Valid Survey Simulations with Limited Human Data: The Roles of Prompting, Fine-Tuning, and Rectification

Stefan Krsteski[1], **Giuseppe Russo**[1,2], **Serina Chang**[3], **Robert West**[1], **Kristina Gligorić**[4]
[1]EPFL
[2]Stanford University
[3]University of California, Berkeley
[4]Johns Hopkins University
Correspondence: stefan.krsteski@epfl.ch

Abstract

Surveys provide valuable insights into public opinion and behavior, but their execution is costly and slow. Large language models (LLMs) have been proposed as a scalable, low-cost substitute for human respondents, but their outputs are often biased and yield invalid estimates. We study the interplay between synthesis methods that use LLMs to generate survey responses and rectification methods that debias population estimates, and explore how human responses are best allocated between them. Using two panel surveys with questions on nutrition, politics, and economics, we find that synthesis alone introduces substantial bias (24–86%), whereas combining it with rectification reduces bias below 5% and increases effective sample size by up to 14%. Overall, we challenge the common practice of using all human responses for fine-tuning, showing that under a fixed budget, allocating most to rectification results in more effective estimation.

1 Introduction

Self-reported surveys are the gold standard for capturing how people think, feel, and behave across domains such as public policy, economics, and health. However, they are costly, time-consuming, and logistically complex (Groves et al., 2011). Recent works at the intersection of survey research and natural language processing have explored using LLMs as proxies for human respondents (Gao et al., 2024; Lira et al., 2022; Argyle et al., 2023; Anthis et al., 2025; Bail, 2024).

Despite their potential (Manning et al., 2024; Shah et al., 2025), LLMs are not reliable out-of-the-box as survey respondents (Gao et al., 2024). Empirical studies document demographic and positional biases (Cheng et al., 2023; Wang et al., 2025a), sensitivity to prompt wording, lexical features, and option order (Atreja et al., 2025; Gligoric et al., 2024), desirability bias (Sharma et al., 2023; Cheng et al., 2025), and hallucinations or self-contradictions (Tjuatja et al., 2024; Dominguez-Olmedo et al., 2024; Pezeshkpour and Hruschka, 2023; Huang et al., 2025). Naïve use can therefore distort population estimates. Training-time adaptations such as fine-tuning on survey responses demand extensive human annotation and remain vulnerable to domain shift, whereas inference-time techniques like demographic prompting or persona-based generation are highly prompt-sensitive (Sun et al., 2025).

Methods such as *Prediction-Powered-Inference* (PPI) and *Design-based Supervised Learning* (DSL) (Angelopoulos et al., 2023a; Egami et al., 2023) have been proposed as a post-hoc correction approach, but they have not been rigorously evaluated for large-scale survey simulation or in combination with training- and inference-time adaptations often used in practice. Moreover, these approaches guarantee validity only for corrected estimates, not for the generated responses themselves. Ensuring that generations are unbiased remains important when follow-up questions are posed (Wang et al., 2024a; Shaikh et al., 2024) or when other quantities beyond the corrected estimate need to be inferred.

Consequently, the effectiveness of training-time, inference-time, and post-hoc methods for valid LLM-based survey simulation is still unclear, as are their potential interactions. A priori, the optimal allocation of limited human data across these methods is not evident. For example, dedicating all data to fine-tuning precludes effective post-hoc correction, while allocating none may compromise the quality of generated responses. Nonetheless, these interactions remain uncharacterized and few guidelines exist to date.

To address these gaps, we conduct an evaluation examining supervised fine-tuning, persona-guided prompting, and rectification methods, and how to allocate gold-standard human responses for maxi-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10887–10906

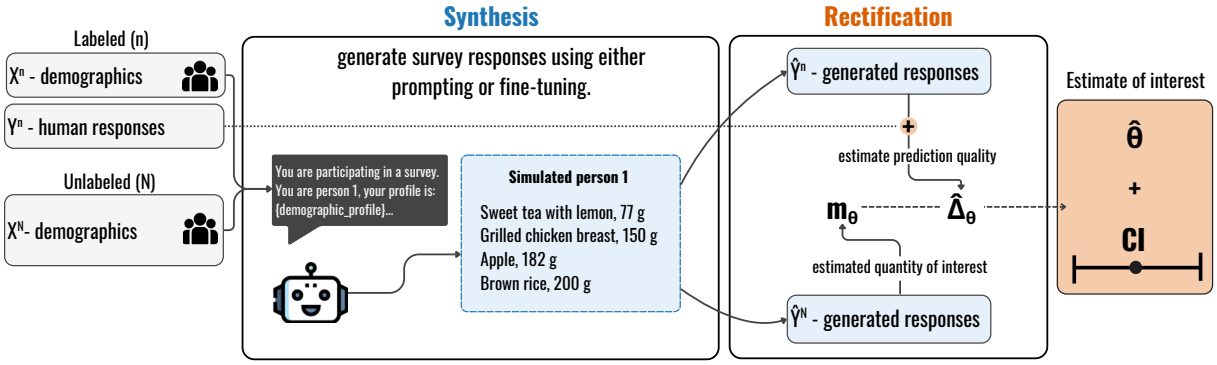

Figure 1: **Evaluation setup: Overview of synthesis and rectification.** Given a small human dataset (X^n, Y^n) and a disjoint, large demographic only dataset X^N, `Synthesis` produces responses $\hat{Y}^n, \hat{Y}^N$ using either prompting or fine-tuning. `Rectification` then combines model predictions $\hat{Y}^n$ with human responses Y^n to compute a correction term. Then, this term is combined with $\hat{Y}^N$ to produce a final estimate $\hat{\theta}$ of the target θ^*, with corresponding confidence intervals.

mizing performance. Our evaluation is carried out on two large-scale surveys: the NHANES dietary recall survey (for Health Statistics et al., 2017) and the American Trends Panel (ATP) (Pew Research Center). Studying longitudinal data allows us to address the most general setting, from which simpler survey designs can be derived. Through this setup, we contribute a format-agnostic, budget-aware study of training-time (fine-tuning), inference-time (prompting), and post-hoc rectification methods for survey simulation. Our contributions are the following:

1. **A rigorous evaluation** of four common synthesis strategies combined with two rectification methods, grounded in two large-scale surveys with focal tasks in food, politics, and economics (§4, §5.1).

2. **Quantitative insights** showing that population and subpopulation estimates can be corrected to within 5% bias using as few as 100 responses (≈1% of a full survey) (§5.2).

3. **Evidence-based guidelines** for allocating responses across synthesis and rectification (Table 4). For example, while dedicating all data to fine-tuning may seem natural, we show that under a fixed budget, allocating the majority of responses to rectification yields the best bias–variance trade-off.

All code and data are released to support development of new methods and benchmarks[1].

[1]Code: github.com/skrsteski/survey-simulations

2 Related work

Training-time adaptation. LLMs have been proposed as proxies for human respondents, enabling low-cost, large-scale survey simulation in domains such as public opinion, voting (Argyle et al., 2023; Santurkar et al., 2023). A common way to improve domain performance is fine-tuning. Recent work aligns response distributions by minimizing a forward KL divergence between model probabilities and human distributions for a given (question, subgroup) pair (Suh et al., 2025; Cao et al., 2025; Davidson et al., 2024). First-token probability alignment is typical for multiple-choice questions (MCQs). Although LoRA and other PEFT techniques reduce computation, fine-tuning still requires unpredictable amounts of gold-standard data and meticulous preprocessing. (Hu et al., 2022; Suh et al., 2025; Wang et al., 2025b). Moreover, survey-specific artifacts require careful format alignment (Dominguez-Olmedo et al., 2024; Wang et al., 2024b). This raises the question of where scarce human responses are most valuable, but existing studies provide no principled answer. Our evaluation aims to address this gap.

Inference-time adaptation. Prompting and in-context learning adapt the model at inference without parameter updates, making them a popular choice for survey simulation (Brown et al., 2020; Latona et al., 2024). Early work such as "silicon sampling" shows that sociodemographic conditioning can elicit responses with high population-level fidelity to humans (Argyle et al., 2023; Santurkar et al., 2023; Sun et al., 2024). Later methods introduce synthetic personas and steering (Hu and

Collier, 2024), including steering vectors (Kim et al., 2025; Russo et al., 2025b), soft-prompt control (Li et al., 2023, 2025), and probabilistic mixtures of persona-conditioned agents (Bui et al., 2025). Despite this progress, evaluations report prompt sensitivity, question- and option-order artifacts, and demographic skew, casting doubt on reliability for population-level inference (Dominguez-Olmedo et al., 2024; Tjuatja et al., 2024; Geng et al., 2024). A related concern is format-induced bias: many studies focus on MCQ/Likert formats, whereas prior work shows that responses under constrained generation can diverge from open-ended ones (Röttger et al., 2024; Zhang et al., 2025). If minor setup changes (e.g., prompt or model) lead to substantially different responses, the validity of conclusions from such simulations is questionable. In this work, we show that incorporating even a small amount of human responses via rectification can mitigate such biases.

