Prebunking Elections Rumors: Artificial Intelligence Assisted Interventions Increase Confidence in American Elections

Mitchell Linegar[1*], Betsy Sinclair[2], Sander van der Linden[3]
R. Michael Alvarez[1]
[1]Linde Center for Science, Society, and Policy, California Institute of Technology, Pasadena & 91125, USA.
[2]Department of Political Science, Washington University in St. Louis, St. Louis & 63130, USA.
[3]Department of Psychology, Cambridge University, Cambridge & CB2 3EB, UK.
*Corresponding author. Email: mlinegar@caltech.edu

TexMeth, Feb 21, 2025

Election Rumors and Conspiracies

- Widespread and widely believed
- Strong partisan trends
- Actively endorsed by political elites
- Undermine democracy, democratic participation
- Discourage peaceful transition of power

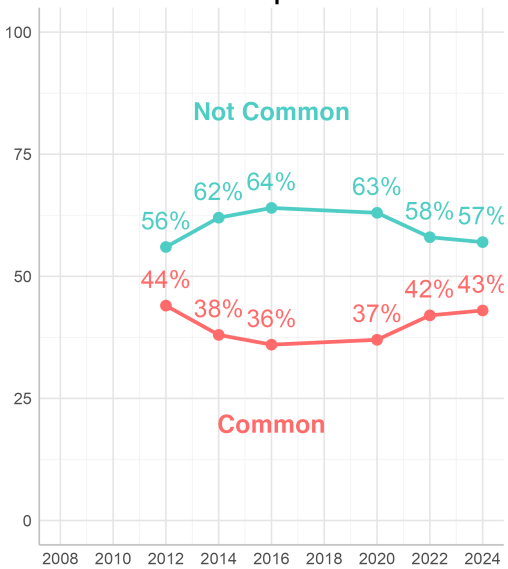
A Need for Tools to Combat Disinformation

- Enormous space of misinformation
- Debunking and prebunking can reduce belief in misinformation
- Individually pushing back against false allegations is costly and slow
- Enter AI

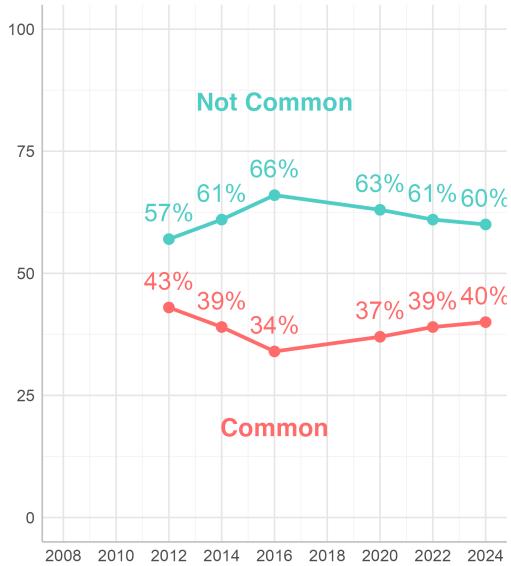
This paper

- Preregistered, two-wave experimental study of U.S. registered voters
- YouGov panel ($N = 4,293$)
- Goal: prebunk / inoculate against election disinformation
- Test five common and widespread election myths
- Use AI to automatically produce inoculation doses
- AI-generated prebunks reduce belief in election conspiracies, increase belief in election integrity

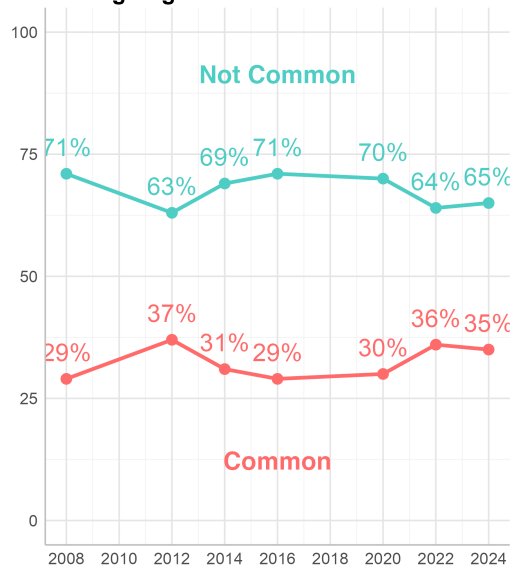
People voting an absentee ballot intended for another person



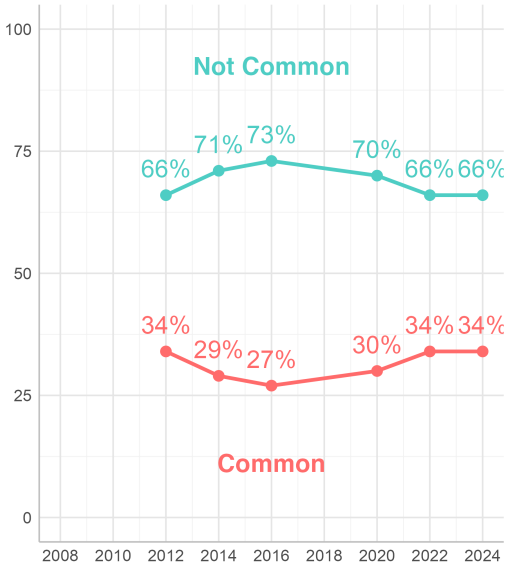
People voting who are not U.S. citizens



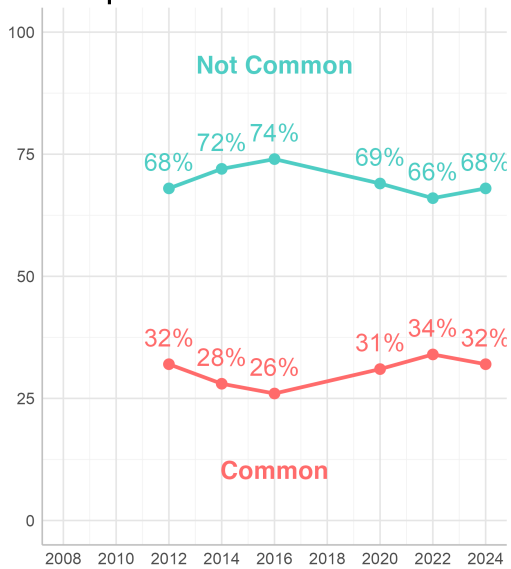
People pretending to be someone else when going to vote



