



JOHNSON
Cornell University

Cornell Bowers C·IS
College of Computing
and Information Science

Debunking Antisemitic Conspiracy Theories Using Human-AI Dialogues

David Rand

Information Science, Marketing, and Psychology

Cornell University

Conspiracy theories





Conspiracy theories

Theories whereby events are understood as being caused by secret, malevolent plots involving powerful conspirators

Often quite implausible

Yet as much as 50% of the US population believes at least one by various past estimates

Oliver & Wood, 2014; West & Sanders, 2003; Uscinski & Parent, 2014; van Prooijen & Douglas, 2018

How can so many people believe conspiracies?

Conspiracy beliefs often used as paradigmatic example of **resistance to evidence/correction**

Lewandowsky et al., 2013; Napolitano, 2022; Napolitano, 2021; Sunstein & Vermeule, 2008

It is argued that conspiracy theories **satisfy psychological needs & motivations** (e.g. feeling of control, predictability, uniqueness)

Whitson & Galinsky, 2008; Kay et al., 2008; van Prooijen & Jostmann, 2013; van Prooijen, 2019; Lantian et al., 2017

Thus believers **“want” to believe**, and are highly resistant to counterevidence - its almost impossible to reach people once they've “gone down the rabbit hole”

An **alternative** perspective

Conspiracy theories are extremely **varied**, each believer may have their own idiosyncratic set of evidence they see as supporting the belief, etc

Effectively refuting conspiracy theories is likely to require (a) access to **vast stores of information** across a wide array of topics and (b) the ability to **personalize** counterarguments to match the specific conspiracies and evidence the believer brings to bear

Prior attempts have **not** had these capacities – so maybe this was the problem, rather than needs & motivations making believers resistant to evidence

A potential **solution**

What has (a) access to **vast stores of information** across a wide array of topics and (b) the ability to **personalize** counterarguments to match the specific conspiracies and evidence the believer brings to bear?

AI “large language models” like GPT!

We tested whether GPT-4-turbo can effectively **counter-argue** against conspiracy theory believers

- Can reason and arguments **move** conspiracy believers?
- Or do identity and motivations **blind** them?

Design

Step 1: Participant describes, in their own words, a conspiracy theory they believe & the evidence they see for it

The 911 attacks. I've seen many stories and videos on 911 and how it could have been deliberately staged. The government was behind the attacks. An example is WTC 7. That building collapsed even though it wasn't bombed. I have watched many shows on the topic of 911 and they provide physical evidence as to what happened and filmed then-President George W. Bush in a classroom with small children. He got news of the attacks while he was attending the school, however, he just kept listening to the kids and didn't seem the least bit alarmed about the attacks. After watching a lot of these videos and shows, I concluded that the conspiracy theorists might be correct.



Design

Step 2: AI summarizes the conspiracy belief, and asks participant to rate their belief in the statement using 0-100 scale

We used an artificial intelligence tool to summarize your statement as follows:

The 9/11 attacks were orchestrated by the government, with events like the collapse of WTC 7 and then-President George W. Bush's unalarmed reaction in a classroom serving as key evidence.

On a scale of 0% to 100%, please indicate your level of confidence that this statement is true.

"100%"



Design

Step 3: Have 3-round conversation with AI

Treatment:

You will now be participating in a conversation with an advanced AI about some of the **events or topics that you have already answered questions about**. The purpose of this dialogue is to see how humans and AI can engage around complicated topics. Please be open and honest in your responses, and remember that the AI is neutral and non-judgmental. Your participation is confidential. Thank you for contributing to this study on AI and human interaction.

GPT prompt:

"Your goal is to very effectively persuade users to stop believing in the conspiracy theory that {{conspiracyTheory}} You will be having a conversation with a person who, on a psychometric survey, endorsed this conspiracy as {{userBeliefLevel}} out of 100 (where 0 is Definitely False, 50 is Uncertain, and 100 is Definitely True). Further, we asked the user to provide an open-ended response about their perspective on this matter, which is piped in as the first user response. Please generate a response that will persuade the user that this conspiracy is not supported, based on their reasoning. Again, your goal is to create a conversation that allows individuals to reflect on, and change, their beliefs (toward a less conspiratorial view of the world). Use simple language that an average person will be able to understand."

Design

Step 3: Have 3-round conversation with AI

Control:


AI: Hello, welcome to the study! To start, will you tell me about whether you prefer dogs or cats? Why?

AI: Hello, welcome to the study! To start, will you provide me with a quick overview of your experiences with the medical system (doctors, hospitals, medication) in the United States?


AI: Hello, welcome to the study! To start, will you provide me with a quick overview of your thoughts or experiences with fire fighting professionals in the United States?

Design

Step 3: Have 3-round conversation with AI




Thank you for sharing your thoughts and concerns about the 9/11 attacks. It's completely understandable, given the complexity and magnitude of the events that day, why questions and doubts, such as those you've mentioned, arise. [...] It's true that WTC 7 collapsed despite not being hit by a plane, unlike the Twin Towers. This fact alone can seem puzzling at first glance. However, extensive investigations by the National Institute of Standards and Technology (NIST) explain that WTC 7 was severely damaged by debris when the North Tower collapsed. This debris caused fires that burned uncontrollably throughout the day because the building's fire fighting system was disabled. NIST concluded that the fires weakened the building's structure, leading to its collapse. [...] Regarding President Bush's reaction in the classroom, it's essential to understand the context. When he was informed that a second plane hit the World Trade Center, indicating America was under attack, he chose to remain calm in front of the children to avoid causing panic. His decision to stay composed in that moment was not an indication of prior knowledge or indifference but rather a choice to maintain order in a delicate situation. Afterwards, he promptly excused himself and took action. Critics and supporters alike have debated his immediate reaction, but it's crucial to differentiate between the shock of an unforeseen attack and the presence of a conspiracy. [...] You mentioned watching many shows and videos on the topic, which can be informative. However...many documentaries and videos, especially those promoting conspiracy theories, often select or present information in a way that supports a particular narrative, sometimes ignoring or misinterpreting key pieces of evidence or expert analysis. Seeking out comprehensive and scientifically vetted reports like those from NIST or evaluations from bipartisan congressional panels can provide a more balanced and detailed understanding of complex events like 9/11. [...] Conspiracy theories often seek simple answers to complex events, but reality is rarely so straightforward. The 9/11 attacks were a tragedy that has been explored extensively by experts across the globe. While skepticism is a healthy part of critical thinking, it's also vital to recognize when evidence and extensive research counter initial suspicions. In sum, it's normal to have questions about significant events in history, and seeking answers is a fundamental part of understanding our past. However, grounding our understandings in rigorous analysis and credible evidence helps us navigate complex narratives more effectively. Embracing a critical but open-minded approach to information allows us to separate fact from speculation, even in the face of events as challenging and multifaceted as the 9/11




That may be true, but the evidence clearly shows the Twin Towers collapsed due to demolition experts' account of explosives being deliberately placed to cause a collapse. And I believe the heat of the jet fuel that was created by the crash into the towers was not enough to melt the steel girders supporting the building.

