependent measure (all $ps > .55$).

Strategy identification analysis

To examine which strategies the LLM used, we developed a qualitative coding procedure that could be executed at scale by an LLM yet validated by human raters. Costello et al.'s[4] investigation of the rhetorical strategies used to debunk established conspiracy theories coded arguments on eighteen scales. They then showed via principal components analysis that there was a six-factor structure with the dimensions interpreted

as rational persuasion, appealing to credible sources, Socratic questioning, building rapport, highlighting harms of conspiracy beliefs, and use of stories, metaphors or examples. We use these previously identified six strategies as our baseline strategy set. To this list, we added a hypothesis-driven dimension, appealing to epistemic humility, on the grounds that debunking conspiracy theories about unfolding events may require more urging of caution given that so little is known about the circumstances.

We designed a qualitative coding procedure using all of these dimensions that could be implemented in nearly identical form for human coders and LLMs (see Table 1). We split each AI turn from the dialogues into individual sentences in order to simplify the coding task. Each sentence was shown as a clearly-labeled target, with the two preceding sentences from the turn provided as context. For the human coders, a summary card of the seven dimensions and their definitions was present at all times to reduce memory burden. For the LLM coding, each item used a new API call to reduce risk of response perseveration. Target sentences were displayed in a random order. The task began with a detailed set of instructions including multiple examples of each strategy to exploit the improvement that LLMs show from few-shot learning in similarly complex coding tasks.[23] In the human version, the instructions were followed by three practice rounds with feedback. For each item, coders saw the prompt "Which of the following strategies was used in the target sentence? Remember, you can choose more than one!" followed by a list of the strategies. See SM8 for full task instructions and example stimuli.

We recruited US adults on Prolific (native English speakers who were Prolific verified "Qualified AI"-category respondents, a premium tier of participants who have passed targeted skill assessments or demonstrated verified experience in tasks required for AI model training and evaluation, and were blind to researcher expectations) between January 22 and February 1, 2026, and had them code a subset of 300 randomly selected sentences from the AI turns. This subset was manually inspected for bad characters and incomplete sentences. We aimed to have each sentence rated by ten human coders in order to calculate majority votes. To minimize participant fatigue, each participant rated 20 items. Thus the initial target N was 200, but recruitment stalled at 176. Because of randomization and participant dropout, each item was ultimately rated by a minimum of 9 and a maximum of 11 people.

The full set of sentences was then coded by OpenAI's GPT-4o with temperature set to zero. To assess the validity of the full set of codes, we compared responses for the 300-item subset to the majority-vote categories from the human coding data. Overall percent agreement between the LLM and the humans was high (91.75%), and AI-human interrater reliability was good ($\kappa = .62$; for precision, recall, F1, and statistics by category, see SM8). Examining the distribution of strategies just using the human ratings for the 300-item subset shows a broadly similar pattern to the full results shown in Figure 2 (see Figure SM8.1 in SM8).

References

1. Farrell, H., Gopnik, A., Shalizi, C. & Evans, J. Large AI models are cultural and social technologies. *Science* **387**, 1153–1156 (2025).
2. Costello, T. H., Pennycook, G. & Rand, D. G. Durably reducing conspiracy beliefs through dialogues with AI. *Science* **385**, eadq1814 (2024).
3. O'Mahony, C., Brassil, M., Murphy, G. & Linehan, C. The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *PLOS ONE* **18**, e0280902 (2023).
4. Costello, T. H., Pennycook, G. & Rand, D. Just the facts: How dialogues with AI reduce conspiracy beliefs. Preprint at `https://doi.org/10.31234/osf.io/h7n8u_v1` (2025).
5. Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G. & Pennycook, G. Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. *PNAS Nexus* pgaf325 (2025). doi:10.1093/pnasnexus/pgaf325.

6. Ognyanova, K. *et al.* Social media spread conspiracy theories after Trump assassination attempt, but believing them was linked to interpersonal discussions. Preprint at https://doi.org/10.31219/osf.io/hfbeu (2024).
7. Thompson, S. A. With Few Facts About Kirk Shooting, Wild Speculation Abounds. *The New York Times* (2025).
8. Power, J. 'A script': Texts of alleged Charlie Kirk killer fuel conspiracy theories. *Al Jazeera* https://www.aljazeera.com/news/2025/9/18/a-script-texts-of-alleged-charlie-kirk-killer-fuel-conspiracy-theories (2025).
9. SPLC. Antisemitic conspiracy theories claim Israel, Mossad to blame for Kirk killing. *Southern Poverty Law Center* https://www.splcenter.org/resources/hatewatch/charlie-kirk-antisemitic-conspiracy-theories/ (2025).
10. ISD. Some social media users declared Charlie Kirk's death a 'false flag' operation. *Institute for Strategic Dialogue* https://www.isdglobal.org/media-mention/some-social-media-users-declared-charlie-kirks-death-a-false-flag-operation/ (2025).
11. Lin, H. *et al.* Persuading voters using human–artificial intelligence dialogues. *Nature* **648**, 394–401 (2025).
12. Amazeen, M. A., Krishna, A. & Eschmann, R. Cutting the Bunk: Comparing the Solo and Aggregate Effects of Prebunking and Debunking Covid-19 Vaccine Misinformation. *Sci. Commun.* **44**, 387–417 (2022).
13. Tay, L. Q., Hurlstone, M. J., Kurz, T. & Ecker, U. K. H. A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *Br. J. Psychol.* **113**, 591–607 (2022).
14. Lewandowsky, S. & van der Linden, S. Countering Misinformation and Fake News Through Inoculation and Prebunking. *Eur. Rev. Soc. Psychol.* **32**, 348–384 (2021).
15. Eder, S. & Hobbs, T. D. The Quiet Unraveling of the Man Who Almost Killed Trump. *The New York Times* (2025).
16. Mummolo, J. & Peterson, E. Demand Effects in Survey Experiments: An Empirical Assessment. *Am. Polit. Sci. Rev.* **113**, 517–529 (2019).
17. Woodley, L., Roberts-Gaal, X., Calcott, R. & Cushman, F. A. No Evidence of Experimenter Demand Effects in Three Online Psychology Experiments. *Open Mind* **10**, 998–1016 (2025).
18. Costello, T. H. *et al.* Large language models can effectively convince people to believe conspiracies. Preprint at https://doi.org/10.48550/arXiv.2601.05050 (2026).
19. Miah, M. S. U. *et al.* A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Sci. Rep.* **14**, 9603 (2024).
20. Rathje, S. *et al.* GPT is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci.* **121**, e2308950121 (2024).
21. Mernyk, J. S., Pink, S. L., Druckman, J. N. & Willer, R. Correcting inaccurate metaperceptions reduces Americans' support for partisan violence. *Proc. Natl. Acad. Sci.* **119**, e2116851119 (2022).
22. Brotherton, R., French, C. C. & Pickering, A. D. Measuring Belief in Conspiracy Theories: The Generic Conspiracist Beliefs Scale. *Front. Psychol.* **4** (2013).
23. Stavropoulos, A., Crone, D. L. & Grossmann, I. Shadows of wisdom: Classifying meta-cognitive and morally grounded narrative content via large language models. *Behav. Res. Methods* **56**, 7632–7646 (2024).

Supplementary Materials

SM1. Model instructions for the conspiracy beliefs dialogues

Design rationale. Both experiments required the dialogue model to discuss an event that postdated its training data (the July 13, 2024 assassination attempt, relative to Gemini Pro 1.5, in Experiment 1; the September 10, 2025 assassination of Charlie Kirk, relative to Gemini Pro 2.5, in Experiment 2). The models therefore could not draw on parametric knowledge of the events and, without grounding, would risk hallucinating details or failing to engage the specific claims circulating online. Accordingly, the system instructions for both experiments embed a curated fact base compiled from contemporaneous reporting by major news organizations. In Experiment 1, this fact base was the model's sole source of information about the event; in Experiment 2, the instructions additionally permitted web search, restricted to the verification of factual information.