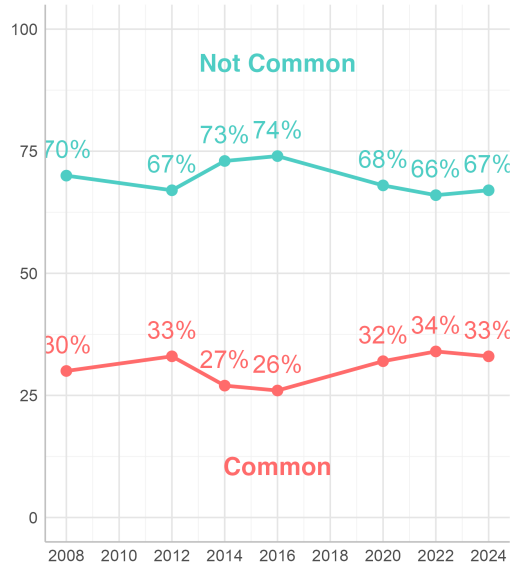
People voting more than once in an election

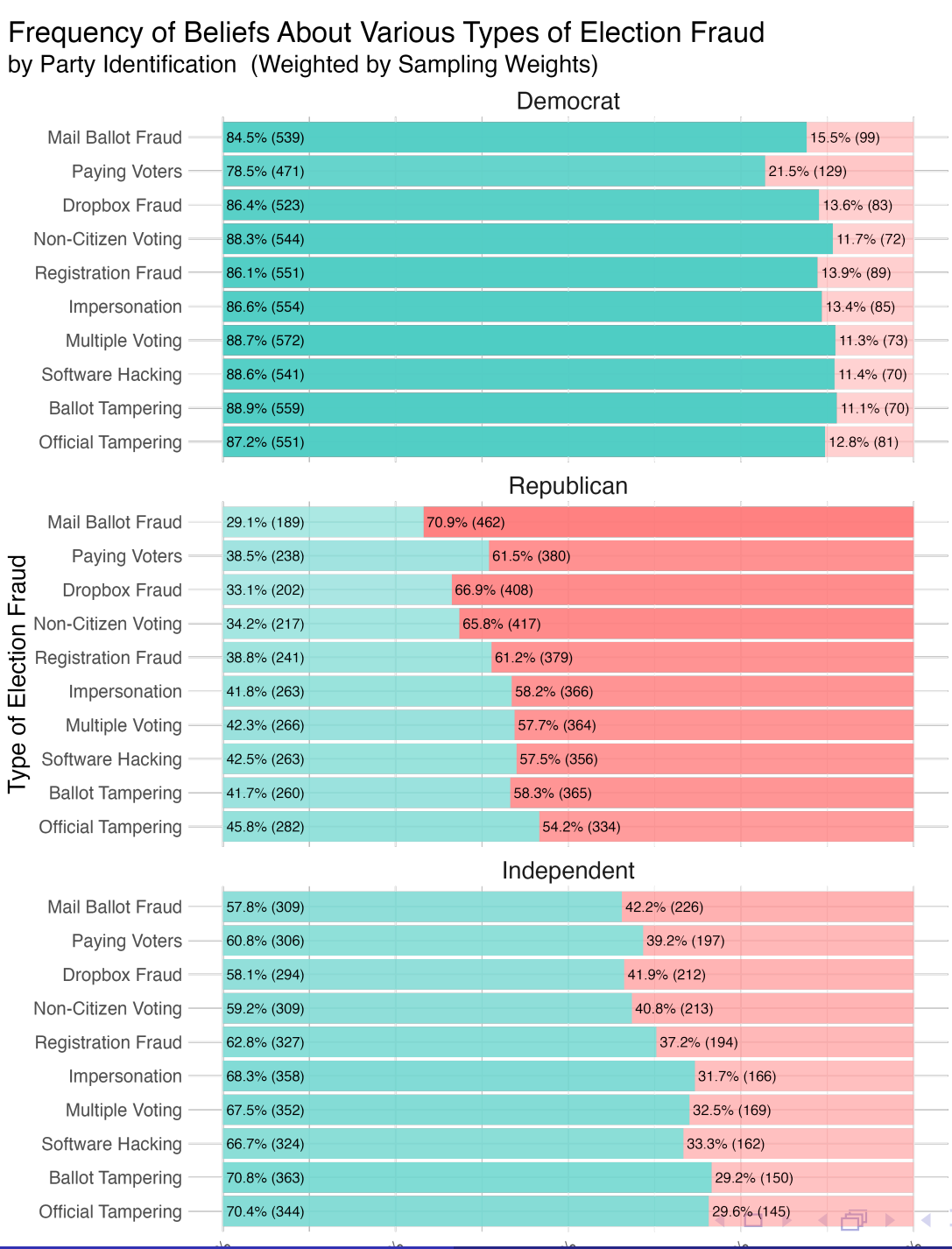


Officials fraudulently changing the reported vote count



People stealing or tampering with ballots that have been voted




Frequency of Beliefs About Various Types of Election Fraud
by Party Identification (Weighted by Sampling Weights)
Democrat
Mail Ballot Fraud 84.5% (539) 15.5% (99)
Paying Voters 78.5% (471) 21.5% (129)
Dropbox Fraud 86.4% (523) 13.6% (83)
Non-Citizen Voting 88.3% (544) 11.7% (72)
Registration Fraud 86.1% (551) 13.9% (89)
Impersonation 86.6% (554) 13.4% (85)
Multiple Voting 88.7% (572) 11.3% (73)
Software Hacking 88.6% (541) 11.4% (70)
Ballot Tampering 88.9% (559) 11.1% (70)
Official Tampering 87.2% (551) 12.8% (81)
Republican
Type of Election Fraud
Mail Ballot Fraud 29.1% (189) 70.9% (462)
Paying Voters 38.5% (238) 61.5% (380)
Dropbox Fraud 33.1% (202) 66.9% (408)
Non-Citizen Voting 34.2% (217) 65.8% (417)
Registration Fraud 38.8% (241) 61.2% (379)
Impersonation 41.8% (263) 58.2% (366)
Multiple Voting 42.3% (266) 57.7% (364)
Software Hacking 42.5% (263) 57.5% (356)
Ballot Tampering 41.7% (260) 58.3% (365)
Official Tampering 45.8% (282) 54.2% (334)
Independent
Mail Ballot Fraud 57.8% (309) 42.2% (226)
Paying Voters 60.8% (306) 39.2% (197)
Dropbox Fraud 58.1% (294) 41.9% (212)
Non-Citizen Voting 59.2% (309) 40.8% (213)
Registration Fraud 62.8% (327) 37.2% (194)
Impersonation 68.3% (358) 31.7% (166)
Multiple Voting 67.5% (352) 32.5% (169)
Software Hacking 66.7% (324) 33.3% (162)
Ballot Tampering 70.8% (363) 29.2% (150)
Official Tampering 70.4% (344) 29.6% (145)

Hypotheses

- **H1**: Participants exposed to prebunking of a specific election-related rumor will report lower confidence in that rumor compared to the control group.
- **H2**: Participants exposed to prebunking of a specific election-related rumor will report higher confidence that their votes will be accurately counted in the next election compared to the control group.

