The Zeus Protocol: Automating Survey Experimental Social Science Research Discovery with Large Language Models

Eduardo Tamaki and Levente Littvay (others)

2nd Budapest Methods Workshop on Large Language Models and Generative AI

HUN REN INSTITUTE FOR POLITICAL SCIENCE
CENTRE FOR SOCIAL SCIENCES


Public Opinion
80%
55%
57%
59%
50%

Political Analysis

Article contents

Abstract

Footnotes

References

Out of One, Many: Using Language Models to Simulate Human Samples

Published online by Cambridge University Press: **21 February 2023**

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate

Show author details

Article | Supplementary materials | Metrics

Get access | Share | Cite | Rights & Permissions

Abstract

We propose and explore the possibility that language models can be studied as effective proxies for specific human subpopulations in social science research. Practical and research applications of artificial intelligence tools have sometimes been limited by problematic biases (such as racism or sexism), which are often treated as uniform properties of the models. We show that the "algorithmic bias" within one such tool—the GPT-3 language model—is instead both fine-grained and demographically correlated, meaning that proper conditioning will cause

European Research Council

Established by the European Commission

HUN REN tkpti

Broad Application

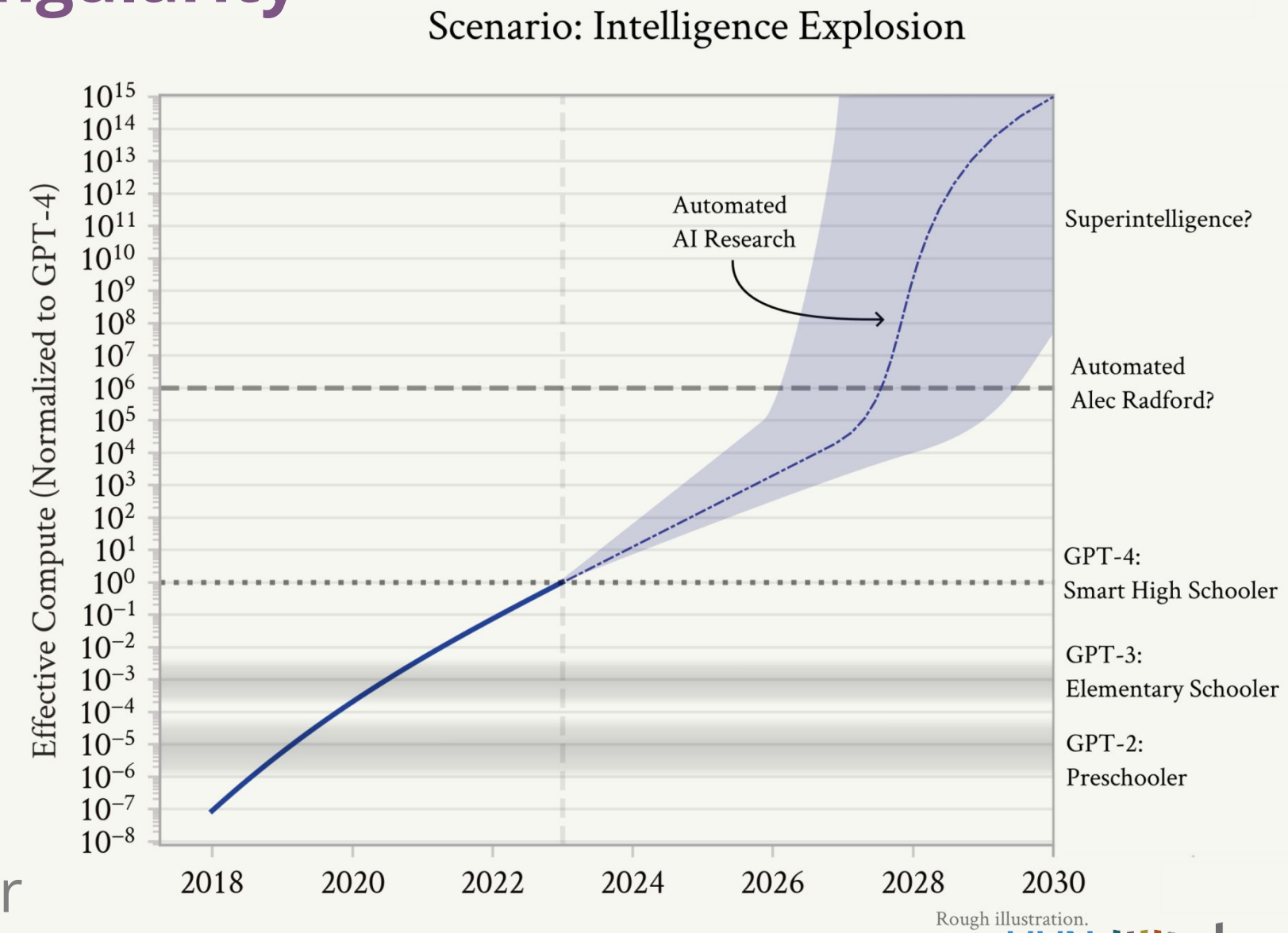
- God of Time - **Chrono-sampling:** Time Machine to Study Public Opinion Research in the Past (See tomorrow !!!)
- Goddess of Fame - **Pheme-sampling:** Silicon Elite Surveys
- God of Prophecy - **Apollo Causal Model:** Alternative Realities, Like Policy Regimes
- Titaness of Vision - **Theia-imputation:** Missing Data Imputation
- God of Foresight - **Prometheus Protocol:** Survey Pretesting

