LLM Human Response Alignment: A Multi-Sample Debiasing Framework

Li Jiang [1] **Xiao Liu** [2]

Abstract

Large Language Models (LLMs) are increasingly used as proxies for human respondents in social science, behavioral, and marketing experiments, yet their synthetic responses diverge systematically from human ones. We study the post-hoc debiasing route that keeps the LLM frozen and learns an external debiasing estimator to align the LLM responses with the human responses. We model both human and LLM responses as latent distributions to capture their inherent heterogeneity. However, the two distributions are observed asymmetrically, with the collection cost capping the human side at a few observations per individual, while the LLM is queryable at negligible cost. Existing pipelines collapse the LLM side to a scalar before correction and discard the distributional signal that repeated cheap queries could otherwise capture. Motivated by this missed signal, we employ the full multi-sample vector of LLM responses as the input feature to a debiasing module on both population-level and individual-level debiasing tasks. Under squared loss, we give an information-theoretic motivation for retaining the full vector, showing it weakly Bayes-dominates any compression and strictly so on supra-mean targets. Across three benchmarks, our method attains the best value on the majority of metric cells at both the population and individual levels, reducing prediction error over the uncorrected Base LLM by over 50% on multiple tasks, and yields better distributional alignment to the human responses against the baselines.

[1]Desautels Faculty of Management, McGill University [2]Stern School of Business, New York University. Correspondence to: Xiao Liu <xl23@stern.nyu.edu>.

Workshop on Pluralistic-Alignment, Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Large Language Models (LLMs) are increasingly used as low-cost stand-ins for human respondents in social science surveys, marketing studies, behavioral experiments, and economic decision tasks (Horton et al., 2023; Brand et al., 2023; Argyle et al., 2023; Aher et al., 2023; Park et al., 2026; Hewitt et al., 2024; Cui et al., 2025; Lippert et al., 2024; Yeykelis et al., 2024; Anthis et al., 2025; Li et al., 2024; Maier et al., 2025). However, those synthetic responses diverge systematically from human ones, exhibiting opinion misalignment, distorted preferences, demographic flattening, and persona homogenization (Santurkar et al., 2023; Goli & Singh, 2024; Wang et al., 2025a; Tjuatja et al., 2024; Jiang et al., 2025; Li et al., 2025). To close this gap, prior work pursues two broad strategies, modifying the LLM itself (e.g., fine-tuning) or debiasing its outputs externally. We focus on the latter, the post-hoc route, which keeps the LLM fixed and assumes only query access, learning an external debiasing module that aligns LLM responses toward human ones.

Faithfully modeling this problem requires treating responses as distributional for both humans and LLMs, rather than as deterministic scalars. A human respondent's answer varies across time and context, and is best modeled as a draw from a latent response distribution: within-person test-retest accuracy on repeated questions reaches about 80% (Toubia et al., 2025; Park et al., 2026). Analogously, the persona (e.g., demographic) sent to an LLM cannot fully determine the response, and the LLM's output is itself a draw from a stochastic distribution that varies with decoding hyperparameters and model size. Yet beyond this symmetry in framing, the two distributions are often observed asymmetrically. Repeated collections from human respondents are costly and time-consuming, so most consumer-research datasets provide only a single scalar answer per respondent and question, leaving the underlying human distribution effectively latent. The LLM distribution, by contrast, can be queried as many times as desired at negligible marginal cost.

Constrained by what available datasets supply, we therefore retain a single scalar human response as the scalar ground truth. Existing post-hoc debiasing pipelines, however, ex-

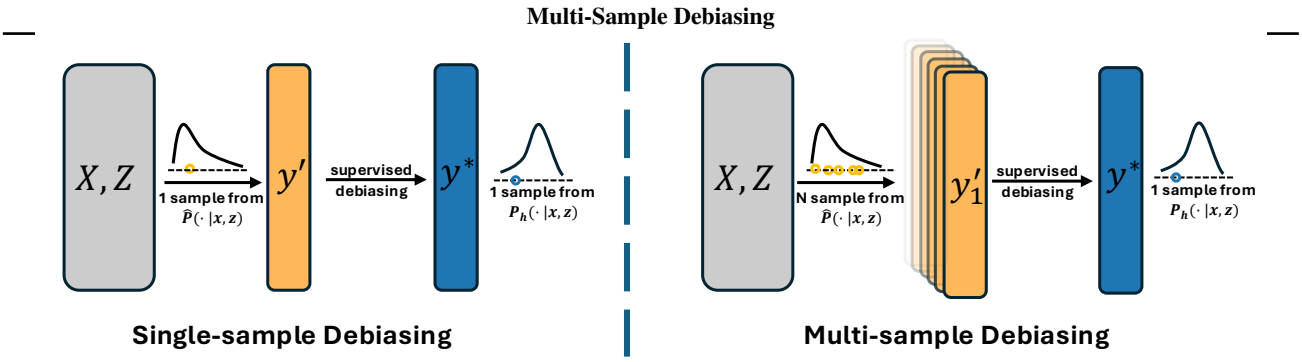


Figure 1. Multi-sample debiasing pipeline. Whereas single-sample debiasing collapses the LLM-side input to one response y', our multi-sample approach uses the multi-sample vector $\mathbf{Y}'_N$ of N LLM responses as input to a downstream debiasing module.

tend this scalar collapse to the LLM side before any correction, feeding only *one* LLM response for the following debiasing module (Wang et al., 2026; Zhang et al., 2025a; Audinet de Pieuchon et al., 2025; Wang et al., 2025b). While unavoidable on the human side, such a collapse is statistically inefficient on the LLM side, since the LLM admits cheap queries, and the resulting multiple samples reveal distributional structure (e.g., variance) that a single-sampling method cannot capture. We confirm this argument on a motivating toy benchmark (Section 3.2), where multi-sampling debiasing reduces single-sample test error by roughly $4\times$ compared to the approaches without distributional information.

We therefore propose to use the multi-sample LLM responses as an input vector to the downstream debiasing module. We forgo the explicit distribution modeling with a parametric (e.g., Gaussian) head because survey response distributions are routinely bimodal, bounded, or heavily skewed (Toubia et al., 2025; Santurkar et al., 2023; Kolluri et al., 2025), and rare human observations leave any variance head unsupervised. The same input feature template instantiates two complementary tasks: population-level debiasing predicts the average response of the human population to a given question, while individual-level debiasing predicts a specific respondent's answer to a given question, additionally conditioning on a respondent embedding. We provide an information-theoretic motivation, rather than a finite-sample guarantee, for retaining the full vector; under squared loss, an estimator on the full multi-sample vector weakly Bayes-dominates any estimator on a measurable summary of it, strictly so when the target depends on supra-mean features (Section 4).

We evaluate our multi-sample vector debiasing algorithm on three benchmarks spanning Likert-scale surveys (Twin-2K-500, OpinionQA) and middle-school educational assessment (EEDI), at both the population and individual level. Compared with other LLM-input variants, our method attains the best performance across most metric cells at both task levels, reducing prediction error over the uncorrected Base LLM by up to 58% at the population level and up to 71% at the individual level (Section 5.2). The individual-level gain is most pronounced on EEDI, with smaller but consistent improvements on Twin-2K-500 (-24%) and OpinionQA (-27%). Beyond pointwise accuracy, our method achieves better distributional alignment than other approaches, shown in Section 5.3.

2. Related Work

LLM-generated responses diverge from human ones systematically, surfacing across a wide range of simulation tasks, e.g., opinions (Santurkar et al., 2023; Tjuatja et al., 2024; Durmus et al., 2023; Dominguez-Olmedo et al., 2024; Bisbee et al., 2024; Peng et al., 2025), decision-making (Goli & Singh, 2024; Bini et al., 2025; Chen et al., 2025; Horton et al., 2023; Aher et al., 2023; Mei et al., 2024; Wang et al., 2025c), and persona-conditioned behavior (Wang et al., 2025a; Kang et al., 2025; Li et al., 2025; Argyle et al., 2023; Cheng et al., 2023; Gupta et al., 2024; Chen et al., 2026). These distortions share a common generative origin: pretraining via next-token prediction pulls outputs toward the modes of internet-scale corpora that already encode demographic stereotypes (Huang & Rust, 2025; Jiang et al., 2025; Guilbeault et al., 2025; Peng et al., 2025), and post-training (e.g., RLHF) further narrows the response distribution. Efforts to address this fall into two broad strategies—in-model interventions, or post-hoc output correction.

Model-side adaptation via fine-tuning and persona/agent design. One family of debiasing approaches intervenes upstream of generation, modifying the model's weights, conditioning inputs, or surrounding agent system. One thread fine-tunes LLM weights on human behavioral data from surveys, social-science experiments, personal traces, and cognitive-psychology corpora (Argyle et al., 2023; Brand et al., 2023; Kolluri et al., 2025; Lei et al., 2026; Binz et al., 2025). A second thread engineers personas, either by deepening individual personas through interviews, backstories, or psychometric calibration, or by broadening the persona pool through diverse genera-

tion and synthetic populations (Park et al., 2026; Kang et al., 2025; Huang et al., 2026; Paglieri et al., 2026; Lin et al., 2025). A third thread embeds LLMs in agent frameworks for specific behavioral settings (Lu et al., 2025; Bhattacharyya et al., 2026). These approaches are computationally expensive and require weight access or heavy persona engineering, yet remain inconsistently effective—fine-tuned models and digital twins often fail to outperform well-configured base LLMs (Peng et al., 2025).

Post-hoc debiasing with human-anchored estimators. A second family combines a small human-labeled sample with abundant LLM-generated annotations, rooted in semi-parametric missing-data inference (Robins et al., 1994; Chernozhukov et al., 2018) and extended to LLMs through prediction-powered inference and design-based supervised learning (Egami et al., 2023; Angelopoulos et al., 2023; Audinet de Pieuchon et al., 2025), with applications across conjoint market research (Wang et al., 2026), demand and treatment-effect estimation (Zhang et al., 2025a), and fine-tuning with post-hoc rectification (Wang et al., 2025b). These methods target unbiased estimation of a downstream scalar parameter (regression coefficient, treatment effect, or population mean) with valid confidence intervals, and use a single LLM annotation per unit because the inferential machinery treats the LLM as a noisy surrogate label. Our framework instead targets the response itself, evaluated by pointwise predictive metrics on held-out human responses; treating the LLM as a feature rather than a surrogate label lets us retain repeated cheap queries as a vector input that captures the distributional structure scalar-surrogate pipelines discard.

Distributional and uncertainty-aware perspectives. A complementary thread argues that LLM outputs should be understood as samples from a distribution rather than point estimates (Brand et al., 2023; Park et al., 2026). Leng et al. (2024) operationalize this view by reweighting LLM-generated samples against a small human reference to compose human-like ensembles, thereby reducing preference disparity at the distribution level. Huang et al. (2025) take a different angle: they treat repeated LLM responses as draws from a synthetic population and adaptively choose the simulation sample size needed to achieve nominal coverage of human-population parameters. Their target, however, is uncertainty quantification for population parameters; ours is point prediction at the individual and population level. To our knowledge, no prior work has used the multi-sample repeated LLM draws as input to a downstream debiasing module.

3. A Multi-Sample Debiasing Framework

3.1. Setup and Task Definition

This paper studies the post-hoc debiasing task. The LLM is frozen, and we learn an external debiasing module g_θ that maps the LLM's response evidence to a bias-corrected estimate of the human target. Let $\mathcal{X}$ denote the space of questions and $\mathcal{Z}$ the space of respondent profiles (e.g., demographics, psychographics, or past behaviors). Given a question $x \in \mathcal{X}$ and a respondent with profile $z \in \mathcal{Z}$, our goal is to predict how the respondent would answer x. We model both human and LLM responses as distributions rather than fixed scalars, capturing within-person variability on the human side and across-model variability with stochastic decoding on the LLM side. Each elicited human answer is one sample from a latent $P_h(\cdot \mid x, z)$, while LLM outputs form $\hat{P}(\cdot \mid x, z)$.

$\hat{P}$ is systematically biased relative to P_h, and the two distributions are further observed asymmetrically. P_h is typically available as a single scalar response per (respondent, question) pair, since repeated elicitation from the same respondent is costly and rarely feasible at survey scale, whereas $\hat{P}$ admits cheap repeated queries and therefore yields additional observable response evidence. Let $y^* \sim P_h(\cdot \mid x, z)$ denote the scalar human target induced by the latent human response distribution, and let $\mathcal{E}_{\text{LLM}} = \Phi(\hat{P}(\cdot \mid x, z))$ denote observable evidence extracted from the LLM response distribution. The post-hoc debiasing task is to learn a mapping $g \in \mathcal{G}$ that uses question, respondent profile, and LLM-generated responses to predict y^* under loss ℓ:

$$\min_{g \in \mathcal{G}} \mathbb{E}[\ell(g(x, z, \mathcal{E}_{\text{LLM}}), y^*)]\,. \tag{1}$$

3.2. A Motivating Toy Benchmark

Before specifying the method, we isolate the central methodological question in a controlled synthetic setting: when the target is a scalar induced from a latent distribution, how should a predictor use N repeated samples drawn from a related observable distribution? A latent mean $\mu_{\text{latent}} \sim U[-1, 1]$ and an ambiguity score $a \sim U[0, 1]$ govern the data. Ambiguity determines a response scale $\sigma(a) = 0.1 + 0.5\,a$ through which $N = 8$ repeated responses $y'_1, \ldots, y'_N \sim \mathcal{N}(\mu_{\text{latent}}, \sigma(a)^2)$ are drawn. The synthetic human target is

$$y^* = \mu_{\text{latent}} + \beta\, a + \gamma\, (\mu_{\text{latent}}\, a),$$

with $\beta = 1$ and $\gamma = 1$. A control case sets $\beta = \gamma = 0$, so the target depends only on the latent mean. All methods use the same two-hidden-layer MLP ($[32, 16]$, ReLU) and are averaged over five seeds. We compare five instantiations of the LLM-generated responses $\mathcal{E}_{\text{LLM}}$ from Section 3.1, each

constructed from $\mathbf{Y}'_N$ and paired with a regression head g_θ, split by how much of the LLM response distribution they retain. Two *scalar-summary* baselines collapse $\mathbf{Y}'_N$ to a single number: `One` uses a single draw y'_1, discarding every other sample; and `Mean` uses the first moment $\bar{y}'_N = \frac{1}{N}\sum_{n=1}^{N} y'_n$, which aggregates the N draws but retains only their mean and discards variance, tails, and higher-order shape. Three *distribution-level* strategies retain information beyond the mean: `Vector` uses the full vector $\mathbf{Y}'_N = (y'_1, \ldots, y'_N)$; `Stats` uses six order-invariant statistics $(\bar{y}'_N, s^2_N, y'_{\min}, y'_{\max}, q_{0.25}, q_{0.75})$, with s^2_N the sample variance and q_α the α-quantile; and `Dist` shares $(y'_1, \ldots, y'_N)$ but emits a Gaussian $\hat{y}^* \sim \mathcal{N}(\mu_\theta, \exp\rho_\theta)$ over y^*. `One`, `Mean`, `Vector`, and `Stats` train with MSE $\mathcal{L} = (g_\theta(\cdot) - y^*)^2$; `Dist` trains with Gaussian NLL $\mathcal{L} = \frac{1}{2}\rho_\theta + (y^* - \mu_\theta)^2/(2\exp\rho_\theta)$.

