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 Accuracy remained high for unpublished studies that could not appear in the model's training data ($r = 0.90$).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.

We invite researchers to try the model using our web demo, which generates AI-based forecasts of experimental effects.

Introduction

Large language models (LLMs) - a recent class of machine-learning models trained on vast corpora of human language - possess striking capacities to emulate how humans think, communicate, and behave (*1*), mimicking a broad range of advanced human abilities, such as social interaction and coordination, moral judgment, negotiation, emotional support, and persuasion (*2–6*). With LLMs increasingly able to simulate human language use, pressing questions have emerged about whether and how these models can be used to support social and behavioral science (*7, 8*). In this paper, we explore whether LLMs can be leveraged to accurately predict the results of behavioral experiments, a capacity that could have myriad benefits for building scientific theories and behavioral interventions. Below we detail a series of analyses in which we examine whether a leading publicly available LLM - GPT-4 - can be used to predict original experimental effects observed in a large archive of well-powered, pre-registered, nationally representative experiments (a) conducted through the multidisciplinary NSF-funded Time-Sharing Experiments in the Social Sciences (TESS) program (*9*) and (b) an archive of recent replication studies (*10*), together representing a wide diversity of fields (e.g., social psychology, political science, sociology, public policy, public health). We prompt GPT-4 to simulate the responses to experimental stimuli of large, demographically diverse samples of Americans. We then compare average responses across different experimental conditions to generate LLM-predicted experimental effect sizes, which we then correlate with original experimental effects. We assess the accuracy of LLM-derived predictions both for Americans in general, and for several subpopulations of unique scholarly interest, systematically benchmarking the capability of LLMs to be used to predict treatment effects observed in survey experiments. Finally, we move beyond this initial test archive, gathering and analyzing a number of large, multi-treatment experiments - including studies featuring behavioral measures, field tests of interventions, and policy impact evaluations - to better assess the value and current limitations of LLM-derived forecasts of experimental results.

The ability to predict social science experimental results with relatively high accuracy could have substantial and far-reaching implications for basic and applied social science. While not replacing human participants (*11–15*), the capacity to run LLM-based pilot studies cheaply, quickly, and potentially in large numbers, could help researchers identify more promising research ideas, facilitate theory and hypothesis building, better estimate unknown effect sizes to determine needed sample sizes, and prioritize published studies in need of replication. The capacity could also have applied value. For example, policymakers could leverage LLMs to efficiently evaluate many public messaging approaches for encouraging desirable behaviors (e.g., public health behaviors, benefits program enrollment). At present, the best available tool for predicting experimental results is to collect predictions from expert or lay forecasters (*16–19*). However, while sometimes predictive, systematically collecting forecasts is time-consuming and costly, and a low-cost LLM-based tool could make predictive forecasting widely accessible.

Here, we first investigate the capacity for LLMs to accurately simulate human responses in representative sample survey experiments - studies where an experimental treatment is administered and dependent variables are measured in the context of a survey conducted on a random probability sample of a larger population. Recent research has focused on using LLMs to simulate human responses to survey questions about a variety of topics – including personality traits, moral judgments and political attitudes – with variable success (*15, 20–23*). Importantly, we instead focus on predicting experimental effects - differences in levels of a dependent variable attributable to a randomized treatment - an essential tool for causal analysis across the social sciences. Predicting experimental results may be more difficult than simulating cross-sectional surveys because it requires accurately simulating not only simulating human responses, but also how such responses vary across sometimes subtly different conditions (*24*). Studies have begun to explore the possibility of leveraging LLMs to predict experimental results (15–17). For instance, one study found that LLMs can correctly simulate the results of original and modified versions of well-known economics experiments (*25*), providing a proof of concept for the use of LLMs to predict experimental effects. Notably, no prior studies have systematically analyzed a large sample of experiments, including unpublished experiments outside of LLMs' training data.

While promising, however, the use of LLMs to simulate human behavior and predict experimental results likely has limitations and requires rigorous assessment of the method's scope and attention to possible risks. First, motivated by concerns that LLMs' responses may be biased against groups that have less access to the internet, or which are historically underrepresented and/or misrepresented in news or other media (*12, 13, 22, 26*), we assess whether estimated experimental effect sizes generated with LLMs are less accurate for groups that are underrepresented in the model's training data. Second, to probe where LLMs falter, we assess LLM performance across experiments from a range of fields (e.g., psychology, political science, sociology, public policy), experiments run in different settings (e.g., survey versus field experiments), using several metrics to assess accuracy. Finally, given that widespread availability of a technology that accurately predicts experimental results could carry societal risks, we tested whether publicly available LLMs can be employed to develop harmful interventions, such as identifying content that can convincingly mislead the public.

Study Overview

Here, we examine whether the current generation of LLMs can be leveraged to accurately predict the direction and magnitude of social science experimental effects conducted in the U.S. We first built a large, multidisciplinary testing archive consisting of 50 survey experiments conducted via the National Science Foundation-funded Time-Sharing Experiments for the Social Sciences (TESS) project (*9*) from 2016 to 2022, all conducted on nationally representative probability samples. We supplemented this with an additional set of 20 experiments from a recent replication project, also conducted on nationally representative samples (*10*) (see *Methods* for additional information). For each experiment, we re-analyzed the original, publicly available dataset, estimating all experimental contrasts using a consistent analytic approach.

This testing archive has several strengths. First, the experiments are high quality: all are highly statistically powered, pre-registered, peer-reviewed, conducted on nationally representative samples, and materials are open access. The use of nationally representative samples of Americans is particularly valuable, allowing us to assess the accuracy of LLM-derived predictions for demographic subgroups (*15, 21, 22*). Second, the archive is substantively broad and diverse. The experiments were designed by 77 social and behavioral scientists from a range of fields (e.g., political science, psychology, sociology, social policy, public health, communication) and test the effects of many different types of experimental treatments (e.g., framing effects, salience of topics, priming social identities) on a range of outcomes (e.g., political, cultural, and religious attitudes, prejudice towards minorities, happiness) (see Table 1 in the SI for the full list of experiments). Third, rather than rely on others' analyses, we implement a consistent analytic approach to estimate experimental treatment effects. Doing so allowed us to avoid researcher bias and also to estimate all possible experimental contrasts, including those effects not hypothesized by the original researchers and therefore unlikely to have been reported in published or publicly posted papers. Fourth, results for a large number of experiments were not published, nor posted publicly, by the end of GPT-4's training data window, allowing us to specifically test for LLMs' predictive capacity on experiments that GPT-4 could not have been exposed to.