HUN REN tkpti

Technological Singularity

Leopold Aschenbrenner

Situational Awareness

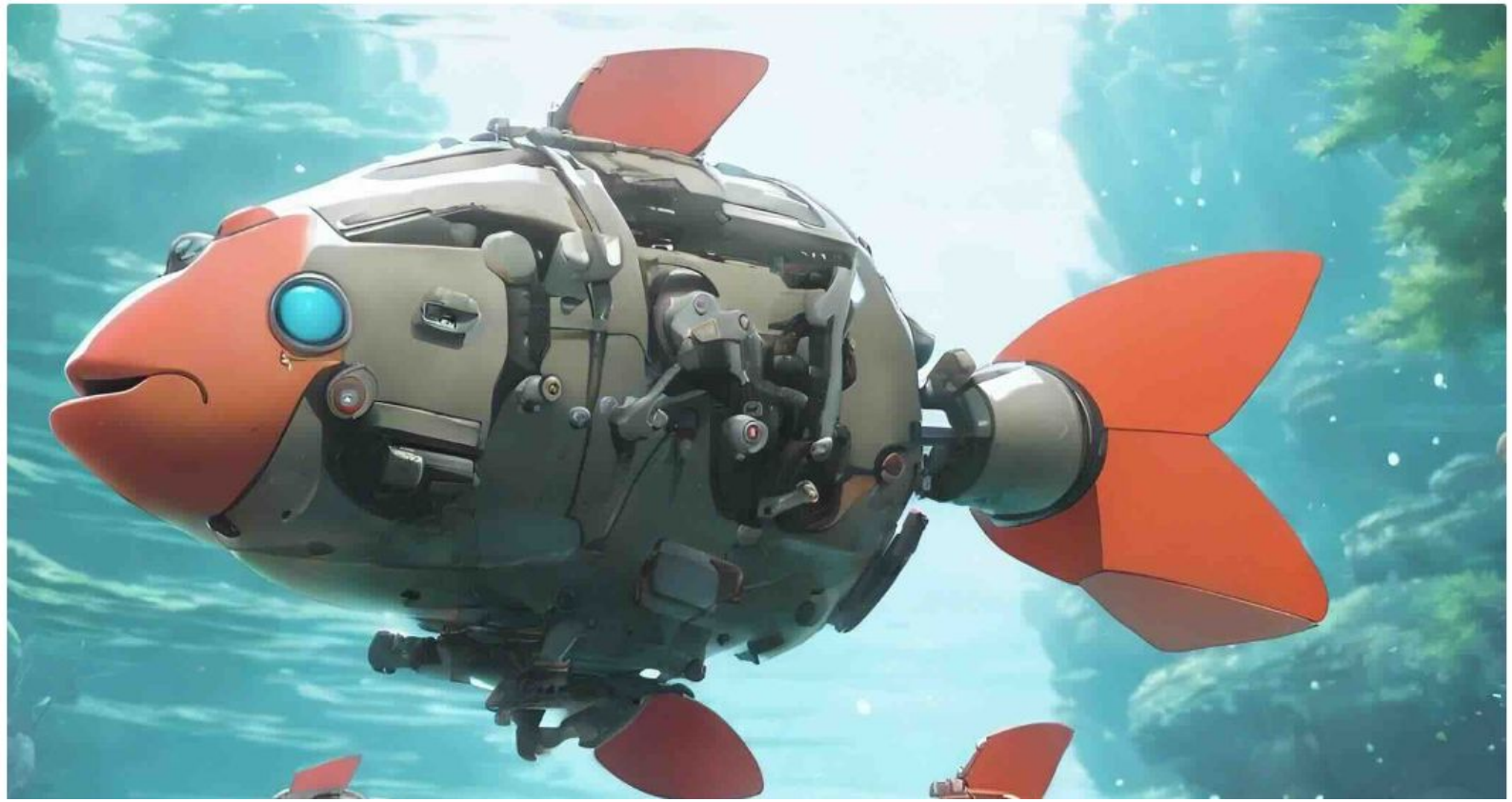


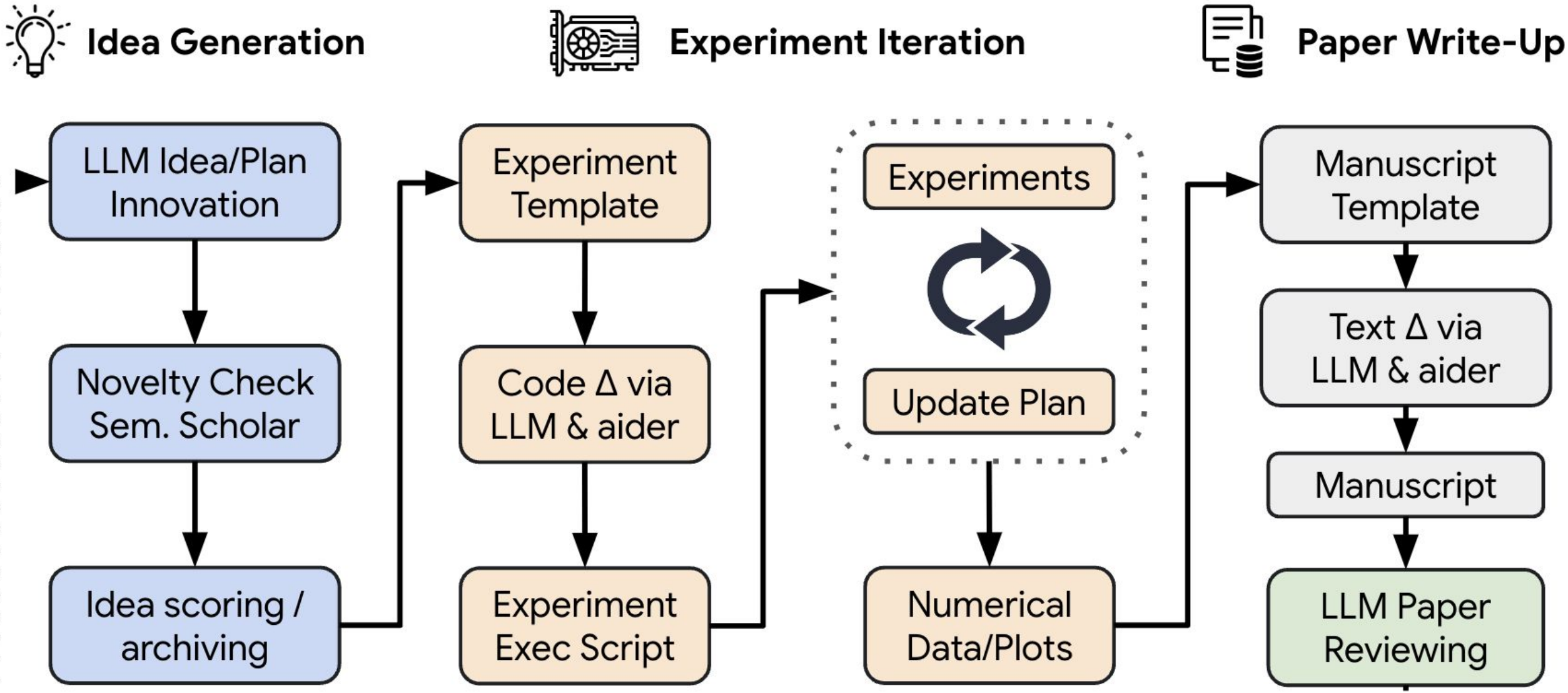
HUN REN tkpti

sakana.ai

The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

August 13, 2024




Idea Generation
Experiment Iteration
Paper Write-Up
LLM Idea/Plan Innovation
Novelty Check Sem. Scholar
Idea scoring / archiving
Experiment Template
Code Δ via LLM & aider
Experiment Exec Script
Experiments
Update Plan
Numerical Data/Plots
Manuscript Template
Text Δ via LLM & aider
Manuscript
LLM Paper Reviewing

Experimental Research on Depolarization

Experimental Research on Depolarization

Predicting Results of Social Science Experiments Using Large Language Models

Luke Hewitt[*1] Ashwini Ashokkumar[*2] Isaias Ghezae[1] Robb Willer[1]

