Model-Free Assessment of Simulator Fidelity via Quantile Curves

Garud Iyengar [*1] **Yu-Shiou(Willy) Lin** [*1] **Kaizheng Wang** [*1]

Abstract

As generative AI models are increasingly used to simulate real-world systems, quantifying their "sim-to-real" gap is critical. We study this gap across a population of input settings, called *scenarios*, e.g. survey questions or operating conditions. The population-level quantities defining the real-world and simulated system are only observed through finite, often heterogeneous, samples. Consequently, the population-level discrepancy cannot be directly computed, and standard predictive inference methods that target observable outputs are ill-suited. We propose a model-agnostic framework that constructs confidence sets for the latent parameters, forms a conservative proxy for the sim-to-real discrepancy, and estimates its quantile function across scenarios. The resulting calibrated risk profile allows for inference on a new scenario, tail-risk summaries such as Conditional Value-at-Risk (CVaR), and principled comparisons across simulators. Our method applies to a range of output spaces, including categorical survey responses and continuous outcomes. We demonstrate its utility by evaluating the alignment of four major LLMs with human populations on the WorldValueBench dataset.

Keywords: Simulator fidelity, quantile function, LLM human–simulation, digital twins

1. Introduction

The recent advances in artificial intelligence (AI) have accelerated the adoption of simulators across operations and manufacturing, agent-based modeling in social-science research, user surveys, and education (Zhang et al., 2019; Macal, 2016; Argyle et al., 2023; Aher et al., 2023). At the platform level, industry ecosystems such as NVIDIA Omniverse and Earth-2 exemplify the push toward high-fidelity, AI-enabled digital twins, and LLMs are increasingly used to build generative agents that emulate human responses (Park et al., 2023; Lu et al., 2025). Collectively, these developments are driving growth in simulation across domains.

Against this backdrop, quantifying the *distributional* discrepancy between simulated and real-world, or ground truth, outcomes is critical. This "sim-to-real" gap is a pervasive challenge in domains ranging from LLMs to computing systems and robotics (Gao et al., 2025; Durmus et al., 2023; Tobin et al., 2017; Peng et al., 2018). Crucially, the gap depends on the particular *scenario* (e.g., operating conditions or survey questions). Thus, the gap in a particular scenario is a sample from *latent scenario-level distribution* for both the real and simulated systems. Assessing fidelity therefore requires characterizing this discrepancy distribution, rather than a single average error, which can hide rare but consequential failures. This distributional characterization is the primary objective of our work.

Existing uncertainty quantification methods are insufficient for this task since they focus on assessing *individual* outcomes (e.g., conformal prediction), and do not quantify the *distribution* of the discrepancy between the simulator and real system across different scenarios. Conflating outcome validity with distributional fidelity may harm downstream applications. For instance, in survey research, an LLM-based simulator might generate coherent individual responses while yielding systematically biased *population-level aggregates* (Westwood, 2025), and recent work has identified distributional alignment as a key objective (Cao et al., 2025). We provide theoretical guarantees for sim-to-real discrepancy assessment without making structural assumptions about the simulator, treating it strictly as a black box.

Contributions We characterize the quantile function of the discrepancy between real world and simulated output distributions across possible scenarios. Since the real-world distribution is observed only through finite, often heterogeneous samples, the true quantile function cannot be computed. We construct per-scenario confidence sets for the latent real-world parameters, use them to form a proxy for the sim-to-real discrepancy, and establish finite-sample guar-

[1]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA. Correspondence to: Yu-Shiou(Willy) Lin <yl5782@columbia.edu>.

Accepted to the Workshop on Statistical Frameworks for Uncertainty in Agentic Systems at ICML 2026, Seoul, South Korea.
Copyright 2026 by the author(s).

antees for the quantile curve at any desired level. Moreover, we allow the coverage of the confidence set to be chosen as a function of the number of available samples. This yields an adaptive procedure that produces tighter quantile curves in data-rich scenarios, while preserving reliability and interpretability when data are scarce.

We illustrate our methodology using the dataset *WorldValueBench* (Zhao et al., 2024), constructed from the World Values Survey (Haerpfer et al., 2020), by generating synthetic response profiles and benchmarking four LLMs against human survey data. Leveraging the large human sample size, we also study tightness and propose a data-driven procedure to construct confidence bands for the population quantile function.

Our main contributions are:

1. **Model-free, finite-sample fidelity profiling.** A black-box assessment framework with finite-sample, distributional guarantees under heterogeneous sample sizes across scenarios.
2. **Quantile-based risk summaries.** A calibrated discrepancy quantile curve enabling the estimation of tail-risk measures, such as CVaR, and other summaries.
3. **Inference and comparison.** Statistical inference for real-world parameters in new scenarios and principled comparisons across simulators.

Related Literature The sim-to-real discrepancy has been a central concern in the uncertainty quantification (UQ) literature. Following the taxonomy of Roy & Oberkampf (2011), we distinguish *input* uncertainty from *output* uncertainty. Work on input uncertainty (Chen et al., 2024; Barton et al., 2014; Lam, 2022) typically considers a simulator with parameters learned from finite real-world input data, and studies how the uncertainty of learned parameters, together with Monte Carlo noise, propagates to output quantities. These methods usually need refitting on bootstrap samples of the input data. Our setting is different: modern simulators such as LLMs or digital twins are often black boxes and may be systematically misspecified. Moreover, we do not have the training data, and it can be computationally prohibitive to refit. We therefore study *output-level fidelity*: the discrepancy between real and simulated output distributions across a population of scenarios. This is related to UQ using simulation output (Jeon et al., 2024) and simulation-based inference (Cranmer et al., 2020; Papamakarios et al., 2019), but our target is not Monte Carlo error or posterior inference on simulator parameters; it is a finite-sample assessment of scenario-level sim-to-real discrepancy. Two-sample testing is also model-agnostic and sample-based (Gretton et al., 2012; Schrab et al., 2023), but addresses a different question: whether real and simulated samples are distinguishable, rather than the distribution of the sim-to-real gap across scenarios.

A second line of work studies LLM-based human simulation and opinion alignment, asking whether synthetic responses reproduce human responses (Santurkar et al., 2023; Huang et al., 2025). Existing approaches often report aggregate alignment metrics or confidence sets for specific population quantities. Our goal is different: rather than summarizing fidelity by a single metric or validating one estimand, we estimate a quantile curve that describes how simulator error varies across scenarios.

Finally, our guarantees are related to conformal and predictive inference (Vovk et al., 2005; Bates et al., 2021), which provide finite-sample, distribution-free guarantees for black-box predictors. Standard conformal methods target future *observable* outcomes, whereas our target is a *scenario-level distributional parameter* observed only through finite, often heterogeneous batches. Recent conformal variants for distributional objects (Snell et al., 2022; Budde et al., 2025) are closer in spirit, but do not directly handle latent scenario-level targets with heterogeneous real and simulated sample sizes. We address this gap by calibrating quantile-based sim-to-real risk profiles under heterogeneous batching.

The remainder of the paper proceeds as follows. In Section 2 we present a motivating example, and formally state the problem. In Section 3 we present the main theoretical results, and in Section 4 we discuss an application of our methodology. We also address the tightness of our methodology to the actual quantile curve in Section 5. In Section 6 we conclude by discussing future directions.

2. Framework and Motivating Example

We begin with a motivating use case of a survey-style LLM and then formalize the quantile estimation problem. Although we introduce the notation using an LLM example, the framework itself is general. In Appendix E, we illustrate how the same formulation applies to other settings, including discrete-event and other simulators with both categorical and continuous outputs.

Suppose a media research company plans to survey customers on a particular question with multinomial outcomes (e.g., 'Agree', 'Neutral', 'Disagree'). The company wants to estimate customer opinion before committing resources to an expensive population study, and is considering whether to directly use a digital twin based on pre-trained LLM or develop its own fine-tuned LLM. To make this decision, the company needs to evaluate the distribution of the sim-to-real gap across the range of questions that it is likely to encounter. While pre-trained LLMs can appear to provide coherent predictions, in Appendix F we show examples where several pre-trained LLMs underperform the baseline that randomly

picks answers. The methods developed in this work assist such assessments.

Next, we formalize the setting and the associated theoretical challenges. There are two sources of uncertainty here: variation across scenarios and sampling noise within each scenario. First, there is a probability distribution Ψ of scenarios (e.g. questions, operating conditions, etc.), and the dataset consists of m i.i.d. scenarios $\{\psi_j\}_{j=1}^m$ from Ψ. For each scenario ψ, the ground-truth samples are drawn from the conditional real-world distribution $Q^{rw}(\cdot \mid \psi)$, and simulated samples are drawn from the conditional simulator distribution $Q^{sim}(\cdot \mid \psi)$. [1] In the LLM-survey example, the distributions $Q^{rw}(\cdot \mid \psi)$ and $Q^{sim}(\cdot \mid \psi)$ are obtained by marginalizing over human/synthetic profiles; see Appendix A.1 for details. For each scenario $j \in [m]$, we observe a heterogeneous number of finite ground-truth samples:

$$y_{j,i}^{rw} \sim Q^{rw}(\cdot \mid \psi_j), \qquad i = 1, \ldots, n_j.$$

For simplicity, we assume that we have the same number of synthetic samples k from the simulator:

$$y_{j,\ell}^{sim} \sim Q^{sim}(\cdot \mid \psi_j), \qquad \ell = 1, \ldots, k.$$

Let $p(\psi) \in \Theta$ denote a quantity of interest corresponding to the real world distribution $Q^{rw}(\cdot \mid \psi)$, and let $q(\psi) \in \Theta$ denote the same quantity corresponding to the simulator distribution $Q^{sim}(\cdot \mid \psi)$. In the context of the media research company, $p(\psi)$ (resp. $q(\psi)$) denotes the real-world (resp. simulator) probability distributions of the responses in scenario ψ. For simplicity, we write $p_\psi := p(\psi)$ and $q_\psi := q(\psi)$; for past scenarios ψ_j, we write $p_j := p(\psi_j)$ and $q_j := q(\psi_j)$. The samples $\{y_{j,i}^{rw}\}_{i=1}^{n_j}$ and $\{y_{j,\ell}^{sim}\}_{\ell=1}^{k}$ can be used to produce plug-in estimates $\hat{p}_j$ and $\hat{q}_j$ of p_j and q_j, respectively.

Table 1. Illustrative Example of multinomial distribution $\mathcal{D}$

Scenario	**True Distribution** (p_j)	n_j	**Real Output** (y^{rw})
1	(0.8, 0.1, 0.1)	10	(1, 1, 2, 3, 1, 2, 1, 1, 1, 1)
2	(0.3, 0.4, 0.3)	4	(3, 1, 2, 2)
$\vdots$	...	$\vdots$	...
m	(0.05, 0.05, 0.9)	8	(1, 3, 3, 3, 3, 2, 2, 3)

The dataset is $\mathcal{D} = \left\{(\psi_j, \hat{p}_j, \hat{q}_j, n_j, k)\right\}_{j=1}^m$. For concreteness, Table 1 provides an example of $\mathcal{D}$ under our survey example. p_j is a multinomial parameter and the human sample sizes n_j may vary across scenarios, so the realized outcomes $\{y_{j,i}^{rw}\}_{i=1}^{n_j}$ differ in sample size. We record outcomes as category indices. For instance, when $j = 2$ the true distribution is $p_2 = (0.3, 0.4, 0.3)$ and a sample of size $n_2 = 4$ yields $\hat{p}_2 = (0.25, 0.5, 0.25)$. The simulator analog $\hat{q}_j$ is constructed from k simulator samples.

The discrepancy for a question ψ between simulated and real output distributions is defined as

$$\Delta_\psi := L(p_\psi, \hat{q}_\psi),$$

where $L : \Theta \times \Theta \to [0, \infty)$ is a user-chosen discrepancy function. We define the discrepancy in terms of $\hat{q}_\psi$ to reflect a fixed simulator query budget; when simulator sampling is effectively unlimited, one may replace $\hat{q}_\psi$ with q_ψ to study the intrinsic discrepancy $L(p_\psi, q_\psi)$. Our method works with any function L and allows practitioners to choose one that suits their application, e.g. Kullback–Leibler (KL) divergence for categorical outputs or a Wasserstein distance for empirical distributions.

Now, we can formally state our task:

Construct a function $\hat{V}_m(\cdot, \mathcal{D}) : (0,1) \to \mathbb{R}$ such that, for a new $\psi \sim \Psi$ and all $\tau \in (0,1)$,

$$\mathbb{P}_{\psi \sim \Psi}\left(\Delta_\psi \le \hat{V}_m(\tau, \mathcal{D}) \middle| \mathcal{D}\right) \ge \tau - \varepsilon_m,$$

holds with high probability over the dataset $\mathcal{D}$ and $\lim_{m\to\infty} \varepsilon_m = 0$.

Note that $\hat{V}_m$ is indexed by both the prescribed quantile level τ and the dataset $\mathcal{D}$. Establishing our desired result requires controlling two sources of uncertainty: (i) *scenario uncertainty*, since we observe only finitely many scenarios $\{\psi_j\}_{j=1}^m \sim \Psi$, and (ii) *finite-sample uncertainty* for each scenario, since p_j and q_j are only observed through finite samples that produce $\hat{p}_j$ and $\hat{q}_j$. Our construction of $\hat{V}_m(\tau, \mathcal{D})$ explicitly accounts for both, and the main theorem will yield non-vacuous, finite-sample coverage guarantees. We summarize the notations introduced in Appendix A.2.

3. Theoretical Result

In this section, we construct $\hat{V}_m$, establish its theoretical guarantees, and conclude by discussing the benefits of distribution-level fidelity guarantees.

3.1. Methodology

The dataset $\mathcal{D}$ consists of m scenarios with finite sample estimates $\hat{p}_j, \hat{q}_j$ of the latent targets p_j, q_j. Our procedure has two steps.

1. For each j, define the pseudo-discrepancy and its cor-

[1] We assume the LLM setting is fixed during calibration so that variation in simulation reflects scenario differences rather than encoding drift, hyperparameter choices, etc.; choosing a different setting effectively defines a different simulator.

responding γ_j-confidence set $\mathcal{C}_j(\hat{p}_j, \gamma_j)$

$$\hat{\Delta}_j := \sup_{u \in \mathcal{C}_j(\hat{p}_j, \gamma_j)} L\big(u, \hat{q}_j\big),$$
$$\mathbb{P}\big(p_j \in \mathcal{C}_j(\hat{p}_j, \gamma_j) \mid \psi_j, n_j\big) \geq \gamma_j. \tag{1}$$

We will provide guidance on choosing the confidence level γ_j.

2. For $\tau \in (0, 1)$, let

$$\hat{V}_m(\tau) := \text{the } \tau\text{-quantile of} \{\hat{\Delta}_j\}_{j=1}^m. \tag{2}$$

We will use $\hat{V}_m(\tau)$ to construct a calibrated quantile curve for the discrepancy distribution, up to explicit finite-sample corrections. Note that the key design choice in our method is the confidence set $\mathcal{C}_j(\cdot)$ that quantifies the uncertainty of $\hat{p}_j$. This uncertainty set is what links per-scenario sampling error to distribution-level calibration. Next, we show how to construct confidence sets using concentration bounds and compute the respective $\hat{\Delta}_j$ efficiently. We, again, adopt our multinomial example setting.

Example 3.1 (Multinomial Confidence Set)**.** Suppose the outcomes are multinomial with parameter $p_j \in \Theta = \mathcal{P}^d := \{u \in [0,1]^d : \sum_{i=1}^d u_i = 1\}$, i.e., $Y_{j,i} \mid \psi_j \sim \text{Categorical}(p_j)$ and $\hat{p}_j$ are the empirical average vectors from n_j samples. Let $\text{KL}(\cdot\|\cdot)$ denote the KL divergence. Then the γ_j-confidence set for p_j and the pseudo-discrepancy $\hat{\Delta}_j$ for loss $L(u, \hat{q}_j) = \text{KL}(u\|\hat{q}_j)$ is

$$\hat{\Delta}_j = \sup_{u \in \mathcal{C}_j(\hat{p}_j, \gamma_j)} \text{KL}\big(u\|\hat{q}_j\big), \quad \text{where}$$
$$\mathcal{C}_j(\hat{p}_j, \gamma_j) := \Big\{u \in \mathcal{P}^d : \text{KL}(\hat{p}_j\|u) \leq \frac{d-1}{n_j} \ln\Big(\frac{2(d-1)}{1-\gamma_j}\Big)\Big\}.$$

The confidence set results from a variant of Chernoff Bound (Lemma B.5). With strictly positive $\hat{p}_j$ and $\hat{q}_j$, the maximizer is attained on the boundary $\text{KL}(\hat{p}_j\|u) = r_j$; computing this can be solved by standard constrained nonlinear solver.

More generally, p_j may be any functional of the conditional output distribution, or the distribution itself; see Appendix E.2 for additional examples. We now turn to the calibration guarantee for $\hat{V}_m$.

3.2. Calibrated Quantile Curve Theory

Our theoretical guarantee relies on the following two assumptions.

Assumption 3.2 (Independent data)**.** The scenarios $\psi_j \sim \Psi$ are i.i.d. Conditional on $(\psi_1, \ldots, \psi_m)$, the real and simulated sample batches are independent across scenarios, and within each scenario the real and simulated batches are independent. A new scenario and its simulator samples are independent of $\mathcal{D}$. The human sample sizes $\{n_j\}_{j=1}^m$ are treated as fixed deterministic design quantities.

Assumption 3.3 (Regular Discrepancy)**.** The discrepancy $L : \Theta \times \Theta \to [0, \infty)$ is jointly continuous on $\Theta \times \Theta$ and satisfies $L(u, u) = 0$ for all $u \in \Theta$.

