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

Anonymous Authors[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 broader population? Using a COVID-19 misinformation survey, we benchmark three families of approaches: prompting, PPI (Prediction-Powered Inference) rectification, and PEFT (parameter-efficient fine-tuning). We decompose recovery along three axes: marginal fidelity, defined as cross-respondent distributional similarity; structural fidelity, defined as alignment in predictor–outcome relationships; and individual fidelity, defined as agreement on per-respondent summaries. PEFT applying LoRA adapters with an MLP classifier head performed best across nearly all axes. These findings suggest that fine-tuning on small pilot samples offers a balanced approach for achieving multiple forms of fidelity.

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 intervention design) rely on *predictor–outcome* relations – which demographic and psychological predictors drive susceptibility – not just on marginal accuracy. 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 (Wang et al., 2025; Li 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 three 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: *marginal* fidelity (do simulated marginals match the human ones?), *structural* fidelity (do predictor–outcome relations match?), and *individual* fidelity (does each simulated respondent track their human counterpart?). **Second**, we introduce a calibrated evaluation protocol that maps each axis to a specific metric: cross-respondent Earth Mover's Distance (EMD) on per-respondent scalar summaries for the marginal axis; Lin's Concordance Correlation Coefficient (CCC), decomposed into a sign-agreement rate and a magnitude ratio, for the structural axis; and paired Pearson r_d (relative) and MAE_d

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(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 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. **LoRA + MLP** dominates the structural axis and is the lowest-EMD method on all three scalar summaries. **ZS-PI** is a strongest prompt-only competitor without using any pilot data. **LoRA** fine-tuning shows the most pronounced magnitude compression – its predictor–outcome slopes are systematically attenuated relative to GT. **PPI** pulls the most biased simulator's mean score but is algebraically degenerate on the fine-tuning family.

2. Related Work

We organize prior work into calibration families relevant to the small-pilot setting: prompt-based conditioning, test-time and statistical interventions, and parameter-efficient fine-tuning. We then position our work against each other.

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 richer schemes incorporate social-network and peer features (Choi et al., 2026). Park et al. (2024) uses full two-hour interview transcripts as the persona and recovers 85% of human test–retest accuracy on the GSS. 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.

Pilot conditioning as base-rate injection. A natural response to these audits is to inject a small amount of *Contextual Calibration* via 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). To our knowledge, no prior work has systematically isolated the effect of a small pilot (∼5%) on *both* marginal and structural recovery within a single misinformation survey, nor compared zero-shot per-item, few-shot per-item, PPI rectification, and fine-tuning head-to-head on the same data under three-axis fidelity metrics.

2.2. Test-Time and Statistical Interventions

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. *Audience Segmentation* (Qin et al., 2026) restores within-group heterogeneity by varying identifier granularity. 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 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. Two practical

issues are largely unresolved in this literature: (i) what happens with a small pilot (we use only 5% of respondents)? and (ii) does the choice of *output parametrization* – autoregressive token generation vs. a discriminative classification head – matter for the recovery axes that actually matter to downstream uses? We address both head-on.

A recent systematic review of *LLM psychometrics* (Ye et al., 2025) stresses that LLMs need to be evaluated with classical reliability/validity standards, not only marginal accuracy. We follow this prescription: we report marginal, structural, and individual metrics across all method families on the *same* pilot.

Position of this work. The closest study to ours is Krsteski et al. (2025), which compares prompting, fine-tuning, and PPI rectification in a survey simulation. 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 a calibrated structural-fidelity metric that penalizes deviations from the y=x line; (iii) we report how much individual predictions are correlated to each of the human participants' responses, providing a benchmark of individual fidelity as well.

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 a DISCERNMENT scalar per respondent, $d_i = \overline{y_{i,\text{Tru}}} - \overline{y_{i,\text{Mis}}}$, and use it for both individual and structural analyses.

3.2. Pilot Sampling

We draw a 5% pilot D_p (n=74) with a fixed seed (42), 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-PI). For each held-out respondent i, the LLM is given X_i and asked to predict Y_i using only its pretrained *Universal Prior*. **ZS** (batch) elicits all 36 ratings in a single prompt; **ZS-PI** (per-item) 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).