[1]Stanford University [2]New York University

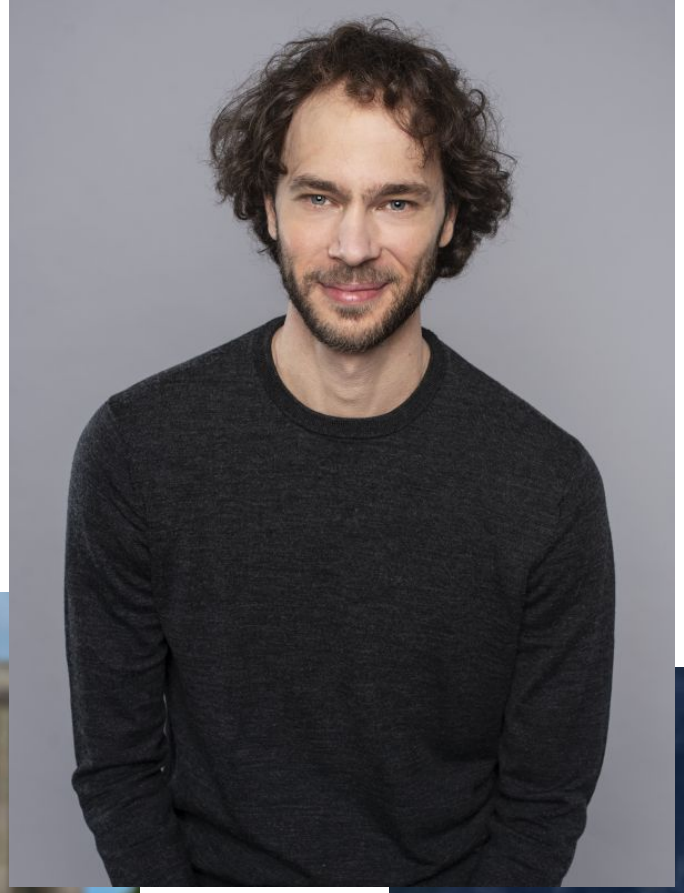
[*]Equal contribution, order randomized

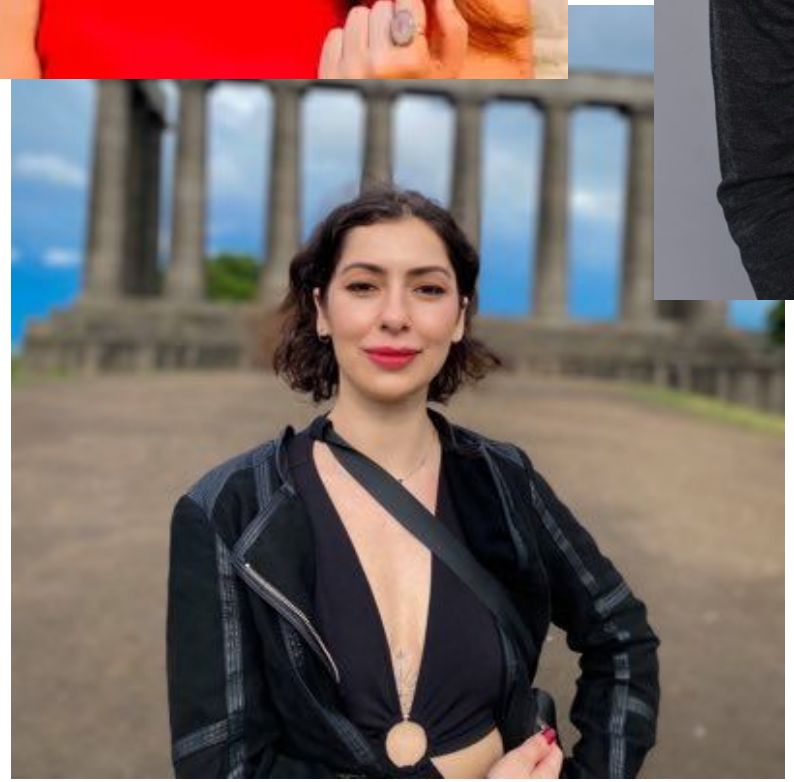
August 8, 2024

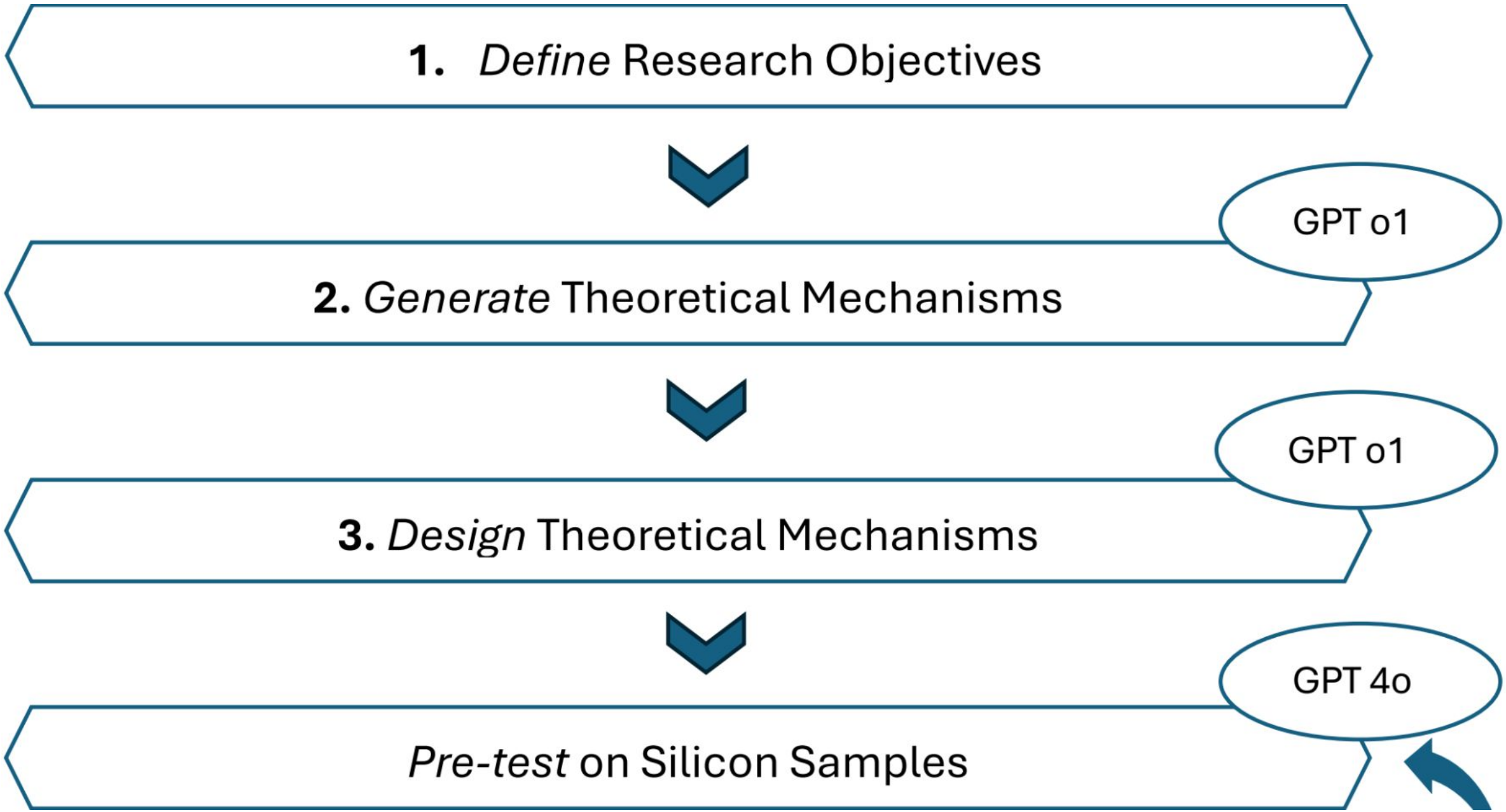
r = 0.85

Abstract

To evaluate whether large language models (LLMs) can be leveraged to predict the results of social science experiments, we built an archive of 70 pre-registered, nationally representative, survey experiments conducted in the United States, involving 476 experimental treatment effects and 105,165 participants. We prompted an advanced, publicly-available LLM (GPT-4) to simulate how representative samples of Americans would respond to the stimuli from these experiments. Predictions derived from simulated responses correlate strikingly with actual treatment effects ($r = 0.85$), equaling or surpassing the predictive






1. *Define* Research Objectives
2. *Generate* Theoretical Mechanisms
GPT o1
3. *Design* Theoretical Mechanisms
GPT o1
Pre-test on Silicon Samples
GPT 4o

General Protocol: Researcher Supplies

- Theme: In a few words (Political Polarization)
- Conceptual Definition: One paragraph
- Goal: Increase / Decrease
- Survey Question: Thinking about people whose political views are very different from your own, to what extent do you feel their views pose an existential threat to the United States?
- Response Categories: not at all, a little, quite a bit, a great deal
- Is it reverse coded? (No)

General Protocol: Idea Generation and Review

- Generates Theoretically Sound Mechanisms of Impact
  - Produce Ideas That Could Achieve Researcher's Goals
  - Accept top idea
  - Produce More (one by one) until total of 3 (parameterized) are accepted
- "Peer" Review (set up as two-round grant review)
  - Strengths and Weaknesses: Textual
  - Originality: Novel and Creative?
  - Quality: Will It Produce Externally Valid Results?
  - Promise: Potential Impact and Relevance?
  - Ethical Concerns: If any, reject
  - Contribution: Will It Advance the State of the Art