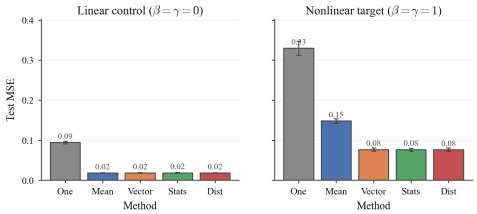


Figure 2. Toy benchmark test MSE across five methods ($N = 8$, five seeds). Under the linear control, `Mean` matches the three distribution-level strategies (`Vector`, `Stats`, `Dist`) and only `One` lags; on the nonlinear target, the distribution-level family jointly achieves the best performance while both scalar-summary baselines (`One`, `Mean`) fall behind.

Figure 2 reports the results. Under the linear control, where the target depends only on the latent mean μ_{latent}, `Mean` and the three distribution-level strategies (`Vector`, `Stats`, `Dist`) all drop to ≈ 0.02 test MSE while `One` remains at 0.09. Once the target becomes nonlinear, `Mean` falls behind, with the distribution-level family clustering at $\approx$ 0.077 test MSE (halving `Mean`'s 0.148) while `One`'s 0.330 is roughly $4\times$ larger. Retaining distribution-level information about $\mathbf{Y}'_N$ thus matters precisely when the target depends on features beyond the mean, while a first-moment summary suffices otherwise. The rest of this section formalizes the three method families.

3.3. From Scalar to Vector: Multi-Sample Debiasing

Existing LLM-debiasing pipelines feed the corrective signal as a single scalar $\mathcal{E}_{\text{LLM}} = y'_1$ (Aher et al., 2023; Wang et al., 2026; Zhang et al., 2025a; Audinet de Pieuchon et al., 2025), the `One` baseline in our terminology. Replacing this with $\mathcal{E}_{\text{LLM}} = \mathbf{Y}'_N$ adds queries at negligible cost relative to human-response collection, and Section 3.2 confirms the replacement is informative whenever the human target depends on distributional features of $\hat{P}$ beyond the mean.

We adopted `Vector` instead of other distributional modeling (e.g., `Dist`) for two reasons. First, $P_h(\cdot \mid x, z)$ routinely violates Gaussian assumptions, with bimodal Likert opinions and bounded skewed responses (Toubia et al., 2025; Santurkar et al., 2023; Kolluri et al., 2025), so any parametric head incurs a misspecification floor. Second, only one human observation per (x, z) pair leaves a Gaussian's variance head without direct supervision, making `Dist` a strictly harder inference problem than the data supports.

Let $j \in \{1, \ldots, J\}$ index questions and $i \in \{1, \ldots, I\}$ index respondents, so that a question-respondent pair is written (x_j, z_i). Let ψ denote the frozen text encoder, with $\mathbf{x}_{q,j} = \psi(x_j)$ and $\mathbf{z}_{d,i} = \psi(z_i)$ shared across training and test, so only the head g_θ carries trainable parameters. We instantiate two debiasing tasks at complementary levels of granularity. *Population-level* debiasing targets the population-average response $\mu_{h,j} := \mathbb{E}[y_h \mid x_j]$ to question x_j, while *individual-level* debiasing targets a specific respondent's answer $y_{h,ij} \sim P_h(\cdot \mid x_j, z_i)$, with the respondent's demographics z_i as an additional feature. Both tasks share a single upstream sampling step and no additional LLM queries are made beyond it: for every pair (i, j) we collect an N-sample vector

$$\begin{aligned} \mathbf{Y}'_{N,ij} &= (y'_{ij,1}, \ldots, y'_{ij,N}), \\ y'_{ij,n} &\sim \hat{P}_n(\cdot \mid x_j, z_i), \quad n = 1, \ldots, N. \end{aligned}$$

The N draws can come from repeated queries to a single LLM, from N different LLMs once each, or any mixture, and the method is agnostic. Our main setting adopts the second scheme, indexing $n \in \{1, \ldots, N\}$ as a *model index* where draw n is the response from the n-th of N distinct models. Cross-model sampling yields less correlated draws and broader output coverage than single-model repeats (Kirk et al., 2024; Mohammadi, 2024; Zhang et al., 2025b); heterogeneous models additionally let $\mathbf{Y}'_{N,ij}$ encode between-model epistemic uncertainty rather than biases specific to one model (Abbasi-Yadkori et al., 2024; Xia et al., 2025; Herrera-Poyatos et al., 2025; Hamidieh et al., 2026).

For *individual-level* debiasing, the total number of samples is $I \times J$, where the unit of observation is a (respondent, question) pair (i, j) with target $y^*_{ij} = y_{h,ij}$. The features used for debiasing $\mathbf{f}_{ij}$ include the question and respondent embeddings and the LLM output `Vector` $\mathbf{Y}'_{N,ij} \sim \mathbb{R}^N$:

$$\begin{aligned} \mathbf{f}_{ij} = \big[&\psi(\mathbf{x}_{q,j}), \\ &\psi(\mathbf{z}_{d,i}),\ \mathbf{Y}'_{N,ij}\big]. \end{aligned} \tag{2}$$

For *population-level* debiasing, the total number of samples is J, where the unit of observation is question $j \in \mathcal{J}$ with target $y^*_j = \mu_{h,j}$. To isolate respondent-driven variation from cross-model disagreement, we fix a model index $n_0 \in \{1, \ldots, N\}$ and draw a subset of $K \le I$ respondents

$i_1, \ldots, i_K$, taking the n_0-th (i.e., the specific LLM source) entry $y'_{i_k j, n_0}$ from the already-collected $\mathbf{Y}'_{N, i_k j}$ and stacking them into

$$\mathbf{Y}'_{K,j} \;=\; (y'_{i_1 j,\, n_0},\; y'_{i_2 j,\, n_0},\; \ldots,\; y'_{i_K j,\, n_0}).$$

We ablate K in Appendix C.3, whether to pool across LLM sources in Appendix C.2, and the choice of a single source n_0 (e.g., GPT-4o) in Appendix C.4. The feature concatenates the question embedding with this stacked sample,

$$\mathbf{f}_j \;=\; \big[\psi(\mathbf{x}_{q,j}),\; \mathbf{Y}'_{K,j}\big]. \tag{3}$$

We treat both tasks as scalar regression on human and LLM responses normalized to $[0, 1]$. Classification is unsuitable because responses are ordinal and the admissible option count varies across items even within a single dataset (Toubia et al., 2025; Augenstein et al., 2018; Kim et al., 2018), while a continuous-space regressor handles heterogeneous targets with a single shared head (Spyromitros-Xioufis et al., 2016). The head g_θ is trained on T labeled units (a training split of J questions for population-level, or of $I \times J$ pairs for individual-level) by ℓ_2-regularized mean squared error,

$$\mathcal{L}_{\text{MSE}}(\theta) \;=\; \frac{1}{T}\sum_{t=1}^{T}\big(g_\theta(\mathbf{f}_t) - y_t^*\big)^2 \;+\; \lambda\,\|\theta\|^2, \tag{4}$$

where the training unit t indexes questions (subset of J) for population-level debiasing and respondent-question pairs (subset of $J \times I$) for individual-level debiasing. Under this squared-loss objective, Section 4 provides an information-theoretic motivation for the design choice via weak Bayes-risk dominance of raw-vector estimators over any measurable summary of $\mathbf{Y}'_N$, including the Gaussian head's two-parameter output, holding for any joint distribution of $\mathbf{Y}'_N$ and therefore covering the heterogeneous-source settings used in Section 5.

4. An Information-Theoretic View

We motivate the design choice between retaining the multi-sample vector and compressing it. Under squared loss with scalar supervision y^*, the relevant object is the conditional mean of y^* given the available features. Let C denote the shared non-LLM covariates (question-side at the population level, additionally including respondent-side at the individual level), $\mathbf{Y}'_N := (y'_1, \ldots, y'_N)$ the multi-sample vector, and $\bar{y}'_N := \frac{1}{N}\sum_n y'_n$ its mean. Define

$$F_{\text{V}} := (C, \mathbf{Y}'_N), \qquad F_{\text{M}} := (C, \bar{y}'_N), \qquad F_1 := (C, y'_1),$$

and the population Bayes risk $\mathcal{R}^*(F) := \inf_{g\ \text{measurable}} \mathbb{E}\big[(y^* - g(F))^2\big]$, an information-level quantity independent of any finite-sample procedure.

The following is a direct consequence of the tower property of conditional expectation; we state it explicitly because it pinpoints exactly when the full vector strictly improves over a summary.

Theorem 4.1 (Bayes dominance of the full response vector)**.** *Assume $y^* \in L^2$. Let $S = S(\mathbf{Y}'_N)$ be any measurable summary, $F_S := (C, S(\mathbf{Y}'_N))$, $m_{\text{V}} := \mathbb{E}[y^* \mid F_{\text{V}}]$, and $m_S := \mathbb{E}[y^* \mid F_S]$. Then*

$$\mathcal{R}^*(F_S) - \mathcal{R}^*(F_{\text{V}}) = \mathbb{E}\big[(m_{\text{V}} - m_S)^2\big] = \mathbb{E}[\text{Var}(m_{\text{V}} \mid F_S)] \geq 0,$$

with equality if and only if m_{V} is $\sigma(F_S)$-measurable up to null sets. In particular, $\mathcal{R}^*(F_{\text{V}}) \leq \mathcal{R}^*(F_{\text{M}})$ and $\mathcal{R}^*(F_{\text{V}}) \leq \mathcal{R}^*(F_1)$.

The proof imposes no distributional assumption on $\mathbf{Y}'_N$, so the inequality holds for any joint law, whether i.i.d., heterogeneous, or dependent. The gap is strict whenever y^* depends on supra-mean features (variance, disagreement, tails, model-specific deviations); in our experiments, entries of $\mathbf{Y}'_N$ come from N distinct LLMs (individual level) or K distinct respondent profiles (population level), so such between-source variation is present by construction. The full proof is given in Appendix B.

Mean versus One are not nested. A more interesting structural observation is that, beyond the vector-versus-summary comparison, neither F_{M} nor F_1 is a measurable function of the other, so no unconditional ordering between Mean and One exists. The risk gap depends on which moment of the response distribution carries the predictive signal. $\mathcal{R}^*(F_{\text{M}}) \leq \mathcal{R}^*(F_1)$ holds under the orthogonality condition $\mathbb{E}[m_{\text{V}} - m_{\text{M}} \mid F_1] = 0$ a.s. (Theorem B.3), and this condition fails on simple constructions in which $|y'_1|$ determines y^* yet $\bar{y}'_N$ does not, giving $\mathcal{R}^*(F_1) < \mathcal{R}^*(F_{\text{M}})$ (Theorem B.4); intuitively, the magnitude of a single draw can encode dispersion that the sample mean averages away. Conversely, when the y'_n are conditionally i.i.d. noisy measurements of a latent scalar, averaging reduces variance and Mean dominates One. Real LLM rollouts span both regimes, since heterogeneous prompts and decoding can induce either mean-informative or dispersion-informative responses. Hence committing to a fixed compression a priori is unsafe, which is what motivates retaining the full vector and letting the downstream debiasing module learn the relevant projection from data. Formal statements are deferred to Appendix B.

5. Experiments

We evaluate our framework on three benchmarks at both the population and individual levels. Section 5.1 introduces the benchmarks, tasks, and evaluation metrics. Section 5.2 reports the main pointwise comparison of `Vector` against four LLM-input variants, and Section 5.3 extends

it to distributional alignment. Additional experiments are deferred to Appendix C, including an SFT comparison (Appendix C.1), multi-LLM rollouts beyond the canonical `gpt-4o` source (Appendix C.2), a rollout-size K ablation (Appendix C.3), and a per-source robustness study (Appendix C.4); the full benchmark and training setup is in Appendix D.

5.1. Benchmarks and Setups

We evaluate on three public benchmarks, Twin-2K-500 (Toubia et al., 2025), OpinionQA (Santurkar et al., 2023), and EEDI (Wang et al., 2021), each instantiated at both the population-level and individual-level tasks of Section 3. The frozen encoder ψ is OpenAI `text-embedding-3-small` (OpenAI, 2024) with output dimension 256, the number of source LLMs is N=8, and both tasks use a random $80/20$ train-test split. Full preprocessing detail is deferred to Appendix D.

Twin-2K-500. Twin-2K-500 collects 665 questions from 2,058 U.S. respondents across four survey waves, for a total of 1,222,466 respondent-question responses. We filter the item pool to Likert-style questions, dropping open-ended and numeric items whose admissible range exceeds 9 categories, leaving 166,710 respondent-question samples for the individual-level task and 105 question-level samples for the population-level task.

OpinionQA. OpinionQA, drawn from Pew Research's American Trends Panel, contains 385 five-point opinion items answered by 32,864 respondents (1,476,868 human answers in total). Following Huang et al. (2025), we restrict the persona pool to the 200 respondents with complete LLM responses support, leaving 9,437 respondent-question samples for the individual-level task and 385 question-level samples for the population-level task.

EEDI. EEDI originates from the NeurIPS 2020 Education Challenge and consists of multiple-choice math diagnostics with options $\{1, 2, 3, 4\}$. We start from the processed release of Huang et al. (2025) (412 items with ≥ 100 responses and no figure-dependent content), cluster students by profile, and take the per-(profile, question) mean answer as the individual-level target y^*, yielding 3,479 individual-level samples and 411 population-level samples.