Zero-Shot Per-Item Prompt

System: Predict a participant's perceived accuracy rating for the claim. Return strict JSON: {"answer": "<label>"}.
User: Claim: *Claim A*
Profile: Male, 53, Independent.
Response:

Family 2 – Few-Shot Prompting (FS, FS-PI). We inject D_p as in-context examples in two flavors that mirror Family 1. **FS** (batch) appends a fixed pool of (X_j, Y_j) pairs from D_p to a single 36-item prompt; **FS-PI** (per-item) instead draws a fresh handful of (X_j, Y_j) pairs for every (target respondent, item) and queries items individually. Together with Family 1 these form a 2×2 factorial of {no pilot, with pilot}×{batch, per-item}.

Few-Shot Per-Item Prompt

System: *(identical to ZS, single-item version)*
User: Examples of real human responses on this claim:
[1] Male, 36, Independent → *Not accurate at all*
[2] Female, 62, Independent → *Not very accurate*
⋮
Now predict for: Male, 53, Independent.
Response:

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** – standard LoRA adapters on attention and MLP projections, autoregressively emitting the

gold 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 parametrization, isolating the effect of the discriminative 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 $\lambda_q^{\text{opt}} = \text{Cov}(y, \hat{y})/\text{Var}(\hat{y})$ 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 under that counterfactual substitution 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. Bootstrap uncertainty intervals on every headline metric are reported in App. B; details of the resampling procedure are in App. A.

Marginal fidelity. *Do the simulated cross-respondent distributions match the human ones?* For each held-out respondent we compute three scalar summaries – discernment d_i, Misinfo mean m_i, and Trueinfo mean t_i – and report the 1-Wasserstein *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).

Structural fidelity. *Do the 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	Pop.↓ EMD-d_i	Respondent r_d ↑	Respondent MAE$_d$ ↓	Structural ↑ CCC$_{\text{biv }r}$	Structural ↑ CCC$_{\text{OLS }\beta}$
ZS	0.72	0.12	0.79	0.56	0.50
ZS-PI	0.48	0.31	0.58	0.80	0.71
FS	0.23	0.18	0.67	0.51	0.59
FS-PI	0.41	0.27	**0.58**	0.61	0.63
LoRA	0.63	0.17	0.75	0.60	0.41
LoRA + MLP	**0.17**	**0.37**	0.61	**0.85**	**0.78**

Table 1. **Headline metrics across the three fidelity axes** on the held-out evaluation set ($n_{\text{eval}} \approx 1{,}300$ per method). **Population**: cross-respondent EMD on d_i (lower = better). **Respondent**: r_d is the cross-respondent Pearson between simulator and GT d_i; MAE$_d$ is mean $|d_i^{\text{sim}} - d_i^{\text{GT}}|$. **Structural**: Lin's CCC on K=12 paired predictor–d_i correlations and 12 standardized OLS coefficients. CIs in App. B; **bold** = best non-baseline value per column.

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 Recovery

Table 1 reports the headline numbers; Fig. 1 renders the structural axis as a forest plot. **LoRA + MLP** dominates the bivariate-r (CCC 0.85) and the OLS-β (CCC 0.78). **ZS-PI** 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 CCC reveals different failure modes hidden behind similar headline numbers. **LoRA + MLP** and **FS-PI** tie for highest sign-agreement on bivariate r (0.92, 11 of 12 predictors),

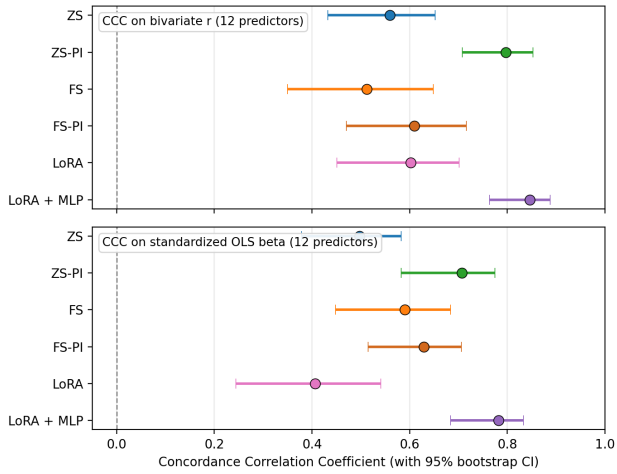

Figure 1. **Structural-fidelity forest plot** on the 12 predictors. Per method: point estimate of Lin's CCC with bootstrap uncertainty intervals (App. B). Dashed line at 0. Top panel = bivariate r; bottom panel = standardized OLS β. LoRA + MLP dominates and ZS-PI is a strong prompt-only second.