Before stating our main result, we explain the intuition by considering the special case $\gamma_j \equiv \frac{1}{2}$ and considering the asymptotic regime $m \to \infty$, in which our guarantee (5) implies the following:

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi \leq \hat{V}_m\Big(\tfrac{1+\tau}{2}\Big)\Big|\mathcal{D}\Big) \geq \tau, \text{ for all quantile level } \tau \in (0,1). \tag{3}$$

In words, the τ-quantile of the discrepancy Δ_ψ is approximated by the empirical pseudo-quantile curve at an *inflated index* $(1+\tau)/2$. This inflation induces an inherent conservativeness: for $\tau \approx 0$, the index $(1+\tau)/2$ is at least the median of the pseudo-discrepancy distribution, leading to substantial inflation in the lower tail. To solve this, our methodology allows γ_j to be different, and we obtain:

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(\bar{\gamma}_m\tau + (1-\bar{\gamma}_m)\big)\Big|\mathcal{D}\Big) \geq \tau, \tag{4}$$

where $\bar{\gamma}_m = \frac{1}{m}\sum_{j=1}^m \gamma_j$. Now, to achieve τ-level coverage we evaluate the empirical pseudo-quantile at an index that is affine in τ with slope $\bar{\gamma}_m$. One would ideally want $\bar{\gamma}_m = 1$ to control the index inflation. However, increasing γ_j typically enlarges the set $\mathcal{C}_j$, and thus, increases the values of the pseudo-discrepancies $\hat{\Delta}_j$ and shifts the curve $\hat{V}_m$ upward. Therefore, one needs to balance index adjustment with increased conservativeness of the pseudo-quantile curve.

This trade-off motivates choosing γ_j adaptively as a function of the human sample size n_j. When n_j is large, p_j is typically more concentrated around $\hat{p}_j$; therefore, γ_j can be set high without significantly enlarging $\mathcal{C}_j$. Therefore, we set $\gamma_j = g(n_j)$ for an increasing function $g : \mathbb{R}_+ \to (0, 1)$. See Section 4 for numerical evidence illustrating the conservativeness of fixed γ and the benefits of adaptive coverage. We are now ready to state the main theorem rigorously.

Theorem 3.4. *Suppose Assumptions 3.2–3.3 hold and $\gamma_j \in (0,1)$ for each $j \in [m]$. Let $\mathcal{C}_j(\hat{p}_j, \gamma_j)$ denote any γ_j-confidence set, $\hat{\Delta}_j$ denote the pseudo-discrepancy in (1), and let $\hat{V}_m$ denote the quantile function of the pseudo-discrepancy in (2). Then, with probability at least $1-\delta$ over $\mathcal{D}$, simultaneously for all $\tau \in (0,1)$,*

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(\bar{\gamma}_m\tau + (1-\bar{\gamma}_m)\big)\Big|\mathcal{D}\Big) \geq \tau - \varepsilon_m(\delta), \tag{5}$$

where $\bar{\gamma}_m := \frac{1}{m}\sum_{j=1}^m \gamma_j$ and $\varepsilon_m(\delta) := \sqrt{\frac{\log(6/\delta)}{2m}} + \frac{1}{m} + \frac{1}{\bar{\gamma}_m}\sqrt{\frac{2\log(3m/\delta)}{m}}$. In particular, $\varepsilon_m(\delta) \to 0$ as $m \to \infty$.

Unlike standard conformal inference, which targets future observable outcomes, Theorem 3.4 provides uncertainty

quantification for latent scenario-level population quantities, such as conditional output distributions or their functionals. Thus the guarantee is distributional across scenarios, rather than a prediction guarantee for a single future draw. We next provide a proof sketch.

Proof Sketch: Suppose we have access to the true p_j for each scenario j. Then $\{\Delta_j\}_{j=1}^m$ are i.i.d. , and the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality yields a uniform deviation bound for the empirical CDF of discrepancy, which can be translated into upper bounds on its quantile function. In our setting, however, (i) p_j is latent and (ii) each scenario is observed with a heterogeneous sample size n_j. Therefore, we aim to use empirical quantiles of the observable proxy $\{\hat{\Delta}_j\}$ to approximate the quantile function of the unobservable $\{\Delta_j\}$. For simplicity, we fix $\gamma_j \equiv \gamma$. The key steps are:

(i) Let x_τ denote the population τ-quantile of Δ_j, and define the corresponding hardest scenarios $S_\tau := \{j : \Delta_j \geq x_\tau\}$. By definition, $\frac{|S_\tau|}{m} \approx 1 - \tau$.

(ii) By construction, $\hat{\Delta}_j \geq \Delta_j$ holds with probability γ for each j. Consequently, concentration implies that, with high probability, a γ-fraction of the indices in S_τ, satisfy $\hat{\Delta}_j \geq \Delta_j \geq x_\tau$.

(iii) Hence, at least a $\gamma(1-\tau)$-fraction of the pseudo-discrepancies exceed x_τ. In other words, the τ-quantile of Δ_j is dominated by the $\big(1-\gamma(1-\tau)\big)$-quantile of $\hat{\Delta}_j$.

Therefore, for any quantile level τ, taking the $\big(\gamma\tau+(1-\gamma)\big)$-quantile of the pseudo-discrepancies yields an upper bound on the τ-quantile of the oracle discrepancies. This is why the quantile index must shift: to upper bound the oracle τ-quantile of Δ, we need to use a more conservative pseudo-quantile level so that the resulting pseudo-threshold is high enough to cover the target quantile despite the miscoverage. This index adjustment is the first source of conservativeness; the second source is the DKW deviation that links empirical and population CDFs. Combining these two ingredients yields the theorem. We prove a sharper theorem in Appendix C.1. The proof of Theorem 3.4 then follows as a clean-index consequence of that sharper result and is provided in Appendix C.2.

□

Remark 3.5. A subtle modeling choice is that Theorem 3.4 actually characterizes the distribution of $\Delta^{(k)} = L(p, \hat{q}^{(k)})$, where $\hat{q}^{(k)}$ is estimated using k simulator samples, instead of the actual simulator parameter q. Conceptually, $\Delta^{(k)}$ is the correct target when one wants to characterize performance for a fixed query budget k. In contrast, Δ is more natural when simulator queries are cheap, and the goal also includes quantifying the simulator's inherent bias. For brevity, we ignored the k superscript in our main text. A discussion to the latter case is presented in Corollary C.3.

3.3. Benefits of Distribution-Level Characterization

We now turn to the benefits of obtaining a distribution-level characterization of sim-to-real discrepancy. For illustrative purposes we set $\bar{\gamma}_m = \frac{1}{2}$ and define the index–adjusted (calibrated) curve $\hat{V}_m^{\mathrm{cal}}(\tau) := \hat{V}_m\left(\frac{1+\tau}{2}\right)$, for all $\tau \in [0,1]$, which matches the coverage guarantee in Equation 3 in asymptotic regime. With the calibrated quantile, for an unseen scenario ψ we can derive a confidence set for p_ψ given a simulated $\hat{q}_\psi$. For any confidence level $\bar{\alpha} \in (0,1)$, define

$$S_{\bar{\alpha}} := \big\{u \in \Theta : L(u, \hat{q}_\psi) \leq \tau_{\bar{\alpha}}\big\}, \quad \tau_{\bar{\alpha}} := \hat{V}_m\big(1 - \tfrac{\bar{\alpha}}{2}\big)\,.$$

By Theorem 3.4, $p_\psi \in S_{\bar{\alpha}}$ with probability at least $1-\bar{\alpha}$ up to an $o_m(1)$ remainder.

In addition, the calibrated quantile curve can be compressed to provide summary statistics and is therefore more informative than any single risk summary. Returning to the motivating example, a media research firm can use a calibrated AUC to assess average simulator bias, while a risk-averse product team may track a calibrated CVaR_α to bound rare but consequential misreads of public sentiment. We can summarize overall average sim-to-real discrepancy and tail risk CVaR for any $\alpha \in (0,1)$ by,

$$\mathrm{AUC}_{\mathrm{cal}} := \int_0^1 \hat{V}_m^{\mathrm{cal}}(\tau)\, d\tau,$$
$$\mathrm{CVaR}_\alpha^{\mathrm{cal}} := \frac{1}{\alpha}\int_{1-\alpha}^1 \hat{V}_m^{\mathrm{cal}}(u)\, du.$$

In words, $\mathrm{AUC}_{\mathrm{cal}}$ aggregates the calibrated curve into a single average-bias summary, while $\mathrm{CVaR}_\alpha^{\mathrm{cal}}$ averages over the worst α fraction of scenarios under the same index-adjusted quantile curve.

4. Application: LLM Fidelity Profiling

4.1. Dataset and Methodology

We evaluate our procedure on the *WorldValueBench* dataset, curated by (Zhao et al., 2024) and constructed from the World Values Survey (Haerpfer et al., 2020). The dataset comprises survey questions asked across 64 countries, covering 12 thematic categories (e.g., social values, security, migration). Individual-level covariates are available for each respondent, including gender, age, migration status, education, marital status, and related demographics.

Preprocessing. After data cleaning, we retain 235 distinct questions and responses from 96,220 individuals. Each

Q199. How interested would you say you are in politics? Are you
1 Very interested
2 Somewhat interested
3 Not very interested
4 Not at all interested

Figure 1. Example of World Value Question. Source: (Haerpfer et al., 2020).

question offers a categorical set of possible answers. Examples of the questions are provided in Figure 1, and more in Appendix D. To place heterogeneous answer sets on a common scale, we map each question's categories to the interval $[-1, 1]$, producing a bounded real-valued outcome for every scenario and respecting the natural ordering of response options when available (e.g., disagree–neutral–agree). Individual-level covariates are then used to construct synthetic profiles and prompts for simulators. Details on preprocessing and the dataset are provided in Appendix D.

Methodology We generate synthetic responses and estimate $\{\hat{q}_j\}_{j=1}^{235}$ for four LLMs: GPT-4o (`gpt-4o`), GPT-5 MINI (`gpt-5-mini`), LLAMA 3.3 70B (`Llama-3.3-70B-Instruct-Turbo`), and QWEN 3 235B (`Qwen3 235B A22B Thinking 2507 FP8`). For each LLM, we simulate 200 responses per question to estimate $\hat{q}_j$. On the human side, we subsample 500 respondents per question to compute $\{\hat{p}_j\}_{j=1}^{235}$ and construct the confidence sets $\mathcal{C}_j$; due to nonresponses, the effective sample sizes n_j vary across questions, typically between 450 and 500.

We include two baseline benchmarks: a *uniform* baseline and a K-nearest-neighbor (KNN) predictor that uses OpenAI's publicly available embedding API (`text-embedding-3-small`) to embed questions.[2]. The uniform baseline samples uniformly from the available answers for each question. The KNN predictor estimates $\hat{q}_j$ from semantically similar questions: we use five-fold cross-validation over the 235 questions, and within each training split we tune the KNN hyperparameter and predict each held-out question's $\hat{q}_j$ using the human-side estimates $\{\hat{p}_\ell\}$ from its nearest training questions in embedding space. These benchmarks also illustrate the generality of our framework: as long as a method produces a simulator estimate $\hat{q}_j$, the same quantile-curve construction applies.

Since the outcome is bounded in $[-1, 1]$, we construct per-scenario confidence sets $\mathcal{C}_j$ using Example E.1. We set $\gamma_j = 1 - n_j^{-1/3}$, which is increasing in sample size n_j,[3] yielding an average coverage level of $\bar{\gamma}_m \approx 0.87$. We use squared loss $L(p, q) = (p - q)^2$. Applying the procedure in Section 3 (see (1), (2)) to each candidate simulator yields a calibrated fidelity profile, which we plot using the asymptotic index adjustment $\tau \mapsto \bar{\gamma}_m \tau + (1 - \bar{\gamma}_m)$ in (4). For completeness, we also apply the same workflow to different datasets, including Bernoulli setting (EEDI), multinomial setting (OpinionQA) and continuous temperature prediction setting (UMAR dataset from NEST); see Appendix F. The full dataset, code, and reproduction materials are available in an anonymized repository for review at: `https://anonymous.4open.science/r/simulator_fidelity-E3F4/`.

[2]This API is no longer free to access, but it was at the time of writing.

[3]We adopt a generic form $\gamma_j = 1 - n_j^{-\beta}$ with $\beta = 1/3$, and conduct a sensitivity analysis on β in Appendix D.5.

4.2. Fidelity Profiling

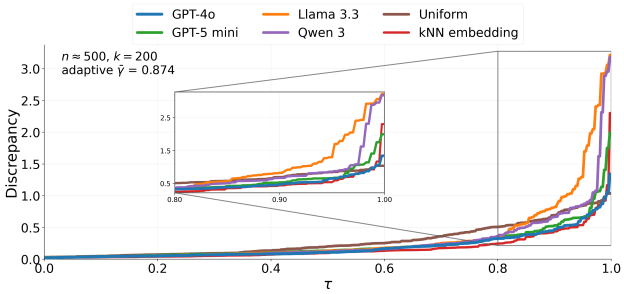


Figure 2. Calibrated discrepancy quantile curves $\hat{V}(\tau)$ across LLM simulators with zoom-in on upper 80-100 quantile.

Figure 2 plots the calibrated discrepancy quantile curves. Lower, flatter curves indicate smaller scenario-level sim-to-real gaps, while sharp elbows reveal rare but severe failures. Among the generative LLM simulators, GPT-4o has the most stable profile, with GPT-5-MINI close behind; LLAMA 3.3 and QWEN-3 show substantially larger upper-tail discrepancies. All simulators improve over the uniform benchmark on over 80% of questions, but fail to capture tail scenarios. We conducted robustness checks under different human sample sizes n_j in Section D.4, and this pattern is stable.

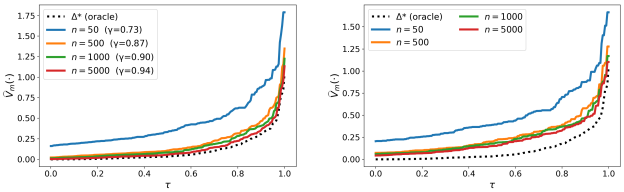
The two baselines serve as diagnostics rather than competing black-box simulators. The uniform baseline is not intended to be a poor estimator, since for near-balanced questions a uniform response distribution can have a small sim-to-real gap. Its role is to provide a structure-free benchmark without using any semantic information about the question. The embedding-based KNN baseline plays a different diagnostic role: by predicting each held-out question from human response profiles of semantically nearby questions, it tests how much scenario-level response structure is captured by question embeddings alone. Its strong performance over most quantiles suggests that semantic similarity carries substantial information about human response distributions. Thus, the KNN result indicates semantic recoverability: question embeddings capture substantial scenario-level distributional structure, and our quantile profile assesses whether genera-

tive simulators recover it.

5. Tightness Analysis of Adjusted Quantiles

A natural concern with our approach is the tightness of our calibrated quantile curve. Since our method uses off-the-shelf concentration bounds and worst-case pseudo-discrepancies, the calibrated curve may be conservative. We assess this empirically using the large human sample in WorldValueBench. For each question, we treat the response distribution estimated from all available respondents, over 90,000, as a proxy for the true p_j, which gives an oracle discrepancy $\Delta_j^\star := L(p_j, \hat{q}_j)$ and hence an oracle quantile curve. We then fix the simulator budget at $k = 200$, subsample $n_j \in \{50, 500, 1000, 5000\}$ respondents per question, recompute the confidence sets and pseudo-discrepancies, and compare the calibrated curves with the oracle curve. The results plotted in Figure 3a indicate that the calibrated envelope is substantially conservative when $n_j = 50$, but becomes tighter when $n_j \geq 500$. This behavior is consistent with our construction: larger n_j yields concentrated estimates $\hat{p}_j$ and smaller confidence sets $\mathcal{C}_j(\hat{p}_j, \gamma_j)$, which reduces the inflation induced by the supremum defining $\hat{\Delta}_j$.

Figure 3a uses the adaptive choice $\gamma_j = 1 - n_j^{-1/3}$. To isolate the role of this choice, as we discussed in Section 3, we repeat the experiment with fixed coverage $\gamma_j \equiv 1/2$; the results are shown in Figure 3b. Unlike the adaptive case, the calibrated curves remain separated from the oracle curve even for large n_j. This persistent gap comes from the index adjustment. When $\bar{\gamma}_m = 1/2$, the asymptotic calibration evaluates the pseudo-quantile at $\hat{V}_m(0.5\tau + 0.5)$. Thus, even near the lower tail ($\tau \approx 0$), the procedure uses the median pseudo-discrepancy, creating conservativeness that does not vanish as n_j grows. This highlights the value of choosing adaptive coverage levels that changes with n_j.



(a) *(b)*

Figure 3. Tightness analysis under GPT-4o. The dotted black curve is the oracle quantile curve for $\Delta^\star$; colored curves are computed with $n_j \in \{50, 500, 1000, 5000\}$, and the adaptive average coverage levels are $\bar{\gamma}_m \approx 0.73, 0.87, 0.90$, and 0.94, respectively. Panel (b) uses fixed $\gamma_j = 1/2$.

While WorldValueBench provides a large pool of respondents that facilitates an oracle tightness check, most applications do not enjoy such extensive ground-truth data. We next develop a complementary tightness assessment: a confidence band that sandwiches the oracle quantile curve.