Main Metrics. We report four metrics, each an expectation over the evaluation set $\mathcal{D}_{\text{eval}}$, Mean Absolute Error (`MAE`), Normalized Accuracy (`NAcc`), Hit Accuracy (`HA`), and Soft Accuracy (`SA`), with admissible category set $\mathcal{C} = \{a, \dots, b\}$ varying per question. All four are reported on the normalized $[0, 100]$ scale (per-question responses rescaled from $\mathcal{C}$ to $[0, 1]$ then multiplied by 100), making values directly comparable across datasets with different Likert lengths (Twin-2K-500 $b{-}a \in \{4, 6, 8\}$, OpinionQA $b{-}a$=4, EEDI $b{-}a$=3). For `SA`, let $\rho(\hat{y}) = \arg\min_{c \in \mathcal{C}} |\hat{y} - c|$ and $c_1, c_2 \in \mathcal{C}$ be the two nearest admissible categories to $\hat{y}$; set $w(\hat{y}; \hat{y})$=1 when $\hat{y} \in \mathcal{C}$, otherwise $w(c_1; \hat{y})$=$|\hat{y} - c_2|$ and $w(c_2; \hat{y})$=$|\hat{y} - c_1|$ with zero on the rest (these sum to one since $\mathcal{C}$ is unit-spaced).

$$
\begin{aligned}
\texttt{MAE}\,(\downarrow) &= \mathbb{E}_{\mathcal{D}_{\text{eval}}}\big[|\hat{y} - y|\big], \\
\texttt{NAcc}\,(\uparrow) &= \mathbb{E}_{\mathcal{D}_{\text{eval}}}\Big[1 - \tfrac{|\hat{y}-y|}{b-a}\Big], \\
\texttt{HA}\,(\uparrow) &= \mathbb{E}_{\mathcal{D}_{\text{eval}}}\big[\mathbf{1}\{\rho(\hat{y}) = \rho(y)\}\big], \\
\texttt{SA}\,(\uparrow) &= \mathbb{E}_{\mathcal{D}_{\text{eval}}}\Big[\textstyle\sum_{c \in \mathcal{C}} w(c; \hat{y})\, w(c; y)\Big].
\end{aligned}
$$

`MAE` and `NAcc` capture numerical fit; `HA` and `SA` capture agreement with bounded discrete categories, with `SA` giving partial credit when $\hat{y}$ falls between adjacent categories.

5.2. Main Population and Individual-level Results

Prior post-hoc LLM-debiasing methods feed a single LLM response as the corrective signal (Wang et al., 2026; Zhang et al., 2025a; Audinet de Pieuchon et al., 2025; Wang et al., 2025b) and therefore reduce to our single-sample `One` baseline. We highlight that these methods target downstream parameter estimation rather than response prediction, so *are not directly comparable* on the metrics we evaluate (Section 2). Since accurate response prediction is positively correlated with accurate downstream parameter estimation, the natural axis of comparison within the response-prediction setting is how much LLM-side signal each method retains. We therefore compare four feature constructions, `w/o LLM`, `One`, `Mean`, and `Vector`, sharing a common downstream head and differing only in their LLM-side input. Table 1 shows that `Vector` attains the best or tied-best result in 22 of 24 cells (3 datasets $\times$ 2 levels $\times$ 4 metrics), with 2 ties against `Mean`.

Population-level results. The left block of Table 1 shows population-level results with LLM source `gpt-4o` (Hurst et al., 2024) and compares our method against four baselines. We study the LLM source choice in Appendix C.2 and ablate the rollout size N in Appendix C.3. `Base LLM` is the *pre-debiasing* reference: the scalar average of N=50 persona-conditioned `gpt-4o` responses, used directly as the population-mean prediction without any debiasing process. `w/o LLM` denotes the learned predictor with no LLM feature appended. `One` appends a single `gpt-4o` response drawn from a population-level prompt (see Appendix D.2) as the LLM feature. `Mean` appends the mean of the K=50 persona-conditioned responses. Our method `Vector` appends the full K=50 persona-conditioned response vector as the LLM feature. `Vector` attains the best (one tied-best) performance on four metrics. The largest margin is on Twin-2K-500, where `MAE` drops by 44.0% (62.0 $\rightarrow$ 34.7)

Table 1. Main results at the population and individual levels. Cells are mean ± s.d. across 5 seeds. Gray rows mark `Base LLM` before correction; blue cells mark the best value. `Vector` attains the best or tied-best result in 22 of 24 cells (3 datasets × 2 levels × 4 metrics), with 2 ties against `Mean`.

		Population-level				Individual-level (80% subsample)			
Dataset	Representation	MAE (↓)	NAcc (↑)	HA (↑)	SA (↑)	MAE (↓)	NAcc (↑)	HA (↑)	SA (↑)
Twin-2K-500	Base LLM	62.0 ± 0.0	86.3 ± 0.0	42.7 ± 0.0	42.5 ± 0.0	58.5 ± 0.0	76.4 ± 0.0	59.7 ± 0.0	57.5 ± 0.0
	w/o LLM	67.4 ± 2.0	87.4 ± 0.4	43.1 ± 3.5	40.7 ± 1.1	47.6 ± 0.5	81.8 ± 0.7	64.2 ± 0.4	62.0 ± 2.1
	One	62.5 ± 7.2	88.3 ± 1.5	49.2 ± 4.8	43.1 ± 2.5	46.1 ± 0.3	83.5 ± 0.2	65.8 ± 0.5	63.6 ± 1.9
	Mean	57.2 ± 9.5	89.4 ± 1.9	50.0 ± 4.7	44.7 ± 3.1	45.3 ± 1.1	84.0 ± 0.9	67.5 ± 0.4	64.4 ± 0.3
	Vector	34.7 ± 4.2	93.3 ± 0.9	65.8 ± 7.1	52.2 ± 2.6	44.2 ± 0.2	84.5 ± 0.1	68.4 ± 0.3	66.2 ± 0.2
OpinionQA	Base LLM	62.1 ± 0.0	84.5 ± 0.0	44.9 ± 0.0	40.9 ± 0.0	89.0 ± 0.0	77.7 ± 0.0	43.3 ± 0.0	43.3 ± 0.0
	w/o LLM	46.8 ± 0.3	88.3 ± 0.1	59.5 ± 2.8	47.5 ± 1.0	68.5 ± 0.2	82.9 ± 0.5	44.6 ± 0.6	43.1 ± 0.8
	One	44.3 ± 1.1	88.9 ± 0.3	61.3 ± 4.0	50.0 ± 0.6	69.1 ± 1.1	82.7 ± 1.3	43.4 ± 0.7	41.8 ± 1.3
	Mean	35.5 ± 2.1	91.1 ± 0.5	69.3 ± 1.7	53.8 ± 0.6	70.2 ± 2.3	82.4 ± 2.1	44.4 ± 1.1	42.5 ± 1.1
	Vector	30.5 ± 1.4	92.4 ± 0.3	69.3 ± 2.0	57.0 ± 1.1	65.3 ± 0.7	84.3 ± 1.1	46.4 ± 0.9	45.1 ± 0.8
EEDI	Base LLM	61.7 ± 0.0	79.4 ± 0.0	46.4 ± 0.0	41.3 ± 0.0	34.5 ± 0.0	88.5 ± 0.0	68.4 ± 0.0	55.4 ± 0.0
	w/o LLM	38.4 ± 1.4	87.2 ± 0.5	53.7 ± 7.8	47.4 ± 0.7	10.5 ± 0.2	96.5 ± 0.8	87.6 ± 0.7	62.6 ± 1.3
	One	36.2 ± 0.6	87.9 ± 0.2	58.8 ± 3.8	50.0 ± 0.6	11.9 ± 0.2	96.0 ± 1.4	85.6 ± 1.6	62.5 ± 1.1
	Mean	30.7 ± 1.6	89.8 ± 0.5	71.1 ± 5.6	54.3 ± 1.4	11.0 ± 0.1	96.3 ± 0.8	85.5 ± 1.1	63.0 ± 0.8
	Vector	25.6 ± 1.3	91.5 ± 0.4	74.5 ± 5.8	58.1 ± 2.0	10.1 ± 0.3	96.9 ± 0.3	85.5 ± 1.3	63.1 ± 0.9

and HA rises by 54.1% (42.7 → 65.8) relative to Base LLM. Replacing Mean with Vector improves all four metrics on every dataset, with the only exception of OpinionQA HA, demonstrating the valuable information beyond the mean for our debiasing tasks.

Individual-level results. The right block of Table 1 compares our method against four baselines at the individual level. Base LLM (*pre-debiasing* reference) is a single persona-conditioned gpt-4o response for each (respondent, question) pair, used directly as the prediction of $y_{h,ij}$, and w/o LLM retains only respondent and question features. One appends a single persona-conditioned gpt-4o response, while Mean and our method Vector use the mean and the full vector of N=8 persona-conditioned responses drawn from a suite of eight distinct LLMs (one per model). Vector attains the best or tied-best result in 11 of the 12 individual-level cells, with the sole loss to w/o LLM on EEDI HA. The largest relative gain is on EEDI, where MAE drops from 34.5 (Base LLM) to 10.1 (−70.7%); comparable but smaller drops appear on Twin-2K-500 (58.5 → 44.2, −24.4%) and OpinionQA (89.0 → 65.3, −26.6%). Against Mean, Vector edges every metric on Twin-2K-500 and OpinionQA, and three of four on EEDI (with HA tied at 85.5).

The monotonic One→Mean→Vector ordering holds on Twin-2K-500 and EEDI, matching the Bayes-risk ranking of Theorem 4.1 and the toy benchmark of Section 3.2; on OpinionQA, Mean dips slightly below One on MAE (70.2 vs. 69.1) and NAcc (82.4 vs. 82.7) before Vector recovers to the strict best. The absolute Vector versus Mean gap is also notably tighter at the individual level (e.g., Twin-2K-500 MAE gap of −22.5 at population versus −1.1 at individual), because the N=8 same-persona draws used for individual-level debiasing are more correlated than the K=50 heterogeneous persona draws used in population-level debiasing and leave less supra-mean signal for Vector to recover, while individual prediction is inherently harder because the predictor must additionally condition on the respondent embedding $\mathbf{z}_d$, further capping any LLM-side feature gain.

5.3. Distributional Alignment Beyond Pointwise Accuracy

To assess whether a predictor preserves the cross-respondent response distribution for each question, we in-

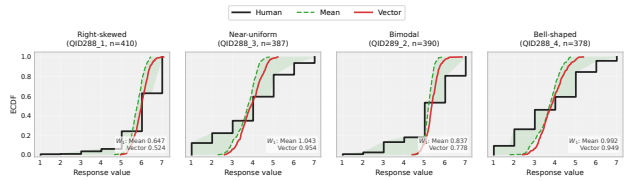


Figure 3. Per-question ECDF of predicted vs. human responses on four Twin-2K-500 questions (diverse distribution shapes, $n \approx 400$ per question). `Vector` (red solid) tracks the human step function more closely than `Mean` (green dashed); shading marks the Mean–Human gap. W_1 values in each panel confirm the improvement.

troduce three complementary metrics capturing respondent ordering, mean bias, and full distributional discrepancy. The per-question Pearson correlation $\bar{r}$ measures whether the predictor preserves the relative ordering of respondents. The mean bias $|\overline{\Delta}|$ measures the absolute difference between predicted and human question-level means. The 1-Wasserstein distance measures the empirical gap between the predicted and human response distributions for each question.

Figure 3 shows per-question ECDFs for four questions from Twin-2K-500, chosen to represent diverse distribution shapes. Across right-skewed, near-uniform, bimodal, and bell-shaped human distributions, `Vector` (red) tracks the human step function more closely than `Mean` (green), with a consistently lower Wasserstein distance in each panel. The shaded region shows the gap between `Mean` and `Human` that `Vector` partially closes. Full numerical results across all three datasets and all metrics are given in Table 6 (Appendix C.5). `Vector` attains the best value on 8 of the 9 dataset–metric cells; the largest gains are on EEDI ($\bar{r}$ rises from 0.197 for `Mean` to 0.750 for `Vector`). The full table and metrics formulas and details are provided in Appendix C.5.

6. Conclusion

We recast post-hoc LLM debiasing as a regression problem over the multi-sample vector of LLM responses, exploiting the structural asymmetry that the human distribution is observed once whereas the LLM distribution is freely queryable. Under squared loss, a vector estimator weakly dominates any estimator on a measurable summary, with strict improvement when the target depends on features beyond the mean. Across three benchmarks, our method attains the best value on the majority of metric cells at both the population and individual levels, with `MAE` reductions over the uncorrected Base LLM of up to 58% at the population level and up to 71% at the individual level (EEDI), and consistent distributional-alignment gains over scalar-input baselines.

Limitations and future work: First, existing datasets supply only a single scalar response per respondent, constraining the target to a scalar and introducing bias relative to the latent distribution; collecting repeated responses would reduce this bias but is prohibitively costly, a limitation shared by all current post-hoc methods. Second, how the population-level rollout size K should be chosen remains open. Third, the text encoder ψ is frozen and task-agnostic, and jointly adapting it on human responses could close part of the remaining gap.

References

Abbasi-Yadkori, Y., Kuzborskij, I., György, A., and Szepesvári, C. To believe or not to believe your LLM: iterative prompting for estimating epistemic uncertainty. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6aebba00fff5b6de7b488e496f80edd7-Abstract-Confere html.

Aher, G. V., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 337–371. PMLR, 2023. URL https://proceedings.mlr.press/v202/aher23a.html.

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., Evans, J., Brynjolfsson, E., and Bernstein, M. LLM Social Simulations Are a Promising Research Method, 2025. URL https://arxiv.org/abs/2504.02234.

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.

Audinet de Pieuchon, N., Daoud, A., Jerzak, C. T., Johansson, M., and Johansson, R. Benchmarking debiasing methods for llm-based parameter estimates. *arXiv e-prints*, pp. arXiv–2506, 2025.

Augenstein, I., Ruder, S., and Søgaard, A. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1896–1906, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1172. URL https://aclanthology.org/N18-1172.

Bhattacharyya, A., Borah, A., Singla, Y. K., Shah, R. R., Chen, C., and Krishnamurthy, B. Social agents: Collective intelligence improves llm predictions. In *The Fourteenth International Conference on Learning Representations*, 2026.

Bini, P., Cong, L. W., Huang, X., and Jin, L. J. Behavioral economics of ai: Llm biases and corrections. *Available at SSRN 5213130*, 2025.

Binz, M., Akata, E., Bethge, M., Brändle, F., Callaway, F., Coda-Forno, J., Dayan, P., Demircan, C., Eckstein, M. K., Éltető, N., et al. A foundation model to predict and capture human cognition. *Nature*, 644(8078):1002–1009, 2025.

