Supplementary Information for "Predicting Results of Social Science Experiments Using Large Language Models"

Contents

1 Materials and Methods

1.1 Primary experimental archive details

Coding process. The primary archive included 70 experiments: 50 experiments from TESS, and 20 experiments from a meta-analysis by Coppock and colleagues [1]. The 50 TESS experiments were directly downloaded from the TESS website. For each experiment, the TESS website provides the original survey materials, a codebook, and original data in a consistent format across studies. We downloaded all data files from the website and directly read them into our R code for analysis without requiring any manual coding, ensuring zero chance of error. We manually coded study materials (i.e., condition text, outcome measure wording, outcome scale and labels) from the website. We also manually identified variable names corresponding to the independent and dependent/outcome variables, which we coded in a consistent format. The availability of study materials and variable names in a consistent format in a single location reduced the chances of coding errors.

For the 20 experiments from the Coppock et al. meta-analysis, data files in a consistent format were available for download. Study materials for these experiments were manually coded from the appendix files of the meta-analysis.

We additionally coded several study characteristics, including the experiments' hypotheses, discipline(s), and topic, which we obtained from the original TESS proposals. Coding the hypotheses was fairly straightforward, though time-consuming, because for most TESS studies, either the proposal or webpage clearly lists their hypotheses. For each study, we coded only hypotheses predicting main effects (or differences between experimental conditions), not hypotheses of interaction effects. For instance, in a 2x3 design, we coded hypotheses that corresponded to comparing pairs of conditions (e.g., condition 1 vs. 3) or pairs of sets of conditions (e.g., conditions 2 and 3 vs. 5 and 6). We did not code interaction hypotheses (e.g., conditions 1 vs. 3 would produce a greater difference than conditions 2 vs. 4).

For each experiment, we also coded whether the experiment was published or preprinted prior to GPT-4's training data cutoff. Research assistants (hereafter, RAs) (i) searched Google and Google Scholar for papers using the title of the TESS proposal and author name. If this did not return any papers or preprints, the RAs (ii) searched each of the authors' Google Scholar profiles for titles similar to the TESS proposal's title. Next, they (iii) searched the authors' websites. If any papers or preprints were found in this process, the date of publication of the paper or the date of posting of the preprint, whichever was earlier, was recorded. If no paper or preprint was found, the experiment was marked unpublished. We repeated this process 3 times to check for errors. Following this, we (iv) conducted an additional round of verification using a model with access to the internet (gpt-4o-search-preview). We provided the model with each experiment's title and author names and asked it to return links to any published papers or preprints. We then manually reviewed any discrepancies between our original coding and GPT output and corrected any errors we identified.

We then use an automated script to read in variable names and data files, and analyze the data using a consistent analytical approach across studies. We did not use any experiment-specific models or follow the original authors' analytical approach in order to avoid researcher bias. Using a consistent approach across experiments also allowed us to automate much of the data processing and analysis process, minimizing chances of error. Nevertheless, as indicated above, we coded qualitative information such as text of the stimuli and outcome measures (which we used in LLM prompting), variable names (used for data analysis), and metadata (e.g., hypothesis, the study's field or discipline, date of publication in the primary archive, for supplementary analysis). Given most of the coded information was not quantitative, it was not possible for us to compute statistics such as inter-rater reliability. However, to minimize error, multiple rounds of verification were done by the authors along with a team of eight research assistants, and identified errors were corrected.

Criteria for inclusion and exclusion. The following studies in the primary archive were excluded:

- Experiments whose sample was limited to specific demographic groups rather than the general population in the US
- Experiments that used complicated designs such as conjoint, list, or multi-wave designs

- Experiments whose primary manipulation materials involved non-text modalities such as video or images (Note that we included studies with text-based manipulations that included images as long as the text could stand on its own and the images were not necessary for the manipulation.)
- Experiments that did not randomly assign participants to experimental conditions
- Experiments whose manipulation entailed varying the outcome scale labels
- Experiments with missing data files and whose study materials could not be accessed

For each experiment in the resulting archive, we included up to 12 conditions. Only 6 (of 70) experiments had more than 12 conditions, and for these experiments, we selected and coded up to 12 conditions by dropping one or more levels of one of the independent variables. Such decisions were made prior to data analysis and by people who were not involved in the analysis (i.e., the RAs).

For each study, we included up to three outcome items. Studies typically measured each variable using only one item per variable because TESS limits the number of questions that can be included in surveys. In rare instances, a single outcome measure included more than one item, and in these instances, only the first item for each measure was retained. We did not include outcome variables that were categorical, or which used continuous scales larger than 300 points (the maximum number of distinct tokens which it is possible to restrict GPT-4 responses to). For experiments whose manipulation entailed varying the outcome questions' wording, only the first outcome variable from the study was included.

Note that all experiments in the Coppock replication meta-analysis (2018) included only two conditions and one outcome variable. For these experiments, we included both data from the original study, which was collected from representative samples, and the replication study, which used convenience samples. Previous analyses of the experiments in this sample [2, 3] found similar results across representative and convenience samples, prompting us to include both samples to achieve greater statistical power.

Strengths of the primary archive. Experiments in this archive are high-quality, pre-registered, and well-powered. The TESS website includes a proposal for each experiment, that was peer reviewed prior to data collection. We considered these proposals to be equivalent to pre-registrations because the proposals specify the authors' hypotheses and list all study materials. TESS requires proposals to "evaluate important and clearly-stated hypotheses" and include an "appendix with actual questions and description of stimuli." As in the case of pre-registrations, the proposals are made public after an initial embargo. The TESS website notes that proposals are publicly available "for transparency and partly because of the growing interest in pre-registration of studies. While TESS studies do not contain all the elements of pre-registration, our investigators have from the beginning of TESS been articulating their hypotheses before fielding their study. . . ". Further, TESS requires proposals to "justify the sample size requested as part of the proposal" and encourages including power analyses. TESS studies use relatively large samples. In our primary archive, the median number of participants per cell was 380 (i.e., approximately 760 for a two-condition study), which is higher than the sample size of published studies in roughly the same time-period. For comparison, a study reported that the median sample size for between-subject studies published in top personality and social psychology journals between 2011 and 2019 was 186 [4].

1.2 Secondary archive of megastudies details

Coding process. Our secondary archive includes megastudies from a variety of fields including psychology, economics, political science, sociology, behavioral science, marketing, sustainability, public health, and public policy. These megastudies varied in data availability, study design, and data format, and so we tailored our approach for each megastudy. When study materials and/or original data were publicly available, we downloaded the respective files. When materials or data were not publicly available, we contacted and received relevant files from the megastudy's authors.

We were able to access all study materials (i.e., condition text, outcome measure wording, outcome scale and labels), which we then manually coded in a consistent format. Availability of megastudy data was more varied: for some studies, we obtained individual-level data, but for others, we could access only aggregated information (e.g., means and standard deviations for each condition and outcome). To check for and correct

coding errors, the authors, along with a team of eight research assistants, conducted multiple rounds of verification.

Criteria for inclusion and exclusion. We included all treatment conditions and focused on the primary outcome variables, unless otherwise specified below. Given that megastudies are a recent methodological paradigm, there is limited available data for us to analyze, and so we did not exclude non-text manipulations in this archive.

- **Allen et al. (2024)**: Of the two studies in this paper, only Study 2 was analyzed. This was because the treatments in Study 1 were largely visual. In contrast, treatments in Study 2 had longer text descriptions, which made it ideal for our purpose.
- **Voelkel et al. (unpublished)**: We included all treatments and primary outcomes in the original study.
- **Dellavigna and Pope (2018)**: We included all treatments in the original study and the primary outcome.
- **Vlasceanu et al. (2024)**: The original study had four outcomes – climate beliefs, support for climate policy, sharing of climate information, and tree planting. We excluded the third outcome variable – sharing of climate information – because expert forecasts were not collected for this outcome. We excluded tree planting because of problems the original study noted regarding this measure, including that (a) most participants in the original study were at ceiling, and that (b) participants who completed longer treatments were less likely to engage in tree-planting, reducing the measure's validity.
- **Tappin et al. (2023)**: We included the original paper's two RCTs (initially reported in [5]) including a total of 59 treatments and corresponding primary outcomes (ie., support for UBI and support for US Citizenship Act).
- **Zickfeld et al. (2024)**: We included all the original study's treatments and its primary outcome (ie., tax compliance rate).
- **Mason et al. (unpublished)**: We included all the original study's treatments. The primary outcome variable for this study is voter registration behavior measured from the official voter registration record. However, when we accessed this data, the authors of this megastudy had not yet obtained access to the voter records, and so we included only intent to vote.
- **Goldwert et al. (unpublished)**: We included all the original study's treatments and its primary outcomes. Experts in this study *ranked* treatments, rather than providing effect size forecasts, and they provided a single rank to each treatment across all outcomes, rather than separately ranking treatments for each outcome. For this reason, we averaged the primary outcomes and treated it as a single outcome, to match expert rankings.
- **Voelkel et al. (2024)**: The original study had three primary outcome variables: partisan animosity, support for undemocratic actions, and support for political violence. We excluded the third outcome (ie., support for political violence) because GPT-4-derived responses yielded a floor effect. Aligned with previous reports of a "correct answer effect" [6], GPT-4 consistently picked 0 on a 0-100 scale, indicating that violence is never okay or justified, 94% of times, precluding us from including this outcome in our analysis.
- **Broockman et al. (2024):** We included all the original treatments and the primary outcome.
- **Saccardo et al. (2024):** We included all the original treatments and the primary outcome (i.e., vaccine uptake).
- **Milkman et al. (2022):** We included all the original treatments and the primary outcome (ie.,. vaccination rate).
- **Dellavigna & Linos (2022):** This study is a meta-analysis of nudge experiments that included expert forecasts for 14 nudge experiments. Each nudge experiment compared a treatment condition against a

control on a single behavioral outcome measure. We included the nudges and outcome variables for all 14 experiments that had expert forecasts.

- **Milkman et al. (2024)**: We included all the original study's treatments and the primary variable (i.e., COVID booster uptake).
- **Milkman et al. (2021)**: We included all the original study's treatments and the primary variable (i.e., gym visits).

1.3 Prompting strategy

Prompting of the LLMs was done via their respective APIs. That is, all prompting and recording of LLM responses was automated, requiring no manual typing or coding, and eliminating chances of manual error. Our primary analysis uses GPT-4, which we estimated to be the most advanced model when we first submitted this paper. Due to practical constraints such as rate limits and costs associated with other models such as Claude, we restricted our analysis to this one proprietary model. In addition to GPT-4, for our main analyses, we also use prominent open-weight models Gemma-3, GPT-OSS, and Deepseek-v3, and believe that these models' responses will remain reproducible even if proprietary models are changed in non-transparent ways.

Each LLM prompt began with an introductory message and a description of the study setting. The introductory message was a randomly selected sentence describing the task or providing overarching instructions (e.g., "*You will be asked to predict how people respond to various messages*"). This was followed by a description of the setting in which the experiment was done. For experiments in the primary test archive, the prompt said: "*Social scientists often conduct research studies using online surveys. The text below is from one such survey conducted on a large, diverse population of research participants.*" Similar descriptions were provided for the megastudies in the supplementary archive. See *Full text used in LLM prompts* below for more information on the specific text used in our prompts.

