Finetuning LLMs for Human Behavior Prediction in Social Science Experiments

Akaaash Kolluri[1*], **Shengguang Wu**[1*], **Joon Sung Park**[1], **Michael S. Bernstein**[1]

[1]Stanford University
{akaash,shgwu}@stanford.edu

Abstract

Large language models (LLMs) offer a powerful opportunity to simulate the results of social science experiments. In this work, we demonstrate that finetuning LLMs directly on individual-level responses from past experiments meaningfully improves the accuracy of such simulations across diverse social science domains. We construct SOCSCI210 via an automatic pipeline, a dataset comprising 2.9 million responses from 400,491 participants in 210 open-source social science experiments. Through finetuning, we achieve multiple levels of generalization. In completely unseen studies, our strongest model, SOCRATES-QWEN-14B, produces predictions that are 26% more aligned with distributions of human responses to diverse outcome questions under varying conditions relative to its base model (Qwen2.5-14B), outperforming GPT-4o by 13%. By finetuning on a subset of conditions in a study, generalization to new unseen conditions is particularly robust, improving by 71%. Since SOCSCI210 contains rich demographic information, we reduce demographic parity difference, a measure of bias, by 10.6% through finetuning. Because social sciences routinely generate rich, topic-specific datasets, our findings indicate that finetuning on such data could enable more accurate simulations for experimental hypothesis screening. We release our data, models and finetuning code at stanfordhci.github.io/socrates.

1 Introduction

Large language models have shown impressive potential to simulate human behavior (Park et al., 2024; Hewitt et al., 2024; Kim and Lee, 2023). For social science experiments, simulations enable researchers to screen and iterate on hypotheses before committing to costly studies (Rothschild et al., 2024; Hewitt et al., 2024; Wang et al., 2025b). Accordingly, LLM-based simulation methods have been explored across various social science disciplines (Argyle et al., 2023; Horton, 2023; Brand et al., 2023).

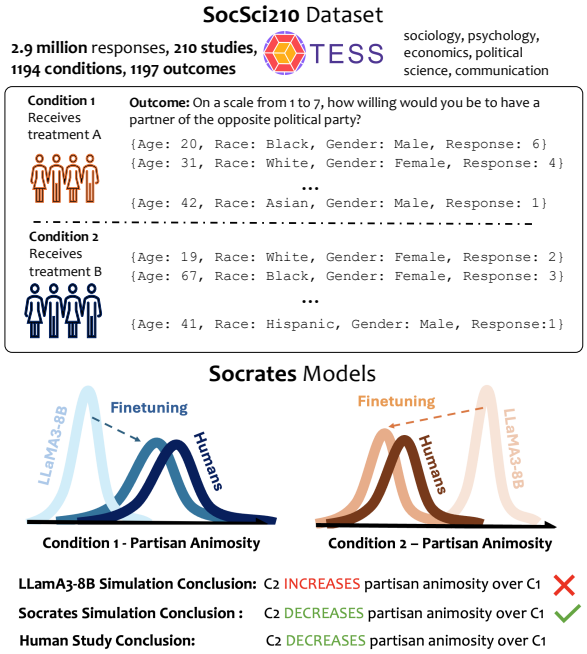

Figure 1: We release SOCSCI210, a large-scale dataset built from open-source social science experiments. Through finetuning, we create behavioral prediction models SOCRATES-LLAMA-8B and SOCRATES-QWEN-14B, which predict responses that are 12.1% and 13.2% respectively more aligned with human response distributions to outcomes under diverse experimental conditions, relative to GPT-4o.

Previous work simulating human responses has used direct prompting such as with demographic personas (Hewitt et al., 2024), human conversations (Cho et al., 2024), and detailed life narratives (Park et al., 2024; Moon et al., 2024). Still, LLMs routinely distort opinion distributions (Bisbee et al., 2024; Gao et al., 2024), overestimate effect sizes in experimental manipulations by **2 to 10** times (Park et al., 2024; Hewitt et al., 2024), and incorrectly predict significant effect directions **10 - 32% of**

* Denotes equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30085–30100
November 4-9, 2025

the time (Hewitt et al., 2024; Bisbee et al., 2024). LLMs further introduce biases that flatten variation across demographic groups (Wang et al., 2025a). These error cases currently limit the viability of effective LLM-simulations for social science experiments.

Recent work has begun exploring the viability of finetuning language models for improved human response prediction (Suh et al., 2025; Chu et al., 2023; Lu et al., 2025; Binz et al., 2024) and has demonstrated generalization on their specific tasks (*e.g.*, cognitive science, public opinion).

In this work, we broaden the domain and scope of prior fine-tuning work in pursuit of a general purpose, domain-agnostic human behavior prediction model. To enable this, we first construct SocSci210, a standardized, large-scale dataset comprising **2.9 million** individual responses from over **400,000** participants across **210** social science studies. All studies were drawn from NSF's Time-sharing Experiments for the Social Sciences—peer-reviewed, high-powered experiments spanning multiple disciplines (*e.g.*, economics, political science, behavioral psychology) and conducted on nationally representative samples with rich demographic reporting ("TESS", 2025). We design an LLM agent to convert each study's data into a consistent text-based representation describing respondent demographic profiles, the experimental questions, and the recorded responses.

Using this dataset, we provide a comprehensive comparison of finetuning methods (supervised fine-tuning, augmenting with reasoning traces, contrastive preference optimization) against various prompting baselines (reasoning and in-context learning) on both proprietary `GPT-4o` and open-source `LLaMa3-8B` (Grattafiori et al., 2024), `Qwen2.5-14B` (Yang et al., 2024a) LLMs. Through our evaluations, we highlight that supervised fine-tuning greatly improves distributional alignment between predicted responses and human responses, while contrastive preference optimization leads to the best prediction accuracy for individual responses. Notably, relative to their base models, fine-tuning improves alignment with human response distributions in unseen studies by 30% for `LLaMa3-8B` and 26% for `Qwen2.5-14B`. We further demonstrate robust generalization to unseen participants, conditions and outcomes. Finally, we highlight finetuning improve demographic bias in predictions by 10+%.

Our main contributions are:

1. We release SocSci210, a standardized, large-scale dataset comprising 2.9 million individual responses from over 400,000 participants (5× the number participants of prior work's datasets) with rich demographic reporting across 210 social-science studies spanning multiple disciplines.

2. We present Socrates-LLaMA-8B and Socrates-Qwen-14B finetuned on SocSci210, which, relative to `GPT-4o`, generate predictions that align 12.1% and 13.2% better to human response distributions, reflecting 26+% performances gains relative to their base models.

3. Motivated by practical use cases of social scientists with in-domain data, we demonstrate robust generalization at various levels. Finetuning on as little as 10% of an experiment's data reduces prediction error by 13% on unseen participants, and training on subsets of experimental conditions or outcomes boosts generalization to unseen conditions by 71% and unseen outcomes by 49%.

2 Related Work

Datasets for Human Response Finetuning. Recent works have assembled large-scale public-opinion datasets and used them to finetune LLMs. For instance, Santurkar et al. (2023) train on opinion distributions from 60 U.S. demographic groups over 500 contentious questions. Likewise, Suh et al. (2025) compile 3,362 survey questions with the responses distributions from 70,000 demographics, demonstrating a great breadth of topic diversity.