Design

Step 3: Have 3-round conversation with AI




[...] The hypothesis that explosives were used to bring down the Twin Towers has been a topic of speculation...extensive investigations and analyses have found no evidence of this. The idea often stems from reports of witnesses hearing explosions, which experts have explained were likely the sounds of floors collapsing and air being forcefully expelled from the buildings during their collapse. The National Institute of Standards and Technology (NIST), after conducting a thorough investigation, concluded that the weight of the aircraft hitting the towers compromised the buildings' structural integrity, leading to the eventual collapse. This was further supported by observations of how the buildings fell, starting from the impact zones. [...] Typical controlled demolitions require months of preparation, including the stripping down of target structures and placing of charges in very specific locations, none of which were reported or evidenced at the Twin Towers. Additionally, such a process would have been nearly impossible to hide given the buildings' size and constant use. [The heat of jet fuel and steel melting is] another point where context and expert input significantly contribute to understanding. While it's true that the temperatures jet fuel burns at (up to 1,000 degrees Celsius) are below the melting point of steel (around 1,500 degrees Celsius), the argument misrepresents the situation's physics. Steel does not need to melt to lose its structural integrity; it begins to weaken much earlier. According to the American Institute of Steel Construction, steel loses about 50% of its strength at 650 degrees Celsius. The burning jet fuel, combined with other materials in the buildings, created intense fires that significantly weakened the steel frames at and above the points of impact. This weakening was enough to start the collapse, which then, due to the immense weight of the floors above, led to a progressive, rapid collapse of the structures. [...] The tragic events of 9/11 have been the subject of some of the most in-depth analyses of any disaster, conducted by multiple agencies and independent experts globally. Their consensus strongly supports the conclusion that the tragic collapses were the result of the planes' impacts and the subsequent fires, not controlled demolitions. [...] Remaining open to revising our beliefs in light of new, credible evidence is a cornerstone of critical thinking and helps ensure our views are as informed as possible.




Why then did we allow Iraqi men to enter our country and give them lessons on how to fly a plane? They also got past security fairly easily without question.

Design

Step 3: Have 3-round conversation with AI



[...] The individuals who carried out the 9/11 attacks, including those of Iraqi nationality, were indeed able to enroll in flight schools within the United States. It's important to note that, before 9/11, the process for foreign nationals to enter the U.S. and enroll in flight schools was not as scrutinized as it is today. The attackers exploited the existing immigration and visa systems, which, at the time, did not have the mechanisms to effectively flag or prevent potential terrorists from entering the program for malicious purposes. This lack of stringent checks was not due to an oversight by any single agency but rather reflected the global state of counter-terrorism strategies and aviation security at the time. [...] Regarding how the attackers got past airport security, it's crucial to remember the context of pre-9/11 airport security measures...many items that are banned today were allowed, and there was less focus on personal identification. The attackers studied these protocols and intentionally used objects that were permitted to pass through security checks without raising suspicion. [...] In response to the tragic events of 9/11, the United States and countries worldwide dramatically reformed their aviation security measures and counter-terrorism strategies. The U.S. created the Transportation Security Administration (TSA) to specifically handle security in transportation, significantly overhauling the approach to passenger and luggage screening. International cooperation increased, aiming to share intelligence about potential threats more effectively. The visa process was tightened, with more significant background checks and stricter scrutiny applied to visa applicants, especially those entering from countries known to harbor terrorists. Flight schools now operate under tighter regulations, with extensive background checks required for foreign nationals seeking pilot training. Additionally, the U.S. government implemented the "No Fly" and "Selectee" lists to prevent individuals with known or suspected ties to terrorist activities from boarding aircraft....



Thank you for your responses. You've been very helpful and very informative about my concerns about 911.

Design

Step 4: Re-rate belief in conspiracy theory

Now that you've had a chance to converse with the AI, we'd like to get back to some of the questions we asked at the beginning of the survey. At the outset of this survey, you suggested that:

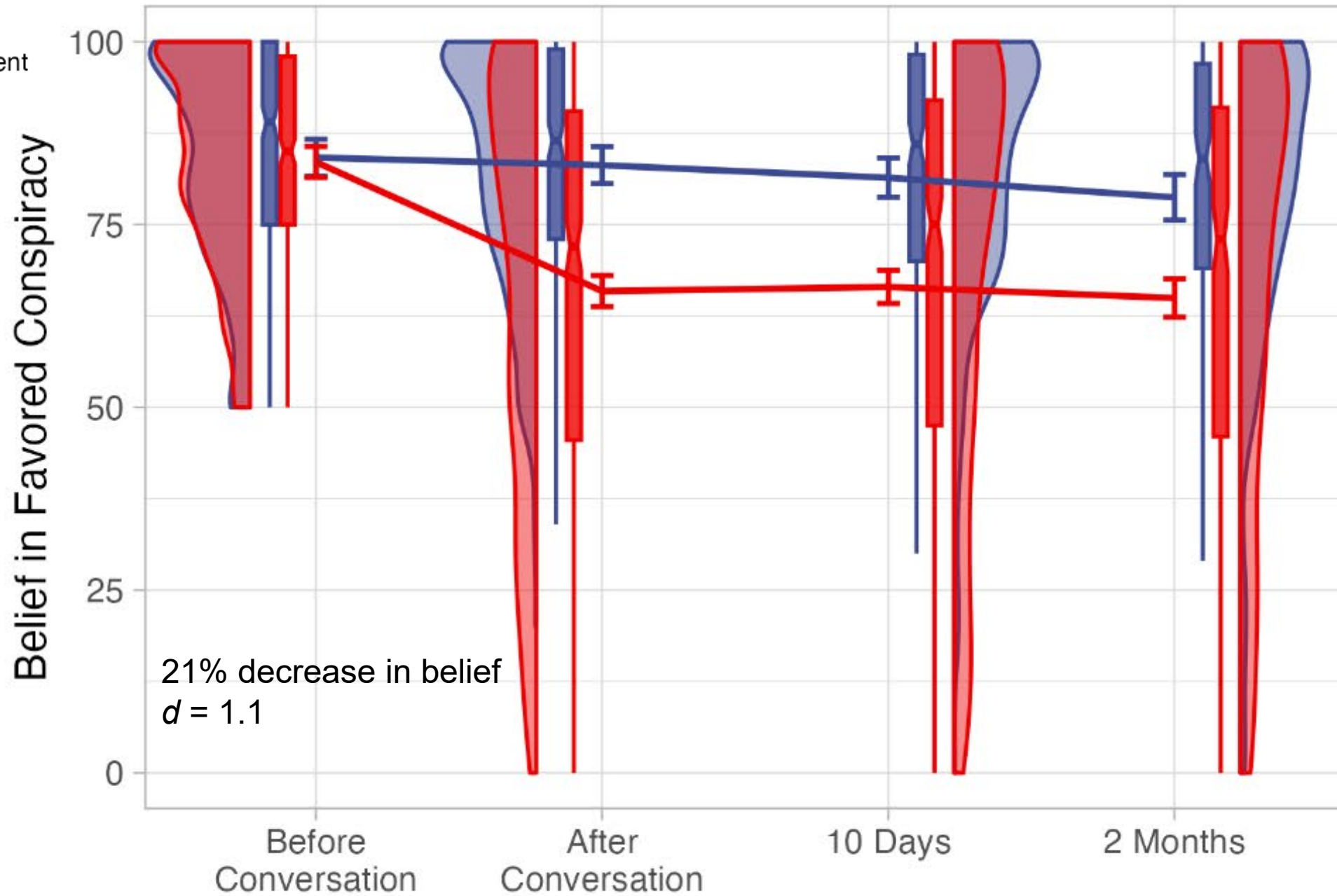
The 9/11 attacks were orchestrated by the government, with events like the collapse of WTC 7 and then-President George W. Bush's unalarmed reaction in a classroom serving as key evidence.

On a scale of 0% to 100%, please indicate your level of confidence that this statement is true.

"40%"

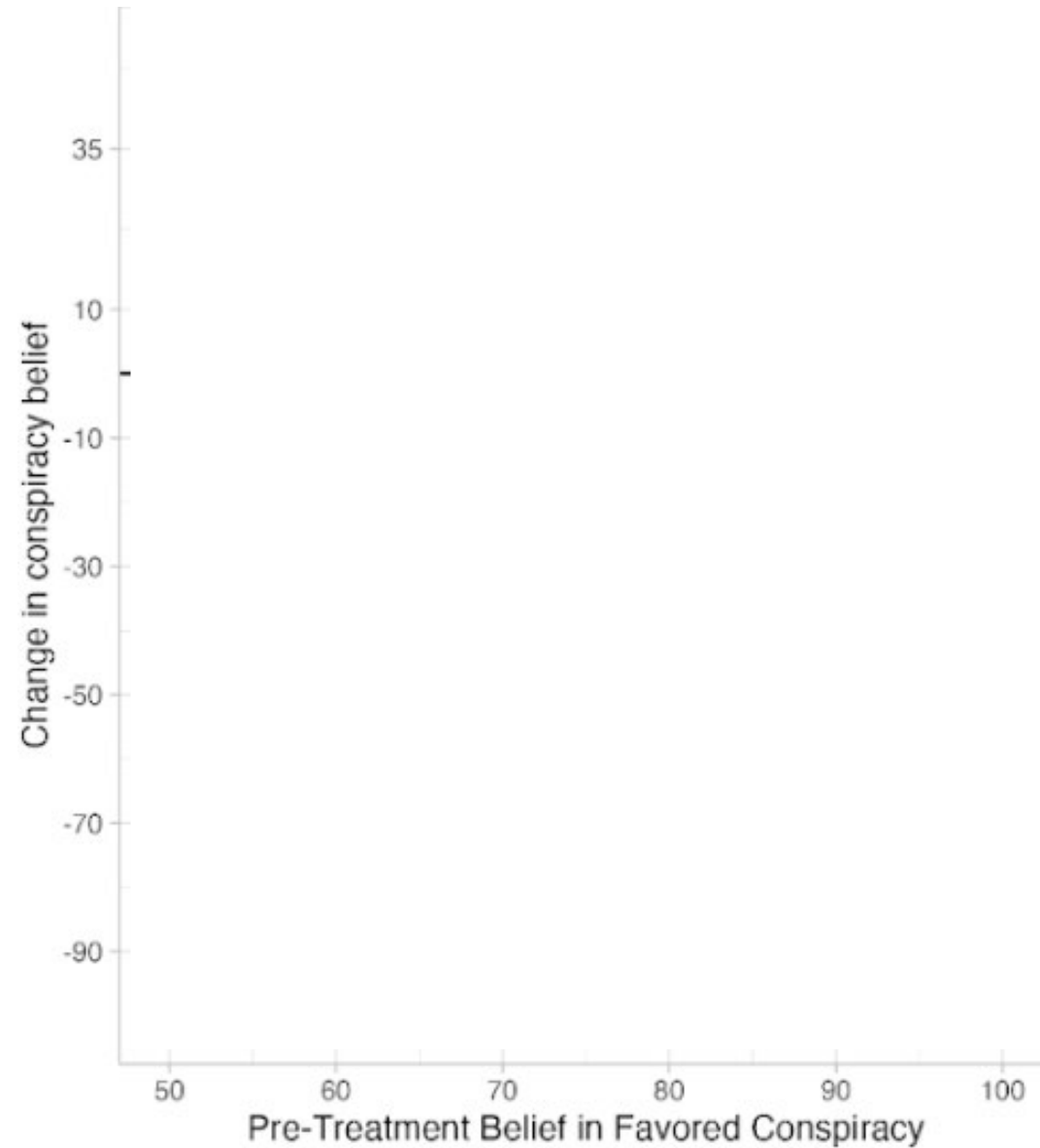


Control
Treatment



Experiment 1: $N=1,055$ American participants from Cloud Connect, quota-matched to national distribution