Experimental Design

- Participants answer pre-treatment questions
- Pick five salient election rumors, each participant assigned to one
- Choose Breitbart articles endorsing each rumor
- All participants read rumor-relevant article ("Full Exposure Article")
- Prior to this, read either AI-written "Inoculation Article" or AI-written "Placebo Article"
- Participants answer post-treatment questions, and again one week later

AI-Written Inoculation Articles

- Input: article endorsing false rumor, CISA fact sheet, prompt
- Human-in-the-loop process: iterate initial prompt until can produce satisfactory inoculation articles for a single rumor
- Use same prompt for all other rumors
- Randomly assign HITL and purely LLM generated articles

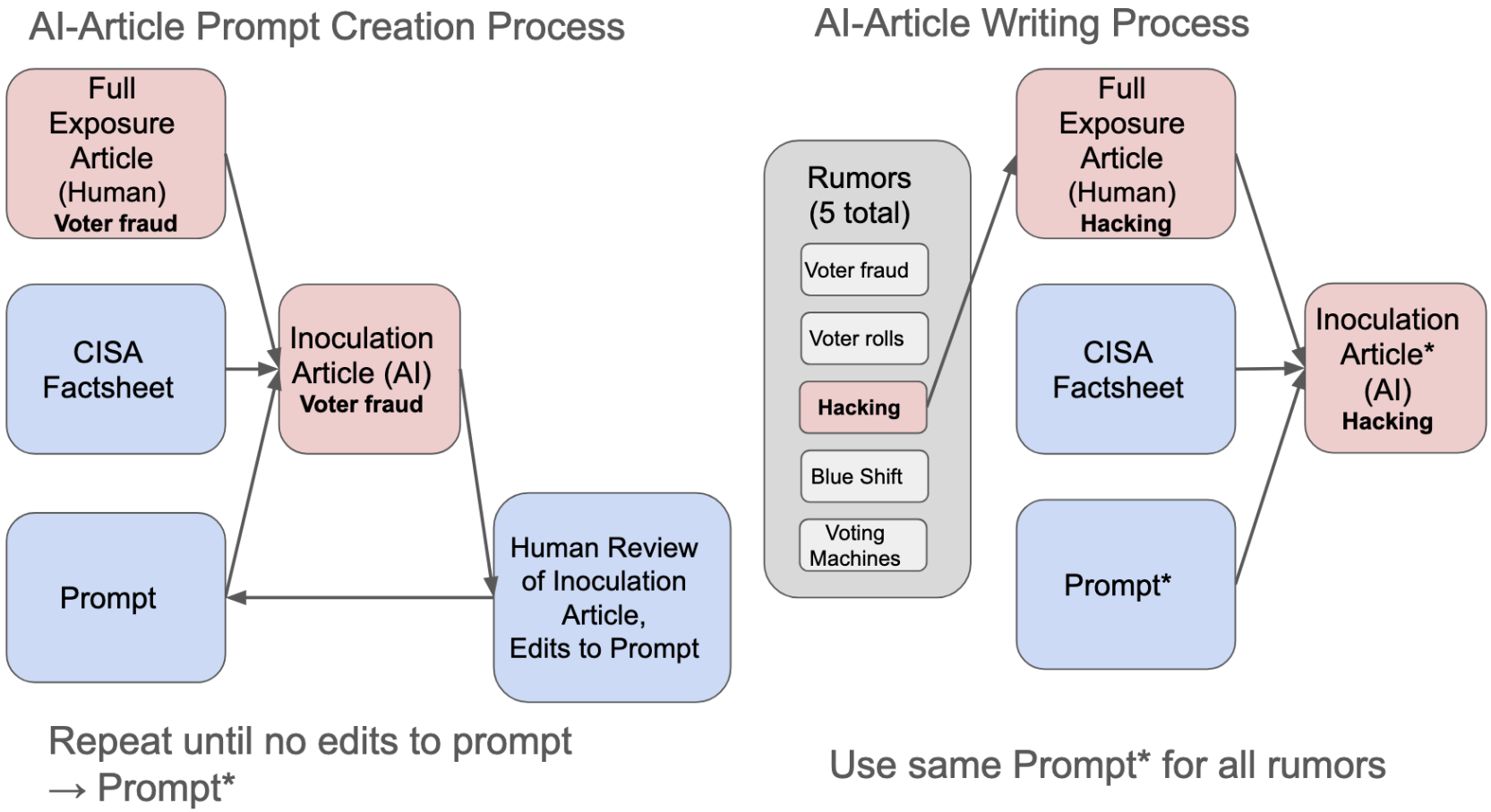


Figure: **Prompt and article writing procedure.**

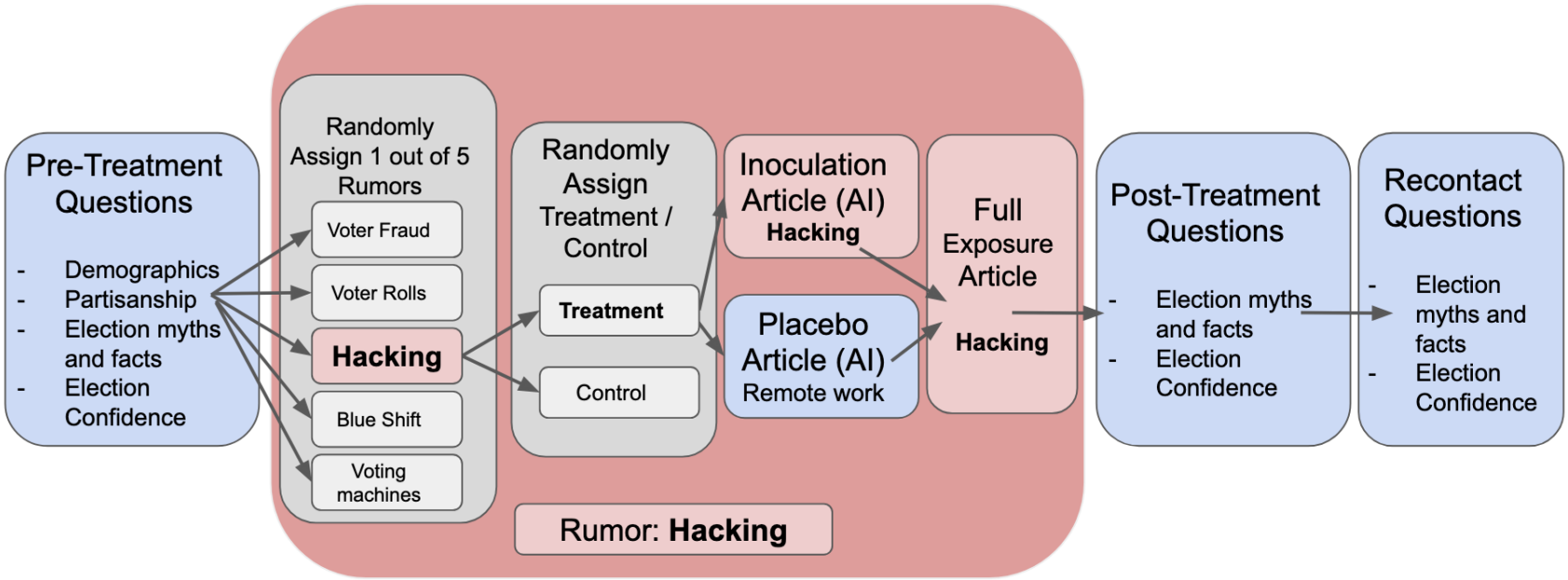
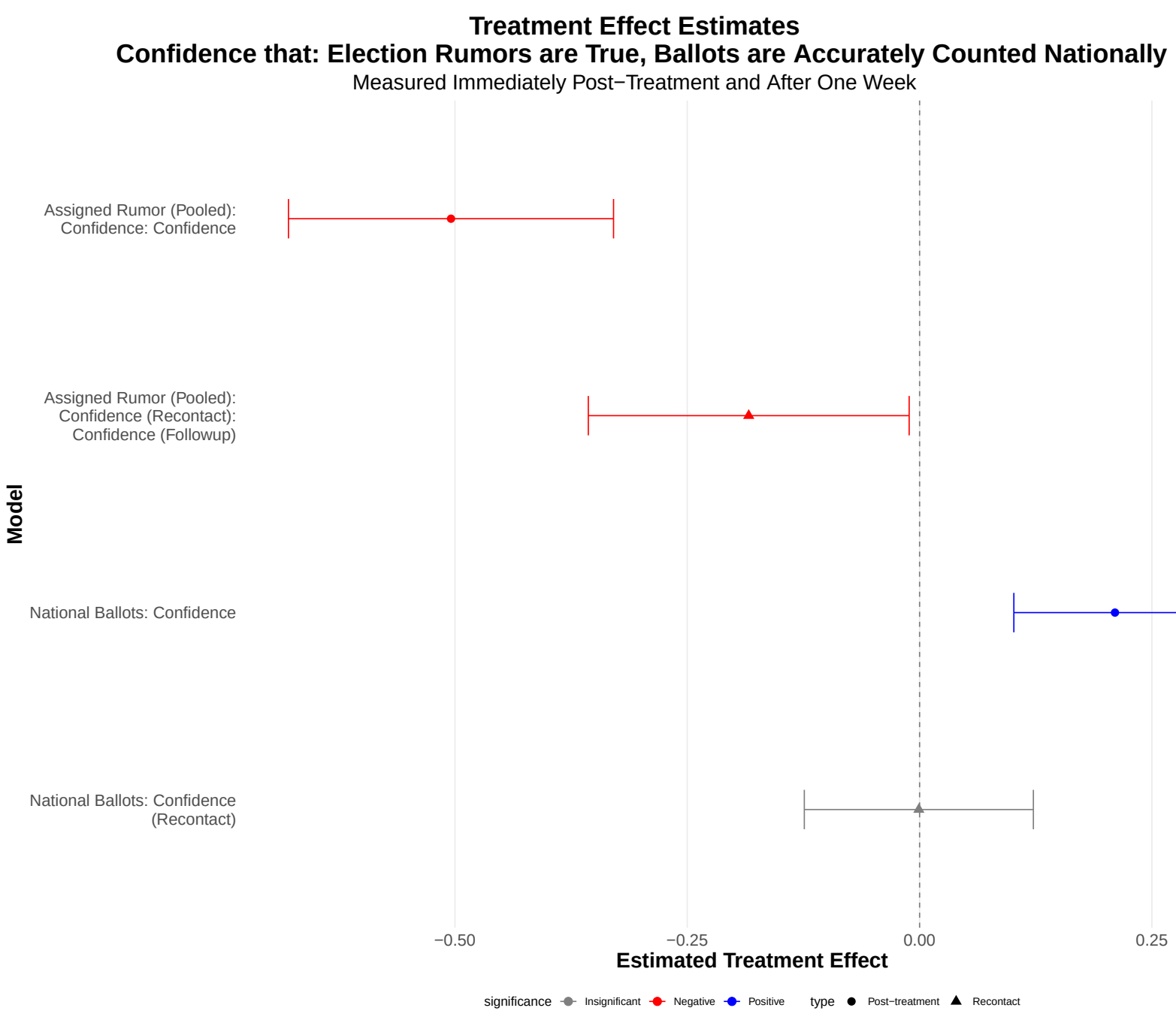
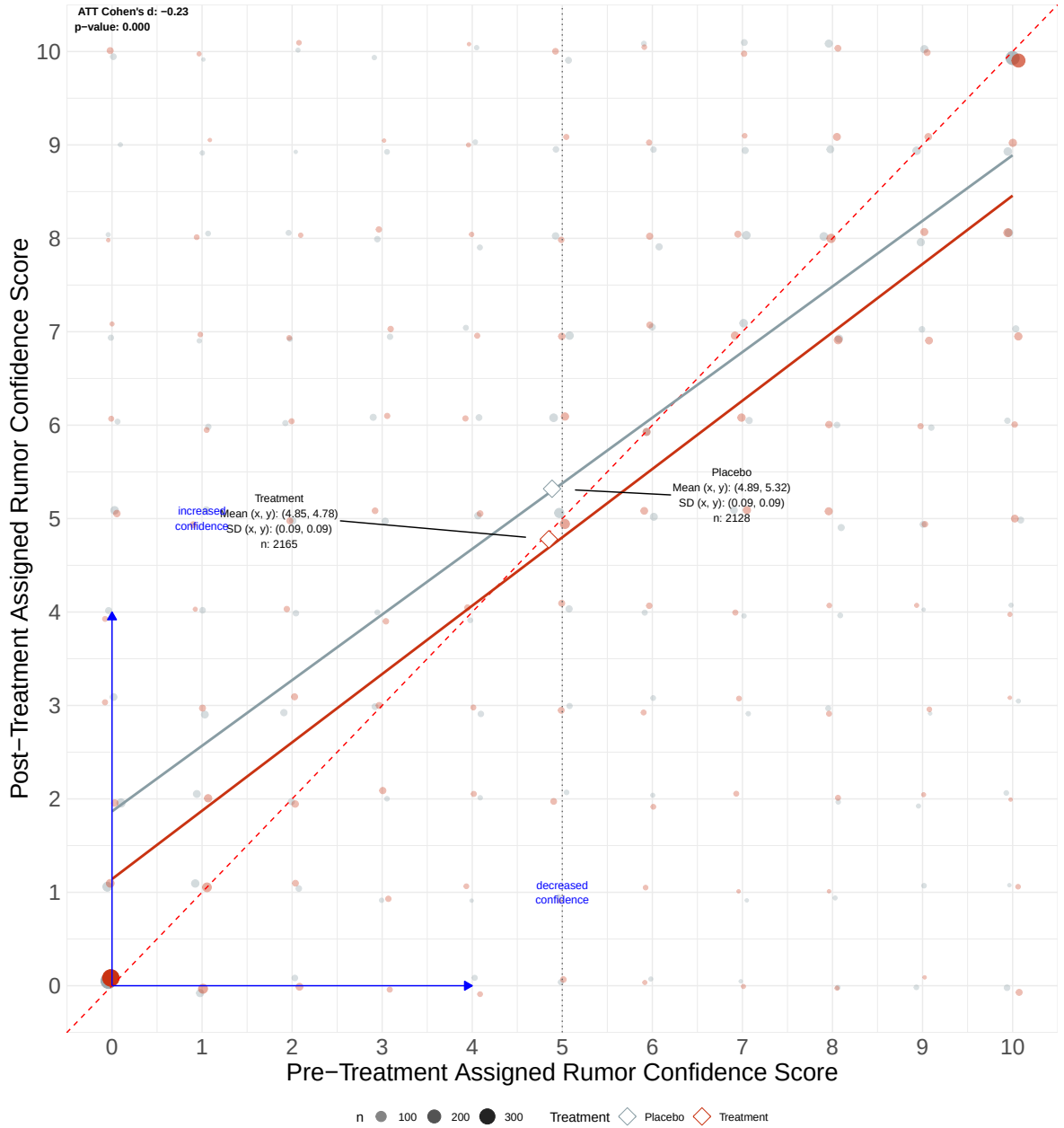