Other works have focused on datasets of individual decision-making across different behavioral contexts. Binz et al. (2024), for example, introduce Psych-101, which contains over 10 million choices from 60,000 participants across 160 cognitive-science experiments. Orlikowski et al. (2025) collect and finetune on 60,000 individuals reactions to different texts to explore how sociodemographic factors shape perception, and Lu et al. (2025) finetune on 230,965 logged decisions from 3,526 users to predict web-action generation. Zhu et al. (2025) explores reinforcement fine-tuning for reasoning trace generation from a dataset of 13,000 risky human choices produced by Peterson et al. (2021). Xie et al. (2025) combines multiple types of datasets for fine-tuning, aggregating data from 17,667 individual surveys and economic games played by 68,790 individuals. (they also include titles and abstracts from 2,703 behavioral science publications). Although rich in behavioral

Dataset	Size			Feats.		Domain
	Source	Individuals	Total Data Points	D	I	
Psych-101 (Binz et al., 2024)	160 experiments	60,000	10,000,000	✗	✓	Psychology
SubPOP (Orlikowski et al., 2025)	3,362 questions	—	70,000	✓	✗	Public Opinion
E-commerce (Lu et al., 2025)	31,865 sessions	3,526	230,965	✗	✓	E-commerce
OpinionQA (Santurkar et al., 2023)	∼1,500 questions	—	90,000	✓	✗	Public Opinion
	50 questions	17,667	883,350	✓	✓	Big-5 Personality
Be.FM (Xie et al., 2025)	6 games	68,790	82,057	✓	✓	Behavioral Econ.
	2,703 abstracts	—	2,703	✗	✗	Behavioral Science
SocSci210 (Ours)	210 experiments	**400,491**	2,900,000	✓	✓	Social Sciences

Table 1: Comparison of our SocSci210 to dataset characteristics used in prior finetuning work. Under "Feats (features)" column, "D" indicates if the dataset includes participant demographics, and "I" refers to if training samples are done at the individual level (as opposed to the aggregate distribution level). When training occurs at the aggregate level, we omit the total number of individuals used to construct the dataset as they do not maintain individual granularity.

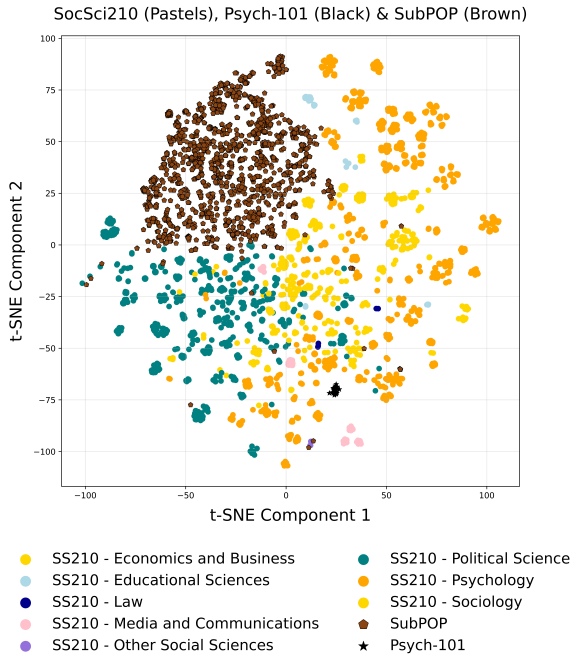

Figure 2: t-SNE projected embedding space of questions in SocSci210, compared to SubPop (Suh et al., 2025) and Psych101 (Binz et al., 2024). SocSci210 shows much broader topic diversity across social science disciplines.

detail, these datasets do not yet capture the full diversity of social science disciplines that simulations could enable at a granular, individual level.

Our work attempts to bridge these gaps by constructing a dataset that expands upon the coverage of human behavioral sciences covered in prior datasets while enabling granular, individual-level responses in behavioral contexts. Increased scientific topic breadth is a key enabler of engineering a shared agent model that can enable any simulation across any social science context.

LLM Finetuning Methods. Finetuning adapts pretrained LLMs for specific tasks, such as following user instructions (Ouyang et al., 2022; Wang et al., 2022; Zhang et al., 2023) or learning social skills (Liu et al., 2023; Yang et al., 2024b; Wu et al., 2024). Apart from supervised finetuning (SFT), reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2019), as well as the simplified DPO (Rafailov et al., 2024) and SimPO (Meng et al., 2024), all use paired data of `{preferred response, dispreferred response}` for finetuning, which aligns model outputs to annotated human preferences. More recently, reasoning models (DeepSeek-AI, 2025; Abdin et al., 2025; Team, 2025b) trained with RL show strong performance in improving LLM capability by exploring chain-of-thought (Wei et al., 2022a), especially for solving complex tasks. The reasoning abilities of larger teacher models can also be distilled into smaller models via finetuning on the teacher-generated reasoning traces (Zhao et al., 2025; Team, 2025a).

In this work, we provide a comparative evaluation of SFT and DPO—contrasting the scenarios in which each is most effective (see §4). Prior finetuning studies in this domain have explored direct SFT (Suh et al., 2025; Binz et al., 2024), SFT augmented with reasoning (Lu et al., 2025), and reinforcement fine-tuning with GRPO (Zhu et al., 2025).

3 Task Formulation

3.1 Task Description

We finetune an LLM to predict the responses of an individual (of some demographic) to a stimulus in an experiment with multiple treatments. Fig. 3

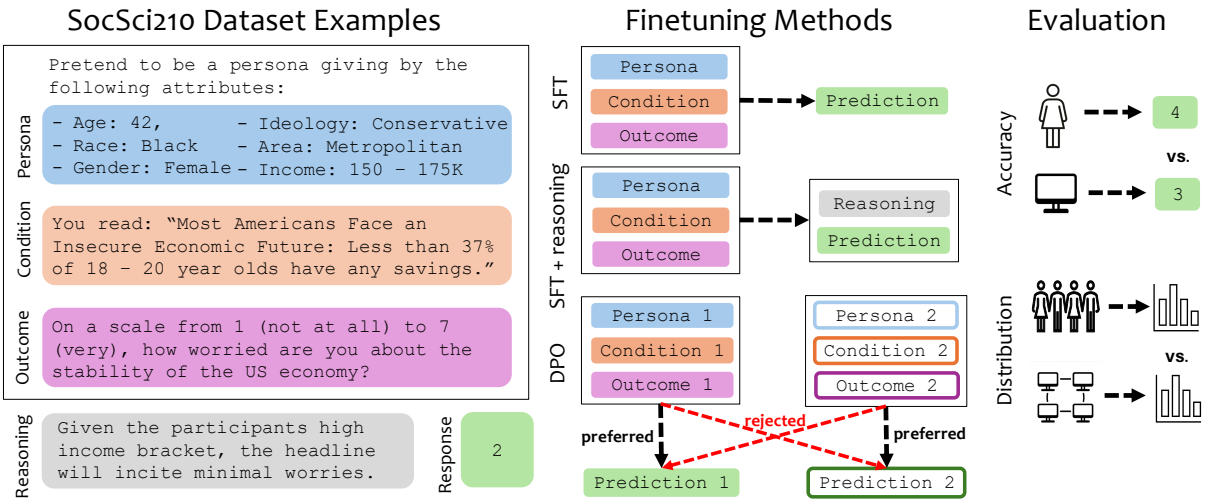

Figure 3: Overview of our task formulation, methods, and evaluation. Our dataset contains information on personas, conditions, outcomes, and predictions. We compare SFT, SFT on reasoning traces, and DPO. Our evaluation measures performance gains on both predicting individual accuracy and aggregate distributions under conditions.

shows an example prediction. We consider questions that are ordinal (*e.g.*, "On a scale from 1 to 7, how satisfied are you with your life?") or binary (*e.g.*, "Would you buy this? Answer yes or no."). In the dataset of experiments we draw from, outcomes were primarily ordinal or binary; restricting to only these response types standardized our model evaluations, with minimal data size reduction.