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.

Brand, J., Israeli, A., and Ngwe, D. Using GPT for market research. *Available at SSRN 4395751*, 2023.

Chen, J., Xu, R., Cao, B., Pan, R., Zhang, Y., Hu, Y., Du, Y., Gao, T., Lu, Y., Sun, Y., et al. Towards real-world human behavior simulation: Benchmarking large language models on long-horizon, cross-scenario, heterogeneous behavior traces. *ArXiv preprint*, abs/2604.08362, 2026. URL https://arxiv.org/abs/2604.08362.

Chen, Y., Kirshner, S. N., Ovchinnikov, A., Andiappan, M., and Jenkin, T. A manager and an ai walk into a bar: does chatgpt make biased decisions like we do? *Manufacturing & Service Operations Management*, 27(2):354–368, 2025.

Cheng, M., Piccardi, T., and Yang, D. CoMPosT: Characterizing and evaluating caricature in LLM simulations. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10853–10875, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.669. URL https://aclanthology.org/2023.emnlp-main.669.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Cui, Z., Li, N., and Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8):627–634, 2025.

Dominguez-Olmedo, R., Hardt, M., and Mendler-Dünner, C. Questioning the survey responses of large language models. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/515c62809e0a29729d7eec26e2916fc0-Abstract-Confere html.

Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., et al. Towards measuring the representation of subjective global opinions in language models. *ArXiv preprint*, abs/2306.16388, 2023. URL https://arxiv.org/abs/2306.16388.

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d862f7f5445255090de13b825b880d59-Abstract-Confere html.

Goli, A. and Singh, A. Frontiers: Can large language models capture human preferences? *Marketing Science*, 43 (4):709–722, 2024. doi: 10.1287/mksc.2023.0306.

Guilbeault, D., Delecourt, S., and Desikan, B. S. Age and gender distortion in online media and large language models. *Nature*, 646(8087):1129–1137, 2025.

Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., and Khot, T. Bias runs deep: Implicit reasoning biases in persona-assigned llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May*

7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=kGteeZ18Ir.

Hamidieh, K., Thost, V., Gerych, W., Yurochkin, M., and Ghassemi, M. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lOoRJo8xWy.

Herrera-Poyatos, D., Peláez-González, C., Zuheros, C., Herrera-Poyatos, A., Tejedor, V., Herrera, F., and Montes, R. An overview of model uncertainty and variability in llm-based sentiment analysis. challenges, mitigation strategies and the role of explainability, 2025. URL https://arxiv.org/abs/2504.04462.

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

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

Huang, C., Wu, Y., and Wang, K. How many human survey respondents is a large language model worth? an uncertainty quantification perspective. *ArXiv preprint*, abs/2502.17773, 2025. URL https://arxiv.org/abs/2502.17773.

Huang, M., Zhang, X., Soto, C., and Evans, J. Designing ai-agents with personalities: A psychometric approach. *Personality Science*, 7:27000710251406471, 2026.

Huang, M.-H. and Rust, R. T. The genai future of consumer research. *Journal of Consumer Research*, 52(1):4–17, 2025.

Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. GPT-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Jiang, L., Chai, Y., Li, M., Liu, M., Fok, R., Dziri, N., Tsvetkov, Y., Sap, M., Albalak, A., and Choi, Y. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *ArXiv preprint*, abs/2510.22954, 2025. URL https://arxiv.org/abs/2510.22954.

Kang, M., Moon, S., Lee, S. H., Raj, A., Suh, J., Chan, D. M., and Canny, J. Deep binding of language model virtual personas: a study on approximating political partisan misperceptions. *ArXiv preprint*, abs/2504.11673, 2025. URL https://arxiv.org/abs/2504.11673.

Kim, D.-J., Choi, J., Oh, T.-H., Yoon, Y., and Kweon, I. S. Disjoint multi-task learning between heterogeneous human-centric tasks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1699–1708. IEEE, 2018.

Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., and Raileanu, R. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

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*, 2025.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Lei, Y., Wang, T., Lian, J., Hu, Z., Lian, D., and Xie, X. Humanllm: Towards personalized understanding and simulation of human nature. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 621–632, 2026.

Leng, Y., Sang, Y., and Agarwal, A. Reduce disparity between llms and humans: Optimal llm sample calibration. *Available at SSRN 4802019*, 2024.

Li, A., Chen, H., Namkoong, H., and Peng, T. LLM generated persona is a promise with a catch. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL https://openreview.net/forum?id=qh9eGtMG4H.

Li, P., Castelo, N., Katona, Z., and Sarvary, M. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43 (2):254–266, 2024.

Lin, R. F., Tian, K., Zheng, H., Zhang, C., Zeng, L., and Huang, S. Crowdllm: Building llm-based digital populations augmented with generative models. *ArXiv preprint*, abs/2512.07890, 2025. URL https://arxiv.org/abs/2512.07890.

Lippert, S., Dreber, A., Johannesson, M., Tierney, W., Cyrus-Lai, W., Uhlmann, E. L., Pfeiffer, T., Collaboration, E. E., et al. Can large language models help predict

results from a complex behavioural science study? *Royal Society Open Science*, 11(9), 2024.

Lu, Y., Huang, J., Han, Y., Yao, B., Bei, S., Gesi, J., Xie, Y., He, Q., Wang, D., et al. Can llm agents simulate multi-turn human behavior? evidence from real online customer behavior data. *ArXiv preprint*, abs/2503.20749, 2025. URL https://arxiv.org/abs/2503.20749.

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*, abs/2510.08338, 2025. URL https://arxiv.org/abs/2510.08338.

Mei, Q., Xie, Y., Yuan, W., and Jackson, M. O. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.

Mohammadi, B. Creativity has left the chat: The price of debiasing language models. *ArXiv preprint*, abs/2406.05587, 2024. URL https://arxiv.org/abs/2406.05587.

OpenAI. Vector embeddings. https://developers.openai.com/api/docs/guides/embeddings, 2024. OpenAI API documentation for text-embedding-3-small and text-embedding-3-large; accessed 2026-04-22.

Paglieri, D., Cross, L., Cunningham, W. A., Leibo, J. Z., and Vezhnevets, A. S. Persona generators: Generating diverse synthetic personas at scale. *ArXiv preprint*, abs/2602.03545, 2026. URL https://arxiv.org/abs/2602.03545.

Park, J. S., Zou, C. Q., Kamphorst, J., Egan, N., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., Liang, P., Willer, R., and Bernstein, M. S. Llm agents grounded in self-reports enable general-purpose simulation of individuals, 2026. URL https://arxiv.org/abs/2411.10109.

Peng, T. et al. Digital twins as funhouse mirrors: Five key distortions, 2025. URL https://arxiv.org/abs/2509.19088.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. HybridFlow: A flexible and efficient RLHF framework, 2024. URL https://arxiv.org/abs/2409.19256.

Spyromitros-Xioufis, E., Tsoumakas, G., Groves, W., and Vlahavas, I. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98, 2016.

Tjuatja, L., Chen, V., Wu, T., Talwalkwar, 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.

Toubia, O., Gui, G. Z., Peng, T., Merlau, D. J., Li, A., and Chen, H. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *ArXiv preprint*, abs/2505.17479, 2025. URL https://arxiv.org/abs/2505.17479.

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, 2025a.

Wang, L., Ye, Z., and Zhao, J. Efficient inference using large language models with limited human data: Fine-tuning then rectification. *ArXiv preprint*, abs/2511.19486, 2025b. URL https://arxiv.org/abs/2511.19486.

Wang, M., Zhang, D. J., and Zhang, H. Large language models for market research: A data-augmentation approach. *Marketing Science*, 2026.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, Y., Hernández-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Barton, C., Jones, S. P., Woodhead, S., and Zhang, C. Results and insights from diagnostic questions: The NeurIPS 2020 education challenge. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pp. 191–205, 2021.

Wang, Z., Lu, Y., Li, W., Amini, A., Sun, B., Bart, Y., Lyu, W., Gesi, J., Wang, T., Huang, J., Su, Y., Ehsan, U., Alikhani, M., Li, T. J.-J., Chilton, L., and Wang, D. OPeRA: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation. *ArXiv preprint*, abs/2506.05606, 2025c. URL https://arxiv.org/abs/2506.05606.

Xia, Z., Xu, J., Zhang, Y., and Liu, H. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21381–21396, 2025.

Yeykelis, L., Pichai, K., Cummings, J. J., and Reeves, B. Using large language models to create ai personas for replication, generalization and prediction of media effects: An empirical test of 133 published experimental research findings. *ArXiv preprint*, abs/2408.16073, 2024. URL https://arxiv.org/abs/2408.16073.

Zhang, B., Li, J., Hortaçsu, A., Ye, X., Chernozhukov, V., Ni, A., and Huang, E. W. Agentic economic modeling. *ArXiv preprint*, abs/2510.25743, 2025a. URL https://arxiv.org/abs/2510.25743.

Zhang, J., Yu, S., Chong, D., Sicilia, A., Tomz, M. R., Manning, C. D., and Shi, W. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *ArXiv preprint*, abs/2510.01171, 2025b. URL https://arxiv.org/abs/2510.01171.

A. Broader Impacts

Positive impacts. Post-hoc debiasing improves the alignment of LLM proxies with human response distributions, supporting more accurate and lower-cost behavioral, marketing, and education research. Better distributional alignment also reduces the demographic flattening and persona homogenization that under-represent minority opinions in single-shot LLM elicitation, with possible downstream benefits for fairness in survey, polling, and pedagogical applications.

Negative impacts and mitigations. Better-aligned synthetic respondents could lower the perceived cost of substituting LLMs for genuine human studies, with risks including displacement of real participant voice, fabricated survey reporting, and false confidence in alignment quality outside the evaluated domains. We mitigate by training only on publicly released benchmarks under their original licenses, framing the released artifacts as a debiasing module on top of frozen LLMs rather than a primary respondent generator, and recommending validation against held-out human samples before any downstream deployment.

B. Additional Theory and Proofs

Throughout the appendix, all random variables are defined on a common probability space. Let C denote the shared non-LLM covariate-side representation used by the debiasing module. In the group-level task, C is the question-side representation; in the individual-level task, C additionally includes the respondent-side representation. Let

$$\mathbf{Y}'_N := (y'_1, \dots, y'_N),$$

and for any measurable summary $S = S(\mathbf{Y}'_N)$, define

$$F_{\mathrm{V}} := (C, \mathbf{Y}'_N), \qquad F_S := (C, S(\mathbf{Y}'_N)).$$

For any feature representation F, define the population Bayes risk under squared loss by

$$\mathcal{R}^*(F) := \inf_{g \text{ measurable}} \mathbb{E}\left[(y^* - g(F))^2\right],$$

where estimators with infinite risk are ignored in the infimum. We write

$$m_F := \mathbb{E}[y^* \mid F]$$

for the squared-loss Bayes estimator associated with F.

Lemma B.1 (Bayes risk under squared loss)**.** *Assume $y^* \in L^2$. For any feature representation F,*

$$\mathcal{R}^*(F) = \mathbb{E}\left[(y^* - m_F)^2\right] = \mathbb{E}[(y^*)^2] - \mathbb{E}[m_F^2] = \mathbb{E}[\mathrm{Var}(y^* \mid F)]\,.$$

Moreover, m_F is the unique Bayes estimator up to almost-sure equivalence.

Proof. Let $g(F)$ be any measurable estimator with finite second moment. Since $m_F = \mathbb{E}[y^* \mid F]$, we have

$$\mathbb{E}[y^* - m_F \mid F] = 0.$$

Expanding the square gives

$$\mathbb{E}[(y^* - g(F))^2] = \mathbb{E}[(y^* - m_F)^2] + \mathbb{E}[(m_F - g(F))^2] + 2\mathbb{E}[(y^* - m_F)(m_F - g(F))].$$

The cross term is zero because $m_F - g(F)$ is F-measurable:

$$\mathbb{E}[(y^* - m_F)(m_F - g(F))] = \mathbb{E}[\mathbb{E}[y^* - m_F \mid F]\,(m_F - g(F))] = 0.$$

Therefore,

$$\mathbb{E}[(y^* - g(F))^2] = \mathbb{E}[(y^* - m_F)^2] + \mathbb{E}[(m_F - g(F))^2].$$

The risk is minimized by $g(F) = m_F$ almost surely, proving the Bayes optimality and almost-sure uniqueness of m_F.