Figure: **Experimental design.** Blue: common to all participants, regardless of assigned rumor. Grey: where randomization occurs. Red: articles participants are assigned to.

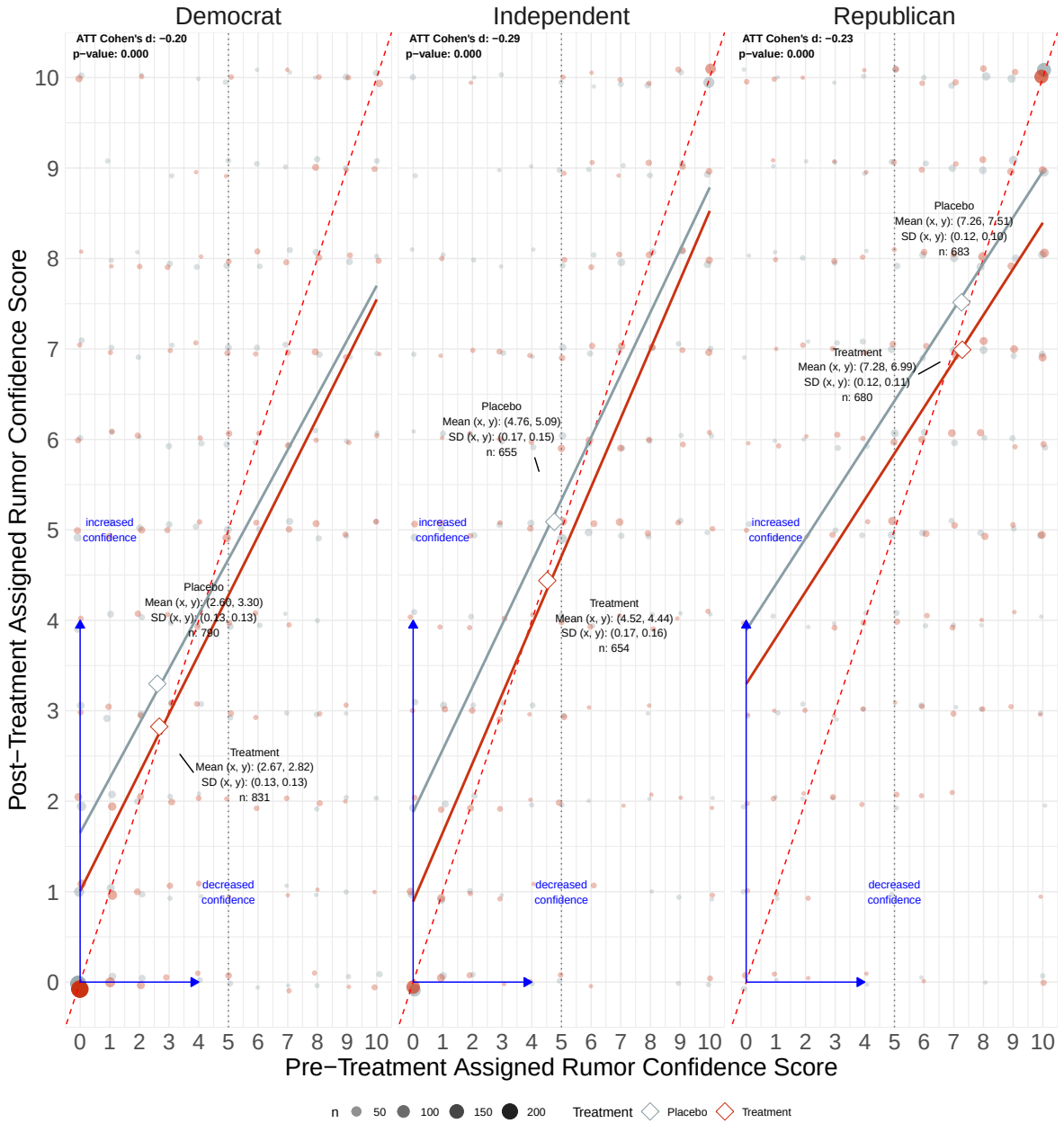

Treatment Effect Estimates
Confidence that: Election Rumors are True, Ballots are Accurately Counted Nationally
Measured Immediately Post−Treatment and After One Week
Assigned Rumor (Pooled): Confidence: Confidence
Assigned Rumor (Pooled): Confidence (Recontact): Confidence (Followup)
National Ballots: Confidence
National Ballots: Confidence (Recontact)
Model
−0.50
−0.25
0.00
0.25
Estimated Treatment Effect
significance
Insignificant
Negative
Positive
type
Post−treatment
Recontact

