Beyond the Mean: Three-Axis Fidelity for Aligning LLM-Based Survey Simulators from Small Pilot Data

Eun Cheol Choi [1] **Youngrae Kim** [1] **Prabhu Pugalenthi** [1] **Hong-En Chen** [1] **Bo-Ruei Huang** [1]

Abstract

Large language models (LLMs) are increasingly used to simulate social survey responses, yet their outputs exhibit systematic biases: marginal distributions are skewed, response variance is poorly calibrated, and predictor–outcome relationships are attenuated. We ask a simple question: given a small pilot sample of human responses, can an LLM recover the statistical characteristics of a broader population? We decompose recovery along three axes: structural fidelity, marginal fidelity, and individual fidelity. Using a COVID-19 misinformation survey as a case study, we benchmark three families of approaches: prompting, rectification, and fine-tuning. The findings suggest that fine-tuning on small pilot samples offers a balanced approach for achieving multiple forms of fidelity, but the levels of such fidelity can vary across subsamples, potentially threatening pluralistic alignment.

1. Introduction

Large language models (LLMs) are now routinely used as proxies for human respondents – ranging from *silicon samples* of voters (Argyle et al., 2023) to interview-grounded "generative agents" that simulate 1,000 real Americans (Park et al., 2024), to fine-tuned models that predict experiment-level outcomes (Kolluri et al., 2025) and prompted GPT-4 used to forecast effect sizes from social-science experiments (Hewitt et al., 2024). In parallel, social surveys remain the dominant instrument for measuring beliefs and attitudes, but recruiting representative respondents is expensive and slow, so most large surveys begin with a small *pilot* sample to gauge feasibility (Van Teijlingen & Hundley, 2001). This raises a concrete computational question: *can an LLM, given a small human pilot, recover the statistical structure of the full population it was drawn from?*

This question is especially important in the domain of *misinformation belief*, where (i) populations differ in which false claims they have even *encountered* (Lee et al., 2023); (ii) accuracy ratings depend on local information ecosystems that pretraining may not fully capture (Choi et al., 2026); and (iii) downstream uses (e.g. targeted interventions based on certain psychosocial features) rely on *predictor–outcome* relations, and not just on marginal distributions. Recent audits show that LLM-generated survey responses (i) approximate marginals but flatten variance, with effect-size signs flipping in roughly a third of cases (Bisbee et al., 2024); (ii) exhibit topic-specific "machine bias" that is socially inconsistent across topics (Boelaert et al., 2025); (iii) homogenize and structurally distort minority groups (Li et al., 2025; Wang et al., 2025); (iv) are highly sensitive to seemingly innocuous prompt perturbations (Rupprecht et al., 2025; Tjuatja et al., 2024); and (v) display large analytic flexibility (Cummins, 2025). Together, these pathologies suggest that LLMs should not be treated as drop-in replacements for human respondents, but rather as conditional generative or estimation models whose outputs must be evaluated against ground-truth data.

We make four contributions. **First**, we reframe LLM-based survey simulation as a *recoverability* task – given a small pilot D_p of human responses, how much of a held-out population's structure can be reconstructed – and decompose recovery along three axes: *structural* fidelity (do predictor–outcome relations match?), *marginal* fidelity (do simulated marginals match the human ones?), and *individual* fidelity (does each simulated respondent track their human counterpart?). **Second**, we introduce a calibrated evaluation protocol that maps each fidelity axis to a specific metric: Lin's Concordance Correlation Coefficient (CCC), decomposed into a sign-agreement rate and a magnitude ratio, for the structural axis; cross-respondent Earth Mover's Distance (EMD) on per-respondent scalar summaries for the marginal axis; and paired Pearson r_d (relative) and MAE_d (absolute) on per-respondent discernment for the individual axis (Sec. 3.4). **Third**, on the same 5% pilot of an N=1,466 COVID-19 misinformation survey, we run a head-to-head

All authors contributed equally to this work. [1]University of Southern California, CA, USA. Correspondence to: Eun Cheol Choi <euncheol@usc.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

comparison of {ZS, FS}×{batch, per-item} prompting, PPI rectification (Angelopoulos et al., 2023) applied uniformly to every prompt-based simulator, and LoRA / LoRA + MLP fine-tuning. **Fourth**, we audit fidelity *within* each demographic subgroups: statistical recovery degrades for certain types of identity subgroups, calling for the attention of the pluralistic alignment community.

2. Related Work

We organize prior work into calibration families relevant to the small-pilot setting: prompt-based conditioning, test-time intervention and statistical rectification, and parameter-efficient fine-tuning.

2.1. Prompt-Based Conditioning

The earliest line of work conditions LLMs on textual descriptions of demographics, attitudes, or context. *Silicon Sampling* (Argyle et al., 2023) showed that GPT-3, prompted with ANES-style backstories, can approximate group-level voting distributions, and *Random Silicon Sampling* (Sun et al., 2024) extended this to demographic role-playing using group-level marginals. *LLM-Mirror* (Kim et al., 2024) added pre-existing responses and psychological traits, and other richer schemes (Choi et al., 2026) incorporate social-network and peer features. Park et al. (2024) uses full two-hour interview transcripts as the persona and recovers 85% of human test–retest accuracy on the GSS. *Audience Segmentation* (Qin et al., 2026) restores within-group heterogeneity by varying identifier granularity. The common assumption is that the LLM's *Universal Prior* – general world knowledge acquired during pretraining – is rich enough that, conditioned on the right description, the model produces calibrated human-like responses.