Theorem 5.1 (Informal Version)**.** *Suppose the setup of Theorem 3.4 and Assumptions 3.2–3.3 hold. Let $\{\gamma_{L,j}, \gamma_{U,j}\}_{j=1}^m \subset (0, 1]$ be coverage levels and define the lower and upper pseudo-discrepancies*

$$\hat{\Delta}_j^- := \inf_{u \in \mathcal{C}_j(\hat{p}_j, \gamma_{L,j})} L(u, \hat{q}_j), \; \hat{\Delta}_j^+ := \sup_{u \in \mathcal{C}_j(\hat{p}_j, \gamma_{U,j})} L(u, \hat{q}_j),$$

where $\mathcal{C}_j(\hat{p}_j, \gamma) \subset \Theta$ are γ-confidence sets. Let $\hat{V}_m^-(\cdot)$ (resp. $\hat{V}_m^+(\cdot)$) denote the empirical quantile function of $\{\hat{\Delta}_j^-\}_{j=1}^m$ (resp. $\{\hat{\Delta}_j^+\}_{j=1}^m$). Let $V : [0, 1] \to \mathbb{R}$ denote the quantile function of $\Delta(\psi)$. Then, as $m \to \infty$, for any fixed $\tau \in (0, 1)$,

$$\hat{V}_m^-\big(\bar{\gamma}_{L,m}\tau\big) \lesssim V(\tau) \lesssim \hat{V}_m^+\big(\bar{\gamma}_{U,m}\tau + (1 - \bar{\gamma}_{U,m})\big), \quad (6)$$

where $\bar{\gamma}_{L,m} = \frac{1}{m}\sum_{j=1}^m \gamma_{L,j}$, and $\bar{\gamma}_{U,m} = \frac{1}{m}\sum_{j=1}^m \gamma_{U,j}$.

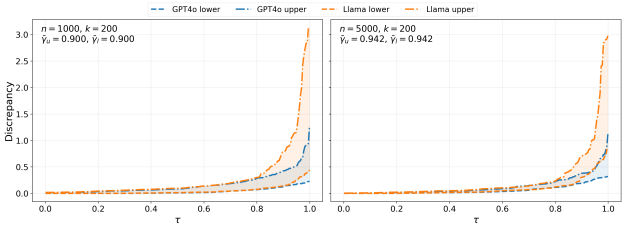


Figure 4. Confidence bands for GPT-4o and LLAMA 3.3. The left panel uses $n_j = 1000$, and $\bar{\gamma}_L = \bar{\gamma}_U = 0.9$; the right panel uses $n_j = 5000$, and $\bar{\gamma}_L = \bar{\gamma}_U = 0.94$, both with $k = 200$.

The formal finite-sample statement is given in Theorem C.4. Together, Theorems 3.4 and C.4 give complementary distributional fidelity guarantees: Theorem 3.4 provides a one-sided calibrated upper envelope for the population quantile curve $V(\tau)$, while Theorem C.4 adds a lower envelope, yielding a confidence band for $V(\tau)$. Figure 4 illustrates these bands for $n_j \in \{1000, 5000\}$, comparing GPT-4O with LLAMA 3.3. Across both sample sizes, the GPT-4O band lies below the LLAMA 3.3 band, especially in the upper tail. This suggests that GPT-4O has smaller scenario-level discrepancies and smaller extreme failures, and the conclusion is stable as n_j increases. Importantly, the comparison is fully data-driven: the band construction does not require access to ground-truth discrepancies and can be carried out using only the observed heterogeneous batches.

6. Discussion

We develop a framework for profiling simulator fidelity through the quantile function of the sim-to-real discrepancy. Our approach makes no parametric assumptions about both the simulator/real systems, accounts for finite and heterogeneous sampling across scenarios, and yields finite-sample guarantees at any target quantile level. The resulting fidelity profile supports a range of summary statistics, from

average discrepancy to tail-risk measures that capture rare but critical failures. Finally, we demonstrate the practical value of the framework on real survey data, evaluating multiple LLM-based simulators and illustrating distribution-level differences in their alignment with human populations.

Despite its broad applicability, several extensions remain. The current framework targets marginal scenario-level population quantities, not respondent-level dependence across questions. It also assumes static i.i.d. scenarios and does not yet cover dynamic simulators with temporal dependence or distribution shift among scenarios. Moreover, generic concentration and DKW-type bounds can be conservative when the number of scenarios or per-scenario samples is small. Finally, on the computational side, the main bottleneck is simulator-side generation: once real and simulated batches are available, our methodology is computationally lightweight, but expensive simulators may make sample generation itself the limiting factor.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Aher, G., Arriaga, R. I., and Kalai, A. T. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31 (3):337–351, 2023. doi: 10.1017/pan.2023.2.

Bardenet, R. and Maillard, O.-A. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361 – 1385, 2015. doi: 10.3150/14-BEJ605. URL https://doi.org/10.3150/14-BEJ605.

Barton, R. R., Nelson, B. L., and Xie, W. Quantifying input uncertainty via simulation confidence intervals. *INFORMS Journal on Computing*, 26(1):74–87, 2014. doi: 10.1287/ijoc.2013.0548.

Bates, S., Angelopoulos, A., Lei, L., Malik, J., and Jordan, M. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

Budde, C. E., Hartmanns, A., Meggendorfer, T., Weininger, M., and Wienhöft, P. Statistical model checking beyond means: Quantiles, cvar, and the dkw inequality (extended version). *arXiv preprint arXiv:2509.11859*, 2025.

Cao, Y., Liu, H., Arora, A., Augenstein, I., Röttger, P., and Hershcovich, D. Specializing large language models to simulate survey response distributions for global populations. arXiv preprint, 2025. URL https://arxiv.org/abs/2502.07068.

Chen, M., Lam, H., and Liu, Z. Quantifying distributional input uncertainty via inflated kolmogorov-smirnov confidence band. *arXiv preprint arXiv:2403.09877*, 2024.

Cranmer, K., Brehmer, J., and Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117.

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

Gao, Y., Lee, D., Burtch, G., and Fazelpour, S. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025. doi: 10.1073/pnas.2501660122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2501660122.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., Puranen, B., et al. World values survey: Round seven – country-pooled datafile (2017–2020), 2020. URL https://doi.org/10.14281/18241.1.

He-Yueya, J., Ma, W. A., Gandhi, K., Domingue, B. W., Brunskill, E., and Goodman, N. D. Psychometric alignment: Capturing human knowledge distributions via language models. *arXiv preprint arXiv:2407.15645*, 2024.

Heer, P., Derungs, C., Huber, B., Bünning, F., Fricker, R., Stoller, S., and Niesen, B. Comprehensive energy demand and usage data for building automation. *Scientific Data*, 11(1):469, 2024. doi: 10.1038/s41597-024-03292-2. URL https://doi.org/10.1038/s41597-024-03292-2.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Huang, C., Wu, Y., and Wang, K. Uncertainty quantification for LLM-based survey simulations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=nY1Ge2wxtP.

Jeon, Y., Chu, Y., Pasupathy, R., and Shashaani, S. Uncertainty quantification using simulation output: Batching as an inferential device, 2024. URL https://arxiv.org/abs/2311.04159.

L.A., P. and Bhat, S. P. A wasserstein distance approach for concentration of empirical risk estimates. *Journal of Machine Learning Research*, 23(238):1–61, 2022. URL http://jmlr.org/papers/v23/20-965.html.

Lam, H. Cheap bootstrap for input uncertainty quantification. In *Proceedings of the 2022 Winter Simulation Conference*. IEEE, 2022. doi: 10.1109/WSC57314.2022.10015450.

Lu, Y., Huang, J., Han, Y., Yao, B., Bei, S., Gesi, J., Xie, Y., He, Q., Wang, D., et al. Prompting is not all you need! evaluating llm agent simulation methodologies with real-world online customer behavior data. *arXiv preprint arXiv:2503.20749*, 2025.

Macal, C. Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, 10: 144–156, 05 2016. doi: 10.1057/jos.2016.7.

Mardia, J., Jiao, J., Tánczos, E., Nowak, R. D., and Weissman, T. Concentration inequalities for the empirical distribution of discrete distributions: beyond the method of types. *Information and Inference: A Journal of the IMA*, 9(4):813–850, 11 2019. ISSN 2049-8772. doi: 10.1093/imaiai/iaz025. URL https://doi.org/10.1093/imaiai/iaz025.

Massart, P. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Papamakarios, G., Sterratt, D., and Murray, I. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 837–848. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/papamakarios19a.html.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Peng, X. B., Andrychowicz, M., Zaremba, W., and Abbeel, P. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 3803–3810. IEEE, 2018.

Richner, P., Heer, P., Largo, R., Marchesi, E., and Zimmermann, E. Nest – a platform for the acceleration of innovation in buildings. *Informes de la Construcción*, 70 (550):e255, 2018. doi: 10.3989/ic.17.062. pp. 222.

Roy, C. J. and Oberkampf, W. L. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25):2131–2144, 2011. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2011.03.016. URL https://www.sciencedirect.com/science/article/pii/S0045782511001290.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. Mmd aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023. URL http://jmlr.org/papers/v24/21-1289.html.

Snell, J. C., Zollo, T. P., Deng, Z., Pitassi, T., and Zemel, R. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. *arXiv preprint arXiv:2212.13629*, 2022.

Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

Wang, Z., Lamb, A., Saveliev, E., Cameron, P., Zaykov, J., Hernandez-Lobato, J. M., Turner, R. E., Baraniuk, R. G., Craig Barton, E., Peyton Jones, S., Woodhead, S., and Zhang, C. Results and insights from diagnostic questions: The neurips 2020 education challenge. In Escalante, H. J. and Hofmann, K. (eds.), *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine*

Learning Research, pp. 191–205. PMLR, 06–12 Dec 2021. URL https://proceedings.mlr.press/v133/wang21a.html.

Westwood, S. J. The potential existential threat of large language models to online survey research. *Proceedings of the National Academy of Sciences*, 122(47):e2518075122, 2025. doi: 10.1073/pnas.2518075122. Epub 2025-11-20.

Zhang, L., Zhou, L., Ren, L., and Laili, Y. Modeling and simulation in intelligent manufacturing. *Computers in Industry*, 112:103123, 2019. ISSN 0166-3615. doi: https://doi.org/10.1016/j.compind.2019.08.004. URL https://www.sciencedirect.com/science/article/pii/S0166361519303239.

Zhao, W., Mondal, D., Tandon, N., Dillion, D., Gray, K., and Gu, Y. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models, 2024. URL https://arxiv.org/abs/2404.16308.

A. Setup Discussions

A.1. Profile-Level Interpretation of the Survey Application

The main text works with the scenario-level conditional output distributions $Q^{rw}(\cdot \mid \psi)$ and $Q^{sim}(\cdot \mid \psi)$. In the survey application, these distributions can be obtained by marginalizing over respondent profiles. To include the respondent profile into the notation, let the real-world system be characterized by a latent profile $z \in \mathcal{Z}$ distributed according to the population distribution $\mathcal{P}$. For each pair (ψ, z), the real-world system produces an outcome Y^{rw} with conditional distribution $Q^{rw}(\cdot \mid z, \psi)$. The scenario-level real output distribution used in the main text is the population mixture

$$Q^{rw}(A \mid \psi) := \mathbb{E}_{z\sim P}\left[Q^{rw}(A \mid z, \psi)\right],$$

for any measurable output set A.

In the categorical survey example, if $Q^{rw}(\cdot \mid z, \psi) = \mathrm{Categorical}(\Pi^{rw}(\psi, z))$, then the population response vector is

$$p_\psi = \mathbb{E}_{z\sim P}\left[\Pi^{rw}(\psi, z)\right].$$

For scalar responses, p_ψ may instead denote the corresponding marginal mean or another population-level functional of $Q^{rw}(\cdot \mid \psi)$.

Similarly, the simulator draws a synthetic profile $z^{sim} \in \mathcal{Z}^{sim}$ from a simulator population distribution P^{sim} and generates an outcome according to $Q^{sim}(\cdot \mid z^{sim}, \psi)$. The scenario-level simulator distribution is

$$Q^{sim}(A \mid \psi) := \mathbb{E}_{z^{sim}\sim P^{sim}}\left[Q^{sim}(A \mid z^{sim}, \psi)\right].$$

In the categorical case,

$$q_\psi = \mathbb{E}_{z^{sim}\sim P^{sim}}\left[\Pi^{sim}(\psi, z^{sim})\right],$$

with analogous definitions for means, functionals, or full distributions.

A.2. Table of Notation

Table 2. Notation summary.

Symbol	Meaning
$\psi \sim \Psi$	Scenario drawn from the scenario population.
$z \sim \mathcal{P}$	Real-world respondent drawn from the population distribution.
$p(\psi)$(or p_ψ)	Real-world population parameter for scenario ψ.
$q(\psi)$(or q_ψ)	Simulator-side population parameter for scenario ψ under fixed simulator settings.
$\hat{p}_\psi, \hat{q}_\psi$	Finite-sample estimators of p_ψ and q_ψ.
n_j	Human sample size for scenario j.
k (or k_j)	Simulator sample size for scenario j.
$L(\cdot,\cdot)$	Discrepancy function between real and simulated distribution.
Δ_ψ	True sim-to-real discrepancy for scenario ψ.
γ_j	Per-scenario confidence-set coverage level.
$\bar{\gamma}_m$	Average coverage level across m scenarios.
$\mathcal{C}_j(\hat{p}_j, \gamma_j)$	γ_j-confidence set for the latent parameter p_j.
$\hat{\Delta}_j$	Pseudo-discrepancy for scenario j, defined via a confidence set around $\hat{p}_j$.
$\hat{V}_m(\tau)$	Pseudo-discrepancy quantile curve built from m scenarios, indexed by quantile level τ.
τ	Quantile-function argument, counted in the usual direction from lower to upper quantiles.
α	Upper-tail fraction / exceedance level used when discussing tail scenarios.

B. Technical Lemmas

Lemma B.1. *(Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality via (Massart, 1990))*

Let $X_1, X_2, \ldots, X_n$ be i.i.d. real-valued random variables with cumulative distribution function (CDF) F^*, and let $\hat{F}_n$ be the empirical distribution function defined by

$$\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le x\}.$$

Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x) - F^*(x)\right| > \varepsilon\right) \le 2e^{-2n\varepsilon^2}.$$

Equivalently, for any confidence level $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{x\in\mathbb{R}}\left|\hat{F}_n(x) - F^*(x)\right| \le \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}.$$

Lemma B.2. *[Hoeffding (additive) via (Hoeffding, 1963)]*

Let $Z_1, \ldots, Z_n \in [0,1]$ be independent, $T = \sum_{i=1}^n Z_i$, and $\mu = \mathbb{E}[T]$. For any $t \in [0,\mu]$,

$$\mathbb{P}\big(T \le \mu - t\big) \le \exp\left(-\frac{2t^2}{\sum_{i=1}^n (1-0)^2}\right) = \exp\left(-\frac{2t^2}{n}\right).$$

Lemma B.3. *[Chernoff–Hoeffding for one-parameter exponential family]*

Let $X_1, \ldots, X_n$ be i.i.d. with density (or mass) in the one-parameter canonical exponential family

$$p_\theta(x) \;=\; \exp\{\theta T(x) - A(\theta)\}\, h(x), \qquad \theta \in \Theta,$$

where $T(x)$ is the sufficient statistic, $A(\theta)$ is the log-partition function (convex, differentiable on Θ), and the mean map is $\mu(\theta) := \mathbb{E}_\theta[T(X)] = A'(\theta)$. Assume Θ is an open interval and all quantities below are finite.

Define the empirical mean of the sufficient statistic

$$\bar{T}_n \;=\; \frac{1}{n}\sum_{i=1}^n T(X_i).$$

For each t in the range of $\bar{T}_n$ let θ_t be the (unique) canonical parameter satisfying $\mu(\theta_t) = t$. Then for any $\varepsilon > 0$,

$$\mathbb{P}\big(D(p_{\theta_{\bar{T}_n}} \| p_\theta) > \varepsilon\big) \le 2e^{-n\varepsilon}$$

Proof:

Define the shifted log-MGF under p_θ,

$$\psi_\theta(\lambda) := \log \mathbb{E}_\theta\big[e^{\lambda T(X)}\big] = A(\theta+\lambda) - A(\theta),$$

where the displayed equality follows from the exponential-family form (for λ in the domain where the expectation is finite). For an attainable mean m denote θ_m as the unique solution of $\mu(\theta_m) = m$.