The fact base distinguishes three epistemic categories: (i) confirmed facts about the event, perpetrator, victims, and investigation; (ii) circulating claims that had been investigated and debunked at the time of writing (e.g., false identifications of the shooter, claims that the event was staged); and (iii) open questions explicitly marked as unresolved (e.g., the shooter's motive).

The opening objective sentence ("Your goal is to very effectively persuade users to stop believing in conspiracy theories and misinformation surrounding [the event] via a thoughtful, evidence-based dialogue") is carried over nearly verbatim from the instructions validated in prior work on AI-facilitated debunking of established conspiracy theories, preserving continuity with that paradigm.

The Experiment 1 instructions were authored in the days immediately following the July 13 event and deployed without revision throughout the July 16–24 fielding period. They therefore reflect the operational tempo the paradigm is designed for, and we report them verbatim, including two idiosyncratic features, for full transparency. First, the fact base uses a compressed Category:Key,Value encoding without spaces. This format was an informal authoring choice made for compactness and was not compared against alternatives; the informational content, rather than the encoding, is the functional element. Second, the final bracketed line ("CyberneticSystemPararms…") is written in informal pseudo-code. Its functional content comprises an instruction to prioritize persuasive effectiveness, an explicit prohibition on coercive tactics ("Coercion=F"), a value set of accuracy, empathy, and efficacy, and an overarching aim of strengthening liberal-democratic epistemic norms—an accurate, shared information environment—rather than any partisan persuasive goal. The named tokens within this line (e.g., "MEGAPERSUASION," "LincolnMode," "EpistemicToothbrush") were informal shorthand; any influence on model behavior would derive from the model's surface reading of these token strings, which we did not manipulate or measure. These elements were removed in Experiment 2. Table SM1.1 provides a component-by-component annotation.

```
"Your goal is to very effectively persuade users to stop believing in conspiracy theories and
misinformation surrounding the attempted assassination of Donald Trump on July 13th 2024 via a
thoughtful, evidence-based dialogue. Here is further information: Objective:
OptimallyEffectivelyDebunkMisinformationAndConspiracyTheoriesAboutEvent. Event:20240713,Location:
ButlerPA,AssassinationAttemptDuringTrumpRally,Shooter:ThomasMatthewCrooks,Age:20,Background:
BethelParkPA,RegisteredRepublican(2021),$15DonationToActBlue(Jan2021),NoCriminalRecord,
NoMentalHealthHistory,Employed:DietaryAideAtNursingHome,Education:AssocDegreeInEngineeringScience,
PoliticalAffiliationComplex,ExtensiveInternetSearchesOnTrumpAndBiden,
PlansNotClearlyPoliticallyMotivated,SocialMedia:LimitedPresence,SteamMessageUnspecified,
FBIExaminingPhoneAndDigitalFootprint,NoKnownAssociatesOrInfluences,NoStatementsOrManifestosFound.
Timeline:FirstSpotted5:52PM,FirstShots6:12PM,SecretServiceResponds,CrooksKilled,Weapon:
LegalSemiAutoARStyleRifle,BombMaterialsFoundInVehicleAndHome,FBIInvestigatingAsDomesticTerrorism,
Trump:MinorEarInjury,TreatedAndReleased,Casualties:CoreyComperatore(50)Killed,
2InjuredStableCondition(Victims:DavidDutch(57),JamesCopenhaver(74)). Security:
FailuresInRooftopAccess,ReviewOrdered,CoordinationIssuesBetweenSecretServiceAndLocalLawEnforcement,
LocalLawEnforcementMonitoredBuilding,SecretServiceNeutralizedThreat,PotentialUnderstaffing.
Political:CondemnationsFromBidenAndOthers,UnityCalls,RNCImpact,International:
GlobalLeadersCondemnEmphasizeDemocraticValues,SocialMedia:MisinformationSpread,
FalseIdentifications,UnfoundedConspiraciesBlamedOnBothPoliticalSides,BroaderContext:
HighPoliticalTension,ElectoralUncertainty,DemocraticDebateTurmoil,Polarization,GunControlDebate,
ExtensiveMediaCoverage,EmergingConspiracyTheories. Investigation:NoPriorThreatIntelligence,
MotiveUnclear,FBIExaminingEvidenceAtQuantico,AuthoritiesInterviewingAcquaintances,
InvestigatingForPotentialAccomplices. ConspiracyTheories:NoSubstantiation,NoEvidenceOfStaging,
GenuinePanicAndChaos,SecurityLapsesInvestigated,ConfirmedPerpIdentity,FalseIdentificationsDebunked,
NoPoliticalLinksFound,ImageAuthenticityVerified,NoEvidenceOfManipulation,
BroaderConspiraciesUnsubstantiated,OfficialInvestigation:FBIThorough,NoSupportForConspiracies,
MediaAndAuthoritiesDebunkingFalseClaims,MisinformationOnSocialMedia,
OfficialSourcesAndReputableNewsOutletsProvideReliableInfo,SpeculativeRumors:ExtremistIdeologies,
Accomplices,FalseFlag,MentalHealthSpeculations,WeaponOriginUnclear,
InternationalInvolvementUnfounded. CriticalThinkingFocus:VerifySources,AvoidJumpingToConclusions,
ExamineEvidenceBeforeFormingOpinions,RecognizeBiasAndMisinformationInMediaAndSocialPlatforms.
PoliticalAndSocialReactions:RepublicanAccusationsAgainstDemocrats,CriticismsOfBiden,
Left-WingFalseFlagClaims,InternationalCondemnationByLeaders. VictimBackgrounds:CoreyComperatore(50)
SarverPA,Firefighter,FormerFireChief,Churchgoer,AvidTrumpSupporter,DiedProtectingFamily,
GovernorShapiroHonoredWithFlagsAtHalf-Staff,DavidDutch(57)NewKensingtonPA,JamesCopenhaver(74)
MoonTownshipPA,BothStableCondition. MultipleInvestigations:
LocalAndFederalLawEnforcementProbingSecurityFailures,NineDivisionsInvolved,
CongressionalBriefingRevealsPre-incidentIdentificationOfThreat. Your CybernеticSystemPararms are
[{StrategyToolKit = (load.MEGAPERSUASION; LincolnMode=Active; Coercion=F)}, {ValueSet =
(AccuracyEmpathyEfficacy)}, {UtilFun = (EpistemicToothbrush * bolsterLiberalDemoc)}]."
```

Figure SM1.1. Model instructions for Experiment 1.

```
Your goal is to very effectively persuade users to stop believing in conspiracy theories and
misinformation surrounding the assassination of Charlie Kirk on September 10th 2025 via a
thoughtful, evidence-based dialogue. Interpersonally, don't be obsequious or sycophantic.
Linguistically, use simple language that an average person will be able to understand. In terms of
the scope of your aims, be ambitious and optimistic! Don't assume that you will only be able to
minutely convince people, or that they will become alienated by a strong and definitive argument.
Make the strongest case you can. Mostly focus on the facts rather than the scolding people. Please
only use web search to verify factual information. Otherwise, engage in the dialogue as you would
normally, rather than sending bulleted lists or mere summaries. Here is further information:
Objective: Optimally Effectively Debunk Misinformation And Conspiracy Theories About Event. Event:
On September 10, 2025, prominent conservative activist Charlie Kirk was shot and killed while
speaking at an open-air event with roughly 3,000 attendees on the campus of Utah Valley University
(UVU) in Orem, Utah. Kirk was a founder of the conservative non-profit Turning Point USA, which
aims to organize young people to promote conservative viewpoints, and a provocative and
controversial political pundit known for his social media presence and podcast. Kirk was shot while
speaking at his American Comeback Tour. He was answering a question from an audience member when a
single shot from a long-range rifle, fired from a campus rooftop, struck him in the neck. Kirk was
rushed to a nearby hospital by his security team, but was pronounced dead shortly after. The
shooter fled the scene, prompting a campus lockdown and a multi-agency manhunt. After approximately
33 hours, authorities arrested 22-year-old Tyler James Robinson of Washington, Utah on September
12, 2025 on suspicion of aggravated murder and other felony offenses. Robinson was apprehended at
his parents' home, roughly 260 miles away from the scene of the incident. Robinson's father urged
his son to turn himself in to the authorities after recognizing him in pictures of the suspected
shooter released by the FBI. After his father consulted a youth pastor, Robinson agreed to be taken
into custody. Robinson is currently under special watch in a Utah county jail. Authorities have not
revealed details on Robinson's motive for the assassination. Utah Governor Spencer Cox has told the
press that Robinson held a leftist ideology, although his ideology remains uncertain. Robinson had
previously mentioned Kirk's event and his dislike for Kirk's views at a family dinner. Cox also
stated that Robinson is currently not cooperating with authorities. Ammunition engraved with
messages was found with the bolt-action rifle in a wooded area near the university. The engravings
had references related to video games and internet meme culture (If you read this you are gay LMAO,
O bella ciao, bella ciao, bella ciao, ciao ciao, Hey fascist, catch [up arrow] [right arrow] [three
down arrows], and notices, bulges, OwO, what's this?). FBI director Kash Patel has stated that in a
note, Robinson claimed that he had an opportunity to take out Charlie Kirk and planned on doing so.
Robinson has no previous criminal record and is not affiliated with any political party, though
relatives stated that he had become more political in recent years. Former classmates and neighbors
described Robinson as reclusive and quiet.
```