A complementary literature audit tests that assumption and finds it fragile. Bisbee et al. (2024) shows that ChatGPT-generated feeling thermometers compress variance and flip effect-size signs in ∼32% of cases on ANES items; Boelaert et al. (2025) documents opinion-poll "machine bias" that is socially inconsistent across topics; Wang et al. (2025) shows that identity-prompted LLMs harmfully misportray and flatten minority groups; Li et al. (2025) formalizes this as "Das Man" homogenization driven by accuracy-maximizing decoding. Zhou et al. (2025) report that, even with repeated random sampling from GPT, the resulting silicon population overrepresents some demographic groups and is far more deterministic than humans on attitudinal items. Chapala et al. (2025) test prompt-based mitigations of social-desirability bias including neutral third-person reformulation and reverse-coding. Cummins (2025) stress-tests the entire pipeline, showing that 252 plausible analyst configurations yield strikingly different conclusions, motivating multiverse-style robustness checks.

A natural response to these audits is to inject a small amount of in-context examples drawn from the same population. This idea underlies recent few-shot demonstrations (Argyle et al., 2023), persona-pretest pipelines (Kim et al., 2024), and audience-segmentation prompting (Qin et al., 2026).

2.2. Test-Time Intervention and Rectification

A second family of methods accepts that prompting alone is insufficient and instead modifies elicitation or post-processing. *Semantic Similarity Rating* (Maier et al., 2025) avoids regression-to-the-mean on Likert scales by eliciting free text and projecting it into a similarity space. A statistically grounded sub-family treats LLM outputs as biased predictors and corrects them with a small gold sample. *Prediction-Powered Inference* (PPI) (Angelopoulos et al., 2023) debiases synthetic population estimates using a held-out human sample, and Krsteski et al. (2025) adapts PPI for survey simulation with a power-tuned per-item λ. We use this PPI variant as our rectification family, applied uniformly to every prompt-based simulator.

2.3. Fine-Tuning for Survey Simulation

A third, rapidly growing family fine-tunes the LLM directly on human responses. Kolluri et al. (2025) fine-tune LLaMA3-8B and Qwen2.5-14B on ∼2.9M responses from over 400,000 participants in the SocSci210 corpus, reducing prediction error on unseen experiments by 30% and 26% respectively relative to GPT-4o. Cao et al. (2025) fine-tune LLMs to match country-level WVS / Pew distributions using a first-token-probability objective and generalize to unseen questions and countries. Huang et al. (2025) introduces *Distribution Shift Alignment*, a two-stage scheme that explicitly aligns subgroup-conditional shifts and reduces required real-data volume by 53–69%.

Krsteski et al. (2025) provides the closest direct comparison of *prompting, fine-tuning, and rectification* under limited data, showing that the methods are complementary rather than substitutes. We differ in three ways: (i) our domain is COVID-19 misinformation, with predictor–outcome structure rich enough to expose multivariate failures invisible in marginal accuracy alone (Choi et al., 2026); (ii) alongside the population's marginal fidelity, we report structural-fidelity metrics that penalizes deviations from the bivariate and regression coefficients; (iii) we report how much individual predictions are correlated to each of the human participants' responses, providing a benchmark of individual fidelity as well. We also address whether the choice of *output head* – autoregressive token generation vs. a discriminative classification head – matter for the simulation fidelity.

3. Methodology

We frame the problem as a recoverability task. Given a full survey dataset D, we observe a small pilot $D_p \subset D$ and aim to generate a synthetic dataset D_s over the held-out respondents $D \setminus D_p$ such that D_s preserves the statistical properties of D.

3.1. Data

The survey is the COVID-19 misinformation belief study of Lee et al. (2023), conducted in South Korea in May 2020 (N=1,466). Each respondent i is described by a profile X_i comprising demographics (age, gender, education, income, political orientation), psychometric scales (open-mindedness, faith in intuition, need for evidence, truth-as-political, skepticism), and exposure measures (info exposure, emotional response). Each respondent answers Y_i on 36 belief items: 18 MISINFO (false claims) and 18 TRUEINFO (true claims), with 9 political and 9 scientific items in each subset. Responses use a 4-point Likert scale (*Not accurate at all*, *Not very accurate*, *Somewhat accurate*, *Very accurate*) plus a separate *Have not seen it* (HNS) option that we treat as missing throughout. The empirical HNS rate is 13.3%. We define three per-respondent scalar summaries – the MISINFO mean $m_i = \overline{y_{i,\text{Mis}}}$, the TRUEINFO mean $t_i = \overline{y_{i,\text{Tru}}}$, and a DISCERNMENT score $d_i = t_i - m_i$ – and use them throughout the structural, marginal, and individual fidelity analyses.

3.2. Pilot Sampling

We draw a 5% pilot D_p (n=74) with a fixed seed, holding out the remaining 1,392 respondents as the evaluation set. The same pilot is reused across all calibration pipelines so that differences in performance reflect differences in method rather than data.

3.3. Calibration Pipelines

We benchmark four families. All upstream simulators predict the same 36 Likert ratings on the same held-out respondents.

Family 1 – Zero-Shot Persona (ZS, ZS-perItem). For each held-out respondent i, the LLM is given X_i and asked to predict Y_i. **ZS** (batch) elicits all 36 ratings in a single prompt; **ZS-perItem** elicits them one item at a time. Comparing the two isolates the effect of cross-item conditioning independently of pilot examples, ablating one of the analytic-flexibility knobs flagged by Cummins (2025). An example prompt is presented in App. A.

Family 2 – Few-Shot Prompting (FS, FS-perItem). We randomly inject a subset (five rows) of D_p as in-context examples. Together with Family 1 these form a 2×2 factorial of {no pilot, with pilot}×{batch, per-item}.