Next,

$$\mathbb{E}[(y^* - m_F)^2] = \mathbb{E}[(y^*)^2] - 2\mathbb{E}[y^* m_F] + \mathbb{E}[m_F^2].$$

Since m_F is F-measurable,

$$\mathbb{E}[y^* m_F] = \mathbb{E}[\mathbb{E}[y^* m_F \mid F]] = \mathbb{E}[m_F \, \mathbb{E}[y^* \mid F]] = \mathbb{E}[m_F^2].$$

Hence,

$$\mathcal{R}^*(F) = \mathbb{E}[(y^*)^2] - \mathbb{E}[m_F^2].$$

Finally,

$$\mathbb{E}[(y^* - m_F)^2 \mid F] = \mathrm{Var}(y^* \mid F),$$

and taking expectations yields

$$\mathcal{R}^*(F) = \mathbb{E}[\mathrm{Var}(y^* \mid F)].$$

□

Proof of Theorem 4.1. Let $S = S(\mathbf{Y}'_N)$ be any measurable summary of the repeated LLM-response vector. Since $F_S = (C, S(\mathbf{Y}'_N))$ is a measurable function of $F_{\mathrm{V}} = (C, \mathbf{Y}'_N)$, we have

$$\sigma(F_S) \subseteq \sigma(F_{\mathrm{V}}).$$

Therefore, by the tower property,

$$m_S = \mathbb{E}[y^* \mid F_S] = \mathbb{E}[\mathbb{E}[y^* \mid F_{\mathrm{V}}] \mid F_S] = \mathbb{E}[m_{\mathrm{V}} \mid F_S].$$

By Lemma B.1,

$$\mathcal{R}^*(F_S) - \mathcal{R}^*(F_{\mathrm{V}}) = \Big(\mathbb{E}[(y^*)^2] - \mathbb{E}[m_S^2]\Big) - \Big(\mathbb{E}[(y^*)^2] - \mathbb{E}[m_{\mathrm{V}}^2]\Big),$$

so

$$\mathcal{R}^*(F_S) - \mathcal{R}^*(F_{\mathrm{V}}) = \mathbb{E}[m_{\mathrm{V}}^2] - \mathbb{E}[m_S^2].$$

Using $m_S = \mathbb{E}[m_{\mathrm{V}} \mid F_S]$, we obtain

$$\mathbb{E}[m_{\mathrm{V}} m_S] = \mathbb{E}[\mathbb{E}[m_{\mathrm{V}} m_S \mid F_S]] = \mathbb{E}[m_S \, \mathbb{E}[m_{\mathrm{V}} \mid F_S]] = \mathbb{E}[m_S^2].$$

Hence,

$$\mathbb{E}[(m_{\mathrm{V}} - m_S)^2] = \mathbb{E}[m_{\mathrm{V}}^2] - 2\mathbb{E}[m_{\mathrm{V}} m_S] + \mathbb{E}[m_S^2] = \mathbb{E}[m_{\mathrm{V}}^2] - \mathbb{E}[m_S^2].$$

Therefore,

$$\mathcal{R}^*(F_S) - \mathcal{R}^*(F_{\mathrm{V}}) = \mathbb{E}[(m_{\mathrm{V}} - m_S)^2] \geq 0.$$

Moreover, since $m_S = \mathbb{E}[m_{\mathrm{V}} \mid F_S]$,

$$\mathbb{E}[(m_{\mathrm{V}} - m_S)^2 \mid F_S] = \mathrm{Var}(m_{\mathrm{V}} \mid F_S).$$

Taking expectations gives

$$\mathcal{R}^*(F_S) - \mathcal{R}^*(F_{\mathrm{V}}) = \mathbb{E}[\mathrm{Var}(m_{\mathrm{V}} \mid F_S)].$$

The equality condition follows immediately. The gap is zero if and only if

$$\mathbb{E}[(m_{\mathrm{V}} - m_S)^2] = 0,$$

which holds if and only if

$$m_{\mathrm{V}} = m_S \qquad \text{a.s.}$$

Since m_S is $\sigma(F_S)$-measurable, this is equivalent to m_{V} being $\sigma(F_S)$-measurable up to null sets. This proves the theorem.
□

Corollary B.2 (Vector versus Mean and One)**.** *Let*

$$F_{\mathrm{M}} := (C, \bar{y}'_N), \qquad F_1 := (C, y'_1),$$

where

$$\bar{y}'_N := \frac{1}{N}\sum_{n=1}^{N} y'_n.$$

Define

$$\mathcal{R}_{\mathrm{V}} := \mathcal{R}^*(F_{\mathrm{V}}), \qquad \mathcal{R}_{\mathrm{M}} := \mathcal{R}^*(F_{\mathrm{M}}), \qquad \mathcal{R}_1 := \mathcal{R}^*(F_1),$$

and

$$m_{\mathrm{M}} := \mathbb{E}[y^* \mid F_{\mathrm{M}}], \qquad m_1 := \mathbb{E}[y^* \mid F_1].$$

Then

$$\mathcal{R}_{\mathrm{M}} - \mathcal{R}_{\mathrm{V}} = \mathbb{E}[(m_{\mathrm{V}} - m_{\mathrm{M}})^2] = \mathbb{E}[\mathrm{Var}(m_{\mathrm{V}} \mid F_{\mathrm{M}})],$$

and

$$\mathcal{R}_1 - \mathcal{R}_{\mathrm{V}} = \mathbb{E}[(m_{\mathrm{V}} - m_1)^2] = \mathbb{E}[\mathrm{Var}(m_{\mathrm{V}} \mid F_1)].$$

Consequently,

$$\mathcal{R}_{\mathrm{V}} \le \mathcal{R}_{\mathrm{M}}, \qquad \mathcal{R}_{\mathrm{V}} \le \mathcal{R}_1.$$

Proof. Apply Theorem 4.1 with $S(\mathbf{Y}'_N) = \bar{y}'_N$ and then with $S(\mathbf{Y}'_N) = y'_1$. □

Proposition B.3 (A sufficient condition for Mean to dominate One)**.** *Assume*

$$\mathbb{E}[m_{\mathrm{V}} - m_{\mathrm{M}} \mid F_1] = 0 \qquad \textit{a.s.}$$

Then

$$\mathcal{R}_1 - \mathcal{R}_{\mathrm{M}} = \mathbb{E}\Big[\big(m_{\mathrm{M}} - \mathbb{E}[m_{\mathrm{M}} \mid F_1]\big)^2\Big] \ge 0.$$

Consequently,

$$\mathcal{R}_{\mathrm{M}} \le \mathcal{R}_1.$$

Equality holds if and only if m_{M} is $\sigma(F_1)$-measurable up to null sets.

Proof. By Corollary B.2,

$$m_1 = \mathbb{E}[y^* \mid F_1] = \mathbb{E}[m_{\mathrm{V}} \mid F_1].$$

The assumed orthogonality condition gives

$$\mathbb{E}[m_{\mathrm{V}} - m_{\mathrm{M}} \mid F_1] = 0.$$

Therefore,

$$\mathbb{E}[m_{\mathrm{V}} \mid F_1] = \mathbb{E}[m_{\mathrm{M}} \mid F_1],$$

and hence

$$m_1 = \mathbb{E}[m_{\mathrm{M}} \mid F_1].$$

Using Lemma B.1,

$$\mathcal{R}_1 - \mathcal{R}_{\mathrm{M}} = \Big(\mathbb{E}[(y^*)^2] - \mathbb{E}[m_1^2]\Big) - \Big(\mathbb{E}[(y^*)^2] - \mathbb{E}[m_{\mathrm{M}}^2]\Big),$$

so

$$\mathcal{R}_1 - \mathcal{R}_{\mathrm{M}} = \mathbb{E}[m_{\mathrm{M}}^2] - \mathbb{E}[m_1^2].$$

Substituting $m_1 = \mathbb{E}[m_{\mathrm{M}} \mid F_1]$, we get

$$\mathcal{R}_1 - \mathcal{R}_{\mathrm{M}} = \mathbb{E}[m_{\mathrm{M}}^2] - \mathbb{E}\big[\mathbb{E}[m_{\mathrm{M}} \mid F_1]^2\big].$$

As in the proof of Theorem 4.1,

$$\mathbb{E}[m_{\mathrm{M}}^2] - \mathbb{E}\big[\mathbb{E}[m_{\mathrm{M}} \mid F_1]^2\big] = \mathbb{E}\Big[\big(m_{\mathrm{M}} - \mathbb{E}[m_{\mathrm{M}} \mid F_1]\big)^2\Big].$$

This proves the nonnegative risk gap.

The equality condition follows because the gap is zero if and only if

$$m_{\mathrm{M}} = \mathbb{E}[m_{\mathrm{M}} \mid F_1] \qquad \text{a.s.},$$

which is equivalent to m_{M} being $\sigma(F_1)$-measurable up to null sets. □

Example B.4 (Mean and One are not ordered in general)**.** *This example shows that no unconditional ordering between Mean and One follows from the setup alone.*

Let C be trivial and let $N = 2$. Let

$$y^* \sim \mathrm{Unif}\{0, 1\}.$$

Conditional on $y^* = 0$, let

$$y'_1, y'_2 \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1, +1\},$$

and conditional on $y^* = 1$, let

$$y'_1, y'_2 \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-2, +2\}.$$

Then y^* is determined by $|y'_1|$. Specifically,

$$|y'_1| = 1 \iff y^* = 0, \qquad |y'_1| = 2 \iff y^* = 1.$$

Thus,

$$m_1 = \mathbb{E}[y^* \mid y'_1] = y^*,$$

and therefore

$$\mathcal{R}_1 = 0.$$

However, the mean summary

$$\bar{y}'_2 = \frac{y'_1 + y'_2}{2}$$

does not always determine y^*. In particular,

$$\bar{y}'_2 = 0$$

occurs with positive probability under both $y^* = 0$ and $y^* = 1$. Hence

$$\mathrm{Var}(y^* \mid \bar{y}'_2) > 0$$

on an event of positive probability, which implies

$$\mathcal{R}_{\mathrm{M}} = \mathbb{E}[\mathrm{Var}(y^* \mid \bar{y}'_2)] > 0.$$

Therefore,

$$\mathcal{R}_1 < \mathcal{R}_{\mathrm{M}}.$$

This shows that Mean does not always dominate One.

Conversely, there are also standard settings in which Mean dominates One. For example, if y^* depends on a latent scalar signal and $y'_1, \ldots, y'_N$ are conditionally independent noisy measurements of that signal with symmetric noise, then averaging can reduce noise relative to a single draw. Thus, Mean and One are generally not comparable without additional assumptions.

Remark B.5 (Exchangeability alone is not enough)**.** *Conditional exchangeability of the repeated LLM draws does not imply the orthogonality condition in Proposition B.3. Let C be trivial, $N = 2$, and*

$$y^* \sim \mathrm{Unif}\{0, 1\}.$$

Conditional on $y^* = 0$, let

$$(y'_1, y'_2) \in \{(1, -1), (-1, 1)\}$$

with equal probability. Conditional on $y^* = 1$, let

$$(y_1', y_2') \in \{(2, -2), (-2, 2)\}$$

with equal probability. Then (y_1', y_2') is exchangeable conditional on y^*. However,

$$\bar{y}_2' = 0 \qquad a.s.$$

so

$$m_{\mathrm{M}} = \mathbb{E}[y^* \mid \bar{y}_2'] = \frac{1}{2}.$$

On the other hand, $|y_1'|$ determines y^*, so

$$m_1 = m_{\mathrm{V}} = y^*.$$

Therefore,

$$\mathbb{E}[m_{\mathrm{V}} - m_{\mathrm{M}} \mid F_1] = y^* - \frac{1}{2} \neq 0 \qquad a.s.$$

Thus, the orthogonality condition in Proposition B.3 is a genuine additional assumption and does not follow merely from exchangeability.

C. Additional Experiments

C.1. Comparison with SFT on Open-Source LLMs

A natural alternative to post-hoc debiasing is to fine-tune an open-source LLM directly on human responses, modifying the model's weights rather than learning an external module on top of frozen LLM outputs. This route is also operationally relevant for privacy-sensitive deployments where proprietary API access (e.g., `gpt-4o`) is not permitted. To assess this trade-off, we fine-tune `Qwen3-8B` (Qwen Team, 2025) on the same train split of each dataset at both task levels and evaluate on the same test split used for `Vector`.

Table 2. `Vector` vs. supervised fine-tuned `Qwen3-8B` on the same train/test splits, with the raw `Base LLM` (`gpt-4o`) for reference. Cells are mean ± s.d. (`Vector` across 5 seeds. Gray cells mark `Base LLM` and blue cells mark the best within each block, respectively.

		MAE (↓)			NAcc (↑)			HA (↑)			SA (↑)		
Level	Dataset	Base	SFT	Vector	Base	SFT	Vector	Base	SFT	Vector	Base	SFT	Vector
Population	Twin-2K-500	62.0 ± 0.0	82.5 ± 0.0	34.7 ± 4.2	86.3 ± 0.0	82.3 ± 0.0	93.3 ± 0.9	42.7 ± 0.0	44.7 ± 0.0	65.8 ± 7.1	42.5 ± 0.0	44.7 ± 0.0	52.2 ± 2.6
	OpinionQA	62.1 ± 0.0	66.2 ± 0.0	30.5 ± 1.4	84.5 ± 0.0	83.4 ± 0.0	92.4 ± 0.3	44.9 ± 0.0	51.9 ± 0.0	69.3 ± 2.0	40.9 ± 0.0	51.9 ± 0.0	57.0 ± 1.1
	EEDI	61.7 ± 0.0	104.6 ± 0.0	25.6 ± 1.3	79.4 ± 0.0	65.1 ± 0.0	91.5 ± 0.4	46.4 ± 0.0	48.2 ± 0.0	74.5 ± 5.8	41.3 ± 0.0	48.2 ± 0.0	58.1 ± 2.0
Individual	Twin-2K-500	58.5 ± 0.0	59.1 ± 0.0	44.2 ± 0.2	76.4 ± 0.0	67.7 ± 0.0	84.5 ± 0.1	59.7 ± 0.0	53.9 ± 0.0	68.4 ± 0.3	57.5 ± 0.0	53.9 ± 0.0	66.2 ± 0.2
	OpinionQA	89.0 ± 0.0	100.1 ± 0.0	65.3 ± 0.7	77.7 ± 0.0	75.0 ± 0.0	84.3 ± 1.1	43.3 ± 0.0	42.3 ± 0.0	46.4 ± 0.9	43.3 ± 0.0	42.3 ± 0.0	45.1 ± 0.8
	EEDI	34.5 ± 0.0	38.7 ± 0.0	10.1 ± 0.3	88.5 ± 0.0	87.1 ± 0.0	96.9 ± 0.3	68.4 ± 0.0	74.7 ± 0.0	85.5 ± 1.3	55.4 ± 0.0	74.7 ± 0.0	63.1 ± 0.9

Table 2 shows that `Vector` achieves the best result in all population-level cells (12/12) and in 11 of the 12 individual-level cells, attaining the best `MAE`, `NAcc`, and `HA` on every dataset at both levels; the sole loss is on EEDI individual `SA`, where SFT's categorical output lands directly on a human-response category while `Vector`'s regression output typically falls between adjacent categories and is penalized by `SA`'s two-nearest-category weighting. SFT itself fails to outperform the raw `Base LLM` on individual-level `MAE` for OpinionQA (100.1 vs. 89.0) and EEDI (38.7 vs. 34.5), indicating that supervised fine-tuning of an open-source LLM does not automatically transfer the closed-source quality of `gpt-4o`. In deployment terms, `Vector` dominates both alternatives on the main pointwise metrics whenever proprietary API access is available; when only locally hosted models are permitted, fine-tuning small open-source models is a viable fallback that recovers most of the `Base LLM` headroom and occasionally surpasses it (e.g., EEDI individual `SA`).

