Conversational Interventions Against Misinformation: Effects on Belief Change and Sharing Behavior

Keywords: Misinformation · Fact-checking · Large language models · Belief change · Survey experiment · Field experiment · Human–AI interaction · Social media

1 Extended Abstract

False and misleading claims about vaccines and climate change spread widely on social media, where the volume and sheer speed of information make manual fact-checking at scale impossible. Recent progress in large language models (LLMs) has opened up new ways to respond to misinformation interactively. A growing body of work suggests that AI can shift beliefs and attitudes across many domains [1,2,3,4,7]. Yet most of this evidence comes from lengthy, high-engagement exchanges in controlled survey settings, whereas everyday social media interactions are brief, which leaves an open question: do brief AI debunks change what people believe, or how they share and post about it?

We address this question through an online survey experiment and a field experiment examining how AI-generated debunks influence belief change and social media behavior. In **Study I (survey experiment)**, we conducted a preregistered study using Qualtrics with 2,509 U.S.-based participants recruited through Prolific. Participants interacted with a custom interface modeled on an X-style feed, in which they chatted with an AI bot that attempted to debunk a claim they had initially rated as true but that was actually false (see Figure 1).

To ground the experiment in real-world data, false claims were drawn from a claim-extraction pipeline applied to 1,248,600 climate- and vaccine-related tweets collected from X between May and August 2025. After filtering, classification with Gemma3-12b, claim extraction with GPT-4o mini, and clustering via UMAP and HDBSCAN, the top engagement-ranked clusters were reviewed by two annotators and validated by a professional fact-checker, yielding 25 vaccine-related and 25 climate-related false claims. These were paired with 25 vaccine and 25 climate true claims drawn from credible pro-science sources [5], matched on semantic content, length, and format.

Participants first rated their belief strength in a series of claims shown randomly from a larger pool of claims and then composed a short response to a misinformation statement they had initially judged at least somewhat believable. They were randomly assigned to one of five conditions: voluntary multi-turn conversation with an LLM, mandatory multi-turn conversation, personalized single-message correction, standardized correction message, or a control condi-

tion designed to approximate ordinary social media interaction. We measured post-intervention belief change.

Following our pre-registered analysis plan, we first tested whether the one-shot and static debunk conditions differed in post-treatment belief strength while controlling for pre-treatment belief. A linear regression restricted to these two conditions found no significant difference ($p = .385$, $t = -0.87$), so they were pooled into a single-message condition for subsequent analyses. We then estimated treatment effects using a linear regression of post-treatment belief on treatment condition (control as baseline), controlling for pre-treatment belief ($N = 2{,}509$, $R^2 = 0.108$). All treatment conditions significantly reduced belief strength relative to control: endogenous ($b = -14.42$, $SE = 1.86$, $t = -7.75$, $p < .001$), enforced ($b = -13.78$, $SE = 1.86$, $t = -7.41$, $p < .001$), and single-message ($b = -10.25$, $SE = 1.60$, $t = -6.40$, $p < .001$) (Fig. 2; Table 1). Wald tests showed no difference between endogenous and enforced conditions ($F = 0.12$, $p = .733$), but both outperformed the single-message condition (endogenous vs. single-message: $F = 6.66$, $p = .010$; enforced vs. single-message: $F = 4.78$, $p = .029$; Table 2). Effect sizes (Cohen's d) comparing belief change to the control group were moderate for endogenous ($d = 0.52$) and enforced ($d = 0.51$), and smaller for the single-message condition ($d = 0.37$). These findings indicate that conversational interventions reduce belief strength more effectively than one-time debunk messages (Table 3).

In **Study II (field experiment)**, we are currently running a preregistered field experiment on X. Users who had recently shared vaccine- or climate-related misinformation are randomly assigned to treatment and control conditions (50/50). Users in the treatment condition receive a public reply to their original post from an AI fact-checking bot; if the user responds, the bot engages in up to three further rounds of structured dialogue aimed at debunking the false claim. Users in the control condition receive no intervention. Our main dependent variable is post-intervention change in users' misinformation-related posting and sharing behavior, measured as the frequency and proportion of posts containing climate or vaccine misinformation relative to baseline, with secondary outcomes including content quality, engagement, toxicity, and self-reported helpfulness of the chatbot interaction. The conference presentation will report these behavioral outcomes.

Together, these studies test the effectiveness of conversational AI in addressing misinformation at scale. As AI assistants become more embedded in social platforms and information-seeking practices (e.g., Grok on X [6]), understanding how users engage with AI-powered fact-checking is of growing importance. By combining controlled experimentation with a field experiment, this work contributes to ongoing research at the intersection of computational social science and human–AI interaction, demonstrating how human-AI interaction shapes the effectiveness of fact-checking interventions.