Family 3 – Parameter-Efficient Fine-Tuning (LoRA, LoRA + MLP). We fine-tune Qwen3-8B (Yang et al., 2025) with LoRA (Hu et al., 2022) on the same 5% pilot in two configurations. **LoRA** – adapters on attention and MLP projections, autoregressively emitting the label string. **LoRA + MLP** – the same LoRA adapters plus a trained 5-way MLP classification head on the final hidden state, with cross-entropy over the four Likert classes plus a fifth class for HNS. The two configurations differ only in output head.

Family 4 – PPI Rectification. We rectify each prompt-based simulator's per-item population mean using Prediction-Powered Inference (Angelopoulos et al., 2023), following Krsteski et al. (2025)'s adaptation to survey simulation. For each item q, the rectified per-item population estimate is

$$\hat{\theta}_q = \bar{y}_{\text{pilot},q} + \lambda_q \left(\bar{\hat{y}}_{\text{held},q} - \bar{\hat{y}}_{\text{pilot},q}\right),$$

with $\lambda_q = \text{Cov}(y, \hat{y})/\text{Var}(\hat{y})$ on the pilot per item. PPI emits a scalar per item, not individual predictions, so it is evaluable only on per-subset population means (Tab. 3). For our LoRA configurations the simulator was trained on the pilot, so its in-distribution pilot predictions reduce to memorized GT and the correction term collapses; we report PPI and discuss the algebraic degeneracy in §4.2.

3.4. Evaluation Metrics

We decompose recovery into three complementary axes. Each axis answers a different question, and a method that excels on one can fail on another. Details of the bootstrapping procedure and uncertainty intervals on every headline metric are reported in App. B.

Structural fidelity. *Do predictor–outcome relations match?* We regress d_i on twelve predictors – the seven psychometric / exposure scales plus five demographics (age, gender, education, income, political orientation) – and assemble two paired vectors per method: the bivariate predictor–d_i correlations, $\{(r_k^{\text{GT}}, r_k^{\text{sim}})\}_{k=1}^{12}$, and the standardized OLS coefficients. We calculate the GT–Sim Concordance Correlation Coefficient (CCC), which penalizes deviations from the y=x line and so jointly captures direction and magnitude:

$$\rho_c = \frac{2\,\text{Cov}(x, y)}{\text{Var}(x) + \text{Var}(y) + (\bar{x} - \bar{y})^2}.$$

Alongside CCC we report a sign-agreement rate (fraction of the 12 predictors with matching sign) and a magnitude ratio $\overline{|x^{\text{sim}}|}/\overline{|x^{\text{GT}}|}$ (>1 inflates, <1 compresses), giving a direction-vs-magnitude decomposition.

Method	Structural ↑		Marginal↓	Individual	
	$\text{CCC}_{\text{biv } r}$	$\text{CCC}_{\text{OLS } \beta}$	EMD-d	r_d ↑	MAE_d ↓
ZS	0.56	0.50	0.72	0.12	0.79
ZS-perItem	0.80	0.71	0.48	0.31	0.58
FS	0.51	0.59	0.23	0.18	0.67
FS-perItem	0.61	0.63	0.41	0.27	**0.58**
LoRA	0.60	0.41	0.63	0.17	0.75
LoRA + MLP	**0.85**	**0.78**	**0.17**	**0.37**	0.61

Table 1. **Headline metrics across the three fidelity axes** on the held-out evaluation set. **Structural**: Lin's CCC on K=12 paired predictor–d_i correlations and 12 standardized OLS coefficients. **Marginal**: cross-respondent EMD on d_i. **Individual**: r_d is the cross-respondent Pearson between simulator and GT d_i; MAE_d is mean $|d_i^{\text{GT}} - d_i^{\text{sim}}|$. Bootstrapped CIs in App. B.

Marginal fidelity. *Do simulated marginals match the human ones?* For each held-out respondent we take the three scalar summaries d_i, m_i, t_i and report the Wasserstein-1 *Earth Mover's Distance* (EMD) between the simulator's and GT's cross-respondent distributions of each summary. All three EMDs are distances between the same kind of object (cross-respondent distribution of a per-respondent scalar), so they are directly comparable. We supplement the EMDs with the per-subset population means μ_{Mis} and μ_{Tru} before and after PPI rectification (Tab. 3).

Individual fidelity. *Does each simulated respondent track their human counterpart?* We summarize each respondent by the discernment scalar d_i and report two paired statistics across respondents: a relative agreement metric $r_d = \text{Pearson}(d_i^{\text{GT}}, d_i^{\text{sim}})$, which asks "do high-discernment respondents get high-discernment predictions," and an absolute agreement metric $\text{MAE}_d = \overline{|d_i^{\text{GT}} - d_i^{\text{sim}}|}$ across respondents, which asks "how far off is each respondent's predicted discernment." The two answer different questions: a simulator can rank respondents correctly while inflating magnitudes (high r_d, large MAE_d) or hit absolute values close on average without preserving the ranking (low MAE_d, low r_d). The same pair (r, MAE) is also reported on m_i and t_i in App. B.

4. Results

4.1. Structural Fidelity

Table 1 reports the headline numbers; Fig. 1 renders the structural axis as a forest plot. **LoRA + MLP** shows the highest point estimates of bivariate-r (CCC 0.85) and the OLS-β (CCC 0.78). **ZS-perItem** is the strongest prompt-only competitor (CCC 0.80 on bivariate r; 0.71 on OLS β) – with no pilot examples or fine-tuning.