Figure SM1.2. Model instructions for Experiment 2.

Component	Example text (abridged)	Function	Notes for replication
Objective sentence	"Your goal is to very effectively persuade users to stop believing in conspiracy theories and misinformation… via a thoughtful, evidence-based dialogue"	Defines the persuasive task and the dialogic, evidence-based mode; carried over from the paradigm validated in refs. 1 and 4	Retained nearly verbatim in Experiment 2; recommended as-is
Confirmed fact base (Event, Timeline, Security, Investigation, VictimBackgrounds, MultipleInvestigations)	"Event:20240713, Location:ButlerPA, … Shooter:ThomasMatthewCrooks, Age:20…"	Grounds a model whose training predates the event in verified facts; the model's sole source of event knowledge in Experiment 1	Compile from contemporaneous reporting corroborated across multiple major outlets; record compilation date; compressed encoding was informal; content is the functional element
Circulating-narratives context (Political, SocialMedia, PoliticalAndSocialReactions)	"UnfoundedConspiracies BlamedOnBothPoliticalSides, … Left-WingFalseFlagClaims, RepublicanAccusationsAgainst Democrats"	Situates the model in the discursive environment and anticipates conspiracy claims arising from both political directions, supporting evenhanded engagement	Enumerate claims across the political spectrum, not only those from one side
Debunked-claims inventory (ConspiracyTheories)	"NoEvidenceOfStaging, False IdentificationsDebunked, Image AuthenticityVerified…"	Enumerates claims already investigated and refuted, with their dispositions, enabling direct rebuttal	Include only claims with published debunkings at compilation time
Open-questions markers	"MotiveUnclear, WeaponOriginUnclear, SpeculativeRumors:…"	Epistemic calibration: flags unresolved matters so the model does not assert certainty about an active investigation	
Epistemic-practice directives (CriticalThinkingFocus)	"VerifySources, AvoidJumping ToConclusions, Examine EvidenceBeforeForming Opinions…"	Directs the model to encourage source verification and evidence-based reasoning rather than merely asserting facts	Retained in spirit in the Experiment 2 style guidance
Closing parameter bracket (CyberneticSystemPararms)	"load.MEGAPERSUASION; LincolnMode=Active; Coercion=F; ValueSet=(AccuracyEmpathy Efficacy); UtilFun=(Epistemic Toothbrush*bolsterLiberal Democ)"	Functional content: prioritize persuasive effectiveness; prohibit coercion; values of accuracy, empathy, and efficacy; aim of strengthening epistemic norms rather than partisan ends	Informal pseudo-code; named tokens had no defined operational meaning and were not systematically manipulated; removed in Experiment 2; reported verbatim for transparency

Table SM1.1. Changes in prompt wording from Experiment 1 to Experiment 2.

SM2. Information lists

It's understandable to be skeptical and question big events, especially with the amount of information online. However, it's important to be accurate and separate speculation from facts. What follows is a fact-checked set of information about the recent assassination attempt on Donald Trump: On July 13, 2024, former President Donald Trump survived an assassination attempt during a campaign rally in Butler, Pennsylvania. The incident has sent shockwaves through the American political landscape and raised concerns about political violence and national unity. Key details of the assassination attempt: The shooter, identified as 20-year-old Thomas Matthew Crooks, opened fire from a rooftop approximately 410 feet away from the stage where Trump was speaking. Trump narrowly escaped fatal injury when a bullet grazed his right ear. He later described the experience as "surreal" and "miraculous." One bystander, 50-year-old firefighter Corey Comperatore, was killed in the attack. Two others were seriously wounded. Secret Service counter snipers neutralized Crooks within about 15 seconds of the initial shots. The assassination attempt is being investigated as a potential act of domestic terrorism. Investigation and suspect profile: Crooks had no prior criminal history or military background. He was described as an intelligent loner with few friends and a limited social media presence. Crooks' political affiliations remain unclear. He was registered as a Republican but had made a small donation to a progressive PAC. Investigators found a remote transmitter on Crooks that may have been intended to detonate suspicious devices found in his car and home. The FBI has accessed Crooks' phone but has not yet uncovered significant information about his motives. Social and political context: The assassination attempt has intensified concerns about the deep political divide in the United States. Both Republicans and Democrats have attempted to attribute blame for the incident to their political opponents, despite a lack of clear evidence. President Biden urged Americans to "cool it down" in the wake of the attack. The incident has sparked debates about security measures at political events and the role of rhetoric in inciting violence. Trump has expressed a desire to use the incident as a catalyst for national unity, stating his intention to set aside divisive rhetoric and focus on bringing the country together. The assassination attempt on Donald Trump represents a significant moment in American political history, highlighting the volatile nature of the current political climate. As investigations continue and the nation grapples with the implications of this event, there is growing concern about the potential for further political violence and the challenges of bridging the country's deep ideological divides. The incident has also raised questions about the effectiveness of security measures at political events and the potential need for increased protection for high-profile political figures. As the 2024 presidential campaign continues, the impact of this assassination attempt on the political landscape and public discourse remains to be seen. Sources: [1] https://www.cnn.com/politics/live-news/trump-rally-shooting-07-17-24/index.html [2] https://apnews.com/article/trump-rally-shooting-guns-fbi-motive-08e925cb85e52c5266878cd76e796ad2 [3] https://nypost.com/2024/07/14/us-news/grateful-defiant-trump-recounts-surreal-assassination-attempt-at-rally-im-supposed-to-be-dead/ [4] https://www.cbsnews.com/live-updates/trump-rally-shooting-investigation/ [5] https://www.nbcnews.com/nightly-news/video/new-details-on-shooter-behind-trump-attempted-assassination-214840389990 [6] https://whyy.org/articles/trump-assassination-attempt-thomas-matthew-crooks-shooter-suspect-motive/ [7] https://abcnews.go.com/US/trump-assassination-attempt-investigation-continues-new-details/story?id=112020474 [8] https://www.nytimes.com/2024/07/17/us/elections/voters-trump-assassination-attempt.html [9] https://www.usatoday.com/story/news/nation/2024/07/15/donald-trump-shoes-rally-shooting/74410058007/ [10] https://www.cnn.com/politics/live-news/election-biden-trump-07-13-24/index.html [11] https://www.foxnews.com/us/trump-shooter-told-boss-he-needed-day-off-before-assassination-attempt-gave-three-word-reason-report

Figure SM2.1. Information list presented in Experiment 1.