Post-hoc adaptation. A complementary strategy bypasses model editing, treating the LLM as an informative, but potentially biased black-box predictor whose outputs are statistically corrected. Frameworks such as PPI (Angelopoulos et al., 2023a) and confidence-driven inference (CDI) (Gligorić et al., 2024) provide finite-sample–valid estimators and confidence intervals by combining abundant model predictions with a small human subset. Although previously applied to annotation tasks (Fan et al.; Calderon et al., 2025; Rister Portinari Maranca et al., 2025), this paradigm fits survey workflows where instrument design is costly, but collecting a small subsample of human answers is easy. In social-science methodology, design-based semi-supervised learning (DSL) combines model predictions with a small human sample via a doubly robust estimator (Egami et al., 2023). Similarly, the mixed-subjects design incorporates LLM predictions as additional observations alongside human responses (Broska et al., 2025). However, existing validations focus on controlled behavioral tasks rather than large-scale survey simulation, and they do not examine interactions with different synthesis strategies or data-allocation trade-offs. Here, we provide guidance on how to get the best of both worlds, simultaneously debiasing synthesis models and population-level estimates. We fill this gap with a multi-dataset evaluation that pairs several synthesis choices (prompting/personas and fine-tuning) with post-hoc correction across open-ended and MCQ questions.

3 Methodology

The task of interest is producing accurate population-level estimates that reflect human survey responses. To this end, we evaluate strategies that synthesize responses conditioned on demographics, correct estimates using a small subset of human responses, and combine both to balance their strengths.

Problem formalization. We frame the problem in the general setting of panel surveys, from which cross-sectional designs arise as a special case. Thus, the formulation leverages histories when available and reduces to the cross-sectional case when not. A panel survey is a sequence of responses collected from the same N individuals over T discrete time points. A single time point t (a "wave") may represent, for example, a monthly opinion poll. For each wave t, let $q_t \in \mathcal{Q}$ denote the survey question asked. Each participant $i \in \{1, \ldots, N\}$ provides a response $y_{i,t}$ to q_t, respecting the response space $\mathcal{Y}_{q_t}$ (e.g., multiple-choice options or free-text answers). Given a history window of length $T-1$, we generate a synthetic response at wave T using an LLM:

$$\hat{y}_{i,T} \;=\; f\big(\mathbf{x}_i,\, y_{i,1:T-1}\big).$$

Our objective is to recover the population estimand at wave T as the finite-population mean over the unlabeled, demographic-only frame:

$$\theta^* := \frac{1}{N}\sum_{i=1}^{N} \phi(y_{i,T})\,,$$

where $\phi : \mathcal{Y} \to \mathbb{R}$ maps responses to a common scale (e.g., numeric coding for Likert, scalar extraction for open-ended). Population-level estimates are, by definition, question-specific and require a single target question.

Synthesis. We use LLMs to generate survey responses, considering both inference-time adaptations (e.g., demographic or persona prompting) and training-time adaptations (e.g., domain-specific fine-tuning on survey data). At a high level, synthesis strategies divide into two categories:

Prompt-based methods rely on conditioning without parameter updates. `Demo-only` (demographic conditioning) prompts models with participant demographics $\mathbf{x}_i$ alone. `Persona-guided`

extends demographic prompting by incorporating behavioral patterns from past responses $(y_{i,1:T-1})$. An auxiliary LLM analyzes each participant's past responses to generate natural language personas capturing recurring behavioral tendencies, which then condition response generation at time T.

Fine-tuning methods adapt model parameters using training data, typically through supervised fine-tuning (SFT) as in instruction-following setups (Ouyang et al., 2022). `Domain-FT` fine-tunes on historical responses from our target survey across time points 1 to $T-1$, learning from (question, demographics, response) triplets via standard cross-entropy loss. `SubPOP-FT` fine-tunes on the SubPOP auxiliary dataset (Suh et al., 2025), which contains 3,229 questions from American Trends Panel with response distributions across 22 subpopulations. This method applies first-token alignment to minimize KL divergence between model logits and empirical response distributions for each (question, subgroup) pair, using external survey data rather than the target survey's history. We include it due to its demonstrated generalization to unseen surveys and subpopulations (Suh et al., 2025).

Rectification. In the synthesis step, a language model f generates predictions $\hat{y}_{i,T}$ for each participant. However, these predictions can be biased by factors such as the model's training data or the chosen prompt (Bender et al., 2021). An alternative is to collect human responses for the same survey question to estimate θ^*. While more reliable, such human data are costly to obtain (Groves et al., 2011), and far less abundant than LLM responses. This trade-off motivates correction frameworks such as PPI and DSL (Angelopoulos et al., 2023a; Egami et al., 2023), which combine cheap, plentiful model predictions with a small set of human answers.

We therefore assume access to a small set of human responses $\mathcal{H} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n}$ at wave T. For each $j \in \mathcal{H}$, we also compute a model prediction $\hat{y}_j = f_\pi(\mathbf{x}_j)$ using our synthesis setup. A general correction estimator takes the form

$$\hat{\theta}_\lambda = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \lambda\, \hat{y}_i}_{\text{synthetic mean}} + \underbrace{\frac{1}{n}\sum_{j=1}^{n} \big(y_j - \lambda\, \hat{y}_j\big)}_{\text{bias correction}}, \quad (1)$$

where $\lambda \in [0, 1]$ is a scalar ("power-tuning" parameter) interpolating between ignoring model predictions ($\lambda = 0$) and using them fully ($\lambda = 1$). This formulation corresponds to the PPI estimator (Angelopoulos et al., 2023a), with DSL (Egami et al., 2023) recovered as the special case $\lambda = 1$, which we refer to as $\text{Rec}_{\lambda=1}$. When λ is not specified, it is chosen from the human set $\mathcal{H}$ using the PPI++ power-tuning rule (Angelopoulos et al., 2023b), which minimizes the estimated variance of the estimator; we denote this as $\text{Rec}_{\lambda_{\text{opt}}}$.

A key benefit is variance reduction. If the synthetic mean (first term) is computed on a set $\mathcal{U}$ disjoint from the human responses set $\mathcal{H}$, the variance decomposes as

$$\text{Var}\big(\hat{\theta}_\lambda\big) = \frac{\lambda^2\, \text{Var}(\hat{y})}{N} + \frac{\text{Var}\big(y - \lambda\, \hat{y}\big)}{n}. \quad (2)$$

The first term is the variability of synthetic predictions, while the second term reflects the prediction error variance on the human responses set. According to Eq. (2), two conditions make this estimator more effective than using human data alone: (i) access to a large set of demographics and (ii) reasonably accurate predictions. Or, formally:

$$\frac{\lambda^2\, \text{Var}(\hat{y})}{N} + \frac{\text{Var}\big(y - \lambda\, \hat{y}\big)}{n} < \frac{\text{Var}(y)}{n}. \quad (3)$$

The first condition is typically satisfied in survey research, since demographic covariates can be collected at scale (e.g., from census data) without requiring human responses to substantive questions. The second condition is equally important: an accurate model means the prediction error variance, $\text{Var}(y - \hat{y})$, is small. As a result, the second term in Eq. (3) becomes negligible, and the estimator's variance is dominated by the first term $\frac{\text{Var}(\hat{y})}{N}$. Since this term shrinks as the synthetic sample size N increases, rectification can produce significantly tighter confidence intervals than estimators that rely solely on the small set of human responses. We refer readers to Appendix B.1 for more details.

4 Experiments

Our evaluation uses two longitudinal panel surveys chosen to span different domains (diet, economics, politics), response formats (open-ended, multiple-choice), and response distributions (approximately normal, skewed) (Table 1). Following our problem formulation, we evaluate at the question level: NHANES contributes one repeated dietary-intake item across two waves, while ATP contributes two distinct opinion items across four waves.

	NHANES (diet)	ATP Q1 (economics)	ATP Q2 (politics)
Response format	Open-ended (24h recall)	Multiple choice (6)	Multiple choice (4)
Participants	8.5k	691	643
Target	Mean daily energy intake (kcal)	Mean Likert score	Mean Likert score
Target mean	1766 kcal	3.16 (scale 1–6)	3.57 (scale 1–4)
Covariates	12 demo./lifestyle	25 demo./political	25 demo./political
Waves (T) and repeat	2 (repeated)	4 (not repeated)	4 (not repeated)

Table 1: **Datasets used in our evaluation.** NHANES includes two waves (T=2) asking the same food choice question, so responses are directly comparable across waves. The selected tasks vary in domain, format, and response distribution to test robustness across heterogeneous settings. ATP Q1 and Q2 are observed over four waves (T=4), with unique (non-repeated) focal questions at wave T. We nevertheless treat ATP as a panel on covariates and prior responses: cross-wave trajectories (waves 1:$T-1$) are used to construct personas and for fine-tuning.

NHANES. The U.S. National Health and Nutrition Examination Survey 2015–2016 (for Health Statistics et al., 2017) is a food-consumption survey with over 16,000 full-day dietary recall entries from 8,500 participants across two waves (T=2). Each entry records participants' food intake over the previous 24 hours in an open-ended format (e.g., 'oatmeal 100g, rice 150g, banana 45g'), along with total daily energy intake. Participant metadata includes 12 demographic and lifestyle covariates (e.g., age, sex, income). This survey allows us to evaluate open-ended generation with substantial individual- and temporal-level variation (its mapping function ϕ is detailed in Appendix B.2). The target is the mean daily energy intake (kcal) per participant, with a dataset mean of 1,766.

American Trends Panel (ATP). ATP is the Pew Research Center's longitudinal panel for U.S. public-opinion research (Pew Research Center), consisting of approximately 10,000 randomly selected adults nationwide. We select two focal questions from waves 146–149, differing in domain (economics vs. politics) and distribution (normal vs. skewed), providing a controlled yet diverse setting to assess the robustness of the methods.