Formally, in our dataset D, a person is represented as P, a set of unique attributes. In our dataset, these attributes are demographic characteristics. Each experiment is set up as $E = ([c_1, ..., c_j], [o_1, ..o_i])$ with j different conditions and i outcome questions. The goal of the experiment is to see how different conditions impact participant responses to a given outcome. Studies may have a between-subject design (i.e., participants are randomized into a single condition and answer multiple outcome questions (between-subject studies) or they may have an within-subject design (participants are randomized into multiple conditions and answer the same outcome). In either setup, each participant gives k responses, r, each corresponding to a question representing a (c, o) pair, which we refer to as the stimuli. Note, the randomization process occurred when the study of the original dataset was conducted, so we do not randomize the participants into the conditions, we simply tag the participants' responses with the condition they were in in the original study. For every individual participant in a study, we have k tuples of (P, c, o, r). We finetune an LLM, F', to learn $F(P, c, o) \implies r$.

3.2 Evaluation

We evaluate the performance of our finetuned model F' at two levels.

Individual Response Accuracy. Given the ordinal property of all predictions (§3.1), we compute a normalized accuracy between predicted and actual responses as

$$\text{Acc.} = 1 - \frac{1}{N} \sum_{(P,c,o)} \frac{\left| F'(P, c, o) - r \right|}{r_{\max} - r_{\min}}$$

where $r_{\max}, r_{\min}$ represent the maximum and minimum value of r under a specific condition and outcome in the ground truth responses (*i.e.*, the bounds of the response scale). This value is computed for each study, and then averaged across all studies to get a final individual response accuracy score.

Distribution Alignment Under Conditions. While predicting individual responses is important, accuracy is upper-bounded by the non-deterministic nature of $F(P, c, o)$ (*i.e.*, the same demographic may have different responses to the same stimuli, so $F(P, c, o)$ is a distribution). The primary goal of social-science experiments is to compute an end statistic that represents how responses to outcome questions vary across stimulus conditions. The exact statistical analyses in these contexts (*e.g.*, t-tests, ANOVA, regressions) depend on the type of experiment being conducted (Maravelakis, 2019), but all statistics depend on the underlying distribution of the responses to outcome questions under each condition. Because of this,

distributional alignment for this context is more important than measuring just accuracy (we provide further examples and justification of this in appx.B). Thus, we also measure whether the distribution of responses to each outcome under each experimental condition aligns with corresponding human responses.

Specifically, for each (condition, outcome) pair $(c, o) \in E$, we compare the distribution of predicted responses across all participants assigned to c against the empirical distribution of actual responses. Similar to Suh et al. (2025), we compute distributional alignment using the Wasserstein distance, which estimates both the shape and mean of distributions. For consistency, we first standardize all distributions to be between $[0, 1]$ by subtracting $r_{\min}$ and dividing by $r_{\max} - r_{\min}$. We average the Wasserstein distance across all condition outcomes pairs in a study to get an aggregate score for the study. We then average each study's score across all studies to get the final score. A lower Wasserstein distance indicates better alignment between predicted and empirical response distributions.

3.3 Dataset Construction

Data Source. We collect studies from NSF's Time-sharing Experiments for the Social Science (TESS) project ("TESS", 2025), a repository of peer-reviewed experiments across various social disciplines (*e.g.*, psychology, political science, and economics). TESS studies are nationally representative and high powered (studies in SocSci210 have mean 1907 and median 1954.5 participants).

Reconstruction. We employ a data-construction agent – powered by OpenAI's `o4-mini-high` – to automatically parse the source data into `{persona, stimuli, response}` formats from each study, where the stimuli is tagged by its respective condition and outcome question. We detail the reconstruction workflow in Appx.A. In total, our agent successfully reconstructs 210 studies.

Final Dataset Statistics. SocSci210 comprises 2.9 million individual responses spanning 1197 outcomes and 1194 conditions (yielding collectively 5,998 unique stimuli) from 400,491 participants. Our SocSci210 includes responses from **five times as many individuals as prior finetuning work**. Tab. 1 offers an explicit comparison with prior datasets used for human behavior finetuning. Apart from being large in scale, SocSci210 is also diverse across disciplines. Fig. 2 shows the embedding space of our stimuli with other large-scale fine-tuning datasets (Binz et al., 2024; Suh et al., 2025), illustrating this broad topic diversity.

4 Finetuning Methods

To finetune LLMs for simulating responses, we experiment with supervised finetuning (SFT), SFT on oracle reasoning traces, and contrastive preference tuning via DPO (Rafailov et al., 2024).

Supervised Finetuning (SFT). Given our dataset $\mathcal{D}$, an individual persona P, experiment condition c, and outcome question o, we form a prompt q that asks the model to predict the individual's response (see Appx.D for the template). Let F' denote the model being finetuned. The SFT objective minimizes the negative log-likelihood (cross-entropy) of the ground-truth response r:

$$\mathcal{L}_{\text{SFT}}(F') = -\,\mathbb{E}_{(q,r)\sim\mathcal{D}}\big[\log F'(r \mid q)\big] \qquad (1)$$

Augmentation with Oracle Reasoning Traces. Prior work has used explanatory reasoning for predicting human behavior (Park et al., 2024) and improving fine-tuning performance (Lu et al., 2025), so we also augment SFT with oracle reasoning. Specifically, given the LLM prompt and the corresponding human response, we query `GPT-4o-mini` to generate reasoning traces explaining the human decision from a social scientist's perspective (see Appx.D). These oracle-generated reasoning traces are incorporated into the target output, enriching the response r with explicit rationales[1].

Contrastive Finetuning via Preference Optimization. To enhance the model's ability to differentiate responses based on variations in conditions c, demographics P, or outcome questions o, we construct paired data by varying these components and contrasting the corresponding responses.

For demographic contrastive pairs (see §5.5), we take a focal persona p_{pos} and randomly sample a contrasting persona p_{neg} from the dataset $\mathcal{D}$ under the same condition c and outcome question o, while ensuring their responses differ (r_{neg}). Each pair specifies that (p_{pos}, c, o, r_{pos}) is preferred over (p_{pos}, c, o, r_{neg}). Following Rafailov et al. (2024), the DPO objective is

[1]While we use `GPT-4o` for our main evaluation, we opted to generate the oracle reasoning traces with `GPT-4o-mini` due to practical cost constraints. Since these traces are produced oracle-style, we believe any difference in reasoning quality between the models is minimal.

$$\mathcal{L}_{\text{DPO}}(F'; F) = - \mathbb{E}_{(q, r_{pos}, r_{neg}) \sim \mathcal{D}} \Big[\log \sigma \Big(\beta \log \frac{F'(r_{pos} \mid q)}{F(r_{pos} \mid q)} - \beta \log \frac{F'(r_{neg} \mid q)}{F(r_{neg} \mid q)} \Big) \Big], \quad (2)$$

where q concatenates p_{pos}, c, and o into a single prompt (see Appx.D); F' is the finetuned model, F the fixed reference model, σ is the sigmoid function, and β scales the preference impact.