Direction vs. magnitude (Tab. 2). Decomposing the direction and the magnitude of relationships among variables reveals different success/failure modes hidden behind similar headline numbers. **LoRA + MLP** and **FS-perItem** tie for the highest sign-agreement on bivariate r (0.92, 11 of 12 predictors), but differ on magnitude: FS-perItem sits closest to the y=x line while LoRA + MLP mildly inflates. At the other end, **LoRA** compresses slopes most severely, with both direction and magnitude off. The broader pattern is that high CCC requires *both* directional agreement and matched magnitude, and that fine-tuning on the pilot appears to swing magnitudes in either direction (compression for LoRA, inflation for LoRA + MLP). Some of the patterns are consistent with the structural failures (variance flattening, effect-size attenuation) documented in prior LLM-survey audits (Bisbee et al., 2024; Choi et al., 2026). Per-method sign-agreement and magnitude ratios are in Tab. 2 and App. B with bootstrapped CIs.

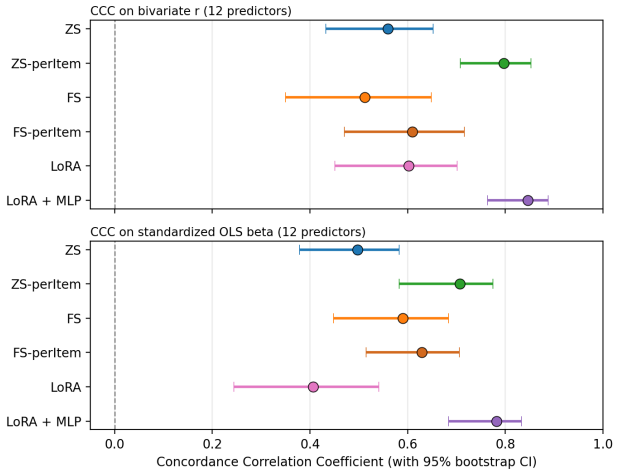

Figure 1. **Structural-fidelity forest plot** on the 12 predictors. Per method: point estimate of Lin's CCC with bootstrapped CIs (higher = better; App. B). Dashed line at 0. Top panel = bivariate r; bottom panel = standardized OLS β.

Method	Bivariate r		Standardized β	
	sign-agree	|sim|/|GT|	sign-agree	|sim|/|GT|
ZS	0.58	1.07	0.58	1.03
ZS-perItem	0.83	1.48	0.75	1.50
FS	0.58	0.56	0.50	0.63
FS-perItem	0.92	0.85	0.67	0.99
LoRA	0.50	0.50	0.50	0.58
LoRA + MLP	0.92	1.28	0.75	1.32

Table 2. **Structural-fidelity decomposition.** Sign-agreement = fraction of the K=12 predictors with matching sign across simulator and GT (higher = better). |sim|/|GT| is the ratio of mean absolute coefficients (>1 inflates, <1 compresses). Bootstrap CIs are in App. B.

4.2. Marginal Fidelity

Figure 2 plots the three cross-respondent EMDs. **LoRA + MLP** achieves the lowest EMD on all three summaries, with **FS** second. Batch **ZS** is the weakest on discernment EMD,

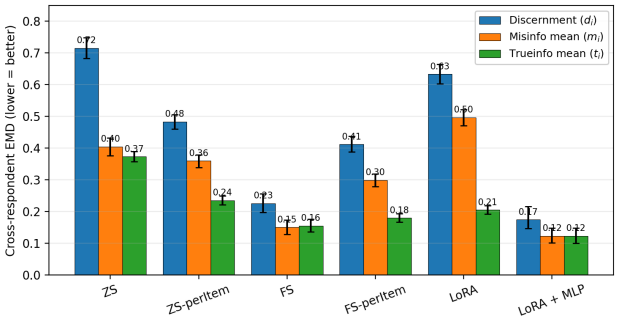

Figure 2. **Cross-respondent EMD on per-respondent scalar summaries.** Wasserstein-1 between simulator and GT distributions of, respectively, discernment d_i, Misinfo mean m_i, and Trueinfo mean t_i. Lower is better; error bars are bootstrap uncertainty intervals (App. B).

Method	Raw simulator μ_{Mis}	Raw simulator μ_{Tru}	+ PPI rectification μ_{Mis}	+ PPI rectification μ_{Tru}
GT	2.04	3.13	2.04	3.13
ZS	1.77	3.45	2.12 ↓	3.08 ↓
ZS-perItem	2.01	2.98	2.12 ↑	3.08 ↓
FS	2.06	3.28	2.11 ↑	3.09 ↓
FS-perItem	2.02	3.13	2.11 ↑	3.07 ↑
LoRA†	1.57	3.21	1.58	3.19
LoRA + MLP†	2.01	3.23	2.03	3.21

Table 3. **PPI rectification.** Per-subset population means: raw simulator output (left) and PPI-rectified estimates aggregated over the 18 items in each subset (right). ↓ marks subsets where PPI brings the estimate closer to GT than the raw simulator; ↑ marks the opposite. † marks rows where the LoRA family was trained on the pilot, so the rectifier reduces algebraically to the raw held-out mean in the case of perfect memorization (see §4.2).

tending to move the two subset means in opposite directions (under-rating Misinfo, over-rating Trueinfo) and inflating the cross-respondent variance of d_i. Full per-summary values with CIs are in App. B.

PPI rectification. Prediction-Powered Inference treats each upstream simulator's per-item population estimate as a biased predictor and rectifies it using the pilot, $\hat{\theta}_q = \bar{y}_{\text{pilot},q} + \lambda_q(\bar{\hat{y}}_{\text{held},q} - \bar{\hat{y}}_{\text{pilot},q})$, with $\lambda_q = \text{Cov}(y, \hat{y})/\text{Var}(\hat{y})$ on the pilot. In our run, PPI helps the most biased simulator (**ZS**, both subset means moving toward GT), is roughly neutral for already-calibrated methods, and can modestly hurt the Misinfo estimate of methods whose raw output is already near GT. For the LoRA family the rectifier is algebraically degenerate: when $\hat{y}_{\text{pilot}} = y_{\text{pilot}}$ (the limit of training-set memorization), $\lambda_q \equiv 1$ and $\hat{\theta}_q = \bar{\hat{y}}_{\text{held},q}$, so PPI reduces to the simulator's raw held-out mean; we mark these rows with † in Tab. 3. The reading we draw is that rectification is a *conditional* tool whose benefit depends on where the simulator's bias variance sits relative to the pilot's sampling variance. Per-item values with CIs are in App. B.