By adopting the one-sided Chernoff bound, for any real λ such that expectations exist and any $m \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}_\theta(\bar{T}_n \ge m) &= \mathbb{P}\big(e^{\lambda n \bar{T}_n} \ge e^{\lambda n m}\big)\\ &\le e^{-\lambda n m}\, \mathbb{E}_\theta\big[e^{\lambda n \bar{T}_n}\big]\\ &= \exp\big(-n(\lambda m - \psi_\theta(\lambda))\big).\end{aligned}$$

Optimizing over λ gives the Chernoff bound

$$\mathbb{P}_\theta(\bar{T}_n \ge m) \le \exp\big(-n\psi_\theta^*(m)\big), \implies \psi_\theta^*(m) := \sup_{\lambda\in\mathbb{R}}\{\lambda m - \psi_\theta(\lambda)\},$$

where $\psi_\theta^*(m)$ is the Fenchel-Legendre transform of log-MGF. A symmetric argument with $\lambda < 0$ yields the lower-tail bound

$$\mathbb{P}_\theta(\bar{T}_n \le m) \le \exp\big(-n\psi_\theta^*(m)\big).$$

We next link the Fenchel-Legendre transform to KL-Divergence using exponential tilting. First, adopting change of variables $\eta = \theta + \lambda$. Then

$$\begin{aligned}\psi_\theta^*(m) &= \sup_\eta \{\langle \eta - \theta, m\rangle - (A(\eta) - A(\theta))\} \\ &= A(\theta) + A^*(m) - \langle \theta, m\rangle,\end{aligned}$$

where $A^*(m) = \sup_\eta\{\langle \eta, m\rangle - A(\eta)\}$ is the convex conjugate of A. When m is attainable, the supremum is achieved at $\eta = \theta_m$, and therefore

$$\psi_\theta^*(m) = \langle \theta_m - \theta, m\rangle - (A(\theta_m) - A(\theta)).$$

But for exponential-family densities one has the following direct algebraic identity for the KL:

$$\begin{aligned}D\big(p_{\theta_1}\|p_{\theta_2}\big) &= \mathbb{E}_{\theta_1}\Big[\log \frac{p_{\theta_1}(X)}{p_{\theta_2}(X)}\Big] \\ &= \mathbb{E}_{\theta_1}\big[(\theta_1 - \theta_2)T(X) - (A(\theta_1) - A(\theta_2))\big] \\ &= (\theta_1 - \theta_2)\,\mathbb{E}_{\theta_1}[T(X)] - \big(A(\theta_1) - A(\theta_2)\big).\end{aligned}$$

Taking $\theta_1 = \theta_m$ and $\theta_2 = \theta$ (so $\mathbb{E}_{\theta_1}[T] = m$) yields

$$\psi_\theta^*(m) = D\big(p_{\theta_m}\|p_\theta\big).$$

Combining this with the one-sided Chernoff-bound above yields the one-sided KL-form Chernoff bounds

$$\begin{aligned}\mathbb{P}_\theta(\bar{T}_n \geq m) &\leq e^{-nD\left(p_{\theta_m}\|p_\theta\right)}, \\ \mathbb{P}_\theta(\bar{T}_n \leq m) &\leq e^{-nD\left(p_{\theta_m}\|p_\theta\right)}.\end{aligned}$$

Fix $\varepsilon > 0$. Because $A'' > 0$ the function $m \mapsto D(p_{\theta_m}\|p_\theta)$ is continuous, strictly convex and has a unique minimum 0 at $m = A'(\theta)$. Thus the sublevel set $\{m : D(p_{\theta_m}\|p_\theta) < \varepsilon\}$ is an open interval (m_-, m_+); equivalently

$$\{m : D(p_{\theta_m}\|p_\theta) \geq \varepsilon\} = (-\infty, m_-] \cup [m_+, \infty).$$

Hence

$$\{D(p_{\theta_{\overline{T}_n}}\|p_\theta) \geq \varepsilon\} \subseteq \{\overline{T}_n \leq m_-\} \cup \{\overline{T}_n \geq m_+\}.$$

Applying the one-sided KL bounds at $m_\pm$ (each equals ε) and using the union bound gives

$$\mathbb{P}\big(D(p_{\theta_{\overline{T}_n}}\|p_\theta) \geq \varepsilon\big) \leq e^{-n\varepsilon} + e^{-n\varepsilon} = 2e^{-n\varepsilon},$$

which proves the lemma.

Lemma B.4. *(Chernoff-Hoeffding Inequality)*

Let $X_1, \ldots, X_n \sim Ber(\tilde{p})$ be i.i.d. Bernoulli random variables with unknown mean $\tilde{p}$, and define the empirical mean as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i.$$

Then for any $\varepsilon > 0$,

$$\mathbb{P}\left(D\left(\bar{X}_n \,\|\, \tilde{p}\right) > \varepsilon\right) \leq 2e^{-n\varepsilon},$$

where $D(p \,\|\, q)$ is the Kullback-Leibler divergence between Bernoulli distributions with parameters p and q.

Proof: Via Lemma B.3.

Lemma B.5 (Multinomial Chernoff-Hoeffding Bound via (Mardia et al., 2019))**.** *For all $d \leq (\frac{nC_0}{4})^{\frac{1}{3}}$ and $P \in \mathcal{M}_k$, the following holds with the universal constants $C_0 = \frac{e^3}{2\pi} \approx 3.1967$, for any $\epsilon > 0$,*

$$\Pr\big(D(\widehat{P}_{n,d} \,\|\, P) \geq \epsilon\big) \leq 2(d-1)e^{-\frac{n\epsilon}{d-1}}.$$

Lemma B.6 (Wasserstein distance bound via (L.A. & Bhat, 2022))**.** *Let X be a sub-Gaussian r.v. with parameter σ. Let F denote the CDF of X. Then, for every $n \geq 1$ and ε such that $\frac{512\sigma}{\sqrt{n}} < \varepsilon < \frac{512\sigma}{\sqrt{n}} + 16\sigma\sqrt{e}$, we have*

$$\mathbb{P}\big(W_1(F_n, F) > \varepsilon\big) \;\leq\; \exp\!\left(-\frac{n}{256\sigma^2 e}\left(\varepsilon - \tfrac{512\sigma}{\sqrt{n}}\right)^2\right),$$

where e is Euler's number.

Lemma B.7 (Order-statistic thresholding)**.** *Let $x_1, \ldots, x_m \in \mathbb{R}$ and let $x_{(1)} \leq \cdots \leq x_{(m)}$ be their order statistics. Fix a threshold $\tau \in \mathbb{R}$ and an integer $N \in \{1, \ldots, m\}$. If at least N of the sample values are at least τ, i.e. $\big|\{j : x_j \geq \tau\}\big| \geq N$, then*

$$x_{(m-N+1)} \;\geq\; \tau.$$

(Equivalently, if at least N of the x_j are strictly larger than τ, the same conclusion holds.)

Proof:

Suppose, for contradiction, that $x_{(m-N+1)} < \tau$. Then all of the first $m - N + 1$ order statistics are strictly less than τ, so there are at most $m - (m - N + 1) = N - 1$ indices j with $x_j \geq \tau$, contradicting $\big|\{j : x_j \geq \tau\}\big| \geq N$. Hence $x_{(m-N+1)} \geq \tau$.

C. Theorems Proof

Theorem C.1. *Suppose Assumptions 3.2–3.3 hold and $\gamma_j \in (0,1)$ for each $j \in [m]$. Let $\mathcal{C}_j(\hat{p}_j, \gamma_j)$ denote any γ_j-confidence set, $\hat{\Delta}_j$ denote the pseudo-discrepancy (see (1)), and let $\hat{V}_m$ denote the quantile function of the pseudo-discrepancy (see (2)). For $\tau \in (0,1)$, let $[1-\tau]_m := m^{-1}\lceil m(1-\tau)\rceil$ denote the rounded empirical grid point associated with the upper-tail level $1-\tau$, so that $[1-\tau]_m \in [1-\tau, 1-\tau+\frac{1}{m}]$, and let $[x]_+ := \max\{x, 0\}$. Set*

$$\tau_{\text{eff}}(\tau) := 1 - \Big[\big(\bar{\gamma}_m - \varepsilon_{\text{al}}(\tau,\delta)\big)[1-\tau]_m - b_m(\delta)\sqrt{[1-\tau]_m}\Big]_+, \tag{7}$$

where $\bar{\gamma}_m := \frac{1}{m}\sum_{j=1}^m \gamma_j$, $\varepsilon_{\text{al}}(\tau,\delta) = \sqrt{\frac{\log(3m/\delta)}{2m[1-\tau]_m}}$, and $b_m(\delta) = \sqrt{\frac{1}{2m}\log\frac{3m}{\delta}}$.[4] Then, with probability at least $1-\delta$ over $\mathcal{D}$, simultaneously for all $\tau \in (0,1)$,

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(\tau_{\text{eff}}(\tau)\big)\Big|\mathcal{D}\Big) \geq \tau - \varepsilon_m(\delta), \tag{8}$$

where $\varepsilon_m(\delta) := \sqrt{\frac{\log(6/\delta)}{2m}} + \frac{1}{m}$. In particular, for fixed δ, $\varepsilon_m(\delta) \to 0$ and $\tau_{\text{eff}}(\tau) \to \bar{\gamma}_m\tau + (1-\bar{\gamma}_m)$ as $m \to \infty$ uniformly over $\tau \in (0,1)$.

C.1. Proof for Theorem C.1

We condition on $\mathcal{G}_j := \sigma(\psi_j, \hat{q}_j, n_j)$ and $\mathcal{G} := \sigma(\{\mathcal{G}_j\}_{j=1}^m)$. By Assumption 3.2, conditional on $\mathcal{G}$ the additional randomness used to construct $\hat{p}_j$ (hence $\hat{\Delta}_j$ and the indicators below) is independent across j. For brevity, we abuse notation and write $p(\psi_j), q(\psi_j)$ as p_j, q_j, and similarly $\hat{p}, \hat{q}$. We also ignore the superscript k to simplify the proof, and discussions on the dependence on k can be found in Section 3. Split $\delta = \delta_{\text{DKW}} + \delta_{\text{al}} + \delta_{\text{cnt}}$ with $\delta_{\text{DKW}} = \delta_{\text{al}} = \delta_{\text{cnt}} = \delta/3$, and union bound the three events defined below. Moreover, we prove the theorem in the form of probability of upper tail $\alpha := 1-\tau$ for simplicity, and will reparameterize at the end of the proof. For $\alpha \in (0,1)$, define the rounded upper-tail grid point $\alpha_+(\alpha) := \frac{\lceil m\alpha\rceil}{m}$. Define

$$\varepsilon_{\text{al}}(\alpha,\delta) := \sqrt{\frac{\log(3m/\delta)}{2m\alpha_+(\alpha)}}, \quad b_m(\delta) := \sqrt{\frac{1}{2m}\log\frac{3m}{\delta}},$$

and

$$\alpha_{\text{eff}}(\alpha) := \Big[\big(\bar{\gamma}_m - \varepsilon_{\text{al}}(\alpha,\delta)\big)\alpha_+(\alpha) - b_m(\delta)\sqrt{\alpha_+(\alpha)}\Big]_+.$$

[4] We use the extended endpoint convention that, when the expression inside $[\cdot]_+$ is zero, $\hat{V}_m(1) = +\infty$. Equivalently, under the usual finite empirical convention $\hat{V}_m(1) = \max_j \hat{\Delta}_j$, the non-vacuous finite-threshold guarantee below is asserted only for levels satisfying $\big(\bar{\gamma}_m - \varepsilon_{\text{al}}(\tau,\delta)\big)[1-\tau]_m - b_m(\delta)\sqrt{[1-\tau]_m} > 0$.

For grid values $\alpha_r = r/m$, this reduces to

$$\alpha_{\text{eff}}(\alpha_r) = \left[\left(\bar{\gamma}_m - \varepsilon_{\text{al}}(\alpha_r, \delta)\right)\alpha_r - b_m(\delta)\sqrt{\alpha_r}\right]_+ .$$

We will show that these adjustment terms will coincide with the terms presented in the theorem statement after replacing α with $1 - \tau$.

We first show an idealized case where we have access to the oracle p_j, and hence can obtain the actual sim-to-real discrepancy Δ_j. Under this setting, we can define $F(t) := \mathbb{P}_{\psi\sim\Psi}(\Delta(\psi) \le t)$ and

$$\bar{F}_m(t) := \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{\Delta_j \le t\}, \quad \bar{V}_m(u) := \inf\{t : \bar{F}_m(t) \ge u\}.$$

Set $\varepsilon_{\text{DKW}}(\delta) := \sqrt{\frac{\log(6/\delta)}{2m}}$. By the DKW inequality (Lemma B.1), the event

$$\mathcal{E}_{\text{DKW}} := \left\{ \sup_t |\bar{F}_m(t) - F(t)| \le \varepsilon_{\text{DKW}}(\delta) \right\}$$

has probability at least $1 - \delta_{\text{DKW}}$. On $\mathcal{E}_{\text{DKW}}$, for every $u \in (0, 1)$,

$$F\big(\bar{V}_m(u)\big) \ge \bar{F}_m\big(\bar{V}_m(u)\big) - \varepsilon_{\text{DKW}}(\delta) \ge u - \varepsilon_{\text{DKW}}(\delta).$$

Equivalently,

$$\mathbb{P}_{\psi\sim\Psi}\big(\Delta(\psi) \le \bar{V}_m(u)\big|\mathcal{D}\big) \ge u - \varepsilon_{\text{DKW}}(\delta). \tag{9}$$

Now in our dataset, we do not have access to the above oracle Δ_j. Therefore, we control for Δ_j using the pseudo-discrepancy $\hat{\Delta}_j$. Define $Y_j := \mathbf{1}\{\hat{\Delta}_j \ge \Delta_j\}$. By the coverage of $\mathcal{C}_j(\hat{p}_j, \gamma_j)$ and the definition of $\hat{\Delta}_j$,

$$\mathbb{P}(Y_j = 1 \mid \mathcal{G}_j) \ge \gamma_j.$$

For $r = 1, \ldots, m-1$, recall that $\alpha_r := r/m$ and we define the oracle right-tail set

$$S_r := \{j \in [m] : \Delta_j \ge \bar{V}_m(1 - \alpha_r)\}, \quad s_r := |S_r|.$$

Let $\Delta_{(1)} \le \cdots \le \Delta_{(m)}$ denote the order statistics of $\{\Delta_j\}_{j=1}^m$. Since

$$\bar{V}_m(1 - \alpha_r) = \Delta_{(m-r)},$$

at most $m - r - 1$ observations are strictly below $\bar{V}_m(1 - \alpha_r)$. Hence $s_r \ge r + 1$.

We then link the oracle set to the pseudo-discrepancies by defining

$$T_r := \sum_{j\in S_r} Y_j, \quad U_r := \big|\{j : \hat{\Delta}_j \ge \bar{V}_m(1 - \alpha_r)\}\big|.$$

By construction, $U_r \ge T_r$, and U_r will give us the number of $\hat{\Delta}$ required to achieve the tail count of the empirical oracle quantile function, hence we will try to create a lower bound for U_r via T_r.

Since $\gamma_j = g(n_j)$ and the sample sizes n_j are fixed design quantities under Assumption 3.2, the vector $(\gamma_1, \ldots, \gamma_m)$ is fixed independently of the exchangeable vector $(\Delta_1, \ldots, \Delta_m)$. Hence, conditional on s_r and on the realized multiset $\{\gamma_j\}_{j=1}^m$, the set S_r is a uniformly random subset of $[m]$ of size s_r. Therefore, the collection $\{\gamma_j : j \in S_r\}$ is a simple random sample without replacement from the finite population $\{\gamma_j\}_{j=1}^m$. By Hoeffding–Serfling (Bardenet & Maillard, 2015), for each r and each $t > 0$,

$$\mathbb{P}\Big(\frac{1}{s_r}\sum_{j\in S_r} \gamma_j \le \bar{\gamma}_m - t\Big) \le \exp(-2s_r t^2).$$

Set

$$t_r := \sqrt{\frac{\log(m/\delta_{\mathrm{al}})}{2s_r}} \le \sqrt{\frac{\log(3m/\delta)}{2r}} = \sqrt{\frac{\log(3m/\delta)}{2m\alpha_r}} = \varepsilon_{\mathrm{al}}(\alpha_r, \delta).$$

A union bound over $r = 1, \dots, m-1$ gives an event $\mathcal{E}_{\mathrm{al}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{al}}) \ge 1 - \delta_{\mathrm{al}}$ on which, for all r,

$$\sum_{j \in S_r} \gamma_j \;\ge\; s_r\big(\bar{\gamma}_m - \varepsilon_{\mathrm{al}}(\alpha_r, \delta)\big). \tag{10}$$

Conditional on $\mathcal{G}$, the indicators $\{Y_j\}_{j=1}^m$ are independent and bounded in $[0,1]$, so Lemma B.2 implies that for any $t > 0$,

$$\mathbb{P}\Big(T_r \le \mathbb{E}[T_r \mid \mathcal{G}] - t \Big| \mathcal{G}\Big) \le \exp\Big(-\frac{2t^2}{s_r}\Big).$$

Since $\mathbb{E}[T_r \mid \mathcal{G}] = \sum_{j \in S_r} \mathbb{E}[Y_j \mid \mathcal{G}] \ge \sum_{j \in S_r} \gamma_j$, choose

$$c_m := \sqrt{\frac{1}{2}\log\frac{m}{\delta_{\mathrm{cnt}}}} = \sqrt{\frac{1}{2}\log\frac{3m}{\delta}}, \quad t_r := c_m\sqrt{s_r}.$$

Then $\mathbb{P}(T_r \ge \sum_{j \in S_r} \gamma_j - c_m\sqrt{s_r} \mid \mathcal{G}) \ge 1 - \delta_{\mathrm{cnt}}/m$. A union bound over $r = 1, \dots, m-1$ yields an event $\mathcal{E}_{\mathrm{cnt}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{cnt}}) \ge 1 - \delta_{\mathrm{cnt}}$ on which, for all $r = 1, \dots, m-1$,

$$T_r \ge \sum_{j \in S_r} \gamma_j - c_m\sqrt{s_r}. \tag{11}$$

Combining (10) and (11) gives, for all $r = 1, \dots, m-1$,

$$T_r \ge s_r\big(\bar{\gamma}_m - \varepsilon_{\mathrm{al}}(\alpha_r, \delta)\big) - c_m\sqrt{s_r}$$

on $\mathcal{E}_{\mathrm{al}} \cap \mathcal{E}_{\mathrm{cnt}}$. We now have a computable lower bound for T_r up to s_r. Let

$$A_r := \bar{\gamma}_m - \varepsilon_{\mathrm{al}}(\alpha_r, \delta), \qquad f_r(s) := A_r s - c_m\sqrt{s},$$

and

$$\beta_r := \alpha_{\mathrm{eff}}(\alpha_r) = \left[A_r\alpha_r - b_m(\delta)\sqrt{\alpha_r}\right]_+.$$