5 Experiments

5.1 Training Configurations

We train `LLaMA3-8B-Instruct` (Grattafiori et al., 2024) and `Qwen2.5-14B-Instruct` (Yang et al., 2024a) as representative base LLMs of different sizes. Appx.C has complete training details.

5.2 Baselines

Bounds on Metrics. When computing the Wasserstein distance, we treat the responses in our dataset as the ground-truth distribution of human responses. Because individual responses naturally vary, this empirical sample may not perfectly capture the true distribution of outcomes under each condition. Our source experiments were highly powered to estimate a treatment effect of a certain size, not to robustly estimate the full distribution. To estimate an empirical upper bound on performance given this variance, we perform bootstrapping: we generate 100 resampled datasets (with replacement) from the original responses. For each bootstrapped dataset, we compute the Wasserstein distance between that resample and the full original sample, then average these distances across all 100 iterations. If our models Wasserstein distance meets or exceeds this, our error is no larger than the variability inherent in our data. We label this the "**Empirical Best**".

To establish a lower bound, we compare the standardized distribution of our predicted responses to a uniform distribution across $[0, 1]$. We use this uniform baseline "**Uniform Guess**" to calculate a corresponding lower bound on both Wasserstein distance and accuracy.

Comparing Metrics. Because the Wasserstein distance has a narrow range (e.g., 0.2 for a uniform guess, 0.1 for the empirical best), we report results as *relative change* versus a baseline (either the base model with one-shot prompting or `GPT-4o` with one-shot prompting). For each method, we compute $\frac{|a_{\text{method}} - a_{\text{base}}|}{|a_{\text{base}}|} \times 100\%$, assigning the sign so that positive values indicate improvement and negative values indicate regression. Here, a_{method} is the metric (e.g., Wasserstein distance) for the method under evaluation, and a_{base} is the metric for the baseline.

Prompting Baselines. In our task formulation, we already incorporate all available demographic information from participants, which has proven an effective prompting mechanism in prior work (Hewitt et al., 2024). Prior work has also shown that reasoning over intermediate decisions (Wei et al., 2022b) or using in-context-prompting (Dong et al., 2022) improves LLM prediction accuracy. Accordingly, we evaluate three baselines: (1) direct prediction prompting; (2) prompting to generate explicit reasoning traces before prediction; and (3) in-context prompting with few-shot examples. We select few-shot examples by finding the closest prompt stimuli neighbor via cosine similarity of embeddings, then choosing examples from five random participants. We use OpenAI's `text-embedding-3-large` model to embed stimuli. Prompt templates are provided in Appx.D and Appx.E.

5.3 Generalization Across Unseen Studies

First, we consider the case when there is **no in-domain studies data are available** to finetune on. Specifically, we assess the study-wise generalization of our finetuning methods by evaluating on studies that are completely out of domain (i.e.,, not seen during training).

Setup. From the 210 studies in SocSci210, we split 170 studies as our train-studies and 40 studies as test-studies. We train over 100% of the training studies data, and evaluate on the 40 test studies. We compare the performance of our three finetuning objectives (SFT, SFT+Reasoning, and Contrastive DPO (§4)) across two open-source models, to three prompting baselines on `GPT-4o`.

Results. Tab. 2 details the results on our evaluation metrics (§3.2). We find that finetuning meaningfully generalizes to unseen studies and improves the distributional alignment metric. Relative to the `GPT-4o` baseline, after SFT, `LLaMA3-8B` outperforms `GPT-4o` by **12.1%**, and `Qwen2.5-14B` outperforms GPT4o by **13.2%**, reflecting relative gains of 30.1% and 26.3% from fine-tuning. In open-source models, prompting through reasoning

Model Variant	Accuracy ↑			Distribution ↓		
	Score	%Δ vs Base	vs GPT-4o	Score	%Δ vs Base	vs GPT-4o
Proprietary Models						
GPT-4o Base	72.9	–	–	0.174	–	–
+ Few-shot (5)	73.2	0.4%	0.4%	0.161	7.5%	7.5%
+ Reasoning	73.1	0.3%	0.3%	0.169	2.9%	2.9%
Open-Source Models						
LLaMA3-8B Base	70.3	–	-3.6%	0.219	–	-25.9%
+ Few-shot (5)	68.9	-2.0%	-5.5%	0.212	3.2%	-21.8%
+ Reasoning	69.8	-0.7%	-4.3%	0.174	20.6%	–
+ SFT	69.1	-1.7%	-5.2%	**0.153**	30.1%	12.1%
+ SFT w/ Reasoning	67.5	-4.0%	-7.4%	0.165	24.7%	5.2%
+ DPO	**72.6**	3.3%	-0.4%	0.185	15.5%	-6.3%
Qwen2.5-14B Base	72.9	–	–	0.205	–	-17.8%
+ Few-shot (5)	71.9	-1.4%	-1.4%	0.196	4.4%	-12.6%
+ Reasoning	72.7	-0.3%	-0.3%	0.166	19.0%	4.6%
+ SFT	69.5	-4.7%	-4.7%	**0.151**	26.3%	13.2%
+ SFT w/ Reasoning	67.6	-7.3%	-7.3%	0.164	20.0%	5.7%
+ DPO	**74.0**	1.4%	1.4%	0.181	11.7%	-4.0%
Bounds						
Uniform Guess	61.2	–	-16.1%	0.203	–	-16.7%
Empirical Best	–	–	–	0.125	–	28.2%

Table 2: Comparison of model variants on accuracy and distribution distance metrics across unseen studies (§5.3). In each scenario, **best scores are in boldface**, second-best underlined. Percent changes are relative.

achieves significant gains, though our fine-tuned model still outperforms this baseline on distribution alignment. The best distributional alignment score (0.151) is achieved by fine-tuning `Qwen2.5-14B`; given an empirical best bound of 0.125 on this metric, our model **closely approximate actual human response distributions**.

Interestingly, although distributional alignment improves for unseen studies, response-level accuracy does not necessarily increase. This suggests that individual predictions may become less precise as we better approximate the distribution of how responses should look (for example, if user responses under condition follow a distribution $N(0, 1)$, then a model that more accurately captures that distribution can incur higher error than one that always predicts the mean). We include an in-depth discussion analyzing this in appx.B.

For predicting individual accuracy metric, contrastive DPO outperforms all other methods achieving 73.9% accuracy. This is potentially due to our demographic-focused contrastive pair construction (§4) that enables models to learn detailed distinction in simulating individual decisions, leading to more precise predictions of each individual.

5.4 Generalization to Unseen Conditions and Outcomes

Researchers often have topic-specific dataset they can leverage for finetuning. For example, a political polarization researcher may have existing data on how an intervention shifts feelings toward the opposing party and want to run studies to predict either (a) how this intervention influences a different outcome, such as respondents' confidence in government or (b) how a new intervention will affect that same outcome. Thus, in this section, we examine: *How does finetuning help generalization to unseen conditions/outcomes within the same study?*

Setup. We subset SocSci210 to all studies that contain at least 4 conditions / outcomes. We split the dataset into two splits: one for testing condition generalization, and one for testing outcome generalization. For the condition split, we randomly selected 75% of the conditions across studies, and take all questions under those conditions for train, and hold out the other 25% for test. For outcome splits, we repeat the same process but sampling on outcomes. For each, we finetune on the train set, then run our evaluation on the test set. `LLaMA3-8B-Instruct` is used as an example base model.