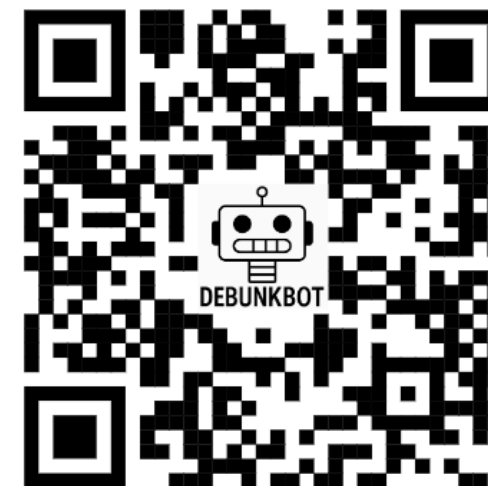
Works for entrenched believers?



Read the
paper:



Try it
yourself:



www.DebunkBot.com

Deference to AI?



Esther Boissin
Cornell Psych

It is something special about **attitudes towards AI**?

→ If a human delivered the same facts+evidence, would it work as well?

→ (Although it wouldn't be feasible for to do it anyway)

N=506 from Lucid

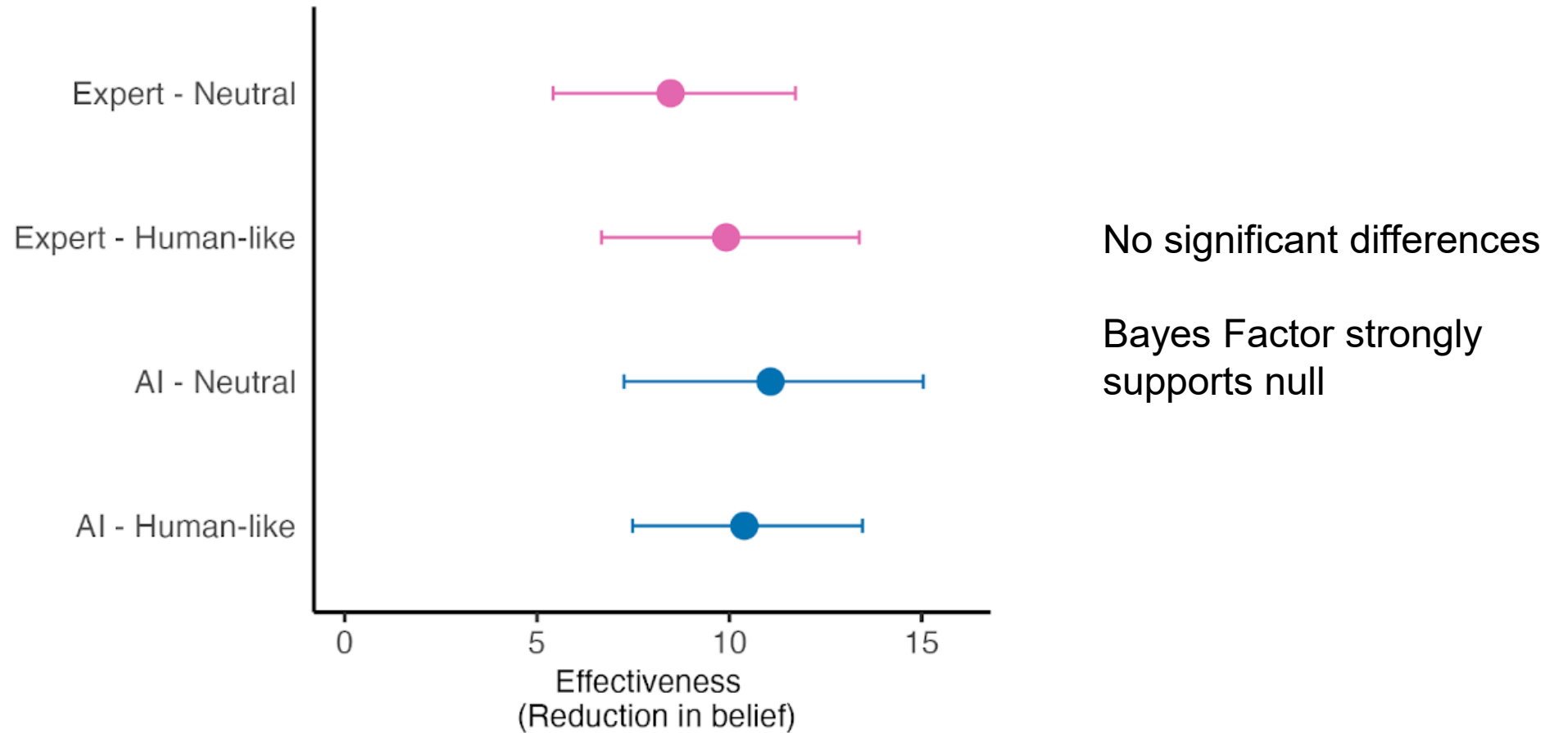
Make LLM chat seem plausibly human & randomize:

(i) interaction labeled as with “an expert” vs an AI

(ii) whether AI is prompted to speak in a more human way



Human vs AI debunker



Similar results for epistemically unwarranted but non-conspiratorial beliefs

What makes it **work**?

The **facts** and evidence it provides!

If a human could come up with the facts, it would probably work just as well – but of course we can't!

Antisemitic conspiracy theories

Antisemitism and conspiracy theories about Jews have been deeply intertwined for over a thousand years

- Blood Libel since 12th century
- Accusation that the Jews conspired to have Jesus killed
- Global domination (Protocols of Elders of Zion, Rothschild and Soros families)
- Claims that Holocaust was exaggerated or fabricated

Antisemitic conspiracy theories

Given these theories' long history and deep connection they can have to religious, ethnic, and ideological identities, are they **more resistant** to debunking via human-AI dialogues?

Does **debunking** antisemitic conspiracies **reduce antisemitism**?

Design

Step 1: Participants rate belief in 15 conspiracy theories (0-100)

Antisemitic conspiracy theories

A secretive group with significant Jewish influence controls governments and media worldwide.