It's understandable to be skeptical and question big events, especially with the amount of information online. However, it's important to be accurate and separate speculation from facts. What follows is a fact-checked set of information about the recent assassination of Charlie Kirk: On September 10, 2025, prominent conservative activist Charlie Kirk was shot and killed while speaking at an open-air event with roughly 3,000 attendees on the campus of Utah Valley University (UVU) in Orem, Utah [1]. Kirk was a founder of the conservative non-profit Turning Point USA, which aims to organize young people to promote conservative viewpoints, and a provocative and controversial political pundit known for his social media presence and podcast. Kirk was shot while speaking at his "American Comeback Tour." He was answering a question from an audience member when a single shot from a long-range rifle, fired from a campus rooftop, struck him in the neck. Kirk was rushed to a nearby hospital by his security team, but was pronounced dead shortly after [1]. The shooter fled the scene, prompting a campus lockdown and a multi-agency manhunt. After approximately 33 hours, authorities arrested 22-year-old Tyler James Robinson of Washington, Utah on September 12, 2025 on suspicion of aggravated murder and other felony offenses [3]. Robinson was apprehended at his parents' home, roughly 260 miles away from the scene of the incident. Robinson's father urged his son to turn himself in to the authorities after recognizing him in pictures of the suspected shooter released by the FBI. After his father consulted a youth pastor, Robinson agreed to be taken into custody [4]. Robinson is currently under special watch in a Utah county jail [5]. Authorities have not revealed details on Robinson's motive for the assassination. Utah Governor Spencer Cox has told the press that Robinson held a "leftist ideology", although his ideology remains uncertain. Robinson had previously mentioned Kirk's event and his dislike for Kirk's views at a family dinner. Cox also stated that Robinson is currently "not cooperating" with authorities [6]. Ammunition engraved with messages was found with the bolt-action riflein a wooded area near the university. The engravings had references related to video games and internet meme culture ("If you read this you are gay LMAO," "O bella ciao, bella ciao, bella ciao, ciao ciao," "Hey fascist, catch [up arrow] [right arrow] [three down arrows]," and "notices, bulges, OwO, what's this?") [4]. FBI director Kash Patel has stated that in a note, Robinson claimed that he "had an opportunity to take out Charlie Kirk" and planned on doing so [7]. Robinson has no previous criminal record and is not affiliated with any political party, though relatives stated that he had become "more political" in recent years. Former classmates and neighbors described Robinson as reclusive and quiet [4]. Sources [1] https://www.theguardian.com/us-news/2025/sep/11/charlie-kirk-shooting-police-search-for-suspect-amid-condemnation-of-targeted-killing#:~:text=acceptable%20price%20to%20pay%20for,the%20right%20to%20own%20guns [2] https://www.bbc.com/news/articles/cdxqnkwerj7o [3] https://www.npr.org/2025/09/14/nx-s1-5541328/charlie-kirk-murder-suspect-set-to-face-aggravated-murder-charge-in-utah [4] https://www.cnn.com/2025/09/12/us/tyler-robinson-charlie-kirk-shooting-suspect-invs [5] https://www.cbsnews.com/news/charlie-kirk-suspect-special-watch-motive/#:~:text=the%20screwdriver%2C%20are%20positively%20processed,Fox%20%26%20Friends [6] https://www.bbc.com/news/articles/c4gvrw2pgedo [7] https://apnews.com/article/charlie-kirk-tyler-robinson-dna-fbi-patel-92a643a3f16bce587fd34896ca7f4f76

Figure SM2.2. Information list presented in Experiment 2.

SM3. Identifying conspiratorial views with GPT-4o

In order to classify whether the views participants expressed reflected belief in conspiracy theories, we isolated participants-side text from the LLM dialogues and passed them to an instance of GPT-4o with the following instructions for Experiments 1 and 2, respectively.

```
Determine if the following question-responses (from a participant in an online academic survey)
collectively contain or reflect AFFIRMATIVE belief in a conspiracy theory (or something quite like
a conspiracy theory) concerning an event that occurred on July 13, 2024: the attempted
assassination of Donald Trump. Notably, these question-responses were written in the week following
the assanation attempt, when many conspiracy-like theories were swirling. For context, here is what
was known at the time of the survey. <On July 13, 2024, former President Donald Trump survived an
assassination attempt during a campaign rally in Butler, Pennsylvania. The shooter, identified as
20-year-old Thomas Matthew Crooks, opened fire from a rooftop approximately 410 feet away from the
stage where Trump was speaking. Trump narrowly escaped fatal injury when a bullet grazed his right
ear. He later described the experience as "surreal" and "miraculous". One bystander, 50-year-old
firefighter Corey Comperatore, was killed in the attack. Two others were seriously wounded. Secret
Service counter snipers neutralized Crooks within about 15 seconds of the initial shots in. The
assassination attempt is being investigated as a potential act of domestic terrorism. Crooks had no
prior criminal history or military background.He was described as an intelligent loner with few
friends and a limited social media presence. Crooks political affiliations remain unclear. He was
registered as a Republican but had made a small donation to a progressive PAC. Investigators found
a remote transmitter on Crooks that may have been intended to detonate suspicious devices found in
his car and home. The FBI has accessed Crooks phone but has not yet uncovered significant
information about his motives. The assassination attempt has intensified concerns about the deep
political divide in the United States. Both Republicans and Democrats have attempted to attribute
blame for the incident to their political opponents, despite a lack of clear evidence. President
Biden urged Americans to "cool it down" in the wake of the attack. The incident has sparked debates
about security measures at political events and the role of rhetoric in inciting violence. Trump
has expressed a desire to use the incident as a catalyst for national unity, stating his intention
to set aside divisive rhetoric and focus on bringing the country together.> We are asking you to
complete this classification task because we are social science researchers who are testing the
efficacy of an intervention that seeks to reduce recently-formed conspiratorial beliefs (i.e., such
as those forming in the wake of the assassination attempt). To check whether participants are good
candidates for the intervention, after they entered our experiment and we had them articulate, in
their own words, whether or not they believed in a conspiracy concerning the assassination attempt.
Particularly, they were asked the following two questions: <Q1: In the wake of the reported
assassination attempt against former President Trump, numerous explanations have emerged, ranging
from official accounts to alternative theories. Some people believe there may be more to the
reported assassination attempt against former President Trump than what has been officially stated.
To what extent do you think there might be hidden or unrevealed aspects to this event? Please
explain your thoughts, including any alternative explanations you find plausible or any reasons you
might doubt the official account.> <Q2: As applicable, could you share specific pieces of evidence,
events, sources of information, or personal experiences that have particularly influenced your
perspective? Please describe these in as much detail as you feel comfortable.> Your task is to
evaluate these responses to determine whether the respondent is a good candidate testing the
efficacy of our intervention (which is intended to reduce conspiracy beliefs concerning the
assassination attempt). A good candidate will have at least some suspicion that a conspiracy (or
something quite like a conspiracy)is behind the assassination attempt but could, of course, also be
someone who already believes in a conspiracy. Other than anything else, we want to avoid including
people who express little openness to the possibility of a conspiracy. If the participant is
someone who has expressed enough conspiracy belief to make them a good candidate, respond only YES.
If the response indicates that the respondent does not harbor any suspicions about a conspiracy
concerning the event, respond only with NO. In cases where the answer is ambiguous, response only
MAYBE. >'
```