Question 1 (Economic well-being): "Compared to your parents when they were the age you are now, do you think your own standard of living now is..." Options: (1) Much better, (2) Somewhat better, (3) About the same, (4) Somewhat worse, (5) Much worse, (6) Not sure. We analyze 691 complete cases. The target is the mean Likert score on a 1–6 scale (dataset mean: 3.16).

Question 2 (Political opinion): "How would you rate the job Supreme Court justices are doing in keeping their own political views out of how they decide major cases?" Options: (1) Excellent, (2) Good, (3) Only fair, (4) Poor. We analyze 643 complete cases. The target is the mean Likert score on a 1–4 scale with a mean of 3.57.

Both ATP items include 25 demographic and political covariates (e.g., age, gender, education, race, party ID, income, region). The two questions have different answer distributions (Q1 is approximately normal, Q2 left-skewed), allowing us to test performance across distinct response patterns.

Evaluation setup. We compare four synthesis methods across multiple datasets and models, applying two post-hoc correction strategies uniformly to each. For each dataset and model, we generate synthetic responses at wave T using one of the four synthesis methods described in §3. All models use a fixed sampling temperature of $\tau = 0.7$, with prompts held constant (Appendix B.3). We evaluate across four language models: Qwen2.5 8B, Llama 3.1 8B, Mistral v0.3 7B, and GPT-4o mini. Rectification methods (PPI and DSL) are applied on top of each synthesis strategy. At wave T, we draw $n_{\text{human}} = 100$ participants as the gold-standard set and treat the remainder as unlabeled. As baselines, we use previous-day responses for NHANES and, for ATP, random responses obtained by uniformly sampling answer options independently of covariates.

We assess performance using two complementary metrics capturing bias and variance. Bias is measured as the relative error between estimated and true population parameters:

$$\Delta_{\%} = \frac{|\hat{\theta} - \theta^*|}{\theta^*} \times 100, \tag{4}$$

where $\hat{\theta}$ is our estimator and θ^* is the ground-truth population parameter from full human responses.

Variance reduction is summarized by ESS gain,

$$\mathrm{ESS}_{\mathrm{gain}\%} = \left(\frac{\mathrm{Var}(\hat{\theta}_{\mathrm{human}})}{\mathrm{Var}(\hat{\theta}_{\mathrm{method}})} - 1 \right) \times 100, \quad (5)$$

so, for example, an ESS gain of 50% means the method achieves the same precision as having 1.5× more human data. This is equivalent to getting more information out of each human response.

5 Results

We first compare synthesis and rectification methods across datasets, focusing on population-level bias and variance (Table 2). Then, we turn to deeper analyses, examining how to best allocate human responses between fine-tuning and correction, as well as how rectification affects subgroup bias.

5.1 Bias and efficiency across methods

Unrectified synthesis is biased, whereas rectification consistently fixes it. Pure LLM synthesis shows large and inconsistent bias (top block). Averaged across datasets, the baseline has 24.11% bias, while `Domain-FT`, `Persona-guided`, `Demo-only`, and `SubPOP-FT` have 34.66%, 50.76%, 55.60%, and 86.23%, respectively. Per-dataset behavior is heterogeneous: e.g., on ATP Q2 (skewed/heavy-tailed answer distribution), `Persona-guided` reduces the unrectified bias to 19.99% vs. a 62.41% baseline, but performs poorly elsewhere. We hypothesize that low performance is due to misalignment between the personas generated and the true characteristics of the respondents. Furthermore, the variability in bias across methods and datasets confirms that synthesis-only methods cannot be trusted to make valid claims about population preferences or behaviors. Applying rectification collapses bias to single digits (bottom block). $\mathrm{Rec}_{\lambda=1}$ achieves some reduction but leaves significant residual bias, making it less reliable than $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$. By contrast, $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ consistently drives bias even lower: the lowest average bias is achieved by `Domain-FT` | $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ (2.82%), followed closely by `SubPOP-FT` | $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ (3.45%), both training-based methods, while all approaches under $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ on average remain below 5.5% bias. Importantly, training-based methods with $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ reduce bias to statistically insignificant levels across all datasets.

$\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ rectification enables positive ESS gains. ESS is meaningful only when confidence intervals achieve nominal coverage, i.e., when they contain the true population value at the expected frequency. All unrectified methods fail this criterion. However, when combined with $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$, every method achieves positive ESS gains, confirming a significant reduction in variance relative to the human-only estimator. `Persona-guided` + $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ attains the highest average ESS gain (6.92%) and peaks at 14.19% on NHANES, equivalent to a ≈14% increase in human sample size without additional data collection. In contrast, $\mathrm{Rec}_{\lambda=1}$ consistently produces negative ESS gains. As expected, the most biased methods deliver the largest ESS gains, a direct manifestation of the bias–variance tradeoff. Based on these findings, we focus on $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ in all subsequent analyses.

ESS gains shrink as the number of human responses grow. We next vary the number of human responses used in rectification $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$, $n_{\mathrm{human}} \in \{50, 100, 150, 200\}$, and track ESS (Fig. 4, Appendix A.2). Similarly, `Persona-guided` maintains the largest ESS gains across settings, while all methods show an expected decline in ESS gain as n_{human} increases (the human-only estimator improves). In practice, correction is most beneficial when n_{human} is limited and the remaining pool N is large. A reasonable stopping rule in terms of variance is to discontinue using synthetic responses once Eq. (3) is not satisfied.

5.2 Subgroup effects and allocation strategies

Rectification reduces bias per subgroup. A key concern is whether rectification reduces error consistently across sub-populations, rather than only at the population level. To test this, we evaluate subgroup error on NHANES before and after applying global $\mathrm{Rec}_{\lambda_{\mathrm{opt}}}$ rectification to a fine-tuned model (Table 3). Specifically, we re-center each response around the estimate $\theta_{\mathrm{Rec}_{\lambda_{\mathrm{opt}}}}$ and compare subgroup-level bias. Most subgroups (6 of 7) show reductions (e.g., Mexican-Americans: –37%; income \$35–45k: –49%), though bias increases for female respondents (+58%). Thus, while population-level re-centering renders such heterogeneity expected, careful validation is required to ensure no subgroup suffers increased bias. Nevertheless, the overall trend suggests that rectification can mitigate performance disparities across groups.

Rectification for bias, fine-tuning for efficiency. Since training-based methods achieve the best rectified performance, a natural question arises: with a limited human set, should responses be allocated to fine-tuning the synthesis model or reserved for

Method	Bias (%)↓				ESS Gain (%) ↑			
	NHANES	ATP Q1	ATP Q2	Avg.	NHANES	ATP Q1	ATP Q2	Avg.
`Baseline`	7.61	2.30	62.41	24.11	†	†	†	†
Synthesize only								
`Domain-FT | None`	7.03	34.27	62.69	**34.66**	†	†	†	†
`SubPOP-FT | None`	180.78	44.82	33.08	86.23	†	†	†	†
`Demo-only | None`	88.10	47.53	31.18	55.60	†	†	†	†
`Persona-guided | None`	82.08	50.22	19.99	50.76	†	†	†	†
Synthesize + Rectify								
`Domain-FT |` $\text{Rec}_{\lambda=1}$	9.41	8.46	4.11	**7.33**	-32.95	-16.97	-62.23	-37.38
`SubPOP-FT |` $\text{Rec}_{\lambda=1}$	37.57	8.74	10.98	19.10	-72.91	-13.57	-41.69	-42.72
`Demo-only |` $\text{Rec}_{\lambda=1}$	13.38	16.94	3.12	11.15	-50.51	-22.27	-24.05	**-32.28**
`Persona-guided |` $\text{Rec}_{\lambda=1}$	12.32	11.10	4.50	9.31	-60.61	-20.47	-28.34	-36.47
`Domain-FT |` $\text{Rec}_{\lambda_{\text{opt}}}$	3.33	1.73	3.40	**2.82**	2.81	1.32	1.01	1.71
`SubPOP-FT |` $\text{Rec}_{\lambda_{\text{opt}}}$	3.23	4.06	3.07	3.45	1.28	1.74	1.85	1.34
`Demo-only |` $\text{Rec}_{\lambda_{\text{opt}}}$	1.75	6.75	4.52	4.34	6.07	4.83	2.62	3.97
`Persona-guided |` $\text{Rec}_{\lambda_{\text{opt}}}$	3.99	4.42	7.65	5.35	14.19	5.57	1.01	**6.92**

Table 2: **Bias (%) ↓ and effective sample size (ESS) gain (%) ↑ on three datasets**. Top block: unrectified LLM synthesis (*Synthesize only*). Bottom block: synthesis combined with rectification. `X | Y` indicates synthesis method `X` with rectification method `Y`. All results are computed at a human sample size of $|\mathcal{H}| = n_{\text{human}}$=100. Green indicates 95% bootstrap CIs where bias includes 0 and ESS > 0; averages are macro-averages across datasets. † = ESS not reported because nominal coverage (0.95) is not achieved for unrectified methods. For $\text{Rec}_{\lambda_{\text{opt}}}$, λ was automatically selected with average values of $\lambda \approx 0.15$ (NHANES), 0.3 (ATP Q1), and 0.05 (ATP Q2).