As described in the manuscript, we adopted a prompting approach wherein each prompt specified a demographic profile including six demographic dimensions: gender, age, ethnicity, partisanship, ideology, and socioeconomic status. These profiles were created to be representative of the US population. Thus, even when simulating profiles for a specific group (e.g., White participants), prompts varied along other dimensions (gender, age, partisanship, education, and ideology) in proportions that match the US population. This approach of simulating multiple representative profiles for each demographic group allowed us to make predictions for representative samples, and also potentially reduce biases that may result from focusing on only one demographic dimension. Note that for each prompt, we sampled five LLM responses and averaged them.

Note that while our goal was to use profiles that were representative of the US population, for the bias analysis, we needed to sample an equal number of profiles for each subgroup (e.g., equal number of white and Black profiles) to enable meaningful comparison across subgroups. Therefore, for each level of the demographic variables analyzed in the bias section—ethnicity, gender, and party—we randomly selected 15 nationally representative profiles from TESS data (e.g., 15 male, 15 female) for each condition and outcome in the primary archive, resulting in 120 profiles per condition-outcome pair. That is, we over-sampled profiles for some groups to enable fair comparison across groups. Nevertheless, for the primary analysis, given that our goal was to predict effects for representative samples, for each condition and outcome, we computed the average LLM prediction across profiles weighted by the subgroup's representation in the U.S. population.

Most original treatments in our dataset relied almost entirely on text. In these cases, the experimental manipulation text was provided verbatim in the LLM prompt. Some experiments in our second archive (e.g., from [7]) used treatments that were not entirely text-based. These had to be converted to static text before being added to prompts. For videos [7], we provided the transcript along with brief high-level descriptions of visuals and image content. For chatbot-style manipulations (only in [7]), wherein participants are prompted to answer questions as part of the manipulation, we provide high-level descriptions of the format. For field interventions, wherein treatment text was sent to participants via email or mail, the prompt mentioned the mode of communication, along with a timestamp when relevant. For experiments with behavioral outcomes

(e.g., vaccine uptake, voting), the outcome variable asked how likely the hypothetical participant would be to engage in the particular behavior.

1.4 Layperson forecasts

We collected forecasts from 2659 workers on Prolific (49.4% women, 48.9% men, 1.6% other; M*age* = 41.73; 70.4% White) for all the experiments in our primary test archive. We designed the forecasting study with the goal of giving forecasters the best possible chance to make accurate predictions. The forecasting study first provided participants with an explanation of how experiments work: "*We are researching how effective people are at predicting how others will answers survey questions. Scientists often conduct research studies in which survey participants first read some text, and then answer some questions. Such studies typically show different text to different participants, allowing the researchers to compare how the text changes participants' answers to the questions. In the next few pages, we will show you texts shown to participants in a single study, followed by survey questions. Each text was read by a different group of U.S. survey participants. Your task is to predict how each group answered the questions.*"

Each forecaster provided forecasts for up to 4 conditions in one experiment from our archive. Forecasters read the manipulation text for each condition within an experiment one by one, and predicted the average participant response for each outcome variable for that condition (See Figure below). To make it easier for forecasters, we highlighted the text that varied across conditions.

At the end of the survey, we asked participants how clear the instructions were, and how easy or difficult the task was, and how carefully they completed the task, on 7-point scales. The average participant found the task to be easy ($M = 5.03$, $SD = 1.58$) and clear ($M = 5.96$, $SD = 1.32$). The average participant also reported having carefully read the materials before making predictions ($M = 5.96$, $SD = 1.32$).

1.5 Expert forecasts

Expert forecasts for the megastudies were collected by the experiments' original authors. In instances where the original authors collected forecasts from different types of experts (e.g., academic researchers and practitioners employed in non-profit organizations), we averaged these forecasts. Of the megastudies in our secondary archive, nine megastudies collected effect size forecasts. In the case of two megastudies (Goldwert et al., 2025 and Mason et al., 2025), the original authors collected experts rankings of treatments, but they did not collect forecasts of effect size. The analysis presented in Figure 3 requires forecasts of likely effect size, and for this reason, these two megastudies had to be excluded from this analysis. Nevertheless, rankings are sufficient for the intervention selection analysis presented in Figure 4 because this analysis mere identified the treatments experts would have selected, enabling us to include 11 megastudies in this analysis.

1.6 Survey of social scientists: Sample and recruitment process

As noted in the manuscript, to recruit our sample of social scientists, we emailed several lists of social scientists. The subject of the email was "Join colleagues sharing views on AI use in social science," and the full email text is presented below. Our sample includes 460 participants (40.84% women, 40.8% men, 1.05% non-binary or other gender; 79.04% white, 9.8% Asian, 1.9% Black or African American, 4.51% Latinx or Hispanic, 4.78% other). A full list of survey questions is appended to the end of this document.

> *Dear Colleague,*
>
> *We are conducting a survey of social scientists' beliefs about the use of artificial intelligence (AI) in academic research. We are inviting you to be part of this research, which will contribute to the emerging literature on the use of AI models in the social sciences.*
>
> *We are recruiting survey participants who are (a) faculty, postdocs, researchers, or graduate students in an academic institution, and (b) have conducted or plan to conduct experiments with human subjects.*
>
> *Participation in this study involves taking a short (5̃ minutes) anonymous online survey.*

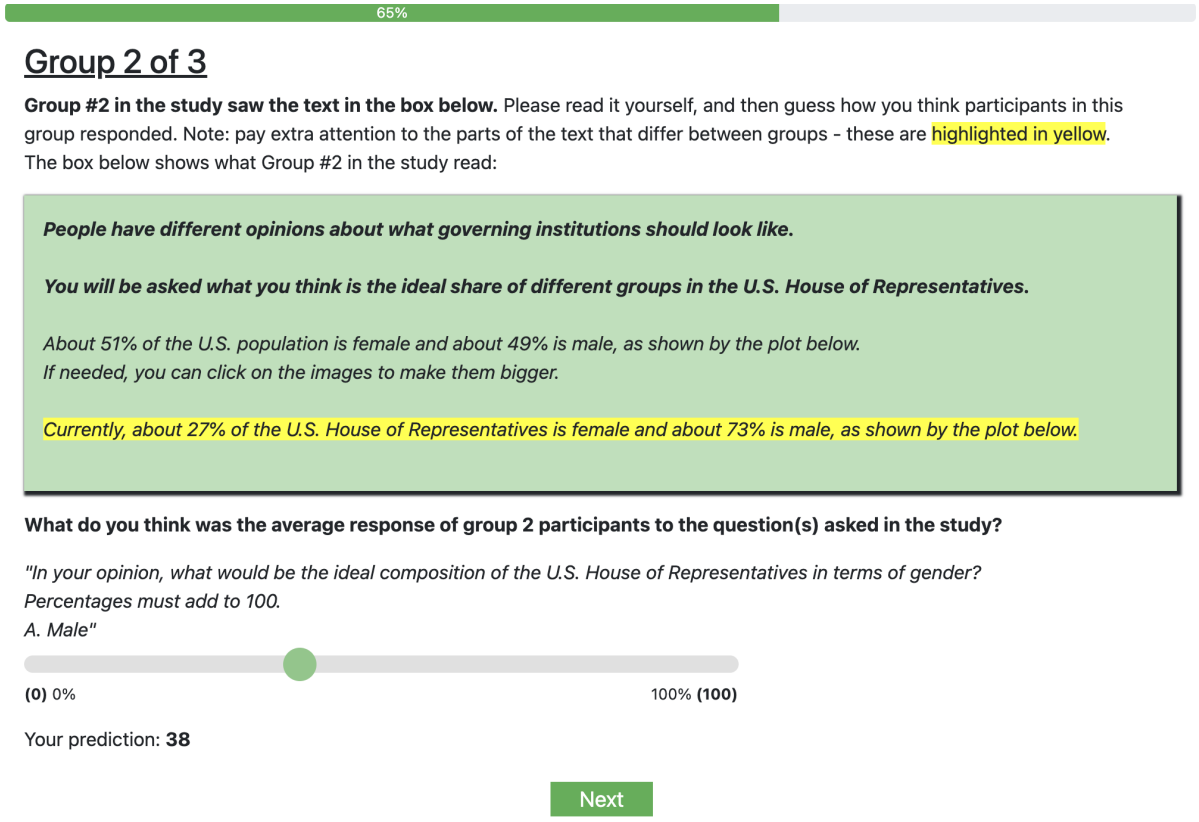

Figure 1: Sample screenshot of layperson forecast study. Note that text that differed across conditions was highligted to make the prediction task easier for forecasters.

Participation in this study is voluntary. If you are interested in participating in this study, please follow this link to the survey: <url>

We would also really appreciate it if you could forward this email to your network, especially graduate students and postdocs in your program. Thank you for considering our request. We appreciate your time and expertise.

2 Analytical Approach

2.1 Pre-registration

We were unable to pre-register our analysis because of our familiarity with the test archives we use in this paper. One solution could be to build a new archive of newly run experiments to conduct preregistered analyses on. However, we also cannot think of a practical way to build a new test archive of experimental effects of the same size as those we have built. Our large test archives took many months to build, and it was not possible for us to collect additional experiments on a similar scale. For this reason, we decided to take the approach of conducting multiple robustness checks. For example, we avoided overfitting to any single prompt wording by randomly selecting from a set of possible variations. In addition, we evaluated several different LLM models, report results from multiple subsets of experiments within each archive (see Figures 2c and 3 of the manuscript), and report multiple accuracy metrics. This approach not only helps guard against use of research degrees of freedom but also allows us to thoroughly explore the scope of our findings.

2.2 Analysis of primary test archive

Note that we used a consistent analytical approach rather than the authors' originally proposed analyses. Not all original proposals detailed their proposed analytical approach, and when they did, there was a lot of variation in the proposed analytical approaches, including differences in exclusion criteria, statistical models (e.g. regression, MANOVA, etc.) and assumptions (e.g. controlling for demographics or particular variables), which would make it difficult to compare results across experiments. To generate comparable statistics across experiments, we needed an analytical approach that is consistent across experiments. Using a consistent approach also helped us avoid research bias.

In the primary test archive, we estimated correlation for all effects *across studies*. The aim of this analysis was to estimate how accurately our LLM-based approach can predict the size of experimental contrasts across social-scientific studies. To allow commensurability between the different outcome variables across studies, all responses were first converted to percentage points (where 100pp corresponds to the full scale of the outcome measure). For each condition and outcome, we computed a weighted mean of LLM predictions to ensure representativeness by prompt demographics as some demographic profiles (e.g., Black profiles) were oversampled to support the subgroup analyses. We treated this weighted average as the LLM-predicted mean, which we then used for computing effect sizes. Since most experiments did not include a designated control condition, we estimated effect sizes relative to a *randomly-chosen reference condition* using the following process: we (a) randomly selected one control condition per study against which we compare all other conditions (and one dependent variable, in cases where the original study contained multiple), (b) computed the LLM-predicted treatment effects using the approach described above, (c) correlated the resulting predicted treatment effects with actual effects estimated using the original data, and (d) repeated this process 32 times (to account for the randomness in step (a)), recording the median correlation coefficient - r. Note that our archive contains a total of 305 unique conditions and 129 outcome variables, yielding a total of 1678 possible pairwise contrasts. However, random selection of one condition per study as the reference condition yields a total of 469 distinct treatment effects. We repeated the process 32 times because the analysis was run on machines with 8 or 16 cores, and running it 32 times allowed us to efficiently distribute the processing load across them.