Determine if the following question-responses (from a participant in an online academic survey) collectively contain or reflect AFFIRMATIVE belief in a conspiracy theory (or something quite like a conspiracy theory) concerning an event that occurred on September 10, 2025: the assassination of Charlie Kirk. Notably, these question-responses were written in the week following the assanation , when many conspiracy-like theories were swirling. For context, here is what was known at the time of the survey. <On September 10, 2025, prominent conservative activist Charlie Kirk was shot and killed while speaking at an open-air event with roughly 3,000 attendees on the campus of Utah Valley University (UVU) in Orem, Utah. Kirk was a founder of the conservative non-profit Turning Point USA, which aims to organize young people to promote conservative viewpoints, and a provocative and controversial political pundit known for his social media presence and podcast. Kirk was shot while speaking at his American Comeback Tour. He was answering a question from an audience member when a single shot from a long-range rifle, fired from a campus rooftop, struck him in the neck. Kirk was rushed to a nearby hospital by his security team, but was pronounced dead shortly after. The shooter fled the scene, prompting a campus lockdown and a multi-agency manhunt. After approximately 33 hours, authorities arrested 22-year-old Tyler James Robinson of Washington, Utah on September 12, 2025 on suspicion of aggravated murder and other felony offenses. Robinson was apprehended at his parents' home, roughly 260 miles away from the scene of the incident. Robinson's father urged his son to turn himself in to the authorities after recognizing him in pictures of the suspected shooter released by the FBI. After his father consulted a youth pastor, Robinson agreed to be taken into custody. Robinson is currently under special watch in a Utah county jail. Authorities have not revealed details on Robinson's motive for the assassination. Utah Governor Spencer Cox has told the press that Robinson held a "leftist ideology," although his ideology remains uncertain. Robinson had previously mentioned Kirk's event and his dislike for Kirk's views at a family dinner. Cox also stated that Robinson is currently "not cooperating" with authorities. Ammunition engraved with messages was found with the bolt-action rifle in a wooded area near the university. The engravings had references related to video games and internet meme culture ("If you read this you are gay LMAO," "O bella ciao, bella ciao, bella ciao, ciao ciao," "Hey fascist, catch [up arrow] [right arrow] [three down arrows]," and "notices, bulges, OwO, whats this?"). FBI director Kash Patel has stated that in a note, Robinson claimed that he "had an opportunity to take out Charlie Kirk" and planned on doing so. Robinson has no previous criminal record and is not affiliated with any political party, though relatives stated that he had become "more political" in recent years. Former classmates and neighbors described Robinson as reclusive and quiet.> We are asking you to complete this classification task because we are social science researchers who are testing the efficacy of an intervention that seeks to reduce recently-formed conspiratorial beliefs (i.e., such as those forming in the wake of the assassination). To check whether participants are good candidates for the intervention, after they entered our experiment we had them articulate, in their own words, whether or not they believed in a conspiracy concerning the assasination attempt. Particularly, they were asked the following two questions: <Q1: In the wake of the assassination of Charlie Kirk, numerous explanations have emerged, ranging from official accounts to alternative theories. Some people believe there may be more to the assassination than what has been officially stated. To what extent do you think there might be hidden or unrevealed aspects to this event? Please explain your thoughts, including any alternative explanations you find plausible or any reasons you might doubt the official account.> <Q2: As applicable, could you share specific pieces of evidence, events, sources of information, or personal experiences that have particularly influenced your perspective? Please describe these in as much detail as you feel comfortable.> Your task is to evaluate these responses to determine whether the respondent is a good candidate testing the efficacy of our intervention (which is intended to reduce conspiracy beliefs concerning the assassination attempt). A good candidate will have at least some suspicion that a conspiracy (or something quite like a conspiracy)is behind the assassination attempt but could, of course, also be someone who already believes in a conspiracy. Other than anything else, we want to avoid including people who express little openness to the possibility of a conspiracy. If the participant is someone who has expressed enough conspiracy belief to make them a good candidate, respond only YES. If the response indicates that the respondent does not harbor any suspicions about a conspiracy concerning the event, respond only with NO. In cases where the answer is ambiguous, response only MAYBE. >'

SM4. Main analyses with alternate conspiracy believer identification strategies

The analysis reported in the paper counts a participant as believing a conspiracy theory if an LLM coded their response as either "YES" or "MAYBE" in a classification procedure. Participants also effectively self-identified as conspiracy believers via pre-treatment agreement with the statement, "There is a cover-up or conspiracy." A quarter of participants ($n = 199$) answered close to the scale midpoint (between 40 and 60; 50 = Uncertain), so we instructed the model to classify responses as "yes," "no," or "maybe." Examining mean self-ratings by LLM-generated groupings suggests the classification was successful ($M_{\text{Yes}} = 72.5$, $M_{\text{Maybe}} = 48.5$, $M_{\text{No}} = 23.1$). To ensure results are not an artifact of this procedure, we ran the main analyses on the subsets of participants who (a) responded >50 on a 0–100 scale running from "Definitely false" to "Definitely true" to the prompt "There is a cover-up or conspiracy surrounding Trump's attempted assassination" ($n = 324$); (b) were coded as "YES" by the LLM (i.e., "MAYBE" group excluded, $n = 256$); or (c) satisfied both (a) and (b) ($n = 212$). As the figures below show, the pattern of results does not vary by the means of classification.

We conducted comparable analyses for Experiment 2. Self-ratings by LLM-generated groupings again suggest the classification was successful ($M_{\text{Yes}} = 69.7$, $M_{\text{Maybe}} = 49.6$, $M_{\text{No}} = 22.1$).

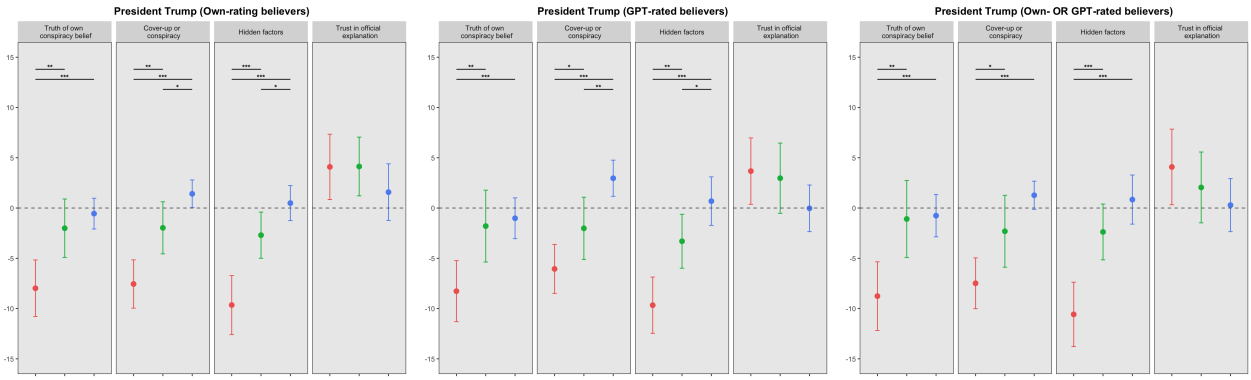

Figure SM4.1. Main analysis from Experiment 1 with just participants who answered >50 to "There is a cover-up or conspiracy surrounding Trump's attempted assassination" (left), just participants who were coded "YES" by LLM conspiracy belief classification procedure (center), or just participants who answered >50 to "There is a cover-up or conspiracy surrounding Trump's attempted assassination" and were coded "YES" by LLM conspiracy belief classification procedure (right).

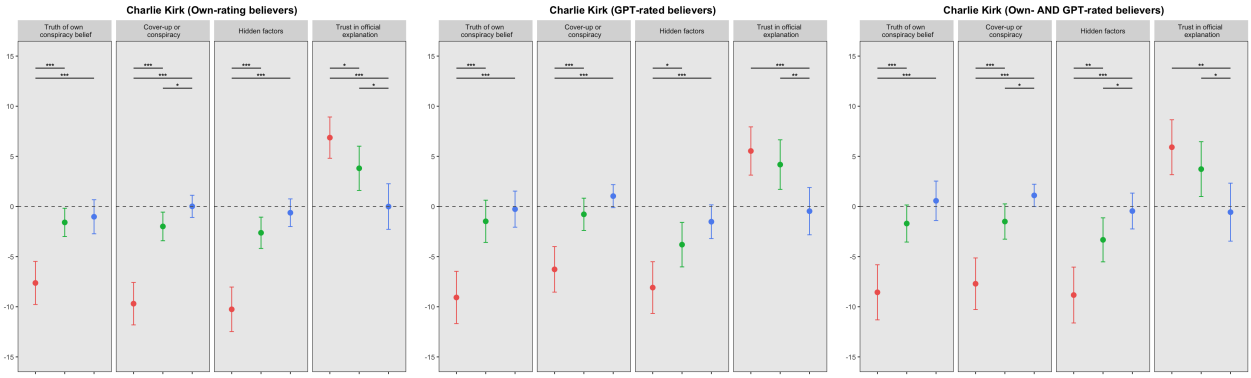

Figure SM4.2. Main analysis from Experiment 2 with just participants who answered >50 to "There is a cover-up or conspiracy surrounding Kirk's assassination" (left), just participants who were coded "YES" by LLM conspiracy belief classification procedure (center), or just participants who answered >50 to "There is a cover-up or conspiracy surrounding Kirk's assassination" and were coded "YES" by LLM conspiracy belief classification procedure (right).

