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

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 accuracy of human forecasters. Accuracy remained high for unpublished studies that could not appear in the model's training data ($r = 0.90$). We further assessed predictive accuracy across demographic subgroups, various disciplines, and in nine recent megastudies featuring an additional 346 treatment effects. Together, our results suggest LLMs can augment experimental methods in science and practice, but also highlight important limitations and risks of misuse.