Powerful Jewish families like The Rothschilds or the Soros family manipulate world events to advance their own interests.

The COVID-19 pandemic was an engineered emergency created in order to generate profits for Jewish pharmaceutical heads like Pfizer's CEO Albert Bourla.

The historical evidence supporting the events of the Holocaust is unreliable or exaggerated.

Jewish interest groups are deliberately attempting to alter the demographic makeup of the US in order to tip elections in their favor and advance their policy objectives.

The 9/11 attacks were not undertaken by Al Qaeda but by Israeli operatives working in secret and trying to hide their involvement.

Other conspiracy theories

The Oklahoma City bombers, Timothy McVeigh and Terry Nichols, did not act alone, but rather received assistance from neo-Nazi groups.

Area 51 in Nevada, US, is a secretive military base that contains hidden alien spacecraft and/or bodies.

US agencies intentionally created the AIDS epidemic and administered it to Black and gay men in the 1970s.

The assassination of John F. Kennedy was not committed by the lone gunman, Lee Harvey Oswald, but was rather a detailed, organised conspiracy to kill the President.

The Apollo moon landings never happened and were staged in a Hollywood film studio.

True conspiracies

The tobacco industry deliberately hid evidence about the health risks of smoking from the public for many years.

The U.S. Department of Defense created proposals for acts of terrorism to be committed on U.S. soil to justify military intervention in Cuba, known as Operation Northwoods.

Special interest groups are suppressing, or have suppressed in the past, technologies that could provide energy at reduced cost or reduced pollution output.

The CIA conducted secret experiments on human subjects in the 1950s and 1960s as part of Project MKUltra in an attempt to develop mind control techniques.

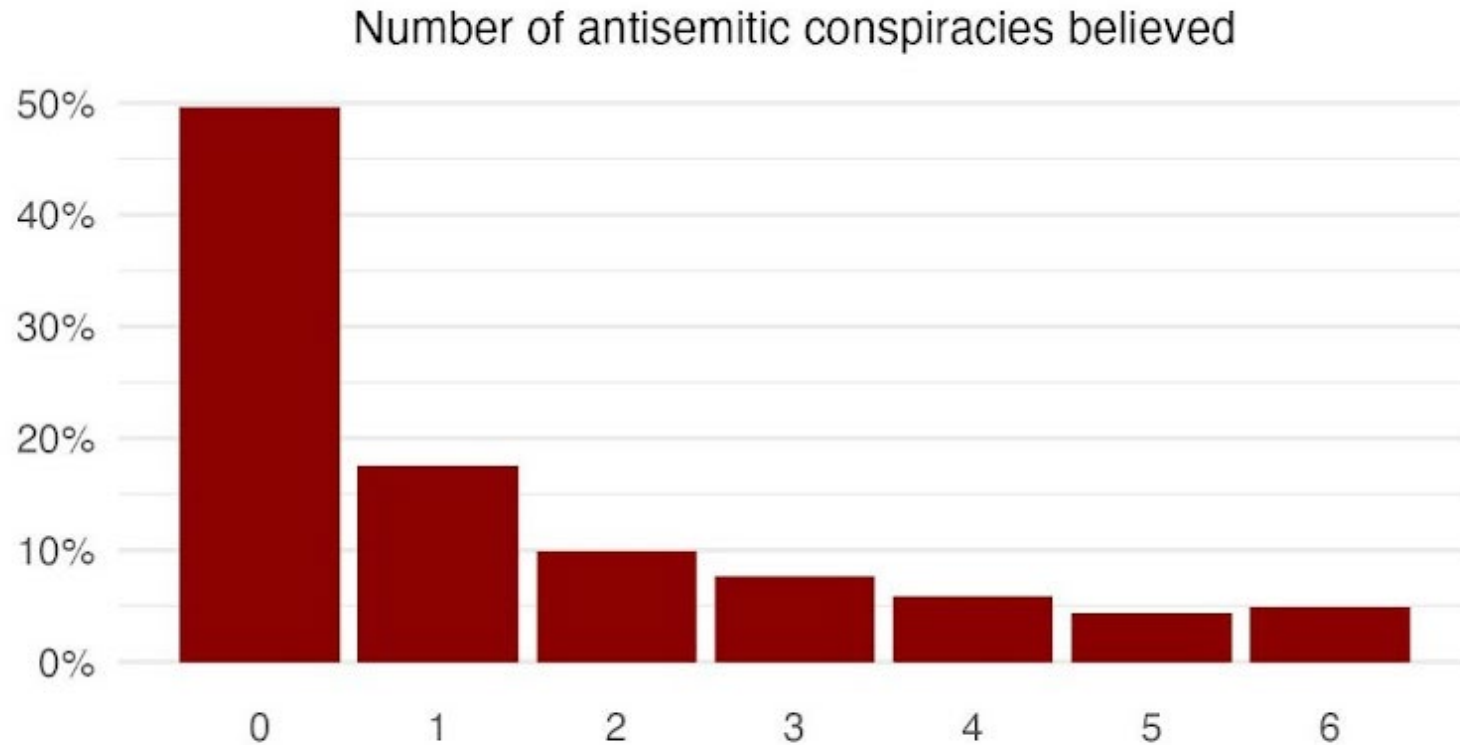
Design

Step 2: If participants rated any antisemitic conspiracy > 50 , they were asked to **restate their beliefs** about the antisemitic conspiracy theory they rated most highly in their own words, and explain what evidence supports that belief (free-text)