Then

$$U_r \ge T_r \ge f_r(s_r).$$

If $\beta_r = 0$, then $1 - \beta_r = 1$, and the desired comparison $\hat{V}_m(1-\beta_r) \ge \bar{V}_m(1-\alpha_r)$ is immediate under the extended endpoint convention $\hat{V}_m(1) = +\infty$. Hence assume $\beta_r > 0$. Since $r = m\alpha_r$ and $b_m(\delta) = c_m/\sqrt{m}$,

$$f_r(r) = rA_r - c_m\sqrt{r} = m\beta_r > 0.$$

Thus $A_r\sqrt{r} > c_m$, and for all $s \ge r$,

$$f_r'(s) = A_r - \frac{c_m}{2\sqrt{s}} \ge A_r - \frac{c_m}{2\sqrt{r}} > 0.$$

Since $s_r \ge r + 1$, by the positive derivative, we obtain

$$U_r \ge T_r \ge f_r(s_r) > f_r(r) = m\beta_r.$$

As U_r is integer,

$$U_r \ge \lfloor m\beta_r \rfloor + 1.$$

Therefore at least $\lfloor m\beta_r \rfloor + 1$ of the $\hat{\Delta}_j$'s are at least $\bar{V}_m(1-\alpha_r)$. By the order-statistic thresholding lemma B.7, with $N = \lfloor m\beta_r \rfloor + 1$,

$$\hat{\Delta}_{(m-N+1)} = \hat{\Delta}_{(m-\lfloor m\beta_r \rfloor)} \ge \bar{V}_m(1-\alpha_r).$$

Since

$$\lceil m(1-\beta_r)\rceil = m - \lfloor m\beta_r \rfloor,$$

this is exactly

$$\hat{V}_m\big(1-\alpha_{\text{eff}}(\alpha_r)\big) \geq \bar{V}_m(1-\alpha_r), \qquad r = 1,\dots,m-1. \tag{12}$$

Finally, work on $\mathcal{E}_{\text{DKW}} \cap \mathcal{E}_{\text{al}} \cap \mathcal{E}_{\text{cnt}}$, which has probability at least $1-\delta$ by the union bound. Fix $\alpha \in (0,1)$ and set $\alpha_+ := \frac{\lceil m\alpha \rceil}{m}$. If $\alpha_+ = 1$, then $1-\alpha-\varepsilon_{\text{DKW}}(\delta) - \frac{1}{m} < 0$, so the desired lower bound is trivial. Hence assume $\alpha_+ < 1$ and let $r = \lceil m\alpha \rceil \in \{1,\dots,m-1\}$. By the definition of $\alpha_{\text{eff}}(\alpha)$ through the rounded grid point, $\alpha_{\text{eff}}(\alpha) = \alpha_{\text{eff}}(\alpha_+)$. Applying (12) at this r gives

$$\hat{V}_m\big(1-\alpha_{\text{eff}}(\alpha)\big) = \hat{V}_m\big(1-\alpha_{\text{eff}}(\alpha_+)\big) \geq \bar{V}_m(1-\alpha_+).$$

Applying (9) with $u = 1-\alpha_+$ gives

$$\begin{aligned}\mathbb{P}_{\psi\sim\Psi}\big(\Delta(\psi) \leq \bar{V}_m(1-\alpha_+) \,\big|\, \mathcal{D}\big) &\geq 1-\alpha_+ - \varepsilon_{\text{DKW}}(\delta) \\ &\geq 1-\alpha-\varepsilon_{\text{DKW}}(\delta) - \frac{1}{m}.\end{aligned}$$

By monotonicity in the threshold,

$$\mathbb{P}_{\psi\sim\Psi}\big(\Delta(\psi) \leq \hat{V}_m\big(1-\alpha_{\text{eff}}(\alpha)\big) \,\big|\, \mathcal{D}\big) \geq 1-\alpha-\varepsilon_{\text{DKW}}(\delta) - \frac{1}{m}.$$

Define

$$\varepsilon_m(\delta) := \varepsilon_{\text{DKW}}(\delta) + \frac{1}{m} = \sqrt{\frac{\log(6/\delta)}{2m}} + \frac{1}{m}.$$

This proves the α-form of the guarantee.

It remains only to reparameterize from the upper-tail level α to the coverage level τ. Set $\alpha = 1-\tau$. Then $\alpha_+(\alpha) = \frac{\lceil m(1-\tau)\rceil}{m} = [1-\tau]_m$. Moreover,

$$\alpha_{\text{eff}}(1-\tau) = \Big[\big(\bar{\gamma}_m - \varepsilon_{\text{al}}(\tau,\delta)\big)[1-\tau]_m - b_m(\delta)\sqrt{[1-\tau]_m}\Big]_+, \text{ where } \varepsilon_{\text{al}}(\tau,\delta) = \sqrt{\frac{\log(3m/\delta)}{2m[1-\tau]_m}}.$$

By definition, $\tau_{\text{eff}}(\tau) := 1-\alpha_{\text{eff}}(1-\tau)$. Therefore,

$$\hat{V}_m\big(1-\alpha_{\text{eff}}(1-\tau)\big) = \hat{V}_m\big(\tau_{\text{eff}}(\tau)\big),$$

and the α-form guarantee becomes

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(\tau_{\text{eff}}(\tau)\big)\Big|\mathcal{D}\Big) \geq \tau - \varepsilon_m(\delta),$$

where

$$\varepsilon_m(\delta) = \sqrt{\frac{\log(6/\delta)}{2m}} + \frac{1}{m}.$$

This is exactly (8).

□

C.2. Proof for Theorem 3.4

Proof. By Theorem C.1, with probability at least $1-\delta$ over $\mathcal{D}$, simultaneously for all $\alpha \in (0,1)$,

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(1-\alpha_{\text{eff}}(\alpha)\big)\Big|\mathcal{D}\Big) \geq 1-\alpha-\varepsilon_m(\delta),$$

where $\varepsilon_m(\delta) = \sqrt{\frac{\log(6/\delta)}{2m}} + \frac{1}{m}$. Our goal is to get a simpler form as in Theorem 3.4, and we observe that from the definition of $\alpha_{\text{eff}}(\cdot)$,

$$\alpha_{\text{eff}}(\alpha') \geq \bar{\gamma}_m \alpha'_+ - \sqrt{\frac{2\alpha'_+}{m} \log \frac{3m}{\delta}} \geq \bar{\gamma}_m \alpha' - \sqrt{\frac{2}{m} \log \frac{3m}{\delta}}.$$

Thus, to guarantee $\alpha_{\text{eff}}(\alpha') \geq \bar{\gamma}_m \alpha$, it is enough to choose $\alpha' = \alpha + \rho_m(\delta)$ with

$$\bar{\gamma}_m \rho_m(\delta) = \sqrt{\frac{2}{m} \log \frac{3m}{\delta}} \implies \rho_m(\delta) = \frac{1}{\bar{\gamma}_m} \sqrt{\frac{2\log(3m/\delta)}{m}}.$$

Now, if $\alpha + \rho_m(\delta) \geq 1$, then $1 - \alpha - \varepsilon_m(\delta) - \rho_m(\delta) \leq 0$, so the desired lower bound is trivial. Hence assume $\alpha + \rho_m(\delta) < 1$, and by applying Theorem C.1 at α' gives

$$\mathbb{P}_{\psi \sim \Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(1 - \alpha_{\text{eff}}(\alpha')\big) \Big| \mathcal{D}\Big) \geq 1 - \alpha' - \varepsilon_m(\delta).$$

The clean pseudo-quantile level should have upper-tail mass $\bar{\gamma}_m \alpha$, since only a $\bar{\gamma}_m$-fraction of the oracle upper tail is certified by the pseudo-discrepancies. Thus the corresponding pseudo-quantile index is $1 - \bar{\gamma}_m \alpha$. By the choice of $\rho_m(\delta)$ above, we have $\alpha_{\text{eff}}(\alpha') \geq \bar{\gamma}_m \alpha$, and hence $1 - \alpha_{\text{eff}}(\alpha') \leq 1 - \bar{\gamma}_m \alpha$. Since $\hat{V}_m(\cdot)$ is nondecreasing,

$$\hat{V}_m\big(1 - \alpha_{\text{eff}}(\alpha')\big) \leq \hat{V}_m\big(1 - \bar{\gamma}_m \alpha\big).$$

Therefore, by monotonicity of the event in the threshold,

$$\mathbb{P}_{\psi \sim \Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(1 - \bar{\gamma}_m \alpha\big) \Big| \mathcal{D}\Big) \geq 1 - \alpha' - \varepsilon_m(\delta) = 1 - \alpha - \rho_m(\delta) - \varepsilon_m(\delta).$$

Finally, return to the coverage notation $\tau = 1 - \alpha$. Then $1 - \bar{\gamma}_m \alpha = \bar{\gamma}_m \tau + (1 - \bar{\gamma}_m)$, and the last display becomes

$$\mathbb{P}_{\psi \sim \Psi}\Big(\Delta_\psi \leq \hat{V}_m\big(\bar{\gamma}_m \tau + (1 - \bar{\gamma}_m)\big) \Big| \mathcal{D}\Big) \geq \tau - \widetilde{\varepsilon}_m(\delta),$$

where

$$\widetilde{\varepsilon}_m(\delta) := \varepsilon_m(\delta) + \rho_m(\delta) = \sqrt{\frac{\log(6/\delta)}{2m}} + \frac{1}{m} + \frac{1}{\bar{\gamma}_m} \sqrt{\frac{2\log(3m/\delta)}{m}}.$$

Since Theorem C.1 holds simultaneously for all $\alpha \in (0, 1)$, the preceding argument holds simultaneously for all $\tau \in (0, 1)$. This proves Theorem 3.4. □

Corollary C.2. *Adopt the setting of Theorem C.1 and set $\gamma_j = \gamma$ for all $j \in [m]$. Then, for any $\alpha \in (0, 1)$, with probability at least $1 - \eta$ over $\mathcal{D}$, we have*

$$\mathbb{P}_{\psi \sim \Psi}\Big(\Delta_\psi^{(k)} \leq \hat{V}_m(1 - \gamma\alpha) \Big| \mathcal{D}\Big) \geq 1 - \alpha - \frac{\varepsilon_\gamma(\alpha, m, \eta)}{\sqrt{m}}, \tag{13}$$

where the remainder ε_γ is $O\big(\sqrt{(\log m)}\big)$ as $m \to \infty$.

C.3. Proof for Corollary C.2:

One method is to directly adopt Theorem C.1's proof and readjust the remainder term. Here we adopt a slightly alternative method. We first consider a fixed level of confidence $\bar{\alpha}$, then extend to $\bar{\alpha}$ holding uniformly across $(0, 1)$. Throughout the proof, $\alpha \in (0, 1)$ denotes a generic quantile index (a function argument) and is distinct from the target coverage level $\bar{\alpha}$.

As a setup, we work conditionally on $\mathcal{G}_j := \sigma(\psi_j, \hat{q}_j, n_j), \mathcal{G} := \sigma\big(\{\mathcal{G}_j\}_{j=1}^m\big)$. Here $\mathcal{G}_j$ is the σ-field collecting all "human-side" information for scenario j. Thus $\mathcal{G} := \sigma(\{\mathcal{G}_j\}_{j=1}^m)$ represents the joint information from all scenarios that we treat as fixed when analyzing the remaining randomness coming from the estimation noise in $\hat{p}_j$. For brevity, we abuse notation and write $p(\psi_j), q(\psi_j)$ as p_j, q_j, and similarly $\hat{p}, \hat{q}$. We also ignore the superscript k to simplify the proof, and discussions on the dependence on k can be found in Section 3. In addition, for any quantities $\{\Delta_j\}_{j=1}^m$, we denote the sorted version as $\{\Delta_{(i)}\}_{i=1}^m$, ie. $\Delta_{(1)} \leq \cdots \leq \Delta_{(m)}$. For any sequence $\{\Delta_j\}_{j=1}^m$, let $\Delta_{(1)} \leq \cdots \leq \Delta_{(m)}$ denote its order statistics.

We seek the quantile function of $\Delta_j = L(p_j, \hat{q}_j)$, but only observe the estimators $(\hat{p}_j, \hat{q}_j)$. Note that the sequence $\{\Delta_j\}_{j=1}^m$ is i.i.d., since the simulator budget k is fixed and each scenarios $\{\psi_j\}$ are i.i.d. as assumed in Section 2.

Let $\bar{V}_m(\alpha) = \inf\{t : \bar{F}_m(t) \geq \alpha\} = \Delta_{(\lceil m\alpha \rceil)}$ be the empirical α quantile of $\{\Delta_j\}_{j=1}^m$, where $\bar{F}_m(x) = \frac{1}{m}\sum_{j=1}^m \mathbb{I}(\Delta_j \leq x)$. By Lemma B.1, with probability $1-\delta$:

$$\mathbb{P}_{\psi\sim\Psi}(L(p(\psi), \hat{q}(\psi)) \leq \bar{V}_m(1-\frac{\bar{\alpha}}{2})|\mathcal{D}) \geq 1 - \frac{\bar{\alpha}}{2} - \sqrt{\frac{\log(2/\delta)}{2m}} \tag{14}$$

and we would have our desired bound.

However, for the above argument to hold, we would need to know p_j, instead we only have the estimated $\hat{p}_j$. Therefore, we instead construct a pseudo-gaps using confidence sets with coverage level γ, and define the two gap terms as

$$\begin{cases} \Delta_j = L(p_j, \hat{q}_j) & \text{, i.i.d., yet unobservable} \\ \hat{\Delta}_j := \sup_{u\in C_j(\hat{p}_j,\gamma)} L(u, \hat{q}_j) & \text{, Not i.i.d., yet observable,} \end{cases}$$

where $\mathcal{C}_j(\hat{p}_j, \gamma) \subset \Theta$ are data-driven confidence sets satisfying $\mathbb{P}(p_j \in \mathcal{C}_j(\hat{p}_j, \gamma) \,|\psi_j, n_j) \geq \gamma$.

By Assumption 3.3 and the compactness of the confidence set $C_j(\hat{p}_j, \gamma) \subset \Theta$, Berge's maximum theorem guarantees that the supremum in the definition of $\hat{\Delta}_j$ is attained, hence $\hat{\Delta}_j$ is well defined. Therefore, by the coverage property of C_j, we have $\mathbb{P}(\hat{\Delta}_j \geq \Delta_j | \mathcal{G}_j \cup \sigma(\gamma)) \geq \gamma$, and we get $\forall j$

$$\mathbb{P}(\hat{\Delta}_j \geq \Delta_j | \mathcal{G}_j) \geq \gamma. \tag{15}$$

Furthermore, by the tower property and $\gamma \perp\!\!\!\perp \mathcal{G}_j$,

$$\mathbb{P}\big(\hat{\Delta}_j \geq \Delta_j\big) = \mathbb{E}\big[\mathbb{P}(\hat{\Delta}_j \geq \Delta_j \big| \mathcal{G}_j)\big] \geq \gamma.$$

Moreover, by the above discussion, conditional on $\mathcal{G}$, the indicators $Y_j = \mathbf{1}\{\hat{\Delta}_j \geq \Delta_j\}$ are independent across j. We will use $\{\hat{\Delta}_j\}_{j=1}^m$ to create an upper bound of $\bar{V}_m(1-\alpha)$, which along side (14) will give us our desired envelope. Intuitively, we want to find a larger quantile of $\hat{\Delta}_j$ and with (15), we can claim an upper bound of $\bar{V}_m(1-\alpha)$ with high probability.

First, define $S_\alpha := \{j \in [m] : \Delta_j \geq \bar{V}_m(1-\alpha)\}$ and $s := |S_\alpha|$. By definition of the empirical $(1-\alpha)$–quantile $\bar{V}_m(1-\alpha)$, at most a fraction $(1-\alpha)$ of the sample can lie strictly below it. More precisely,

$$\big|\{j \in [m] : \Delta_j < \bar{V}_m(1-\alpha)\}\big| \leq \lfloor m(1-\alpha) \rfloor.$$

Therefore the number of indices with $\Delta_j \geq \bar{V}_m(1-\alpha)$ is at least

$$s \geq m - \lfloor m(1-\alpha) \rfloor = \lceil m\alpha \rceil \geq \lfloor m\alpha \rfloor.$$

We next define $Y_j = \mathbb{I}(\hat{\Delta}_j \geq \Delta_j)$. By (15): $\mathbb{P}(Y_j = 1|\mathcal{G}) \geq \gamma$. Next, define $T_\alpha = \{j \in S_\alpha : \hat{\Delta}_j \geq \Delta_j\} = \{j \in S_\alpha : Y_j = 1\} = \{j : \hat{\Delta}_j \geq \Delta_j \geq \bar{V}_m(1-\alpha)\}$. First, we calculate $\mathbb{E}[T_\alpha]$:

$$\mathbb{E}[T_\alpha|\mathcal{G}] = \mathbb{E}\Big[\sum_{j\in S_\alpha} Y_j|\mathcal{G}\Big] = \sum_{j\in S_\alpha} \mathbb{P}(Y_j|\mathcal{G}) \geq \gamma s.$$

Lemma B.2 implies that for any $\delta \in [0, \gamma s]$, we have:

$$\mathbb{P}\Big(|T_\alpha| \leq \gamma s - t|\mathcal{G}\Big) \leq \mathbb{P}\Big(|T_\alpha| \leq \mathbb{E}[T_\alpha] - t|\mathcal{G}\Big) \leq \exp(-\frac{2t^2}{s}),$$

where we applied $\mathbb{E}[T_\alpha|\mathcal{G}] \geq \gamma s$ in the first inequality. By setting $t = c\sqrt{s}$:

$$\mathbb{P}\Big(|T_\alpha| \leq \gamma s - t|\mathcal{G}\Big) \leq \exp(-2c^2) \tag{16}$$

With this bound, we can link the actual set of indices we have interest, ie. $U_\alpha = \{j : \hat{\Delta}_j \geq \bar{V}_m(1-\alpha)\}$ to T_α. By construction, $T_\alpha \subseteq U_\alpha$, hence by (16), with probability greater than $1 - \exp(-2c^2)$:

$$|U_\alpha| \geq |T_\alpha| \geq \gamma s - c\sqrt{s},$$

which implies at least $\gamma s - c\sqrt{s}$ of the $\hat{\Delta}_j$'s are larger than $\bar{V}_m(1-\alpha)$ with high probability.

We now analyze what coverage guarantee can we get for the inner probability via order statistics for any α. Set $N := \lfloor \gamma s - c\sqrt{s} \rfloor$ and define $\hat{V}_m(\alpha) := \inf\{t : \hat{F}_m(t) \geq \alpha\} = \Delta_{(\lceil m\alpha \rceil)}$, where $\hat{F}(x) = \frac{1}{m}\sum_{j=1}^m \mathbb{I}(\hat{\Delta}_j \leq x)$. If at least N sample values exceed $\bar{V}_m(1-\alpha)$, then by order-statistics calculus (Lemma B.7)

$$\hat{V}_m(1-\alpha) = \hat{\Delta}_{(m-\lfloor m\alpha \rfloor)} \geq \bar{V}_m(1-\alpha_{\text{eff}}),$$

whenever α_{eff} is chosen so that $N \geq \lfloor m\alpha \rfloor + 1$ holds when $s \geq \lfloor m\alpha_{\text{eff}} \rfloor + 1$.

A sufficient condition for the above to satisfy is

$$\gamma m \alpha_{\text{eff}} - c\sqrt{m\alpha_{\text{eff}}} - 1 \geq m\alpha.$$

Define $\alpha_{\text{eff}}(\alpha, c, m) := \inf\{x \in (0,1) : \gamma x - c\sqrt{\frac{x}{m}} - \frac{1}{m} \geq \alpha\}$. Writing $y = \sqrt{x}$, this is equivalent to $\gamma y^2 - \frac{c}{\sqrt{m}} y - \left(\frac{1}{m} + \alpha\right) \geq 0$, so the minimal admissible y is

$$y^* = \frac{c/\sqrt{m} + \sqrt{c^2/m + 4\gamma\alpha + 4\gamma/m}}{2\gamma}, \quad \alpha_{\text{eff}}(\alpha, c, m) = (y^*)^2.$$

Thus, we define

$$\mathcal{E}_\alpha := \left\{ \hat{V}_m(1-\alpha) \geq \bar{V}_m\big(1 - \alpha_{\text{eff}}(\alpha, c, m)\big) \right\}, \tag{17}$$
$$\text{then} \qquad \mathbb{P}(\mathcal{E}_\alpha) \geq 1 - e^{-2c^2}.$$

Next, we apply (14) at level $1 - \alpha_{\text{eff}}(\bar{\alpha}, c, m)$:

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi \leq \bar{V}_m(1-\alpha_{\text{eff}}) \Big| \mathcal{D}\Big) \geq 1 - \alpha_{\text{eff}}(\bar{\alpha}, c, m) - \varepsilon_m(\delta).$$

On $\mathcal{E}_{\bar{\alpha}}$ in (17), $\bar{V}_m(1-\alpha_{\text{eff}}) \leq \hat{V}_m(1-\bar{\alpha})$. Hence,

$$\mathbb{P}_{\psi\sim\Psi}\big(\Delta_\psi \leq \hat{V}_m(1-\bar{\alpha}) \big| \mathcal{D}\big) \geq 1 - \alpha_{\text{eff}}(\bar{\alpha}, c, m) - \varepsilon_m(\delta), \tag{18}$$

with outer probability at least $1 - \delta - e^{-2c^2}$.

The exact algebraic form is

$$\begin{aligned}\alpha_{\text{eff}}(\bar{\alpha}, c, m) &= \frac{\left(\frac{c}{\sqrt{m}} + \sqrt{\frac{c^2}{m} + 4\gamma\bar{\alpha} + \frac{4\gamma}{m}}\right)^2}{4\gamma^2} \\ &= \frac{\bar{\alpha}}{\gamma} + \frac{1}{\gamma m} + \frac{c^2}{2\gamma^2 m} + \frac{c}{2\gamma^2\sqrt{m}}\sqrt{4\gamma\bar{\alpha} + \frac{c^2 + 4\gamma}{m}}.\end{aligned}$$

Therefore, for fixed $\bar{\alpha} > 0$ as $m \to \infty$,

$$\alpha_{\text{eff}}(\bar{\alpha}, c, m) = \frac{\bar{\alpha}}{\gamma} + O(m^{-1/2}).$$

We have shown that for any target level α and choice of $c > 0$, the preceding argument yields a high-probability concentration bound based on $\{\hat{\Delta}_j\}_{j=1}^m$. We now extend the guarantee to hold *uniformly* over all α. Fix $c > 0$ and $\delta \in (0,1)$. For the grid $\alpha_r := r/m$ $(r = 1, \ldots, m)$, let

$$\mathcal{E}_{\text{DKW}} := \left\{ \sup_x \left|\hat{F}_m(x) - F^*(x)\right| \leq \varepsilon_m(\delta) \right\}, \quad \mathcal{E}_r := \Big\{\text{(18) holds with } \alpha = \alpha_r\Big\}.$$

By DKW, $\mathbb{P}(\mathcal{E}_{\mathrm{DKW}}) \geq 1-\delta$, and by the fixed-level argument, $\mathbb{P}(\mathcal{E}_r) \geq 1 - e^{-2c^2}$ for each r. Hence, by a union bound,

$$\mathbb{P}\Big(\mathcal{E}_{\mathrm{DKW}} \cap \bigcap_{r=1}^{m} \mathcal{E}_r\Big) \geq 1 - \delta - me^{-2c^2}.$$

For any $\alpha \in (0,1)$ let $r = \lceil m\alpha \rceil$ and denote $\alpha_+ := \alpha_r = r/m \in [\alpha, \alpha + 1/m]$. Since the empirical quantile is piecewise constant on the m-grid,

$$\widehat{V}_m(1-\alpha) = \widehat{V}_m(1-\alpha_+).$$

Applying (18) at level α_+ yields

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta(\psi) \leq \widehat{V}_m(1-\alpha)\Big|\mathcal{D}\Big) \geq 1 - \frac{\alpha_+}{\gamma} - \frac{c}{\gamma\sqrt{m}}\sqrt{4\gamma\alpha_+ + \tfrac{c^2+4\gamma}{m}} - \frac{c^2}{2\gamma^2 m} - \frac{1}{\gamma m} - \varepsilon_m(\delta).$$

Since $\alpha_+ \in [\alpha, \alpha + 1/m]$ and the right-hand side is nonincreasing in α, the same bound holds with α replaced by α_+, and (optionally) one may absorb the rounding slack $\alpha_+ - \alpha \leq 1/m$ into the $O(m^{-1})$ term by a crude inequality

$$\frac{\alpha_+}{\gamma} \leq \frac{\alpha}{\gamma} + \frac{1}{\gamma m}, \quad \sqrt{4\gamma\alpha_+ + \tfrac{c^2+4\gamma}{m}} \leq \sqrt{4\gamma\alpha + \tfrac{c^2+8\gamma}{m}}.$$

Therefore, with probability at least $1 - \delta - me^{-2c^2}$ over $\mathcal{D}$, the guarantee (13) (at α replaced by α_+) holds *uniformly* for all $\alpha \in (0,1)$. The form in the main theorem is a simplification with respect to c and δ while taking $\gamma = \frac{1}{2}$.

C.4. Corollary for simulator budget:

Corollary C.3. *Suppose Assumptions 3.2 and 3.3 hold. For any simulation sample size $k_j \in \mathbb{N}$, define the per-scenario true sim-to-real discrepancy and its pseudo-discrepancy by*

$$\Delta_j^\star := L(p_j, q_j), \qquad \hat{\Delta}_j^{\star,(k)} := \sup_{\substack{u\in C_j^p(\hat{p}_j)\\ v\in C_j^q(\hat{q}_j)}} L(u,v),$$

where $C_j^p(\hat{p}_j) \subset \Theta_p$, $C_j^q(\hat{q}_j) \subset \Theta_q$ are data-driven compact confidence sets such that

$$\mathbb{P}\big(p_j \in C_j^p(\hat{p}_j),\ q_j \in C_j^q(\hat{q}_j)\big|\psi_j, n_j, k_j\big) \geq \tfrac{1}{2}.$$

Let $\hat{V}_m^\star(\alpha)$ denote the empirical α-quantile of $\{\hat{\Delta}_j^{\star,(k)}\}_{j=1}^m$.

Then, for any $\alpha \in (0,1)$ and any $\eta \in (0,1)$, with probability at least $1-\eta$ over $\mathcal{D}$, we have

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi^\star \leq \hat{V}_m^\star\big(1-\tfrac{\alpha}{2}\big)\Big|\mathcal{D}\Big) \geq 1 - \alpha - \frac{\varepsilon(\alpha,m,\eta)}{\sqrt{m}}, \tag{19}$$

where the remainder $\varepsilon(\alpha,m,\eta)$ is the same as in Theorem C.2, in particular $\varepsilon(\alpha,m,\eta) = O\big(\sqrt{\log m}\big)$ and $\varepsilon(\alpha,m,\eta)/\sqrt{m} = O\big(\sqrt{(\log m)/m}\big)$.

Proof. The argument is identical to the proof of Theorem C.2, except that we now construct separate confidence sets for p_j and q_j with marginal coverages γ_p and γ_q chosen so that $\gamma_p\gamma_q = \frac{1}{2}$. For example, we can set $\gamma_p = \gamma_q = \frac{1}{\sqrt{2}}$. This ensures that the joint event $\{p_j \in C_j^p(\hat{p}_j),\ q_j \in C_j^q(\hat{q}_j)\}$ plays exactly the same role as the univariate coverage event in Theorem C.1, i.e., $\mathbb{P}(\hat{\Delta}_j^{\star,(k)} \geq \Delta_j^{\star,(k)} \mid \mathcal{G}_j) \geq \frac{1}{2}$. All subsequent steps then carry over verbatim, yielding the stated bound. □

Theorem C.4. *Suppose the setup of Theorem C.1 and Assumptions 3.2–3.3 hold. Let $\{\gamma_{L,j}, \gamma_{U,j}\}_{j=1}^m \subset (0,1]$ be coverage levels and define the lower and upper pseudo-discrepancies*

$$\hat{\Delta}_j^- := \inf_{u\in\mathcal{C}_j(\hat{p}_j,\gamma_{L,j})} L(u,\hat{q}_j), \quad \hat{\Delta}_j^+ := \sup_{u\in\mathcal{C}_j(\hat{p}_j,\gamma_{U,j})} L(u,\hat{q}_j),$$

where $\mathcal{C}_j(\hat{p}_j, \gamma) \subset \Theta$ are data-driven compact confidence sets satisfying the conditional coverage guarantees

$$\mathbb{P}\big(p_j \in \mathcal{C}_j(\hat{p}_j, \gamma_{L,j}) \mid \psi_j, n_j\big) \geq \gamma_{L,j}$$
$$\mathbb{P}\big(p_j \in \mathcal{C}_j(\hat{p}_j, \gamma_{U,j}) \mid \psi_j, n_j\big) \geq \gamma_{U,j}.$$

Let $\hat{V}_m^-(\cdot)$ denote the empirical quantile function of $\{\hat{\Delta}_j^-\}_{j=1}^m$, and $\hat{V}_m^+(\cdot)$ the empirical quantile function of $\{\hat{\Delta}_j^+\}_{j=1}^m$. Define the empirical mean coverage levels $\bar{\gamma}_{L,m} := \frac{1}{m}\sum_{j=1}^m \gamma_{L,j}, \bar{\gamma}_{U,m} := \frac{1}{m}\sum_{j=1}^m \gamma_{U,j}$. For $\alpha \in (0,1)$ and $\delta \in (0,1)$ define the discretized level $\tilde{\alpha}(\alpha) := \frac{\lceil m\alpha \rceil}{m}$. Set $\varepsilon_{\text{al}}^{\text{bd}}(\alpha, \delta) := \sqrt{\frac{\log(6m/\delta)}{2m\tilde{\alpha}(\alpha)}}$, and $b_m^{\text{bd}}(\delta) := \sqrt{\frac{1}{2m}\log\frac{6m}{\delta}}$. Define the effective indices

$$\alpha_{\text{eff}}^L(\alpha) := \Big[\big(\bar{\gamma}_{L,m} - \varepsilon_{\text{al}}^{\text{bd}}(\alpha,\delta)\big)\tilde{\alpha}(\alpha) - b_m^{\text{bd}}(\delta)\sqrt{\tilde{\alpha}(\alpha)}\Big]_+,$$
$$\alpha_{\text{eff}}^U(\alpha) := \Big[\big(\bar{\gamma}_{U,m} - \varepsilon_{\text{al}}^{\text{bd}}(\alpha,\delta)\big)\tilde{\alpha}(\alpha) - b_m^{\text{bd}}(\delta)\sqrt{\tilde{\alpha}(\alpha)}\Big]_+.$$

Then, with probability at least $1-\delta$ over $\mathcal{D}$, simultaneously for all $\alpha \in (0,1)$,

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta(\psi) \geq \hat{V}_m^-\big(\alpha_{\text{eff}}^L(\alpha)\big)\Big|\mathcal{D}\Big) \geq 1 - \alpha - \varepsilon_m^{\text{bd}}(\delta),$$
$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta(\psi) \leq \hat{V}_m^+\big(1-\alpha_{\text{eff}}^U(\alpha)\big)\Big|\mathcal{D}\Big) \geq 1 - \alpha - \varepsilon_m^{\text{bd}}(\delta),$$

where $\varepsilon_m^{\text{bd}}(\delta) := \sqrt{\frac{\log(12/\delta)}{2m}} + \frac{1}{m}$. We use the endpoint conventions $\hat{V}_m^-(0) = -\infty$ and $\hat{V}_m^+(1) = +\infty$ whenever the corresponding effective index degenerates to zero. Let $V : [0,1] \to \mathbb{R}$ denote the quantile function of $\Delta(\psi)$. In particular, for any fixed $\tau \in (0,1)$, as $m \to \infty$, the band admits the heuristic form

$$\hat{V}_m^-\big(\bar{\gamma}_{L,m}\tau\big) \lesssim V(\tau) \lesssim \hat{V}_m^+\big(\bar{\gamma}_{U,m}\tau + (1-\bar{\gamma}_{U,m})\big) \tag{20}$$

C.5. Proof for Theorem C.4:

We prove the lower-band inequality with failure probability $\delta/2$. The upper-band inequality follows by applying the α-form of Theorem C.1 with failure probability $\delta/2$ to the upper pseudo-discrepancies $\{\hat{\Delta}_j^+\}_{j=1}^m$ and adaptive coverage levels $\{\gamma_{U,j}\}_{j=1}^m$. Combining the two one-sided events by a union bound then gives the claimed probability $1-\delta$.

Recall that we work conditionally on $\mathcal{G}_j := \sigma(\psi_j, \hat{q}_j, n_j)$, $\mathcal{G} := \sigma(\{\mathcal{G}_j\}_{j=1}^m)$. The true per-scenario discrepancy is $\Delta_j := L(p_j, \hat{q}_j)$, for $j = 1, \ldots, m$. Define the lower pseudo-discrepancies and validity indicators by

$$\hat{\Delta}_j^- := \inf_{u \in \mathcal{C}_j(\hat{p}_j, \gamma_{L,j})} L(u, \hat{q}_j), \quad Y_j^- := \mathbf{1}\{\hat{\Delta}_j^- \leq \Delta_j\}.$$

By Assumption 3.3 and compactness of $\mathcal{C}_j(\hat{p}_j, \gamma_{L,j})$, the infimum is attained. Moreover, on the event $\{p_j \in \mathcal{C}_j(\hat{p}_j, \gamma_{L,j})\}$, we have $\hat{\Delta}_j^- \leq \Delta_j$. Therefore, by conditional coverage,

$$\mathbb{P}(Y_j^- = 1 \mid \mathcal{G}_j) \geq \gamma_{L,j}.$$

Conditional on $\mathcal{G}$, the indicators $\{Y_j^-\}_{j=1}^m$ are independent.