Confidence in Assigned Election Rumor:
Pre−Treatment vs. Post−Treatment

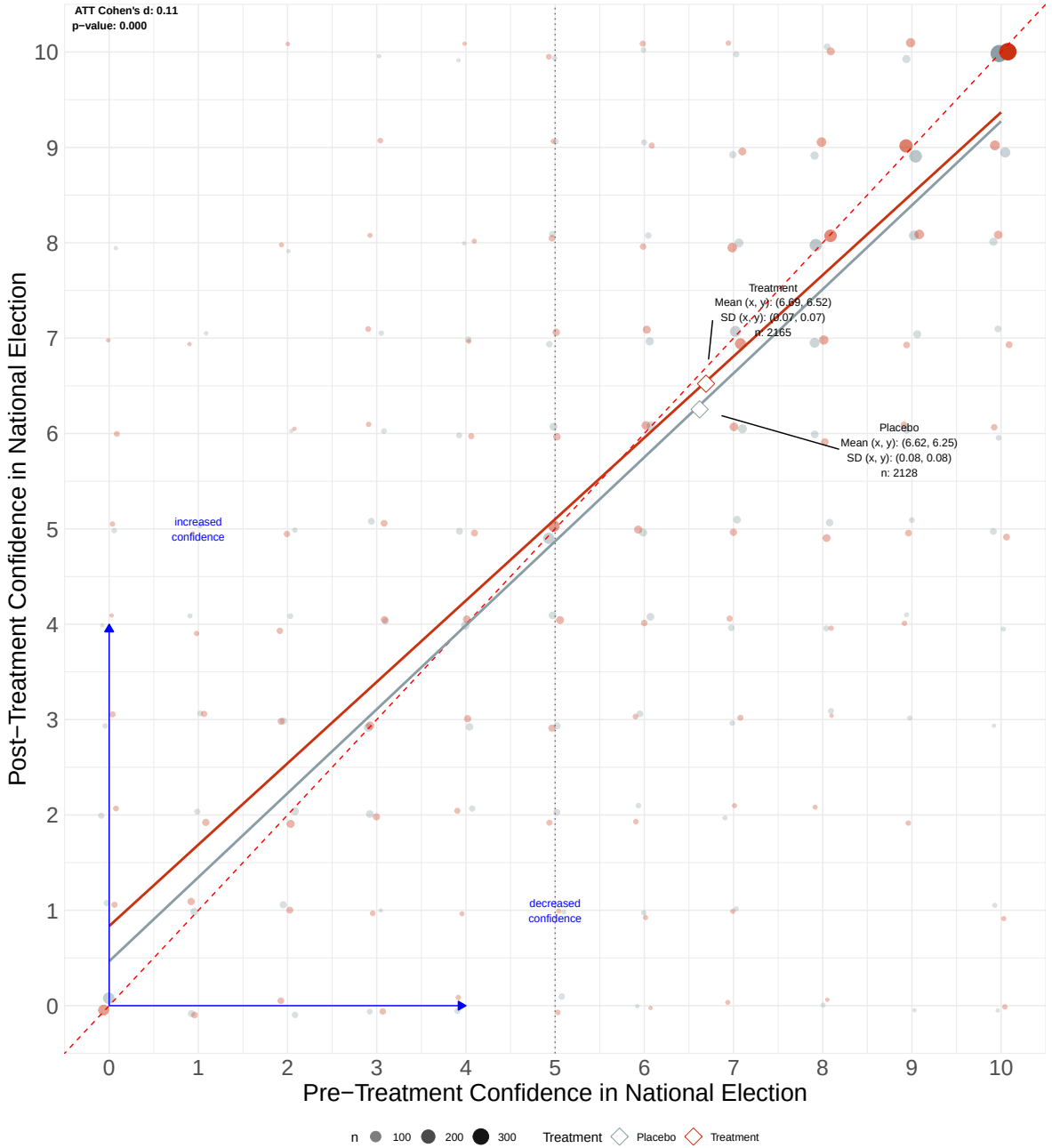


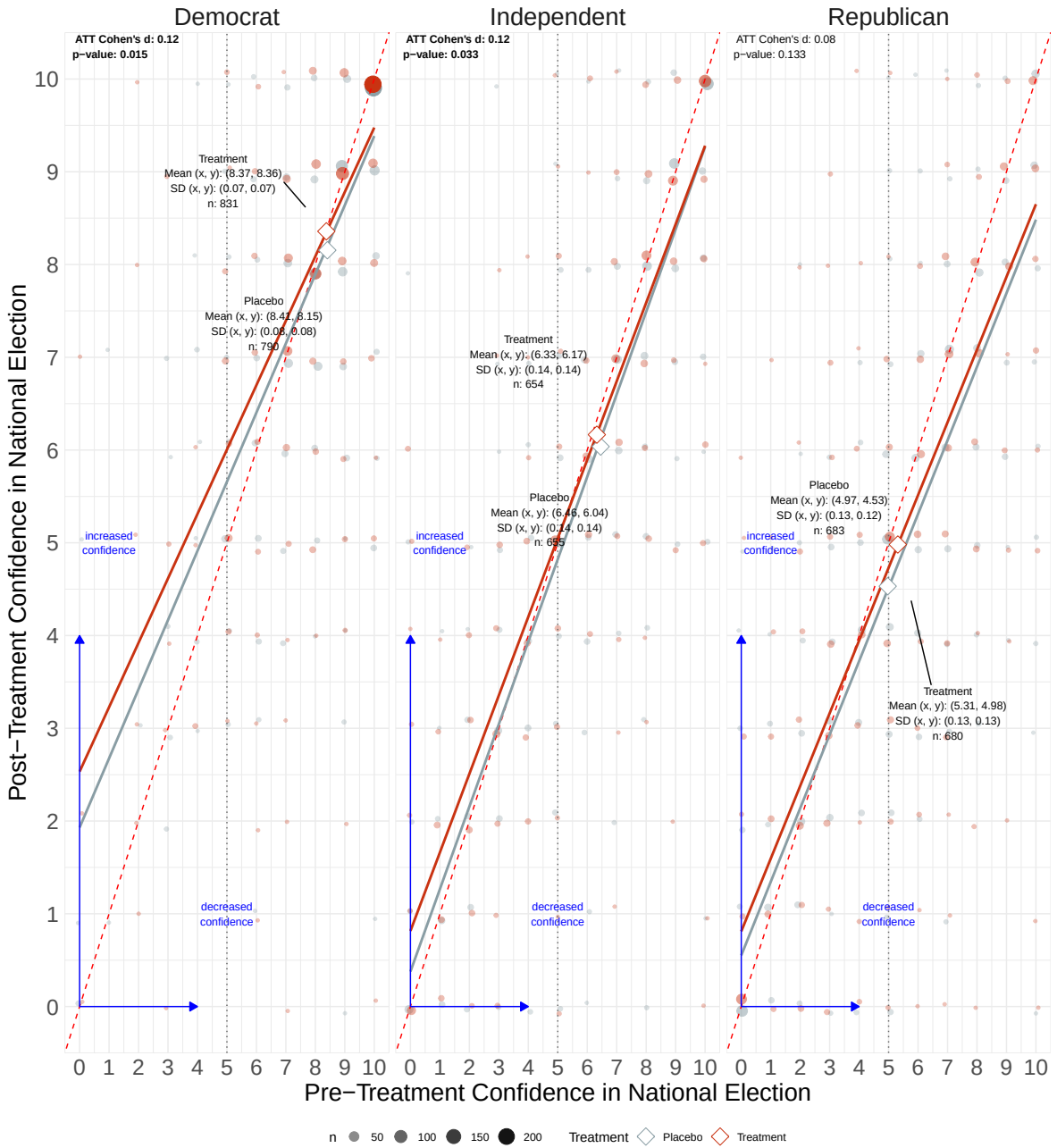
Confidence in Assigned Election Rumor:
Pre−Treatment vs. Post−Treatment