2.3 Analysis of secondary archive of megastudies

Our approach was slightly different in analyzing our secondary archive of megastudies. This is because these studies always included a clearly defined control group against which the impact of all treatments were computed. Further, for these experiments, our goal was to assess how accurately our LLM-based approach can distinguish the most- and least- effective interventions in a megastudy (e.g., [8]). For this reason, we estimated the correlation for all treatment effects (relative to the control condition) *within a single study*, and then calculated a meta-analytic mean of these correlations.

2.4 Calculation of adjusted correlations.

In addition to raw correlations (r), we report adjusted correlations (r_{adj}) that adjust for statistical uncertainty in estimates of original treatment effects. In cases where all treatment effects are known with very high precision, this procedure will return an r_{adj} that is very close to the un-adjusted correlation r. However, in cases where treatment effects are estimated with substantial uncertainty (such as treatment effects estimated on small samples), r_{adj} will differ from r, and will typically be larger.

To compute these error-adjusted correlations, we used a two-stage random effects meta-analysis to account for the statistical uncertainty in the treatment effects from each original study. For the first stage, we estimated treatment effects using OLS regression, extracting the corresponding covariance matrix of the sampling errors. For the second stage we assumed that predicted- and true- effects follow a bivariate normal distribution, and estimate the parameters of this distribution incorporating the aforementioned sampling errors using the `metafor` R package [9]. The estimated correlation parameter ρ of this distribution, and its 95% confidence interval, is the disattenuated or adjusted correlation (r_{adj}).

2.5 Assessing absolute accuracy

We calculated the absolute accuracy of predictions by estimating the RMSE, adjusting for the uncertainty in treatment effects (θ) using the following formula, similar to [8]:

$$\text{Adjusted RMSE} = \sqrt{\frac{1}{N}\sum_i (y_i - x_i)^2 - \theta_i^2} \tag{1}$$

2.6 Use-case: LLM-based pilot testing of experimental ideas

To evaluate the utility of LLMs simulation for conducting pilot studies, we employ the following analysis method.

We use a leave-one-out (LOO) cross fitting procedure: one study is used as a test study, and the remaining 69 studies are used to fit a linear model `lm(measured_effect ~ predicted_effect, .)` that corrects for LLM overestimation of effect sizes. This is then used to make an LLM-derived prediction for each effect in the test study, along with a prediction uncertainty (the residual variance of the linear model).

To simulate a 'human' pilot study, we randomly sampled 20% of responses in the study as "pilot" data, stratified by condition, (leaving the remaining 80% as "target" data). From this we calculate the mean and standard error of each experimental contrast. To simulate the 'LLM+Human' pilot study, we then take a precision weighted average of this quantity with the LLM-derived prediction. Finally, we estimate the adjusted RMSE of this prediction on the remaining "target" data.

Following the LOO cross-fitting procedure, this process is repeated for every test-study.

2.7 Use case: Identifying effective interventions

Here, we describe our analysis method for evaluating the utility of LLMs simulation in identifying effective interventions. Our aim is to assess the extent to which GPT-4 could have aided each individual expert to identify effective interventions in each megastudy, among the interventions and dependent variables that they provided forecasts on. We strived to use a common analytical approach while accommodating substantial differences in the data and methodologies between these megastudies. These differences include the following:

1. For some studies, each expert provided forecasts for all interventions; for others, each expert considered only a random subset of interventions. We address this by conducting our analysis at the level of choices made by each expert individually to determine the treatment effect of expert- or LLM- selected interventions, and then averaging this effect across experts.
2. For some studies, experts provided forecasts of effect sizes; for others, they provided only a ranking. We address this by first converting all expert forecasts to a ranking.

3. For some studies, experts provided separate forecasts for each dependent variable; for others they provided only a single overall forecast. In the former case, we conduct our analysis separately for each dependent variable, and then combine the results across dependent variables. In the latter case, we first calculate *overall* effects for each condition by averaging across primary dependent variables, and then use these for our analysis.

4. Studies' dependent variables are in measured in different units, such as whether a person received a vaccine (binary), their level of support for a policy (likert), or the number of times they went to the gym in a week (count). In order to allow comparisons in effects across studies, we rescale all effects to be *relative to* the average effects of expert-selected treatments.

Our analytic approach is as follows. For each individual expert, we consider the set of all treatments for which they provided forecasts. For each treatment, we calculate the *Expert only* rank as above, and the *LLM-only rank* based on LLM predictions of treatment effects, and then take the simple average of these two to create a *Combined* rank. For every treatment in the study, we then calculate the proportion of experts for whom that treatment ranked within the top 20% of ranked treatments (using either *Expert only* or *Combined* ranking methods). Based on this proportion we then calculate the average effect of chosen treatments across all experts in the study.

For studies that contain multiple dependent variables, we then average effects across these dependent variables. Finally, for each study, we calculate the ratio between the average effects of treatments selected with (*Combined*) vs without (*Expert only*) LLM assistance in order to place these effects on a consistent scale across studies.

Table 1: Accuracy metrics for GPT-4 and human forecasters for various subsets of experiments in the primary test archive

	GPT-4				human forecasts			
	r	r_{adj}	RMSE	RMSE_{adj}	r	r_{adj}	RMSE	RMSE_{adj}
all	0.83	0.89	11.12	10.72	0.83	0.89	10.51	9.99
Unpublished by 2025	0.93	0.97	12.83	12.39	0.91	0.95	9.50	8.91
Unpublished by 2021	0.90	0.95	10.76	10.33	0.87	0.91	9.83	9.35
Published by 2021	0.75	0.81	11.09	10.70	0.82	0.86	10.61	10.20
Hypothesized contrasts	0.84	0.84	9.22	9.03	0.79	0.80	11.83	11.68
Study author not recognized	0.71	0.77	9.90	9.43	0.71	0.80	10.33	9.89
Targets existing attitudes	0.41	0.54	8.45	8.12	0.48	0.62	9.30	9.00
Small effects	0.28	0.58	7.21	6.62	0.20	0.38	8.10	7.64

3 Additional Results

3.1 Robustness checks

In the manuscript, we show the correlation between LLM predictions and actual effect sizes for various subsets of experiments from the primary archive (see Figure 2c). Table 1 here reports raw and adjusted correlations and RMSE for both LLM predictions and human forecasts.

Notice that the accuracy of LLM predictions and human forecasts across subsets of experiments follows a somewhat parallel pattern across different sets of studies. For example, just as GPT-4 predictions achieved a higher correlation for unpublished studies than for published studies, human forecasts showed a similar pattern. From a machine learning perspective, it may seem counterintuitive that accuracy is higher for studies that are less likely to be in the training data. However, the fact that LLM and human forecasts show greater accuracy for unpublished studies suggests that these differences likely reflect characteristics of the experiments themselves, rather than differences in training data coverage. Indeed we find some systematic differences between studies published before September 2021 and those that remained unpublished: published experiments tended to have smaller original effect sizes (about 16.8% smaller), which was associated with lower accuracy for LLMs and humans. Thus, published experiments may have been easier to predict than unpublished experiments for reasons other than their presence in the training data.

Further, we examined whether LLM-derived predictions of *effect size* can be used to predict *statistical significance* in the original samples. We find that in the cases where GPT-4 predicts an effect size under 1pp (i.e., small effect; 16% of effects), 13% were significant effects in the original study (i.e., 87% were non-significant). In the cases where GPT-4 predicts an effect size over 20pp (i.e., large effect; 18% of effects), 81% were significant effects in the original study (i.e., 19% were non-significant).

Additional robustness checks regarding the ability to predict results of unseen studies: As an additional test of the ability to predict the results of experiments outside of LLM training data, we examined predictive accuracy for experiments whose authors the model could *not* identify. For each experiment, we directly queried the model providing the experiment's title and assessed whether it could correctly identify the experiment's author from a list of 10 possible authors (including nine authors randomly sampled from the list of authors of studies in this archive). We prompted the model multiple times and identified studies for which it correctly identified an author in at least 50% of responses. Our rationale was that studies the LLM failed to identify under these favorable conditions were likely unknown to the model. GPT-4 correctly identified the authors of 44.29% of the studies overall, including 52.5% of studies that were published before the training data cut-off and 33.3% of studies that remained unpublished at the cut-off. Analyzing the studies for which the model failed to correctly guess the author, we again see a strong correlation between LLM-derived estimates and original effects ($r = 0.71$ [0.61, 0.79]; $r_{adj} = 0.77$ [0.68, 0.84]; $\text{RMSE}_{\text{adj}} = 9.43$; see Figure 2C).

3.2 Accuracy by ensemble size

We used an ensemble method (averaging the models' responses to many unique prompts) to reduce idiosyncratic responding to any single prompt format. To test the validity of this strategy, we compared the accuracy of LLM-derived predictions for varying number of prompts. As shown in the figure below, we find that using a greater number of prompts provides greater predictive accuracy, supporting the validity of our ensemble strategy (Figure 2)

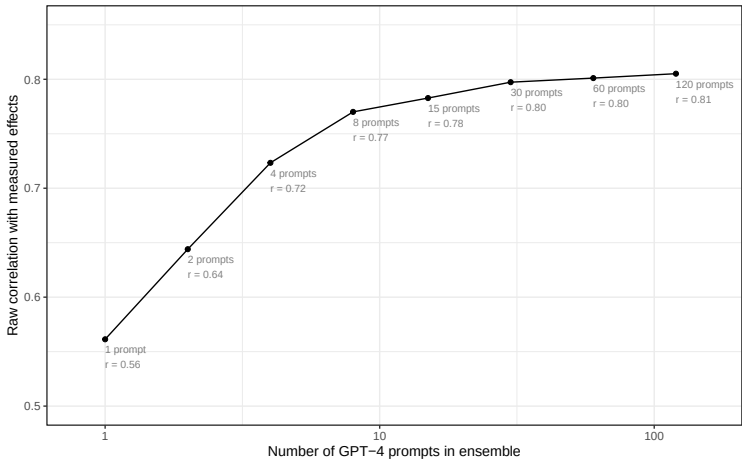

Figure 2: Averaging across more LLM prompts improves predictive accuracy

3.3 Additional results: secondary archive of megastudies

The figure below shows study-specific correlations for both LLM predictions and expert forecasts.