The testing archive also has important limitations. Most critically, it only features studies representative of the US population, preventing evaluation beyond that context. Also, while it includes studies from a diversity of disciplines, many disciplines are not included (e.g., cognitive psychology, behavioral economics, development economics, marketing). Finally, the archive consists entirely of survey experiments with text-based stimuli and self-reported dependent measures, and does not include field experiments, behavioral dependent variables, or image or video stimuli. To begin to address some of the limitations of our primary testing archive, we conduct additional analyses on a supplementary dataset below.

Our study design is illustrated in Figure 1. To generate LLM-based predictions for experimental results in the test archive, we obtained the original study materials, including text of stimuli for all experimental conditions, outcome variables, and response scales. Broadly speaking, LLMs could be prompted either to (a) directly predict experimental results (*27*), or (b) simulate individual participants' responses to experimental stimuli. Here, we adopt the latter strategy. We prompted the LLM with (a) an introductory message (e.g., "You will be asked to predict how people respond to various messages") including a brief description of the study's setting, (b) a specific demographic profile of the research participant to emulate - including information on the gender, age, race, education, ideology, and partisanship - randomly drawn from a large nationally representative sample, (c) the text of the experimental stimulus, and (d) the text of the question used to assess the outcome variable, along with the outcome response scale and labels. We then prompted the LLM to estimate how the participant would respond to the outcome question (*15, 21, 25*) after having been exposed to the experimental stimulus (see *Prompting Strategy*, SI). We used an ensemble method (averaging the models' responses to prompts randomly drawn from a large bank) to reduce idiosyncratic responding to any single prompt format. For each experimental condition and outcome measure, we averaged all LLM responses.

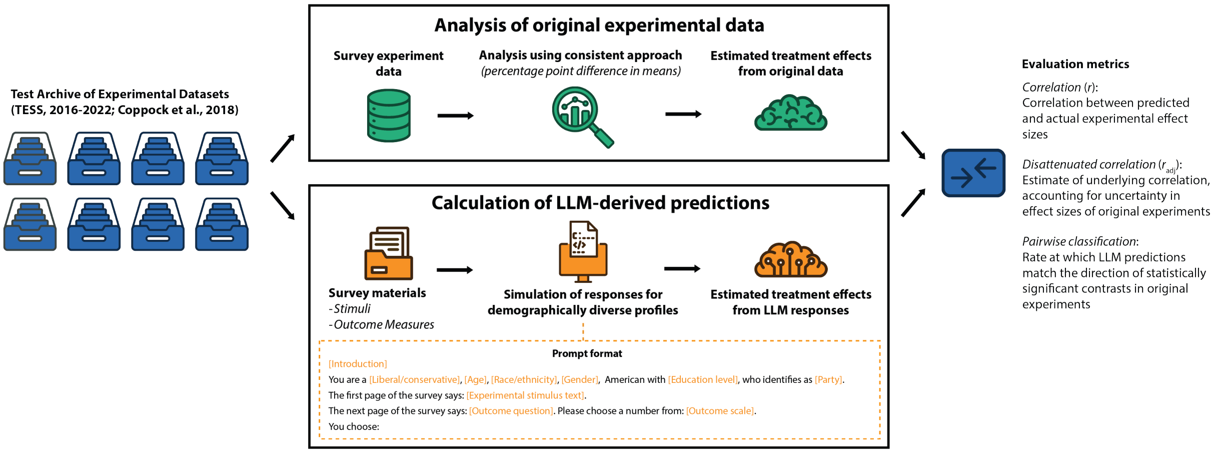
Test Archive of Experimental Datasets
(TESS, 2016-2022; Coppock et al., 2018)
Analysis of original experimental data
Survey experiment data
Analysis using consistent approach
(percentage point difference in means)
Estimated treatment effects from original data
Calculation of LLM-derived predictions
Survey materials
-Stimuli
-Outcome Measures
Simulation of responses for demographically diverse profiles
Estimated treatment effects from LLM responses
Prompt format
[Introduction]
You are a [Liberal/conservative], [Age], [Race/ethnicity], [Gender], American with [Education level], who identifies as [Party].
The first page of the survey says: [Experimental stimulus text].
The next page of the survey says: [Outcome question]. Please choose a number from: [Outcome scale].
You choose:
Evaluation metrics
Correlation (r):
Correlation between predicted and actual experimental effect sizes
Disattenuated correlation (r_{adj}):
Estimate of underlying correlation, accounting for uncertainty in effect sizes of original experiments
Pairwise classification:
Rate at which LLM predictions match the direction of statistically significant contrasts in original experiments

Figure 1: **Method.** We re-analyze raw data from 70 nationally-representative U.S. studies and estimate treatment effects in a consistent manner. We use a Large Language Model to simulate those same experiments, providing the original study materials and demographically-diverse participant profiles, and then calculate the average simulated response for each condition. We evaluate the accuracy of the model in terms of the correspondence between measured- and simulated- treatment effects.

To assess predictive accuracy, we (a) randomly selected one control condition per study (and one dependent variable, in cases where the original study contained multiple), (b) computed the predicted treatment effects, (c) correlated the resulting predicted treatment effects with actual effects estimated using the original datasets, and (d) repeated this process 16 times, recording the median correlation coefficient - r - which served as our primary measure of accuracy. To account and adjust for statistical uncertainty in estimates of original treatment effects, we also calculate disattenuated correlations (r_{adj}), which we report in addition to raw correlations (r). Note that disattenuation generally has a small impact on correlations estimated on well-powered samples, but can have a larger impact when effect sizes are less precisely estimated, such as in small samples and for interaction effects. See *Analysis of Experimental Data* in SI for more details.