Method	Bivariate r		Standardized β	
	sign-agree	$\lvert\text{sim}\rvert/\lvert\text{gt}\rvert$	sign-agree	$\lvert\text{sim}\rvert/\lvert\text{gt}\rvert$
ZS	0.58	1.07	0.58	1.03
ZS-PI	0.83	1.48	0.75	1.50
FS	0.58	0.56	0.50	0.63
FS-PI	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). $\lvert\text{sim}\rvert/\lvert\text{gt}\rvert$ is the ratio of mean absolute coefficients (>1 inflates, <1 compresses). Bootstrap CIs are in App. B.

but they differ on magnitude: FS-PI sits closest to the y=x line ($\lvert\text{sim}\rvert/\lvert\text{gt}\rvert$=0.85), while LoRA + MLP mildly inflates (1.28). **ZS-PI** also has high sign-agreement (0.83 on r, 0.75 on β) but inflates magnitudes more substantially (1.48 on r, 1.50 on β). At the other end, **Vanilla LoRA** compresses slopes most severely (mag 0.50 on r, 0.58 on β, with sign-agreement also at the floor of 0.50) – both directions and magnitudes are off. **FS** shows a similar compression pattern (0.58 sign / 0.56 mag on r). **ZS** occupies a middle ground – magnitudes match GT closely (1.07 on r, 1.03 on β) but sign-agreement is only 0.58. The takeaway is that high CCC requires *both* directional agreement and matched magnitude, and that fine-tuning on the pilot can swing magnitudes in either direction (compressed for vanilla LoRA, inflated for LoRA + MLP) depending on output parametrization. The compression/inflation patterns match the structural failures (variance flattening, effect-size attenuation) documented in prior LLM-survey audits (Bisbee et al., 2024; Choi et al., 2026).

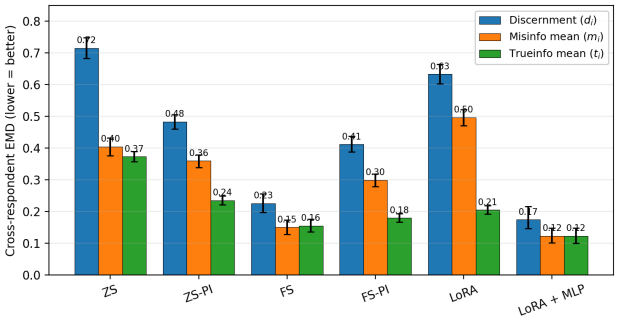

Figure 2. **Cross-respondent EMD on per-respondent scalar summaries.** Wasserstein-1 between simulator and GT distributions of, respectively, discernment d_i, Misinfo mean m_i, and Trueinfo mean t_i. Lower is better; error bars are bootstrap uncertainty intervals (App. B). LoRA + MLP attains the lowest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

4.2. Marginal Recovery

Figure 2 plots the three cross-respondent EMDs. **LoRA + MLP** achieves the lowest EMD on all three (EMD-D 0.17, EMD-Mis 0.12, EMD-Tru 0.12), separated from every other method on each summary. **FS** is second on each (EMD-D 0.23, EMD-Mis 0.15, EMD-Tru 0.16), and **FS-PI** third. **ZS** is the worst on discernment EMD (0.72) – batch zero-shot prompting moves both subset means in the wrong direction (under-rates Misinfo, over-rates Trueinfo), inflating the cross-respondent variance of d_i. **LoRA** (without the MLP head) is also poor on Misinfo EMD (0.50) despite winning the structural sign-agreement axis, again consistent with the magnitude-compression pattern.