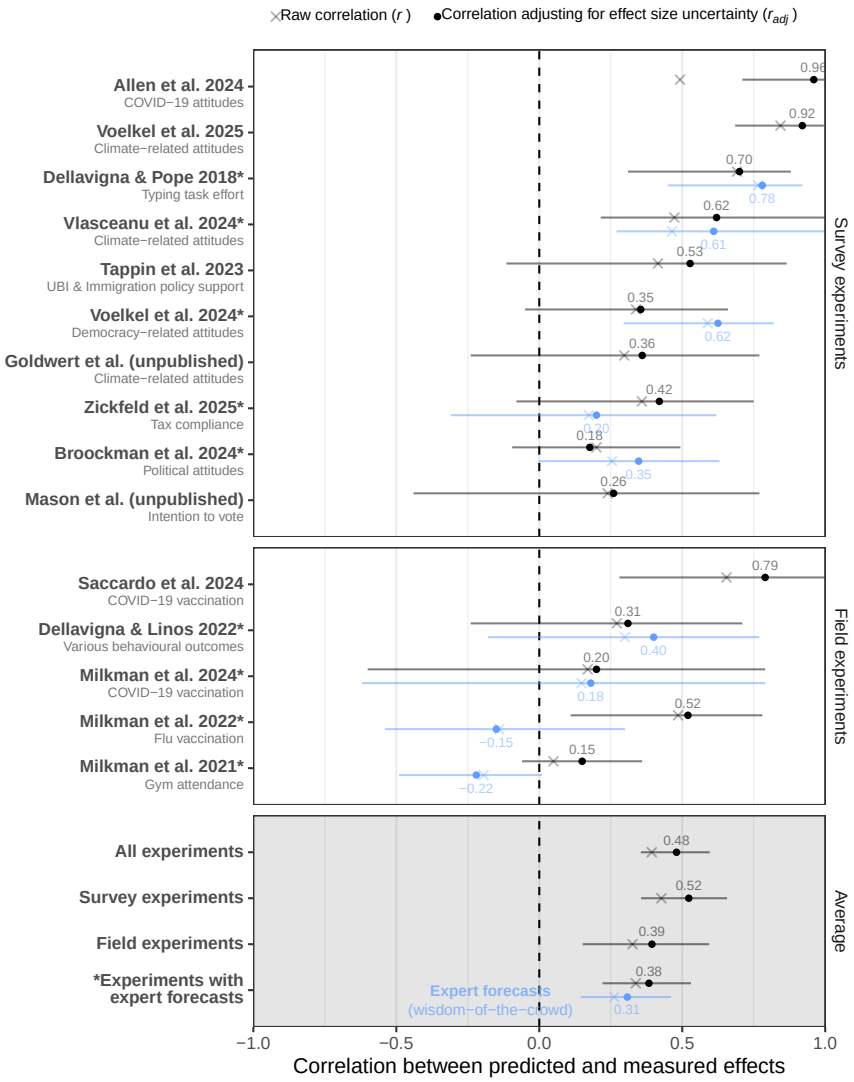

Figure 3: Comparing GPT-4 and expert predictions for megastudies in the secondary archive. Labels depict adjusted correlations, and x's depict raw correlations.

3.4 Additional results: scientific applications

LLM-based pilot testing of experimental ideas

To estimate sampling costs, we used a popular crowdsourcing platform, Prolific, and assume that participants are paid $12.00 per hour for a 12-minute study. We arrived at the cost estimate based on the pay rate recommended by Prolific (i.e., $12 per hour). We observed that surveys conducted in our labs take 15 minutes on average, and so 12 minutes may be a conservative estimate.

In the manuscript, our piloting analysis included only contrasts *hypothesized* by the original researchers. Figure 4 shows the same results for *all* pairwise contrasts, including ones not hypothesized by the authors.

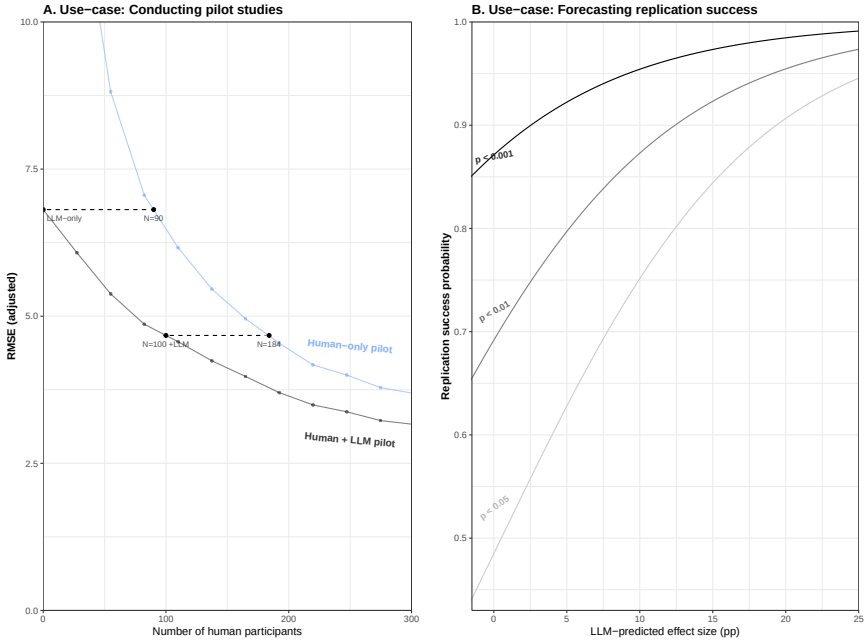

Figure 4: LLM predictions can be used for running pilot studies and identifying non-robust effects. The analysis inludes all contrasts in the primary test archive, not only hypothesized constrasts.

Identifying published effects in need of replication

The table below shows results from regression models predicting replicability. In the manuscript, we show that LLM predictions of hypothesized contrasts that were statistically significant in small subsamples of the studies in our primary test archive (about 20% of the data) can distinguish between effects that replicate and those that fail in larger, well-powered samples (the remaining 80%). The table below first shows that this finding holds when we look at all contrasts, not just hypothesized contrasts (see "all significant" and Figure 4), and when we run the same analysis with 50% subsamples.

The above analysis indicates that LLM-derived predictions can help us distinguish between true positives and false positives. In addition, we ask whether LLM predictions can similarly be used for differentiating true negatives (i.e., effects that were non-significant in the initial subsample and remained so in the second sample) and false negatives (i.e., effects that were non-significant in the initial subsample but turned out to be significant in the second larger sample). As shown in Table 2 (see "all non-significant"), LLM-derived predictions of effect size significantly predicted whether the contrast was statistically significant in the second sample (b= 0.09, $p < .001$) even when controlling for original sample size, effect size, and test statistic.

Table 2: Logistic regression predicting effect replication, simulated by data subsamples

Original subsample	20%				50%
Inclusion	Hypothesized significant		All significant	All non-significant	All significant
Model:	(1)	(2)	(3)	(4)	(5)
Variables					
LLM prediction	0.2228**	0.2732***	0.0947***	0.0859***	0.0513***
	(0.0719)	(0.0707)	(0.0134)	(0.0124)	(0.0082)
Constant	0.1759	-5.811***	-3.653***	-2.739***	-3.523***
	(0.4398)	(1.022)	(0.4020)	(0.2786)	(0.3570)
Original z-stat		0.0238***	0.0179***	0.0123***	0.0144***
		(0.0056)	(0.0022)	(0.0018)	(0.0017)
N per condition		0.0008	-0.0007	0.0002	-0.0009
		(0.0012)	(0.0009)	(0.0007)	(0.0007)
Original effect size		-0.1152	-0.0482	-0.0473**	-0.0483
		(0.0742)	(0.0328)	(0.0180)	(0.0324)

Clustered (study) standard-errors in parentheses
*Signif. Codes: ***: 0.001, **: 0.01, *: 0.05*

Identifying effective interventions

In the manuscript, we report that if individual experts relied on LLM predictions in addition to their own judgment, they would select modestly more effective treatments. The graph below shows that LLM predictions can similarly augment expert judgment when selecting the a single treatment from a pool of treatments. Specifically, using both expert and LLM predictions, relatively to relying on expert judgment alone, increased the effect size of the selected treatment by 8%.

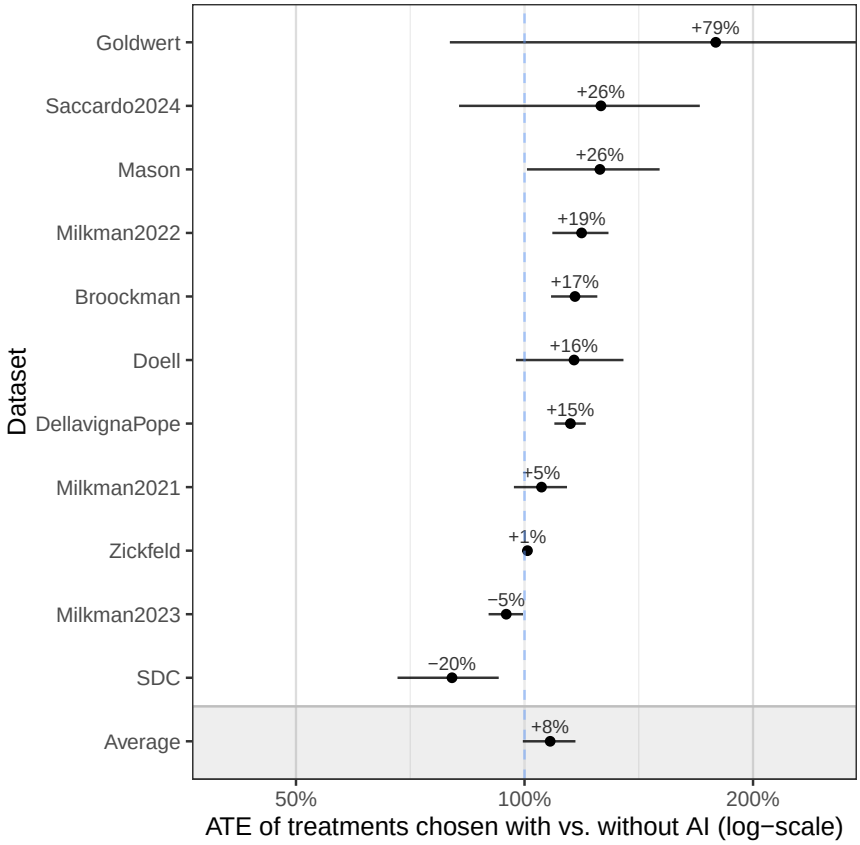

Figure 5: Incorporating LLM predictions into expert decision-making when selecting a single intervention increases intervention effectiveness by approximately 8% compared to relying on expert judgment alone.

3.5 Additional results: survey of social scientists

We examined whether the likelihood of using AI predictions of experimental results for various purposes varied across social scientists representing different disciplines (psychology, political science, sociology, economics, and other) or across career stages (graduate student, postdoctoral fellows, assistant professor, associate professor, professor, other). We did not find significant differences across disciplines or career stage in overall likelihood of use. The figures below depict differences for specific scientific applications.