Results. As shown in Tab. 3, when predicting held-out ***conditions***, our finetuned model with reasoning improves by **71%** in estimating response distributions relative to the base model. Notably the distributional alignment after finetuning **surpass that of the "Empirical Best" threshold**, suggesting that the alignment with predictions in response

Model Variant	Accuracy↑		Distribution↓	
	Score	vs Base	Score	vs Base
Condition Split				
LLaMA3-8B Base	71.0	–	0.219	–
+ SFT	**74.2**	4.5%	0.077	64.8%
+ SFT w/ R.	71.9	1.3%	**0.063**	71.2%
+ DPO	71.2	0.3%	0.208	5.0%
Uniform Guess	62.1	–	0.180	–
Empirical Best	–	–	0.090	–
Outcome Split				
LLaMA3-8B Base	71.7	–	0.224	–
+ SFT	71.7	0.0%	0.125	44.2%
+ SFT w/ R.	69.9	-2.5%	**0.114**	49.0%
+ DPO	**72.6**	1.3%	0.225	-0.5%
Uniform Guess	63.3	–	0.165	–
Empirical Best	–	–	0.086	–

Table 3: Performance metrics of LLaMA3-8B under different training configurations, evaluated on 75% train / 25% held-out splits for both Outcome and Condition scenarios (§5.4). In each scenario, **best scores are in boldface**, second-best underlined. Percent changes are relative.

to predictions is as close as another sample would be. SFT also increases the accuracy on individual predictions from 71.0% to 74.2%.

Across ***outcomes***, finetuning with reasoning also improves distribution distance, leading to 49% relative improvement compared to LLaMA3-8B base. Generalization across unseen conditions tends to be greater than across unseen outcomes. This may be because LLMs grasp the underlying effects of how condition manipulations influence responses, but are more prone to misestimate the initial distribution of outcome questions. Thus, by holding the outcome constant and varying only the stimuli, the model can more effectively learn the resulting effects. This has especially practical value since studies often test many different condition stimuli on the same outcome (*e.g.*, Strand et al. (2024) tests 25 interventions to reduce partisan animosity).

5.5 Generalization to Unseen Participants

Researchers often run pilot experiments on a small set of participants before committing to fully powered studies. Specific to such use case, we consider finetuning directly on a subset of participants and testing generalization to unseen participants in the same study. We examine how little data from a pilot study is needed to accurately predict outcomes for the remaining high-powered sample.

Setup. We reuse the same study-level train-test split from §5.3: 170 studies for training and 40 for testing. For the 170 training studies, we randomly divide the participants **once** into: *participant-train* (50% of all individuals for training); *participant-eval* (the remaining 50% for evaluation of unseen participants *but* seen studies). From the *participant-train* pool we draw progressively larger *pilot* subsets corresponding to $1, 5, 10, 20, 30, 40, 50\%$ of *all* participants (*i.e.*, the 50% split *all* of *participant-train*). For each split size we finetune the model and report performance on (i) the participant-eval splits of the 170 training studies (unseen individuals *but* seen studies) and (ii) the 40 completely held-out studies (unseen individuals *and* unseen studies).

This design reveals how much participant data a pilot must collect to obtain reliable generalization across both new participants and new studies.

Results. Fig. 4 shows the results on both the participants remained in the observed studies and all the participants in the observed studies.

In seen studies (the 50% held-out evaluation set), contrastive DPO tuning outperforms simple SFT for learning individual responses: with just 10% of the data, accuracy rises from 71% to 75% (a 13% relative error reduction). However, SFT estimates the distribution more effectively. When augmented by reasoning traces, SFT+Reasoning brings even further alignment on the distribution. Across all studies, the learning curves in Fig. 4 indicate that saturation of learning is consistently achieved at around **10%** of participant data.

In unseen studies (the completely held-out 40 studies), a similar pattern holds. Contrastive DPO tuning also shows superior performance on the individual response accuracy, surpassing GPT4o with only 10% of data. SFT yields better distribution alignment, whereas reasoning trace augmentation provides little benefit in this setting.

5.6 Demographic Bias and Parity

A principal strength of our dataset is that it captures rich demographic attributes of all participants and how they may affect responses. In this section, we analyze the response distributions for the full-data SFT of LLaMA3-8B, described in §5.3. We further break down our distributional alignment metric (*i.e.*, the Wasserstein distance between model responses under each unique condition and outcome) by demographic category: we subset the responses to each demographic subgroup and compute the average Wasserstein distance. Using this, we also

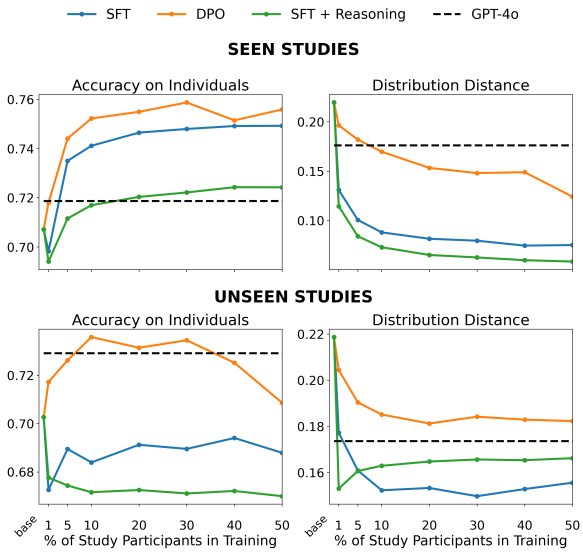

Figure 4: Learning curves on how % of training samples generalizes to held-out participants in seen studies and all participants in unseen studies, across varying participant size across studies (§5.5).

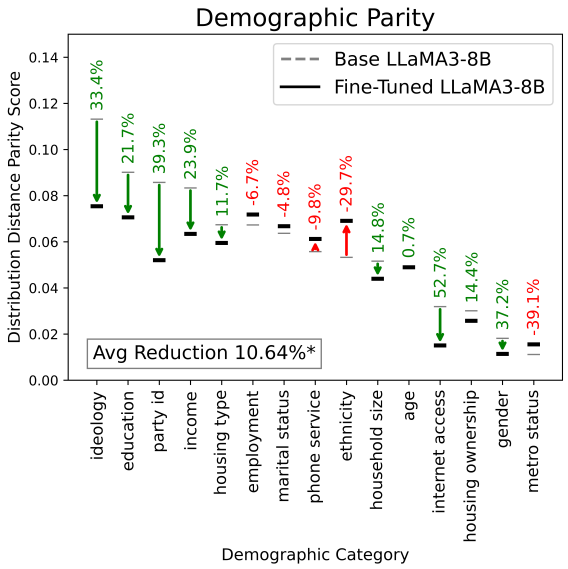

Figure 5: Parity reduction in predicting distributions across demographic categories after finetuning `LLaMA-8B`) (§5.6)

compute demographic parity difference, a measure of bias (Jiang et al., 2022) defined as the absolute gap between the highest- and lowest-performing demographic subgroups.

After finetuning – which improves overall distribution alignment – we observe an average relative improvement of 28.5% in distributional alignment across demographic categories (a full breakdown of improvement across every demographic subgroup is available in Appx.F). More notably, across all demographic categories, parity is reduced by 10.5%—a meaningful decrease in model bias, as shown in Fig. 5.