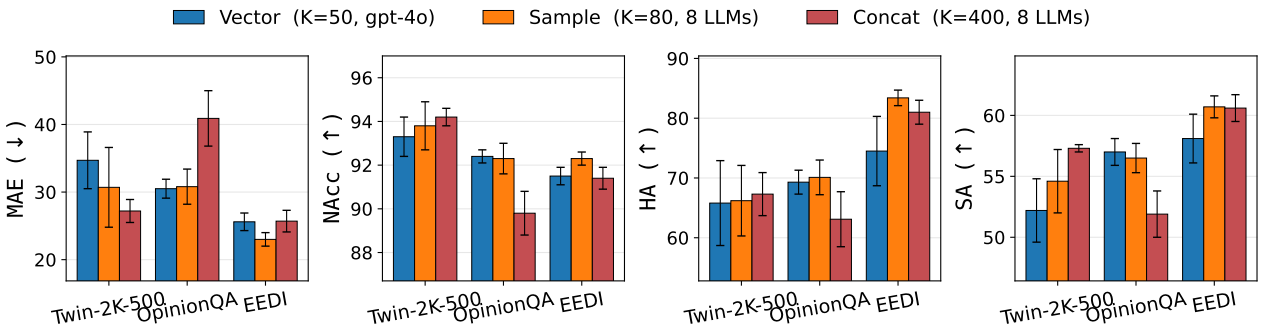


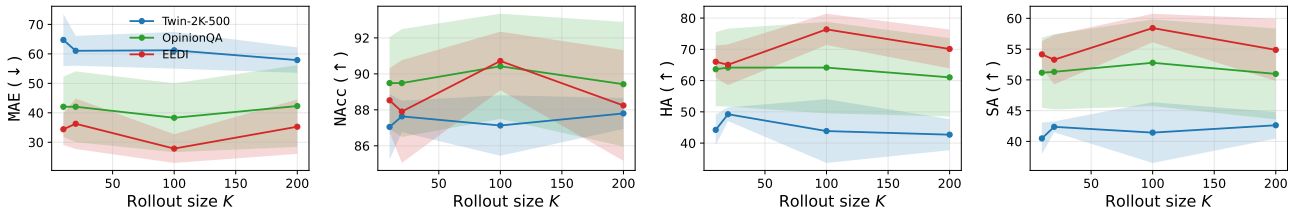
Figure 4. Population-level effect of broadening the LLM-side input over 5 seeds. Legend: `Vector` (K=50), `Sample` (K=80), `Concat` (K=400).

C.2. Multi-LLM Rollouts for Population-level Task

Extending Table 1, we ask whether broadening the LLM-side input across multiple LLMs improves prediction. Three variants share Table 1's head, encoder, and target, differing only in their LLM-side input. `Vector` reuses the single-source K=50 `gpt-4o` draws, `Sample` stratifies K=80 draws across the eight LLMs of Appendix D.1 (10 per LLM), and `Concat` concatenates each LLM's full K=50 vector across all eight LLMs, capped at 400.

Figure 4 reveals two regimes. On Twin-2K-500, every metric improves monotonically from `Vector` to `Sample` to `Concat`. On OpinionQA, `Concat` regresses sharply (e.g., `MAE` 30.5 → 40.9), and on EEDI `Sample` is the strict winner. Following Huang et al. (2025), each LLM source has an effective pool size κ; stacking beyond κ amplifies systematic bias rather than reducing variance. Detailed analysis is deferred to Appendix C.3.

C.3. Ablation: Rollout Size K at Group Level



(a) `MAE` (↓) *(b)* `NAcc` (↑) *(c)* `HA` (↑) *(d)* `SA` (↑)

Figure 5. Population-level metrics as a function of single-source rollout size K. Each curve is the mean across 5 seeds with shaded ±1 s.d. Source is fixed to `gpt-4o` and the head matches `Vector` of Table 1. Legend shown only in panel (a) and shared across panels.

The main population-level experiment fixes K=50 `gpt-4o` draws per question; the multi-LLM ablation in Appendix C.2 additionally suggested that larger K can amplify bias on datasets with limited per-source effective hidden-pool size. To isolate the effect of K alone, we hold the source fixed (`gpt-4o`) and the head architecture identical to `Vector` in Table 1, sweeping $K \in \{10, 20, 100, 200\}$ on the same train/test splits as the main result. Each setting is repeated over 5 seeds.

Figure 5 shows no consistent scaling pattern across datasets. Twin-2K-500 only mildly improves with K, while EEDI and OpinionQA peak near K=100 and slip back toward the smaller-K neighborhood at K=200, so additional draws do not uniformly help. The overall effect of K is also limited in magnitude; `NAcc` stays within 87.0–90.7 across all settings, and the remaining metrics fluctuate by only a few points around the K=50 values of Table 1, far smaller than the `Vector` versus `Base LLM` gaps reported there. The mild non-monotonicity echoes the bias-amplification mechanism of Huang et al. (2025), illustrated here by OpinionQA `MAE` rising from 38.3 at K=100 to 42.3 at K=200. We therefore treat K=50 as a safe default rather than a tuned optimum, and recommend selecting K on a held-out split when deployment data warrants it.

C.4. Ablation: Source-Specific `LLM filed` Choices

The population-level main result in Table 1 fixes one source (`gpt-4o`) to isolate the representation comparison. Here we vary the source-specific `LLM filed`. The three tables below report mean ± s.d. across the five seeds on the same 0–100 scale as the main table. Gray numeric cells denote `Base LLM`; blue boxes mark the best displayed corrected value within each source-metric row among the non-base representations, including ties. The bottom block in each table adds two richer multi-source rows, `Sample` and `Concat`, where the `Mean` and `Vector` columns report the corresponding multi-source representations.

Table 3. Population-level `MLP` robustness table on Twin-2K-500 across source-specific `LLM filed` choices, followed by richer multi-source `Sample` and `Concat` rows.

	MAE (↓)					NAcc (↑)					HA (↑)					SA (↑)				
Source	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector
`claude-opus-4-6`	52.1 ± 0.0	67.4 ± 2.0	61.4 ± 7.6	56.5 ± 9.9	35.6 ± 6.2	88.2 ± 0.0	87.4 ± 0.4	88.5 ± 1.5	89.5 ± 2.0	92.7 ± 1.2	42.3 ± 0.0	43.1 ± 3.5	48.5 ± 5.3	49.2 ± 4.8	53.8 ± 5.8	45.1 ± 0.0	40.7 ± 1.1	43.6 ± 2.5	44.6 ± 2.8	51.3 ± 3.2
`deepseek-r1`	43.1 ± 0.0	67.4 ± 2.0	60.1 ± 8.4	56.4 ± 9.8	33.2 ± 2.2	90.3 ± 0.0	87.4 ± 0.4	88.7 ± 1.7	89.5 ± 2.0	93.2 ± 0.5	59.6 ± 0.0	43.1 ± 3.5	47.7 ± 7.5	49.6 ± 6.6	65.0 ± 5.5	53.0 ± 0.0	40.7 ± 1.1	43.4 ± 2.6	44.7 ± 2.9	53.7 ± 0.8
`doubao-seed-2.0-pro`	60.7 ± 0.0	67.4 ± 2.0	63.7 ± 5.6	56.6 ± 9.9	33.0 ± 5.8	86.9 ± 0.0	87.4 ± 0.4	88.1 ± 1.2	89.5 ± 2.0	93.4 ± 1.1	32.7 ± 0.0	43.1 ± 3.5	46.2 ± 3.6	50.0 ± 5.3	63.5 ± 8.7	38.8 ± 0.0	40.7 ± 1.1	42.9 ± 1.7	44.8 ± 3.3	52.9 ± 2.4
`glm-4.7`	34.7 ± 0.0	67.4 ± 2.0	65.3 ± 4.7	58.2 ± 8.7	36.8 ± 5.3	92.8 ± 0.0	87.4 ± 0.4	87.8 ± 1.0	89.2 ± 1.8	92.8 ± 1.0	63.5 ± 0.0	43.1 ± 3.5	46.2 ± 4.7	49.2 ± 5.2	62.7 ± 6.7	58.2 ± 0.0	40.7 ± 1.1	42.1 ± 1.7	44.2 ± 2.7	52.2 ± 2.2
`gpt-4`	89.0 ± 0.0	67.4 ± 2.0	67.0 ± 4.2	63.6 ± 6.8	52.6 ± 4.6	80.9 ± 0.0	87.4 ± 0.4	87.4 ± 0.9	88.1 ± 1.4	90.0 ± 0.9	26.9 ± 0.0	43.1 ± 3.5	46.9 ± 3.5	47.7 ± 5.3	46.5 ± 3.4	32.5 ± 0.0	40.7 ± 1.1	41.9 ± 1.7	42.5 ± 2.5	45.9 ± 1.0
`gpt-4o`	62.0 ± 0.0	67.4 ± 2.0	62.5 ± 7.2	57.2 ± 9.5	34.7 ± 4.2	86.3 ± 0.0	87.4 ± 0.4	88.3 ± 1.5	89.4 ± 1.9	93.3 ± 0.9	42.7 ± 0.0	43.1 ± 3.5	49.2 ± 4.8	50.0 ± 4.7	65.8 ± 7.1	42.5 ± 0.0	40.7 ± 1.1	43.1 ± 2.5	44.7 ± 3.1	52.2 ± 2.6
`gpt-5.4`	65.9 ± 0.0	67.4 ± 2.0	59.4 ± 8.4	57.1 ± 9.1	33.9 ± 2.5	85.5 ± 0.0	87.4 ± 0.4	88.9 ± 1.7	89.3 ± 1.8	93.0 ± 0.5	40.4 ± 0.0	43.1 ± 3.5	50.4 ± 5.2	50.0 ± 4.7	65.8 ± 4.2	38.7 ± 0.0	40.7 ± 1.1	43.7 ± 2.3	44.3 ± 2.7	51.7 ± 1.5
`qwen3-max`	66.6 ± 0.0	67.4 ± 2.0	59.9 ± 8.8	57.1 ± 9.0	36.8 ± 5.3	85.2 ± 0.0	87.4 ± 0.4	88.8 ± 1.8	89.4 ± 1.8	92.6 ± 0.9	48.1 ± 0.0	43.1 ± 3.5	49.2 ± 5.9	51.5 ± 5.0	63.5 ± 4.1	39.7 ± 0.0	40.7 ± 1.1	43.8 ± 2.8	44.5 ± 2.9	50.9 ± 2.4
`Sample`	47.4 ± 0.4	67.4 ± 2.0	—	56.2 ± 10.0	30.7 ± 5.9	89.3 ± 0.1	87.4 ± 0.4	—	89.6 ± 2.1	93.8 ± 1.1	45.4 ± 1.7	43.1 ± 3.5	—	48.5 ± 5.5	66.2 ± 5.9	45.8 ± 0.4	40.7 ± 1.1	—	45.0 ± 3.3	54.6 ± 2.6
`Concat`	46.8 ± 0.0	67.4 ± 2.0	—	55.7 ± 10.0	27.2 ± 1.7	89.4 ± 0.0	87.4 ± 0.4	—	89.7 ± 2.0	94.2 ± 0.4	44.2 ± 0.0	43.1 ± 3.5	—	48.8 ± 4.8	67.3 ± 3.6	46.0 ± 0.0	40.7 ± 1.1	—	45.1 ± 3.2	57.3 ± 0.3

Table 4. Population-level `MLP` robustness table on OpinionQA across source-specific `LLM filed` choices, followed by richer multi-source `Sample` and `Concat` rows.

	MAE (↓)					NAcc (↑)					HA (↑)					SA (↑)				
Source	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector
`claude-3.5-haiku`	50.1 ± 0.0	46.8 ± 0.3	44.2 ± 1.6	38.6 ± 1.2	34.5 ± 1.6	87.5 ± 0.0	88.3 ± 0.1	89.0 ± 0.4	90.4 ± 0.3	91.4 ± 0.4	44.2 ± 0.0	59.5 ± 2.8	57.9 ± 4.2	64.9 ± 1.8	68.1 ± 4.3	44.5 ± 0.0	47.5 ± 1.0	48.9 ± 0.4	51.5 ± 0.5	54.9 ± 0.7
`deepseek-v3`	47.8 ± 0.0	46.8 ± 0.3	43.1 ± 0.6	36.9 ± 2.2	31.2 ± 3.1	88.0 ± 0.0	88.3 ± 0.1	89.2 ± 0.1	90.8 ± 0.5	92.2 ± 0.8	50.6 ± 0.0	59.5 ± 2.8	61.0 ± 3.8	68.3 ± 3.0	68.8 ± 5.8	47.0 ± 0.0	47.5 ± 1.0	50.5 ± 0.5	52.5 ± 0.7	56.6 ± 1.6
`gpt-3.5-turbo`	52.5 ± 0.0	46.8 ± 0.3	44.0 ± 1.5	40.1 ± 2.0	37.7 ± 2.1	86.9 ± 0.0	88.3 ± 0.1	89.0 ± 0.4	90.0 ± 0.5	90.6 ± 0.5	52.0 ± 0.0	59.5 ± 2.8	60.8 ± 1.1	65.5 ± 3.4	71.7 ± 2.3	47.1 ± 0.0	47.5 ± 1.0	49.5 ± 0.3	51.9 ± 0.6	54.4 ± 1.1
`gpt-4o-mini`	62.2 ± 0.0	46.8 ± 0.3	43.2 ± 1.4	39.8 ± 2.0	34.8 ± 1.6	84.5 ± 0.0	88.3 ± 0.1	89.2 ± 0.4	90.0 ± 0.5	91.3 ± 0.4	48.0 ± 0.0	59.5 ± 2.8	61.0 ± 2.3	65.7 ± 3.0	68.0 ± 2.2	40.5 ± 0.0	47.5 ± 1.0	50.0 ± 0.7	51.1 ± 0.6	53.8 ± 1.5
`gpt-4o`	62.1 ± 0.0	46.8 ± 0.3	44.3 ± 1.1	35.5 ± 2.1	30.5 ± 1.4	84.5 ± 0.0	88.3 ± 0.1	88.9 ± 0.3	91.1 ± 0.5	92.4 ± 0.3	44.9 ± 0.0	59.5 ± 2.8	61.3 ± 4.0	69.3 ± 1.7	69.3 ± 2.0	40.9 ± 0.0	47.5 ± 1.0	50.0 ± 0.6	53.8 ± 0.6	57.0 ± 1.1
`gpt-5-mini`	51.0 ± 0.0	46.8 ± 0.3	46.4 ± 1.8	38.0 ± 1.8	33.1 ± 1.4	87.2 ± 0.0	88.3 ± 0.1	88.4 ± 0.5	90.5 ± 0.4	91.7 ± 0.4	48.0 ± 0.0	59.5 ± 2.8	56.6 ± 2.0	66.8 ± 1.5	70.1 ± 3.0	45.5 ± 0.0	47.5 ± 1.0	48.7 ± 0.4	52.7 ± 0.7	55.6 ± 1.1
`llama-3.3-70B`	50.8 ± 0.0	46.8 ± 0.3	44.1 ± 1.6	37.9 ± 1.9	33.3 ± 1.6	87.3 ± 0.0	88.3 ± 0.1	89.0 ± 0.4	90.5 ± 0.5	91.7 ± 0.4	48.0 ± 0.0	59.5 ± 2.8	59.7 ± 2.8	65.2 ± 5.1	66.0 ± 4.9	45.1 ± 0.0	47.5 ± 1.0	49.5 ± 0.5	51.7 ± 0.5	54.4 ± 0.6
`mistral-7B-v0.3`	57.1 ± 0.0	46.8 ± 0.3	44.3 ± 0.5	43.6 ± 2.3	43.7 ± 2.1	85.7 ± 0.0	88.3 ± 0.1	88.9 ± 0.1	89.1 ± 0.6	89.1 ± 0.5	57.1 ± 0.0	59.5 ± 2.8	61.3 ± 1.9	61.6 ± 3.3	65.7 ± 3.5	44.0 ± 0.0	47.5 ± 1.0	49.9 ± 0.6	49.8 ± 1.0	50.3 ± 0.7
`Sample`	41.5 ± 1.3	46.8 ± 0.3	—	37.3 ± 2.0	30.8 ± 2.6	89.6 ± 0.3	88.3 ± 0.1	—	90.7 ± 0.5	92.3 ± 0.7	55.1 ± 4.7	59.5 ± 2.8	—	67.5 ± 2.8	70.1 ± 2.9	50.4 ± 1.0	47.5 ± 1.0	—	52.5 ± 0.6	56.5 ± 1.2
`Concat`	39.6 ± 0.0	46.8 ± 0.3	—	36.7 ± 2.1	40.9 ± 4.1	90.1 ± 0.0	88.3 ± 0.1	—	90.8 ± 0.5	89.8 ± 1.0	58.4 ± 0.0	59.5 ± 2.8	—	68.6 ± 1.4	63.1 ± 4.6	51.4 ± 0.0	47.5 ± 1.0	—	52.9 ± 0.6	51.9 ± 1.9