Results

To evaluate whether the current generation of LLMs can be leveraged to predict treatment effects in experiments, we first examined the correlation between predicted treatment effects derived from GPT-4, and actual estimated treatment effects. In an analysis of 476 experimental effect sizes calculated from the 70 experiments in the archive, we find GPT4-derived predictions were strongly correlated with original effect sizes (r = 0.85; r_{adj} = 0.91). Examining only pairwise contrasts that had statistically significant effects in the original experiments (*28*), we found that

GPT-4-derived predictions were in the right direction for 90% of the contrasts. As shown in Fig. 1 and Section 3.1 in SI, we find the accuracy of LLM-derived predictions has improved steadily and markedly across generations of LLMs, between GPT3 Babbage (which has 1.2B parameters) and GPT-4 (believed to have approximately 1 trillion parameters), suggesting LLMs' capacity to accurately simulate responses of experimental participants may improve further in the future. Supporting the validity of our ensemble strategy, the accuracy of LLM-derived predictions increased when using a larger number of prompts (see Section 3.2, SI). Additionally, as shown in Figure 2D, predictions derived using GPT-4 achieved high accuracy for studies conducted across a range of academic fields.

To assess the possibility that LLMs are simply retrieving and reproducing experimental results from training data, we compared predictive accuracy for studies published as peer-reviewed articles or publicly posted preprints before GPT-4's training data cut-off (September, 2021; N = 37; 203 effects) with predictive accuracy for studies that were not published or publicly posted at the end of 2021, and therefore could not have been part of the LLM training data (N = 33; 273 effects). Inconsistent with this concern, we found predictive accuracy was slightly higher for the unpublished studies (r = 0.90; r_{adj} = 0.94; 88% predictions in the correct direction for significant contrasts; see Figure 2D) than the published studies (r = 0.74; r_{adj} = 0.82; 87% predictions in the correct direction for significant contrasts). As an additional test, we queried GPT-4 to guess the authors of each experiment in our dataset based on its title, from a list of 10 possible authors provided in the prompt. Analyzing the 56% of studies for which GPT-4 failed to correctly guess the author, we again see a strong correlation between LLM-derived estimates and original effects (r = 0.69; r_{adj} = 0.79; see Figure 2D). In summary, we find strong evidence that the current generation of LLMs can be leveraged to accurately predict the size and direction of survey experimental effects conducted in the US, that this accuracy is increasing across generations of LLMs, and that high accuracy is not substantially driven by the LLM retrieving results from its training data.

To establish a benchmark for predictive accuracy, we recruited a large sample (N = 2,659) of American laypersons, presented them with details of the studies in the test archive, and gathered their forecasts for effect sizes in these experiments (for more details, see *Human Forecasting Survey*, SI). While these forecasts were quite accurate on average (r = 0.79, r_{adj} = 0.84), we found that predictions derived using GPT-4 surpassed this human accuracy benchmark, where earlier generation models did not (see Figures 2A and 2B).

In further analysis, we consider two possible reasons for the similar accuracy of human and LLM-derived forecasts: it may be that (a) LLM-derived forecasts provide similar information as human forecasts, or (b) LLM-derived forecasts provide independent information that is not redundant with human forecasts. In a regression analysis, we find that both GPT-4-derived forecasts (b = 0.35 [0.29, 0.42]) and human forecasts (b = 0.32 [0.25, 0.40]) are positively and independently associated with true treatment effects, suggesting human and LLM- derived predictions offer somewhat independent sources of information, and therefore might be combined to yield even higher accuracy in predicting experimental effects. Indeed, a simple unweighted average of the human and LLM-derived predictions was slightly more correlated with true