6 Conclusion

In this paper, we finetune LLMs to create a general use behavioral model that can accurately predict how individuals respond in social science experiments. We introduce SocSci210, a standardized, large-scale dataset comprising 2.9 million individual responses from more than 400,000 participants across 210 social science experiments. Through fine-tuning `Qwen2.5-14B`, we create Socrates-Qwen-14B, which relative to `GPT-4o`, produces predictions that are 13% more distributionally aligned with real human responses. Given the strong generalization we observe, we recommend that researchers begin finetuning on their existing datasets to yield more accurate and useful simulations. To support this, we will open-source both our dataset (SocSci210), models (Socrates) and finetuning code. These provide a foundation in creating a unified behavioral prediction engine that can power simulations across every discipline.

Limitations

Our participant sweeps in §5.5 show that performance tends to quick plateau with more participants' data. This is likely because we use relatively small parameter models. Given the size of our dataset, we anticipate that scaling to larger models (such as `Llama-70B` or `Llama-405B`) could further improve performance.

We rely on `GPT-4o-mini` to generate our oracle reasoning traces, and show that SFT on this traces does not always help improve performance for our task. Future work might investigate distilling those traces from more powerful reasoning models (*e.g.*, OpenAI's `o3`), which could create more performant models. Our models are trained exclusively on SocSci210. Although our dataset is diverse, it contains only representative samples of the U.S. population for closed-form questions. We do not evaluate the generalization of our training to non-U.S. populations or to open-form questions. Future work should focus on integrating datasets from prior research into our training paradigm, constructing new datasets, and exploring training on open-ended responses, which may further enhance performance and generalization.

Ethics Statement

We publicly release all data and our fine-tuned models. All materials were collected and processed in accordance with the respective data, checkpoint, and API usage policies. The dataset used in this study is drawn from publicly available, peer-reviewed social science experiments from the NSF's TESS repository, all of which comply with established ethical and privacy standards. Our dataset includes stimuli that may be considered contentious, and our fine-tuned models may generate incorrect or unsafe content. While fine-tuning has led to meaningful improvements in model accuracy, it may also lead users to become overconfident in the results. We strongly advise all users to verify outputs carefully before deploying this work in real-world applications.

Acknowledgments

We thank Omar Shaikh for helpful discussions on finetuning for this task. We acknowledge funding support from the Stanford Institute for Human-Centered Artificial Intelligence, Google, AXA, SCB X, Hanwha, and American Express.

References

Marah Abdin, Sahaj Agarwal, Ahmed Awadallah, Vidhisha Balachandran, Harkirat Behl, Lingjiao Chen, Gustavo de Rosa, Suriya Gunasekar, Mojan Javaheripi, Neel Joshi, and 1 others. 2025. Phi-4-reasoning technical report. *arXiv preprint arXiv:2504.21318*.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, and 1 others. 2024. Centaur: a foundation model of human cognition. *arXiv preprint arXiv:2410.20268*.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

James Brand, Ayelet Israeli, and Donald Ngwe. 2023. Using gpt for market research. *Harvard business school marketing unit working paper*, (23-062).

Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. 2024. Llm-based doppelgänger models: Leveraging synthetic data for human-like responses in survey simulations. *IEEE Access*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Eric Chu, Jacob Andreas, Stephen Ansolabehere, and Deb Roy. 2023. Language models trained on media diets can predict public opinion. *arXiv preprint arXiv:2303.16779*.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 2024. Take caution in using llms as human surrogates: Scylla ex machina. *arXiv preprint arXiv:2410.19599*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *Preprint*.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. 2022. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*.

Junsol Kim and Byungkyu Lee. 2023. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. *arXiv preprint arXiv:2305.16960*.

Yuxuan Lu, Jing Huang, Yan Han, Bennet Bei, Yaochen Xie, Dakuo Wang, Jessie Wang, and Qi He. 2025. Beyond believability: Accurate human behavior simulation with fine-tuned llms. *arXiv preprint arXiv:2503.20749*.

Petros Maravelakis. 2019. The use of statistics in social sciences. *Journal of Humanities and Applied Social Sciences*, 1(2):87–97.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M Chan. 2024. Virtual personas for language models via an anthology of backstories. *arXiv preprint arXiv:2407.06576*.

Matthias Orlikowski, Jiaxin Pei, Paul Röttger, Philipp Cimiano, David Jurgens, and Dirk Hovy. 2025. Beyond demographics: Fine-tuning large language models to predict individuals' subjective text perceptions. *arXiv preprint arXiv:2502.20897*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. 2021. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

David M Rothschild, James Brand, Hope Schroeder, and Jenny Wang. 2024. Opportunities and risks of llms in survey research. *Available at SSRN*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Palma Joy Strand, Jan Gerrit Voelkel, Michael Stagnaro, James Chu, Robb Willer, and Malka Kopell. 2024. Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity. *Available at SSRN 5034911*.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. 2025. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*.

Open Thoughts Team. 2025a. Open thoughts. https://github.com/open-thoughts/open-thoughts.

Qwen Team. 2025b. Qwq-32b: Embracing the power of reinforcement learning.

"TESS". 2025. Time-sharing experiments for the social sciences. *www.tessexperiments.org*.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. 2025a. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, pages 1–12.

Qian Wang, Zhenheng Tang, and Bingsheng He. 2025b. From chatgpt to deepseek: Can llms simulate humanity? *arXiv preprint arXiv:2502.18210*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022a. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. 2024. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22583–22599.

Yutong Xie, Zhuoheng Li, Xiyuan Wang, Yijun Pan, Qijia Liu, Xingzhi Cui, Kuang-Yu Lo, Ruoyi Gao, Xingjian Zhang, Jin Huang, and 1 others. 2025. Be. fm: Open foundation models for human behavior. *arXiv preprint arXiv:2505.23058*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024b. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Han Zhao, Haotian Wang, Yiping Peng, Sitong Zhao, Xiaoyu Tian, Shuaiting Chen, Yunjie Ji, and Xiangang Li. 2025. 1.4 million open-source distilled reasoning dataset to empower large language model training. *arXiv preprint arXiv:2503.19633*.

Jian-Qiao Zhu, Hanbo Xie, Dilip Arumugam, Robert C Wilson, and Thomas L Griffiths. 2025. Using reinforcement learning to train large language models to explain human decisions. *arXiv preprint arXiv:2505.11614*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Data Reconstruction Agent

In this section, we overview the workflow of our data construction agent. We intentionally avoided manual intervention to preserve a fully automated reconstruction pipeline. As shown in Fig. 6, we first download and standardize the publicly available data files form TESS [2], (*e.g.*, converting data files to standard `.csvs`, and all `pdfs/docx` to text). Then we feed our data reconstruction agent with each research paper's full context (including description, data files, codebook, and stimuli). The agent then: 1) Identifies all experimental conditions; 2) Identifies the outcome questions; 3) Writes and executes parsing code to merge and clean the repository's CSV files, reconstructing each participant's record, and combining the experimental condition and outcome question into a stimuli that maps to each participant response. This step goes through a generate-and-test cycle, which iterates until the code can reconstruct the dataset.

At the time of scraping, we pulled 443 publicly available project on the TESS OSF account. OSF projects often have a nested structure—where subprojects can inflate this count. After deduplicating these, we identified only 321 unique studies, of which our agent successfully reconstructed 210.

[2]Data available via Open Science Framework repository

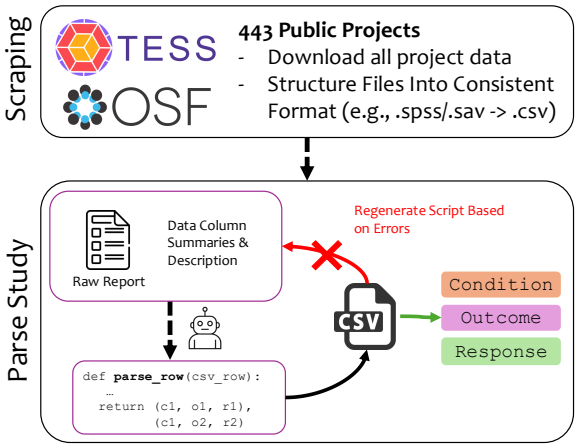

Figure 6: The workflow of our data construction agent for curating SocSci210. The agent generates a script based on all study context and iteratively regenerates it for successful parsing of the data files from the source.

The following conditions were present in our actual reconstruction script: (1) must be able to find a description of the stimuli present for each condition (2) must reconstruct original/binary questions (3) outcome questions correspond to a condition-specific stimuli. If these aren't satisfied the script will skip the study. A bulk of the failure cases were also just the inability of the LLM agent (powered by o4-mini-high) to accurately generate code that could parse the data into natural language, given the long context of the input data. A "successful" scrapes means the agent explicitly verifies two things during the code test-and-verify cycle: (a) The parsing code successfully compiles and executes without rasing errors (b) The parsing code generates non-empty outputs when operated row by row on the dataset.

B Discussion of Metric Evaluation: Accuracy vs. Distributional Alignment

As we discuss in the main text, for our considered experiments, we value accurately modeling the Wasserstein distribution under experimental conditions above achieving high individual-level accuracy. In this appendix, we flesh out the intuition behind this. In treatment effect experiments, verifying the hypothesis involves simply a t-test comparing group means, relying solely on distribution-level attributes (mean and standard deviation) of responses to each condition. Individual responses, since they are parameterized purely based on the stimuli and demographic information, are inherently stochastic.

Formally, the limitation of measuring individual response accuracy is that each response r is parameterized only by a (P, c, o) tuple, where P consists of demographic keys. However, for a given demographic P responding to a question o after seeing stimulus c, we cannot confidently assert that the output should always be r. Instead, our training data merely shows the example r, drawn from the distribution $F(P, c, o)$, where F is the underlying model of human responses. An ideal evaluation of a predictive model F' would therefore compare the distribution $F'(P, c, o)$ to $F(P, c, o)$. However, we do not have enough data for each unique demographic to recover such an exact distribution $F(P, c, o)$. Instead, we compare $F(\cdot, c, o)$ to $F'(\cdot, c, o)$. Furthermore, even if a model were predicting the exact distribution $F(P, c, o)$, such an approach often performs poorly when reduced to single-example accuracy measurements.

As a simple example of this, consider our training data might contain ten 42-year-old males responding to the same stimulus: five individuals answer "0," two answer "1," and three answer "2." A perfect model representing this scenario would predict responses "0," "1," and "2" with probabilities of 50%, 20%, and 30%, respectively. Yet, such predictions yield low accuracy scores when applied on our training data. Conversely, consistently predicting the midpoint ("1") improves accuracy but poorly represents the true distribution.

Our intuition is that `GPT-4o` achieves higher accuracy because its responses often cluster narrowly around mid-scale values (e.g., in our evaluation set the standardized `GPT-4o` predictions have $\sigma = 0.154$, while our SOCRATES-LLAMA-8B predictions have $\sigma = 0.195$ and human responses $\sigma = 0.192$ where σ is sample standard deviation).

To further illustrate the limitations of accuracy, we also compare two predictors: a) always predicting the question response scale mean vs. b) randomly sampling responses from the ground truth distribution of responses in the experimental condition. Across our 40 evaluation studies, the accuracy of method (a) outperforms method (b) for 30% of cases, despite (a) being meaningless and (b) representing a model predicting perfectly.

C Implementation Details

Finetuning Configurations. In all finetuning experiments, we finetune models for 1 epoch with a global batch size of 256 on 8 NVIDIA-A100X80G GPUs. The learning rate (LR) is set to $1e{-}05$ for SFT and $1e{-}06$ for DPO. We adopt cosine LR scheduler with a warm-up ratio of 0.05 and weight decay of 0.1. Training times varied between experiments, but took roughly between 4 and 24 hours. During inference with all open-source models, we uniformly set temperature=0.6, top_p=0.9, max_length=4096.

Prompting Configurations. When prompting OpenAI's proprietary models (e.g. `GPT-4o` for simulation, `GPT-4o-mini` for reasoning trace generation, `o4-mini-high` for dataset consutrction), we use default API parameters (*e.g.*, temperature=1, top_p=1). All metrics reported reflect results from a single experiment run (except when noted otherwise).

D Prompt Templates For Prediction

Prompt for Direct Prediction

[SYSTEM]: You are simulating a survey respondent. Answer exactly as instructed, following the specified response format without additional commentary.

[USER] You are a survey respondent with the following demographic profile:
`{Demographic Info.}`

Read the question below and answer exactly as this person would. Follow the response instructions precisely.
`{Stimuli}`

Example User Message of Direct Prediction

[USER] You are a survey respondent with the following demographic profile:
- Age: 29
- Gender: Female
- Education: Vocational/tech school/some college/associates
- Employment: Employed as paid employee
- Marital Status: Never married
- Housing Ownership: Occupied without payment of cash rent
- Household Size: 6
- Ideology: Somewhat Liberal
- Phone Service: Cellphone only

Read the question below and answer exactly as this person would. Follow the response instructions precisely.

You read 'Emily recently graduated from high school and will attend college in the fall. Her mother and father, both factory workers, are very proud of her. Emily is excited to be attending her first-choice college, a highly-ranked private university. The university provides funding to cover the costs that families cannot pay, so Emily will graduate with no debt.' and then were asked: 'How unlikely or likely would you be to recommend history?' Only return an integer from 1 to 6, nothing else.

Modified System Message For Reasoning-Based Prompting

[SYSTEM]:
You are simulating a survey respondent. You are to answer exactly as instructed, but also include your reasoning (5 sentences or less) before you output your answer.Please follow the exact output format below.
Output format
<trace>
. . . your step-by-step reasoning here. . .
</trace>
PREDICTION: <verbatim answer> (conclude with predicted answer, use exactly the option label/number with no extra commentary)

Modified System Message for Few-Shot Prompting

[SYSTEM]:
You are simulating a survey respondent. Answer exactly as instructed, following the specified response format without additional commentary.

As you answer, consider how the following similar question was answered by other participants:

Question: `{Sampled Similar Stimuli}`

Person 1 Profile:
`{Person 1 Demographic Info.}`
Answer: `{Person 1 Answer}`

Person 2 Profile:
`{Person 2 Demographic Info.}`
Answer: `{Person 2 Answer}`

`......`
`{...More Persona Examples...}`
`......`

Person 5 Profile:
`{Person 5 Demographic Info.}`
Answer: `{Person 5 Answer}`

E Reasoning Trace Generation Prompts

We use `GPT-4o-mini` in order to generate "oracle" reasoning traces from our finetuning data. The model generates what an ideal reasoning trace *should look like* given a prompt, response pair.

Prompt for Oracle Reasoning Traces

[SYSTEM]:You are an expert behavioral scientist asked to write a plausible, forward-looking reasoning trace that *predicts* which answer a survey respondent will give. Draw on knowledge of behavioral and social science theory to explain how and why this person responded the way they did.