Subgroup	n	Bias (FT only) ↓	Bias (FT | $\text{Rec}_{\lambda_{\text{opt}}}$) ↓	Abs. Δ ↑	Rel. Δ (%) ↑
sex: Female	3608	4.62	7.32	-2.70	-58.55
sex: Male	3419	13.86	8.72	5.14	37.08
race: Non-Hispanic White	2342	7.93	4.38	3.54	44.71
race: Non-Hispanic Black	1525	5.79	3.45	2.34	40.37
race: Mexican American	1312	7.77	4.83	2.94	37.87
household income: $35,000 to $44,999	717	8.34	4.25	4.09	49.07
household income: $100,000 and over	1182	8.34	5.54	2.80	33.61

Table 3: **Subgroup bias changes on NHANES.** We compare fine-tuned (FT) models before and after global rectification ($\text{Rec}_{\lambda_{\text{opt}}}$). Green = subgroup bias decreases (improvement); red = subgroup bias increases (deterioration). Fixed human sample size of n_{human}=100. Note: NHANES records "sex" as a self-reported variable (male/female).

post-hoc correction[2]? To study this trade-off, we fix a budget of 1,000 human responses and evaluate allocations between `Domain-FT` and $\text{Rec}_{\lambda_{\text{opt}}}$ in proportions of 10–90, 20–80, 40–60, 60–40, and 80–20. Experiments are run on NHANES, as it offers a larger sample size than ATP [3].

Figure 2 shows that bias is lowest when 20% of responses are allocated to fine-tuning and the remainder to correction (panel a). Larger allocations (40–80%) yield greater efficiency gains (panel b) but at the cost of higher bias and uncertainty. The Pareto frontier (panel c) summarizes this trade-off: points toward the upper left represent more favorable combinations. To illustrate, we group strategies into three regimes (panel d): *Conservative* ($\leq$200 responses for FT), *Balanced* ($\approx$400 for FT), and *Aggressive* ($\geq$600 for FT). These regimes show how different allocation rules trace distinct positions along the frontier, offering interpretable levers for balancing bias and efficiency. We note that these findings are specific to the datasets and questions studied here, and should be treated as directional guidance rather than universal rules.

6 Discussion

Across three longitudinal surveys, we find that all adaptation strategies reduce LLM bias once rectified, with `Domain-FT` achieving the lowest error ($<$3% on average). At the same time, every synthesis method yields significant ESS gains after correction, showing clear improvements over a human-

[2]PPI requires held-out data; reusing fine-tuning data violates its guarantees (Angelopoulos et al., 2023a).

[3]This analysis uses a different data split than Table 2.

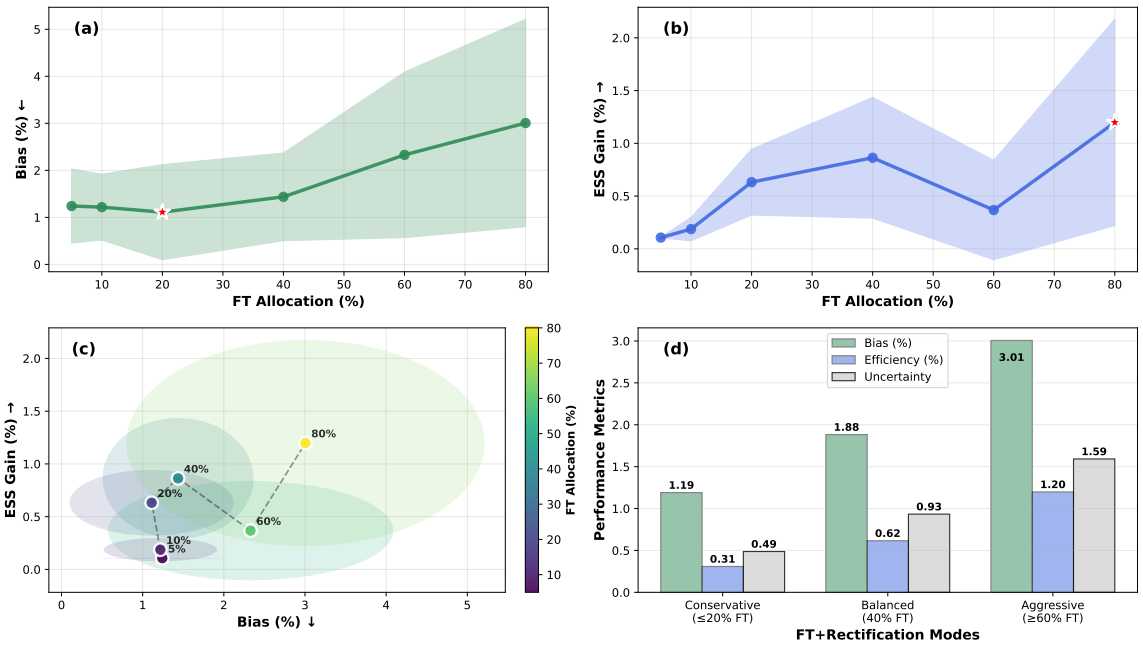

Figure 2: **Fine-tuning and rectification interaction analysis.** Results are averaged over 100 independent runs. (a) Bias vs. FT allocation with confidence bands; 20% allocation minimizes bias (red star). (b) Efficiency peaks at 80% FT allocation but with high uncertainty. (c) Pareto frontier with 95% confidence ellipses where points toward the upper left reflect better trade-offs. (d) Comparison across allocation policies.

only estimator.

Augmenting surveys with LLM responses is therefore not a universal solution but a set of context-dependent choices guided by the analyst's goals. Accordingly, we present our recommendations in Table 4 as practical guidelines. These guidelines are supported by experiments in nutrition, economics, and politics within the U.S., where LLMs show reliable factual grounding. While we expect the main patterns to hold broadly, generalizing to other domains requires awareness of potential differences in performance.

Overall, rectification improves efficiency when a large demographic frame is available and model predictions are reasonably accurate (Eq. 3). Power-tuning further enhances efficiency by avoiding negative ESS outcomes, unlike $\text{Rec}_{\lambda=1}$. Subgroup analysis suggests that it can reduce disparities, but warrants further validation. To contextualize what "reasonably accurate" means in this context, we examine the correlation between available covariates and the target answers. For NHANES, correlations range from 0.006 to 0.220 (mean 0.043), with Gender providing the strongest signal ($r = 0.22$). For ATP Q1 (economics), correlations range from 0.004 to 0.232 (mean 0.058), driven most strongly by Income ($r = 0.23$). For ATP Q2 (politics), correlations range from 0.002 to 0.506 (mean 0.073), with Party providing the strongest signal ($r = 0.51$). Per Mani et al. (2025), PPI++ carries a finite-sample cost when pseudo-label correlation with true labels is low, so gains are not guaranteed. The positive ESS gains we observe suggest the correlation is sufficient in our setting, though this may not hold in all contexts.

Moving forward, promising directions include developing subgroup-aware correction methods, extending evaluations to multilingual and cross-cultural contexts, and testing live deployments within survey infrastructure. Beyond surveys, our findings exemplify how LLM predictions and limited human data can be combined for valid inference in other settings where labeled data are scarce.

7 Conclusion

We presented the first head-to-head evaluation of training-time, inference-time, and post-hoc adaptations for LLM-assisted survey simulation across three longitudinal datasets. Our results show that uncorrected LLM synthesis is consistently biased, but rectification with a small set of human responses reduces bias to below 5% while achieving positive ESS gains. Training-based methods paired with rectification (λ_{opt}) provide the most reliable estimates, while inference-time prompting strategies deliver greater ESS gains at the cost of higher bias.

Objective	Practical recommendations
A. Minimizing the error of the population estimate	• Reserve the majority of the labeling budget for post-hoc correction. • Synthesize with `Domain-FT` (or `SubPOP-FT` if target histories are unavailable).
B. Minimizing the number of needed human responses	• Similarly allocate the majority share to correction. • Afterwards, generate responses with `Persona-guided` prompting for the largest ESS gains.
C. Inference in very low data or compute regimes	• When neither historical data nor fine-tuning compute are available, apply `Demo-only` synthesis plus correction. • Bias collapses below 5% and ESS improves modestly.
D. Having the best synthesis model for follow-up	• If the end-goal is a single best-performing synthesis model (e.g., for follow-up questions or downstream quantities beyond the corrected estimate), skip prompting and `fine-tune` directly on available responses.

Table 4: **Evidence-based guidelines for LLM-assisted survey simulation.**

In our $N{=}1000$ experiments, allocating a majority of responses (60–80%) to correction gave the best trade-off between bias and efficiency, though the precise percentages will vary with data and budget.

Limitations

We identify key limitations of our work. First, our evaluation is restricted to the U.S. context, drawing on NHANES and ATP data with all interactions conducted in English. Performance is likely to vary across languages, cultural norms, and survey methodologies (Shi et al., 2024; Ziegenfuss et al., 2021). Extending benchmarks to multilingual and non-Western contexts is therefore essential before drawing global conclusions.

Second, our experiments rely on simple non-adaptive correction methods. Although accessible due to their simplicity and theoretical guarantees, alternative methods such as confidence-driven inference (CDI) (Gligorić et al., 2024) may offer stronger performance in practice. In particular, adaptive procedures that re-weight based on model confidence could deliver larger efficiency gains, especially when prediction reliability varies across subgroups. However, adaptive approaches during data collection (rather than post hoc) are less accessible to practitioners and require careful validation of the sampling rule. Future work should examine how practitioners balance data-collection simplicity against potential efficiency gains.