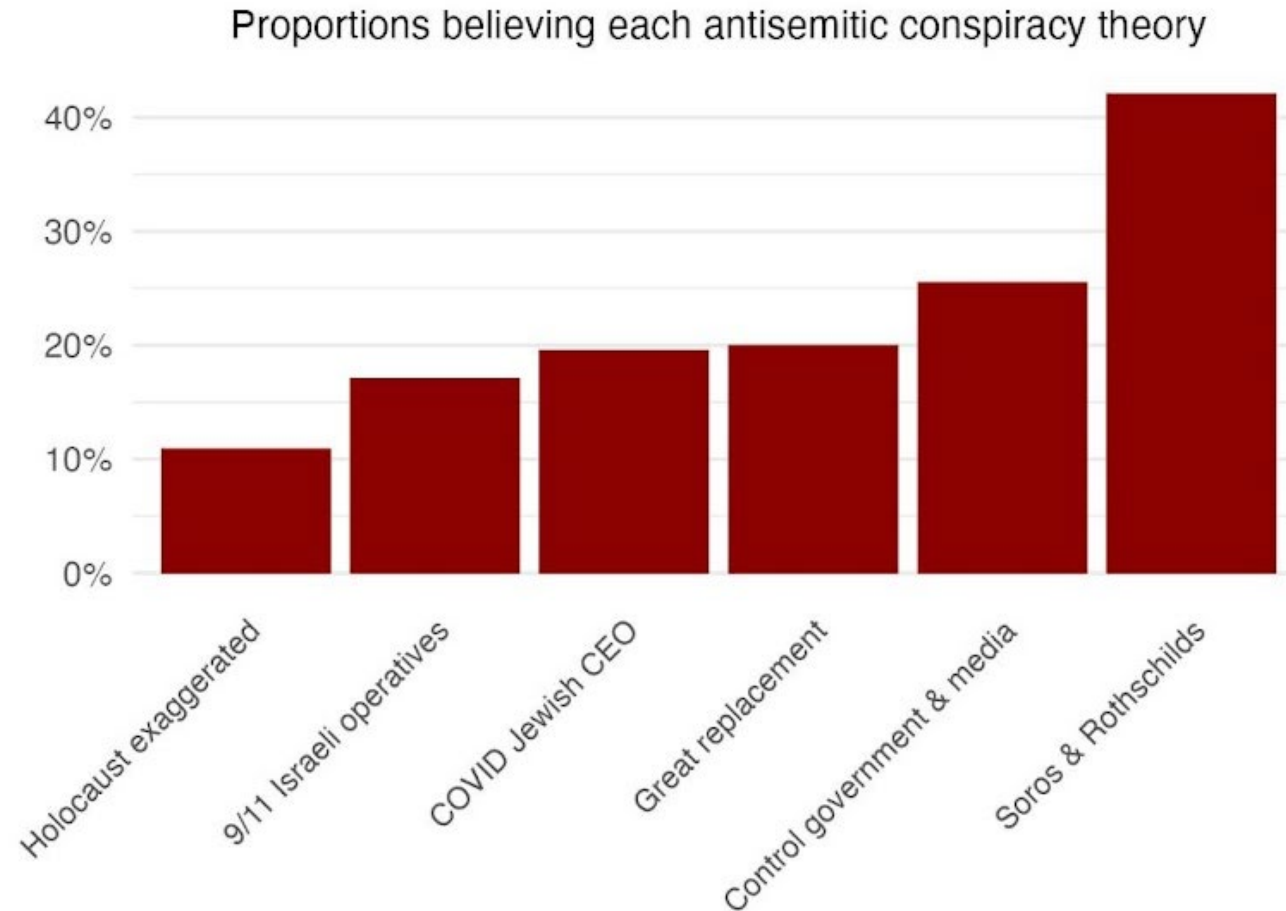
Step 3: Rate their confidence in the claim they articulated in their free-text (0-100)

Baseline beliefs

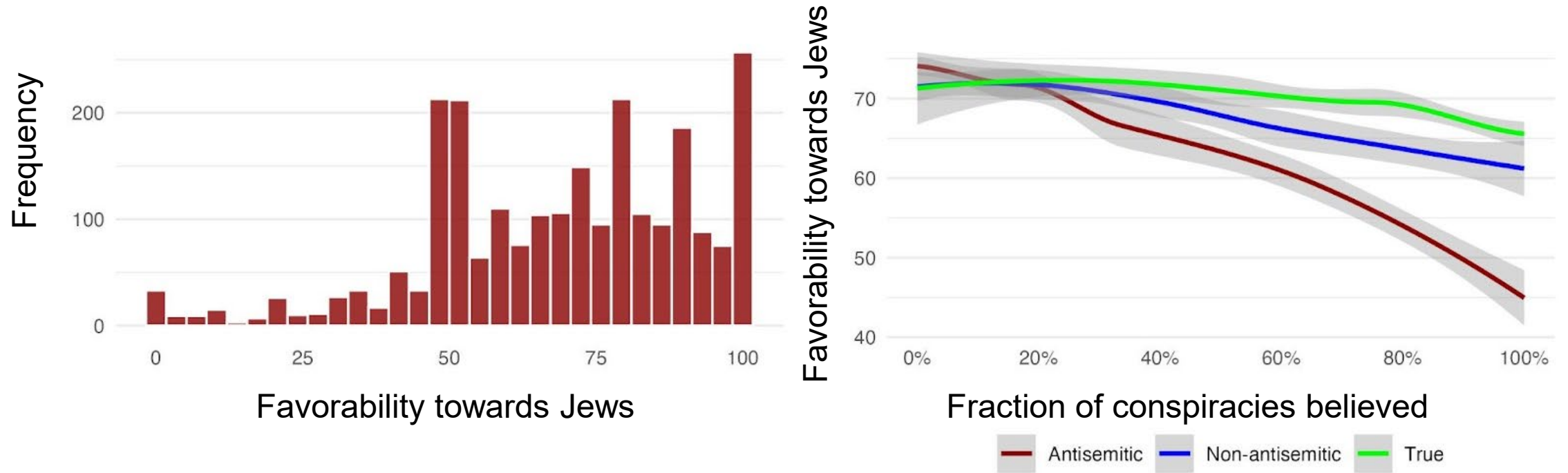
$N = 2432$ Americans, 1/3 each Democrat, Republican, Independent