4.3. Individual Fidelity

The individual-fidelity columns of Tab. 1 report two paired statistics on the held-out respondents' discernment scalars d_i. **LoRA + MLP** wins the relative axis (r_d=0.37), with **ZS-perItem** second (r_d=0.31). The absolute axis is closer: **ZS-perItem** and **FS-perItem** attain the smallest per-respondent deviation (MAE$_d$=0.58), with **LoRA + MLP** close behind. So ZS-perItem's predicted discernment is, on average, closest to GT in absolute terms, but its respondent ranking appears looser than LoRA + MLP's. The full metrics are in App. B.

4.4. Subgroup Fidelity and the Pluralistic-Alignment Stake

Aggregate fidelity can mask systematic failure on the very subpopulations a pluralistic simulator is meant to represent; a model can misrepresent minority groups (Wang et al., 2025; Li et al., 2025). We test this directly by recomputing all three axes within demographic subgroups for the best simulator (LoRA + MLP), slicing along the subgroups based on gender, age, and political orientation (Fig. 3). As the slicing variable is itself one of the 12 predictors, it is dropped from that slice's structural regression (11 predictors); all else is unchanged.

Recovery is *not* uniform. Lin's CCC falls from 0.85 overall to 0.40 in conservative subsample (and to 0.69 in the oldest cohort), and the marginal/individual fidelity in each of the political identity subsample drops as well. The simulator thus recovers the *aggregate* discernment distribution while recovering *how* discernment is organized within some subgroups far less well, which is the risk the pluralistic alignment community warns against (full per-slice CIs in App. B).

5. Discussion

LoRA + MLP leads on both the structural and the marginal axes. LoRA + MLP is the only method in our benchmark that simultaneously (i) clears the no-recovery threshold on both bivariate-r and standardized-β CCC; (ii) attains the lowest cross-respondent EMD on all three per-respondent scalar summaries; and (iii) competitive on both individual fidelity metrics.

Output head appears to shape the simulation fidelity. LoRA and LoRA + MLP share the same backbone and LoRA rank but differ only in output head, and in our single run that design choice coincides with a drastic difference

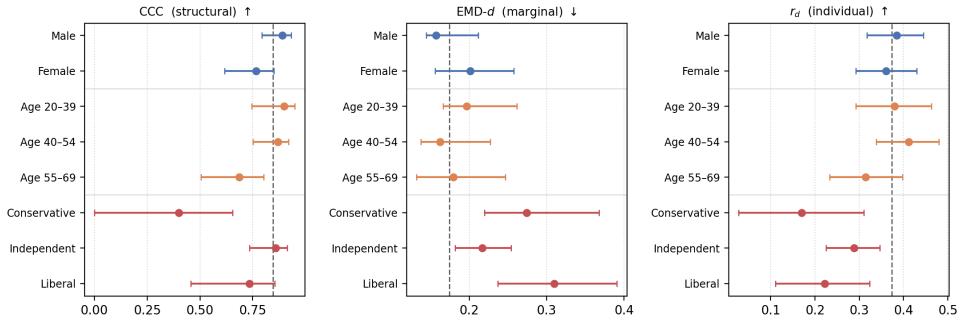

Figure 3. Subgroup fidelity of LoRA + MLP across the structural (CCC), marginal (EMD-d), and individual (r_d) axes. Each point recomputes the fidelity axis by comparing the simulator against ground truth restricted to that same subsample (e.g. sim vs. GT among Conservatives only). Bars are 95% bootstrap CIs; the dashed line marks the full-sample value.

in fidelity. The tentative lesson is not that fine-tuning fails or succeeds on its own, but that the output head appears to play a role.

PPI is conditional on bias variance and algebraically degenerate for fine-tuned simulators. PPI helps simulators that are far off but pulls already-calibrated simulators away from GT. PPI is beneficial mainly when the simulator's bias variance dominates the pilot's sampling variance. Beyond that empirical pattern, PPI has a structural problem with fine-tuned simulators that we make explicit: when $\hat{y}_{\text{pilot}} = y_{\text{pilot}}$ (the limit of training-set memorization), $\lambda_q = \text{Cov}(y, \hat{y})/\text{Var}(\hat{y}) = 1$ identically and the rectifier reduces to $\hat{\theta}_q = \bar{\hat{y}}_{\text{held},q}$, the simulator's own held-out mean. The very thing fine-tuning does (drive pilot error to zero) defeats PPI's bias-estimation step.

Fidelity may not be pluralistically uniform. Subgroup-based evaluation (Sec. 4.4) shows that recovery may not be uniform across subgroups, even though pooled fidelity is high. A simulator can pass an aggregate audit while providing more distorted results on certain types of groups (Wang et al., 2025; Li et al., 2025), so fidelity per subgroup should be closely audited in future works.

Closing the individual-fidelity gap remains an open problem. Across all methods, even the best simulator (LoRA + MLP) reaches only r_d=0.37, accounting for ∼14% of the cross-respondent variance in discernment. This suggests that, at least for this task, LLM silicon samples may not be suitable as direct substitutes for individual-level participant measurement. What level of individual fidelity is adequate and attainable for a given downstream use remains to be established by further empirical work.