HUN REN tkpti

General Protocol: Experimental Design and Review

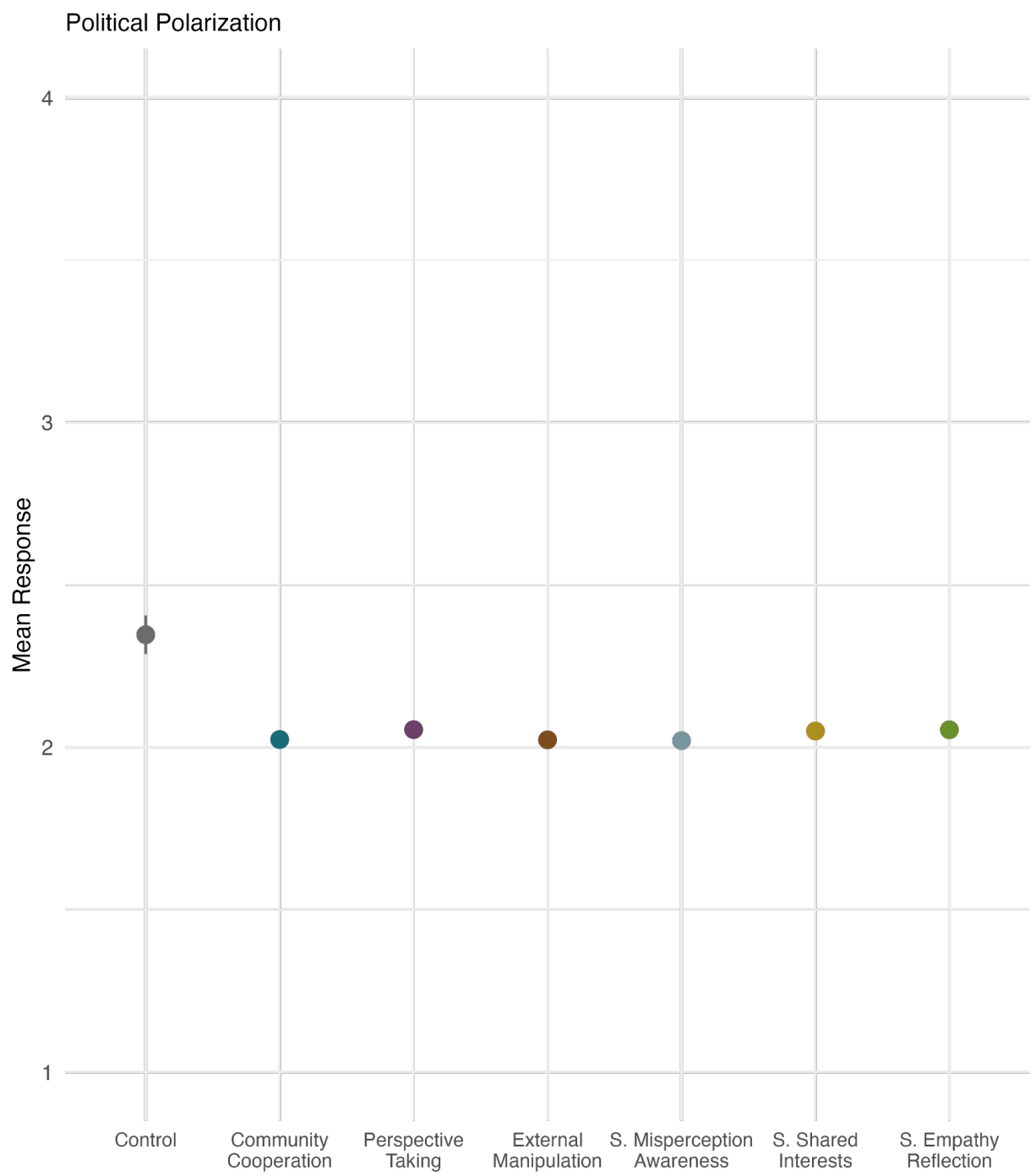
- Write Experimental Treatments

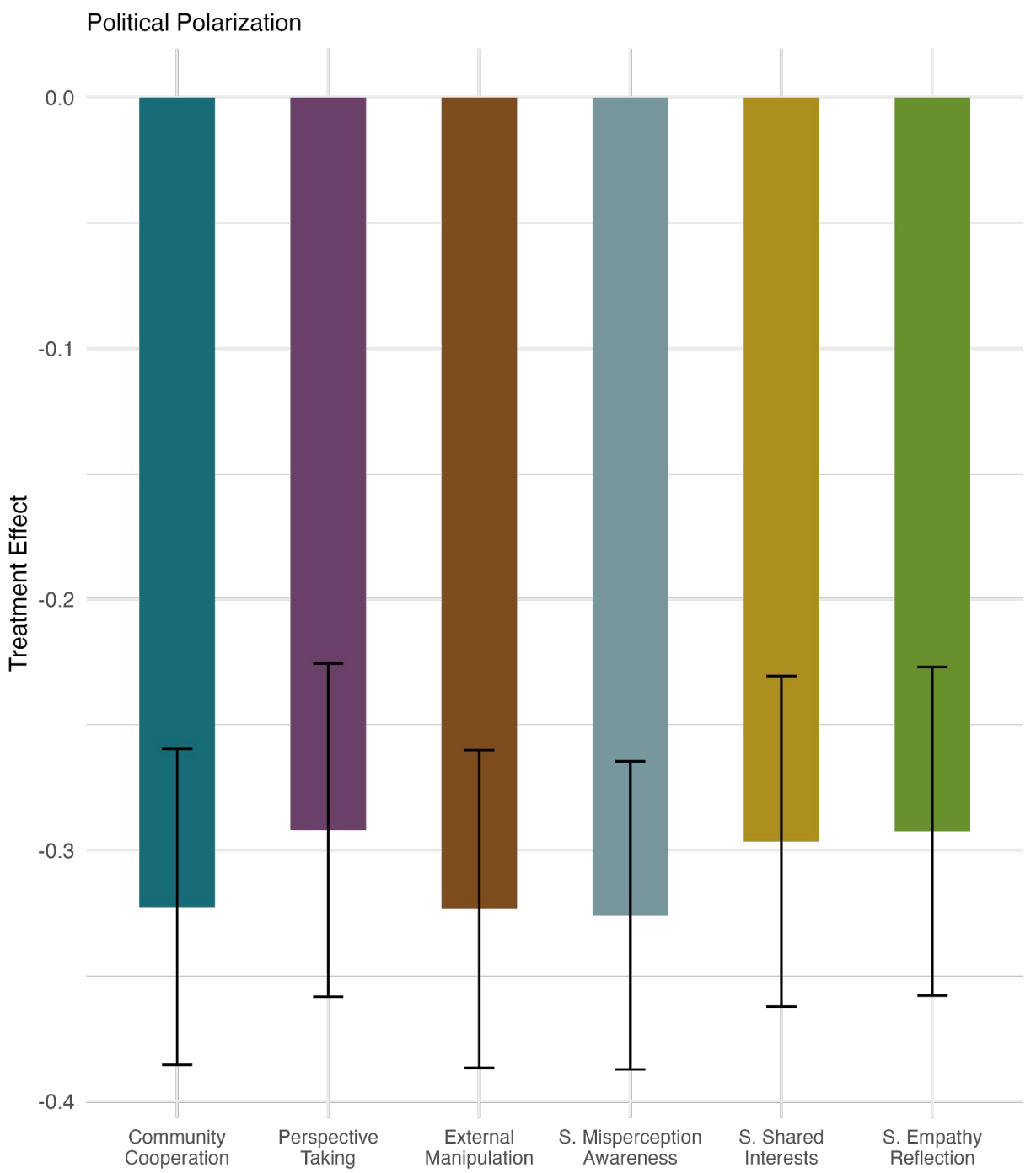
- “Peer” Review of Experimental Treatment (Second Round)
  - Alignment: Match between Proposed Mechanism and Treatment
  - External Validity: Generalizability of Findings Past Experiment
  - Internal Validity: Does the Treatment Elicit Only What Is Intended
  - Originality
  - Contribution

The Ideas

- High Scoring
  - Highlighting Positive Interdependence in Community Projects
  - Drawing upon the power of perspective-taking to reduce affective polarization
  - Leveraging Awareness of External Manipulation to Reduce Affective Polarization
- Poorly reviewed
  - Awareness of Misperception Bias of Out-Group
  - Highlighting Shared Interest in Non-Political Matters
  - General Empathy Trigger
- Generally Underwhelmed

HUN REN tkpti


Political Polarization
Mean Response
4
3
2
1
Control
Community Cooperation
Perspective Taking
External Manipulation
S. Misperception Awareness
S. Shared Interests
S. Empathy Reflection


Political Polarization
Treatment Effect
0.0
-0.1
-0.2
-0.3
-0.4
Community Cooperation
Perspective Taking
External Manipulation
S. Misperception Awareness
S. Shared Interests
S. Empathy Reflection

What's Next

- Implement a Placebo for Control Group (maybe)
- Implement Treatment Reinforcement (maybe)
- Prompt Engineer to Up Creativity (add "think outside the box")
- Rerun the Whole Thing
  - Move to Open Weight Models (Mistral Small 22b, Large 123b Q5, Llama 70b)
- Retool the Review Process
  - Quality / Originality Score
  - Likelihood of Working Score
- Validation
  - Expert Review for Quality / Originality
  - Human Sample Validation for Likelihood of Working
- Two More Examples (Case Studies / Use Cases)
  - 2. Justification of Extreme Actions (Radicalization, Decrease)
  - 3. Risk Taking (Increase / Decrease, not sure)

HUN REN tkpti

Thank you for your attention!