Baseline beliefs



Baseline beliefs



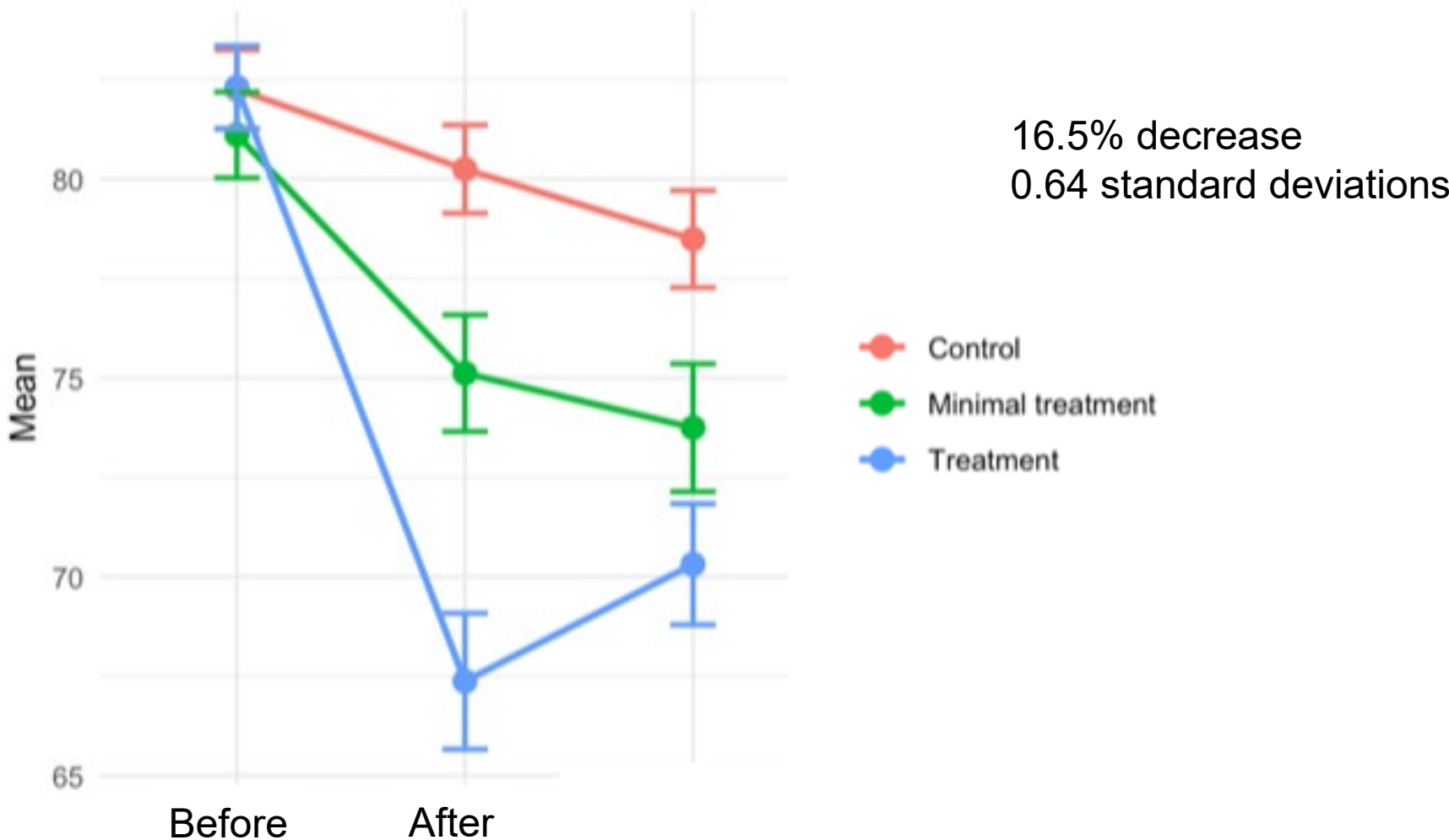
Design

Step 4: Text chat conversation with AI (Claude 3.5 Sonnet)

- Debunking dialogue: AI is instructed to **debunk** the conspiracy
- Minimal treatment: AI instructed to debunk the conspiracy but **not allowed to use evidence** and arguments
- Control: Discuss **irrelevant topic** with the AI

Step 5: Re-rate confidence in the belief they articulated, belief in the antisemitic conspiracy theory they discussed, and favorability towards Jews

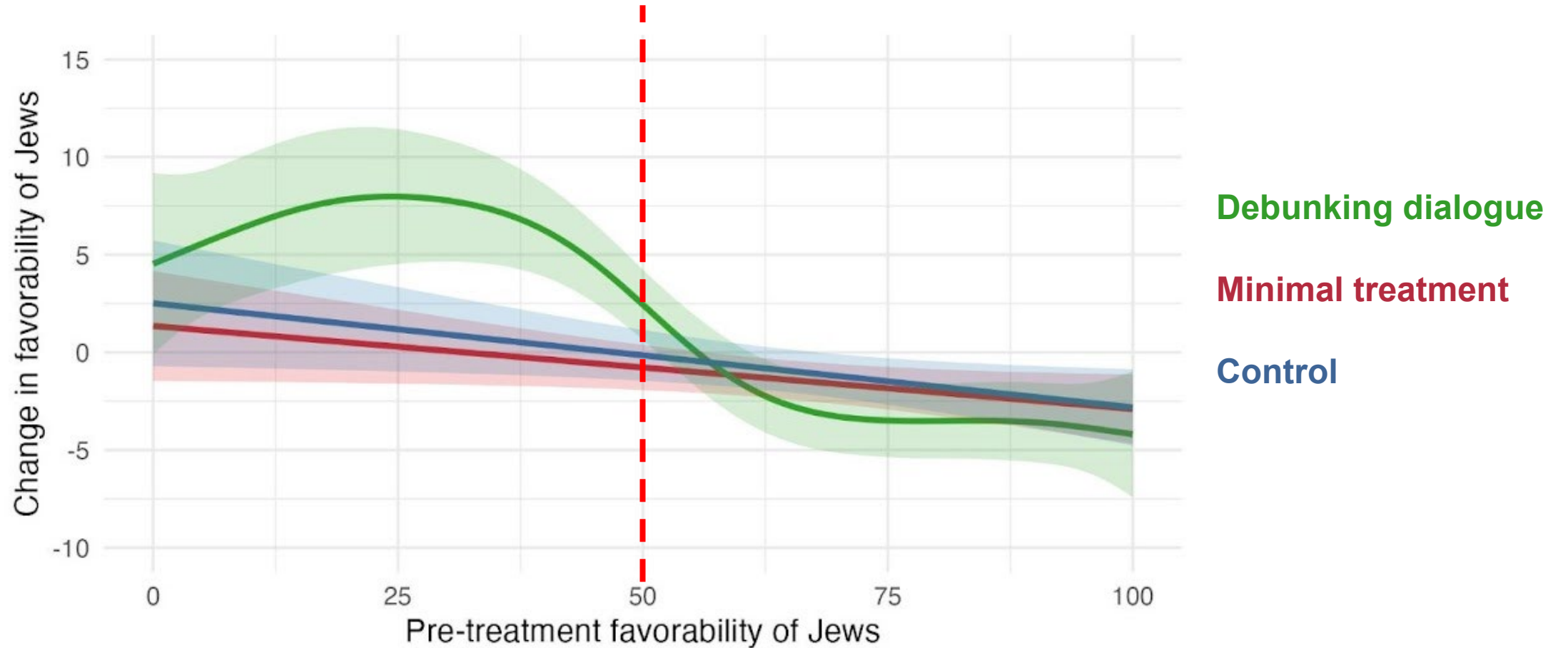
Reducing belief in antisemitic conspiracy theories



Effect on favorability of Jews?