Limitations

Single domain. All experiments are on one survey (COVID-19 misinformation, South Korea, May 2020). Effects of cultural and linguistic context are unmeasured, and the levels of fidelity may differ in domains where the LLM's pretraining prior is more or less aligned. **Single backbone.** LoRA results use Qwen3-8B; closed-source frontier models with stronger zero-shot priors may narrow the prompting–fine-tuning gap. **Pilot composition.** Although we use a fixed-seed pilot, results might still be sensitive to the specific draw (Cummins, 2025). **Recovery, not cognitive process.** High statistical recovery does not imply that LLM outputs reflect human cognitive processes or are appropriate for causal inference (Anthis et al., 2025; Hwang et al., 2025); we recover statistical structure, not psychological mechanism.

Future Work

The most promising next steps are: (i) a multi-domain replication; (ii) comparison among varying pilot sizes; (iii) alternative approaches such as Distribution Shift Alignment (Huang et al., 2025) as a fine-tuning loss that targets distributional rather than per-cell objectives.

Impact Statement

This paper presents work whose goal is to advance the understanding of large language models as instruments for social-science measurement. Our findings suggest that LLM-based survey simulations should not be treated as drop-in replacements for human respondents, and we provide methodology for evaluating their fidelity along multiple axes before they are used downstream. Misuse of synthetic survey data (treating uncalibrated LLM outputs as substitutes for human responses in policy decisions or intervention design) could amplify pre-existing biases or distort minority perspectives, as documented. By emphasizing recovery diagnostics across structural, marginal, and individual axes, our work aims to encourage more cautious and evaluation-driven use of LLM-based simulation in research practice.

AI tools usage disclosure

During the preparation of this work, the authors used Claude for text refinement and code review. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. Prediction-powered inference. *Science*, 382 (6671):669–674, 2023.

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Brynjolfsson, E., Evans, J., and Bernstein, M. S. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4): 401–416, 2024.

Boelaert, J., Coavoux, S., Ollion, É., Petev, I., and Präg, P. Machine bias. how do generative language models answer opinion polls? *Sociological Methods & Research*, 54(3):1156–1196, 2025.

Cao, Y., Liu, H., Arora, A., Augenstein, I., Röttger, P., and Hershcovich, D. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3141–3154, 2025.

Chapala, S., Mironov, M., and Deng, S. Mitigating social desirability bias in random silicon sampling. *arXiv preprint arXiv:2512.22725*, 2025.

Choi, E. C., Young, L., and Ferrara, E. Overstating attitudes, ignoring networks: LLM biases in simulating misinformation susceptibility. In *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 520–547, 2026.

Cummins, J. The threat of analytic flexibility in using large language models to simulate human data: A call to attention. *arXiv preprint arXiv:2509.13397*, 2025.

Hewitt, L., Ashokkumar, A., Ghezae, I., and Willer, R. Predicting results of social science experiments using large language models. *Working paper, Stanford University*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Huang, J., Li, M., and Shao, S. Distribution shift alignment helps LLMs simulate survey response distributions. *arXiv preprint arXiv:2510.21977*, 2025.

Hwang, A. H.-C., Bernstein, M. S., Sundar, S. S., Zhang, R., Horta Ribeiro, M., Lu, Y., Chang, S., Wu, T., Yang, A., Williams, D., Park, J.-s., Ognyanova, K., Xiao, Z., Shaw, A., and Shamma, D. A. Human subjects research in the age of generative AI: Opportunities and challenges of applying LLM-simulated data to HCI studies. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2025.

Kim, S., Jeong, J., Han, J. S., and Shin, D. LLM-mirror: A generated-persona approach for survey pre-testing. *arXiv preprint arXiv:2412.03162*, 2024.

Kolluri, A., Wu, S., Park, J. S., and Bernstein, M. S. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30084–30099, 2025.

Krsteski, S., Russo, G., Chang, S., West, R., and Gligorić, K. Valid survey simulations with limited human data: The roles of prompting, fine-tuning, and rectification. *arXiv preprint arXiv:2510.11408*, 2025.

Lee, S. J., Lee, C.-J., and Hwang, H. The role of deliberative cognitive styles in preventing belief in politicized COVID-19 misinformation. *Health Communication*, 38(13):2904–2914, 2023.

Li, D., Li, L., and Qiu, H. S. Chatgpt is not a man but das man: Representativeness and structural consistency of silicon samples generated by large language models. *arXiv preprint arXiv:2507.02919*, 2025.

Maier, B. F., Aslak, U., Fiaschi, L., Rismal, N., Fletcher, K., Luhmann, C. C., Dow, R., Pappas, K., and Wiecki, T. V. LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings. *arXiv preprint arXiv:2510.08338*, 2025.

Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Willer, R., Liang, P., and Bernstein, M. S. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*, 2024.

Qin, X., Li, Z., and Cheng, X. Restoring heterogeneity in LLM-based social simulation: An audience segmentation approach. *arXiv preprint arXiv:2604.06663*, 2026.

Rupprecht, J., Ahnert, G., and Strohmaier, M. Prompt perturbations reveal human-like biases in large language model survey responses. *arXiv preprint arXiv:2507.07188*, 2025.

Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B. J., and Kim, J. H. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv preprint arXiv:2402.18144*, 2024.

Tjuatja, L., Chen, V., Wu, T., Talwalkar, A., and Neubig, G. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Van Teijlingen, E. and Hundley, V. The importance of pilot studies. *Social Research Update*, (35):1–4, 2001.

Wang, A., Morgenstern, J., and Dickerson, J. P. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., and Qiu, Z. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhou, M., Yu, L., Geng, X., and Luo, L. ChatGPT vs social surveys: Probing objective and subjective silicon population. *arXiv preprint arXiv:2409.02601*, 2025.