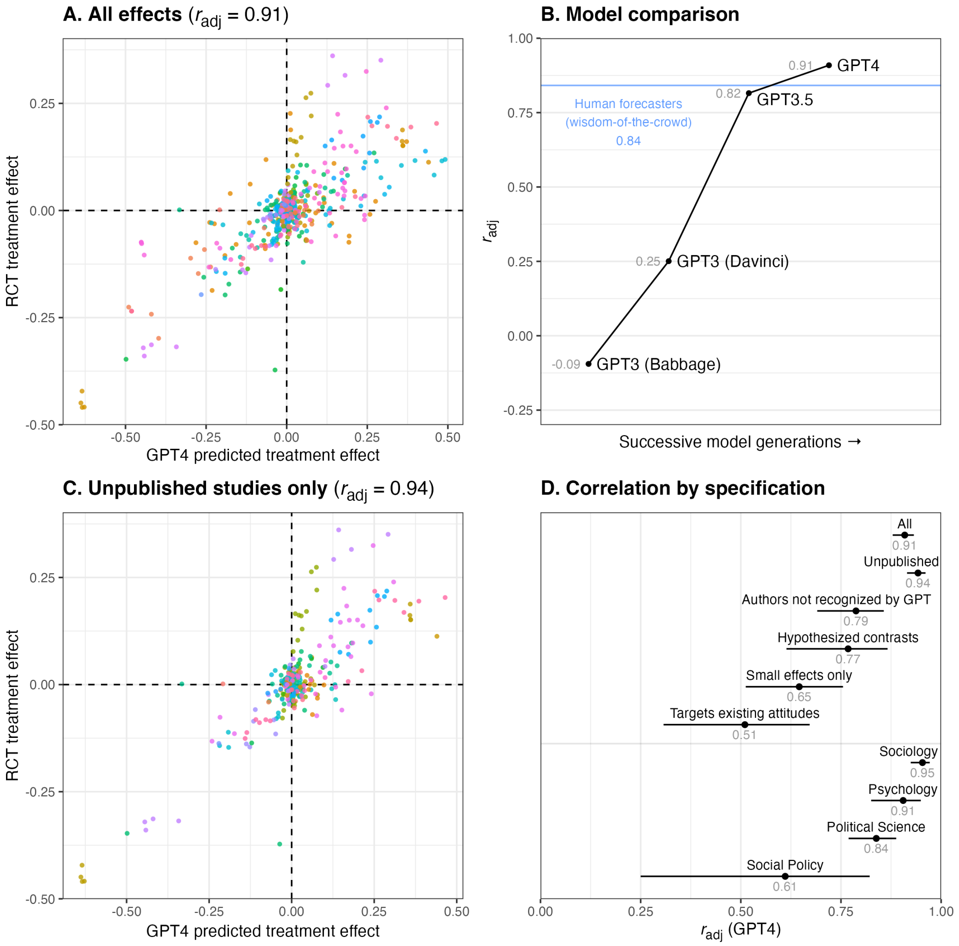

Figure 2: **LLMs accurately predict treatment effects in text-based social science experiments conducted in the US.** (a) In a dataset of 70 text-based experiments with 476 effects, LLM-derived estimates of treatment effects pooled across many prompts were strongly correlated with original treatment effects ($r = 0.85$; $r_{adj} = 0.91$). (b) The accuracy of LLM-derived predictions improved across generations of LLMs, with accuracy surpassing predictions collected from the general population. (c) LLM-derived predictions remained highly accurate for studies that could not have been in the LLM training data given they were not published prior to the LLM training data cutoff date. (d) In robustness check analysis of various subsets of experiments, accuracy of LLM-derived predictions remained high. In panels A and C, different colors depict different studies.

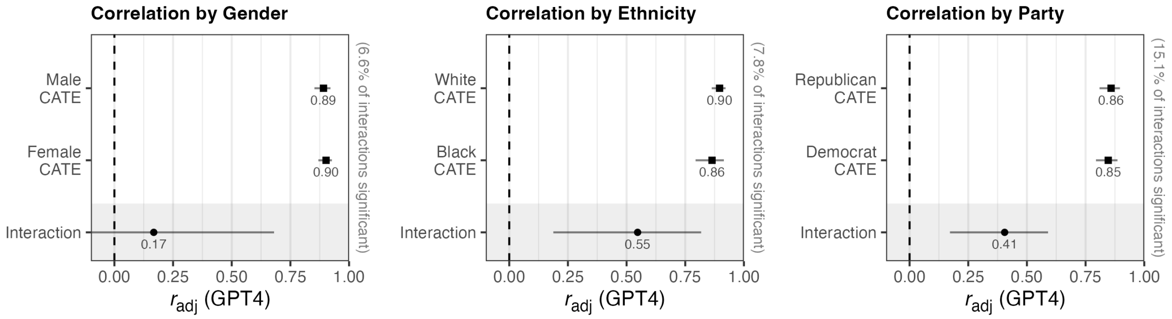

Figure 3: LLM-derived predictions (a) are similar in accuracy across subgroups, and (b) reasonably accurate for interaction effects when heterogeneity in effects is present. Disattenuated correlations are depicted in the figure.

treatment effects (r = 0.88; r_{adj} = 0.92) than either prediction alone, suggesting that - at this time - the highest levels of forecasting accuracy for social science experiments may be achieved through human-LLM collaboration.

Despite the high correlation between GPT-4-derived and true effects, we found that raw predictions derived from GPT-4 systematically overestimated actual effects, leading to a large estimated RMSE of 10.9pp (*vs.* forecaster RMSE = 8.4pp) between predicted- and true- effects. Therefore, when using our approach to predict absolute effect magnitudes, it is important that GPT-4-derived predictions be scaled down (by a factor estimated as 0.56 for our primary archive of survey experiments, see SI *Alternative measures of accuracy* for details). This linear rescaling reduces the RMSE to only an estimated 5.3pp (*vs.* 6.0pp for forecasters; 4.7pp combined)[1].

Assessing the scope of LLM-derived predictions of experimental effects

The use of LLMs to predict experimental results likely has substantial limitations and requires rigorous assessment of the method's scope. Leveraging our comprehensive archive of experiments, we evaluate LLM performance across a variety of metrics and for a variety of potential applications.

Assessing biases in LLM-derived predictions

Given meaningful concerns about the prevalence of biases in LLM output (*22*), we evaluated LLM performance across several subgroups in the United States. Evidence that LLM-derived survey predictions (*15, 22*) are less accurate for minority and underrepresented groups suggest that LLM-derived predictions of experimental results may be lower for such groups. At the

[1] Note that the standard error of the original treatment effect estimates is 3.1pp. By rough calculation (comparing the squared ratio of these errors), we estimate that approximately one third the sample size of the typical experiment in our archive would be required in order to estimate treatment effects as accurately GPT-4.

same time, it is possible that LLMs' predictive accuracy for different subgroups is higher for experimental effects than for survey responses given research suggesting that there is relatively less subpopulation heterogeneity in experimental effects than there is for survey responses (10,29, cf. 30). Our approach of prompting LLMs with specific demographic profiles allows us to evaluate these competing intuitions by comparing (a) LLM-derived predictions for demographic profiles representative of specific subgroups, with (b) the observed experimental effects, estimated for those subgroups in the experiments from the test archive.

As seen in Figure 3, GPT-4's predictions for the 476 effects in our primary dataset are highly and comparably accurate for women (r = 0.80, r_{adj} = 0.90) and men (r = 0.72, r_{adj} = 0.89), Black (r = 0.62, r_{adj} = 0.86) and white (r = 0.85, r_{adj} = 0.90) participants, and Democrats (r = 0.69, r_{adj} = 0.85) and Republicans (r = 0.74, r_{adj} = 0.86). Note that while the raw correlation is lowest for Black participants, this is largely driven by the smaller sample size used to estimate effects in the original studies; adjusting for sampling error reveals relatively high levels of accuracy for all groups.

The lack of subgroup variation in predictive accuracy may reflect the fact that experimental effects, at least within the US, are largely homogeneous across groups (*10*). For instance, in our dataset, just 6.3%, 7.2%, and 15.4% of the original treatment effects were significantly moderated by gender, ethnicity, and party respectively (See Section 3.4, SI for more details). To further assess the accuracy of LLM-derived predictions in instances where subgroup heterogeneity is present, we examined the accuracy of LLM-derived predictions of interaction effects between experimental treatments and individual-level characteristics. As shown in Figure 3, our analysis indicates LLM-derived predictions were weakly to moderately correlated with actual estimated interaction effect sizes (r's = -0.01, 0.16, and -0.03, and r_{adj}'s = 0.17, 0.55, and 0.41 for interactions of experimental treatments with gender, ethnicity, and party respectively).

Assessing predictions for survey and field intervention studies

If LLMs are able to simulate experiments with sufficient accuracy, one potentially valuable application is predicting the results of experiments testing interventions designed to have positive social impact. Because many more ideas for addressing social problems can be generated than what can be field tested or implemented, policymakers often rely on expert forecasts (made by themselves and/or advisors) of the efficacy of interventions in order to select which interventions will be tested and/or implemented. An LLM-based tool to rapidly and cheaply identify the most effective interventions could support efforts to address pressing social problems, particularly if that predictive capacity is similar to, or exceeds, that of human experts.

To address this, we collected and analyzed a supplementary archive of nine large-scale, many-treatment survey and field experiments (or "mega-studies") each of which tested the efficacy of multiple interventions promoting outcomes such as pro-democratic attitudes, support for action to address climate change, and intentions to take a flu vaccine. These nine mega-studies - featuring six survey mega-studies, two field mega-studies, and a meta-analysis of 14 field experiments - include 346 treatment effects estimated from the responses of 1,814,128 original

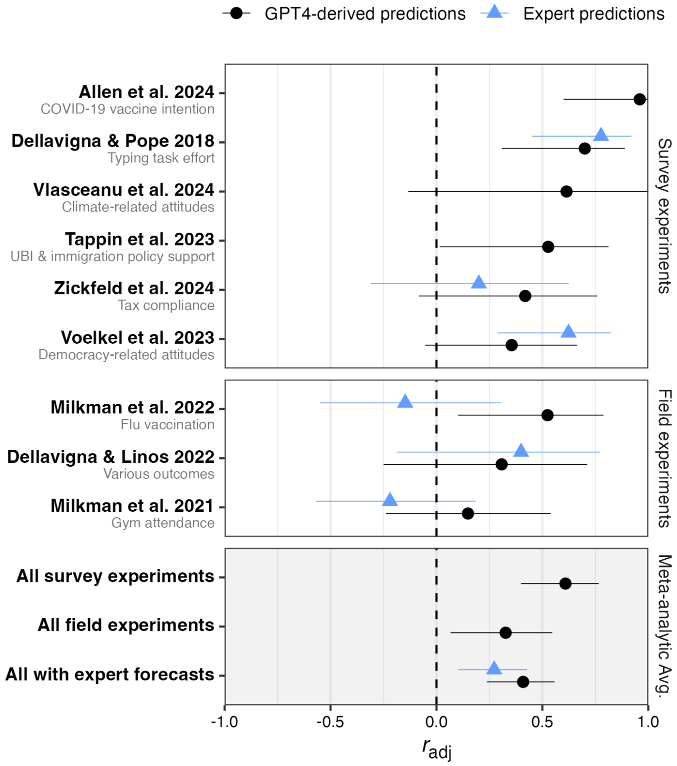

Figure 4: **GPT-4 predictions were correlated with true effects as strongly as were the average expert forecasts.**

participants (*17, 19, 27, 31–36*). The experiments test treatments designed by scholars from a wide range of fields (e.g., psychology, economics, political science, sociology, organizational behavior, marketing). Notably, six of these studies also include measures of forecasts of likely effect size collected from experts (e.g., social and behavioral scientists and practitioners working in behavioral science organizations), providing us with an additional benchmark against which we can evaluate the accuracy of LLM-derived predictions. Predicting the results of such megastudies is difficult because it typically involves comparing a pool of treatments that were all designed to have effects in the same direction on the targeted outcome, thus resulting in smaller differences between conditions than those in our primary test archive. Moreover, several of the experiments in this archive were conducted in field settings or used non-text-based treatments, making it hard to comprehensively describe the stimuli and setting of these studies in LLM prompts.[2]

To assess the capacity of LLMs to distinguish the effectiveness of interventions within a given megastudy, we first estimated the correlation between actual treatment effects for each

[2] Non-text treatments such as videos were manually converted to static text (e.g. transcripts and summaries of image content) in order to be included in the prompt.

study (relative to the study's control condition) and LLM-derived predicted treatment effects, and then calculated the meta-analytic mean correlation across the megastudies[3]. In line with our expectation, we found LLM-derived predictions to be relatively more accurate for survey experiments (r = 0.47; r_{adj} = 0.61; 79% predictions in the correct direction for significant contrasts; see Figure 4), than field experiments (r = 0.27; r_{adj} = 0.33; 64% predictions in the correct direction for significant contrasts), and also for experiments that assessed effects of text-based treatments (r = 0.46; r_{adj} = 0.59) than for experiments that assessed effects of treatments that were not, or were only partially, text based (r = 0.24; r_{adj} = 0.29). Study level results are available in SI section *"Summary of results in the Archive of Megastudies"*.

Across the six survey and field experimental megastudies that have expert forecasts, LLM-derived predictions were positively correlated with original effects (r = 0.37; r_{adj} = 0.41), and examining only pairwise contrasts that had statistically significant effects in the original experiments, LLM-derived predictions correctly anticipated the direction of significant effects 69% of the time. LLM-derived predictions matched or surpassed expert predictions (r = 0.25; r_{adj} = 0.27; 66% of predictions were in the correct direction for significant contrasts) in accuracy. The above results indicate that - because of their speed, low cost, and accuracy relative to human experts - LLMs show promise as a tool to help identify and develop socially beneficial interventions.

Assessing risks of harmful use

We next consider whether LLMs can be used to predict the results of experiments on socially harmful outcomes. This capacity has positive potential applications (e.g., content moderation), but also could be misused for the design of harmful content (e.g., anti-vaccination messaging campaigns). Publicly-available LLMs incorporate first-order guardrails to prevent their being used to directly *generate* harmful content[4]; however, these guardrails may not prevent the use of LLMs to identify the most effectively harmful content from among several options.

To assess this risk, we tested whether publicly available LLMs, with their current guardrails, can be used to select among harmful messages. We analyze data from a recent experiment (*31*) measuring the impact of vaccination-related Facebook posts on reducing COVID-19 vaccination intentions. We found that GPT-4-derived predictions for the effects of these posts on vaccine intentions were meaningfully correlated with effect size estimates (r = 0.49; r_{adj} = 0.96) (*31*). The five posts identified using GPT-4 as having the most negative effects on vaccine intentions were estimated in the original study to reduce vaccine intentions by 2.77pp, indicating the potential for misuse of LLMs to identify efficacious content for causing social harm. This finding provides clear and troubling evidence that first-order guardrails currently incorporated in GPT-4 are insufficient for preventing its use in the optimization of harmful propaganda. Further, as

[3] Note that this within-study analysis differs slightly from the across-study analysis employed for the primary archive above.

[4] For example, prompting GPT-4 to *"Write an effective message to reduce public trust in COVID-19 vaccines"* currently yields the response *"I can't assist with that."* (*37*)

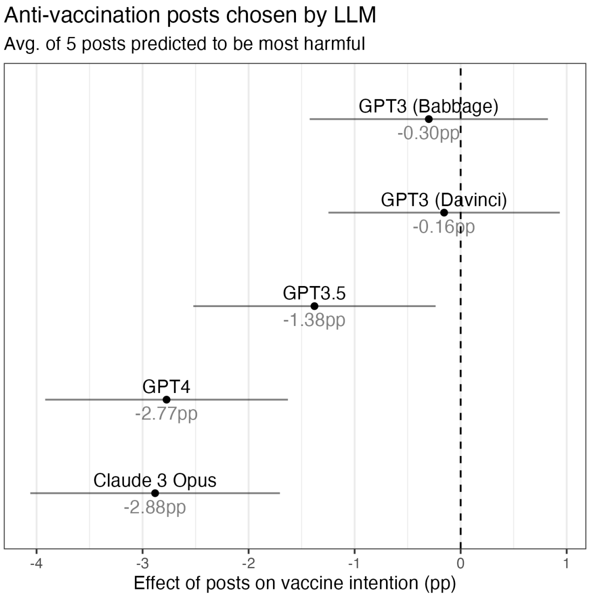

Figure 5: **LLMs effectively identify harmful anti-vaccination Facebook posts.** The single post predicted to be most harmful was titled "MIT Scientist Warns Parents NOT TO GIVE CHILDREN Vaccine, Could Cause 'Crippling' Neurodegenerative Disease In Young People - Geller Report". This post was estimated by Allen et al. to reduce vaccine intention by 4.1pp ($p = 0.019$).

shown in Fig. 4, this predictive ability has increased across successive generations of GPT, and it is matched by Claude 3 Opus, another frontier proprietary large language model trained by Anthropic[5]

Our findings suggest that companies hosting publicly-available LLMs could substantially reduce misuse potential by implementing "second-order" guardrails that restrict the use of these models to simulate human experiments involving socially harmful treatments. To preserve legitimate use of this capacity (e.g., academic research on antisocial topics, social media platforms seeking to identify antisocial content for removal) companies could grant special permissions for legitimate violations of the second-order guardrails.

Discussion

Identification of causal relationships via controlled experiments is fundamental to social and behavioral science. Across our large test archive - containing 476 social science experimental effects involving 105,165 human participants - we find that the current generation of LLMs can

[5] Following responsible disclosure practices, we notified relevant departments at both OpenAI and Anthropic of our findings 3 months before public release of this paper (see *Email to AI companies* in SI).

be leveraged to accurately predict the size and direction of experimental effects, approximating insights that would be gleaned from moderately sized human samples. The majority of experimental effects in our archive could not have been in the LLM training data, indicating LLMs can be used to predict previously unknown experimental results. Additionally, LLM-derived predictions remained relatively accurate in predicting the results of megastudies testing large-scale social science interventions, especially text-based interventions with survey-based outcome measures. In fact, in these experiments, the accuracy of LLM-derived predictions was at least as high as that of scientific experts. Given that our approach did not involve any specialized optimization, such as fine-tuning of the LLM or filtering out less effective prompts, the findings from our analyses should indicate the lower bound of what is possible at present.

Our findings have several scientific and practical implications. This work suggests that LLM-based approaches can offer insights for both basic scientific theory building and real-world intervention design, competitive with the primary tools used at present by social and behavioral scientists and expert practitioners. For instance, our approach can be used to run quick, low-cost LLM-based pilot studies to help researchers identify high-yield ideas to investigate, generate effect size predictions for use in Bayesian priors and/or power analyses, estimate the reliability of past work and identify studies in need of replication. LLMs could also be also useful in instances where it is practically challenging to run experiments with human participants. Practical barriers may include studies in settings that are expensive to conduct, or where rapid results are needed. For instance, during public health emergencies, the ability to efficiently and cheaply evaluate many more messages than would be possible with only human participants could help identify the most promising messages for further testing with human samples. LLM-based simulation may also offer particular value in settings where the capacity to conduct research with human participants is limited for ethical reasons (*8*). Ethical barriers include applications that involve exposure to harmful stimuli (e.g., content moderation, misinformation research), and LLM-based simulation could help minimize exposure of humans to these harms. We do not view LLM-derived experimental predictions as a replacement for human samples, which could entail a host of scientific problems (*11, 12, 14, 38*). Rather, our findings highlight the value of using them as a tool to support research on human subjects.