Set $\delta_{\text{DKW}} = \delta_{\text{al}} = \delta_{\text{cnt}} = \delta/6$ and $c_m := \sqrt{\frac{1}{2}\log\frac{6m}{\delta}}$. For $r = 1, \ldots, m$, let $\alpha_r := r/m$ and define the oracle lower-tail set

$$S_r^- := \{j \in [m] : \Delta_j \leq \bar{V}_m(\alpha_r)\}, \quad s_r^- := |S_r^-|.$$

By the definition of empirical quantiles, $s_r^- \geq r$. Define the counts

$$T_r^- := \sum_{j \in S_r^-} Y_j^-, \quad U_r^- := \big|\{j : \hat{\Delta}_j^- \leq \bar{V}_m(\alpha_r)\}\big|.$$

By construction, $U_r^- \geq T_r^-$. Since the coverage levels $\{\gamma_{L,j}\}_{j=1}^m$ are fixed design quantities and $(\Delta_1, \ldots, \Delta_m)$ is exchangeable, conditional on s_r^- and on the realized multiset $\{\gamma_{L,j}\}_{j=1}^m$, the set S_r^- is a uniformly random subset of $[m]$ of

size s_r^-. Hence $\{\gamma_{L,j} : j \in S_r^-\}$ is a simple random sample without replacement from the finite population $\{\gamma_{L,j}\}_{j=1}^m$. By Hoeffding–Serfling and a union bound over $r = 1, \ldots, m$, there is an event $\mathcal{E}_{\mathrm{al}}^-$ with probability at least $1 - \delta_{\mathrm{al}}$ on which, for all $r = 1, \ldots, m$,

$$\sum_{j \in S_r^-} \gamma_{L,j} \geq s_r^- \Big(\bar{\gamma}_{L,m} - \varepsilon_{\mathrm{al}}^{\mathrm{bd}}(\alpha_r, \delta)\Big).$$

Next, conditional on $\mathcal{G}$, the indicators $\{Y_j^-\}_{j=1}^m$ are independent and bounded. By the additive Chernoff bound and a union bound over $r = 1, \ldots, m$, there is an event $\mathcal{E}_{\mathrm{cnt}}^-$ with probability at least $1 - \delta_{\mathrm{cnt}}$ on which, for all $r = 1, \ldots, m$,

$$T_r^- \geq \sum_{j \in S_r^-} \gamma_{L,j} - c_m \sqrt{s_r^-}\,.$$

On $\mathcal{E}_{\mathrm{al}}^- \cap \mathcal{E}_{\mathrm{cnt}}^-$, define

$$A_r^- := \bar{\gamma}_{L,m} - \varepsilon_{\mathrm{al}}^{\mathrm{bd}}(\alpha_r, \delta), \quad f_r^-(s) := A_r^- s - c_m \sqrt{s},$$

and $\beta_r^- := \alpha_{\mathrm{eff}}^L(\alpha_r) = \big[A_r^- \alpha_r - b_m^{\mathrm{bd}}(\delta)\sqrt{\alpha_r}\big]_+$. Then

$$U_r^- \geq T_r^- \geq f_r^-(s_r^-).$$

If $\beta_r^- = 0$, the desired lower quantile comparison is immediate under the endpoint convention $\hat{V}_m^-(0) = -\infty$. Suppose $\beta_r^- > 0$. Since $r = m\alpha_r$ and $b_m^{\mathrm{bd}}(\delta) = c_m/\sqrt{m}$,

$$f_r^-(r) = rA_r^- - c_m\sqrt{r} = m\beta_r^- > 0.$$

Thus $A_r^- \sqrt{r} > c_m$, and for all $s \geq r$,

$$(f_r^-)'(s) = A_r^- - \frac{c_m}{2\sqrt{s}} \geq A_r^- - \frac{c_m}{2\sqrt{r}} > 0.$$

Since $s_r^- \geq r$, we obtain

$$U_r^- \geq T_r^- \geq f_r^-(s_r^-) \geq f_r^-(r) = m\beta_r^-\,.$$

As U_r^- is integer, $U_r^- \geq \lceil m\beta_r^- \rceil$. Therefore at least $\lceil m\beta_r^- \rceil$ of the lower pseudo-discrepancies are no larger than $\bar{V}_m(\alpha_r)$. Hence, by the order-statistic thresholding lemma,

$$\hat{V}_m^-\big(\alpha_{\mathrm{eff}}^L(\alpha_r)\big) \leq \bar{V}_m(\alpha_r), \quad r = 1, \ldots, m. \tag{21}$$

It remains to transfer this empirical oracle comparison to the population distribution. For the lower-band argument, use the strict CDFs

$$F_\Delta^<(t) := \mathbb{P}_{\psi \sim \Psi}(\Delta(\psi) < t), \qquad \bar{F}_m^<(t) := \frac{1}{m}\sum_{j=1}^m \mathbf{1}\{\Delta_j < t\}.$$

The DKW inequality applies equally to these strict CDFs. Thus, with probability at least $1 - \delta_{\mathrm{DKW}}$,

$$\mathcal{E}_{\mathrm{DKW}}^< := \left\{ \sup_t \big|\bar{F}_m^<(t) - F_\Delta^<(t)\big| \leq \sqrt{\frac{\log(12/\delta)}{2m}} \right\}.$$

Work on $\mathcal{E}_{\mathrm{DKW}}^< \cap \mathcal{E}_{\mathrm{al}}^- \cap \mathcal{E}_{\mathrm{cnt}}^-$. Fix any $\alpha \in (0,1)$ and set $\alpha_+ := \frac{\lceil m\alpha \rceil}{m}$. Let $r = \lceil m\alpha \rceil$. By the definition of $\alpha_{\mathrm{eff}}^L(\alpha)$ through the rounded grid point,

$$\alpha_{\mathrm{eff}}^L(\alpha) = \alpha_{\mathrm{eff}}^L(\alpha_+).$$

Applying (21) at this r gives

$$\hat{V}_m^-\big(\alpha_{\mathrm{eff}}^L(\alpha)\big) \leq \bar{V}_m(\alpha_+).$$

Therefore, by monotonicity and the strict-DKW event,

$$F_{\Delta}^{<}\Big(\hat{V}_m^{-}\big(\alpha_{\text{eff}}^{L}(\alpha)\big)\Big) \le F_{\Delta}^{<}\big(\bar{V}_m(\alpha_+)\big) \le \bar{F}_m^{<}\big(\bar{V}_m(\alpha_+)\big) + \sqrt{\frac{\log(12/\delta)}{2m}}.$$

Since α_+ is a grid point and $\bar{V}_m(\alpha_+) = \Delta_{(\lceil m\alpha_+ \rceil)}$, at most $\lceil m\alpha_+ \rceil - 1$ oracle discrepancies are strictly below $\bar{V}_m(\alpha_+)$. Hence

$$\bar{F}_m^{<}\big(\bar{V}_m(\alpha_+)\big) \le \alpha_+.$$

Thus

$$F_{\Delta}^{<}\Big(\hat{V}_m^{-}\big(\alpha_{\text{eff}}^{L}(\alpha)\big)\Big) \le \alpha + \frac{1}{m} + \sqrt{\frac{\log(12/\delta)}{2m}}.$$

Equivalently,

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta(\psi) \ge \hat{V}_m^{-}\big(\alpha_{\text{eff}}^{L}(\alpha)\big) \,\Big|\, \mathcal{D}\Big) \ge 1 - \alpha - \varepsilon_m^{\text{bd}}(\delta).$$

This proves the adaptive lower-band inequality simultaneously for all $\alpha \in (0,1)$ with failure probability at most $\delta/2$.

The adaptive upper-band inequality follows from Theorem C.1, applied with failure probability $\delta/2$ to the upper pseudo-discrepancies $\{\hat{\Delta}_j^{+}\}_{j=1}^{m}$ and coverage levels $\{\gamma_{U,j}\}_{j=1}^{m}$. Under this substitution,

$$\bar{\gamma}_m \mapsto \bar{\gamma}_{U,m}, \qquad \varepsilon_{\text{al}}(\alpha,\delta/2) = \varepsilon_{\text{al}}^{\text{bd}}(\alpha,\delta), \qquad b_m(\delta/2) = b_m^{\text{bd}}(\delta),$$

and the DKW remainder becomes

$$\sqrt{\frac{\log(12/\delta)}{2m}} + \frac{1}{m} = \varepsilon_m^{\text{bd}}(\delta).$$

Therefore, with probability at least $1 - \delta/2$, simultaneously for all $\alpha \in (0,1)$,

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta(\psi) \le \hat{V}_m^{+}\big(1 - \alpha_{\text{eff}}^{U}(\alpha)\big) \,\Big|\, \mathcal{D}\Big) \ge 1 - \alpha - \varepsilon_m^{\text{bd}}(\delta).$$

Combining the lower and upper one-sided events by a union bound gives the claimed simultaneous band with probability at least $1 - \delta$. □

D. World Value Bench Methodology

D.1. Selection of Survey Questions

The World Values Survey contains 259 questions, grouped into categories such as social values, well-being, economic values, and security. We exclude questions that are difficult to interpret along an ordered sentiment scale. For example, in Figure 5, Question 223 is highly country- and time-specific and does not admit a natural ordering of sentiment, making it hard to align with other questions.

Q223. If there were a national election tomorrow, for which party on this list would you vote? Just call out the number on this card. If DON'T KNOW: Which party appeals to you most?

1. Party 1
2. Party 2
3. Party 3
4. etc.

[INSERT COUNTRY-SPECIFIC LIST OF PARTIES]

Figure 5. Text of Question 223.

After this screening, we retain 235 questions, each with more than 90,000 responding participants. Ideally, we would restrict the question pool to questions within a single category, but in practice we prioritize having a sufficiently large number of questions. Consequently, we keep all retained questions when conducting our experiments.

In your view, how often do the following things occur in this country's elections?

	Very often	Fairly often	Not often	Not at all often
Q224.Votes are counted fairly	1	2	3	4
Q225. Opposition candidates are prevented from running	1	2	3	4
Q226. TV news favors the governing party	1	2	3	4
Q227. Voters are bribed	1	2	3	4
Q228. Journalists provide fair coverage of elections	1	2	3	4
Q229. Election officials are fair	1	2	3	4
Q230. Rich people buy elections	1	2	3	4
Q231. Voters are threatened with violence at the polls	1	2	3	4
Q232. Voters are offered a genuine choice in the elections	1	2	3	4
Q233. Women have equal opportunities to run the office	1	2	3	4

Figure 6. Example questions from the Political Interest category.

Q163. All things considered, would you say that the world is better off, or worse off, because of science and technology? Please tell me which comes closest to your view on this scale: 1 means that "the world is a lot worse off," and 10 means that "the world is a lot better off." (*Code one number*):

A lot worse off									A lot better off
1	2	3	4	5	6	7	8	9	10

Figure 7. Example question from the Science and Technology category.

Q121. Now we would like to know your opinion about the people from other countries who come to live in [your country] - the immigrants. How would you evaluate the impact of these people on the development of [your country]?

Very good	Quite good	Neither good, nor bad	Quite bad	Very bad
5	4	3	2	1

Figure 8. Example question from the Migration category.

D.2. Example Question and Preprocess

We list 3 example questions from the dataset.

As discussed in the main text, to obtain a unified scale across heterogeneous categorical response formats, we map all answers to numerical values in $[-1, 1]$. Specifically, we query GPT-5 to determine the direction of this mapping according to its assessment of the alignment of each response in terms of public opinion. For example, in Figure 7, GPT-5 maps option 1 ("A lot worse off") to -1 and option 10 ("A lot better off") to 1, with intermediate choices linearly spaced between these endpoints.[5]

D.3. Synthetic Profile Generation

In the original dataset, some respondent IDs appear irregularly or are duplicated. We exclude these records and are left with 96,220 unique participants. The dataset contains demographic information for each participant, summarized in Table 3.

We select a subset of these demographic attributes to define our synthetic profiles, including Country, Sex, Age, and Settlement Type, among others. The LLMs are asked to produce a categorical response encoded as `[[1]]`, `[[2]]`, etc. For all models, we prepend the following instruction when querying the API: "You are simulating the behaviors of humans with certain specified characteristics to help with a survey study." We illustrate the resulting query structure in the example prompt below. We set the temperature to be $= 0.2$ and top_p $= 0.9$ across persona generation.

[5]This conversion does not represent the authors' values; it reflects GPT-5's inferred public opinion, which may be influenced by RLHF or other factors unknown to the authors.

Table 3. Master list of demographic variables and categories.

Demographic	Categories
Sex	Male; Female
Country	64 countries/societies
Age	Numeric (years)
Language	Text
Migration status	Native-born; First-generation immigrant; Second-generation; Other/missing
Marital status	Married; Living together as married; Divorced; Separated; Widowed; Never married/Single
Children status	Number of children (numeric)
Education level	ISCED levels 0–8[6]
Employment status	Employed full-time; Employed part-time; Self-employed; Unemployed; Student; Homemaker; Retired; Other
Household income decile	1 (lowest) ; 2; 3; 4; 5; 6; 7; 8; 9; 10 (highest)
Religious affiliation	Christian; Muslim; Hindu; Buddhist; Jewish; No religion; Other religion
Self-identified social class	Lower class; Working class; Lower middle class; Upper middle class; Upper class
Settlement Type	Urban; Rural (with Population Size)

Example prompt:
Pretend that you reside in Mexico. You live in a capital city urban with a town size of 100,000-500,000 area. You are female, your age is between 35-44. and you are Married. You normally speak Spanish; Castilian at home. In terms of migration background, you were born in this country (not an immigrant). In terms of education, you attained Upper secondary education (ISCED 3). Your current employment status is: Housewife not otherwise employed. You work in Not applicable; Never had a job. You belong to Roman Catholic; Latin Church;.
On a scale of 1 to 10, 1 meaning 'A lot worse off' and 10 meaning 'A lot better off', do you think science and technology have improved or worsened the world? Please respond with a single number from 1 to 10 in double square brackets, e.g., [[1]].

D.4. Different n_j quantile curves

We also provide a robustness check on whether the dominance relation is an anomaly of the particular choice for the sample size n_j in Figure 9, which shows performance of the LLMs for $n_j = \{50, 500, 5000, 10000\}$. The results indicate that the dominance relation between LLMs remains the same.

D.5. Adaptive coverage rate γ_j choice

As mentioned in the main text, we focus on the case where we set the coverage rate to the generic functional form

$$\gamma_j = 1 - n_j^{-\beta}.$$

We consider a set of β values in our sensitivity analysis. We conduct a numerical experiment over $\beta \in \{\frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\}$ in Figure 10. The resulting fidelity profiles are largely stable across these γ_j schedules, and we adopt $\beta = \frac{1}{3}$ in the main text for simplicity.

E. Simulation System and Output Functional Examples

E.1. Simulation System Examples

Manufacturing: Factory Production (discrete-event simulation; cycle time).

- Outcome space: $\mathcal{X} = \mathbb{R}_+$ (cycle time or throughput).
- Scenarios: ψ = product mix + scheduling policy.
- Profiles: $\mathcal{Z}$ = machine/operator states, shift team, lot sizes; $\mathcal{P}$ = plant variability.

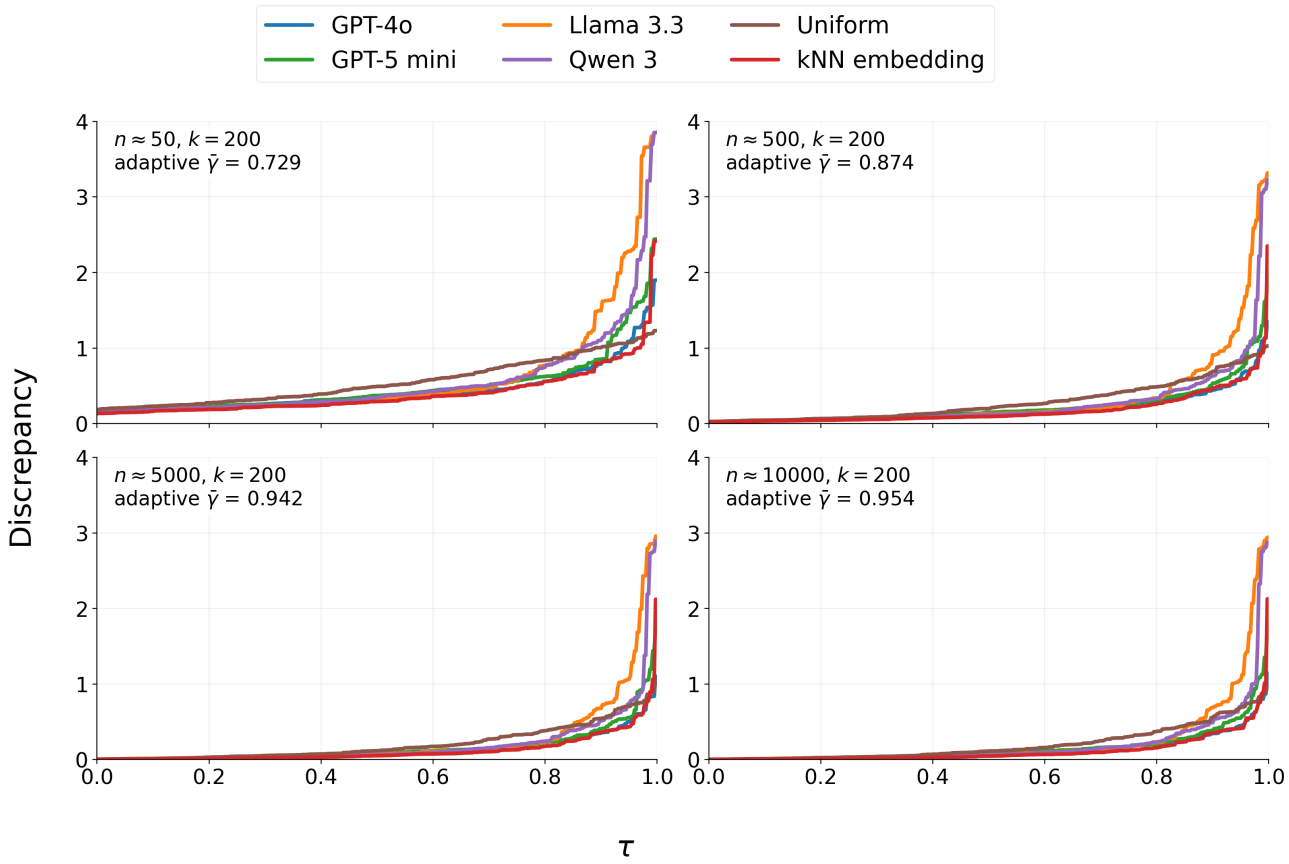


Figure 9. Robustness check of simulator performance under different human sample sizes. Panels correspond to $n_j = 50$ (top left), $n_j = 500$ (top right), $n_j = 5000$ (bottom left), and $n_j = 10000$ (bottom right), with $k = 200$ fixed. Colors match Figure 2.