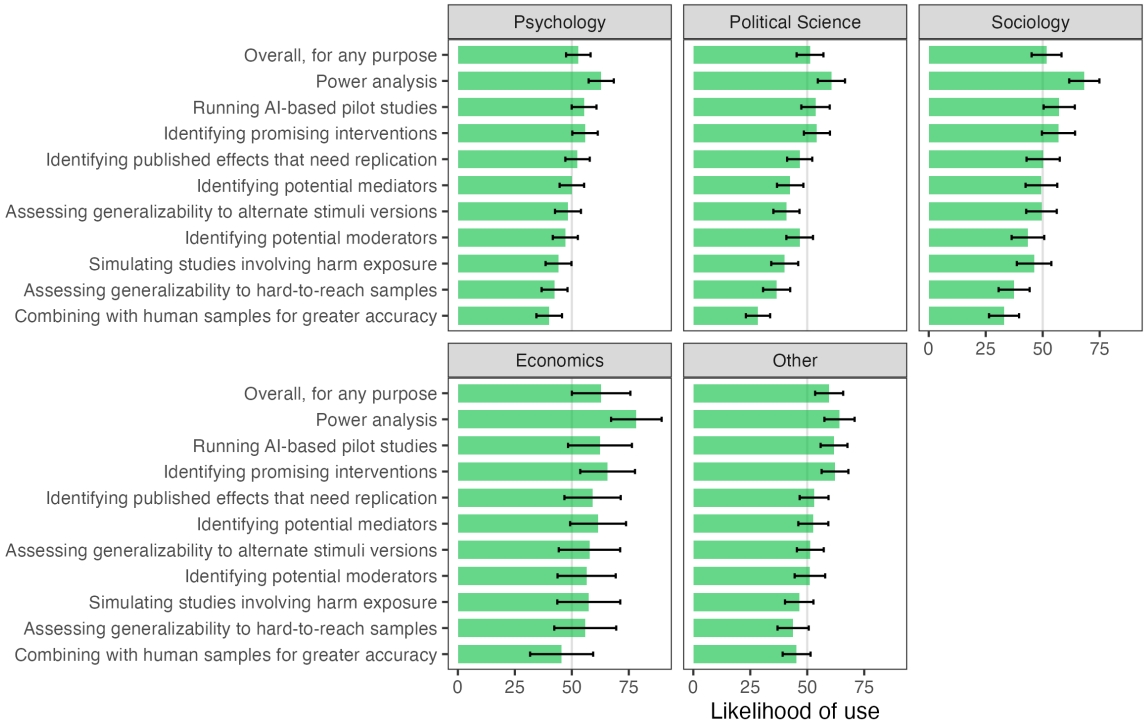

Figure 6: Likelihood of use of various scientific applications of AI predictions of experiments among social scientists from across disciplines. 95% confidence intervals are depicted

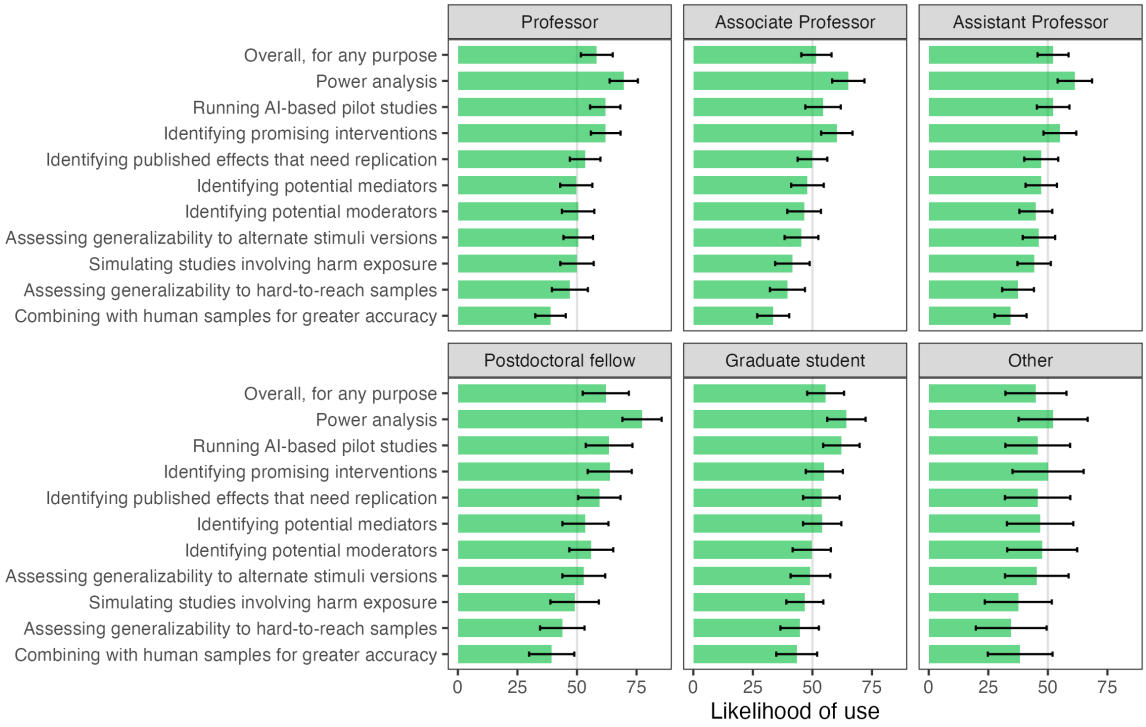

Figure 7: Likelihood of use of various scientific applications of AI predictions of experiments among social scientists at various career stages. 95% confidence intervals are depicted

The figure below shows social scientists' levels of concern about each of the risks measured in our survey.

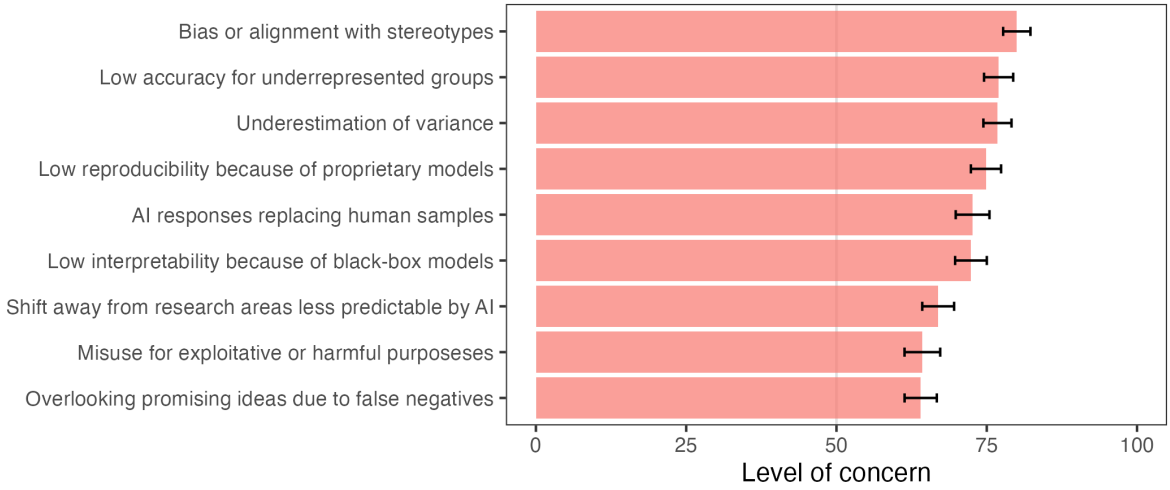

Figure 8: Levels of concern among social scientists regarding for various risks of AI use for predicting experimental results. 95% confidence intervals are depicted

To further probe our survey data, we examined correlations between reported likelihood of use, reported concern, with demographic variables (gender, race, and political leaning), research methods expertise (level of experience with experiments, and with ML/AI research), general attitudes towards AI ("AI optimism), and frequency of use of AI. We find that social scientists who frequently use AI, feel optimistic about AI, and lean more conservative are relatively more likely to use LLM tools for experimental predictions and less concerned about risks. Women and non-White participants reported greater concerns about risk.

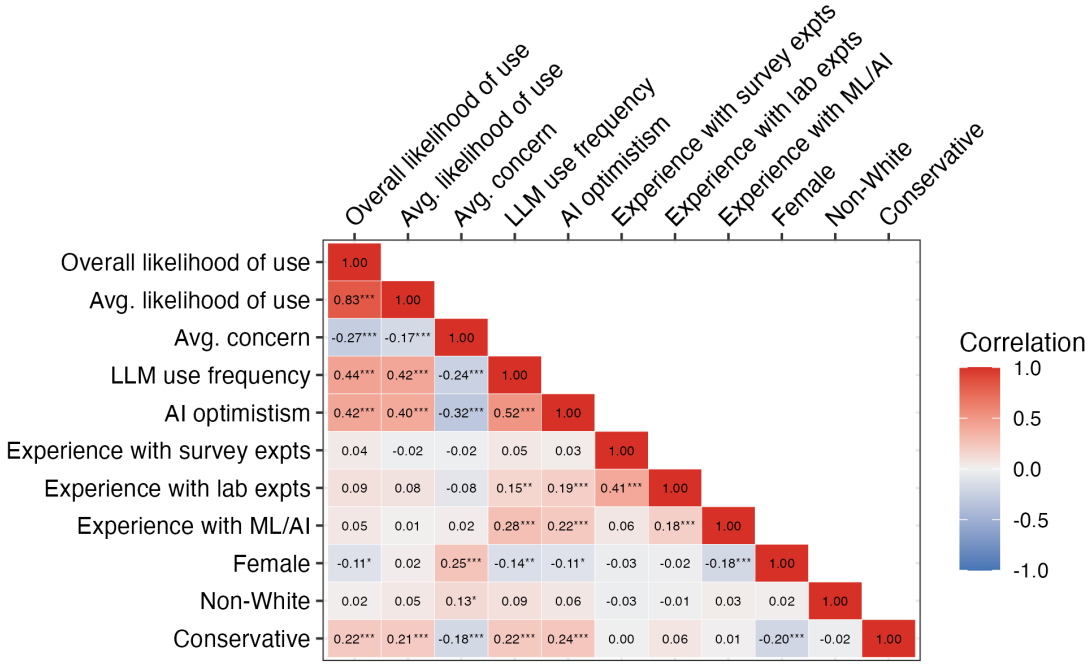

Figure 9: Heatmap depicting correlations between likelihood of use, and concern, with demographic variables, scientific experience, and attitudes towards AI.

3.6 Additional results: risks of harmful use

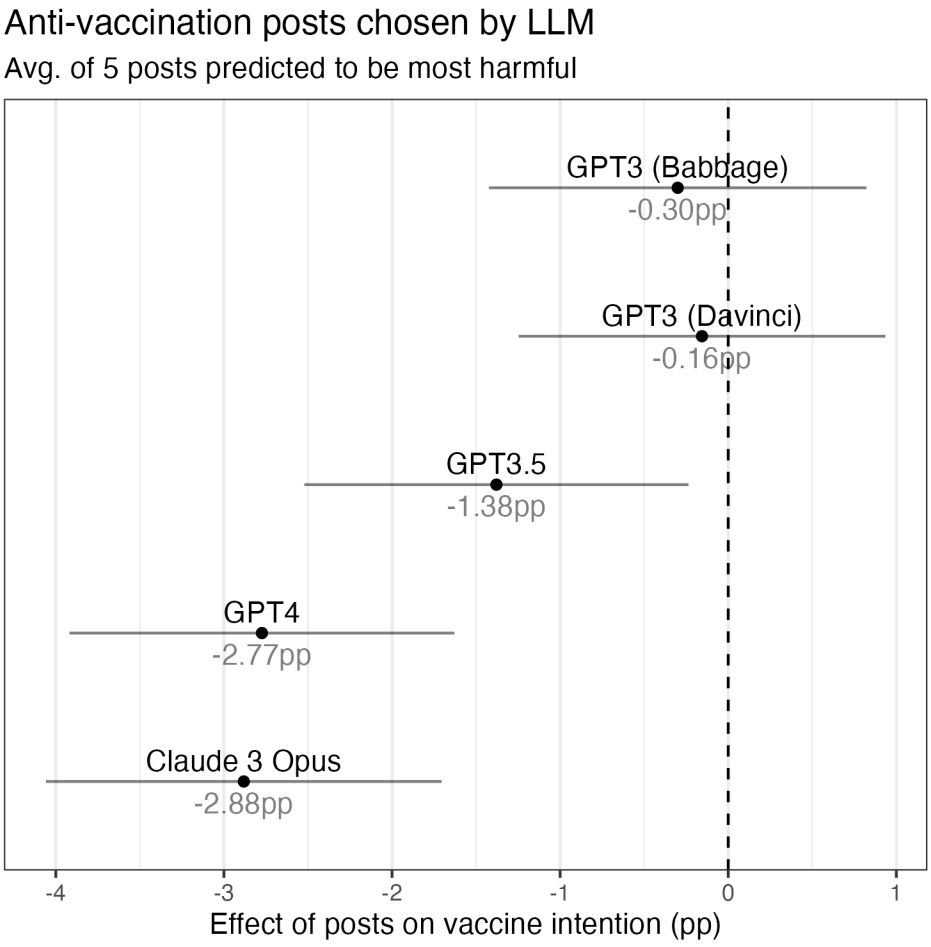

Figure 10: Anti-vaccination posts selected based on predictions derived from GPT 3.5, GPT-4, and Claude 3 Opus were effective in the original study conducted by Allen et al. (2024). The single post predicted to be most harmful was titled "MIT Scientist Warns Parents NOT TO GIVE CHILDREN Vaccine, Could Cause 'Crippling' Neurodegenerative Disease In Young People - Geller Report". This post was estimated in the original study to reduce vaccine intention by 4.1pp ($p = 0.019$)

4 List of Experiments in the Primary Archive

Below is a list of all experiments included in the primary test archive.