A. Example Prompt

The full participant block reproduces the seven psychometric / exposure construct items verbatim with item-text=label pairs. The 36 claims are presented it a per-respondent shuffled order. The per-item variant queries the same model 36 times per respondent, once per claim; the system message changes “exactly 36 labels” to “exactly 1 label” and the user message lists only one claim.

Zero-Shot Batch Prompt

System:

```
You are a model that predicts a participant’s perceived accuracy ratings for
multiple claims based on participant information.
The survey took place in South Korea in May 2020, during the COVID-19 pandemic.
Return strict JSON only with this schema: {"answers": ["<label>", "..."]}
The "answers" array must contain exactly 36 labels in the exact same order as
the provided claims.
Allowed labels: Not accurate at all, Not very accurate, Somewhat accurate,
Very accurate, Have not seen it
Do not include explanations, reasons, claim IDs, or extra keys.
```

User:

```
Claims to evaluate (in order):
- Korea’s method of COVID19 diagnosis is inappropriate.
- COVID19 diagnostic test is free of charge for suspected patients.
- Hand-washing and social distancing is more effective in COVID19 prevention
  than wearing a mask.
- Foreign press including the BBC and the NYT reported that Korea is
  successfully coping with COVID19 through prompt diagnostic tests.
  ... (32 more claims) ...

Participant information:
Participant profile:
- Gender: Male
- Age: 61 years old
- Education: College (2-3 years)
- Household income: KRW 2M-3M
- Political orientation: Conservative

Pre-existing attitudes/perceptions:
- Open-mindedness: A person should always consider new possibilities=Slightly
  agree; People should always take into consideration evidence that goes
  against their beliefs=Slightly agree; ... (6 more items) ...
- Faith in intuition: I trust my gut to tell me what’s true and what’s
  not=Neither agree nor disagree; ... (3 more items) ...
- Need for evidence: ... (4 items, all Neither agree nor disagree) ...
- Truth as political construct: ... (4 items) ...
- Skepticism: I often accept other people’s explanations without further
  thought=Slightly agree; It is easy for other people to convince me=Agree;
  ... (3 more items) ...
- COVID-19 information exposure: Daily newspapers=not at all;
  Television=very frequently; Online news=not at all; Social media=not at all;
  Health or medical professional websites=not at all; People around me
  (family, friends, coworkers)=frequently; Doctors=not at all
- COVID-19 emotional response: I feel fear about COVID-19=Quite a bit;
  I feel worried about COVID-19=Very much; I feel angry about COVID-19=Quite
  a bit; I feel hopeful about prevention and treatment of COVID-19=Very much
```

B. Bootstrap Confidence Intervals for All Body Tables

For each headline metric we resample the per-method eval set with replacement, n_{boot}=1,000, fixed seed. On each resample we recompute the entire statistic from scratch – per-respondent d_i, m_i, t_i scalars; the 12 paired predictor–d_i Pearson correlations and standardized OLS coefficients; CCC, sign-agreement, magnitude ratio; cross-respondent EMDs; per-respondent paired metrics (r_d, MAE$_d$). The reported point estimate is computed once on the full sample (not the bootstrap median); the CI bounds are the 2.5% / 97.5% percentiles of the bootstrap distribution.

The tables in the main paper report point estimates only for brevity. The same numbers with their 95% percentile bootstrap CIs from n_{boot}=1,000 respondent resamples are tabulated below.

B.1. Headline metrics (companion to Tab. 1)

Method	Structural ↑		Marginal↓	Individual	
	$\text{CCC}_{\text{biv } r}$	$\text{CCC}_{\text{OLS } \beta}$	EMD-d	r_d	MAE_d
ZS	0.56 [0.43, 0.65]	0.50 [0.38, 0.58]	0.716 [0.683, 0.749]	0.12 [0.07, 0.16]	0.79 [0.76, 0.82]
ZS-perItem	0.80 [0.71, 0.85]	0.71 [0.58, 0.77]	0.484 [0.460, 0.505]	0.31 [0.25, 0.35]	0.58 [0.56, 0.60]
FS	0.51 [0.35, 0.65]	0.59 [0.45, 0.68]	0.226 [0.197, 0.256]	0.18 [0.13, 0.23]	0.67 [0.64, 0.69]
FS-perItem	0.61 [0.47, 0.72]	0.63 [0.51, 0.71]	0.412 [0.388, 0.438]	0.27 [0.22, 0.32]	0.58 [0.56, 0.61]
LoRA	0.60 [0.45, 0.70]	0.41 [0.24, 0.54]	0.633 [0.604, 0.664]	0.17 [0.11, 0.22]	0.75 [0.73, 0.78]
LoRA + MLP	0.85 [0.76, 0.89]	0.78 [0.68, 0.83]	0.174 [0.148, 0.216]	0.37 [0.33, 0.42]	0.61 [0.59, 0.64]

Table 4. **Headline metrics with 95% bootstrap CIs.** Same data as Tab. 1 but with bracketed gray 95% percentile CIs from resampling held-out respondents.