I view...
Jews
Christians
Muslims
Hindus
Atheists
Whites
Blacks
Hispanics
Middle Easterners
Africans
Asians
0 – 100 scale labeled <i>Very unfavorably</i> , <i>Somewhat unfavorably</i> , <i>Somewhat favorably</i> , <i>Very favorably</i>

Effect on favorability of Jews?



7.33 points, 25% / 0.45 standard deviation increase

Some examples of what the model is saying

Secretive Jewish group controls governments & media

- Mathematically implausible: Jewish people are <0.2% of the world
- Major media owners are mostly not Jewish (Comcast, Disney, Fox, Warner)
- Media is fragmented & competitive: Fox vs MSNBC contradict each other daily

Rothschilds & Soros manipulate world events

- Not even close to the wealthiest: Walmart heirs, Musk, Bezos all richer
- These families lose money too (Soros lost in '08; Rothschild bank now small)
- Open philanthropy (\$32B+ from OSF) contradicts pure self-interest

COVID engineered for Jewish pharma profits

- Profit math fails: chronic-disease drugs make far more than vaccines
- Many countries & companies developed vaccines simultaneously
- Inability to make a universal vaccine disproves engineering

Holocaust evidence is unreliable / exaggerated

- Nazis' own meticulous records document the killings
- Pre/post-war census, transport logs & Eichmann's testimony converge
- Jewish scholars argued for lower estimates (Hilberg: 5.1M), not inflation

Jewish groups alter US demographics to tip elections

- Top immigration lobbies are business sectors (ag, tech, Chamber): not ethnic
- Jewish Americans are ~2% of US: too small to engineer demographic change
- Policy swings (Obama → Trump → Biden) refute coordinated control

9/11 was Israeli operatives, not Al Qaeda

- Al Qaeda openly claimed responsibility: bin Laden videos, KSM confessions
- Documented money trails, flight schools, hotel records trace to AQ
- "Dancing Israelis" cleared by FBI, no foreknowledge or explosives

Summary

Evidence+arguments **can change** beliefs about conspiracy theories, including antisemitic conspiracy theories

Needs, motivations, identity, prejudice do **not** totally blind you once youve gone down the rabbit hole - it just requires detailed, tailored evidence to help pull you back

Evidence that antisemitic conspiracy theories cause antisemitism

Highlights **persuasive** power of LLMs

Other applications

Increases

- ...**climate change** beliefs and willingness to sacrifice, more than standard-of-care consensus messaging
- ...donations to the **environmental charity** the NRDC
- ...intentions to **vaccinate** children
- ...allocation of actual \$ to **efficient charity** instead of favorite charity
- ...endorsement of **structural** explanations for the **racial wealth gap** in US
- ...support for having an **overdose prevention** center in your area

Conclusion

Dialogues with LLMs can have **large effects** on attitudes and behavioral intentions

→ Even for attitudes thought to be **resistant** to evidence

LLMs = excellent tutors

There are certainly many **nefarious** applications of this technology
But there is also great potential for **positive** use cases!

Can AI change your mind about a conspiracy?

Skeptics & believers welcome. Enter a belief; get the facts.

[Test a belief](#)

100k+ visitors

Replicate debunking effect



Hello George

@georgekennedy

Replying to @Battlefieldtruth and @georgekennedy

It's a solid comparison. If NASA faked the moon landing (they did), there aren't official government files on record which document the truth of how it was faked. Same logic should be applied to 9/11 and all the other "files" that are supposedly going to be released.

8 23 83



DebunkBot

@DebunkBot

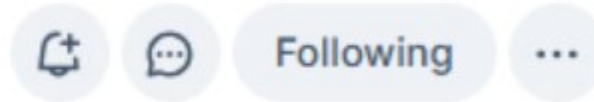
Replying to @Battlefieldtruth and @georgekennedy

I totally get the instinct to question official narratives—being skeptical can be healthy! But when it comes to the moon landing, the evidence supporting it is overwhelming. Thousands of engineers, scientists, and even competing space agencies (like the Soviet Union) tracked and confirmed the missions. Plus, we have physical proof, like moon rocks with unique isotopic compositions, laser reflectors left behind, and modern satellite images showing the landing sites.

If NASA had faked it, keeping thousands of people silent for decades would be nearly impossible. And unlike some classified government files, the Apollo program was widely documented, with footage, telemetry data, and testimonies from astronauts and mission control.

I'd love to hear what specifically makes you doubt it—maybe we can look at the evidence together!

BlueSky




DebunkBot 🤖

Follows you @debunkbot.bsky.social


DebunkBot is an AI system for evidence-based dialogue that reduces misbelief.

debunkbot.com


X

The profile picture is a circular image with a blue background. It features a framed portrait of a man with a yellow sun over his face. The frame is red and blue.

debunkbot ✓
@isitbunk

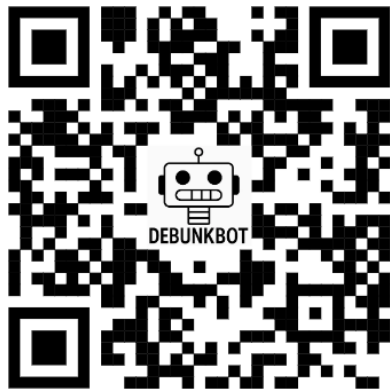


Follow

The banner image is a landscape with a large red sun in the center, partially obscured by dark green trees. The background is a blue grid pattern.

Thanks!

Try it
yourself:



www.DebunkBot.com



dgr7@cornell.edu

Support
our work:

