LLM HUMAN RESPONSE ALIGNMENT: A MULTI-SAMPLE DEBIASING FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) offer scalable synthetic respondents, but their usefulness as human proxies depends on alignment with observed responses. We distinguish two sources of misalignment. Model-side bias arises when LLM and human response distributions differ under the same observed inputs. Latent-context omission limits prediction because relevant respondent information is missing. We propose a multi-sample post-hoc debiasing framework that keeps the LLMs frozen and learns an external correction from human-labeled data. The correction retains a response vector, preserving distributional variation and source identity that a single response can discard. The sampling design adapts to the task: fixed model sources for individual prediction and sampled personas for population-mean prediction. Under squared loss, the vector weakly reduces Bayes-optimal risk relative to either scalar summary, with strict gains when compression loses predictive information. Across three broad benchmarks spanning diverse domains, our method attains the best mean in every main comparison, reducing MAE relative to uncorrected LLM predictions by up to 47% at the population level and 54% at the individual level. We further show that our method preserves respondent rankings and improves revenue prediction. Code and data are available at this link.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as synthetic respondents and evaluators in behavioral simulation, opinion modeling, economic decision tasks, and model evaluation (Aher et al., 2023; Brand et al., 2023; Li et al., 2024; Anthis et al., 2025; Maier et al., 2025). Given a question and an optional respondent profile, they generate responses at a scale that would be costly to achieve through human data collection. Their usefulness depends on whether these outputs predict observed human behavior. Existing studies document systematic opinion misalignment, distorted preference distributions, and reduced behavioral heterogeneity (Santurkar et al., 2023; Goli & Singh, 2024; Wang et al., 2025a; Tjuatja et al., 2024; Li et al., 2025). Generating plausible responses is therefore insufficient: synthetic predictions require calibration against human observations.

We distinguish two sources of misalignment. *Model-side bias* arises when an LLM and humans have different response distributions under the same observed inputs, reflecting differences in pretraining data or post-training objectives. *Latent-context omission* arises when human responses depend on information absent from the observed profile, such as prior experience, domain knowledge, or partisan affiliation. These failures pose different learning problems. Paired human and synthetic observations can support correction of a systematic model-side shift. In contrast, respondents who share the same observed profile can differ in unmeasured context, which repeated queries conditioned only on that profile cannot identify. Under squared loss, the corresponding prediction target is the mean human response within the observed question-profile group. The available improvement lies in estimating this mean more accurately.

Prior work addresses alignment either before generation or through a post-hoc correction. Fine-tuning on behavioral data and enriching prompts with interviews or self-reports can improve simulation, but require weight access, substantial computation, or additional respondent measurement (Kolluri et al., 2025; Binz et al., 2025; Zhang et al., 2026; Park et al., 2024). Post-hoc methods keep the LLM frozen and combine its predictions with limited human-labeled data. Much of this literature develops estimation and uncertainty quantification for downstream population quantities

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Angelopoulos et al., 2023; Egami et al., 2023; Wang et al., 2026). Population-level calibration also learns to combine multiple LLM outputs using aggregate human benchmarks (Leng et al., 2026). We study a complementary prediction problem: how should a supervised correction represent the LLM outputs available for each individual response or population mean?

Our framework retains a vector of K responses instead of reducing the synthetic evidence to one draw or its sample mean. An external prediction network combines that vector with frozen text embeddings and learns from observed human targets. Individual-level vectors query fixed model sources using the same respondent profile, preserving which source produced each answer. Population-level vectors vary personas within one source, representing the simulated panel's response distribution. Averaging removes source identity and distributional features beyond the first moment. Retaining these features allows supervision to learn when particular patterns of agreement or disagreement predict the human target. The method requires additional LLM queries, but no additional human labels or updates to LLM weights.

We make three contributions. First, we formulate a common multi-sample debiasing framework for individual-response and population-mean prediction, using source-indexed responses to the same respondent-question pair for individual correction and persona-sampled responses for population correction. Second, we characterize its information advantage under squared loss and separate that advantage from the uncertainty caused by missing respondent context. The theory compares Bayes-optimal predictors and specifies when output compression loses predictive information. Third, we evaluate the representation across three broad benchmarks spanning diverse domains, retaining matched samples for the Mean versus Vector comparison. Vector achieves the best mean in all main comparisons. Population MAE decreases by 18.1% and 15.8% relative to Mean on the two survey benchmarks. Additional experiments on revenue prediction, respondent ranking, and fixed-budget sampling further demonstrate the benefits of our method.

2 Related Work

Misalignment in synthetic human responses. LLM proxies diverge from human observations in opinion surveys and social simulation (Santurkar et al., 2023; Tjuatja et al., 2024), economic and behavioral tasks (Horton et al., 2023; Goli & Singh, 2024), preference measurement (Li et al., 2024; Brand et al., 2023; Toubia et al., 2025), and LLM-as-a-judge evaluation (Zheng et al., 2023). Persona conditioning can also introduce stereotypes and homogenized behavior (Argyle et al., 2023; Li et al., 2025). Pretraining and post-training may concentrate predictions around corpus or preference-model modes (Jiang et al., 2025). We distinguish this model-side mismatch from missing respondent context: correcting a synthetic response distribution does not reveal which unobserved attributes belong to a particular person.

Pre-generation debiasing. One family of methods changes model weights, conditioning information, or the agent system before responses are generated. Behavioral and cognitive fine-tuning adapts the model toward human responses (Kolluri et al., 2025; Binz et al., 2025; Zhang et al., 2026). Persona design broadens population coverage (Argyle et al., 2023; Bui et al., 2025), while interviews, self-reports, and psychometric measurements supply richer respondent information (Park et al., 2024). Task-specific agent systems model sequential customer behavior and decision processes (Wang et al., 2025c). These approaches can address information missing from basic demographics, but may require additional measurement, access to weights, or specialized system design. Our method works with the existing observed profiles and frozen model sources.

Post-hoc correction and inference. Prediction-powered inference and design-based supervised learning combine machine predictions with human labels for downstream estimation (Angelopoulos et al., 2023; Egami et al., 2023). Related work uses LLM surrogates for survey estimation, uncertainty quantification, and data augmentation (Huang et al., 2025; Wang et al., 2026); two-stage designs combine fine-tuning with post-hoc correction (Wang et al., 2025b). HuMCal learns weights over LLM elicitation variants using aggregate human benchmarks for population-level calibration (Leng et al., 2026). Our individual-level correction uses a response vector for each respondent-question pair, with the corresponding human response as supervision. SYN-DIGITS transfers relationships among synthetic responses to human responses on new questions (Fan et al., 2026). In

our evaluation, One tests a single-response representation rather than reproducing every inferential procedure, and SYN-DIGITS is evaluated under complete-question holdout.

Learning from multiple outputs. Using predictions as features connects our method to stacked generalization (Wolpert, 1992). For LLMs, self-consistency aggregates sampled reasoning paths (Wang et al., 2023), while cross-model disagreement can inform uncertainty estimation (Hamidieh et al., 2026). We learn the relationship between multiple outputs and observed human targets rather than selecting an answer by consensus. A source-indexed vector preserves model identity; persona-sampled outputs provide an empirical response distribution. Neither is assumed to match the human distribution directly. The contribution concerns which generated information the correction retains and how the sampling design changes across prediction tasks.

3 METHOD

3.1 TASK DEFINITION

We learn a post-hoc correction from a labeled sample $\mathcal{D}$ and deploy it on a target sample $\mathcal{D}^{\text{new}}$ without human responses. The LLM and text encoder remain frozen. A query supplies question x and optional profile w to source m, returning

$$y' \sim \widehat{P}_m(\cdot \mid x, w), \tag{1}$$

where $w = z_i$ is respondent i's observed profile, $w = p$ is a sampled persona, and $w = \emptyset$ omits the profile. The synthetic law $\widehat{P}_m$ need not equal the human response law P_h.

Individual and population targets. A prediction unit is $u = (i, x)$ for an individual response, with target $T_{i,x} = y^*_{i,x}$ and encoded features $B_{i,x} = (\psi(x), \psi(z_i))$. At the population level, the unit is a question, the feature is $B_x = \psi(x)$, and the target is the observed panel mean,

$$T_x = \bar{y}^*_x = \frac{1}{n_x} \sum_{i \in I_x} y^*_{i,x}, \tag{2}$$

where I_x indexes available responses and $n_x = |I_x|$. Personas are sampling-design variables supplied to the LLM, not profile features supplied to the population predictor. Table 1 summarizes the task definitions. The main individual evaluation holds out response records; a separate experiment holds out complete questions. EEDI uses aggregated profile targets, as detailed in Section 5.2.

Table 1: The two debiasing tasks.

	Individual-level task	Population-level task
Prediction unit	respondent-question pair (i, x)	question x
Target T	$y^*_{i,x}$, one respondent's response	$\bar{y}^*_x$, panel mean over n respondents
LLM profile slot w	observed profile z_i	researcher-chosen persona or $\emptyset$
Feature B	$B_{i,x} = (\psi(x), \psi(z_i))$	$B_x = \psi(x)$
Generalization axis	respondents and questions	questions

Learning and deployment. Let $\mathbf{y}'_u$ collect the generated responses. For a loss d and predictor class $\mathcal{G}$, we fit

$$\hat{g} \in \arg\min_{g \in \mathcal{G}} \frac{1}{|\mathcal{D}|} \sum_{u \in \mathcal{D}} d(g(B_u, \mathbf{y}'_u), T_u). \tag{3}$$

The fitted rule predicts new units as

$$\hat{T}_v = \hat{g}(B_v, \mathbf{y}'_v), \qquad v \in \mathcal{D}^{\text{new}}. \tag{4}$$

Human targets supervise fitting but are never prediction inputs. Generalization requires the labeled sample to represent the relationship between synthetic and human responses in the target population. An unrepresentative labeled sample can produce a biased correction.

3.2 MULTI-SAMPLE DEBIASING

Source-indexed individual responses. A single response retains one realization of a model's conditional response law. We collect K outputs for the same respondent-question pair (i, x), keeping both the question x and observed profile z_i fixed:

$$\mathbf{y}'_{K,i,x} = (y'_{i,x,1}, \ldots, y'_{i,x,K}), \qquad y'_{i,x,k} \sim \widehat{P}_{m_k}(\cdot \mid x, z_i). \tag{5}$$

The construction allows repeated draws from one source or responses from distinct frozen models. Our individual experiments use $K = 8$, with one response from each of eight sources in a fixed order. Coordinate identity lets the correction learn how different sources relate to human targets. A mean over those responses removes this association. Figure 1 shows the prediction pipeline.

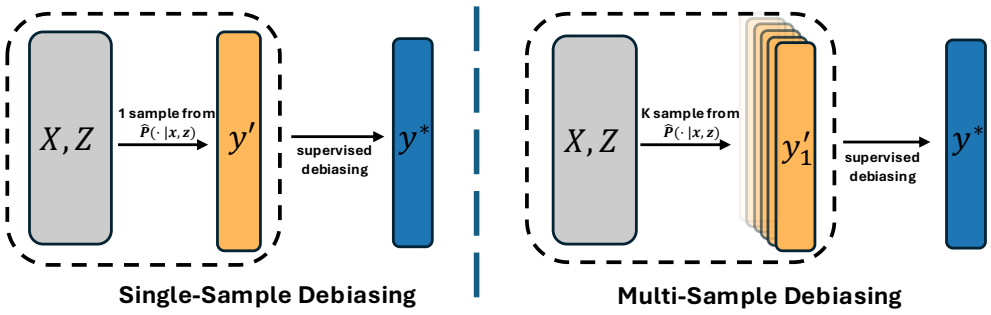


Figure 1: Individual-level multi-sample debiasing. The One baseline supplies one response $y'_{i,x}$ as the LLM-side input, while our approach passes the full vector $\mathbf{y}'_{K,i,x}$ to the debiasing module. The human response $y^*_{i,x}$ is the supervision signal used to fit g and is never an argument of g.

Latent context and the reachable target. Let C denote respondent information absent from $S = (X, Z)$, and write $\mu(S, C) = \mathbb{E}[Y^* \mid S, C]$ and $\bar{\mu}(S) = \mathbb{E}[Y^* \mid S]$. Assume fresh LLM draws are independent of the held-out respondent's (C, Y^*) conditional on S. For a fixed fitted predictor $\hat{Y} = g(B, \mathbf{Y}')$,

$$\begin{aligned}\mathbb{E}[(\hat{Y} - Y^*)^2] = \underbrace{\mathbb{E}[\mathrm{Var}(Y^* \mid S, C)]}_{\text{within-context noise}} + \underbrace{\mathbb{E}[\mathrm{Var}(\mu(S, C) \mid S)]}_{\text{latent-context variance}} \\ + \underbrace{\mathbb{E}[(\hat{Y} - \bar{\mu}(S))^2]}_{\text{estimable prediction error}}.\end{aligned} \tag{6}$$

Only the final term can improve without additional respondent information. Multiple outputs can help estimate the profile-cell mean through response-space coverage, source-specific calibration, and information in the raw-text-conditioned outputs that the frozen encoding does not retain (Appendix A.2). These are possible sources of predictive information, not identification of a person's omitted context. Independently sampled vectors can differ for identical profiles without revealing that context.

Persona-sampled population responses. Population prediction uses a persona pool $\mathcal{P}$ and sampling distribution Π. The main design fixes source m and generates

$$\mathbf{y}'_{K,x} = (y'_{x,1}, \ldots, y'_{x,K}), \qquad y'_{x,k} \sim \widehat{P}_m(\cdot \mid x, p_k), \quad p_k \sim \Pi. \tag{7}$$

We use $K = 50$ and `gpt-4o`. The vector samples the synthetic panel distribution, while the correction learns its relationship to the observed human panel mean. The persona pool may differ from the target panel and the LLM may remain biased even under well-matched personas. Personas are not encoded as features for g. A fixed-budget experiment instead varies sources while holding the persona fixed (Section 5.6).

3.3 ALGORITHM

We instantiate ψ with `text-embedding-3-small`, using 256-dimensional outputs, and g_θ with a multilayer perceptron. The feature vector concatenates B_u with one of three LLM representations:

$$h_{\mathrm{O}}(\mathbf{y}') = y'_1, \qquad h_{\mathrm{M}}(\mathbf{y}') = K^{-1}\sum_{k=1}^{K} y'_k, \qquad h_{\mathrm{V}}(\mathbf{y}') = \mathbf{y}'. \tag{8}$$

These define One, Mean, and Vector; y'_1 is illustrative notation for the response selected by One. Responses are normalized to $[0, 1]$, features are standardized using training statistics, and the scalar head minimizes squared loss with weight decay. Mean and Vector use the same generated responses, embeddings, targets, hidden-layer widths, and training settings. Thus their comparison isolates representation under a fixed query budget. The population Base LLM directly outputs the same sample mean, allowing its comparison with Mean to isolate the learned correction. Architecture, optimization, and alternative prediction heads appear in Appendices D.3 and E.4.

4 INFORMATION-THEORETIC JUSTIFICATION

We analyze the information retained by the response vector under squared loss, using the same observed features and synthetic draws for each representation. The LLM distribution may be biased: the comparison requires no equality between the synthetic and human response laws.

Representations and Bayes risk. For the generic target T and encoded feature B defined in Section 3.1, the task instantiations are

$$\begin{aligned} &\text{individual:} \quad T = y^*_{i,x}, \quad B = (\psi(x), \psi(z_i)); \\ &\text{population:} \quad T = \bar{y}^*_x, \quad B = \psi(x). \end{aligned} \tag{9}$$

Define $F_{\mathrm{V}} = (B, \mathbf{y}'_K)$, $F_{\mathrm{M}} = (B, \bar{y}'_K)$, and $F_{\mathrm{O}} = (B, y'_1)$, with information sets $\mathcal{F}_A = \sigma(F_A)$. Their squared-loss Bayes risks are

$$\mathcal{R}^\star_A = \inf_g \mathbb{E}[(g(F_A) - T)^2], \qquad A \in \{\mathrm{V}, \mathrm{M}, \mathrm{O}\}, \tag{10}$$

where the infimum is over measurable predictors with square-integrable outputs. These predictors use only deployment inputs; the human response remains the supervision target.

Assumption 1 (Basic regularity)**.** *For a fixed $K \geq 1$, $(B, T, \mathbf{y}'_K)$ are defined on a common probability space. The encoder is fixed and measurable, each response is a real-valued measurable random variable, and $\mathbb{E}[T^2] < \infty$.*

Because the mean and first coordinate are measurable functions of the full vector,

$$\mathcal{F}_{\mathrm{M}} \subseteq \mathcal{F}_{\mathrm{V}}, \qquad \mathcal{F}_{\mathrm{O}} \subseteq \mathcal{F}_{\mathrm{V}}. \tag{11}$$

Lemma 1 (Bayes prediction under richer information)**.** *The optimal squared-loss predictor given information $\mathcal{H}$ is $\mathbb{E}[T \mid \mathcal{H}]$. For $\mathcal{H}_0 \subseteq \mathcal{H}_1$,*

$$\mathcal{R}^\star(\mathcal{H}_0) - \mathcal{R}^\star(\mathcal{H}_1) = \mathbb{E}[(\mathbb{E}[T \mid \mathcal{H}_1] - \mathbb{E}[T \mid \mathcal{H}_0])^2] \geq 0. \tag{12}$$

Theorem 1 (Bayes dominance of `Vector`)**.** *Under Assumption 1, at either task level,*

$$\begin{aligned} \mathcal{R}^\star_{\mathrm{M}} - \mathcal{R}^\star_{\mathrm{V}} &= \mathbb{E}[(\mathbb{E}[T \mid B, \mathbf{y}'_K] - \mathbb{E}[T \mid B, \bar{y}'_K])^2] \geq 0, \\ \mathcal{R}^\star_{\mathrm{O}} - \mathcal{R}^\star_{\mathrm{V}} &= \mathbb{E}[(\mathbb{E}[T \mid B, \mathbf{y}'_K] - \mathbb{E}[T \mid B, y'_1])^2] \geq 0. \end{aligned} \tag{13}$$

Each improvement is strict exactly when the corresponding conditional means differ with positive probability.

The theorem follows by applying Lemma 1 to each inclusion in Equation (11); full proofs appear in Appendix B. The result does not order One against Mean. It identifies the possible cost of compressing the vector: dispersion, multimodality, or source-specific disagreement matters when it predicts the target beyond the scalar summary and encoded inputs.

Equality is possible. Under joint permutation invariance of the target and binary response samples, the empirical measure contains the relevant sample information, and the mean determines that measure. With heterogeneous fixed sources, $(0, 1)$ and $(1, 0)$ can instead imply different targets despite equal means. The controlled example in Section 5.1 isolates this distinction. If B already captures all target-relevant input information and synthetic draws are conditionally independent of T given B, additional draws cannot reduce Bayes risk.

These are oracle comparisons, not guarantees for a fitted MLP. Retaining K coordinates increases estimation demands under the same human supervision. Limited data, regularization, model capacity, and optimization can prevent a trained predictor from attaining the information advantage. The empirical comparisons therefore evaluate the three representations under shared training protocols.

5 Experiments

5.1 A controlled example of source identity

Two equally likely profile cells have Bernoulli human responses with means 0.2 and 0.8 but identical embeddings. Two independent LLM sources, conditioned on raw profiles, have response probabilities $(0.1, 0.9)$ in the first cell and $(0.9, 0.1)$ in the second. Their mean has the same distribution in both cells; the ordered patterns $(0, 1)$ and $(1, 0)$ distinguish them.

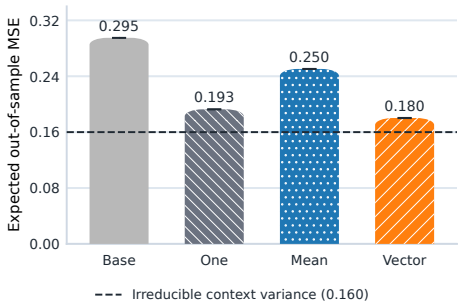


Figure 2: Expected test MSE over five training seeds in the controlled example. Error bars show one standard deviation; the dashed line marks the irreducible error of 0.16.

Using the same smoothed estimator and $2{,}000$ labeled units per seed, Vector achieves expected test MSE of 0.180, versus 0.250 for Mean and 0.193 for One (Figure 2). Source identity helps estimate the human conditional mean, while latent-context variance remains 0.16. For exchangeable binary draws, the mean already determines the empirical distribution. Construction, Bayes risks, and training details appear in Appendix D.1.

5.2 Benchmarks, baselines, and evaluation

Datasets and splits. We use Twin-2K-500 (Toubia et al., 2025), OpinionQA (Santurkar et al., 2023), and EEDI (Wang et al., 2021), covering general surveys, political opinions, and educational responses. Table 2 summarizes the sample construction. Each task uses a random $80/20$ train-test partition. Individual test records can share respondents and questions with training records; population units are questions. EEDI aggregates numerically encoded responses within nine profiles, so its individual-level benchmark predicts profile means. Dataset details and eligibility rules appear in Appendix E.1.

Baselines and implementation. Base LLM uses an uncorrected `gpt-4o` response for individual prediction and the mean of 50 persona-conditioned responses for population prediction. The embedding-only predictor, w/o LLM, measures what encoded inputs explain. One, Mean, and Vector share the prediction architecture and differ in their LLM-side representation. In the main comparisons, One and One Logprob both use `gpt-4o`. One Logprob uses normalized answer-choice probabilities from one generation (Sun & Zhang, 2025); its separate sampling configuration is specified in Appendices E.3 and E.5. HuMCal-Mean learns global persona weights from population

Table 2: Benchmark construction.

Dataset	Full dataset # questions / # respondents	Individual-level sample # questions / # respondents	Individual sample size # records	Population-level sample # questions (samples)
Twin-2K-500	665 / 2,058	60 / 167	10,020	279
OpinionQA	385 / 32,864	385 / 200	9,437	385
EEDI	573 / 2,287	412 / 9 profiles	3,479	412

Notes. Individual sample size is the number of respondent-question records (profile-question records for EEDI); at the population level, each selected question is one sample.

means (Leng et al., 2026), adapting a method designed for distributional supervision. Individual vectors use eight fixed sources; population vectors use 50 sampled personas and one source. EEDI samples its nine profiles with replacement. Source identities and prompts appear in Appendices E.1 and E.2.

Metrics. We report four metrics: mean absolute error (MAE, $[0, \infty)$), range-adjusted accuracy (Acc), hit accuracy (HA, $[0, 100\%]$), and soft accuracy (SA). Response-prediction tables and figures report $100 \times MAE$. Acc and SA lie in $[0, 100\%]$ when predictions and targets are within the admissible response range. Definitions appear in Appendix C. HA compares the admissible categories assigned to prediction and target; for EEDI's encoded aggregates it is not student-level answer correctness. Reported means use five matched training seeds. Significance markers denote unadjusted two-sided paired tests against One with two-way cluster-robust standard errors.

5.3 Main results

Table 3: Main results at both task levels.

Level	Dataset	Metric	`Base LLM`	`w/o LLM`	`One`	`One Logprob`	`Mean`	`HuMCal-Mean`	`Vector`
		LLM queries (K), population / individual	50 / 1	0 / 0	1 / 1	1 / 1	50 / 8	50 / –	50 / 8
Population	Twin-2K-500	MAE (↓)	67.65	66.99*	57.06	57.71	51.92**	59.83	**42.52****
		Acc (↑, %)	83.83	84.19*	86.74	86.39	87.89**	85.62	**90.20****
		HA (↑, %)	39.29	37.50*	45.00	45.00	47.14	44.64	**64.29****
		SA (↑, %)	38.39	37.46	41.11	40.79	44.06**	42.07	**52.15****
	OpinionQA	MAE (↓)	40.67	46.85	44.35	41.05*	36.15***	44.56	**30.45****
		Acc (↑, %)	89.83	88.29	88.91	89.74*	90.96***	88.86	**92.39****
		HA (↑, %)	61.04	60.00	61.30	64.16	68.31*	63.64	**74.29****
		SA (↑, %)	51.69	47.54*	49.95	49.98	53.24**	49.94	**56.26****
	EEDI	MAE (↓)	48.23**	37.57	35.51	33.59	30.86**	52.17**	**25.75*****
		Acc (↑, %)	83.92**	87.48	88.16	88.80	89.71**	82.61**	**91.42*****
		HA (↑, %)	55.42	57.11	60.24	64.82	70.60**	48.19	**77.83****
		SA (↑, %)	47.54	48.16**	50.88	52.69	54.53***	45.22*	**59.39*****
Individual	Twin-2K-500	MAE (↓)	62.30***	58.35***	53.28	58.59***	52.50	–	**50.30*****
		Acc (↑, %)	62.48***	64.11***	69.17	63.83***	69.80	–	**71.88*****
		HA (↑, %)	51.55***	51.93***	57.46	51.71***	58.50	–	**60.18*****
		SA (↑, %)	49.39***	50.89***	56.14	50.76***	56.97	–	**59.02*****
	OpinionQA	MAE (↓)	110.23***	87.60	87.45	87.73	87.32	–	**87.08***
		Acc (↑, %)	72.44***	78.10	78.14	78.07	78.17	–	**78.23***
		HA (↑, %)	34.80***	48.24	48.00	47.58	47.90	–	**48.29**
		SA (↑, %)	32.09***	44.58	44.68	44.29	44.68	–	**44.82**
	EEDI	MAE (↓)	37.88***	18.16	18.29	17.80	17.89	–	**17.35****
		Acc (↑, %)	87.37***	93.95	93.90	94.07	94.04	–	**94.22****
		HA (↑, %)	67.53***	81.47	81.09	81.93	81.81	–	**82.04****
		SA (↑, %)	53.17***	60.35	60.08	60.48	60.34*	–	**60.53****

Notes. Best means over five matched training seeds are in bold. `Base LLM` applies no learned correction. Asterisks denote two-sided paired tests against `One` with two-way cluster-robust standard errors: $^{*}p < .10$, $^{**}p < .05$, $^{***}p < .01$, unadjusted.

Population prediction. Table 3 gives Vector the best mean in all 24 dataset-task-metric comparisons. At the population level, its gains over One are significant in all 12 comparisons ($p < .05$).

Twin-2K-500 MAE falls from 67.65 for Base LLM to 42.52; the relative reduction reaches 47% on EEDI. Mean improves the raw sample mean through supervised correction, whereas Vector additionally retains the individual responses. Relative to Mean, population MAE decreases from 51.92 to 42.52 on Twin-2K-500 and from 36.15 to 30.45 on OpinionQA, reductions of 18.1% and 15.8%. These comparisons use identical samples (Remark 1). HuMCal-Mean is weaker under the shared mean-only supervision, which differs from its original distributional setting.

Individual prediction. The margins are smaller. Gains over One are significant on all four metrics for Twin-2K-500 ($p < .01$) and EEDI ($p < .05$). OpinionQA MAE and Acc improve at $p < .10$, while HA and SA do not reach that threshold. Relative to Mean, MAE falls from 52.50 to 50.30 on Twin-2K-500 and from 87.32 to 87.08 on OpinionQA. The latter's larger improvement over the raw LLM primarily reflects the learned correction. The missing-context decomposition separates an irreducible component from estimable error; it does not imply that all tasks must have equally large representation gains.

5.4 Revenue prediction and offer selection

We evaluate the predictions on 40 products, using the original record-level split. Calibration models include product indicators and relative price alongside respondent features and synthetic responses. Partially pooled logit demand curves convert predicted purchase probabilities into survey-implied revenue at regular prices. Each method selects its top 10 offers; overlap measures agreement with the human benchmark, and regret measures lost human survey-implied revenue per exposure. Full construction and alternative capacities appear in Appendix D.6.

Table 4: Revenue prediction and top-10 offer selection at regular prices for Twin-2K-500.

	Revenue prediction	Top-10 offer selection	
Method	MAE ($/respondent, ↓)	Overlap (%, ↑)	Regret ($/exposure, ↓)
`Base LLM`	1.65***	40	2.514
`w/o LLM`	0.70**	78	0.405
`One`	0.78	80	0.330
`Mean`	0.73**	80	0.305
`Vector`	**0.69****	**82**	**0.303**

Notes. Bolding and significance conventions follow Table 3.

Table 4 shows revenue MAE of $0.69 for Vector versus $0.73 for Mean and $1.65 for Base LLM. Top-10 regret falls from $2.514 for Base LLM to $0.303 for Vector, while Mean already reaches $0.305. Thus, learned correction accounts for most of the regret reduction; retaining source-level responses yields the lowest reported revenue error and regret.

5.5 Respondent rankings within questions

Pointwise accuracy does not determine whether a predictor preserves differences among respondents. We therefore evaluate within-question Spearman correlation, assigning average ranks to ties. Figure 3 reports mean correlations of 0.325 for Base LLM, 0.370 for One, 0.397 for Mean, and 0.425 for Vector. One Logprob reaches 0.243. Vector improves over One on 43 of 60 questions. These gains support retaining multiple-source responses for ranking, without establishing recovery of omitted within-profile context.

5.6 Population-level vector construction

At fixed budget $K = 8$, a shared 8×8 persona-by-source grid compares persona variation within one source with source variation for one persona. Each design yields eight vectors per dataset. Its restricted persona pool differs from the main $K = 50$ setting. In Figure 4, both Vector constructions outperform One on all three datasets. Persona variation improves MAE by 3 to 5 points over source

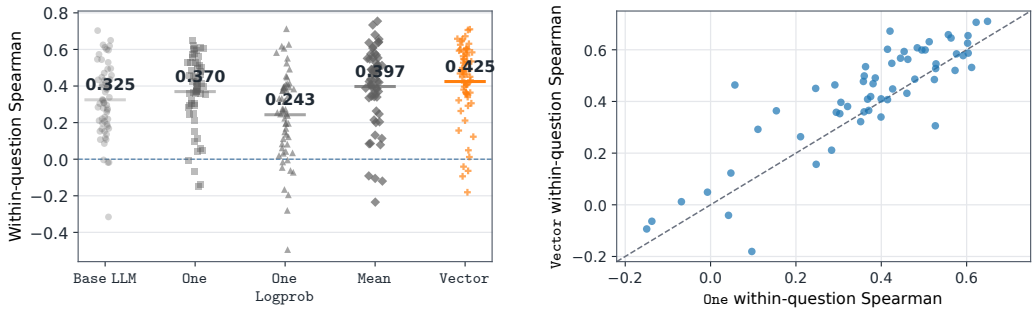


Figure 3: Respondent rankings on the Twin-2K-500 individual test split. Left: within-question Spearman correlations, with mean markers. Right: `Vector` versus `One`; points above the diagonal favor `Vector`.

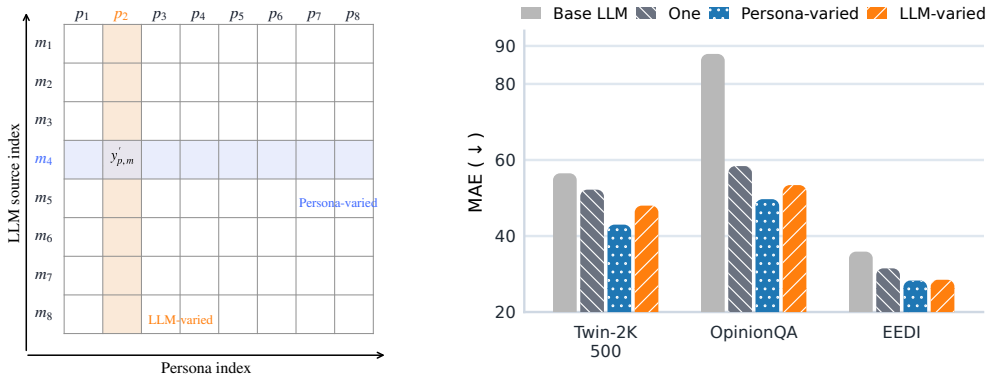


Figure 4: Population-level vector construction at $K = 8$. Left: persona and source sampling within a shared response grid. Right: MAE for `Base LLM`, `One`, and the two `Vector` constructions.

variation on Twin-2K-500 and OpinionQA, favoring persona coverage in this controlled population experiment.

5.7 Additional robustness

Vector outperforms SYN-DIGITS on all four metrics under complete-question holdout (Appendix D.2). The methods use different feature constructions. Further ablations assess response-vector width, source models, and prediction heads (Appendices D.3 to D.5).

6 Conclusion

Multi-sample debiasing retains response variation and source identity for supervised human-response prediction with frozen LLMs. The theory characterizes its Bayes-risk advantage and the limits imposed by missing context. Across three benchmarks, matched gains are clearest in population prediction, with additional benefits for revenue estimation and respondent ranking.

The framework requires representative human supervision and informative synthetic responses. The record-level individual split does not establish generalization to unseen respondents. EEDI's nominally encoded aggregate targets and the oracle theoretical comparison further delimit the claims. Future work includes richer respondent histories, permutation-invariant population representations, and adaptive human-label and LLM-query allocation.

AI Use Statement Generative AI tools assisted with language polishing, including grammar, clarity, and readability. The authors reviewed the suggested edits and take responsibility for the final manuscript.

Ethics Statement This study uses existing human-response datasets and synthetic outputs. Synthetic responses may reproduce demographic biases or obscure differences among respondents. Improved predictive accuracy does not establish fairness or justify consequential decisions about individuals. Downstream applications should respect respondent privacy and dataset access conditions.

Reproducibility Statement We provide detailed dataset construction, prompts, model sources, sampling settings, and training hyperparameters in Appendix E, with additional experimental protocols and results in Appendix D. Code and data are available at this link.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proc. of ICML*, 2023.

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 2023.

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

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

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K Eckstein, Noémi Éltető, et al. A foundation model to predict and capture human cognition. *Nature*, 2025.

James Brand, Ayelet Israeli, and Donald Ngwe. Using GPT for market research. *Available at SSRN 4395751*, 2023.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. Mixture-of-personas language models for population simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24761–24778, 2025. doi: 10.18653/v1/2025.findings-acl.1271. URL https://aclanthology.org/2025.findings-acl.1271/.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Proc. of NeurIPS*, 2023.

Grace Jiarui Fan, Chengpiao Huang, Tianyi Peng, Kaizheng Wang, and Yuhang Wu. SYN-DIGITS: A synthetic control framework for calibrated digital twin simulation, 2026. URL https://arxiv.org/abs/2604.07513.

Ali Goli and Amandeep Singh. Frontiers: Can large language models capture human preferences? *Marketing Science*, 2024.

Kimia Hamidieh, Veronika Thost, Walter Gerych, Mikhail Yurochkin, and Marzyeh Ghassemi. Complementing self-consistency with cross-model disagreement for uncertainty quantification. In *Proc. of ICLR*, 2026.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, 2023.

Chengpiao Huang, Yuhang Wu, and Kaizheng Wang. How many human survey respondents is a large language model worth? an uncertainty quantification perspective. *ArXiv preprint*, 2025.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). *ArXiv preprint*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning LLMs for Human Behavior Prediction in Social Science Experiments. In *Proc. of EMNLP*, 2025.

Yan Leng, Yunxin Sang, and Ashish Agarwal. Reducing disparity between llms and humans: Optimal llm sample calibration. *Marketing Science*, 2026. Forthcoming.

Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. LLM generated persona is a promise with a catch. In *Proc. of NeurIPS*, 2025.

Peiyao Li, Noah Castelo, Zsolt Katona, and Miklos Sarvary. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 2024.

Benjamin F Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C Luhmann, Robbie Dow, Kli Pappas, and Thomas V Wiecki. Llms reproduce human purchase intent via semantic similarity elicitation of likert ratings. *ArXiv preprint*, 2025.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proc. of ICML*, pp. 807–814, 2010.

Joon Sung Park, Carolyn Q. Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, and Michael S. Bernstein. Llm agents grounded in self-reports enable general-purpose simulation of individuals, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proc. of ICML*, 2023.

Hongshen Sun and Juanjuan Zhang. From model choice to model belief: Establishing a new measure for LLM-based research. *ArXiv preprint*, 2025.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of ACL*, 2024.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *ArXiv preprint*, 2025.

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

Lei Wang, Zikun Ye, and Jinglong Zhao. Efficient inference using large language models with limited human data: Fine-tuning then rectification. *ArXiv preprint*, 2025b.

Mengxin Wang, Dennis J Zhang, and Heng Zhang. Large language models for market research: A data-augmentation approach. *Marketing Science*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E. Turner, Richard G. Baraniuk, Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. Results and insights from diagnostic questions: The NeurIPS 2020 education challenge. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, 2021.

Ziyi Wang, Yuxuan Lu, Yimeng Zhang, Jing Huang, and Dakuo Wang. Customer-R1: Personalized simulation of human behaviors via rl-based llm agent in online shopping. *ArXiv preprint*, 2025c.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Feng Zhang, Shijia Li, Chunmao Zhang, Zhanyu Ma, Jun Xu, Jiuchong Gao, Jinghua Hao, Renqing He, Jingwen Xu, and Han Liu. Userlm-r1: Modeling human reasoning in user language models with multi-reward reinforcement learning. *ArXiv preprint*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proc. of NeurIPS*, 2023.

A Additional Task and Representation Details

A.1 Prediction units and sample construction

For individual response prediction, the labeled records are

$$\mathcal{D}_{\text{ind}} = \{(x, z_i, \mathbf{y}'_{K,i,x}, y^*_{i,x}) : (i, x) \in \mathcal{S}\}, \tag{14}$$

and deployment records omit $y^*_{i,x}$. The fitted predictor receives the embeddings $(\psi(x), \psi(z_i))$ and the selected LLM representation. Writing $g(x, z_i, \cdot)$ is shorthand for this encoded-input predictor, not access to the raw text inside the correction network. Training units and test units are separate response records in the main experiment. This does not imply disjoint respondent or question sets. EEDI replaces a respondent's response with an observed mean within a profile-question group, as documented in Appendix E.1.

For population prediction, each question is one unit. Training pairs the question embedding and its synthetic responses with the observed human panel mean; deployment questions omit this mean. Question-specific response coverage determines the set I_x in Equation (2). The persona pool controls synthetic sampling but is not itself a respondent-level feature of the question. A finite human panel mean can differ from the corresponding population expectation even if the synthetic persona distribution were correctly specified.

A.2 Information retained by the sampling designs

Multiple responses can supply two distinct types of evidence. An unordered collection records an empirical response distribution, including variation not determined by its mean when there are more than two possible outcomes. A source-indexed collection also records which source produced each response. Supervision can exploit these features when they predict human targets beyond the encoded inputs. Neither construction makes synthetic frequencies an unbiased estimate of human response frequencies, and neither guarantees recovery of the support or mixture weights of a latent human-response model.

The retained responses can help prediction through three complementary mechanisms. *Response-space coverage* exposes the empirical distribution of generated answers: repeated draws can reveal disagreement or multiple response modes that one realization cannot describe. Human supervision remains necessary because synthetic frequencies need not equal human frequencies. *Heterogeneous-expert calibration* uses the fixed source ordering to learn which output patterns are predictive for a question and profile. Sources may have different systematic biases even when their average response is the same. *Information beyond the encoding* arises when the LLM's raw-text-conditioned outputs retain predictive distinctions missing from B. A compact embedding need not lose relevant information, so this mechanism is conditional on what the encoder preserves. Together, these channels explain why the correction receives the full vector: it can use both empirical response variation and source-specific structure, while a scalar mean makes only their average available.

Remark 1 (Matched population comparisons). The population `Base LLM`, `Mean`, and `Vector` variants use the same generated responses. `Base LLM` outputs their mean directly; `Mean` supplies that mean to a learned predictor; `Vector` supplies the complete vector. Thus, `Mean` versus `Vector` isolates representation while holding the draws and supervision fixed.

Remark 2 (Exchangeability and source identity). If the joint law of $(B, T, \mathbf{y}'_K)$ is invariant to permutations of the response coordinates, the Bayes predictor can be chosen as a function of the empirical response measure. For binary responses at fixed K, the mean determines that measure and the oracle gap is zero. For multiple categories, the mean need not determine the histogram. In a vector with fixed heterogeneous sources, coordinate identity can additionally predict the target. An ordinary MLP is not permutation-invariant by construction; the oracle argument does not impose that property on a fitted network.

Remark 3 (Identical observed profiles). Conditional on a question and observed profile, independently generated LLM vectors have the same law for respondents whose latent context is unobserved. Their realized predictions may differ because of sampling randomness. Such variation does not identify which omitted context belongs to which respondent. Ranking improvements therefore need not imply recovery of within-profile latent differences.

B Proofs and Additional Theory

Proof of Lemma 1. Let $m_{\mathcal{H}} = \mathbb{E}[T \mid \mathcal{H}]$. For any square-integrable $\mathcal{H}$-measurable predictor a, conditional expectation makes the cross term vanish:

$$\mathbb{E}[(T-a)^2] = \mathbb{E}[(T-m_{\mathcal{H}})^2] + \mathbb{E}[(m_{\mathcal{H}}-a)^2]. \tag{15}$$

Consequently $m_{\mathcal{H}}$ minimizes the risk, with

$$\mathcal{R}^\star(\mathcal{H}) = \mathbb{E}[(T-m_{\mathcal{H}})^2]. \tag{16}$$

Apply Equation (15) with $\mathcal{H} = \mathcal{H}_1$ and $a = m_{\mathcal{H}_0}$. This gives Equation (12). □

Proof of Theorem 1. Both y'_1 and $K^{-1}\sum_k y'_k$ are measurable functions of $\mathbf{y}'_K$, giving the inclusions in Equation (11). Apply Lemma 1 first with $\mathcal{H}_0 = \sigma(F_{\mathrm{O}})$, then with $\mathcal{H}_0 = \sigma(F_{\mathrm{M}})$, and use $\mathcal{H}_1 = \sigma(F_{\mathrm{V}})$ in both cases. This proves Equation (13). An expected square is positive exactly when its argument is nonzero with positive probability, proving the strictness statement. The proof uses a generic target T and therefore applies to individual responses, profile averages, and question-level panel averages without changing the argument. □

Proof of the missing-context decomposition. Condition on the fitted predictor and training data. Assume that, for a held-out unit, LLM sampling is independent of (C, Y^*) given S under the target law. Then $\mathbb{E}[Y^* \mid S, \mathbf{Y}'] = \bar{\mu}(S)$, and expansion gives

$$\mathbb{E}[(\hat{Y}-Y^*)^2] = \mathbb{E}[(\hat{Y}-\bar{\mu}(S))^2] + \mathbb{E}[\mathrm{Var}(Y^* \mid S)]. \tag{17}$$

Applying the conditional law of total variance to the last term yields Equation (6). The statement concerns fresh held-out responses under this conditional-independence model. Dependence through shared respondents or questions in an empirical record split must be addressed by the evaluation design; the identity does not certify independence in that split.

Conditional means and latent identification. A human conditional mean can be learned without identifying the latent variable C or each of its mixture components. The role of C here is to clarify what unobserved respondent information prevents. A lower-dimensional encoder is not necessarily insufficient for the prediction task merely because of its dimension. The information gap in Equation (13) requires that the LLM vector supply information about T not already captured by the competing representation.

C Evaluation Metrics

For either level of analysis, let u denote a held-out evaluation unit, t_u its target, and $\hat{t}_u$ its prediction: $(u, t_u, \hat{t}_u) = ((j,x), y^*_{j,x}, \hat{y}_{j,x})$ at the individual level and $(u, t_u, \hat{t}_u) = (x, \bar{y}^*_x, \hat{y}_x)$ at the population level. We write $\mathbb{E}_u$ for the average over $\mathcal{S}_{\mathrm{new}}$ or $\mathcal{X}_{\mathrm{new}}$, respectively. At inference, $\hat{t}_u$ is mapped from the training scale $[0,1]$ back to the original question-specific scale. Predictions and targets may be noninteger on the original response scale; HA maps them to the nearest admissible categories, whereas SA retains their continuous values. Let $\mathcal{A}_u = \{a_u, \ldots, b_u\}$ denote the admissible response categories for the question associated with u.

Mean Absolute Error (MAE) is the mean absolute distance between the prediction and target. Its range is $[0, \infty)$, with lower values preferred. MAE values in the response-prediction tables and figures are reported as $100 \times MAE$. The unscaled metric is:

$$MAE(\downarrow) = \mathbb{E}_u\big[|\hat{t}_u - t_u|\big]. \tag{18}$$

Accuracy (Acc) converts absolute error into range-adjusted accuracy. Higher values are preferred; its range is $[0, 100\%]$ when predictions and targets lie within the admissible response interval. It divides original-scale error by $b_u - a_u$ to make questions with different response ranges comparable:

$$Acc(\uparrow, \%) = 100\,\mathbb{E}_u\Big[1 - \tfrac{|\hat{t}_u - t_u|}{b_u - a_u}\Big]. \tag{19}$$

Hit Accuracy (HA) indicates whether the prediction and target map to the same admissible category. Its range is $[0, 100\%]$, with higher values preferred, and it is computed on the original discrete category set using $\rho_u(r) = \arg\min_{a \in \mathcal{A}_u} |r - a|$:

$$HA(\uparrow, \%) = 100\, \mathbb{E}_u\big[\mathbf{1}\{\rho_u(\hat{t}_u) = \rho_u(t_u)\}\big]. \tag{20}$$

Soft Accuracy (SA) measures overlap between the prediction and target category weights and gives partial credit between adjacent categories. It is expressed as a percentage, with higher values preferred, and lies in $[0, 100\%]$ when predictions and targets lie within the admissible response interval. If $r \in \mathcal{A}_u$, set $\omega_u(r; r)=1$; otherwise, for the two nearest categories $a_1, a_2 \in \mathcal{A}_u$, set $\omega_u(a_1; r)=|r - a_2|$ and $\omega_u(a_2; r)=|r - a_1|$, with zero weight on all other categories:

$$SA(\uparrow, \%) = 100\, \mathbb{E}_u\Big[\textstyle\sum_{a \in \mathcal{A}_u} \omega_u(a; \hat{t}_u)\, \omega_u(a; t_u)\Big]. \tag{21}$$

HA is emphasized for EEDI, but its targets are means of numerically coded nominal choices. Agreement between rounded means is not student-level classification accuracy and is not invariant to relabeling the answer choices. Ordered-response metrics have a more direct interpretation on the survey tasks. Acc is a range-adjusted error score, not classification accuracy; without clipping predictions its formula need not remain nonnegative.

D Additional Experiments

D.1 Toy Example: Construction and Analysis

This appendix provides the full construction, Bayes-risk analysis, and estimation details for the controlled example in Section 5.1.

Setup. As in Section 3.1, respondent i with profile z_i answers question x, and we call (x, z_i) a *profile cell*. Let x ask about vote intention. Within a cell, respondents differ in a latent context $c_i \in \{0, 1\}$, their partisan affiliation ($c_i = 1$ for Republican). This context is not recorded in z_i. Each respondent answers according to their context, $y^*_{i,x} = c_i$, with

$$c_i \mid x, z_i \sim \text{Bernoulli}\big(p(x, z_i)\big). \tag{22}$$

Hence $P_h(\cdot \mid x, z)$ is a two-mode mixture with mixture weight $p(x, z)$ and mixture mean $\bar{\mu}(x, z) = p(x, z)$. Two respondents can share the same cell but differ in c. Their identical observed inputs do not identify which latent context belongs to each respondent.

The module predicts $\hat{y} = g(B_{i,x}, \mathbf{y}'_{K,i,x})$ and is fit on human-labeled rows. Its first input is the frozen encoding $B_{i,x} = \psi(x, z_i)$. Its second input is the response vector $\mathbf{y}'_{K,i,x}$ of K synthetic responses.

Two cells the encoder conflates. Consider two profiles that agree on "age 30, female" and differ only in location. Respondents in cell 1 live in an urban county, and those in cell 2 live in a rural county. Their Republican shares are $p = 0.2$ and $p = 0.8$, respectively. Each cell makes up half of both the labeled and the target sample, so the transportability condition of Section 3.1 holds.

The encoder ψ assigns both cells the same encoding, $B_{i,x} = b_0$. The encoding then carries no information that separates the cells, and a rule based on $B_{i,x}$ alone can do no better than predicting 0.5. Any information distinguishing the cells must therefore come from the synthetic responses, which condition on the raw text of (x, z_i).

Two fixed sources. Let $K = 2$ and tie coordinate k to source $m_k = k$ for every unit. This is the source-indexed construction of Remark 2. Each source returns one mode,

$$y'_{i,x,k} \mid x, z_i \sim \text{Bernoulli}\big(q_k(x, z_i)\big), \qquad q_k(x, z) := q_k(c = 1 \mid x, z), \tag{23}$$

independently across sources. Here $q_k(c = 1 \mid x, z)$ is source k's implicit mixture weight. The draws are independent of the human response conditional on the cell. All respondents in a cell share the same synthetic response law, though realized draws can differ (Remark 3).

Profile cell	Human p	Source 1: q_1	Source 2: q_2
Cell 1 (urban)	0.2	0.1	0.9
Cell 2 (rural)	0.8	0.9	0.1

Neither source's mixture weight matches the human one. Source 1 exaggerates the majority mode in each cell. Source 2 concentrates on the opposite mode, as can happen when sources differ in corpus composition or post-training. Both weights depend on the cell alone and carry no information about any respondent's c_i.

Irreducible and estimable error. Both $B_{i,x}$ and $\mathbf{y}'_{K,i,x}$ depend on (x, z_i) only, so $\hat{y}$ is independent of c_i given the cell. The risk decomposition in Equation (6) then becomes

$$\mathbb{E}\big[(\hat{y} - y^*)^2\big] = \underbrace{p(1-p)}_{\text{irreducible; } = 0.16 \text{ in both cells}} + \underbrace{\mathbb{E}\big[(\hat{y} - \bar{\mu}(x,z))^2\big]}_{\text{estimable}}. \tag{24}$$

No method can predict partisan affiliation within a cell. The only error that any choice of inputs to g can reduce is the error in estimating the cell's mixture mean.

What the vector retains and the mean discards. The probability of each response vector in each cell is:

$(y'_{i,x,1}, y'_{i,x,2})$	Cell 1	Cell 2	$\bar{y}'_K$
$(0,0)$	0.09	0.09	0
$(0,1)$	0.81	0.01	0.5
$(1,0)$	0.01	0.81	0.5
$(1,1)$	0.09	0.09	1

The vector $(0,1)$ almost always signals cell 1, and $(1,0)$ almost always signals cell 2. Both have sample mean 0.5. As a result, $\bar{y}'_K$ has the *same* distribution in both cells: it equals 0, 0.5 and 1 with probabilities 0.09, 0.82 and 0.09. The set of modes returned is equally uninformative, since both modes appear together with probability 0.82 in either cell. Response-space coverage therefore contributes no signal here.

The entire signal lies in *which* source returned which mode. This is coordinate identity, which a source-indexed correction can exploit. Supervised on human responses, g learns that $(0,1)$ indicates a mixture weight near 0.2 and $(1,0)$ indicates one near 0.8. If the two draws were instead exchangeable, as in Remark 2, only their empirical measure would be informative. For a binary pair, that measure is equivalent to $\bar{y}'_K$, so the vector would carry no more information than the mean.

How much this matters. The table below gives the best achievable error for each choice of input. For the calibrated rows, g is the Bayes-optimal rule $\mathbb{E}[\bar{\mu}(x,z) \mid \text{input}]$, which a module trained on enough labeled rows would learn. Rows three to five are the matched comparisons of Remark 1.

Input to g	Prediction $\hat{y}$	Estimable error	Total MSE
$B_{i,x}$ only	0.5	0.090	0.250
Single draw $y'_{i,x,1}$, calibrated	0.26 or 0.74	0.032	0.192
`Base LLM`: $\bar{y}'_K$, no g	$\bar{y}'_K$	0.135	0.295
`Mean`: $\bar{y}'_K$	0.5	0.090	0.250
`Vector`: $\mathbf{y}'_{K,i,x}$	0.21, 0.5, or 0.79	0.020	0.180
Oracle (knows the cell)	$\bar{\mu}(x,z)$	0	0.160

The gap between `Base LLM` and `Mean` measures the value of the learned correction. Here the correction helps only by learning to ignore a biased, uninformative average. The gap between `Mean` and `Vector` isolates the value of cross-draw variation. In this toy, that value comes entirely from coordinate identity.

A single draw also carries some information about the cell. However, as Appendix A.2 notes, it inherits both the full mixture variance and its source's misalignment. When the two sources disagree,

the vector identifies the cell almost perfectly. Only when they agree, which happens with probability 0.18, is g left uninformed.

Throughout, g learns nothing about any respondent's c_i. It learns how source patterns map to human mixture weights. Because this mapping is calibrated to the labeled sample, a labeled sample with a biased mixture would teach g that biased mixture instead.

Estimation and evaluation. `Base` predicts the uncorrected sample mean. The three supervised methods use $F = y'_1$ for `One`, $F = (y'_1 + y'_2)/2$ for `Mean`, and $F = (y'_1, y'_2)$ for `Vector`. All use the same smoothed group-mean estimator,

$$\hat{g}(f) = \frac{1 + \sum_{j:F_j=f} y^*_j}{2 + \#\{j : F_j = f\}}. \tag{25}$$

For each of five random seeds, we generate 2,000 labeled units and fit all three methods to the same responses and human labels. Neither the raw cell type, the mixture weight, nor the source probabilities are supplied to the fitted rule. Training and deployment follow the same law. We evaluate each fitted rule's expected out-of-sample MSE exactly over both cell types and all four possible response pairs.

Figure 2 reports MSEs of 0.295 for `Base`, 0.193 for `One`, 0.250 for `Mean`, and 0.180 for `Vector`. The mean carries no cell information, so its optimal calibrated prediction is the unconditional human mean 0.5, with MSE 0.25. Here even `One` improves on `Mean`: pooling opposing sources discards useful information. Sampling both sources increases coverage of the two response modes, and retaining their identities allows supervision to calibrate that evidence. The resulting vector also supplies cell information lost by the encoder. Its gain is a more accurate estimate of $p(x, z)$; the full irreducible 0.16 remains. This example concerns fixed heterogeneous sources. With exchangeable binary draws, the mean already records the full empirical response distribution and can be as informative as the vector.

D.2 Comparison with SYN-DIGITS under Complete-Question Holdout

Why the main split is incompatible. `SYN-DIGITS` (Fan et al., 2026) uses synthetic control to predict human responses to a new target question. Its individual-level construction requires a human donor matrix in which the same respondents are observed on donor questions, while the entire human target-question column is unavailable. The original Twin-2K-500 individual experiment instead uses a record-level split: training and test cells are interspersed within each of the 60 questions, so every test question retains human labels in training and the human donor matrix is itself partially observed. Directly adding `SYN-DIGITS` to Table 3 would therefore either expose target-question human information that its new-question estimand excludes or discard information available to the other methods. It would not be a like-for-like baseline under the main task.

Compatible benchmark construction. Twin-2K-500 forms a complete 167×60 respondent-by-question matrix. We sample 12 complete question columns (20%) from the sorted question IDs with fixed seed 42 and use the remaining 48 complete columns for training. This produces 8,016 training records (48×167) and 2,004 test records (12×167), matching the sizes of the main record-level split while changing only its missing-data pattern. All 167 respondents are therefore observed on every donor question, no human response to a target question appears in training, and no donor imputation is required. We retrain `One` and `Vector` on these 8,016 records using the same feature definitions, prediction head, hyperparameters, and five model seeds as in the main individual-level experiment. To source-match the direct-response reference to `SYN-DIGITS-EN`, `Base LLM` is the raw persona-conditioned `gpt-4o` response.

For `SYN-DIGITS-EN`, we use one identity-aligned `gpt-4o` response for every respondent-question cell to form a complete synthetic matrix. For each held-out target question, we fit an Elastic Net across the 167 digital-twin rows, regressing its synthetic target column on the 48 synthetic donor columns. We then apply the fitted coefficients to the corresponding 167×48 human donor matrix. Synthetic and human donor columns are standardized separately; the Elastic Net uses $\alpha = 0.01$ and an ℓ_1 ratio of 0.3. This procedure is deterministic for a fixed question split. Unlike `One` and `Vector`, it does not use the shared nonlinear prediction head or the question and profile

embeddings, so the experiment is an estimand-compatible external-method comparison rather than a matched-representation ablation.

Table A.1: Individual-level results on the Twin-2K-500 complete-question holdout benchmark.

Method	MAE (↓)	Acc (↑, %)	HA (↑, %)	SA (↑, %)
Base LLM	70.86	55.95***	43.31***	43.31***
One	70.14	66.79	57.10	56.04
SYN-DIGITS-EN	66.49**	55.92***	45.96***	45.45***
Vector	**61.95*****	**70.80*****	**61.65*****	**60.04*****

Notes. Bolding and significance conventions follow Table 3.

Results. Table A.1 shows that Vector attains the best mean on all four metrics. Two-sided paired tests with two-way cluster-robust standard errors show statistically significant gains of Vector over One in every comparison ($p < .01$). Under the same tests, SYN-DIGITS-EN lowers MAE from 70.14 for One to 66.49 ($p < .05$), but it is worse on Acc, HA, and SA ($p < .01$). Relative to the source-matched raw gpt-4o response, it lowers MAE from 70.86 to 66.49 and raises HA and SA, while Acc is nearly unchanged. Thus, synthetic-control transfer improves some aspects of the single-model signal, but retaining all eight source responses in Vector is substantially more effective on this benchmark.

The metric pattern reflects a limitation of the transferred linear construction. Because the human donor columns are centered separately, the average transferred prediction for each target question equals that question's average synthetic gpt-4o response:

$$\frac{1}{167}\sum_{i=1}^{167}\widehat{y}_{i,x}^{\mathrm{SYN}} = \frac{1}{167}\sum_{i=1}^{167}y_{i,x}^{\prime,\mathrm{gpt\text{-}4o}}. \tag{26}$$

Consequently, the method can transfer respondent-level deviations learned from donor questions but cannot correct a target-question-level mean shift without any target human labels. Its predictions also compress heterogeneity: across the 12 held-out questions, the median ratio of predicted to human within-question standard deviations is 0.616, and, after averaging the neural predictions over seeds, the mean within-question correlation with human responses is 0.257, compared with 0.352 for One and 0.423 for Vector. The issue is especially visible for a binary target on which gpt-4o gives the same response to all 167 profiles; SYN-DIGITS-EN then has no synthetic target variation from which to learn respondent-specific weights. These diagnostics explain why its continuous error can improve while its discrete-response metrics remain weak.

D.3 Robustness to Prediction-Head Choice

We test whether the gains from Vector are robust to the choice of prediction head. The main scalar head in section 3.3 outputs one prediction and minimizes MSE. The Gaussian head instead predicts a conditional mean and a positive, input-dependent variance and minimizes Gaussian negative log-likelihood (NLL). It can represent heteroskedasticity but retains a symmetric distribution with unbounded support. The Beta head predicts two positive concentration parameters and minimizes Beta NLL. Its support is the open interval $(0, 1)$, requiring an explicit treatment of normalized targets at the endpoints, and it can represent the skewness observed in human-response distributions. For evaluation, the Gaussian and Beta heads use their respective distributional means as point predictions.

We hold the Vector feature construction, MLP trunk, data splits, and training hyperparameters fixed and vary only the output head (with its corresponding training loss). All remaining experimental settings follow Appendix E.4. Including One provides a common single-response reference for assessing whether each Vector head retains the benefit of multi-sampling.

Table A.2 shows that the strongest prediction head varies by task. The Beta head is strongest on all four metrics for population-level Twin-2K-500 and OpinionQA and for individual-level Twin-2K-500 and EEDI. The scalar MSE head is strongest throughout population-level EEDI. On individual-level OpinionQA, the Gaussian head is strongest on MAE, Acc, and HA, while the Beta head is

Table A.2: Prediction-head ablation across population- and individual-level tasks.

Level	Dataset	Method	Queries	MAE (↓)	Acc (↑, %)	HA (↑, %)	SA (↑, %)
Population	**Twin-2K-500**	One	1	57.1	86.7	45.0	41.1
		Vector, Gaussian Head (NLL)	50	46.1**	89.3**	58.9**	48.8**
		Vector, Beta Head (Beta NLL)	50	**38.8****	**90.9****	**73.6****	**56.3****
		Vector, Scalar head (MSE)	50	42.5**	90.2**	64.3**	52.2**
Population	**OpinionQA**	One	1	44.4	88.9	61.3	50.0
		Vector, Gaussian Head (NLL)	50	30.6***	92.3***	72.2**	56.8***
		Vector, Beta Head (Beta NLL)	50	**28.6****	**92.9****	**74.8****	**58.3****
		Vector, Scalar head (MSE)	50	30.5**	92.4**	74.3**	56.3**
Population	**EEDI**	One	1	35.5	88.2	60.2	50.9
		Vector, Gaussian Head (NLL)	50	28.3***	90.6***	72.8***	57.1***
		Vector, Beta Head (Beta NLL)	50	27.2***	90.9***	73.3***	58.0***
		Vector, Scalar head (MSE)	50	**25.8****	**91.4****	**77.8****	**59.4****
Individual	**Twin-2K-500**	One	1	53.3	69.2	57.5	56.1
		Vector, Gaussian Head (NLL)	8	49.3***	72.8***	62.0***	60.2***
		Vector, Beta Head (Beta NLL)	8	**48.4****	**73.7****	**62.1****	**61.6****
		Vector, Scalar head (MSE)	8	50.3***	71.9***	60.2***	59.0***
Individual	**OpinionQA**	One	1	87.5	78.1	48.0	44.7
		Vector, Gaussian Head (NLL)	8	**86.8**	**78.3**	**49.3****	46.1***
		Vector, Beta Head (Beta NLL)	8	87.1	78.2	48.3	**47.8****
		Vector, Scalar head (MSE)	8	87.1*	78.2*	48.3	44.8
Individual	**EEDI**	One	1	18.3	93.9	81.1	60.1
		Vector, Gaussian Head (NLL)	8	21.0***	93.0***	79.3***	60.2
		Vector, Beta Head (Beta NLL)	8	**16.7****	**94.4****	**82.7****	**60.6****
		Vector, Scalar head (MSE)	8	17.4**	94.2**	82.0**	60.5**

Notes. Bolding and significance conventions follow Table 3.

strongest on SA. Across the 24 dataset-task-metric comparisons, the Beta head is best in 17, the scalar MSE head in 4, and the Gaussian head in 3.

The scalar MSE head remains a consistent default, improving on One for all four metrics in all six tasks. Two-sided paired tests with two-way cluster-robust standard errors show that every population-level head comparison reaches $p < .05$, and all three heads improve on all four individual-level Twin-2K-500 metrics with $p < .01$. Under the same tests, individual-level OpinionQA is more metric-specific: Gaussian HA and SA and Beta SA reach $p < .01$, while the scalar head's gains in MAE and Acc reach $p < .10$. On individual-level EEDI, the Beta head improves all four metrics ($p < .05$), and the scalar head also improves all four metrics ($p < .05$). The distributional-head gains are not uniform: the Gaussian head is significantly worse than One on EEDI MAE, Acc, and HA ($p < .01$), with no significant SA difference. We therefore retain the scalar MSE head as the task-robust default, while the head-specific gains confirm that the benefit of Vector is not tied to a single prediction-head specification.

D.4 Robustness to Response-Vector Width K in Population-Level Task

This ablation examines how the rollout size K affects our multi-sampling method (Vector), asking whether performance depends on a large single-source rollout pool or whether a modest number of persona-conditioned gpt-4o draws already captures the useful signal. To isolate rollout size from source quality and model capacity, we fix the source to gpt-4o, keep the same prediction head used in table 3, and evaluate $K \in \{1, 8, 25, 50, 100, 200\}$ on the population-level task.

The dominant pattern in Figure A.1 is front-loaded improvement followed by saturation. Performance improves in the correct direction in all twelve dataset-metric panels through $K{=}25$, by which point each panel has already captured 82 to 100% of its maximum improvement over $K{=}1$ anywhere in the evaluated grid. Ten of the twelve panels attain their best displayed mean by $K{=}50$; the only later optima are Twin-2K-500 HA at $K{=}100$ and SA at $K{=}200$, and these do not coincide with consistent gains in its other metrics. Across the full sweep, the largest reduction in MAE is 14.54 points and the largest gain among the accuracy-based metrics is 21.79 percentage points. Thus, $K{=}25$ marks the main knee of the curves, while $K{=}50$ is a conservative common default within the plateau rather than a dataset-specific choice selected from the test curves.

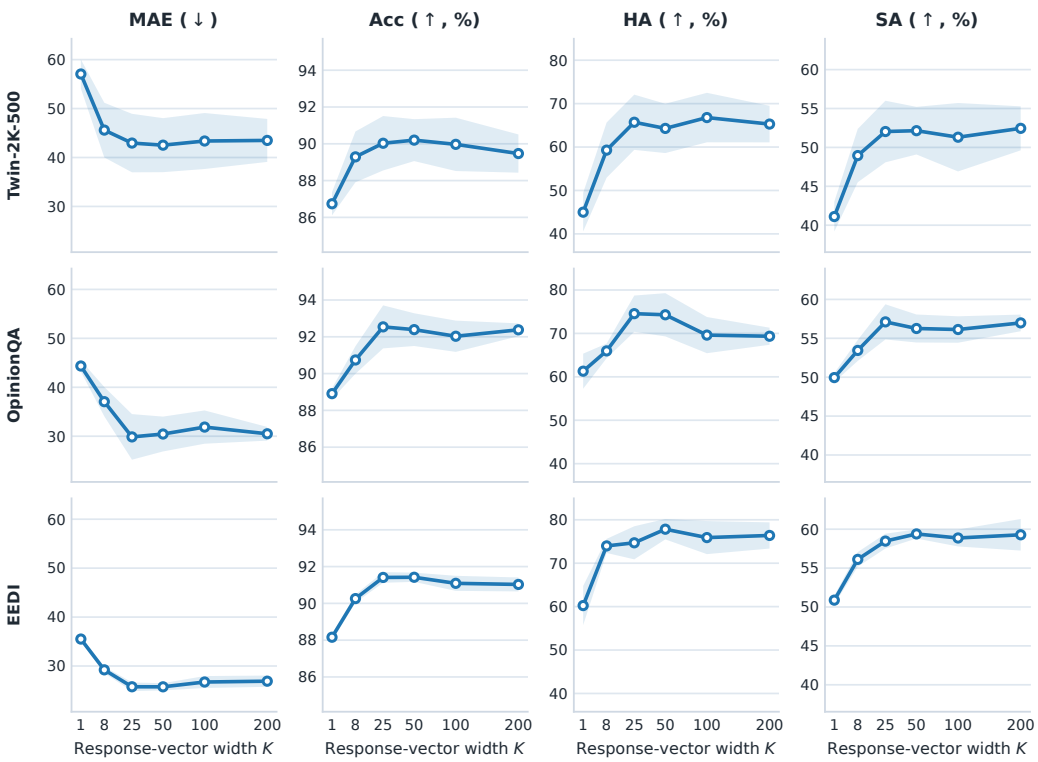


Figure A.1: Population-level `Vector` performance by response-vector width K for single-source `gpt-4o`. Lines report mean performance, and shaded areas indicate one standard deviation over 5 seeds.

EEDI shows a seemingly surprising early scaling effect. As noted in Section 5.2, EEDI contains only 9 distinct observed profiles, so profiles are sampled with replacement when constructing vectors with $K > 9$. Nevertheless, performance improves through K=25 before leveling off. This pattern does not indicate greater profile diversity. Rather, additional draws provide a denser Monte Carlo representation of the response distribution over the same fixed profile support, reducing sampling noise in the empirical mixture presented to the prediction head. The subsequent plateau is consistent with these repeated draws becoming increasingly redundant. Thus, the EEDI result reflects scaling in the number of conditional LLM draws, not in the number of distinct observed profiles.

D.5 Robustness to LLM Source in Population-Level Task

Table A.3: Population-level robustness on Twin-2K-500 across source-specific `LLM field` choices at fixed K=50.

	MAE (↓)					Acc (↑, %)					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	56.8	67.0**	54.4	54.2	**39.9*****	86.5**	84.2**	87.4	87.3	**90.8*****	46.4**	37.5	41.8	45.4**	**68.6*****	44.2**	37.5**	42.4	42.9	**53.2*****
deepseek-v3	45.1*****	67.0**	54.3	51.5**	**37.8*****	89.2*****	84.2**	87.2	87.9**	**91.3*****	67.9*****	37.5	43.6	48.9*	**72.1*****	53.8*****	37.5*	41.8	44.7*****	**54.0*****
gpt-3.5-turbo	80.0*****	67.0	61.2	58.9*	**54.9****	82.2*****	84.2	85.7	86.2*	**87.3****	37.5*	37.5	35.0	39.6**	**44.6****	36.8**	37.5	38.8	40.5*	**42.7***
gpt-4o-mini	79.2*****	67.0**	54.5	53.6	**45.1****	81.3*****	84.2**	87.3	87.4	**89.4****	25.0*****	37.5	45.7	47.1	**60.0*****	30.1*****	37.5**	42.9	43.7	**50.2*****
gpt-4o	67.7	67.0*	57.1	51.9**	**42.5****	83.8	84.2*	86.7	87.9**	**90.2****	39.3	37.5*	45.0	47.1	**64.3****	38.4	37.5	41.1	44.1**	**52.2****
gpt-5-mini	80.8*****	67.0**	53.6	52.4*	**41.3*****	81.3*****	84.2**	87.5	87.8*	**90.3*****	26.8*****	37.5**	42.1	45.7*****	**67.5*****	29.7*****	37.5**	43.2	44.0*	**52.3*****
llama-3.3-70B	87.8*****	67.0**	53.2	51.7*****	**42.6*****	79.2*****	84.2**	87.6	87.9***	**90.0*****	21.4*****	37.5*	43.2	45.4*	**64.3*****	25.3*****	37.5**	43.0	44.1**	**52.2*****
mistral-7B-v0.3	76.7*****	67.0*	56.5	54.5*****	**42.5****	81.2*****	84.2**	86.8	87.3*****	**89.8****	36.8*****	37.5*	42.5	43.6	**63.9****	31.3*****	37.5*	41.5	42.5**	**51.5*****

Notes. Bolding and significance conventions follow Table 3.

Table A.4: Population-level robustness on OpinionQA across source-specific `LLM field` choices at fixed K=50.

	MAE (↓)					Acc (↑, %)					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	52.0*****	46.9**	44.2	39.0*****	**35.2*****	87.0*****	88.3**	89.0	90.2*****	**91.2*****	45.5*****	60.0*	57.9	63.9**	**66.5****	43.9*****	47.5**	48.9	51.0*****	**53.6*****
deepseek-v3	56.9*****	46.9*****	43.1	37.6*****	**31.9*****	85.8*****	88.3*****	89.2	90.6*****	**92.0*****	41.6*****	60.0	61.0	66.8*	**69.9****	42.9*****	47.5*****	50.5	52.0*****	**55.4*****
gpt-3.5-turbo	52.3*****	46.9**	44.0	40.8*****	**37.5****	86.9*****	88.3**	89.0	89.8*****	**90.6****	53.2*****	60.0	60.8	63.4	**68.6****	47.4*****	47.5**	49.5	51.7*****	**53.9****
gpt-4o-mini	63.8*****	46.9*****	43.2	40.6**	**35.1*****	84.1*****	88.3*****	89.2	89.9**	**91.2*****	45.5*****	60.0	61.0	66.0*	**69.1****	40.2*****	47.5*****	50.0	50.9**	**54.3*****
gpt-4o	40.7	46.9	44.4	36.2*****	**30.5****	89.8	88.3	88.9	91.0*****	**92.4****	61.0	60.0	61.3	68.3*	**74.3****	51.7	47.5*	50.0	53.2**	**56.3****
gpt-5-mini	52.6*****	46.9	46.4	38.2*****	**33.5*****	86.9*****	88.3	88.4	90.4*****	**91.6*****	53.2**	60.0**	56.6	67.5*****	**68.3*****	44.2*****	47.5*	48.8	52.3*****	**55.2*****
llama-3.3-70B	50.8*****	46.9*	44.4	38.2*****	**35.3*****	87.3*****	88.3*	88.9	90.4*****	**91.2*****	46.8*****	60.0	60.0	**64.2**	**64.2*****	45.8*****	47.5*	49.3	51.4*****	**52.8*****
mistral-7B-v0.3	56.2*****	46.9*****	44.3	43.4	**43.0**	86.0*****	88.3*****	88.9	89.1	**89.2**	57.1*****	60.0	61.3	62.3	**64.7**	44.6*****	47.5*****	49.9	49.9	**50.7**

Notes. Bolding and significance conventions follow Table 3.