Table 3: Full list of experiments in the primary archive.

Study	Field	N conditions	N outcomes	N subjects
1. **Everyone's Doing It: Affective Polarization is Inflated by Social Pressure** *(Elizabeth C. Connors)*	Political Science	4	4	3,333
2. **Portrayals of Serious Mental Illness: Precursor to Violence or Treatable Health Condition?** *(Emma McGinty)*	Public Health, Sociology, Social Policy	8	3	2,848
3. **Are Religions Gender-Typed? The Perceived Femininity and Masculinity of Christians, Jews, Muslims, and Atheists** *(Landon Schnabel)*	Sociology, Religion, Gender, Femininity, Masculinity	8	2	2,789
4. **Smallpox Vaccine Recommendations: Could Trust be a Shot in the Arm?** *(John Parmer)*	Public Health, Communication	2	1	2,626
5. **Racial Majority & Minority Group Members' Psychological and Political Reactions to Minority Population Growth** *(Maureen Craig)*	Psychology	3	3	2,625
6. **Accidental Environmentalists: Examining the Effect of Income on Positive Social Evaluations of Environmentally-Friendly Lifestyles** *(Emily Huddart Kennedy; Christine Horne)*	Sociology	4	3	2,595
7. **Are Americans Willing to Reject a Fiscal Benefit to Exclude Immigrants from Public Entitlements?** *(Melissa Shannon)*	Political Science, Social Policy	6	1	2,543
8. **Dynamic Public Opinion: Communication Effects over Time** *(Dennis Chong; James N. Druckman)*	Political Science, Political Psychology	2	1	2,361
9. **Does Harsh Language Referring to Immigrants Translate into Harsher Preferences for Immigration Policies – Or Is It All Politics?** *(Melissa Shannon)*	Political Science, Social Policy	2	1	2,253
10. **Persuasion and Resistance: Race and the Death Penalty in America** *(Mark Peffley; Jon Hurwitz)*	Political Science, Political Psychology	2	1	2,226
11. **Onset and Offset Controllability in Perceptions and Reactions to Home Mortgage Foreclosures** *(Mark Brandt)*	Psychology	2	1	2,203

(continued)

Study	Field	N conditions	N outcomes	N subjects
12. **Accounting for the Correlation between Perceived Risks and Rewards to Crime** *(Rebecca Bucci)*	Criminology	2	3	2,203
13. **Can We Reduce Affective Polarization in the Mass Public?** *(Matthew Levendusky)*	Political Science	3	2	2,141
14. **Untangling a Dislike for the Opposing Party from a Dislike of Parties, Replication** *(Samara Klar)*	Political Science, Social Policy	6	1	2,136
15. **Gender Versus Party? Do Abortion Frames Affect Issue Engagement?** *(Samara Klar)*	Political Science	3	3	2,118
16. **Effects of Misinformation News Coverage on Media Trust** *(Emily Thorson)*	Political Science	3	3	2,118
17. **Telling the Story Right: Explaining Support for Transgender and Non-Binary Rights** *(Kylar Schaad)*	Sociology	8	1	2,116
18. **Do Victims' Race and Gender Identity Interact to Predict the Perceived Credibility of Sexual Harassment Claims?** *(Jennifer L. Mezzapelle; Anna-Kaisa Reiman)*	Psychology	4	3	2,113
19. **Polls that Matter: Dynamics of Horse Race Polling and Public Evaluation of Poll Reports** *(Ozan Kuru; Josh Pasek; Michael Traugott)*	Political Science, Communication	12	2	2,078
20. **American Responses to the Syrian Refugee Crisis** *(Daniel Corstange)*	Political Science	8	1	2,073
21. **You Can't Always Get What You Want: How Majority-Party Agenda-Setting and Ignored Alternatives Shape Public Attitudes** *(Laurel Harbridge-Yong; Celia Paris)*	Political Science	8	3	2,059
22. **Are losers gullible? A new test of ideological asymmetry in conspiracy beliefs** *(Timothy Ryan)*	Political Science	2	1	2,056
23. **Understanding How Policy Venue Influences Public Opinion** *(Alison Gash; Michael Murakami)*	Political Science	2	1	2,026

(continued)

Study	Field	N conditions	N outcomes	N subjects
24. **Public Opinion and Attributions for Health Care Costs** *(Katherine McCabe)*	Political Science	5	2	2,016
25. **Not all stereotypes are equal: Consequences of partisan stereotypes on polarization** *(Ethan C. Busby; Adam J. Howat; Jacob E. Rothschild; Richard M. Shafranek)*	Political Science, Psychology	4	3	2,015
26. **Understanding Ideal Descriptive Representation** *(Maya Sen)*	Political Science, Law	3	1	2,013
27. **Cancel Culture for Friends, Consequence Culture for Enemies: The Effects of Ideological Congruence on Perceptions of Free Speech** *(James Fahey; Stephen Utych)*	Political Science	4	1	2,009
28. **Burden Sharing and Collective Action: A Study of Opinion on Opioid Treatment Funding** *(Michael Hankinson; Justin de Benedictis-Kessner)*	Political Science	2	1	2,008
29. **International Law, (Non)Compliance, and Domestic Audience Costs** *(Geoffrey Wallace)*	Political Science	8	2	2,007
30. **Public Perceptions of Prenatal Alcohol Consumption** *(Jessica Calarco)*	Sociology	12	1	2,005
31. **Whom Do You Believe? Assessing Credibility of the Accuser and Accused in Sexual Assault** *(Laura Hamilton; Natasha Quadlin; Brian Powell)*	Sociology	11	1	2,005
32. **The Partisan Gender Gap: Genuine Attachment or Social Motivation?** *(Yanna Krupnikov)*	Political Science, Communication	2	1	2,005
33. **What Triggers Public Opposition to Immigration? Anxiety, Group Cues, and Immigration Threat** *(Ted Brader; Nicholas A. Valentino; Elizabeth Suhay)*	Political Science	2	1	1,991
34. **Politics and purity: The effect of elite partisan cues on pathogen perceptions** *(Daniel Relihan)*	Psychology, Political Science, Public Health, Communication	3	3	1,947

(continued)

Study	Field	N conditions	N outcomes	N subjects
35. **Human Rights Shaming, Compliance, and Nationalist Backlash** *(Rochelle Terman)*	Law, Political Science	5	3	1,912
36. **Equivalency Framing of Societal Problems and Policy Solutions** *(Laura Stoker; Amy E. Lerman; Alexander Sahn)*	Political Science, Public Opinion, Psychology	8	3	1,826
37. **Public Attitudes about Political Equality** *(Patrick Flavin)*	Political Science	2	1	1,785
38. **Terrorist Threat: Overreactions, Underreactions, and Realistic Reaction** *(Suzanne C. Thompson)*	Psychology, Threat, Attitudes toward Terrorism	2	1	1,785
39. **Why Hillary Rodham Became Hillary Clinton: Consequences of Non-Traditional Last Name Choice in Marriage** *(Emily Shafer)*	Sociology	2	1	1,716
40. **To Do, to Have, or to Share? Valuing Experiences and Material Possessions by Involving Others** *(Peter Caprariello)*	Psychology	2	1	1,715
41. **Testing a Theory of Hybrid Femininity** *(Julia Melin)*	Sociology	6	2	1,682
42. **Social Class, College Debt, and the Purpose of College** *(Emma Cohen)*	Sociology	12	3	1,610
43. **Do Messages That Encourage the Use of They/Them Pronouns Influence Public Opinion?** *(Toby Bolsen)*	Political Science	4	3	1,601
44. **Informing the Public or Information Overload? The influence of school accountability data format and on public satisfaction** *(Rebecca Jacobsen)*	Other Discipline, Social Policy, Political Science	2	1	1,547
45. **Beliefs about Racial Discrimination** *(Ingar Haaland; Christopher Roth)*	Economics	2	3	1,542
46. **"The Taxpayer Gap": Perceptions of the Taxpaying Population and Opposition to Welfare Spending** *(Vanessa Williamson)*	Political Science, Social Policy	3	3	1,527

(continued)

Study	Field	N conditions	N outcomes	N subjects
47. **On the Precipice of a "Majority-Minority" America: Perceived Status Threat From the Racial Demographic Shift Affects White Americans' Political Ideology** *(Maureen A. Craig; Jennifer A. Richeson)*	Social Psychology, Political Psychology	2	1	1,467
48. **Issue (Dis)agreement and Intergroup Bias in Affective Polarization** *(Lori Bougher)*	Political Science	3	3	1,447
49. **With God on Our Side** *(Benjamin A. Converse; Nicholas Epley)*	Social Policy, Psychology	2	1	1,407
50. **Can Factual Misperceptions be Corrected? An Experiment on American Public Fears of Terrorism** *(Daniel Silverman; Daniel Kent; Christopher Gelpi)*	Political Science, Terrorism, Foreign Policy, Belief Correction	5	3	1,340
51. **Terrorism Suspect Religious Identity and Public Support for Controversial Detention and Interrogation Practices** *(James Piazza)*	Political Science	2	1	1,321
52. **Understanding the Expense Prediction Bias** *(Chuck Howard; David Hardisty; Abigail Sussman; Melissa Knoll)*	Financial Judgement and Decision Making, Consumer Behaviour, Marketing, Psychology	3	2	1,288
53. **Introducing a Novel Framework for Understanding The Relationships Between Busyness, Idleness, and Happiness** *(Jacqueline Rifkin; Keisha Cutright)*	Psychology	2	1	1,200
54. **Environmental Values, Beliefs, and Behavior** *(Mohana Turaga)*	Social Policy	2	1	1,193
55. **Do Women Make More Credible Threats? Gender Stereotypes and Crisis Bargaining** *(Christopher W. Blair; Joshua A. Schwartz)*	Political Science, Psychology, Women and Gender Studies, Sociology	12	1	1,149