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. The pattern matches Krsteski et al. (2025)'s Eq. 3 sharply: PPI rescues the most biased simulator (**ZS** Mis 1.77→2.12, Tru 3.45→3.08, both moving toward GT), is roughly neutral on Trueinfo for already-calibrated methods, and *actively hurts* the Misinfo estimate of methods whose raw output is already near GT (FS, FS-PI, ZS-PI all see Mis deviation grow modestly). For the LoRA family the rectifier is algebraically degenerate: when $\hat{y}_{\text{pilot}} = y_{\text{pilot}}$ (which holds at the limit of training-set memorization) we have $\lambda_q \equiv 1$ and $\hat{\theta}_q = \bar{\hat{y}}_{\text{held},q}$, so PPI hands back the simulator's raw held-out mean. We mark these rows with † in Tab. 3. The headline reading 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, and that fine-tuning on the pilot mechanically defeats PPI's bias-estimation step.

Method	Raw simulator		+ PPI rectification	
	μ_{Mis}	μ_{Tru}	μ_{Mis}	μ_{Tru}
GT	2.04	3.13	2.04	3.13
ZS	1.77	3.45	2.12 ↓	3.08 ↓
ZS-PI	2.01	2.98	2.12 ↑	3.08 ↓
FS	2.06	3.28	2.11 ↑	3.09 ↓
FS-PI	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 (by ≥ 0.01); ↑ 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 (see §4.2).

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-PI** second (r_d=0.31), **FS-PI** third (r_d=0.27), and **ZS** at the floor (r_d=0.12). The absolute axis tells a slightly different story: **ZS-PI** (MAE$_d$=0.58) ties **FS-PI** (0.58) for the smallest per-respondent deviation, with **LoRA + MLP** (0.61) close behind. The split matters: ZS-PI's predicted discernment is, on average, the closest to GT in absolute terms, but its respondent ranking is meaningfully looser than LoRA + MLP's. We attribute this to the same compression pattern that drives the structural axis – ZS-PI's distribution sits closer to GT in mean but the rank order across respondents is preserved more faithfully by the discriminative head. The full pair of metrics on m_i and t_i is in App. B.

5. Discussion

LoRA + MLP wins both the structural and the population axes. LoRA + MLP is the only method that simultaneously (i) clears the no-recovery threshold by a wide margin on both bivariate-r and standardized-β CCC; (ii) attains the lowest cross-respondent EMD on all three per-respondent scalar summaries; and (iii) wins or ties on both individual metrics. No other method in our benchmark comes close on more than one axis at once – which is what justifies treating LoRA + MLP as the strongest method here.

ZS-PI is a remarkably strong prompt-only second, with no pilot data at all. Per-item zero-shot prompting attains a bivariate-r CCC second only to LoRA + MLP, and is within striking distance on the OLS-β axis. Whatever batch ZS does that ZS-PI undoes, the gap between the two is large; we conjecture that the batch prompt induces a single shared decoding trajectory across items that washes out predictor-conditional structure, while per-item prompting recomputes the conditional from scratch each time. We do not isolate the mechanism here. The empirical fact, however, is that ZS-PI recovers structural relations at a level the few-shot prompting variants do not, despite using zero pilot information.

Output parametrization controls the magnitude direction. Vanilla LoRA and LoRA + MLP share the same backbone and LoRA rank but differ only in output parametrization, and that single design choice flips the structural-recovery profile – from the most compressed in our benchmark (vanilla LoRA, the cleanest illustration in the table of the variance flattening / regression-to-the-mean failure mode documented across LLM-survey audits (Bisbee et al., 2024; Choi et al., 2026)) to the highest sign-agreement with mild over-inflation (LoRA + MLP). The autoregressive token decoder is biased toward modal labels and squashes between-respondent variation, while the discriminative head trained on the empirical class distribution preserves – and slightly amplifies – it. The lesson is not that fine-tuning fails or succeeds on its own, but that the output head controls which direction the magnitudes lean.

PPI is conditional on bias variance and algebraically degenerate for fine-tuned simulators. PPI rescues batch ZS – both subset means moving toward GT – but pulls already-calibrated simulators away from GT on the Misinfo axis, consistent with Krsteski et al. (2025)'s Eq. 3: PPI is beneficial only 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. This paper's "degenerate for LoRA" claim is therefore not just an empirical observation but an algebraic consequence of how PPI uses the pilot.