Third, rectification is effective for population-level estimation but does not solve the harder problem of individual-level simulation. Accurately reproducing a single respondent's answers requires capturing idiosyncratic confounders and latent traits (Shaikh et al., 2025; Belyaeva et al., 2023), which current methods struggle to represent. We explicitly show the difficulty of this problem through the results in Appendix A.1. Progress here will likely require richer behavioral models and new data sources. Moreover, current rectification methods operate as numerical adjustments, while the actual responses themselves remain biased (i.e., we do not directly correct the LLM outputs). Future work could explore approaches that jointly improve both the statistical estimates and the generated responses.

Fourth, we assume simple random sampling and treat survey responses as ground truth. Real surveys use complex designs and suffer from non-response and selection biases, so real-world deployment would require integrating proper survey weights. We also note that "open-ended" responses in NHANES are structured (e.g., a quantity can always be derived); they are not free-text in the explanatory sense.

Ethical considerations

Survey data often include sensitive personal information. If such data are processed by large language models hosted by commercial providers, questions of privacy and consent become paramount (Kalluri et al., 2025). Participants should retain meaningful control over their responses, and safeguards are needed to prevent concentration of power among technology companies (Vincent et al., 2021; Vincent and Li, 2023).

Replacing or reducing human survey participation has economic implications and potential income displacements. Surveys currently provide paid opportunities for respondents (Gray and Suri, 2019), and widespread substitution with LLM-generated data could displace this source of income (Shao et al., 2025; Tiwari, 2023). Any deployment of methods as described in our guidelines must weigh efficiency gains against potential harms

to individuals who rely on survey participation.

Large-scale use of LLMs has broader societal costs, including environmental impacts from training and inference (Wu et al., 2022; Zhong et al., 2024). The studied approach offers a partial mitigation: by rectifying predictions from existing models rather than training new ones from scratch, we reduce the need for additional large-scale model development (Lacoste et al., 2019). In this sense, our evaluation demonstrates how methodological innovation can align with more sustainable AI practices.

Lastly, subgroup bias amplification is an important ethical consideration. If the labeled data is sparse or unrepresentative, rectification can inadvertently amplify subgroup-level errors, correcting toward majority patterns while leaving minority responses systematically misestimated. This risk is particularly salient for survey applications where small subpopulations are of substantive interest. Future work should examine subgroup-aware, stratified, or adaptive rectification strategies that explicitly mitigate disparities (Fogliato et al., 2024).

References

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. 2023a. Prediction-powered inference. *Science*, 382(6671):669–674.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. 2023b. Ppi++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. LLM social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*.

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.

Shubham Atreja, Joshua Ashkinaze, Lingyao Li, Julia Mendelsohn, and Libby Hemphill. 2025. What's in a Prompt?: A Large-Scale Experiment to Assess the Impact of Prompt Design on the Compliance and Accuracy of LLM-Generated Text Annotations. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 122–145.

Christopher A Bail. 2024. Can Generative AI improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y McLean, and Nicholas A Furlotte. 2023. Multimodal llms for health grounded in individual-specific data. In *Workshop on Machine Learning for Multimodal Healthcare Data*, pages 86–102. Springer.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

David Broska, Michael Howes, and Austin van Loon. 2025. The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations. *Sociological Methods & Research*, page 00491241251326865.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025. Mixture-of-personas language models for population simulation. *arXiv preprint arXiv:2504.05019*.

Nitay Calderon, Roi Reichart, and Rotem Dror. 2025. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. *arXiv preprint arXiv:2501.10970*.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. 2025. Specializing large language models to simulate survey response distributions for global populations. *arXiv preprint arXiv:2502.07068*.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked Personas: Using Natural Language Prompts to Measure Stereotypes in Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995*.

Tim R Davidson, Viacheslav Surkov, Veniamin Veselovsky, Giuseppe Russo, Robert West, and Caglar Gulcehre. 2024. Self-recognition in language models. *arXiv preprint arXiv:2407.06946*.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. 2023. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 36:68589–68601.

Shuxian Fan, Adam Visokay, Kentaro Hoffman, Stephen Salerno, Li Liu, Jeffrey T Leek, and Tyler McCormick. From Narratives to Numbers: Valid Inference Using Language Model Predictions from Verbal Autopsies. In *First Conference on Language Modeling*.

Riccardo Fogliato, Pratik Patil, Mathew Monfort, and Pietro Perona. 2024. A Framework for Efficient Model Evaluation through Stratification, Sampling, and Estimation. In *European Conference on Computer Vision*, pages 140–158. Springer.

National Center for Health Statistics and 1 others. 2017. National Health and Nutrition Examination Survey 2015-2016. *Centers for Disease Control and Prevention*.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1–24.

Mingmeng Geng, Sihong He, and Roberto Trotta. 2024. Are large language models chameleons? an attempt to simulate social surveys. *arXiv preprint arXiv:2405.19323*.

Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J Candès, and Dan Jurafsky. 2024. Can Unconfident LLM Annotations Be Used for Confident Conclusions? *arXiv preprint arXiv:2408.15204*.

Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Harper Business.

Robert M Groves, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey methodology*. John Wiley & Sons.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Pratyusha Ria Kalluri, William Agnew, Myra Cheng, Kentrell Owens, Luca Soldaini, and Abeba Birhane. 2025. Computer-vision research powers surveillance technology. *Nature*, pages 1–7.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. *arXiv preprint arXiv:2503.02080*.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert West. 2024. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. 2025. LLM Generated Persona is a Promise with a Catch. *arXiv preprint arXiv:2503.16527*.

Junyi Li, Ninareh Mehrabi, Charith Peris, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. On the steerability of large language models toward data-driven personas. *arXiv preprint arXiv:2311.04978*.

Benjamin Lira, Joseph M O'Brien, Pablo A Peña, Brian M Galla, Sidney D'Mello, David S Yeager, Amy Defnet, Tim Kautz, Kate Munkacsy, and Angela L Duckworth. 2022. Large studies reveal how reference bias limits policy applications of self-report measures. *Scientific reports*, 12(1):19189.

Pranav Mani, Peng Xu, Zachary C Lipton, and Michael Oberst. 2025. No free lunch: Non-asymptotic analysis of prediction-powered inference. *arXiv preprint arXiv:2505.20178*.

Benjamin S Manning, Kehang Zhu, and John J Horton. 2024. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pew Research Center. American Trends Panel (ATP) datasets. https://www.pewresearch.org/american-trends-panel-datasets/. Accessed September 19, 2025.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Alessandra Rister Portinari Maranca, Jihoon Chung, Musashi Hinck, Adam D Wolsky, Naoki Egami, and Brandon M Stewart. 2025. Correcting the Measurement Errors of AI-Assisted Labeling in Image Analysis Using Design-Based Supervised Learning. *Sociological Methods & Research*, page 00491241251333372.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. *arXiv preprint arXiv:2402.16786*.

Giuseppe Russo, Kristina Gligorić, Vincent Moreau, and Robert West. 2025a. Meat-free day reduces greenhouse gas emissions but poses challenges for customer retention and adherence to dietary guidelines. *arXiv preprint arXiv:2504.02899*.

Giuseppe Russo, Debora Nozza, Paul Röttger, and Dirk Hovy. 2025b. The pluralistic moral gap: Understanding judgment and value differences between humans and large language models. *arXiv preprint arXiv:2507.17216*.

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

Anand Shah, Kehang Zhu, Yanchen Jiang, Jeffrey G Wang, Arif K Dayi, John J Horton, and David C Parkes. 2025. Learning from Synthetic Labs: Language Models as Auction Participants. *arXiv preprint arXiv:2507.09083*.

Omar Shaikh, Kristina Gligoric, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296, Mexico City, Mexico. Association for Computational Linguistics.

Omar Shaikh, Shardul Sapkota, Shan Rizvi, Eric Horvitz, Joon Sung Park, Diyi Yang, and Michael S Bernstein. 2025. Creating General User Models from Computer Use. *arXiv preprint arXiv:2505.10831*.

Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. 2025. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the US Workforce. *arXiv preprint arXiv:2506.06576*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.

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*.

Huaman Sun, Jiaxin Pei, Minje Choi, and David Jurgens. 2025. Sociodemographic prompting is not yet an effective approach for simulating subjective judgments with LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 845–854.

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J Jansen, and Jang Hyun Kim. 2024. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*.

Rudra Tiwari. 2023. The impact of AI and machine learning on job displacement and employment opportunities. *International Journal of Engineering Technologies and Management Research*, 7(1):1–9.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Olivier Toubia, George Z Gui, Tianyi Peng, Daniel J Merlau, Ang Li, and Haozhe Chen. 2025. Database report: Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6):1446–1455.

Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 215–227.

Nick Vincent and Hanlin Li. 2023. ChatGPT Stole Your Work. So What Are You Going to Do. *Wired Magazine*, 20.

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.

Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2025b. Parameter-efficient fine-tuning in large language models: a survey of methodologies. *Artificial Intelligence Review*, 58(8):227.

Qianli Wang, Tatiana Anikina, Nils Feldhus, Josef Genabith, Leonhard Hennig, and Sebastian Möller. 2024a. LLMCheckup: Conversational Examination of Large Language Models via Interpretability Tools and Self-Explanations. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 89–104.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. " My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. *arXiv preprint arXiv:2402.14499*.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, and 1 others. 2022. Sustainable ai: Environmental implications, challenges and opportuities. *Proceedings of machine learning and systems*, 4:795–813.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-embed 2.0: Multilingual retrieval without compromise. *arXiv preprint arXiv:2412.04506*.