(continued)

Study	Field	N conditions	N outcomes	N subjects
56. **Understanding Effects of Communicating Different Types of Conflicting Information About Nutrition and Cancer Risk** *(Irina Iles; Arielle S. Gilman; Lauren O'Connor; Rebecca Ferrer; William Klein)*	Psychology, Communication	8	3	1,027
57. **The Effects of Exposure to Women Candidates on Political Attitudes** *(David Campbell; Christina Wolbrecht)*	Political Science	4	3	1,000
58. **The Flexible Correction Model and Party Labels** *(Daniel Bergan)*	Political Science, Psychology, Communication	2	1	997
59. **Does Misery Love Company?: Exploration of a Strategic Intervention to Improve Well-being** *(Kate Farrow)*	Psychology, Behavioral Economics	2	1	995
60. **The Reputational Consequences of International Law and Compliance** *(Geoffrey P. R. Wallace)*	Political Science	2	1	934
61. **Exploring the Role of Constitutional Considerations, Public Support and Personal Preferences in Citizen Assessments of Proposed Congressional Action on Immigration and Gun Control** *(Eileen Braman)*	Political Science, Law, Psychology	12	3	801
62. **Does Transparency Affect Survey Research?** *(Yanna Krupnikov)*	Political Science	3	3	767
63. **Gender and the Disparate Payoffs of Overwork** *(Christin Munsch; Lindsey Trimble O'Connor)*	Sociology	12	1	766
64. **Is Modern Racism Caused by Anti-Black Affect? An Experimental Investigation of the Attitudes Measured by Modern Racism Scales** *(Ryan Enos)*	Political Science	3	1	733
65. **The Emotional Substrates of Support for Authoritarian Populists** *(Carly Wayne; Nicholas A. Valentino; Marzia Oceno)*	Political Science	3	3	716
66. **Patient Responses to Medical Error Disclosure: Does Compensation Matter?** *(Michelle Mello)*	Law, Public Health	2	1	648

(continued)

Study	Field	N conditions	N outcomes	N subjects
67. **The Mechanisms of Labor Market Discrimination: How Sex, Gender Role, and Sexual Orientation Treatments Impact Evaluations of Black Male Job Applicants** *(David Pedulla)*	Sociology	2	1	627
68. **Emotional Reactions to Terrorism and Candidate Evaluations** *(Jennifer Merolla)*	Political Science, terrorism, emotions	3	3	611
69. **Examining the 'Raced' Fatherhood Premium: Workplace Evaluations of Men by Race, Fatherhood Status, and Level of Involvement** *(Kathleen Denny)*	Sociology	2	1	597
70. **Which Economy? Class and Partisan Biases in the Acquisition of Economic Information** *(Ian Anson)*	Political Science	2	2	391

5 List of Experiments in the Secondary Archive of Megastudies

Below is a list of all experiments included in the secondary archive. The table reports accuracy for GPT-4 predictions and expert forecasts for each megastudy.

Table 4: Full list of megastudies.

Study	Treatments	Outcome	N treatments	N effects	N subjects	GPT-4 correlation	Expert correlation
Dellavigna & Linos 2022 *Various outcomes*	Text (mail)	Field	14	14	960,472	0.31	0.40
Milkman et al., 2022 *Flu vaccination*	Text (SMS)	Field	22	22	662,170	0.52	-0.15
Tappin et al., 2023 *UBI & immigration policy support*	Video (in-survey)	Survey	59	59	62,738	0.53	-
Milkman et al., 2021 *Gym attendance*	Interactive (4-week program)	Field	53	53	57,790	0.15	-0.22
Voelkel et al., 2024 *Democracy-related attitudes*	Mixed text / video / interactive (in-survey)	Survey	25	50	25,121	0.35	0.62
Zickfeld et al. 2024 *Tax compliance*	Text (in-survey)	Survey	21	21	20,553	0.42	0.20
Dellavigna & Pope 2018 *Typing task effort*	Text (in-survey)	Survey	15	15	9,321	0.70	0.78
Allen et al. 2024 *COVID-19 attitudes*	Text (in-survey)	Survey	90	90	9,228	0.96	-
Vlasceanu et al. 2024 *Climate-related attitudes*	Text & images (in-survey)	Survey	11	22	6,735	0.62	0.61
Voelkel et al. 2025 *Climate-related attitudes*	Text & images (in-survey)	Survey	10	40	10,638	0.92	-
Mason et al., unpublished *Intention to vote*	Text & images (in-survey)	Survey	10	10	11,733	0.26	-
Goldwert et al., unpublished *Climate-related attitudes*	Text & images (in-survey)	Survey	17	17	27,343	0.36	-
Broockman et al. 2024 *Political attitudes*	Text & images (in-survey)	Survey	172	172	61,869	0.18	0.35
Milkman et al., 2024 *COVID-19 vaccination*	Text (SMS)	Field	7	7	2,677,402	0.20	0.18
Saccardo et al., 2024 *COVID-19 vaccination*	Text (SMS)	Field	14	14	259,249	0.79	-

6 Email to AI companies

Given the risk that LLM users could misuse the capacity of LLMs to predict experimental effect sizes, we sought to follow responsible disclosure practices, notifying the relevant departments of the companies running the two most powerful, publicly-available LLMs. The following email was sent to relevant departments at both OpenAI and Anthropic notifying them of our findings several weeks before we public posted this paper. Additionally, the web tool that we built for academic research purposes includes guardrailing against misuse like the one we identify here.

[Email text]

Dear XXX,

We are reaching out to notify you that we have identified a significant misuse risk in the current generation of LLMs, including OpenAI's GPT-4. We find that using our prompting approach, GPT-4 can be used to aid in the selection of harmful content such as misinformation/propaganda, to maximize its influence on public opinion. We expect the same approach could be implemented on other current generation LLMs, such as Claude3 and Gemini Ultra.

As context, we have been working on a project leveraging LLMs to accurately predict the results of experiments with human participants in the social sciences. We are also assessing related ethical concerns and risks including the possibility that LLMs can be leveraged to identify what messages would be most harmful or manipulative. We find that GPT-4 can be used to accurately estimate the effectiveness of anti-vaccine messages; e.g., we estimate a .96 correlation between (a) GPT-4-derived predictions of the effects anti-COVID vaccine messages on vaccine intentions, and (b) estimated effects of the same messages on Americans' COVID vaccine intentions in recent experimental research (Allen et al, 2023). As a result, the anti-vaccine posts that GPT-4 selects as most persuasive are found in experiments to reduce Americans' vaccine intentions by an average of 2.77pp (a large impact in the context of the relevant literature). Indeed, GPT-4-selected posts were found to be more persuasive than the posts which humans themselves predicted would have the largest negative impact.

Our findings provide clear evidence that existing guardrails that prevent the generation of harmful messages are insufficient for preventing GPT-4's use in harmful persuasion campaigns. While GPT-4 refuses direct requests to generate harmful persuasive messages, such as anti-vaccination misinformation, the impact of these refusals is limited if bad actors can still use GPT-4 to accurately select the most persuasive among a large set of possible alternatives (which are easily generated through other means).

We believe it is likely that the same technique could be applied to identify more persuasive political misinformation, misleading health information, etc, and that the results we obtain here using GPT-4 would be similar for other current generation LLMs, such as Claude3 and Gemini Ultra. Additionally, we find that this capacity of LLMs has increased across generations (see Figure 4 in the draft paper, also pasted below), suggesting the potential for harmful misuse could grow as future generations of LLMs become more powerful.

Here is the draft paper [link] (see section titled "Assessing risks of harmful use") providing more information about our approach and findings. Please do not share the paper as it is not public at this time.

We are motivated to provide whatever resources we can to support efforts to address this identified risk, including sharing our code, our approach to prompting GPT-4 to simulate human experimental participants, and/or scheduling a meeting to provide other information.

Note that the paper has not yet been posted publicly or submitted to a journal. We plan to post our paper as a public working paper in two weeks. However, we are happy to discuss modifying this timeline if it would be helpful for addressing this potential misuse of your platform.

Again, we take this potential misuse very seriously, and we would be happy to support your efforts to address it however we can. Thank you for your time.

7 Prompt format

Full text used in LLM prompts

Prompts were constructed by randomly selecting one introductory sentence, one profile introductory label, one profile description format, and one survey information format. See the below table for the full text used.

Introductory sentence	Profile introductory label	Profile description format (with example content)	Survey format
"(For the following, keep in mind that the public has diverse attitudes and behaviors; people often provide very different answers to the same question.)" "(For the following, keep in mind that people's beliefs and behaviors are malleable; their answers to questions often change based on the framing or information provided.)" "(As you complete the following task, keep in mind what is known about the public's beliefs and behaviors from research in the field of behavioral science)" "(Keep in mind what is known about the public's beliefs and behaviors from research in the field of social psychology)" "(We are interested in predicting how people respond to questions and interventions in social science research studies.)" "(We are interested in predicting the treatment effect of different messages on research participants' attitudes, beliefs, and behaviors.)" "You will be asked to predict how people respond to various messages" "Can reading a message affect people's attitudes and actions?" "Try to be accurate when you make predictions about how different messages affect how people think and behave" "As you complete the following task, take the perspective of a leading expert in social science."	"You are an American." "You are a research participant." "Participant X is an American." "Participant X is a research participant."	*Biographical-style:* "[You are/Participant X is] a liberal, 18-29 year-old, black, female American with college education who identifies as a Democrat." *Survey-style:* "What is your ethnicity?" "[You choose/Participant X chooses] 'White'" …	"Below is a transcript from an online survey. "The first page of the survey says: [introductory message from original study]. "The next page of the survey says: [experimental stimulus text]. "The next page of the survey says: [outcome text]. "Please choose a number on a scale from [outcome scale]. "You choose:"

The prompt also included 1-2 sentences describing the setting in which each experiment was conducted. The table below shows the lines used for (a) our primary test archive, and (b) for each experiment in the archive of megastudies.

Study	Study description
Primary test archive	"Social scientists often conduct research studies using online surveys. The text below is from one such survey conducted on a large, diverse population of research participants."
Archive of mega studies	
Allen et al., 2023	"Allen et al. conducted an RCT experiment aiming to quantify the impact of misinformation and vaccine-skeptical content on Facebook. Study participants were recruited from an online platform and were randomly assigned to receive one of several interventions, administered as an online survey."
Dellavigna & Pope, 2018	"In 2015, researchers conducted an online study to compare how different monetary and non-monetary motivators induce costly effort. Study participants were recruited from an online platform and were randomly assigned to one of 18 treatment arms."
Vlasceanu et al., 2023	"The International Collaboration to Understand Climate Action is an international randomized control trial testing the effectiveness of the state-of-the-art climate action interventions aimed at stimulating collective climate action, using representative samples collected from around the world. Study participants were recruited from an online platform and were randomly assigned to receive one of several interventions, administered as an online survey."
Tappin et al., 2023	"Tappin et al. conducted an RCT experiment aiming to quantify the impact of arguments for and against various political issues. Study participants were recruited from an online platform and were randomly assigned to receive one of several messages, administered as an online survey."
Zickfeld et al., 2024	"In 2023, a team of researchers ran an RCT test comparing 21 honesty oaths. Study participants were recruited from an online survey platform to complete an odd/even number sorting task, and were awarded a bonus based on their performance in the task. Before or after the sorting task, participants were asked to commit to an honesty oath."
Voelkel et al., 2023	"The Strengthening Democracy Challenge is a joint project between academics and practitioners to crowdsource and identify short, scalable interventions to reduce anti-democratic attitudes, support for partisan violence, and partisan animosity among Americans. Study participants were recruited from an online platform and were randomly assigned to receive one of several interventions, administered as an online survey."
Milkman et al., 2022	"In the fall of 2020, a team of researchers tested twenty-two different sets of text messages against a no-text-message control group. Each set of text messages tried to nudge higher flu vaccination rates among customers during that year's flu season (2020-2021). Notably, only customers who had received a flu vaccine from Walmart the previous year, during the 2019-2020 flu season, were included in the study.