6. Conclusion

We benchmarked prompting (ZS / FS × batch / per-item), PPI rectification, and LoRA / LoRA + MLP fine-tuning on the same 5% pilot of a 1,466-respondent COVID-19 misinformation survey, under a calibrated evaluation protocol that reports concordance correlation coefficients on the structural axis and cross-respondent EMDs on per-respondent scalar summaries. **LoRA + MLP** dominates the structural axis and the cross-respondent EMDs on all three summaries. **ZS-PI** is a strong prompt-only second without using any pilot data. **Vanilla LoRA** produces the most pronounced

magnitude compression – the cleanest illustration in our benchmark of the variance flattening / regression-to-the-mean failure mode reported across LLM-survey audits. **PPI** rescues the most biased simulator's marginals but mildly hurts already-calibrated ones; on the LoRA family it is algebraically degenerate because training on the pilot makes $\lambda_q \equiv 1$ and the rectifier returns the raw held-out mean.

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 magnitude of the LoRA + MLP advantage may differ in domains where the LLM's pretraining prior is less aligned with post-hoc scientific consensus (e.g. purchase intent, taste). **Single backbone.** LoRA results use Qwen-3-8B; 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 are still 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) a multiverse analysis varying the pilot seed and pilot size $n \in \{20, 75, 150\}$ (Cummins, 2025); (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 – with calibrated uncertainty – 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 homogenize minority perspectives, as documented in prior works we cite. By emphasizing recovery diagnostics across marginal, structural, and individual axes and by replacing ad hoc structural metrics with the concordance correlation coefficient, 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. D., and Präg, P. Machine bias: How do generative language models answer opinion polls? *Sociological Methods and Research*, 2025. Forthcoming; preprint hal-04849013.

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. *arXiv preprint arXiv:2602.04674*, 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. *Working paper*, 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*, 2025. Also arXiv:2402.01908.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ye, H., Jin, J., Xie, Y., Zhang, X., and Song, G. Large language model psychometrics: A systematic review of evaluation, validation, and enhancement. *arXiv preprint arXiv:2505.08245*, 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. Bootstrap and Pilot-Derivation Details

Bootstrap. For each headline metric we resample the per-method eval set with replacement, n_{boot}=1,000, fixed RNG seed 42. 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.

PPI on the LoRA family. The PPI rectifier $\hat{\theta}_q = \bar{y}_{\text{pilot},q} + \lambda_q(\bar{\hat{y}}_{\text{held},q} - \bar{\hat{y}}_{\text{pilot},q})$ requires simulator predictions on the pilot. Our LoRA configurations were trained on the pilot, so we have no held-out-style predictions for those rows. Setting $\hat{y}_{\text{pilot}} = y_{\text{pilot}}$ (the limit of perfect training-set memorization) yields $\lambda_q = \text{Cov}(y, y)/\text{Var}(y) = 1$ on every item, and $\hat{\theta}_q$ algebraically reduces to the simulator's raw held-out mean. Tab. 3 marks these rows with †. A meaningful PPI evaluation on fine-tuned simulators would require collecting predictions on the pilot from a checkpoint that was *not* trained on those respondents – e.g., a held-out-style validation pilot drawn separately from the training pilot, following Krsteski et al. (2025).

B. Bootstrap Confidence Intervals for All Body Tables

The body tables (Tab. 1, 2, 3) 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	Pop.↓ EMD-d_i	Respondent r_d	Respondent MAE$_d$	Structural ↑ CCC$_{\text{biv }r}$	Structural ↑ CCC$_{\text{OLS }\beta}$
ZS	0.716 [+0.683, +0.749]	0.12 [+0.07, +0.16]	0.79 [+0.76, +0.82]	0.56 [+0.43, +0.65]	0.50 [+0.38, +0.58]
ZS-PI	0.484 [+0.460, +0.505]	0.31 [+0.25, +0.35]	0.58 [+0.56, +0.60]	0.80 [+0.71, +0.85]	0.71 [+0.58, +0.77]
FS	0.226 [+0.197, +0.256]	0.18 [+0.13, +0.23]	0.67 [+0.64, +0.69]	0.51 [+0.35, +0.65]	0.59 [+0.45, +0.68]
FS-PI	0.412 [+0.388, +0.438]	0.27 [+0.22, +0.32]	0.58 [+0.56, +0.61]	0.61 [+0.47, +0.72]	0.63 [+0.51, +0.71]
LoRA	0.633 [+0.604, +0.664]	0.17 [+0.11, +0.22]	0.75 [+0.73, +0.78]	0.60 [+0.45, +0.70]	0.41 [+0.24, +0.54]
LoRA + MLP	0.174 [+0.148, +0.216]	0.37 [+0.33, +0.42]	0.61 [+0.59, +0.64]	0.85 [+0.76, +0.89]	0.78 [+0.68, +0.83]