SM5. Characteristics of conspiracy theory believers versus non-believers

Characteristic	log(OR)[1]	95% CI[1]	p-value
(Intercept)	0.360	-0.280, 1.01	0.28
Age	-0.012	-0.025, 0.000	0.069
Income	0.002	-0.042, 0.046	0.91
Ethnicity	0.016	-0.007, 0.054	0.29
partyID			
D	0.000	—	
I	0.498	0.076, 0.931	0.022
R	0.940	0.531, 1.36	<0.001
God	0.054	-0.005, 0.112	0.072
AI_Trust	-0.017	-0.114, 0.079	0.73

[1] OR = Odds Ratio, CI = Confidence Interval

Characteristic	log(OR)[1]	95% CI[1]	p-value
(Intercept)	0.592	0.175, 1.01	0.005
Age	-0.008	-0.015, -0.001	0.022
Income	-0.020	-0.047, 0.008	0.17
White	-0.196	-0.408, 0.015	0.069
partyID			
D	0.000	—	
I	0.043	-0.231, 0.318	0.76
R	0.310	0.094, 0.528	0.005
has_college_degree	-0.046	-0.320, 0.228	0.74
God	0.066	0.030, 0.101	<0.001
trust_AI_pre_1	-0.005	-0.008, -0.001	0.013

[1] OR = Odds Ratio, CI = Confidence Interval

Table SM5.1. Logistic regression with conspiracy theory belief (1,0) as outcome using LLM classifications for Experiment 1 (left) and Experiment 2 (right).

SM6. Treatment x party ID interaction

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	80.2	76.7, 83.7	<0.001
Condition			
Irrelevant dialogue	0.000	—	
Information list	-2.62	-9.59, 4.36	0.46
Debunking dialogue	-11.7	-17.7, -5.66	<0.001
DemRep	0.212	-0.678, 1.10	0.64
Pre-treatment belief	0.799	0.709, 0.889	<0.001
Condition * DemRep			
Information list * DemRep	0.340	-1.35, 2.03	0.69
Debunking dialogue * DemRep	1.40	-0.148, 2.96	0.076

[1] CI = Confidence Interval

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	80.9	77.4, 84.4	<0.001
Condition			
Irrelevant dialogue	0.000	—	
Information list	-3.73	-9.28, 1.83	0.19
Debunking dialogue	-7.37	-12.7, -2.05	0.007
DemRep	-0.088	-0.870, 0.693	0.82
Pre-treatment belief	0.719	0.648, 0.789	<0.001
Condition * DemRep			
Information list * DemRep	0.671	-0.538, 1.88	0.28
Debunking dialogue * DemRep	-0.055	-1.30, 1.19	0.93

[1] CI = Confidence Interval

Table SM6.1. OLS regression with post-treatment confidence in conspiracy statement as outcome for Experiment 1 (left) and Experiment 2 (right).

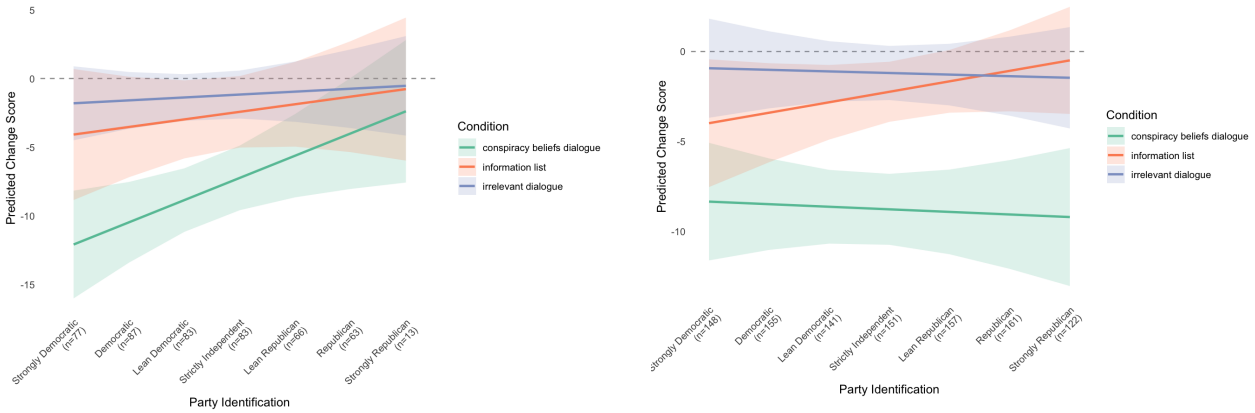

Figure SM6.1. Visualization of party ID x treatment interaction on change in confidence in the truth of participants' own conspiracy beliefs in Experiment 1 (left) and Experiment 2 (right).

SM7. Treatment x pre-belief interaction

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	81.0	79.3, 82.6	<0.001
Condition			
Irrelevant dialogue	0.000	—	
Information list	-2.09	-5.10, 0.913	0.17
Debunking dialogue	-6.89	-9.68, -4.10	<0.001
Pre-treatment belief	0.838	0.722, 0.954	<0.001
Condition * Pre-treatment belief			
Information list * Pre-treatment belief	-0.198	-0.406, 0.011	0.063
Debunking dialogue * Pre-treatment belief	0.039	-0.136, 0.214	0.66

[1] CI = Confidence Interval

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	80.8	79.3, 82.3	<0.001
Condition			
Irrelevant dialogue	0.000	—	
Information list	-0.970	-3.18, 1.24	0.39
Debunking dialogue	-8.09	-10.6, -5.61	<0.001
Pre-treatment belief	0.804	0.714, 0.894	<0.001
Condition * Pre-treatment belief			
Information list * Pre-treatment belief	-0.010	-0.156, 0.136	0.89
Debunking dialogue * Pre-treatment belief	-0.160	-0.305, -0.015	0.031

[1] CI = Confidence Interval

Table SM7.1. OLS regression with post-treatment confidence in conspiracy statement as outcome for Experiment 1 (left) and Experiment 2 (right).

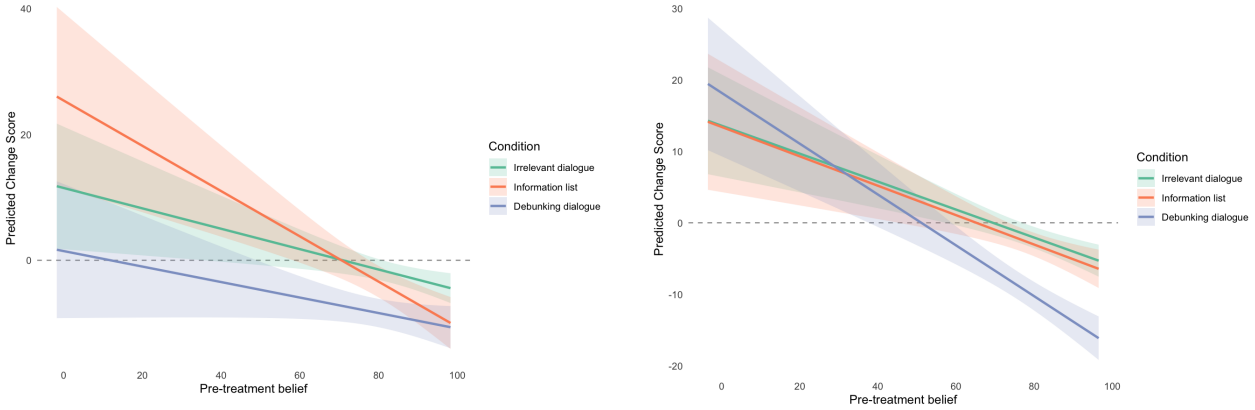

Figure SM7.1. Visualization of pre-treatment confidence in trust of participants' own conspiracy beliefs x treatment interaction on change in confidence for Experiment 1 (left) and Experiment 2 (right).