Correction helps almost universally, with one boundary case. On every (dataset, source) cell except `glm-4.7` on Twin-2K-500, `Vector` achieves lower `MAE` than the corresponding `Base LLM`, with the largest gap precisely where `Base LLM` is weakest (`gpt-4` on Twin: 89.0 → 52.6, gain 36.4; `gpt-4o-mini` on EEDI: 62.9 → 26.4, gain 36.5). The single exception, `glm-4.7` on Twin-2K-500 (34.7 → 36.8), corresponds to a source whose raw outputs are already near-optimal for the dataset, leaving little room for correction; the value of `Vector` thus grows with how much corrective signal is needed.

Table 5. Population-level MLP robustness table on EEDI across source-specific LLM filed choices, followed by richer multi-source Sample and Concat rows.

	MAE (↓)					NAcc (↑)					HA (↑)					SA (↑)				
Source	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector	Base	w/o	One	Mean	Vector
claude-3.5-haiku	42.9 ± 0.0	38.4 ± 1.4	35.0 ± 0.8	32.6 ± 1.6	25.9 ± 0.4	85.7 ± 0.0	87.2 ± 0.5	88.3 ± 0.3	89.1 ± 0.5	91.4 ± 0.1	60.2 ± 0.0	53.7 ± 7.8	62.2 ± 4.4	65.3 ± 6.1	76.6 ± 3.0	50.1 ± 0.0	47.4 ± 0.7	50.8 ± 0.9	52.3 ± 1.3	56.8 ± 0.6
deepseek-v3	39.5 ± 0.0	38.4 ± 1.4	37.5 ± 1.4	31.4 ± 1.6	26.1 ± 1.4	86.8 ± 0.0	87.2 ± 0.5	87.5 ± 0.5	89.5 ± 0.5	91.3 ± 0.5	62.6 ± 0.0	53.7 ± 7.8	55.4 ± 4.5	71.1 ± 2.4	78.8 ± 2.9	51.1 ± 0.0	47.4 ± 0.7	48.6 ± 1.1	53.4 ± 1.3	58.1 ± 1.5
gpt-3.5-turbo	75.2 ± 0.0	38.4 ± 1.4	38.4 ± 1.1	38.0 ± 1.1	39.6 ± 2.6	74.9 ± 0.0	87.2 ± 0.5	87.2 ± 0.4	87.3 ± 0.4	86.8 ± 0.9	39.8 ± 0.0	53.7 ± 7.8	55.2 ± 4.4	55.2 ± 3.6	58.1 ± 6.1	36.9 ± 0.0	47.4 ± 0.7	48.2 ± 0.8	48.0 ± 0.7	47.8 ± 2.0
gpt-4o-mini	62.9 ± 0.0	38.4 ± 1.4	34.4 ± 2.1	31.7 ± 1.6	26.4 ± 1.5	79.0 ± 0.0	87.2 ± 0.5	88.5 ± 0.7	89.4 ± 0.5	91.2 ± 0.5	41.0 ± 0.0	53.7 ± 7.8	62.9 ± 3.2	68.4 ± 3.3	80.5 ± 2.7	38.9 ± 0.0	47.4 ± 0.7	51.2 ± 1.8	53.5 ± 1.6	59.1 ± 1.4
gpt-4o	61.7 ± 0.0	38.4 ± 1.4	36.2 ± 0.6	30.7 ± 1.6	25.6 ± 1.3	79.4 ± 0.0	87.2 ± 0.5	87.9 ± 0.2	89.8 ± 0.5	91.5 ± 0.4	46.4 ± 0.0	53.7 ± 7.8	58.8 ± 3.8	71.1 ± 5.6	74.5 ± 5.8	41.3 ± 0.0	47.4 ± 0.7	50.0 ± 0.6	54.3 ± 1.4	58.1 ± 2.0
gpt-5-mini	64.3 ± 0.0	38.4 ± 1.4	31.9 ± 1.6	30.8 ± 2.1	23.1 ± 1.6	78.6 ± 0.0	87.2 ± 0.5	89.4 ± 0.5	89.7 ± 0.7	92.3 ± 0.5	43.4 ± 0.0	53.7 ± 7.8	69.6 ± 3.8	71.1 ± 4.9	85.8 ± 1.6	39.7 ± 0.0	47.4 ± 0.7	53.1 ± 1.6	54.2 ± 1.9	62.0 ± 2.1
llama-3.3-70B	63.2 ± 0.0	38.4 ± 1.4	33.0 ± 1.7	32.0 ± 1.8	27.6 ± 1.6	78.9 ± 0.0	87.2 ± 0.5	89.0 ± 0.6	89.3 ± 0.6	90.8 ± 0.5	43.4 ± 0.0	53.7 ± 7.8	65.5 ± 6.2	68.4 ± 4.8	78.3 ± 1.7	39.5 ± 0.0	47.4 ± 0.7	52.1 ± 1.5	53.2 ± 1.7	57.4 ± 2.0
mistral-7B-v0.3	78.3 ± 0.0	38.4 ± 1.4	37.8 ± 1.0	37.9 ± 1.0	38.1 ± 0.6	73.9 ± 0.0	87.2 ± 0.5	87.4 ± 0.3	87.4 ± 0.3	87.3 ± 0.2	28.9 ± 0.0	53.7 ± 7.8	55.7 ± 4.8	55.2 ± 6.6	55.9 ± 4.6	30.7 ± 0.0	47.4 ± 0.7	48.7 ± 0.7	48.5 ± 0.7	48.7 ± 0.6
Sample	32.4 ± 0.9	38.4 ± 1.4	—	30.3 ± 1.7	23.0 ± 1.0	89.2 ± 0.3	87.2 ± 0.5	—	89.9 ± 0.6	92.3 ± 0.3	69.4 ± 1.6	53.7 ± 7.8	—	73.0 ± 6.0	83.4 ± 1.3	54.5 ± 0.5	47.4 ± 0.7	—	54.5 ± 1.4	60.7 ± 0.9
Concat	33.8 ± 0.0	38.4 ± 1.4	—	29.7 ± 1.8	25.7 ± 1.6	88.7 ± 0.0	87.2 ± 0.5	—	90.1 ± 0.6	91.4 ± 0.5	66.3 ± 0.0	53.7 ± 7.8	—	74.5 ± 5.8	81.0 ± 2.0	53.8 ± 0.0	47.4 ± 0.7	—	55.0 ± 1.5	60.6 ± 1.1

Base quality and corrected quality are positively correlated, but Vector compresses the spread. Sources with stronger Base LLM predictions tend to yield stronger Vector predictions; on EEDI, the four sources with the lowest Base MAE (deepseek-v3 39.5, claude-3.5-haiku 42.9, gpt-4o 61.7, gpt-4o-mini 62.9) all produce Vector MAE below 27, whereas the two weakest sources (gpt-3.5-turbo 75.2 and mistral-7B-v0.3 78.3) plateau around 38–40. The same ordering holds on Twin-2K-500 and OpinionQA. However, Vector drastically narrows the dynamic range: Twin-2K-500 Base MAE spans 34.7 (glm-4.7) to 89.0 (gpt-4), but Vector MAE spans only 33.0–52.6, indicating that the corrective debiasing module recovers most of the headroom regardless of source quality and only struggles when the source is severely miscalibrated.

Multi-source rollouts (Sample, Concat) generally outperform the best single source. On Twin-2K-500 and EEDI, Concat attains the best MAE, NAcc, HA, and SA of any row, indicating that pooling across heterogeneous LLMs adds information beyond any one source. On OpinionQA the gain is more nuanced: Sample matches the best single source (gpt-4o Vector 30.5 vs Sample 30.8 MAE) but Concat regresses to 40.9, suggesting that simple concatenation can dilute the signal when one source already dominates the others. Across all datasets, Mean from a multi-source pool is consistently weaker than Vector, reaffirming the main paper's claim that distributional features beyond the first moment carry predictive value.

C.5. Distributional Alignment Metrics

This appendix expands the per-question distributional analysis of Section 5.3. Following Peng et al. (2025), every metric is computed across the respondents within a single question, then aggregated across questions. Questions with fewer than three respondents on the test split or zero human variance are dropped (this affects fewer than 1% of questions on each dataset). To make all variants directly comparable to the discrete Base LLM reference and to the integer human responses, we round every continuous prediction to its nearest admissible category before computing any metric.

Metric definitions. Let $\hat{y}_{ij}$ denote the (rounded) prediction and y_{ij} the human response for respondent i on question j. We define three per-question metrics, each aggregated across questions $j \in \mathcal{J}$.

Per-question Pearson. $r_j = \mathrm{cor}_i(\hat{y}_{ij}, y_{ij})$, aggregated via the Fisher-z transform,

$$\bar{r} = \tanh\Big(\tfrac{1}{|\mathcal{J}|}\textstyle\sum_j \tfrac{1}{2}\ln\tfrac{1+r_j}{1-r_j}\Big),$$

which stabilizes the variance of correlation estimates near ± 1. Higher is better.

Per-question Glass's Δ. $\Delta_j = (\bar{\hat{y}}_j - \bar{y}_j)/\mathrm{SD}_i(y_{ij})$, the predicted-vs. human mean shift in units of human standard deviation. We report the magnitude $|\overline{\Delta}| = \frac{1}{|\mathcal{J}|}\sum_j |\Delta_j|$. Lower is better.

Per-question 1-Wasserstein distance. For each question j, let $\hat{\mu}_j = \frac{1}{n_j}\sum_i \delta_{\hat{y}_{ij}}$ and $\mu_j = \frac{1}{n_j}\sum_i \delta_{y_{ij}}$ denote the empirical distributions of predicted and human responses across the n_j respondents. We compute the 1-Wasserstein distance via its CDF integral form,

$$W_j = W_1(\hat{\mu}_j, \mu_j) = \int_{\mathbb{R}} \left| F_{\hat{\mu}_j}(t) - F_{\mu_j}(t) \right| dt,$$

where F_μ denotes the CDF of μ, and average W_j across questions. Lower is better. Unlike the previous two, W_j captures the full distributional gap rather than only its first moment.

Table 6. Distributional alignment metrics at the individual level. Cells are mean ± s.d. across 5 seeds. Gray rows mark `Base LLM` and blue cells mark the best, respectively.

		Pearson $\bar{r}$ (↑)			$\lvert\overline{\Delta}\rvert$ (↓)			Wasserstein (↓)		
	Representation	Twin	Opinion	EEDI	Twin	Opinion	EEDI	Twin	Opinion	EEDI
Individual	`Base LLM`	0.285 ± 0.000	0.158 ± 0.000	0.436 ± 0.000	0.311 ± 0.000	0.576 ± 0.000	0.883 ± 0.000	0.354 ± 0.000	0.764 ± 0.000	0.429 ± 0.000
	`w/o LLM`	0.155 ± 0.005	0.390 ± 0.097	-0.360 ± 0.503	0.307 ± 0.044	0.469 ± 0.032	0.548 ± 0.073	0.382 ± 0.027	0.823 ± 0.071	0.190 ± 0.033
	`One`	0.245 ± 0.009	0.351 ± 0.127	0.135 ± 0.809	0.221 ± 0.018	0.466 ± 0.029	0.509 ± 0.107	0.351 ± 0.015	0.773 ± 0.103	0.197 ± 0.046
	`Mean`	0.294 ± 0.004	0.412 ± 0.129	0.197 ± 1.035	0.204 ± 0.018	0.460 ± 0.023	0.506 ± 0.047	0.337 ± 0.019	0.771 ± 0.097	0.175 ± 0.022
	`Vector`	0.298 ± 0.013	0.465 ± 0.086	0.750 ± 0.438	0.200 ± 0.010	0.449 ± 0.024	0.433 ± 0.026	0.336 ± 0.011	0.769 ± 0.104	0.171 ± 0.016

D. Experiment Details

D.1. Dataset Preprocessing

This appendix documents how each benchmark is constructed end-to-end for both the population- and individual-level tasks. Three preprocessing choices are shared across all three datasets. First, every human and LLM response is mapped to an admissible integer category set $\mathcal{C} = \{a, a+1, \ldots, b\}$ that varies per question (recorded alongside each row), and downstream targets are renormalized to $[0, 1]$ for the regression head. Second, train/test splits are produced by uniformly sampling 80% of the units (respondent-question pairs at the individual level, questions at the population level) at a fixed random seed shared across all dataset-method combinations, so every method is evaluated on the same partition. Third, the LLM-side feature is built from a per-dataset multi-model rollout (eight models per benchmark) used at both the individual and population levels, designed to maximize cross-vendor diversity. `gpt-4o` serves as the canonical `Base LLM` in the main results across all three benchmarks; the remaining seven source models differ between Twin-2K-500 and OpinionQA/EEDI, reflecting two complementary rollout designs (a flagship-tier mix for Twin-2K-500 and a mid-/small-tier mix for OpinionQA and EEDI), and are listed under each benchmark below.