Long Zhang, Meng Zhang, Wei Lin Wang, and Yu Luo. 2025. Simulation as Reality? The Effectiveness of LLM-Generated Data in Open-ended Question Assessment. *arXiv preprint arXiv:2502.06371*.

Junhao Zhong, Yilin Zhong, Minghui Han, Tianjian Yang, and Qinghua Zhang. 2024. The impact of AI on carbon emissions: evidence from 66 countries. *Applied Economics*, 56(25):2975–2989.

Jeanette Y Ziegenfuss, Casey A Easterday, Jennifer M Dinh, Meghan M JaKa, Thomas E Kottke, and Marna Canterbury. 2021. Impact of demographic survey questions on response rate and measurement: A randomized experiment. *Survey Practice*, 14(1).

A Supplementary results and experiments

Both datasets used in our study are publicly available. The U.S. National Health and Nutrition Examination Survey (NHANES) is produced by the National Center for Health Statistics and is in the public domain. The American Trends Panel (ATP) is released by the Pew Research Center for scholarly use under its data use terms. No special licenses or permissions were required for access or use of these datasets in our work.

A.1 Individual-level simulation results

This analysis supports the discussion within the limitations section regarding the difficulty of capturing idiosyncratic behaviors. We report Mean Absolute Error (MAE) between the ground-truth scalar target $y_i = \phi(Y_{i,T})$ and the synthetic value $\hat{y}_i = \phi(\hat{Y}_{i,T})$ for each individual i in NHANES.

These results show that while LLMs can be effectively calibrated to produce unbiased population estimates, accurately predicting an individual's response remains fundamentally difficult. From the LLM's perspective, individuals sharing identical demographic profiles appear indistinguishable, as illustrated in Figure 3. The model does not have enough information to differentiate between a 67-year-old white female who consumes 800 kcal versus another with identical observable characteristics who consumes 2,200 kcal.

Several factors contribute to this issue. First, *identical observable features* mean that LLMs only observe coarse demographic categories, missing subtle but crucial individual differences in metabolism, food preferences, cooking skills, or economic circumstances. Second, *natural daily variation* ensures that even the same individual has substantial day-to-day fluctuations based on work schedule, social events, mood, stress levels, and purely stochastic factors. Finally, *unobserved determinants* such as genetics, medication effects, micronutrient status, food allergies, and personal dietary history remain completely hidden from the model yet strongly influence behavior.

A.2 ESS gains with increasing n_{human}

Figure 4 reports ESS gains under $\text{Rec}_{\lambda_{\text{opt}}}$ as the number of labeled participants increases. We observe the same trend across all three datasets: gains shrink as n_{human} grows, since the human-only estimator improves with more labels.

Synthesis method	MAE (kcal) ↓
`Demo-only`	1257.2
`Persona-guided`	1417.5
`Domain-FT`	815.0
`SubPOP-FT`	920.1

Table 5: Individual-level absolute error (MAE) for daily energy intake (kcal) on NHANES. The high individual-level errors highlight the difficulty of this setting.

A.3 Ablation studies (NHANES)

Context ablation. We ablate the demographic information provided to the model at inference time by incrementally expanding the prompt (e.g., the demographic profile of the person). Each group adds a new set of attributes to the prompt. Group 1 uses only age and sex. Group 2 adds anthropometric variables (height, weight, BMI). Group 3 includes dietary preference (e.g., vegan, vegetarian). Group 4 incorporates race/ethnicity, and Group 5 adds citizenship status. Adding more context about the participant increases performance. The results are summarized in Table 6.

Prompting strategy ablation. We compare four prompting strategies that vary along two axes: (1) single-turn vs. multi-turn prompting, and (2) unconditioned vs. conditioned inputs. In single-turn prompting, the model is asked to recall the entire day's intake in one pass. Multi-turn prompting follows a structured format modeled after real dietary surveys (e.g., the Automated Multiple-Pass Method), where the model is guided through multiple passes and explicitly asked whether anything was forgotten. Conditioning improves accuracy and coverage. Interestingly, multi-turn prompting consistently leads to higher nutrient estimates, which may reflect memory probing (i.e., the model "recalling" additional foods) or acquiescence bias[4].

A.4 Results by base model

Table 8 reports bias and ESS gains separately for each base model and synthesis–rectification combination. Overall, the patterns are consistent with our main findings: unrectified synthesis remains biased and fails to achieve valid coverage (hence ESS is not reported), while applying rectification collapses bias and yields positive ESS gains in most cases.

[4] Acquiescence bias is the tendency to provide affirmative responses regardless of content. When repeatedly asked "Did you forget anything?", LLMs may generate additional foods due to the prompting structure rather than genuine recall.

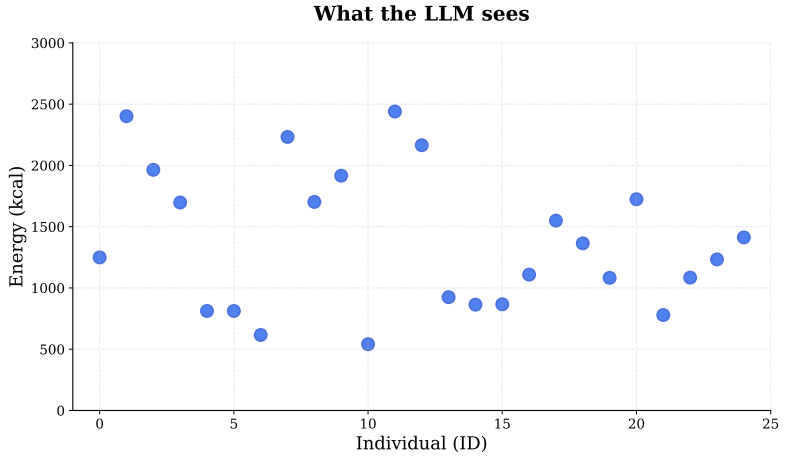

Figure 3: Individual variation in energy intake within identical demographic groups. Each point represents an individual with the same observable characteristics (age 66–70, female, 60–70kg, Non-Hispanic White, no special diet), yet their actual energy consumption varies greatly from 500 to 2,500 kcal.

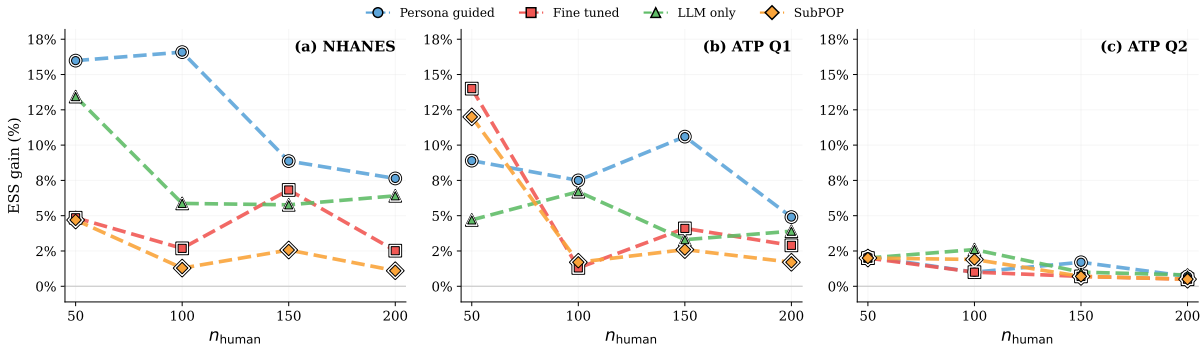

Figure 4: ESS gain under $\text{Rec}_{\lambda_{\text{opt}}}$ across labeled-sample sizes $n_{\text{human}} \in \{50, 100, 150, 200\}$.

Conditioning level	Δ from truth (%)	CI width (%)	CI coverage ↑
Basic: age, sex	13.44	13.42	0.22
+ anthropometrics	12.34	15.79	0.34
+ dietary preference	10.26	15.47	0.45
+ ethnicity	11.51	15.57	0.53
+ citizenship	8.96	14.55	0.60

Table 6: Effect of conditioning level on estimation.

Prompt format	Δ from truth (%)	CI width (%)	CI coverage ↑
Single-turn, no profile	19.84	8.37	0.16
Multi-turn, no profile	34.67	7.85	0.14
Single-turn, full profile	10.00	14.77	0.60
Multi-turn, full profile	25.31	13.71	0.27

Table 7: Comparison of prompting strategies.