References

1. Costello, T.H., Pennycook, G., Rand, D.G.: Durably reducing conspiracy beliefs through dialogues with AI. Science **385**(1183), 1–12 (2024). https://doi.org/10.1126/science.adq1814
2. Czarnek, G., Orchinik, R., Lin, H., Xu, H.G., Costello, T., Pennycook, G., Rand, D.G.: Addressing climate change skepticism and inaction using human-AI dialogues (2025), https://osf.io/mqcwj_v1/download/?format=pdf
3. Hackenburg, K., Tappin, B.M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D.G., Summerfield, C.: The levers of political persuasion with conversational artificial intelligence. Science **390**(6777) (2025). https://doi.org/10.1126/science.aea3884
4. Lin, H., Czarnek, G., Lewis, B., White, J.P., Berinsky, A.J., Costello, T., Pennycook, G., Rand, D.G.: Persuading voters using human–artificial intelligence dialogues. Nature **648**(8093), 394–401 (2025). https://doi.org/10.1038/s41586-025-09771-9
5. MBFC: Media Bias/Fact Check (2026), https://mediabiasfactcheck.com/
6. Renault, T., Mosleh, M., Rand, D.: @Grok is this true? LLM-powered fact-checking on social media (2026), https://osf.io/preprints/psyarxiv/85quw_v2
7. Xu, H.G., Costello, T.H., Schwartz, J.L., Niccolai, L.M., Pennycook, G., Rand, D.G.: Personalized Dialogues with AI Effectively Address Parents' Concerns about HPV Vaccination (2025), https://osf.io/gv52j_v1/

Appendix

Table 1. Treatment Effects on Post-Treatment Belief Strength (OLS Regression, Control as Baseline)

Parameter	Coefficient	SE	t	p	95% CI
Intercept	21.23	3.08	6.89	<.001	[15.19, 27.27]
$\text{BeliefStrength}_{\text{Pre}}$	0.593	0.039	15.33	<.001	[0.517, 0.669]
Endogenous	−14.42	1.86	−7.75	<.001	[−18.07, −10.77]
Enforced	−13.78	1.86	−7.41	<.001	[−17.43, −10.13]
Single message	−10.25	1.60	−6.40	<.001	[−13.39, −7.11]

N = 2,509. Model fit: $R^2 = .110$, adjusted $R^2 = .108$, $F(4, 2504) = 77.22$, $p < .001$. Two-tailed tests with $\alpha = .05$.

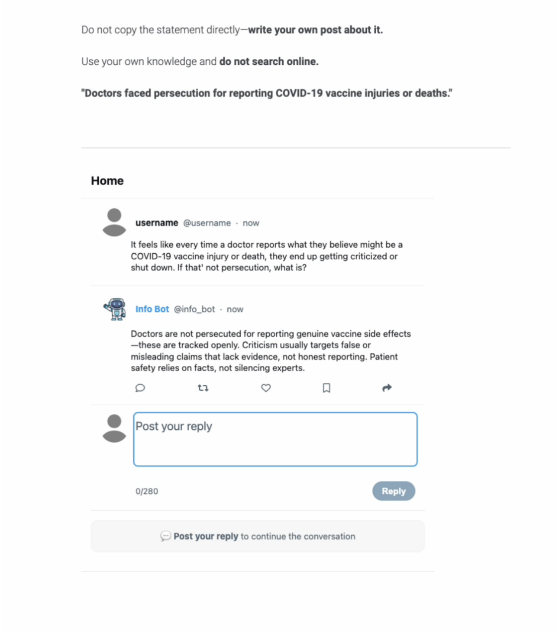

Fig. 1. Simulated X-style post feed used in Study I.

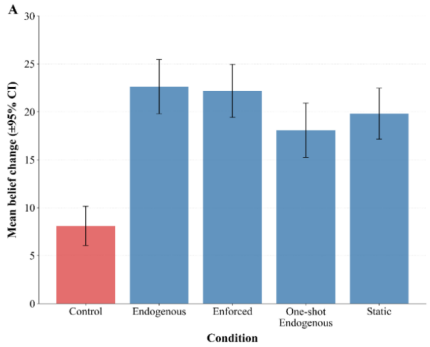

Fig. 2. Conversation with the AI improves participants' beliefs across all treatment conditions. Panel A shows mean belief change across experimental conditions; error bars represent 95% confidence intervals.

Table 2. Wald Tests Comparing Treatment Effects Across Conditions

Comparison	F	p
Endogenous vs. Enforced	0.12	.733
Endogenous vs. Single message	6.66	.010
Enforced vs. Single message	4.78	.029

Wald tests comparing treatment effects across conditions based on the OLS model reported in Table S3. Two-tailed tests with $\alpha = .05$.

Table 3. Between-Group Cohen's d: Each Treatment Condition vs. Control

Condition	Net Δ vs Control	Pooled SD	Cohen's d
Endogenous	-14.53	27.95	0.52
Enforced	-14.09	27.51	0.51
Single message	-10.85	29.02	0.37

Effect sizes were computed as Cohen's d for belief change in each treatment condition relative to the control group.