Twin-2K-500. Twin-2K-500 (Toubia et al., 2025) contains responses from 2,058 U.S. participants across four survey waves and more than 500 questions spanning demographics, psychology, economics, personality, cognition, and behavioral tasks. Because our prediction problem requires bounded, closed-form answers, we exclude open-ended prompts and focus on the closed-form items from waves 1–3, where admissible category sets are either Likert ($\mathcal{C} = \{1, \ldots, b\}$ with $b \in \{5, 7, 9\}$) or binary ($\mathcal{C} = \{0, 1\}$). This yields 2,058 personas, 665 question records, and 1,222,466 response records.

Each individual-level row carries an eight-dimensional LLM vector stacking outputs of `claude-opus-4-6`, `deepseek-r1`, `doubao-seed-2.0-pro`, `glm-4.7`, `gpt-4`, `gpt-4o`, `gpt-5.4`, and `qwen3-max`, each queried once under the persona-conditioned template of Appendix D.2; `gpt-4o` serves as the canonical `Base LLM` in the main results, and the per-source robustness across all eight choices at the population level is shown in Table 3. The individual benchmark contains 166,710 pairs, split 133,368 / 33,342 train/test under the random 80/20 partition.

For the *population-level* task, we average human responses at the question level (per-respondent mean) to obtain $\mu_{h,j}$, and reuse the same eight-model pool by retaining question-level LLM response collections per source. The canonical-source population-level vector is built by taking that source's persona-conditioned response from each respondent sorted by respondent id and keeping the first $K{=}50$ entries, matching the rollout size used in Table 1. We additionally require

the maximum admissible response value to be at most 9 so that population means stay on a comparable Likert scale across questions. Only 105 questions survive this construction (84 / 21 random-split).

OpinionQA. OpinionQA (Santurkar et al., 2023) is built from Pew Research's American Trends Panel. Following the processed benchmark of Huang et al. (2025), we begin from 385 survey questions and 1,476,868 human responses from at least 32,864 respondents; each question has $\mathcal{C} = \{1, 2, 3, 4, 5\}$ ordered categories and at least 400 human answers, and all categories are renormalized to $[0, 1]$. The respondent side is filtered to those with usable LLM responses across our full eight-model rollout (respondents without LLM support are dropped, not backfilled), reducing the persona pool from 32,864 to 200 and leaving 9,437 respondent-question rows.

Each individual-level row carries an eight-dimensional LLM vector stacking outputs of `claude-3.5-haiku`, `deepseek-v3`, `gpt-3.5-turbo`, `gpt-4o-mini`, `gpt-4o`, `gpt-5-mini`, `llama-3.3-70B-instruct-turbo`, and `mistral-7B-instruct-v0.3`, queried once each under the persona-conditioned template. The individual benchmark is split 7,549 / 1,888 under the random 80/20 partition. For the *population-level* task, the human target is the per-respondent mean per question and the LLM side retains question-level response collections per model; the canonical-source population-level vector is built by taking the first K=50 persona-conditioned responses per question (matching Table 1), and all 385 questions remain usable, split 308 / 77.

EEDI. EEDI traces back to the NeurIPS 2020 Education Challenge dataset (Wang et al., 2021) and consists of multiple-choice math diagnostics with $\mathcal{C} = \{1, 2, 3, 4\}$ options. Following the processed survey-simulation version of Huang et al. (2025), we start from 573 mathematics questions and 443,433 student responses from 2,287 students, then retain questions with at least 100 responses and no graphs or diagrams (412 questions); response variables are renormalized to $[0, 1]$.

Unlike Twin-2K-500 and OpinionQA, EEDI groups students by an education profile (gender, premium-pupil status, age) rather than treating each student as an independent persona; the working table contains 13 respondent profiles and 4,084 unique profile-question tuples. Each individual-level unit is one such tuple, and multiple raw student responses sharing the same tuple are pre-averaged into a per-tuple mean used as the individual-level target y^*. The EEDI options $\{1, 2, 3, 4\}$ are nominal rather than ordinal, so the per-tuple mean is interpreted as a soft label encoding the empirical distribution of profile-conditional answer choice (i.e., a 1-D summary of the categorical mixture), not as a numerical value on an ordered scale; we additionally report the order-free `HA` and `SA` metrics, which equal classification accuracy on the rounded prediction and therefore validate the result independently of the averaging step. Keeping only profiles with usable LLM responses leaves 9 profiles and 3,479 tuples, each carrying the same eight-model LLM vector as OpinionQA, split 2,783 / 696 under the random 80/20 partition. For the *population-level* task, we aggregate human responses at the question level and retain model-specific LLM response collections; the canonical-source population-level vector is built from the first K=50 persona-conditioned responses per question (matching Table 1), and the aggregated release contains 411 questions (one removed when a model-response list is missing), split 328 / 83.

D.2. Prompt Templates

We use two prompt templates: a persona-conditioned template for the individual-level task and the population-level `Vector` feature, and a persona-free population-level template for the `Base LLM (One)` population-level baseline. Both instruct the LLM to emit a JSON object mapping question identifiers to integer option numbers so that responses can be parsed without free-text post-processing.

Persona-conditioned template (individual-level; also population-level `Vector`). The placeholder `<demographic string>` is filled with the dataset's native persona fields before the survey questions are appended (e.g., "65-year-old Male identifying as White" for Twin-2K-500, "CREGION: Northeast — AGE: 65+ — SEX: Male — ..." for OpinionQA, or "Gender: 1 — PremiumPupil: 0.0 — Age: 11" for EEDI). The template follows Huang et al. (2025):

```
You are a <demographic string>. You will answer a multi-question survey.

Output requirements:
- For each question shown as "Qk:", include a key "Qk" with your chosen
  option NUMBER as the value.
- Use integers only (not strings). No text, no explanations, no extra keys,
  no arrays, no nested objects.
```

```
- Choose exactly one option number per question, based on the enumerated
  Options list (e.g., 1-..., 2-..., 3-...).

Now, please answer the following questions:
<Qk: question text>
```

At the individual level, `<demographic string>` is fixed to the target respondent's persona $\mathbf{z}_d$ and each respondent-question pair is queried once to each of the eight LLMs listed in Appendix D.1, yielding an $N{=}8$ feature vector. For the population-level `Vector` feature, the LLM is fixed to `gpt-4o` and `<demographic string>` is filled with $K{=}50$ personas sampled from the dataset's persona pool, yielding a $K{=}50$ vector per question.

Persona-free population template (population-level `Base LLM`). The `Base LLM` population-level baseline (`One`) reveals no demographic attribute and asks `gpt-4o` to answer on behalf of a representative target population appropriate to the benchmark. For each question we issue one such query; no respondent embedding $\mathbf{z}_d$ enters the prompt at any stage. A single direct prompt is used rather than aggregating K persona-conditioned draws because the population-level supervision target is the population mean, which the persona-free population prompt is designed to answer in one shot.

Twin-2K-500 and OpinionQA (representative U.S. adults):

```
You are answering a survey question on behalf of a representative sample
of U.S. adults. Return the option NUMBER that the population as a whole
is most likely to select.

Output requirements:
- Respond with a single JSON object of the form {"Qk": <integer>}.
- Use integers only. No text, no explanations, no extra keys.
- Choose exactly one option number per question, based on the enumerated
  Options list (e.g., 1-..., 2-..., 3-...).

Now, please answer the following questions:
<Qk: question text>
```

EEDI (representative secondary-school math students):

```
You are answering a multiple-choice mathematics question on behalf of a
representative sample of secondary-school students (ages 10-14, the
target population of the EEDI diagnostic). Return the option NUMBER
that this student population is most likely to select, including common
misconceptions rather than only the mathematically correct answer.

Output requirements:
- Respond with a single JSON object of the form {"Qk": <integer>}.
- Use integers only. No text, no explanations, no extra keys.
- Choose exactly one option number per question, based on the enumerated
  Options list (A=1, B=2, C=3, D=4).

Now, please answer the following questions:
<Qk: question text>
```

D.3. Debiasing Algorithm Details

This appendix documents the method variants, model families, and hyperparameters used to train every `Vector`, `Mean`, `One`, and `w/o LLM` entry reported in the main text and in Appendix C.4.

Feature variants. Four feature constructions share the same backbone (frozen encoder ψ applied to the question and, at the individual level, the respondent demographics) and differ only in how they consume the N-sample LLM-side rollout

Table 7. `MLP` hyperparameters at the population and individual levels, shared across datasets, source LLMs, and feature variants.

Hyperparameter	Population-level	Individual-level
Hidden layers	$(512, 256, 128)$	$(6144, 3072, 1536, 768, 384)$
Weight decay α	0.01	10^{-4}
Learning rate (init)	5×10^{-4}	2×10^{-4}
Batch size	64	512
Max epochs	1,500	3,500
Dropout	0.0	0.05
Validation fraction	0.1	0.1
Early-stop patience / $\Delta_{\min}$	20 / 10^{-6}	60 / 10^{-6}
LLM-vector transform	raw	raw
Feature standardization	yes	yes

$\mathbf{Y}'_N$. `x_only` (`w/o LLM`) uses the backbone alone with no LLM signal; `x_one_llm` (`One`) appends a single draw $y'_1 \in \mathbf{Y}'_N$; `x_avg_llm` (`Mean`) appends the per-source mean $\bar{y}'_N = \frac{1}{N}\sum_n y'_n$; and `x_all_llm` (`Vector`) appends the full N-dimensional sample vector $\mathbf{Y}'_N$. All four variants are trained from the same train/test split, with the only difference being the LLM-side feature; `Mean` and `Vector` additionally share the same upstream sample so that any gap is attributable to representation alone.

Estimator. We use a multi-layer perceptron (`MLP`) on every (dataset, source, variant) cell at both task levels, with hyperparameters fixed across datasets and source LLMs (Table 7). The population-level settings are intentionally lighter than at the individual level because the population task has only $\mathcal{O}(10^2)$ training questions per dataset, whereas the individual-level training set is two to three orders of magnitude larger and uses a five-layer pyramid backbone with stronger regularization and a longer schedule.

Multi-source population-level rollouts. For the `Sample` and `Concat` rows of Appendix C.4, the LLM input is built from the same eight source LLMs of the corresponding dataset (Appendix D.1) rather than from a single source. `Sample` draws `SAMPLE_PER_LLM`$=10$ persona-conditioned responses per source, yielding a fixed 80-dimensional LLM vector per question. `Concat` concatenates each source's full per-respondent vector; for Twin-2K-500 we cap the total at `CONCAT_LLM_DIM`$=350$ (eight sources $\times$ up to fifty respondents), while OpinionQA and EEDI use their full per-source vectors without capping. All other hyperparameters match the corresponding single-source population-level configuration.

Seeds and reporting. Population-level runs sweep five random seeds $\{0, 1, 2, 3, 4\}$; individual-level runs sweep $\{1, 2, 3, 4, 5\}$. Each cell of every reported table is the mean $\pm$ standard deviation across these five seeds, with the train/test partition rederived under each seed.

D.4. SFT Experiments Details

We supervise-fine-tune `Qwen3-8B` (Qwen Team, 2025) separately on each of the six (dataset, level) configurations: {Twin-2K-500, OpinionQA, EEDI} $\times$ {population, individual}. Training and inference are run on the same train/test splits used by `Vector`, and the full training and rollout pipeline is implemented in the `verl` framework (Sheng et al., 2024).

Data Construction. For each (dataset, level) pair, every train and test row is rendered into a chat-format `messages` list. The system prompt instructs the model to output only a JSON object of the form `{"answer": <integer>}` with no extra text. The user message contains the question text and the admissible integer range $[a, b]$, and at the individual level additionally prepends the respondent's demographic background (joined from the dataset's persona text for that respondent). For the train split, an assistant turn is appended whose content is the JSON-wrapped rounded ground truth $\rho(y)$; the test split contains only the system and user turns. Qwen3 thinking mode is disabled at every row. The population-level target is the rounded population mean response per question, and the individual-level target is the rounded per-respondent answer.

Training. We optimize the autoregressive cross-entropy loss over the assistant turn only (the user turn is masked). Training uses FSDP2 with sharding equal to the number of GPUs, BF16 weights, gradient checkpointing, and a remove-padding

Table 8. SFT training and inference hyperparameters, shared across all six (dataset, level) configurations.

Hyperparameter	Value
Backbone	`Qwen3-8B`
Precision	bf16
Distribution strategy	FSDP2 (sharded across all GPUs)
Gradient checkpointing	enabled
Remove-padding attention	enabled
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.1
LR schedule	cosine with 3% linear warmup
Gradient clipping (ℓ_2)	1.0
Adam state offload	CPU
Epochs	1
Batch size (sequences)	16
Max sequence length	768 tokens
Per-GPU token budget	32,768
Decoding backend	`vLLM`, greedy
Temperature, top-p	0, 1
Prompt / response length	768 / 32
Samples per input	1
GPU memory utilization	0.9

attention kernel for efficient packed-sequence training. The optimizer is AdamW with learning rate 1×10^{-5}, weight decay 0.1, gradient clipping 1.0, cosine schedule with 3% linear warmup, and Adam states offloaded to CPU. We use dynamic batching with a per-GPU token budget of 32,768, a static batch size of 16 sequences, and a maximum sequence length of 768 tokens. We train for one epoch per configuration; longer schedules did not improve held-out scores. Table 8 summarizes the hyperparameters.

Inference. After training, the FSDP shards are merged back to a Hugging Face checkpoint and served with `vLLM` (Kwon et al., 2023). Generation is deterministic with temperature 0, top-p 1, prompt length 768, response length 32, and a single sample per input. The integer answer is extracted from the model output by parsing the first `"answer": <number>` field; outputs that fail to parse are scored as zero.

Scoring. For each test unit we extract the prediction $\hat{y}_{\text{SFT}}$ from the JSON output and report the same four metrics as in the main text (`MAE`, `NAcc`, `HA`, `SA`), computed against the ground truth y on the original response scale. Because SFT outputs a single integer, $\hat{y}_{\text{SFT}} \in \mathcal{C}$ holds by construction, so $\rho(\hat{y}_{\text{SFT}}) = \hat{y}_{\text{SFT}}$ and the soft weight reduces to $w(\hat{y}_{\text{SFT}}; \hat{y}_{\text{SFT}}) = 1$, which makes `SA` numerically equal to `HA` for all SFT rows.