Model: `Synthesis | Rectify`	Bias (%) ↓			ESS Gain (%) ↑		
	NHANES	ATP Q1	ATP Q2	NHANES	ATP Q1	ATP Q2
Unrectified synthesis (LLM-only)						
Baseline: `– | None`	7.49	2.11	59.00	†	†	†
GPT-4o-mini: `Persona-guided | None`	16.85	21.72	11.51	†	†	†
GPT-4o-mini: `Demo-only | None`	18.12	32.52	6.21	†	†	†
Llama: `Persona-guided | None`	186.06	62.67	9.49	†	†	†
Llama: `Domain-FT | None`	9.15	33.16	73.40	†	†	†
Llama: `Demo-only | None`	213.84	82.81	12.31	†	†	†
Llama: `SubPOP-FT | None`	165.58	55.85	17.02	†	†	†
Mistral: `Persona-guided | None`	90.14	52.47	28.08	†	†	†
Mistral: `Domain-FT | None`	2.75	50.98	71.90	†	†	†
Mistral: `Demo-only | None`	108.34	60.75	44.85	†	†	†
Mistral: `SubPOP-FT | None`	359.66	49.65	46.10	†	†	†
Qwen: `Persona-guided | None`	35.10	36.57	32.48	†	†	†
Qwen: `Domain-FT | None`	9.14	20.82	46.13	†	†	†
Qwen: `Demo-only | None`	11.90	14.42	36.66	†	†	†
Qwen: `SubPOP-FT | None`	17.75	28.82	36.77	†	†	†
Synthesize, then Rectify						
GPT-4o-mini: `Persona-guided | Rec`$_{\lambda_{\text{opt}}}$	3.68	4.06	4.54	17.8	15.7	3.8
GPT-4o-mini: `Persona-guided | Rec`$_{\lambda=1}$	4.73	5.71	4.63	-28.6	-23.5	-14.6
GPT-4o-mini: `Demo-only | Rec`$_{\lambda_{\text{opt}}}$	3.65	4.51	3.31	10.2	10.4	6.5
GPT-4o-mini: `Demo-only | Rec`$_{\lambda=1}$	4.19	6.17	4.09	-18.9	-20.5	-11.0
Llama: `Persona-guided | Rec`$_{\lambda_{\text{opt}}}$	3.31	5.12	4.14	10.4	10.5	1.7
Llama: `Persona-guided | Rec`$_{\lambda=1}$	12.46	5.41	4.56	-88.7	-13.8	-29.1
Llama: `Domain-FT | Rec`$_{\lambda_{\text{opt}}}$	3.89	5.26	3.75	5.5	2.3	1.0
Llama: `Domain-FT | Rec`$_{\lambda=1}$	4.80	5.89	7.05	-22.5	-23.6	-60.3
Llama: `Demo-only | Rec`$_{\lambda_{\text{opt}}}$	3.65	5.14	4.11	7.4	3.3	1.9
Llama: `Demo-only | Rec`$_{\lambda=1}$	11.05	4.82	4.62	-86.5	-4.0	-26.2
Llama: `SubPOP-FT | Rec`$_{\lambda_{\text{opt}}}$	3.91	4.80	3.85	2.7	2.4	1.2
Llama: `SubPOP-FT | Rec`$_{\lambda=1}$	9.03	6.15	6.02	-80.9	-29.3	-55.7
Mistral: `Persona-guided | Rec`$_{\lambda_{\text{opt}}}$	3.62	5.73	8.72	11.4	8.0	1.0
Mistral: `Persona-guided | Rec`$_{\lambda=1}$	8.22	5.06	9.62	-74.2	-9.6	-21.2
Mistral: `Domain-FT | Rec`$_{\lambda_{\text{opt}}}$	3.35	4.45	4.23	1.4	1.3	1.3
Mistral: `Domain-FT | Rec`$_{\lambda=1}$	4.69	5.83	7.33	-50.2	-24.9	-65.8
Mistral: `Demo-only | Rec`$_{\lambda_{\text{opt}}}$	3.62	5.34	3.99	3.6	2.5	1.1
Mistral: `Demo-only | Rec`$_{\lambda=1}$	8.88	5.58	5.83	-75.4	-13.7	-48.9
Mistral: `SubPOP-FT | Rec`$_{\lambda_{\text{opt}}}$	3.69	—	3.71	1.4	—	1.4
Mistral: `SubPOP-FT | Rec`$_{\lambda=1}$	39.19	5.36	7.00	-98.7	1.0	-66.8
Qwen: `Persona-guided | Rec`$_{\lambda_{\text{opt}}}$	4.13	5.09	3.53	4.9	7.6	1.3
Qwen: `Persona-guided | Rec`$_{\lambda=1}$	8.20	6.69	3.41	-65.2	-22.4	-11.0
Qwen: `Domain-FT | Rec`$_{\lambda_{\text{opt}}}$	3.95	4.37	3.91	6.5	5.9	1.0
Qwen: `Domain-FT | Rec`$_{\lambda=1}$	4.52	5.44	5.97	-14.7	-10.8	-54.8
Qwen: `Demo-only | Rec`$_{\lambda_{\text{opt}}}$	3.67	4.80	4.01	6.9	5.5	1.0
Qwen: `Demo-only | Rec`$_{\lambda=1}$	5.10	6.35	3.76	-40.5	-39.4	0.3
Qwen: `SubPOP-FT | Rec`$_{\lambda_{\text{opt}}}$	4.13	5.36	—	2.5	4.3	—
Qwen: `SubPOP-FT | Rec`$_{\lambda=1}$	5.21	6.27	3.55	-39.9	-15.3	1.0

Table 8: **Per-model results across datasets.** Bias (%) ↓ and ESS gain (%) ↑ (ESS@100) on NHANES, ATP Q1, and ATP Q2. Notation `X | Y` indicates synthesis method `X` combined with rectification method `Y`; `None` denotes unrectified synthesis. † = ESS not reported because nominal coverage (0.95) is not achieved. Prompting temperature fixed at 0.7 for all generations. Per-model metrics are bootstrap-averaged over 100 random labeled splits (seed=0). Due to bootstrap averaging, individual cell values may differ slightly from the main table. Entries marked "—" indicate cases where the model produced near-constant predictions, preventing valid PPI estimation.

A.5 Additional experiment: Twin2k-500

We replicate our analysis on Twin2k-500 (Toubia et al., 2025), where the focal question is philosophical in nature:

> *"Imagine that a group of research scientists in the School of Medicine are running a laboratory experiment on a vaccine for a rare and fatal virus. The possibility of actually contracting the disease from the vaccine is 1 in 1,000, but once you have the disease there is no known cure. The scientists are seeking volunteers to test the vaccine on. What is the lowest amount (in dollars) that you would have to be paid before you would take part in this experiment?"*

Table 9 reports bias and ESS gain at $n_{\text{human}} = 100$. The pattern from our main results holds: unrectified synthesis yields large bias (60–95%) and fails to achieve nominal coverage, while combining any synthesis method with $\text{Rec}_{\lambda_{\text{opt}}}$ reduces bias below 10% and restores valid coverage. However, ESS gains are modest (0.8–1.8%), in contrast to the main datasets. We attribute this to the nature of the question: willingness-to-pay under a hypothetical risk scenario is unlikely to be well-predicted from demographics alone, and LLM predictions in this setting may carry little correlation with actual responses.

Method	Bias (%) ↓	ESS Gain (%) ↑
`Baseline (prev. day)`	7.80	583.77
Synthesize only		
`Demo-only | None`	80.91	†
`Persona-guided | None`	86.62	†
`Domain-FT | None`	59.71	†
`SubPOP-FT | None`	95.20	†
Synthesize + Rectify		
`Demo-only | Rec`$_{\lambda_{\text{opt}}}$	8.06	1.76
`Persona-guided | Rec`$_{\lambda_{\text{opt}}}$	9.88	1.41
`Domain-FT | Rec`$_{\lambda_{\text{opt}}}$	7.32	1.83
`SubPOP-FT | Rec`$_{\lambda_{\text{opt}}}$	9.64	0.83

Table 9: **Bias (%) and ESS gain (%) on Twin2k-500** at n_{human}=100. † = ESS not reported; nominal coverage not achieved. The strong previous-day baseline reflects high test-retest correlation.

B Methods

B.1 Prediction-Powered Inference (PPI)

Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023a) provides valid confidence intervals for statistical estimands by combining a small labeled dataset with predictions from a machine-learning system on a large unlabeled dataset. Let $\{(X_i, Y_i)\}_{i=1}^n$ denote the labeled data and $\{\tilde{X}_i\}_{i=1}^N$ denote unlabeled covariates with $N \gg n$, both drawn i.i.d. from the same distribution. The predictor f is trained independently of both datasets. PPI applies to estimands θ^* defined as solutions to convex optimization problems of the form:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}[\ell_\theta(X, Y)],$$

where ℓ_θ is a convex loss function. Under mild regularity conditions, θ^* can be characterized by the estimating equation $\mathbb{E}[g_{\theta^*}(X, Y)] = 0$, where $g_\theta = \nabla_\theta \ell_\theta$ is a subgradient of the loss. PPI constructs an estimator by combining two components. The *imputed gradient* uses unlabeled data and model predictions:

$$\hat{g}_\theta^f = \frac{1}{N} \sum_{i=1}^{N} g_\theta\big(\tilde{X}_i, f(\tilde{X}_i)\big).$$

The *rectifier* uses labeled data to correct for prediction bias:

$$\widehat{\Delta}_\theta = \frac{1}{n} \sum_{i=1}^{n} \Big(g_\theta(X_i, Y_i) - g_\theta(X_i, f(X_i)) \Big).$$

The PPI estimator $\hat{\theta}_{\text{PPI}}$ solves:

$$\hat{g}_\theta^f + \widehat{\Delta}_\theta = 0.$$

This estimator is unbiased by construction:

$$\begin{aligned} \mathbb{E}[\hat{g}_{\theta^*}^f + \widehat{\Delta}_{\theta^*}] &= \mathbb{E}[g_{\theta^*}(X, f(X))] \\ &\quad + \mathbb{E}[g_{\theta^*}(X, Y) - g_{\theta^*}(X, f(X))] \\ &= \mathbb{E}[g_{\theta^*}(X, Y)] = 0. \end{aligned}$$

Power tuning. PPI can be extended to include a power tuning parameter $\lambda \in [0, 1]$ that controls the relative weight given to predictions versus labeled data. The rectifier can be scaled as $\lambda\widehat{\Delta}_\theta$, yielding the modified estimating equation $\hat{g}_\theta^f + \lambda\widehat{\Delta}_\theta = 0$. When $\lambda = 1$, this recovers the standard PPI estimator; when $\lambda = 0$, it reduces to pure imputation using predictions; intermediate values interpolate between these extremes. The parameter λ can be chosen to optimize statistical power while maintaining validity, though the default choice $\lambda = 1$ provides valid inference without tuning.