	For this research, on September 25, 2020, customers received text messages from Walmart encouraging them to get a flu shot; the message content and timing were varied experimentally."
Dellavigna & Linos, 2022	"Since 2015, two large behavioral science teams in the United States, one based at the federal level, and the other focusing on local government, have used behavioral nudges with the goal of increasing take-up and compliance in various domains. One such nudge study tested interventions aiming to [goal of nudge study]. Specifically, the study assessed the likelihood that participants " [outcome of nudge study]."
Goldwert et al., unpublished	"In 2024, a team of researchers ran an RCT test comparing 16 interventions aimed at stimulating collective climate action. Study participants were recruited from an online survey platform and were randomly assigned to receive one of the interventions, administered as an online survey."
Mason et al., unpublished	"In 2024, a team of rearchers ran an RCT test comparing 10 interventions designed to increase voter turnout. Study participants were recruited from an online survey platform and were randomly assigned to receive one of the interventions, administered as an online survey."
Saccardo et al., 2024	"In 2023 a team of researchers conducted a megastudy focused on encouraging adults' adoption of the bivalent COVID-19 booster vaccine. As as of 1 August 2023, only 20.5% of adults had received the bivalent booster. The researchers conducted an RCT in partnership with the University of California Los Angeles (UCLA) Health, a large healthcare system in California, aiming to encourage patients who had previously completed the primary COVID-19 vaccine series to receive a bivalent booster dose. They delivered their interventions through text-based reminders."
Broockman et al., 2024	"Broockman et al. conducted an RCT experiment aiming to quantify the impact of arguments for and against various political issues. Study participants were recruited from an online platform and were randomly assigned to receive one of several messages, administered as an online survey."
Milkman et al., 2024	"In the Fall of 2022 A team of researchers conducted a megastudy focused on encouraging adults' adoption of the bivalent COVID-19 booster vaccine. As of mid-November, 2022, only 11% of Americans had received this recommended bivalent booster vaccine. The researchers partnered with CVS Pharmacy to test eight interventions targeting CVS Pharmacy patients who 1) were 18 or older, 2) resided in one of 65 U.S. metropolitan areas selected for study inclusion 3) had previously received at least their primary COVID-19 vaccine series but not the bivalent booster according to CVS Pharmacy records, and 4) had consented in writing to receive text messages from CVS."
Milkman et al., 2021	"Recently, members of 24 Hour Fitness, a national gym chain, were invited to join a digital, 28-day "habit-building, science-based workout program" called StepUp. There were several different versions of the StepUp Program, which was free and offered rewards. The experimental study examined how StepUp influenced gym members' exercise habits."

8 Survey of social scientists: Questions

Below is a list of all questions included in the survey of social scientists. archive.

To help us understand who is participating in the study, please tell us a bit about your background.
Which department(s) and/or discipline(s) are you currently affiliated with? (Check all that apply):

- [ ] Psychology
- [ ] Sociology
- [ ] Political Science
- [ ] Economics
- [ ] Public Health
- [ ] Other (please specify) _____

Which subfield of Psychology are you primarily affiliated with?

- ○ Social psychology
- ○ Cognitive psychology
- ○ Developmental psychology
- ○ Clinical psychology
- ○ Personality psychology
- ○ Other (please specify) _____

Which of the following best describes your current position?

- ○ Pre-doctoral student/researcher
- ○ Graduate student
- ○ Postdoctoral fellow or researcher
- ○ Assistant Professor, or equivalent
- ○ Associate Professor, or equivalent
- ○ Professor, or equivalent
- ○ Retired/emeritus professor
- ○ Researcher in industry or non-academic position
- ○ Other (please specify) _____

Year of PhD / year PhD is expected: ("N/A" if PhD not expected)_____

Have you ever (on your own, or with others) conducted a survey experiment with human subjects for your research? *(Note: a "survey experiment" is a study in which an independent variable is systematically varied and a dependent variable is measured in a survey)*

- ○ Yes
- ○ No

How much experience do you have with each of the following research methods?

	No experience at all 1	2	3	4	5	6	A great deal of experience 7
Experiments with survey outcomes	○	○	○	○	○	○	○
Experiments with behavioral outcomes	○	○	○	○	○	○	○
Computational research (e.g., using machine learning or AI models)	○	○	○	○	○	○	○

How often do you use large language models (LLMs) such as ChatGPT...

	Never	1 time per month or less	2-3 times per month	1 time per week	2-3 times per week	1 time per day	2+ times per day
For **research-related** purposes (e.g., coding text data, literature reviews, data analysis, writing academic text, etc.)	○	○	○	○	○	○	○
For **non-research** purposes	○	○	○	○	○	○	○

Do you think that the **impact of artificial intelligence (AI) on the world** over the next 20 years will be…

- ○ Very negative
- ○ Somewhat negative
- ○ Equally positive and negative
- ○ Somewhat positive
- ○ Very positive

Recently, many researchers have been studying whether Artificial Intelligence (AI) models can be used to **predict the results of human experiments**. On the next page, you will be asked how likely you would be to use such models for a variety of applications.

Below are ways social scientists might use AI models that can predict the results of human experiments. For each of the potential applications, assume that an **AI model has been developed and rigorously validated for the specific application**, and found to be **quite accurate** (i.e., at least as accurate as predictions made by academic experts). How **likely would you be** to use this AI model for each of the applications listed below?

Using AI predictions of effect size to run a statistical power analysis for an upcoming study

I definitely would not use0I am unsure if I would use50I definitely would use100	

Using AI predictions to identify promising intervention(s) to test in an experiment (such as messages promoting vaccination)

I definitely would not use0I am unsure if I would use50I definitely would use100	

Estimating the generalizability of results by using AI to predict results for alternate versions of an experiment (also known as “stimulus sampling”)

I definitely would not use0I am unsure if I would use50I definitely would use100	

Minimizing/avoiding harm to human participants by using AI to predict the results of potentially harmful studies (such as studies involving exposure to misinformation)

I definitely would not use0I am unsure if I would use50I definitely would use100	

Using AI predictions of published effects to identify potentially less robust effects that need to be replicated

I definitely would not use0I am unsure if I would use50I definitely would use100	

Combining effect sizes derived from AI models and human samples to achieve more accurate estimates of the true effect size

I definitely would not use0I am unsure if I would use50I definitely would use100	

Running AI-based pilot studies to gain insight on likely results, or to improve the design of experimental stimuli and measures.

I definitely would not use0I am unsure if I would use50I definitely would use100	

Using an AI model to predict experimental effect sizes for hard-to-reach samples (e.g., minority groups)

I definitely would not use0I am unsure if I would use50I definitely would use100	

Using an AI model to identify potential “mediators” of experimental effects (i.e., dependent variables that are found to explain an experimental effect)

I definitely would not use0I am unsure if I would use50I definitely would use100	

Using an AI model to identify potential demographic or individual-level “moderators” of

experimental effects (e.g., demographic subgroups among whom experimental effects may be larger or smaller)

I definitely would not use0I am unsure if I would use50I definitely would use100	

Are there other applications of AI models that can predict the results of social science experiments that you will be likely to use in the future?

_____

_____

_____

_____

_____

On the whole, how likely would you be to use **an AI model that can predict the results of experiments with human subjects, with relatively good accuracy**, for any purpose?

I definitely would not use0I am unsure if I would use50I definitely would use100	

On the next page, you will be asked about **concerns you may have** about potential uses of AI models that can predict the results of social science experiments.

Considering the use of AI models that can predict the results of social science experiments, to what extent would you be **concerned about** each of the following:

AI-based predictions replacing data from human samples

Not at all concerned0Moderately concerned50Extremely concerned100	

AI predictions being less accurate for under-represented groups

Not at all concerned0Moderately concerned50Extremely concerned100	

AI predictions for certain groups being systematically biased or aligned with stereotypes

Not at all concerned0Moderately concerned50Extremely concerned100	

Threats to interpretability of AI predictions because of use of “black box” AI models

Not at all concerned0Moderately concerned50Extremely concerned100	

AI-based predictions not being reproducible because they used proprietary AI models that may be changed in unknown or non-transparent ways

Not at all concerned0Moderately concerned50Extremely concerned100	

Scientists focusing on research areas that are most predictable using AI, instead of other important research

Not at all concerned0Moderately concerned50Extremely concerned100	

"False negative" results in AI-based pilot studies leading researchers to not pursue what are actually promising ideas

Not at all concerned0Moderately concerned50Extremely concerned100	

AI models failing to reflect the variability of human responses

Not at all concerned0Moderately concerned50Extremely concerned100	

AI-based predictions of human experiments being misused for exploitative or harmful purposes

Not at all concerned0Moderately concerned50Extremely concerned100	

Please describe any other concerns you may have, or risks you anticipate, regarding the possible application of AI models that can predict the results of social science experiments

_____

_____

_____

_____

_____

Thank you! Questions below are optional and may be skipped if you choose. If you would like to be notified when the study is published, you may choose to sign up on the following page.

What is your gender?

- ○ Male
- ○ Female
- ○ Other _____

Please select which race / ethnicity you most identify as.

- ○ White / Caucasian
- ○ Black / African-American
- ○ Latino / Hispanic
- ○ Asian / Asian-American
- ○ Other

In general, do you consider yourself liberal or conservative? Please answer on the following scale.

- ○ ExtremelyLiberal1
- ○ 2
- ○ 3
- ○ 4
- ○ 5
- ○ 6
- ○ Extremely Conservative7

Is there anything else you'd like to share?

_____

_____

_____

_____

_____

References

[1] 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).

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

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

[4] Journal n-pact factors from 2011 to 2019: evaluating the quality of social/personality journals with respect to sample size and statistical power, R. C. Fraley, *et al.*, *Advances in Methods and Practices in Psychological Science* **5**, (2022).

[5] Rank-heterogeneous effects of political messages: Evidence from randomized survey experiments testing 59 video treatments, L. Hewitt, B. M. Tappin (2022).

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

[7] Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity, J. G. Voelkel, *et al.*, *Science* **386**, (2024).

[8] Political practitioners poorly predict which messages persuade the public, D. E. Broockman, J. L. Kalla, C. Caballero, M. Easton, *Proceedings of the National Academy of Sciences* **121**, (2024).

[9] Package 'metafor', W. Viechtbauer, M. W. Viechtbauer, *The comprehensive R Archive network. Package 'metafor* (2015).