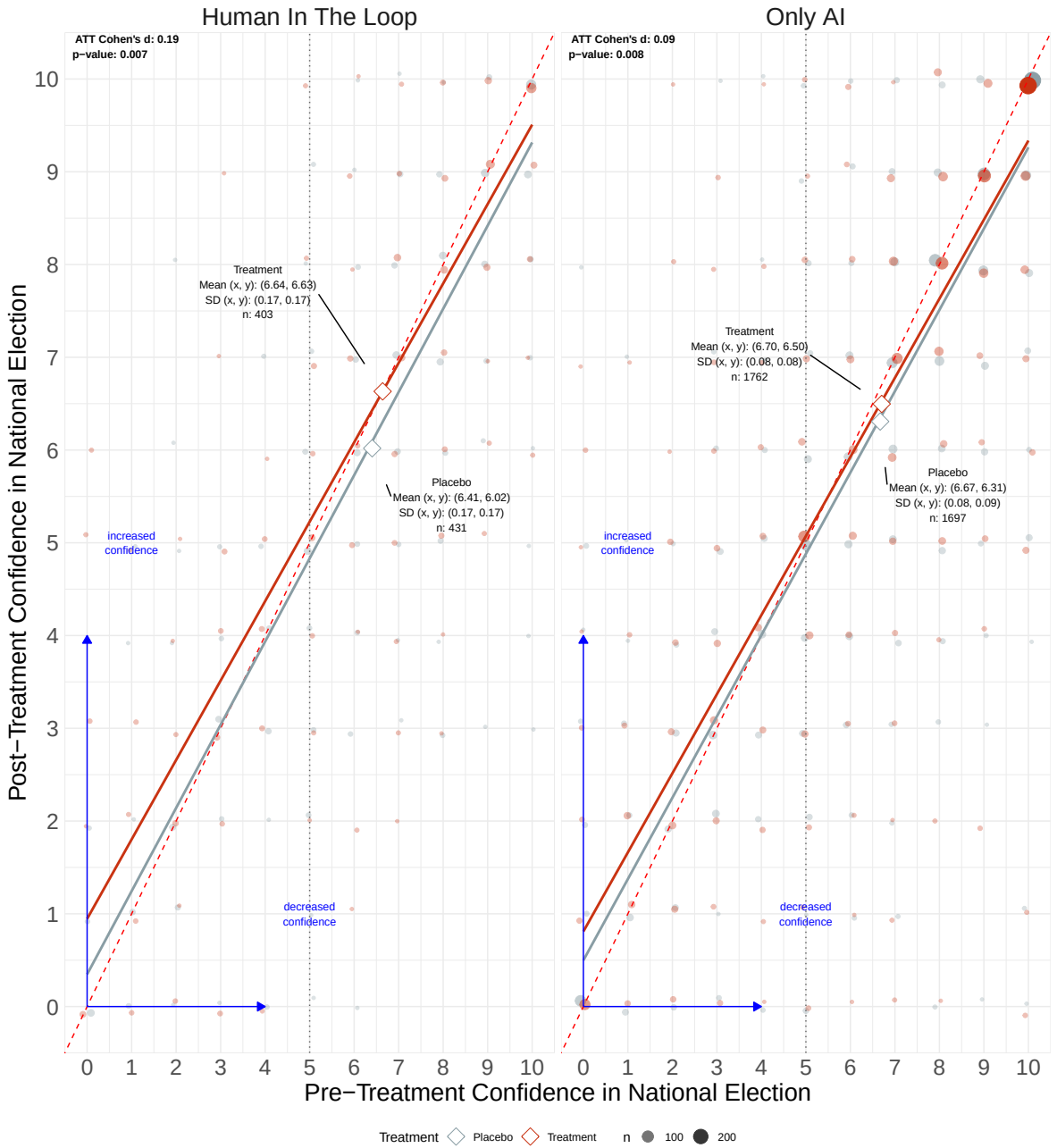
By Party Identification



Confidence in National Election: Pre−Treatment vs. Post−Treatment

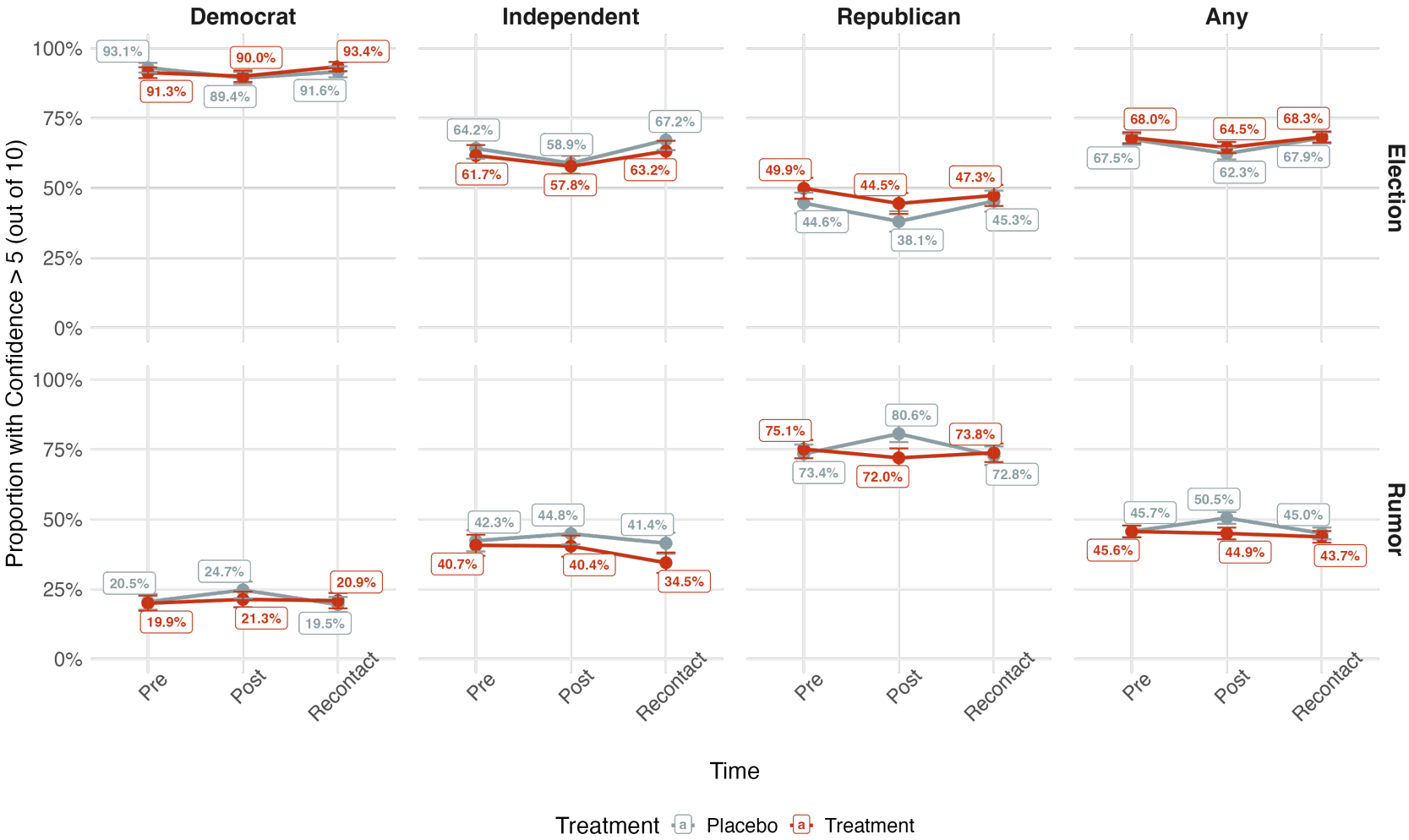



Confidence in National Election: Pre−Treatment vs. Post−Treatment
By Party Identification
Democrat
Independent
Republican
ATT Cohen's d: 0.12
p−value: 0.015
ATT Cohen's d: 0.12
p−value: 0.033
ATT Cohen's d: 0.08
p−value: 0.133
Treatment
Mean (x, y): (8.37, 8.36)
SD (x, y): (0.07, 0.07)
n: 831
Placebo
Mean (x, y): (8.41, 8.15)
SD (x, y): (0.08, 0.08)
n: 790
Treatment
Mean (x, y): (6.33, 6.17)
SD (x, y): (0.14, 0.14)
n: 654
Placebo
Mean (x, y): (6.46, 6.04)
SD (x, y): (0.14, 0.14)
n: 655
Placebo
Mean (x, y): (4.97, 4.53)
SD (x, y): (0.13, 0.12)
n: 683
Treatment
Mean (x, y): (5.31, 4.98)
SD (x, y): (0.13, 0.13)
n: 680
increased confidence
decreased confidence
Post−Treatment Confidence in National Election
Pre−Treatment Confidence in National Election
0 1 2 3 4 5 6 7 8 9 10
n 50 100 150 200
Treatment Placebo Treatment


Confidence in National Election: Pre−Treatment vs. Post−Treatment
By Article Writing Type
Human In The Loop
Only AI
ATT Cohen's d: 0.19
p−value: 0.007
ATT Cohen's d: 0.09
p−value: 0.008
Treatment
Mean (x, y): (6.64, 6.63)
SD (x, y): (0.17, 0.17)
n: 403
Placebo
Mean (x, y): (6.41, 6.02)
SD (x, y): (0.17, 0.17)
n: 431
Treatment
Mean (x, y): (6.70, 6.50)
SD (x, y): (0.08, 0.08)
n: 1762
Placebo
Mean (x, y): (6.67, 6.31)
SD (x, y): (0.08, 0.09)
n: 1697
increased confidence
decreased confidence
Post−Treatment Confidence in National Election
Pre−Treatment Confidence in National Election
Treatment Placebo Treatment
n 100 200

Confidence in Election Integrity and Election Rumors

By Party Identification Over Time, Weighted by Sampling Weights



Conclusion

- Presented experimental method for using AI to prebunk false election rumors and conspiracies
- Labor intensive, Human-in-the-Loop written articles perform similarly as those written purely by AI
- Prebunks are durably effective at reducing belief in specific rumors
- Prebunks are temporarily effective at increasing confidence in elections
- More work is needed to bolster long-term election confidence

Next Steps

- Are more "intensive" interventions more effective?
- Over the 2024 election: ran experiment comparing AI-powered chatbot conversations and AI-written inoculation articles
- Currently analyzing results
- Try the bot out at: https://electionbot.chat