HUN REN

INSTITUTE FOR **POLITICAL** SCIENCE
CENTRE FOR **SOCIAL SCIENCES**

Highlighting Positive Interdependence in Community Projects: Idea

This idea is motivated by the potential for cooperative efforts to bridge divides. The plan is to present a textual stimulus describing a scenario where individuals from different political affiliations collaborate successfully on a community initiative, such as disaster relief or local improvement projects. Design choices involve selecting universally valued projects and demonstrating tangible benefits of cooperation. The ideal outcome is for respondents to see the practical advantages of working together, reducing affective polarization. This differs from existing ideas by focusing on local, concrete collaboration rather than abstract common goals or national identity.

HUN REN tkpti

Highlighting Positive Interdependence in Community Projects: Experimental Treatment

In the aftermath of severe flooding in Riverside County, neighbors of all political affiliations united to assist those affected. Together, they organized relief efforts, collected donations, and helped rebuild damaged homes. Conservatives and liberals worked side by side, focusing on the urgent needs of their community rather than their differences. Their joint efforts restored not only the physical structures but also the spirits of those impacted. This cooperative endeavor demonstrated how collaborative action can lead to meaningful results, highlighting the importance of putting aside political divides to address common challenges.

Drawing upon the power of perspective-taking to reduce affective polarization: Idea

The high-level plan involves presenting a textual stimulus that invites respondents to imagine how their political beliefs might differ if they had been born into a different family, community, or life circumstances. Necessary design choices include crafting a narrative that is neutral, relatable, and encourages self-reflection without inducing defensiveness. The ideal outcome is for respondents to recognize that personal experiences heavily influence political views, fostering empathy and understanding toward those with opposing beliefs. This increased awareness could lead to reduced negative emotions toward the outgroup, thereby decreasing affective polarization.

HUN REN tkpti

Drawing upon the power of perspective-taking to reduce affective polarization: Experimental Treatment

Please imagine that due to different life events or circumstances, you hold political views opposite to those you have now. Think about how factors like family influences, education, or career paths might alter your beliefs. Reflect on the idea that if your experiences had been different, your perspectives might also be different. Keeping this in mind can help in understanding others whose views contrast with your own. As you proceed to the next questions, consider how personal journeys shape beliefs.

Leveraging Awareness of External Manipulation to Reduce Affective Polarization: Idea

This idea is based on the intuition that acknowledging common vulnerabilities can unite individuals across political lines. The high-level plan involves presenting a textual stimulus that explains how external actors, such as foreign governments or misinformation campaigns, deliberately exploit social media and news outlets to amplify divisions within the country. Necessary design choices include using non-partisan language, avoiding blame on any political group, and providing credible examples of such interference that affect all sides equally. The ideal outcome is for respondents to recognize that they and their political opponents are both targets of manipulation, fostering a sense of shared victimhood and encouraging skepticism toward divisive rhetoric. This realization could lead to reduced negative emotions toward the outgroup and decrease affective polarization by highlighting the importance of national unity against external threats.

Leveraging Awareness of External Manipulation to Reduce Affective Polarization: Experimental Treatment

Intelligence agencies have reported that external adversaries are actively attempting to manipulate public discourse in the United States. They use social media bots and fake accounts to spread inflammatory content and exacerbate conflicts between different political groups. These efforts are designed to exploit our differences and distract us from shared values and common goals. Recognizing that these manipulations affect everyone, regardless of political beliefs, can help us remain vigilant against such tactics. By coming together and focusing on common ground, we can counteract these external influences and reduce unnecessary divisions in our society.

HUN REN tkpti

Parameterized Components

- 10 Starting Ideas
- 5 Idea Reviewers (Round One Reviews)
- 1 to Accept from Original Batch
- 3 to Accept Overall
- 3 Treatments Written
- 5 Treatment Reviewers (Round Two Reviews)
- n = 600 Silicon Sample Size

Extra Slides

- Extra Points
- More Extra Points
  - Extra Subpoints
  - More Extra Subpoints

Extra Figure Slide