Key constraints for the reasoning trace
1. **Prospective viewpoint.** Write as if you do *not* know the final choice yet. Describe the mental steps a typical person with the given persona might take when first seeing the stimuli.
2. **No answer leakage inside the trace.** The true answer is supplied only for your private verification. Do **not** quote, paraphrase, or rely on it within the narrative.
3. Be concise but specific in your reasoning and avoid repetition. Keep the reasoning trace 5 sentences or less."

Output format
<trace> ...your step-by-step reasoning here (written as if before 'knowing' the answer)... </trace>
PREDICTION: <verbatim answer> (conclude with predicted answer, use exactly the option label/number with no extra commentary)

[USER]
Persona: `{Demographic Info.}`
Stimuli: `{Stimuli}`
<!– TRUE ANSWER (use only to verify your prediction; do NOT reference inside <trace>): `{Ground-Truth Response}` –>

Write the reasoning trace and final prediction now, following the format above.

Example Prompt and Response with Reasoning Trace

Prompt: You are a survey respondent with the following demographic profile: `{Demographic Info.}`

Read the question below and answer exactly as this person would. Follow the response instructions precisely.

You viewed an image and description of the Hyundai Kona Electric, a 100% electric version of the Hyundai Kona that costs about $40,000, has a 250 mile range and can accelerate from zero to 60 mph in about 6.4 seconds and then were asked: "What is your first reaction to the product?" Only return an integer from 1 to 5, where 1 means Very negative and 5 means Very positive, nothing else.

Reasoning: Upon seeing the Hyundai Kona Electric, the individual may first consider their personal values and preferences, which lean strongly towards traditional vehicles and fossil fuels, aligning with their very conservative ideology. The relatively high price point of $40,000 is likely to evoke concerns about affordability, especially given their household income bracket. Furthermore, the electric vehicle's technological aspects and environmental emphasis may clash with their skepticism towards climate change narratives and push towards eco-friendly solutions. Additionally, a family of four may prioritize practicality and utility over innovation, leading them to view a fully electric vehicle as less suitable for their needs. Overall, these factors suggest a negative response to the product.
Response: 1

F Demographic Subset Results

See Tab. 4 for a breakdown of distribution distance under each demographic category of our dataset.

Attribute	Category	LLaMA-8B	Finetuned	(%) Improvement
age	65-100	0.2353	0.1679	28.63
age	35-50	0.2284	0.1571	31.23
age	50-65	0.2340	0.1599	31.66
age	25-35	0.2190	0.1484	32.23
age	0-25	0.2196	0.1510	31.24
gender	Female	0.1910	0.1342	29.72
gender	Male	0.1814	0.1165	35.80
education	Post grad study/professional degree	0.2381	0.1838	22.84
education	Vocational/tech school/some college/associates	0.2358	0.1612	31.64
education	Bachelor's degree	0.2339	0.1745	25.39
education	High school graduate or equivalent	0.2356	0.1527	35.18
education	Some high school (no diploma)	0.2602	0.1676	35.58
education	Less than high school	0.2578	0.1861	27.79
employment	Self-employed	0.1984	0.1506	24.12
employment	Employed as paid employee	0.1876	0.1380	26.46
employment	Disabled	0.2070	0.1565	24.38
employment	Retired	0.2006	0.1450	27.73
employment	Looking for work	0.1925	0.1437	25.37
employment	Not working for other reasons	0.1950	0.1424	26.97
employment	Temporarily laid off	0.2122	0.1740	18.01
marital status	Divorced	0.2412	0.1769	26.68
marital status	Married	0.2342	0.1652	29.45
marital status	Never married	0.2227	0.1528	31.40
marital status	Living with partner	0.2244	0.1581	29.52
marital status	Widowed	0.2441	0.1882	22.92
marital status	Separated	0.2507	0.2047	18.33
housing ownership	Owned or being bought by you/someone in your household	0.1909	0.1343	29.64
housing ownership	Rented for cash	0.1860	0.1300	30.09
housing ownership	Occupied without payment of cash rent	0.2116	0.1670	21.05
housing type	A one-family house detached from any other house	0.1906	0.1329	30.25
housing type	A one-family house attached to one or more houses	0.1847	0.1354	26.70
housing type	A mobile home or trailer	0.2092	0.1564	25.26
housing type	Boat, RV, van, etc	0.2151	0.1795	16.57
housing type	A building with 2 or more apartments	0.1695	0.0824	51.37
metro status	Metro Area	0.1876	0.1299	30.73
metro status	Non-Metro Area	0.1963	0.1451	26.10
income	50-74K	0.2282	0.1560	31.65
income	40-49K	0.2340	0.1645	29.70
income	20-29K	0.2349	0.1482	36.89
income	200K+	0.2016	0.1648	18.25
income	125-149K	0.2373	0.1708	28.03
income	75-99K	0.2328	0.1671	28.22
income	30-39K	0.2341	0.1584	32.31
income	100-124K	0.2393	0.1761	26.43
income	150-175K+	0.2536	0.1962	22.66
income	175-200K+	0.2484	0.1936	22.05
income	15-19K	0.1677	0.0563	66.45
income	10-14K	0.2158	0.0892	58.66
income	5-9K	0.1914	0.0857	55.25
income	<5K	0.1306	0.0395	69.78
internet access	Internet Household	0.1888	0.1316	30.28
internet access	Non-internet household	0.1932	0.1386	28.29
household size	0-3	0.2310	0.1574	31.86
household size	3-6	0.2266	0.1536	32.21
household size	6-9	0.2311	0.1644	28.83
household size	9-20	0.2851	0.2190	23.18
phone service	Cellphone only	0.1863	0.1312	29.59
phone service	Have a landline, but mostly use cellphone	0.1977	0.1377	30.36
phone service	Have cellphone, but mostly use landline	0.1965	0.1447	26.37
phone service	Landline telephone only	0.2130	0.1607	24.55
phone service	No telephone service	0.2232	0.1889	15.40
party id	Moderate Democrat	0.1784	0.1400	21.54
party id	Don't Lean/Independent/None	0.1822	0.1391	23.64
party id	Strong Democrat	0.1776	0.1554	12.48
party id	Lean Republican	0.1914	0.1456	23.89
party id	Lean Democrat	0.1705	0.1557	8.66
party id	Strong Republican	0.1855	0.1505	18.83
party id	Moderate Republican	0.1858	0.1466	21.10
ideology	Somewhat Liberal	0.1804	0.1514	16.06
ideology	Moderate	0.1878	0.1414	24.73
ideology	Liberal	0.1877	0.1357	27.68
ideology	Somewhat Conservative	0.2063	0.1568	23.98
ideology	Conservative	0.2093	0.1591	24.00
ideology	Extremely Liberal	0.2002	0.1494	25.36
ideology	Declined to Answer	0.2379	0.1592	33.08
ideology	Extremely Conservative	0.2342	0.1949	16.80
ideology	Very Conservative	0.1888	0.1649	12.66
ideology	Very Liberal	0.1809	0.1613	10.82
ethnicity	Hispanic	0.2642	0.1792	32.17
ethnicity	White	0.2679	0.1808	32.52
ethnicity	2+ Race	0.2849	0.2044	28.24
ethnicity	Other	0.2652	0.1895	28.54
ethnicity	Black	0.2749	0.1850	32.70

Table 4: Prediction improvement in Wasserstein distance under each demographic category.