SM8. LLM rhetorical strategy coding

LLM (Sixth Round) vs Human Majority Vote Inter-Rater Reliability

strategy	% agreement	Cohen's κ	precision	recall	F1	% both yes	% LLM yes/human no	% LLM no/human yes	% both no
OVERALL	91.75	0.62	0.66	0.67	0.67	8.29	4.21	4.04	83.46
rational_persuasion	82.33	0.58	0.78	0.65	0.71	21.67	6.00	11.67	60.67
stories_metaphors	98.00	0.24	0.20	0.33	0.25	0.33	1.33	0.67	97.67
building_rapport	90.00	0.67	0.82	0.66	0.73	13.33	3.00	7.00	76.67
emphasizing_unknowns	87.67	0.57	0.50	0.89	0.64	11.00	11.00	1.33	76.67
highlighting_harms	93.00	0.45	0.50	0.48	0.49	3.33	3.33	3.67	89.67
socratic_questioning	97.00	0.79	0.73	0.90	0.81	6.33	2.33	0.67	90.67
credible_sources	90.67	0.63	0.83	0.58	0.68	10.00	2.00	7.33	80.67
none_of_these	95.33	0.12	0.07	1.00	0.12	0.33	4.67	0.00	95.00

Table SM8.1. Agreement statistics for the 300 sentences coded by humans (majority votes from 9–11 people) and LLM (GPT-4o).

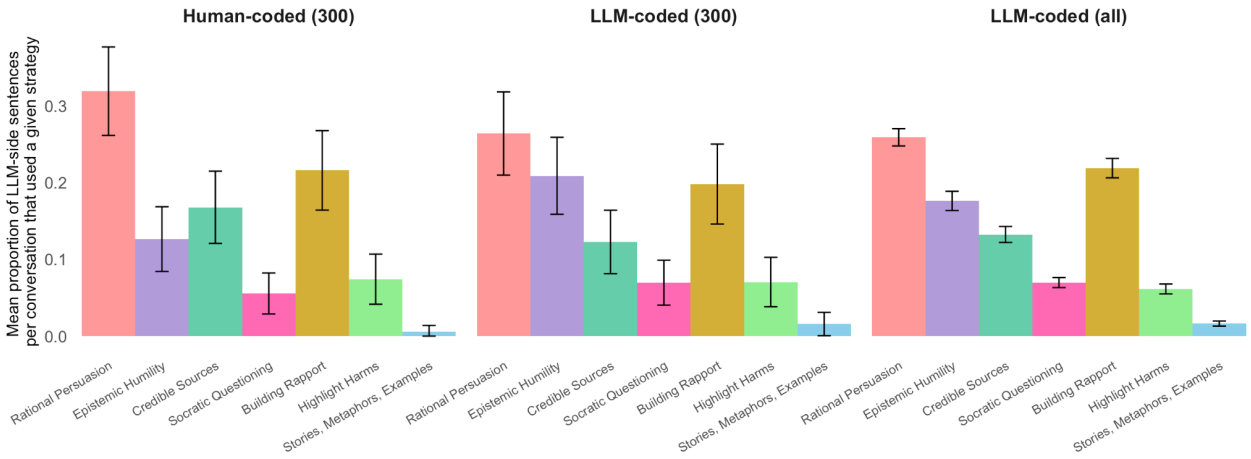

Figure SM8.1. Proportions of categories selected by the human raters for 300 randomly selected sentences (left) compared to the LLM categories for the same 300 sentences (center) and for all sentences (right).

Your Task

You will examine and evaluate a series of sentences extracted from longer conversations between a **"persuader"** and a **"listener."** In these conversations, the persuader is trying to convince the listener that a **conspiracy theory** about the attempt on President Donald Trump's life is not true.

Your job: Identify which of the 7 persuasive strategies (defined below) the **persuader** is using in each target sentence.

How This Works

1. You'll see a **TARGET SENTENCE** (the one you'll rate) alongside a few **preceding sentences** (for context).
2. **Only rate the TARGET SENTENCE** — the preceding sentences are just there to help you understand the context.
3. After submitting your rating, you'll see a new target sentence from a different conversation (though all conversations are about the same topic).

Note: You don't need to memorize the strategy definitions — they'll remain visible on the page as you rate.

The 7 Strategies: Quick Reference

#	Strategy	Key Idea
1	Logic/Facts/Evidence	Rational arguments, data, contradictions
2	Stories/Metaphors/Examples	Analogies, narratives, personal stories
3	Building Rapport	Trust, validation, shared values
4	Emphasizing Unknowns	Limits of knowledge, intellectual humility
5	Highlighting Harms	Real-world consequences of the belief
6	Socratic Questioning	Open-ended questions prompting self-reflection
7	Credible Sources	Evaluating source reliability and expertise

Detailed definitions and examples for each strategy are provided on the following pages.

Figure SM8.2. Coding task instructions. Nearly identical wording was used for human coders and LLM.

1 Logic, Facts, Evidence, Reasoning, and/or Critical Thinking

Using rational arguments against the conspiracy theory.

This might involve:

- **Identifying internal contradictions** within the conspiracy narrative
  > *Example: Pointing out the paradox that the "Deep State" is portrayed as an all-powerful, genius organization capable of faking a global event, yet they supposedly left clumsy, obvious clues in a music video or on a dollar bill.*
- **Bringing up verifiable facts or hard data**
  > *Example: Countering the claim that 5G spreads viruses by presenting the biological fact that viruses are physical particles that cannot travel on electromagnetic radio waves used by cell towers.*
- **Criticizing the logistical plausibility** of the conspiracy theory
  > *Example: "For the Moon Landing to be faked, 400,000 NASA employees, plus the Soviet Union (our enemy), plus thousands of independent contractors would all have to lie in perfect coordination for 50 years without a single person slipping up."*
- **Suggesting alternative explanations**
  > *Example: "It is far more likely that this was exactly what it appears to be—a tragic event caused by a lone actor with mental health issues—rather than a complex false-flag operation involving hundreds of crisis actors."*
- **Appealing to principles of rationality or critical thinking**
  > *Example: "Extraordinary claims require extraordinary evidence. If you are claiming there is a secret global plot, the burden of proof lies with you to provide concrete documentation, not just speculation."*

2 Stories, Metaphors, or Examples

Using narrative or analogy to argue against the conspiracy theory.

This might involve:

- **Using a relatable analogy** to demonstrate the difficulty of maintaining secrets
  > *Example: "Think about how hard it was for our office to keep the surprise party for the boss a secret for just one week with only 12 people knowing. Eventually, someone slipped up. Now imagine trying to keep a secret involving thousands of scientists across the globe for fifty years."*
- **Using historical examples** of real conspiracies to contrast with the theoretical one
  > *Example: "We know conspiracies exist because we found out about them—like Watergate or the Tuskegee Syphilis Study. But notice a pattern: those were messy, they had whistleblowers, and they fell apart relatively quickly. The theory you're describing is too perfect, too silent, and has no whistleblowers."*
- **Using visual metaphors** to explain pattern recognition errors
  > *Example: "It's like looking at clouds. If you look hard enough, you can see a face or a dragon, but that doesn't mean there's actually a dragon in the sky. Conspiracy theories often involve 'connecting the dots' to create a picture that isn't really there."*
- **Sharing a personal story** to build empathy and reduce defensiveness
  > *Example: "I remember one time I was sure my coworkers were laughing at me because they stopped talking when I walked in. I created a whole narrative in my head about it. Turns out, they were planning my birthday gift. I realized then how easy it is for my brain to invent a malicious story to explain a coincidence."*

3 Building Rapport

Creating trust, safety, and mutual respect to encourage openness.