Example: population mean estimation. For estimating the population mean $\theta^* = \mathbb{E}[Y]$, the PPI estimator takes the form:

$$\hat{\theta}_{\text{PPI}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i)}_{\text{prediction average}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))}_{\text{bias correction}}$$

Under the assumption that f is independent of the inference data and both samples are i.i.d. from the same distribution, unbiasedness follows:

$$\mathbb{E}[\hat{\theta}_{\text{PPI}}] = \mathbb{E}[f(X)] + \mathbb{E}[Y - f(X)] = \mathbb{E}[Y] = \theta^*$$

B.2 Mapping $\phi : \mathcal{Y} \to \mathbb{R}$

NHANES. We use the USDA FNDDS database to obtain nutrient profiles and define a mapping function ϕ that returns a *scalar* target from open-ended text (e.g., energy in kcal for a food mention). For each simulated food item, we first retrieve the top 40 candidate foods by cosine similarity over FNDDS (using `snowflake-arctic-embed` (Yu et al., 2024)). These candidates and the query are then passed to GPT-4o-mini with temperature 0 in a RAG-style setup (Lewis et al., 2020; Russo et al., 2025a) to select the best match and we retrieve its kcal density. Using the retrieved kcal density and the reported grams for each food item, we compute the total daily caloric intake. Importantly, this mapping function must remain deterministic so as not to introduce unnecessary variance into the correction step.

ATP. For ATP, ϕ maps ordinal multiple choice responses to numerical values. Specifically, we assign A↦1, B↦2, C↦3, D↦4, and so forth.

B.3 Prompts

NHANES. The prompt for simulating responses for NHANES is included in the box below.

ATP. The prompt for simulating responses for ATP questions is included in the boxes below.

Persona-guided prompt template. To generate detailed persona descriptions from demographic and survey response data, we used the following prompt template:

NHANES prompt template

System message:

You are participating in a dietary recall survey. Describe what you ate and drank in the past 24 hours, based on your memory. Below is your personal profile:
`{demographic_profile}`

Answer honestly and realistically as you are recalling from memory.

User prompt:

Please list everything you ate and drank in the past 24 hours. Include meals, snacks, drinks, and small bites, in the order you consumed them.

Use this exact format on each line:
`[Food name] - [Short description] - [Grams as a number only]`

Instructions:

- One item per line.
- Use a single hyphen and space (`" - "`) to separate fields.
- Grams must be a number only, no units like “g” or “grams”.
- Do not add summaries or explanations—only the list.

ATP prompt template

System message: You are a survey respondent. Adopt this profile: `{demographic_profile}`.

Answer as yourself in first person. Pick exactly one option from the list. Output only one uppercase letter from {letter choices}. No words, no punctuation, no explanations, no qualifiers. Do not discuss ethics, study design, or what most people would do.

User prompt: `{survey_question}`
Choices: `{multiple_choice_options}`

Estimand	Prediction-based score $\hat{g}_\theta^{\text{pred}}$	Rectifier $\widehat{\Delta}_\theta$	Procedure
Mean	$\theta - \frac{1}{N}\sum_{i=1}^{N} f(\tilde{X}_i)$	$\frac{1}{n}\sum_{i=1}^{n}\big(f(X_i) - Y_i\big)$	Alg. 1
Median ($q = \frac{1}{2}$)	$\frac{1}{2} - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{f(\tilde{X}_i) \le \theta\}$	$\frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}\{f(X_i) \le \theta\} - \mathbf{1}\{Y_i \le \theta\}\big)$	Alg. 2
q-quantile	$q - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{f(\tilde{X}_i) \le \theta\}$	$\frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{1}\{f(X_i) \le \theta\} - \mathbf{1}\{Y_i \le \theta\}\big)$	Alg. 2
Logistic reg.	$\frac{1}{N}\sum_{i=1}^{N} \tilde{X}_i\Big(\sigma(\theta^\top \tilde{X}_i) - f(\tilde{X}_i)\Big)$	$\frac{1}{n}\sum_{i=1}^{n} X_i(f(X_i) - Y_i)$	Alg. 3
Linear reg.	$\frac{1}{N}\sum_{i=1}^{N}\big(\tilde{X}_i\tilde{X}_i^\top\theta - \tilde{X}_i f(\tilde{X}_i)\big)$	$\frac{1}{n}\sum_{i=1}^{n} X_i(f(X_i) - Y_i)$	Alg. 4
Convex minimizer	$\frac{1}{N}\sum_{i=1}^{N} \nabla\ell_\theta(\tilde{X}_i, f(\tilde{X}_i))$	$\frac{1}{n}\sum_{i=1}^{n}\big(\nabla\ell_\theta(X_i, Y_i) - \nabla\ell_\theta(X_i, f(X_i))\big)$	Alg. 5

Table 10: Prediction-powered estimating equations for common estimands. Here σ denotes the logistic sigmoid.

`Persona-guided` prompt for synthesizing personas based on historical data

System message: You are an expert at synthesizing detailed persona descriptions from demographic and behavioral data. Your output must be a single, coherent paragraph.
User prompt: Here is a person's demographic information and a log of what they answered in the past.
Demographics `{demo}`
History `{context}`
Your task: Based on this information, write a single, detailed paragraph describing this person's habits, lifestyle, and personality in general. Your description should be a coherent narrative that synthesizes all the available evidence. Focus on inferring patterns, routines, and constraints that are supported by the provided information. Do not use lists, bullet points, or scores. Do not include extra explanations or reasoning. Just provide the narrative persona description.

B.4 Fine-tuning details

We describe the fine-tuning procedures for `Domain-FT` and `SubPOP-FT` to support reproducibility.

Domain-FT. Out of the box, a general-purpose language model likely has no knowledge of a specific survey's question wording, response scale, or the population of respondents. Fine-tuning addresses this by continuing to train the model on historical responses from the target survey, so that it learns to produce answers that match the style and distribution of that particular instrument and population.

Concretely, `Domain-FT` performs full-parameter supervised fine-tuning (meaning all model weights are updated) on the target survey using `axolotl`[5]. Each training example is formatted as a two-turn conversation: a system message containing the participant's demographic profile, and a user message containing the survey question with labeled answer choices (see Appendix B.3 for the full prompt templates). The model is trained to predict the human-provided answer, and is penalized via the standard cross-entropy loss when its predicted response differs from it. Training uses waves $t \in \{1, \ldots, T-1\}$ as the fine-tuning corpus; wave T is held out entirely. A 10% held-out split of the training data is used for validation.

We fine-tune three open-weight instruction-tuned models: Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen2.5-7B-Instruct. All three models share identical training hyperparameters: learning rate 10^{-5} with a cosine decay schedule, 800 warmup steps, 1 epoch, weight decay 0.001, gradient checkpointing, and a maximum sequence length of 1,024 tokens. The only difference between the NHANES and ATP configurations is the micro-batch size (1 vs. 3, respectively), due to the longer response sequences in NHANES. All runs use a fixed random seed of 42. Training was performed on a single NVIDIA A100 GPU.

SubPOP-FT. Rather than adapting to a specific target survey, `SubPOP-FT` (Suh et al., 2025) fine-tunes the model on an external dataset of 3,229 survey questions drawn from ATP waves, each paired with empirical response distributions across 22 demographic subpopulations. The intuition is that exposure to a large, diverse set of opinion questions teaches the model to better reflect how different demographic groups tend to respond, even on surveys it has never seen.

[5] https://axolotl.ai/

The training objective differs from standard SFT. Instead of learning from individual (question, response) pairs, the model is trained to match the empirical distribution of human answers across subpopulations via a KL divergence loss:

$$\mathcal{L}_{\mathrm{KL}}(q, g) = \mathrm{KL}\Big(p_{q,g}^{\mathrm{human}} \,\Big\|\, p_{\theta}^{\mathrm{model}}(\cdot \mid c_{q,g})\Big)\,, \tag{6}$$

where $p_{q,g}^{\mathrm{human}}$ is the empirical response distribution of subgroup g on question q, and $p_{\theta}^{\mathrm{model}}$ is the model's predicted distribution over answer tokens given the subgroup-conditioned prompt $c_{q,g}$. We use the original SubPOP training code and configuration, as released by (Suh et al., 2025). It is important to note that our evaluation questions are drawn from ATP (waves 146-149), which are disjoint from the SubPOP training corpus (waves 61-132), thus eliminating any risk of data leakage.

C Data and computational resources

C.1 Dataset licensing

Both datasets used in our study are publicly available. The U.S. National Health and Nutrition Examination Survey (NHANES) is produced by the National Center for Health Statistics and is in the public domain. The American Trends Panel (ATP) is released by the Pew Research Center for scholarly use under its data use terms. No special licenses or permissions were required for access or use of these datasets in our work.

C.2 Computational resources

All experiments were conducted using models with approximately 7-8 billion parameters. Training and evaluation were performed on a single NVIDIA A100 GPUs. The total training duration across all runs was approximately 3 days (≈72 GPU hours). This includes all fine-tuning, evaluation, and validation steps.