Table A.5: Population-level robustness on EEDI across source-specific `LLM field` choices at fixed K=50.

	MAE (↓)					Acc (↑, %)					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*****	37.6*****	34.0	30.9*****	**26.4*****	85.7*****	87.5*****	88.7	89.7*****	**91.2*****	60.2	57.1*	62.9	70.4**	**75.9****	50.1*****	48.2*****	51.8	53.8*****	**57.1*****
deepseek-v3	39.5*****	37.6	36.3	30.1*****	**26.0*****	86.8*****	87.5	87.9	90.0*****	**91.3*****	62.7	57.1	59.0	69.6*****	**74.9*****	51.1*****	48.2*****	49.5	54.5*****	**57.4*****
gpt-3.5-turbo	75.2*****	37.6	37.9	37.9	**36.2**	74.9*****	87.5	87.4	87.4	**88.5**	39.8*****	57.1	55.4	54.7	**57.3**	36.9*****	48.2	48.4	48.4	**48.6**
gpt-4o-mini	62.9*****	37.6*****	32.6	30.2*****	**24.8*****	79.0*****	87.5*****	89.1	89.9*****	**91.7*****	41.0*****	57.1**	64.6	71.3*****	**82.4*****	38.9*****	48.2*****	53.0	55.0*****	**61.1*****
gpt-4o	48.2**	37.6	35.5	30.9**	**25.8*****	83.9**	87.5	88.2	89.7**	**91.4*****	55.4	57.1	60.2	70.6**	**77.8****	47.5	48.2**	50.9	54.5*****	**59.4*****
gpt-5-mini	64.3*****	37.6*****	30.4	29.7*****	**23.4*****	78.6*****	87.5*****	89.9	90.1*****	**92.2*****	43.4*****	57.1**	70.6	73.0*	**86.5*****	39.7*****	48.2*****	54.5	55.2*****	**61.4*****
llama-3.3-70B	63.2*****	37.6*****	32.3	30.7*****	**25.9*****	78.9*****	87.5*****	89.2	89.8*****	**91.4*****	43.4*****	57.1**	66.5	71.3*****	**80.0*****	39.5*****	48.2*****	53.2	54.3*****	**59.3*****
mistral-7B-v0.3	78.3*****	37.6	37.6	37.5	**37.0**	73.9*****	87.5	87.5	87.5	**88.3**	28.9*****	57.1	**58.3**	56.9	56.6	30.7*****	48.2**	**49.3**	48.9	48.9*

Notes. Bolding and significance conventions follow Table 3.

This robustness check asks whether the population-level advantage of `Vector` depends on the particular LLM source used to generate persona-conditioned responses. We keep the single-source rollout size fixed at K=50, use the same prediction head as in the main population-level experiment, and vary the source-specific `LLM field`. Each row uses the same source for `Base LLM`, `One`, `Mean`, and `Vector`; `w/o LLM` is shared within each dataset because it does not use an LLM input. The tables report source-specific robustness over model seeds 0 to 4 and use the same significance convention as table 3.

Results. Tables A.3 to A.5 show that the main pattern is generally stable across LLM sources rather than being specific to `gpt-4o`. All reported p-values below come from two-sided paired tests against the source-matched `One` with two-way cluster-robust standard errors and are unadjusted. On Twin-2K-500, `Vector` improves on the source-matched `One` and has the best displayed mean in all 32 source-metric comparisons, with $p < .05$ in 31 and $p < .10$ in the remaining comparison. On OpinionQA, `Vector` again improves on `One` in all 32 comparisons and has the best displayed mean in all 32 (tied with `Mean` once), with 28 improvements reaching $p < .05$; the four nonsignificant comparisons all use `mistral-7B-v0.3`. On EEDI, `Vector` has the best displayed mean in 30 of 32 comparisons and improves on `One` in 30, including 24 with $p < .05$. The two unfavorable EEDI comparisons are `HA` and `SA` under `mistral-7B-v0.3`; in particular, the starred `SA` result is a marginal difference in the unfavorable direction, not an improvement. Overall, `Vector` has the best displayed mean in 94 of 96 comparisons and improves on `One` in 94, including 83 with $p < .05$. Thus, source quality affects the absolute performance and the strength of the gain, while the overall multi-sample pattern is not driven solely by the canonical `gpt-4o` source.

D.6 Pricing Application: Design, Revenue Metrics, and Offer Selection

Design and model inputs. The pricing application in Section 5.4 uses 40 products, each evaluated by 167 respondents at an assigned price from 0% to 200% of its regular price in 20% increments. We retain the original 8,016/2,004 record-level training/test partition; the pricing subset contains 5,318 training and 1,362 test records. We identify each product and assigned price from the respondent's actual question, and aggregate by product rather than presentation position. The calibration models use product indicators and relative price for pricing questions, together with the existing respondent features and LLM responses. We retain the original MLP settings and the five training seeds. `Base LLM` and `One` use GPT-4o.

Demand estimation and revenue. Let r denote the assigned relative price and p_q the regular dollar price of product q. For positive-price records, we fit partially pooled demand curves,

$$\operatorname{logit} D_q(r) = a_q + b_q(r-1). \tag{27}$$

Product intercepts and slopes receive Gaussian shrinkage toward shared coefficients. We minimize the mean Bernoulli cross-entropy loss with equal weight per record, using binary human purchases or a method's predicted purchase probabilities, clipped to $[0,1]$, as the targets. Five respondent folds within the training human records select the shrinkage penalty; the selected value is fixed for all demand fits. The human benchmark D_q^{H} is fit to held-out human purchases, while $\widehat{D}_q^{(m,s)}$ is fit to method m's held-out predictions under training seed s.

All offers are evaluated at their regular prices. Human and predicted survey-implied revenues per exposed respondent are

$$R_q^{\mathrm{H}} = p_q D_q^{\mathrm{H}}(1), \qquad \widehat{R}_q^{(m,s)} = p_q \widehat{D}_q^{(m,s)}(1). \tag{28}$$

With $Q = 40$ products, revenue prediction error is

$$\text{Revenue MAE}^{(m,s)} = \frac{1}{Q}\sum_{q=1}^{Q}\left|\widehat{R}_q^{(m,s)} - R_q^{\mathrm{H}}\right|. \tag{29}$$

Zero-price records are excluded from demand fitting and revenue evaluation. The 118 free-product test responses provide a separate purchase-probability diagnostic, with MAE of 0.195 for `Base LLM` and 0.193 for `Vector`.

Offer selection at different capacities. For $k \in \{5, 10, 15, 20\}$, let $\widehat{\mathcal{S}}_k^{(m,s)}$ contain the k products with the highest predicted regular-price revenues, and let $\mathcal{S}_k^{\mathrm{H}}$ contain the k products with the highest human survey-implied revenues. Ties are broken by stable product ID. Selection overlap is

$$\text{Overlap}_k^{(m,s)} = 100 \times \frac{|\widehat{\mathcal{S}}_k^{(m,s)} \cap \mathcal{S}_k^{\mathrm{H}}|}{k}. \tag{30}$$

For a fixed exposure budget allocated equally among the selected products, regret per exposure is

$$\text{Regret}_k^{(m,s)} = \frac{1}{k}\left(\sum_{q\in\mathcal{S}_k^{\mathrm{H}}} R_q^{\mathrm{H}} - \sum_{q\in\widehat{\mathcal{S}}_k^{(m,s)}} R_q^{\mathrm{H}}\right). \tag{31}$$

Both selections are evaluated using the same human demand curves. For the additional capacities, we select and evaluate offers separately for each of the five matched training seeds, then report the mean metrics. The top-10 entries follow Table 4. These quantities measure survey-implied revenue under the specified exposure allocation.

Table A.6: Offer selection at regular prices under capacities of different offers.

	$k = 5$		$k = 10$		$k = 15$		$k = 20$	
Method	Overlap (%, ↑)	Regret ($/exposure, ↓)	Overlap (%, ↑)	Regret ($/exposure, ↓)	Overlap (%, ↑)	Regret ($/exposure, ↓)	Overlap (%, ↑)	Regret ($/exposure, ↓)
`Base LLM`	40.0	2.735	40	2.514	40.0	1.961	40.0	1.700
`w/o LLM`	76.0	0.334	78	0.405	**69.3**	**0.307**	**77.0**	**0.211**
`One`	76.0	0.459	80	0.330	66.7	0.354	74.0	0.255
`Mean`	**80.0**	**0.283**	80	0.305	68.0	0.360	**77.0**	0.285
`Vector`	**80.0**	0.369	**82**	**0.303**	**69.3**	0.315	76.0	0.268

Notes. Best displayed values within each column are bold.

Table A.6 extends the main comparison to different exposure capacities. `Vector` lowers regret relative to `Base LLM` at all four capacities, while the relative ranking of the calibration variants depends on the number of selected offers.

E Experiment Details

E.1 Dataset Details

Across the three datasets, we use the same ordered set of eight LLM sources as in the OpinionQA rollout: `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`. Holding the source identities and ordering fixed makes the multi-source response vectors comparable across datasets; item-level coverage exceptions in the final artifacts are noted below. The statistics below are computed from the final population- and individual-level data files used in our experiments.

Twin-2K-500 provides a large U.S. general survey panel for evaluating whether LLM responses can help recover both aggregate opinion and respondent-level heterogeneity. The full dataset contains 665 question records from 2,058 U.S. respondents, for a total of 1,222,466 respondent-question responses. For the individual-level benchmark, we retain all 60 questions with complete respondent coverage among those supported by the source individual-response pipeline. To keep its sample size broadly comparable to the other two individual-level benchmarks, we randomly sample 167 respondents from the full respondent pool. This sample closely matches the full 2,058-respondent pool on key demographic marginals. Pairing these respondents with the 60 questions produces a complete $167 \times 60 = 10{,}020$ respondent-by-question matrix with nonmissing human responses and aligned 8 LLM responses. For the population-level benchmark, we independently audit the official closed-choice catalog and retain all 279 questions satisfying prespecified strict Likert-type criteria, including at least three monotonically ordered response categories, explicit verbal anchors, and an opinion- or judgment-oriented construct. Table A.7 reports the exact task-level composition of the individual- and population-level question sets, following the task-and-format presentation of the original dataset.

We randomly draw 167 respondents from the full pool and pair them with the 60 individual-level questions, producing 10,020 respondent-question observations. To assess whether this draw materially changes the observable sample composition, Table A.8 compares its empirical marginal distributions with those of all 2,058 respondents. The selected sample closely tracks the full pool: the largest category-level absolute difference is 0.2 percentage points for sex, 1.5 for age, 2.0 for education, and 3.3 for race or origin.

OpinionQA is a public opinion dataset drawn from Pew Research's American Trends Panel, covering politically salient issues for which respondent attitudes can be difficult to infer from basic demographic profiles. The full dataset contains 385 five-point opinion items answered by 32,864 respondents, yielding 1,476,868 human responses. For the individual-level benchmark, we use the

Table A.7: Composition of the retained Twin-2K-500 questions. A dash indicates that a task is not included in that benchmark.