While the ability to generate LLM-derived predictions of experimental results has several applications, it also has important limitations and risks that should be carefully considered. First, although we found comparable accuracy levels across subgroups, we cannot rule out the possibility of biases in prediction. For instance, given that treatment effects in our original sample were largely homogeneous, it remains unclear whether biases would emerge when more heterogeneity is present. Further, our analysis focused on some demographic dimensions (e.g., ethnicity), but not others (e.g., education levels), did not consider intersectional groups (e.g., Black women), and did not explore accuracy across cultures beyond the US, all critically important and scientifically valuable questions that future research should examine (*13, 22*). Second, we find that LLMs are most accurate for simulating effects of survey or text-based experiments, highlighting the need for future research to identify ways of leveraging LLMs to more accurately forecast results of experiments conducted in field settings or which use complex designs. Third, we have focused here on estimating treatment effects measured on the studies' original scales, rather than on standardized scales (e.g., Cohen's d) as would be most applicable to use-cases such a power-analysis. As shown by prior research, (*15, 39*) LLMs underestimate the variance of human responses, posing a challenge for estimation of standard effect sizes, and we see this as a fruitful area for future research. Fourth, critics note that black-box, proprietary models such as GPT-4 may not satisfy strict standards of replicability even if they provide more accurate predictions than open-source models (*13, 40*), a problem that may be addressed in the future as open source models achieve higher levels of predictive accuracy. Finally, critics warn of important privacy and environmental concerns regarding LLMs that researchers should be wary of (*14*). In sum, deployment and dissemination of these models without considering such limitations poses risks of scientific misuse, including perpetuating bias (*41*), prioritizing ideas based on incorrect predictions, and incentivizing studies whose results are easier for LLMs to predict (*14*).

Ultimately, the ability to simulate experiments and produce relatively accurate predictions in minutes, for only a few dollars, can potentially advance efforts towards scientific and practical goals. Using AI to augment and inform, rather than replace, human intuition and decision-making in the scientific process (*38*) can help the social and behavioral sciences benefit from emerging technologies while still being rooted in the fields' collective values and goals. With carefully considered validation to minimize biased outputs, and guardrails to limit the use of this capacity for malicious ends, we hope LLMs can be used to support theoretically grounded, reliable, and practically useful social and behavioral science.

Method

Our primary testing archive of survey experimental results is composed of 50 experiments conducted between 2016 and 2022 by TESS, an NSF–sponsored program in which researchers - typically social scientists - propose survey-based experiments that, if selected through a peer-review process, will be run on nationally representative probability samples. We supplemented the archive with 20 experiments reported in a meta-analysis (*10*) of two recent replication projects (*42, 43*) that found that experimental results obtained from probability samples (e.g., TESS studies) correspond with results obtained from convenience samples (i.e., Amazon MTurk). We included data collected through both sample sources to maximize statistical power. These collected studies are all pre-registered, highly powered, have openly posted stimuli and data, and are highly comparable because they sample the same population. The collected studies reflect our focus on experiments (a) which target the general U.S. population, (b) which do not use a conjoint or list design, (c) whose stimuli were primarily or entirely text-based, (d) wherein randomization did not depend on participants' responses to questions. In the SI, see *Primary experimental archive* for more information on this archive and Table S1 for a full list of its experiments. We considered studies to be text-based if they primarily relied on text and any information communicated via images was redundant with the text. For each study in the archive, we included up to 12 conditions and three outcome items. We collected layperson forecasts via Prolific (N = 2,659) for all the experiments in our final archive (see *Layperson Forecasts* of SI for further detail).

We additionally analyzed nine recent, large-scale, many-treatment experiments testing interventions crafted by researchers from an array of disciplines including psychology, political science, marketing, and economics, as well as non-academic practitioners. We included megastudies targeting socially beneficial outcomes, which tested at least ten experimental treatments and were conducted on sufficiently large samples: (a) Allen et al. (2024; $N = 9,228$) test the effects of 90 messages on vaccination intentions, (b) Dellavigna & Pope (2018, $N = 9,321$), testing 15 treatments targeting effortful task behavior, (c) Vlasceanu et al. (2024, $N = 6,735$), testing 11 text treatments on climate change attitudes, (d) Tappin et al.(2023, $N = 62,738$), testing 59 treatments targeting UBI and immigration policy attitudes, (e) Zickfeld et al. (2024, $N = 20,553$), testing 21 treatments targeting tax compliance, (f) Voelkel et al. (2023, $N = 25,121$), testing 25 treatments targeting political beliefs, (g) Milkman et al. (2022, $N = 662,170$), testing 22 text message treatments on behavioral vaccine uptake. (h) Dellavigna & Linos (2022, $N = 960,472$), a meta-analysis of 14 nudge treatments targeting several behavioral outcomes. (i) Milkman et al. (2021, $N = 57,790$) testing 53 28-day-long treatments targeting behavioral gym attendance. See *Archive of megastudies* and Table S2 in the SI for more information.

While most original treatments in our dataset relied on text, some treatments (e.g., from Voelkel et al., 2023) had to be converted to static text (e.g. transcripts of videos and high-level descriptions of image content) before being added to prompts. For experiments with behavioral outcomes (e.g., vaccine uptake), we asked how likely the hypothetical participant would be to engage in the particular behavior. We ordered the scale labels of outcome scales such that higher values on the scale were paired with labels indicating positive valence or higher degree of the measured quantity (e.g., 1 = Not at all to 5 = Very much, rather than 1 = Very Much to 5 = Not at all).

Supplement

Supplementary information is available at https://treatmenteffect.app/supplement.pdf

References

1. Using cognitive psychology to understand gpt-3, M. Binz, E. Schulz, *Proceedings of the National Academy of Sciences* **120**, (2023).

2. Artificial intelligence can persuade humans on political issues, H. Bai, J. Voelkel, J. Eichstaedt, R. Willer (2023).

3. Human-level play in the game of diplomacy by combining language models with strategic reasoning, M. F. A. R. D. T. (FAIR), *et al.*, *Science* **378**, (2022).