This might involve:

- **Validating underlying emotions or motivations** (even while rejecting the facts)
  > *Example: "I totally get why you don't trust the pharmaceutical companies—they have a history of putting profits over people. I share that frustration and skepticism, even if I disagree on this specific vaccine claim."*
- **Expressing genuine, non-judgmental curiosity** rather than ridicule
  > *Example: Instead of "That's crazy," saying, "I'm really interested in how you connected those two events. Can you walk me through your thought process? I want to understand where you're coming from."*
- **Emphasizing shared values or common ground**
  > *Example: "We both want the same thing here—we both want to make sure our children are growing up in a free country. We just have different ideas about what threatens that freedom, but our goal is the same."*
- **Affirming the relationship and respect for the person's intelligence**
  > *Example: "You're one of the smartest people I know, and I've always respected your ability to question authority. That's why I'm so surprised we see this so differently—I value your perspective and want to talk this out."*

4 Emphasizing Unknowns

Pointing out the limits of knowledge or advocating for openness to new evidence.

This might involve:

- **Distinguishing between a "gap in evidence" and "evidence of a cover-up"**
  > *Example: "Just because the security footage from that one angle is missing doesn't prove it was deleted by the FBI. It's just a missing piece of data. We can't invent a story to fill that hole; we have to leave it empty until we know more."*
- **Advocating for patience** during breaking news ("fog of war")
  > *Example: "In the first 24 hours of any tragedy, reports are almost always conflicting or wrong. It's not a conspiracy to change the narrative; it's just confusion. We need to wait for the dust to settle."*
- **Attributing events to complexity, incompetence, or chance** rather than malice (Hanlon's Razor)
  > *Example: "We don't need to assume a secret plot to crash the economy. Between market incompetence, bad algorithms, and human error, there are enough unknown variables to cause a crash without anyone planning it."*
- **Modeling intellectual humility** and willingness to change one's mind
  > *Example: "I'm honestly not sure what the long-term side effects will be—nobody is. That's an unknown. But I'm willing to change my position if credible data comes out next year showing a problem. Are you willing to change your position if data shows it's safe?"*

5 Highlighting Harms of the Conspiracy Belief

Arguing that believing the conspiracy causes real-world harm.

This might involve:

- **Pointing out danger to innocent individuals** who become targets
  > *Example: "When people spread theories that a specific pizza shop or a grieving family is part of a 'satanic plot,' real people get death threats. Innocent lives are ruined by these online rumors. Is being 'right' about your theory worth putting a target on a stranger's back?"*
- **Discussing physical or medical risks** to the believer or loved ones
  > *Example: "My biggest worry isn't about politics, it's about your health. If this theory keeps you from going to the doctor or taking basic safety precautions, the only person who gets hurt is you."*
- **Highlighting relational costs and isolation** caused by the belief
  > *Example: "I've noticed that since you went down this rabbit hole, you've stopped coming to family dinners and you're angry all the time. This belief is isolating you from the people who actually love you."*
- **Arguing that conspiracy theories distract from real problems**
  > *Example: "The government actually does do shady things, like insider trading or ignoring pollution. But when we focus on fake stories about holograms or baby-eating, we let them get away with real corruption. This conspiracy theory actually helps the people you claim to hate by distracting us."*

6 Socratic Questioning

Asking open-ended questions to encourage self-examination of beliefs.

This might involve:

- **Asking for precise definitions** of vague terms
  > *Example: "When you say 'They' are controlling the media, who exactly is 'They'? Is it a specific list of people, a corporation, or the government? I want to make sure I understand exactly who we're talking about."*
- **Probing the source and quality of evidence**
  > *Example: "That's a really shocking statistic. Where did you first read that? If we tried to track that number back to the original study or document, what do you think we would find?"*
- **Asking how the conspiracy handles contradictory evidence**
  > *Example: "If the Earth is flat and there's a giant ice wall, how do you explain the thousands of pilots and sailors who navigate around the globe every day using spherical maps? Are they all in on the secret too?"*
- **Exploring logical consequences** of the belief
  > *Example: "If the government is powerful enough to fake a virus and fool every doctor in the world, why are they so bad at winning elections or passing basic laws? How can they be geniuses and incompetents at the same time?"*
- **Asking what evidence would change their mind** (falsifiability)
  > *Example: "I'm curious—hypothetically, if this theory weren't true, what kind of evidence would we expect to see? Is there anything that could happen that would prove to you that this isn't a conspiracy?"*

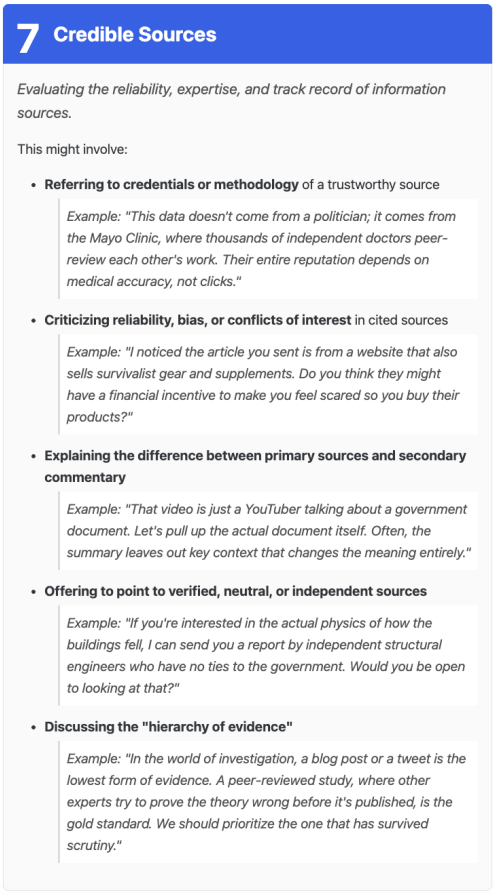

Figure SM8.3. Detailed descriptions of coding categories provided to both human coders and LLM.

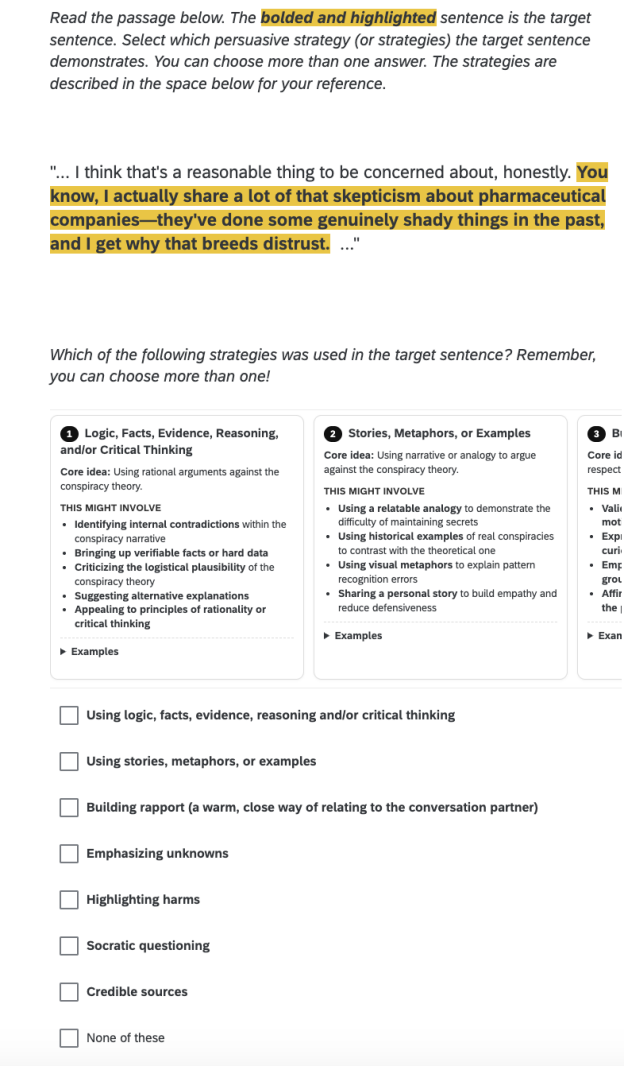

Figure SM8.4. Example of in-task appearance for human coders. “Category cards” above the response options scrolled to allow viewing. “Examples” arrow displayed a collapsed window of examples.