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 (n_{boot}=1,000, fixed seed).

B.2. Individual fidelity per subset

Method	Discernment d_i r ↑	Discernment d_i MAE↓	Misinfo m_i r ↑	Misinfo m_i MAE↓	Trueinfo t_i r ↑	Trueinfo t_i 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-PI	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-PI	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. **Respondent-level fidelity per subset, with 95% bootstrap CIs.** For each per-respondent scalar (discernment d_i, Misinfo mean m_i, Trueinfo mean t_i), r is the cross-respondent Pearson correlation between simulator and GT and MAE is the mean absolute deviation $|x_i^{\text{sim}} - x_i^{\text{GT}}|$.

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

Method	Bivariate r		Standardized β	
	sign-agree	|sim|/|gt|	sign-agree	|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-PI	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-PI	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		+ PPI rectification	
	μ_{Mis}	μ_{Tru}	μ_{Mis}	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-PI	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-PI	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.** 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 (by ≥ 0.01); ↑ 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 (see §4.2). Bracketed gray values are 95% bootstrap CIs from resampling held-out respondents (n_{boot}=1,000, fixed seed).

C. Worked-Out Prompts

The body sketches the prompt structure with abbreviated examples; this appendix reproduces a complete prompt for one respondent (demographics altered to protect privacy).

C.1. ZS / Batch – full 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. (Truncated to first 4 of 36 claims and the participant block; the actual prompt lists all 36 claims in shuffled order.)

```
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
```

The full participant block reproduces the seven psychometric / exposure construct items verbatim with item-text=label pairs. The 36 claims are presented in a per-respondent shuffled order (with a fixed seed derived from the responseID) so that order effects do not co-vary with item subset.

C.2. ZS-PI / Per-item

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:

System.

```
You are a model that predicts a participant's perceived accuracy rating for
ONE claim based on participant information.
The survey took place in South Korea in May 2020, during the COVID-19 pandemic.
Return strict JSON only: {"answer": "<label>"}
Allowed labels: Not accurate at all, Not very accurate, Somewhat accurate,
Very accurate, Have not seen it
```

User.

```
Claim: Korea's method of COVID19 diagnosis is inappropriate.

Participant information:
(same participant block as in App. E.1)
```

C.3. FS-PI / Per-item with five real pilot examples

Few-shot per-item adds an examples block listing real pilot respondents' GT answers on the *same* target claim. The five examples below are pilot rows:

System. (same as ZS-PI)

User.

```
Examples of real human responses on this claim:
[1] Male, 67, Conservative   -> Not very accurate
[2] Male, 62, Independent    -> Not very accurate
[3] Female, 44, Liberal      -> Not accurate at all
[4] Female, 55, Conservative -> Somewhat accurate
[5] Male, 35, Independent    -> Not very accurate

Now predict for:
Male, 61, Conservative.

Claim: Korea's method of COVID19 diagnosis is inappropriate.
Participant information:
(same participant block as in App. E.1)

Response:
```

A fresh handful of (X_j, Y_j) pairs is drawn for every (target respondent, item) call, so different items see different example sets. The pool is the 74-respondent pilot set; we draw five examples per call without replacement, sampled with a deterministic seed derived from the (target responseID, item code) pair so that the same pairing always sees the same examples.

C.4. FS / Batch

The FS / batch variant places the same examples block once at the top of the user message and then appends the full 36-claim list (as in App. E.1). The system message reverts to "exactly 36 labels." Because the example block sits before the claim list and the participant block, it conditions the entire 36-item generation rather than a single item; this is the design distinction between FS and FS-PI documented in §3.