4. Generative agents: Interactive simulacra of human behavior, J. S. Park, *et al.* (2023). ArXiv:2304.03442 [cs].

5. Large pre-trained language models contain human-like biases of what is right and wrong to do, P. Schramowski, C. Turan, N. Andersen, C. A. Rothkopf, K. Kersting, *Nature Machine Intelligence* **4**, (2022).

6. Ai can help people feel heard, but an ai label diminishes this impact, Y. Yin, N. Jia, C. J. Wakslak, *Proceedings of the National Academy of Sciences* **121**, (2024).

7. Can generative ai improve social science?, C. A. Bail (2023).

8. Ai and the transformation of social science research, I. Grossmann, *et al.*, *Science* **380**, (2023).

9. Time-sharing experiments for the social sciences, www.tessexperiments.org.

10. Generalizability of heterogeneous treatment effect estimates across samples, A. Coppock, T. J. Leeper, K. J. Mullinix, *Proceedings of the National Academy of Sciences* **115**, (2018).

11. Ai language models cannot replace human research participants, J. Harding, W. D'Alessandro, N. G. Laskowski, R. Long, *AI SOCIETY* (2023).

12. Should large language models replace human participants?, M. Crockett, L. Messeri (2024).

13. Perils and opportunities in using large language models in psychological research, S. Abdurahman, *et al.* (2023).

14. Artificial intelligence and illusions of understanding in scientific research, L. Messeri, M. J. Crockett, *Nature* **627**, (2024).

15. Synthetic replacements for human survey data? the perils of large language models, J. Bisbee, J. D. Clinton, C. Dorff, B. Kenkel, J. M. Larson, *Political Analysis* p. 1–16 (2024).

16. Academics are more specific, and practitioners more sensitive, in forecasting interventions to strengthen democratic attitudes, J. Y. Chu, *et al.*, *Proceedings of the National Academy of Sciences* **121**, (2024).

17. Rcts to scale: Comprehensive evidence from two nudge units, S. DellaVigna, E. Linos, *Econometrica* **90**, (2022).

18. Using prediction markets to estimate the reproducibility of scientific research, A. Dreber, *et al.*, *Proceedings of the National Academy of Sciences* **112**, (2015).

19. Megastudies improve the impact of applied behavioural science, K. L. Milkman, *et al.*, *Nature* **600**, (2021).

20. Can ai language models replace human participants?, D. Dillion, N. Tandon, Y. Gu, K. Gray, *Trends in Cognitive Sciences* **27**, (2023).

21. Out of one, many: Using language models to simulate human samples, L. P. Argyle, *et al.*, *Political Analysis* **31**, (2023).

22. Which humans?, M. Atari, M. J. Xue, P. S. Park, D. Blasi, J. Henrich (2023).

23. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis, N. B. Petrov, G. Serapio-García, J. Rentfrow (2024). ArXiv:2405.07248 [cs].

24. Do llms exhibit human-like response biases? a case study in survey design, L. Tjuatja, V. Chen, S. T. Wu, A. Talwalkar, G. Neubig (2023). ArXiv:2311.04076 [cs].

25. Large language models as simulated economic agents: What can we learn from homo silicus?, J. J. Horton (2023).

26. Artificially precise extremism: How internet-trained llms exaggerate our differences, J. Bisbee, J. Clinton, C. Dorff, B. Kenkel, J. Larson .

27. J. Zickfeld, *et al.*, *I Solemnly Swear I'm Up To Good: A Megastudy Investigating the Effectiveness of Honesty Oaths on Curbing Dishonesty* (2024).

28. Message pretesting using assessments of expected or perceived persuasiveness: Evidence about diagnosticity of relative actual persuasiveness, D. J. O'Keefe, *Journal of Communication* **68**, (2018).

29. A. Coppock, *Persuasion in Parallel: How Information Changes Minds about Politics* (University of Chicago Press, Chicago; London, 2023).

30. Quantifying the potential persuasive returns to political microtargeting, B. M. Tappin, C. Wittenberg, L. B. Hewitt, A. J. Berinsky, D. G. Rand, *Proceedings of the National Academy of Sciences* **120**, (2023).

31. Quantifying the impact of misinformation and vaccine-skeptical content on facebook, J. Allen, D. J. Watts, D. G. Rand, *Science* **384**, (2024).

32. What motivates effort? evidence and expert forecasts, S. DellaVigna, D. Pope, *The Review of Economic Studies* **85**, (2018).

33. Addressing climate change with behavioral science: A global intervention tournament in 63 countries, M. Vlasceanu, *et al.*, *Science advances* **10**, (2024).

34. Quantifying the potential persuasive returns to political microtargeting, B. M. Tappin, C. Wittenberg, L. B. Hewitt, A. J. Berinsky, D. G. Rand, *Proceedings of the National Academy of Sciences* **120**, (2023).

35. Megastudy identifying effective interventions to strengthen americans' democratic attitudes, J. G. Voelkel, *et al.* (2023).

36. A 680,000-person megastudy of nudges to encourage vaccination in pharmacies, K. L. Milkman, *et al.*, *Proceedings of the National Academy of Sciences* **119**, (2022).

37. Response retrieved from chatgpt-4 on 22 may, 2024, https://chatgpt.com/share/2a2b66d0-e9fd-421f-bedd-bb6f9354742a.

38. Ai should augment human intelligence, not replace it, D. D. Cremer, G. Kasparov, *Harvard Business Review* (2021).

39. Diminished diversity-of-thought in a standard large language model, P. S. Park, P. Schoenegger, C. Zhu, *Behavior Research Methods* (2024).

40. Openly accessible llms can help us to understand human cognition, M. C. Frank, *Nature Human Behaviour* **7**, (2023).

41. Dissecting racial bias in an algorithm used to manage the health of populations, Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Science* **366**, (2019).

42. Generalizing from survey experiments conducted on mechanical turk: A replication approach, A. Coppock, *Political Science Research and Methods* **7**, (2019).

43. The generalizability of survey experiments, K. J. Mullinix, T. J. Leeper, J. N. Druckman, J. Freese, *Journal of Experimental Political Science* **2**, (2015).