B.2. Individual fidelity per subset

Method	Discernment d_i		Misinfo m_i		Trueinfo t_i	
	r ↑	MAE↓	r ↑	MAE↓	r ↑	MAE↓
ZS	0.12 [0.07, 0.16]	0.79 [0.76, 0.82]	0.14 [0.09, 0.19]	0.54 [0.52, 0.56]	-0.01 [-0.05, 0.05]	0.40 [0.38, 0.41]
ZS-perItem	0.31 [0.25, 0.35]	0.58 [0.56, 0.60]	0.16 [0.11, 0.21]	0.45 [0.43, 0.47]	0.18 [0.11, 0.23]	0.32 [0.30, 0.33]
FS	0.18 [0.13, 0.23]	0.67 [0.64, 0.69]	0.16 [0.10, 0.21]	0.51 [0.49, 0.53]	0.04 [-0.01, 0.08]	0.38 [0.36, 0.39]
FS-perItem	0.27 [0.22, 0.32]	0.58 [0.56, 0.61]	0.20 [0.13, 0.25]	0.45 [0.43, 0.47]	0.11 [0.06, 0.16]	0.30 [0.29, 0.32]
LoRA	0.17 [0.11, 0.22]	0.75 [0.73, 0.78]	0.15 [0.09, 0.19]	0.58 [0.56, 0.60]	0.09 [0.04, 0.14]	0.31 [0.30, 0.33]
LoRA + MLP	0.37 [0.33, 0.42]	0.61 [0.59, 0.64]	0.34 [0.29, 0.39]	0.46 [0.44, 0.48]	0.15 [0.10, 0.20]	0.38 [0.37, 0.40]

Table 5. **Individual fidelity with 95% bootstrap CIs.** For each per-respondent scalar (discernment d_i, Misinfo mean m_i, Trueinfo mean t_i), r is the Pearson correlation between simulator and GT and MAE is the mean absolute deviation $|x_i^{\text{GT}} - x_i^{\text{sim}}|$.

B.3. Structural decomposition (companion to Tab. 2)

Method	Bivariate r sign-agree	Bivariate r |sim|/|GT|	Standardized β sign-agree	Standardized β |sim|/|GT|
ZS	0.58 [0.50, 0.83]	1.07 [0.92, 1.26]	0.58 [0.33, 0.75]	1.03 [0.88, 1.14]
ZS-perItem	0.83 [0.67, 1.00]	1.48 [1.31, 1.68]	0.75 [0.58, 0.92]	1.50 [1.28, 1.62]
FS	0.58 [0.33, 0.83]	0.56 [0.46, 0.73]	0.50 [0.33, 0.75]	0.63 [0.55, 0.83]
FS-perItem	0.92 [0.58, 1.00]	0.85 [0.73, 0.98]	0.67 [0.50, 0.92]	0.99 [0.84, 1.13]
LoRA	0.50 [0.42, 0.83]	0.50 [0.43, 0.67]	0.50 [0.33, 0.75]	0.58 [0.48, 0.75]
LoRA + MLP	0.92 [0.67, 1.00]	1.28 [1.14, 1.46]	0.75 [0.50, 0.92]	1.32 [1.12, 1.44]

Table 6. **Structural-fidelity decomposition with 95% bootstrap CIs.** Same data as Tab. 2 with respondent-level bootstrap intervals.

B.4. PPI rectification (companion to Tab. 3)

Method	Raw simulator μ_{Mis}	Raw simulator μ_{Tru}	+ PPI rectification μ_{Mis}	+ PPI rectification μ_{Tru}
GT	2.04	3.13	2.04	3.13
ZS	1.77	3.45	2.12 [2.12, 2.12] ↓	3.08 [3.08, 3.08] ↓
ZS-perItem	2.01	2.98	2.12 [2.12, 2.12] ↑	3.08 [3.08, 3.08] ↓
FS	2.06	3.28	2.11 [2.10, 2.11] ↑	3.09 [3.09, 3.09] ↓
FS-perItem	2.02	3.13	2.11 [2.11, 2.11] ↑	3.07 [3.07, 3.07] ↑
LoRA†	1.57	3.21	1.58 [1.57, 1.59]	3.19 [3.18, 3.20]
LoRA + MLP†	2.01	3.23	2.03 [2.00, 2.05]	3.21 [3.19, 3.22]

Table 7. **PPI rectification.** Same data as Tab. 3 with respondent-level bootstrap intervals.

B.5. Subgroup analysis (companion to Fig. 3)

Subgroup	n	EMD-d ↓	$\text{CCC}_{\text{biv } r}$ ↑	r_d ↑	MAE_d ↓
All respondents	1392	0.17 [0.15, 0.22]	0.85 [0.76, 0.89]	0.37 [0.33, 0.42]	0.61 [0.59, 0.64]
Gender					
Male	719	0.16 [0.14, 0.21]	0.89 [0.79, 0.93]	0.38 [0.32, 0.45]	0.62 [0.59, 0.66]
Female	673	0.20 [0.16, 0.26]	0.77 [0.62, 0.85]	0.36 [0.29, 0.43]	0.60 [0.57, 0.63]
Age group					
Age 20–39	475	0.20 [0.17, 0.26]	0.90 [0.75, 0.95]	0.38 [0.29, 0.46]	0.63 [0.58, 0.67]
Age 40–54	483	0.16 [0.14, 0.23]	0.87 [0.75, 0.92]	0.41 [0.34, 0.48]	0.60 [0.56, 0.64]
Age 55–69	434	0.18 [0.13, 0.25]	0.69 [0.51, 0.80]	0.31 [0.23, 0.40]	0.61 [0.57, 0.66]
Political orientation					
Conservative	215	0.27 [0.22, 0.37]	0.40 [0.00, 0.65]	0.17 [0.03, 0.31]	0.61 [0.55, 0.68]
Independent	812	0.22 [0.18, 0.25]	0.86 [0.74, 0.91]	0.29 [0.23, 0.35]	0.59 [0.56, 0.62]
Liberal	365	0.31 [0.24, 0.39]	0.73 [0.46, 0.85]	0.22 [0.11, 0.32]	0.65 [0.60, 0.71]

Table 8. **Three-axis fidelity of the best simulator (LoRA + MLP) within demographic subgroups** (95% bootstrap CI). When the slicing variable is itself a structural predictor, it is dropped from the structural regression for that slice (11 predictors).