Task or scale	Response format	Individual	Population
Personality scales			
Big Five personality	5-point Likert	–	44
Need for cognition	5-point Likert	–	18
Agentic vs. communal values	9-point Likert	–	24
Consumer minimalism	5-point Likert	–	12
Empathy	5-point Likert	–	20
Green values	5-point Likert	–	6
Conscientiousness	9-point Likert	–	8
Anxiety	4-point Likert	–	21
Individualism vs. collectivism	5-point Likert	–	16
Regulatory focus	7-point Likert	–	10
Tightwad vs. spendthrift	5-/11-point ordered choice	–	4
Need for uniqueness	5-point Likert	–	12
Self-monitoring	6-point Likert	–	13
Self-concept clarity	5-point Likert	–	12
Need for closure	5-point Likert	–	15
Maximization	5-point Likert	–	6
Opinion and behavioral tasks			
Outcome bias	7-point ordered rating	–	2
Disease framing	6-point program preference	–	2
Conjunction problem	6-point probability rating	–	4
Myside bias	6-point policy rating	–	2
Less-is-more judgments	5-point agreement rating	–	3
Proportion dominance	5-/6-point ordered rating	–	6
False consensus	5-point support rating	10	10
Risk–benefit judgments	7-point rating	8	8
Omission bias	4-point behavioral intention	1	1
Denominator neglect	Binary choice	1	–
Product pricing	Binary purchase choice	40	–
Total		**60**	**279**

Table A.8: Demographic marginal distributions in the full Twin-2K-500 respondent pool and the randomly selected individual-level sample.

Dimension	Category	Full pool (%) ($N = 2{,}058$)	Selected (%) ($N = 167$)	Difference (percentage points)
Sex	Female	50.73	50.90	0.17
	Male	49.27	49.10	0.17
Age	18–29	18.85	20.36	1.51
	30–49	35.71	35.93	0.21
	50–64	31.97	31.14	0.84
	65+	13.46	12.57	0.88
Education	Less than high school	0.83	0.00	0.83
	High school graduate	13.22	13.77	0.56
	Some college, no degree	22.74	23.95	1.21
	Associate's degree	12.29	10.78	1.52
	College graduate/some postgraduate	35.71	37.72	2.01
	Postgraduate	15.21	13.77	1.44
Race or origin	White	66.13	69.46	3.33
	Black	12.20	11.98	0.22
	Hispanic	9.43	8.38	1.04
	Asian	6.80	6.59	0.22
	Other	5.44	3.59	1.85

LLM response data from Huang et al. (2025) and retain records with an observed human response, an available demographic representation, and aligned responses from all eight source LLMs. This restriction yields 9,437 response records spanning all 385 questions and 200 respondents. For the population-level benchmark, we retain all 385 questions and construct question-level targets by aggregating the available human responses from the full respondent pool.

The question identifier records the source survey wave. Table A.9 groups the 385 questions by this wave and reports the corresponding topic for both benchmark levels. Because the individual-level and population-level benchmarks contain the same 385 questions, their topic counts coincide.

Table A.9: Composition of the retained OpinionQA questions. Both benchmark levels contain the same 385 questions.

Wave	Main topic	Response format	Individual	Population
26	Guns	5-point opinion scale	17	17
27	Automation and driverless vehicles	5-point opinion scale	18	18
29	Views on gender	5-point opinion scale	16	16
32	Community types and sexual harassment	5-point opinion scale	27	27
34	Biomedical and food issues	5-point opinion scale	10	10
36	Gender and leadership	5-point opinion scale	4	4
41	Views of America in 2050	5-point opinion scale	21	21
42	Trust in science	5-point opinion scale	46	46
43	Race in America	5-point opinion scale	16	16
45	Misinformation	5-point opinion scale	41	41
49	Privacy and surveillance	5-point opinion scale	29	29
50	American families	5-point opinion scale	26	26
54	Economic inequality	5-point opinion scale	49	49
82	Global attitudes	5-point opinion scale	51	51
92	Political typology	5-point opinion scale	14	14
Total			**385**	**385**

EEDI provides a non-survey response setting in which the target is middle-school students' encoded responses to four-choice mathematics diagnostics rather than stated opinion. The full dataset contains 573 questions and 443,433 responses from 2,287 students. Following the processed benchmark of Huang et al. (2025), we retain 412 questions with at least 50 responses and no graphs or diagrams. Because the available student fields are limited to gender, premium-pupil status, and age, many students share the same observed profile; we therefore aggregate students into profile clusters. For the individual-level benchmark, we retain profile-question records with observed human targets and usable LLM responses, yielding 3,479 records over all 412 questions and 9 profiles, where $y^*_{i,x}$ is the mean answer within each profile-question pair. For the population-level benchmark, we retain the same 412 eligible questions and aggregate the available human responses at the question level. Each retained question maps to the hierarchical EEDI mathematics taxonomy (Wang et al., 2021). Table A.10 reports the second-level topic under each of the three first-level subject areas. Because the individual-level and population-level benchmarks contain the same 412 questions, their topic counts coincide. All 412 EEDI questions appear at the population level. At the individual level, eight profiles cover all 412 questions, while the ninth covers 183, producing the 3,479 profile-question observations reported above.

Table A.10: Composition of the retained EEDI questions. Both benchmark levels contain the same 412 questions.

Topic or skill	Response format	Individual	Population
Number			
Negative Numbers	Four-option multiple choice	89	89
Basic Arithmetic	Four-option multiple choice	86	86
Factors, Multiples and Primes	Four-option multiple choice	63	63
Fractions	Four-option multiple choice	16	16
Indices, Powers and Roots	Four-option multiple choice	9	9
Algebra			
Writing and Simplifying Expressions	Four-option multiple choice	21	21
Formula	Four-option multiple choice	16	16
Co-ordinates	Four-option multiple choice	9	9
Factorising	Four-option multiple choice	8	8
Solving Equations	Four-option multiple choice	7	7
Straight Line Graphs	Four-option multiple choice	2	2
Inequalities	Four-option multiple choice	1	1
Sequences	Four-option multiple choice	1	1
Geometry and Measure			
Units of Measurement	Four-option multiple choice	15	15
Transformations	Four-option multiple choice	15	15
Angles	Four-option multiple choice	13	13
Perimeter and Area	Four-option multiple choice	12	12
2D Names and Properties of Shapes	Four-option multiple choice	9	9
Similarity and Congruency	Four-option multiple choice	8	8
Construction, Loci and Scale Drawing	Four-option multiple choice	7	7
Symmetry	Four-option multiple choice	5	5
Total		**412**	**412**

E.2 PROMPT TEMPLATES

We use a separate persona-conditioned prompt for each individual-level benchmark. Twin-2K-500 uses the multi-question JSON template in our current generation pipeline. OpinionQA and EEDI preserve the released prompts of Huang et al. (2025), including their original double-square-bracket answer formats. All three blocks are templates with angle-bracketed runtime fields. The gray headers and bracketed role labels are display annotations and are not literal model input.

Twin-2K-500. The full respondent background is inserted before a batch of questions, each of which already contains its enumerated options or response scale.

TWIN-2K-500 PROMPT

```
[USER]
You are a survey respondent with the following background:
<demographic profile>

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 or scale in the question.

Valid example output:
{
  "Q1": 2,
  "Q2": 5,
  "Q3": 1
}

Important: Do not include option labels or write anything outside
    the
single JSON object.

Now, please answer the following questions:
Q1: <question text with enumerated options or scale>
...
Qm: <question text with enumerated options or scale>
```

OpinionQA. OpinionQA supplies the respondent's region, age, sex, education, citizenship, marital status, religion, religious attendance, political party, income, political ideology, and race through a demographic string.

OPINIONQA PROMPT

```
[SYSTEM]
You are simulating the behaviors of humans with certain specified
    characteristics to help with a survey study.

[USER]
<demographic string> Please answer the following question:
```

```
[How much, if at all, do you think what happens to black people in
    the country overall affects what happens in your own life? ['1.
    A lot', '2. Some', '3. Not much', '4. Not at all', '5. Refused
    '].]

Please provide your answer choice (a single number from 1 to 5) in
    double square brackets.
```

EEDI. EEDI supplies age, gender, and premium-pupil status through a demographic string, then asks the model to emulate whether a student with that profile would solve the mathematics problem. The released prompt requests a plausible incorrect option when the simulated student is likely to struggle.

EEDI PROMPT

```
[SYSTEM]
You are simulating the behaviors of humans with certain specified
    characteristics to help with a survey study.

[USER]
<demographic string> Given your characteristics, is it likely that
    you would be able to solve the following problem?

Problem: What time will it be 17 hours after 11:00?

A. 04:00
B. 06:00
C. 16:00
D. 28:00

If yes, put the final answer choice (a single letter) in double
    square brackets. If you are likely to struggle with this problem
    , put a plausible incorrect answer choice (a single letter) in
    double square brackets.
```

For each target profile z, the parsed responses are converted to the dataset's admissible integer categories. OpinionQA retains the bracketed integer, while EEDI maps A, B, C, and D to 1, 2, 3, and 4. One response from each of the eight source LLMs is retained in a fixed order, producing $\mathbf{y}'_K = (y'_1, \ldots, y'_K)$ with $K = 8$.

E.3 One Logprob Construction

Following the model-belief construction proposed by Sun & Zhang (2025), `One Logprob` replaces the scalar response used by `One` with an answer-choice probability vector extracted from one `gpt-4o` generation. For each question and, at the individual level, each respondent profile, we instruct the model to return one admissible integer answer and request the top 20 token log probabilities at the first generated position. Let $\mathcal{A}_x = \{a_x, \ldots, b_x\}$ be the admissible choices and $\mathcal{T}_{20,x}$ the returned candidates. For each $c \in \mathcal{A}_x$, we collect the returned probability mass assigned to tokens that encode c,

$$m_x(c) = \sum_{t \in \mathcal{T}_{20,x}:\, \text{choice}(t)=c} \exp\big(\ell_x(t)\big), \qquad p_x(c) = \frac{m_x(c)}{\sum_{r \in \mathcal{A}_x} m_x(r)}. \tag{32}$$

If no returned candidate encodes a given admissible choice, that choice receives zero mass before normalization. Thus, the construction normalizes only the visible mass assigned to admissible choices and does not treat a missing choice as a failed generation.

Each benchmark uses a fixed-width feature whose width equals the maximum number of admissible choices within that dataset and task level. The entries for the current question occupy the first

Table A.11: Debiasing MLP hyperparameters at the population and individual levels, shared across datasets and feature variants.

Hyperparameter	Population-level	Individual-level
Hidden layers	$(512, 256, 128)$	$(6144, 3072, 1536, 768, 384)$
Activation / output	ReLU / linear	ReLU / linear
Optimizer	Adam	Adam
Weight decay λ	0.01	10^{-6}
Initial learning rate	5×10^{-4}	2×10^{-4}
LR scheduler	ReduceLROnPlateau	
LR patience / factor / floor	10 / 0.5 / 10^{-6}	
Batch size	64	512
Max epochs	1,500	3,500
Early-stop patience / $\Delta_{\min}$	20 / 10^{-6}	30 / 10^{-6}
Feature standardization	yes	yes

$|\mathcal{A}_x|$ coordinates in choice order, and all remaining coordinates are set to zero. This construction keeps the input dimension constant when questions have different numbers of choices. The resulting probability vector is appended to the question embedding at the population level and to the question and profile embeddings at the individual level. Because the vector comes from one generation, `One Logprob` has $K = 1$.

E.4 Calibration Algorithm Details

We use the MLP g_θ described in section 3.3 for every dataset and feature variant. Each hidden layer is followed by a ReLU activation (Nair & Hinton, 2010), and the final layer is a scalar linear output. Human and LLM responses are normalized to $[0, 1]$. Input features are standardized using statistics fitted on the training split, while the raw ordering of $\mathbf{y}'_K$ is preserved. The training target is the normalized human response y^* rather than a residual target. We minimize the ℓ_2-regularized mean squared error using Adam (Kingma & Ba, 2015), with regularization strength λ implemented as weight decay. A plateau scheduler halves the learning rate after ten epochs without validation-loss improvement, down to a floor of 10^{-6}. The population-level MLP is smaller because each dataset provides only a few hundred training questions, whereas the individual-level task contains many more respondent-question pairs. The complete training hyperparameters are reported in table A.11.

E.5 Evaluation Sampling Settings

All source LLMs use temperature 0.7 and a default output-token limit of $1{,}024$ per request. Top-p and, where supported, top-k are omitted from the requests, leaving their values at the provider or serving-engine defaults.

The `One Logprob` baseline uses a separate configuration: OpenAI Chat Completions with temperature 1.0, default top-p, and a one-token completion limit. Only this baseline requests `logprobs=true` with `top_logprobs=20`, following the official implementation (Sun & Zhang, 2025). The resulting answer-choice probability vector is constructed as described in Appendix E.3.