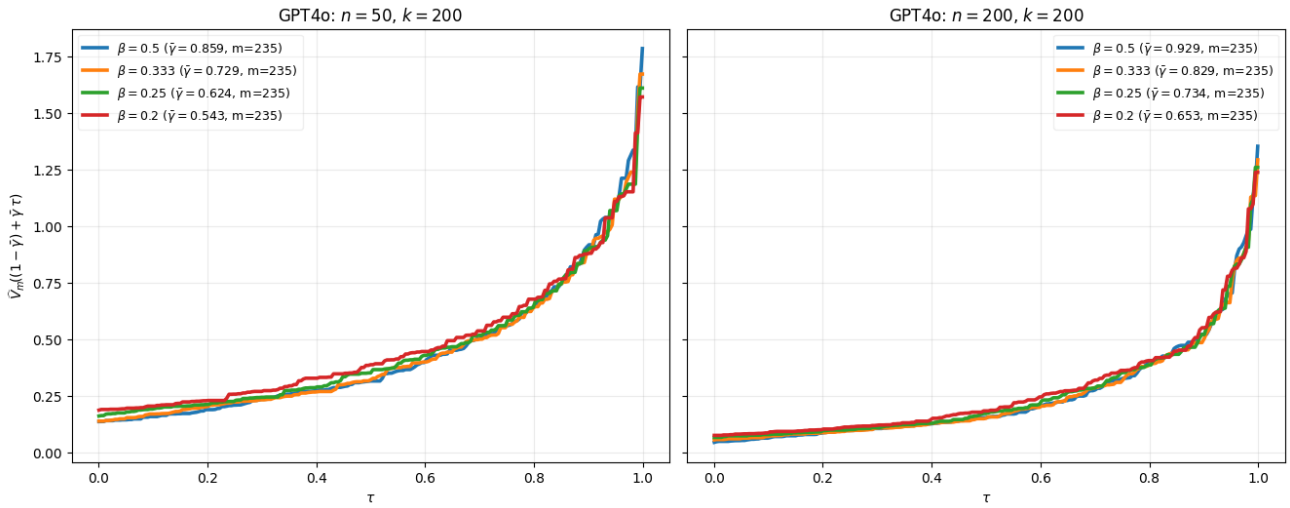


Figure 10. β sensitivtiy analysis under $n = \{50, 200\}$.

- Laws: $Q^{\mathrm{gt}}(\cdot \mid z, \psi)$ = empirical cycle-time distribution on the floor; $Q^{\mathrm{sim}}(\cdot \mid a(z), \psi, r)$ = DES output under mirrored inputs.
- Parameters: $\Theta = \mathbb{R}$ (mean cycle time) or $\mathbb{R}^2$ (mean, variance).

- Discrepancy: L = difference of means, Gaussian KL.
- Sampling: n_j production runs logged; k simulated replications per scenario.

Environment: Urban decarbonization (technology choice; multinomial).

- Outcome space: $\mathcal{X} = \{1, \ldots, K\}$ with mean $p(\psi) \in \Delta^{K-1}$ (for example, gas furnace, heat pump, variable refrigerant flow, other).
- Scenarios: ψ consists of city, season, rebate level, carbon price path, and policy bundle.
- Profiles: $\mathcal{Z}$ contains household and building attributes such as income, occupants, roof area, and baseline electricity use, or exogenous drivers including weather and demand shocks.
- Laws: $Q^{\mathrm{gt}}(\cdot \mid z, \psi)$ denotes the empirical technology-choice distribution, and $Q^{\mathrm{sim}}(\cdot \mid a(z), \psi, r)$ denotes the simulator output under mirrored inputs.
- Parameters: $\Theta = \Delta^{K-1}$ for category probabilities, or a low-dimensional reparameterization such as multinomial logistic parameters.
- Discrepancy: L on the simplex, for example the total-variation distance $\frac{1}{2}\|p - q\|_1$, the multiclass Kullback–Leibler divergence $\sum_{c=1}^{K} p_c \log\bigl(p_c/q_c\bigr)$.
- Sampling: n_j human records per scenario and k synthetic replications per scenario.

Building control: room-temperature response (stochastic emulator; short-horizon thermal dynamics).

- Outcome space: $\mathcal{X} = \mathbb{R}$, where the outcome is a short-horizon room-temperature response.
- Scenarios: ψ denotes a coarse operational context across which simulator fidelity is compared, for example building or unit identity together with hour-of-day, day-type, and an ambient-temperature regime.
- Profiles: $\mathcal{Z}$ collects the within-scenario thermal state and disturbance variables that vary even after ψ is fixed, such as current and lagged room temperature, solar irradiance, current heating setpoint, and other occupant- or disturbance-related variation.
- Laws: $Q^{\mathrm{gt}}(\cdot \mid z, \psi)$ denotes the real conditional law of future room temperature given the current profile z and operational context ψ. $Q^{\mathrm{sim}}(\cdot \mid a(z), \psi, r)$ denotes the emulator output under mirrored pre-decision inputs $a(z)$, context ψ, and simulator randomness r.
- Parameters: $\Theta = \mathbb{R}$ if the target is the scenario-level mean future room temperature, or $\Theta = \mathbb{R}^2$ if one summarizes the law by mean and variance; alternatively, one may target a small collection of temperature quantiles.
- Discrepancy: L may be the absolute difference of scenario-level means, a Wasserstein distance on the scalar room-temperature distribution, or a Gaussian Kullback–Leibler divergence when a parametric approximation is adopted.
- Sampling: n_j real windows observed in scenario j, each contributing one realized future room-temperature outcome; k emulator draws generated under the same scenario.

E.2. Additional Confidence Sets and Pseudo-Discrepancy Computations

The main text presents the multinomial case. We now give two additional examples illustrating the same recipe for scalar functionals and distribution-valued targets.

Example E.1 (Bounded Mean Functional)**.** Suppose $Y_{j,i} \in [a, b]$ are i.i.d. with population mean $p_j := \mathbb{E}[Y_{j,i}]$ for scenario j, and let

$$\hat{p}_j := \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{j,i}.$$

Then, by Hoeffding's inequality,

$$\mathcal{C}_j(\hat{p}_j, \gamma_j) := \left\{ u \in [a,b] : |u - \hat{p}_j| \le (b-a)\sqrt{\frac{\log(2/(1-\gamma_j))}{2n_j}} \right\}$$

is a γ_j-confidence set for p_j.

For the squared discrepancy $L(u, \hat{q}_j) = (u - \hat{q}_j)^2$, write

$$\varepsilon_j := (b-a)\sqrt{\frac{\log(2/(1-\gamma_j))}{2n_j}}, \qquad \ell_j := \max\{a, \hat{p}_j - \varepsilon_j\}, \qquad u_j := \min\{b, \hat{p}_j + \varepsilon_j\}.$$

Then the corresponding pseudo-discrepancy is available in closed form:

$$\hat{\Delta}_j = \sup_{u \in \mathcal{C}_j(\hat{p}_j, \gamma_j)} (u - \hat{q}_j)^2 = \max\{(\ell_j - \hat{q}_j)^2, (u_j - \hat{q}_j)^2\}.$$

Example E.2 (Distribution-Valued Target under W_1)**.** Suppose the scenario-level target is the full outcome distribution P_j, and samples are drawn as $Y_{j,1:n_j} \overset{\text{iid}}{\sim} P_j$. Assume P_j is σ-sub-Gaussian, and let $\widehat{P}_j$ denote the empirical distribution from the n_j samples. Define

$$\mathcal{C}_j^{W_1}(\widehat{P}_j, \gamma_j) := \left\{ Q : \; W_1(\widehat{P}_j, Q) \le r_j(n_j, \gamma_j, \sigma) \right\},$$

where

$$r_j(n_j, \gamma_j, \sigma) := \frac{512\sigma}{\sqrt{n_j}} + \sigma\sqrt{\frac{256e}{n_j} \log \frac{1}{1-\gamma_j}}.$$

Then $\mathcal{C}_j^{W_1}(\widehat{P}_j, \gamma_j)$ is a γ_j-confidence set for P_j by the concentration inequality of L.A. & Bhat (2022).

For the Wasserstein discrepancy $L(Q, \widehat{Q}_j) = W_1(Q, \widehat{Q}_j)$, the triangle inequality gives

$$\sup_{Q \in \mathcal{C}_j^{W_1}(\widehat{P}_j, \gamma_j)} W_1(Q, \widehat{Q}_j) \le W_1(\widehat{P}_j, \widehat{Q}_j) + r_j(n_j, \gamma_j, \sigma).$$

Thus one may use the conservative pseudo-discrepancy

$$\hat{\Delta}_j^{W_1} := W_1(\widehat{P}_j, \widehat{Q}_j) + r_j(n_j, \gamma_j, \sigma),$$

which preserves the calibration guarantee because it upper-bounds the worst-case discrepancy over the confidence set.

F. Additional Simulator Applications

F.1. EEDI Dataset

Our dataset is EEDI (He-Yueya et al., 2024), built on the NeurIPS 2020 Education Challenge (Wang et al., 2021), which consists of student responses to mathematics multiple-choice questions collected on the Eedi online education platform. The full corpus includes 573 distinct questions and 443,433 responses from 2,287 students, and each question has four options A-D that we binarize as "correct/incorrect" based on the student's or simulator's choice, consistent with Lemma B.4. We adopt the preprocessed version curated by (Huang et al., 2025), which retains questions with at least 100 student responses and excludes items with graphs or diagrams, yielding 412 questions. EEDI also provides individual-level covariates such as gender, age, and socioeconomic status, which the authors of (Huang et al., 2025) use to construct synthetic profiles. Under the same problem formulation, they compute $\{\hat{p}_j, \hat{q}_j\}_{j=1}^{412}$ for seven LLMs: GPT-3.5-TURBO (`gpt-3.5-turbo`), GPT-4O (`gpt-4o`), and GPT-4O-MINI (`gpt-4o-mini`); CLAUDE 3.5 HAIKU (`claude-3-5-haiku-20241022`); LLAMA 3.3 70B (`Llama-3.3-70B-Instruct-Turbo`); MISTRAL 7B (`Mistral-7B-Instruct-v0.3`); DEEPSEEK-V3 (`DeepSeek-V3`), and constructed a benchmark random simulator that selects uniformly among the available answer choices. A more detailed exploration into the EEDI dataset and the simulation procedure can be found in (Huang et al., 2025).

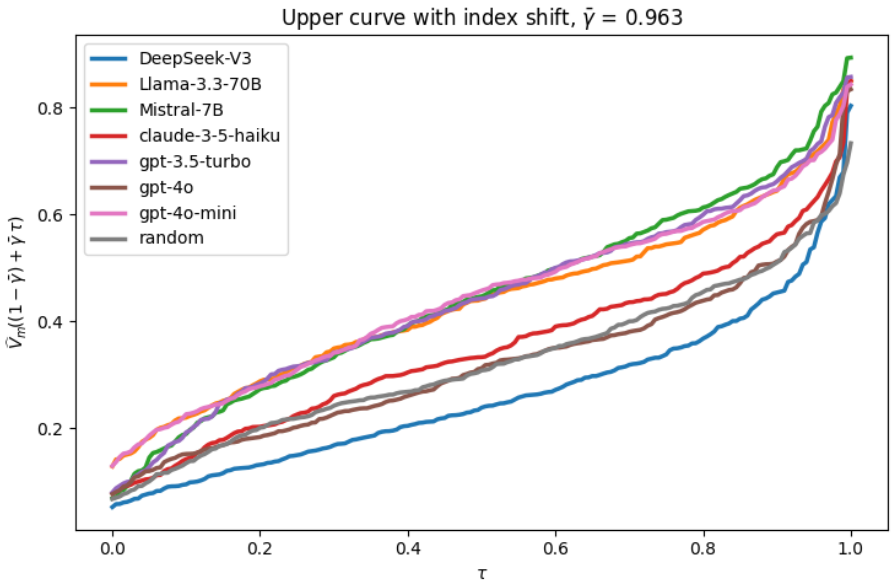


Figure 11. Quantile fidelity profiles $\hat{V}(\tau)$ across LLMs (Discrepancy: Absolute loss, $k = 50$).

We apply our methodology to produce a fidelity profile for each candidate LLM. We use absolute error as the loss, $L(p, q) = |p - q|$ and set the simulation budget $k = 50$. We set the adaptive coverage to be $\gamma_j = 1 - n_j^{-1/2}$ in this example.

Figure 11 compares models by how tightly their synthetic outcomes track the human distribution across items. We plot $\hat{V}_\ell(\tau)$ against τ, where lower-flatter curves indicate uniformly small discrepancies, while elbows reveal rare but severe misses. DEEPSEEK-V3 lies lowest across most quantiles, indicating the most reliable alignment, with the random benchmark and GPT-4O close behind. Notably, several models do not outperform the random baseline, suggesting they may be ill-suited for agent-based simulation under this discrepancy function.

F.2. OpinionQA

Our dataset is OpinionQA (Santurkar et al., 2023), built from the Pew Research's American Trends Panel, which consists of the US population's responses to survey questions spanning topics such as racial equity, security, and technology. We adopt the preprocessed version curated by (Huang et al., 2025), which includes 385 distinct questions and 1,476,868 responses from at least 32,864 people. Each question has 5 choices, corresponding to the order sentiments, which is a multinomial setting. We can construct confidence sets $\mathcal{C}_j$ for multinomial vectors by adopting Example 3.1 with $d = 5$. OpinionQA also provides individual-level covariates such as gender, age, socioeconomic status, religious affiliation, and marital status, and more, which are used to construct synthetic profiles. Under the same problem formulation, the authors of (Huang et al., 2025) compute $\{\hat{p}_j, \hat{q}_j\}_{j=1}^{385}$ for eight LLMs: GPT-3.5-TURBO (`gpt-3.5-turbo`), GPT-4O (`gpt-4o`), and GPT-4O-MINI (`gpt-4o-mini`); CLAUDE 3.5 HAIKU (`claude-3-5-haiku-20241022`); LLAMA 3.3 70B (`Llama-3.3-70B-Instruct-Turbo`); MISTRAL 7B (`Mistral-7B-Instruct-v0.3`); DEEPSEEK-V3 (`DeepSeek-V3`), and constructed a baseline random simulator that selects uniformly among the available answer choices. A more detailed exploration into the OpinionQA dataset and the simulation procedure can be found in (Huang et al., 2025). We also provide the same procedure onto a Bernoulli setting using the EEDI dataset, details can be found in Appendix F.

We apply our methodology to produce a fidelity profile for each candidate LLM. We use total variation as the discrepancy measure, $L(p, q) = \frac{1}{2}\|p - q\|_1$. In addition, we set the simulation budget $k = 100$, the result is presented in Figure 12.

Figure 12 compares models by how tightly their synthetic outcomes track the human distribution across items. We plot $\hat{V}_\ell(\tau)$ against τ, where lower-flatter curves indicate uniformly small discrepancies, while elbows reveal rare but severe misses. GPT-4O lies lowest across most quantiles, indicating the most reliable alignment. Notably, the simulator curves are steeper than the random benchmark, indicating question-dependent alignment and less uniform discrepancies. This suggests

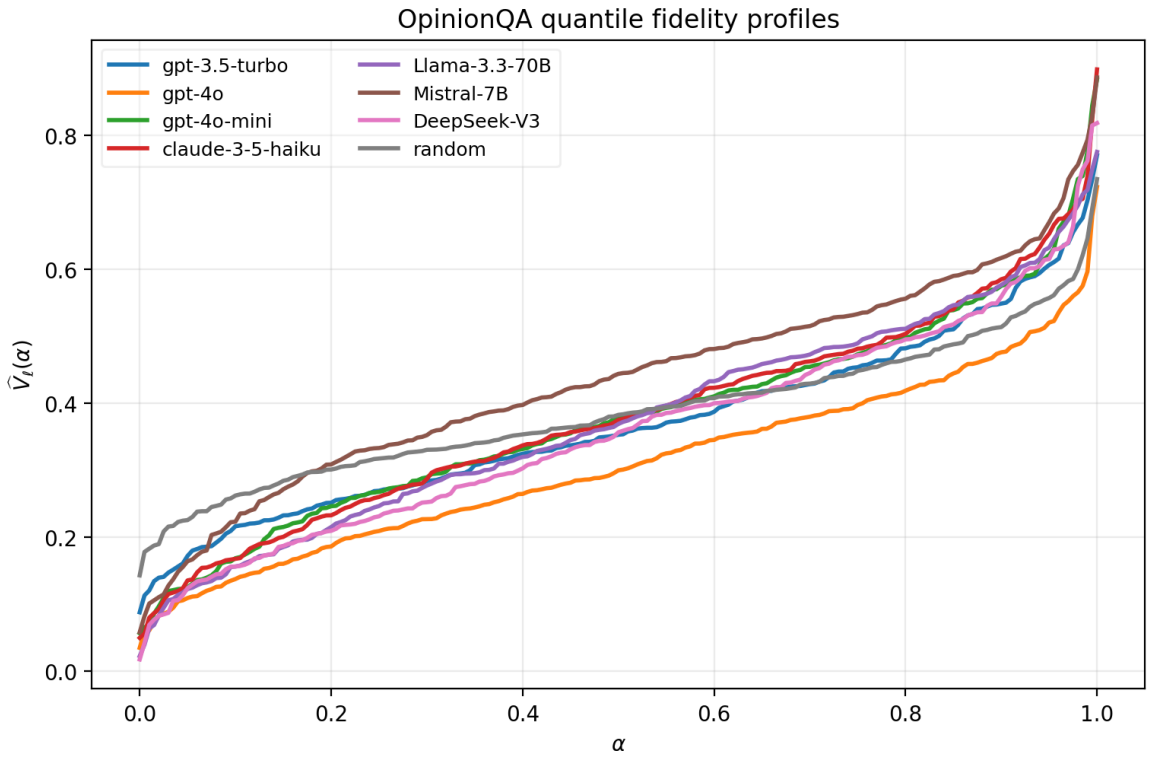


Figure 12. Quantile fidelity profiles $\hat{V}(\tau)$ across LLMs.

the simulators may require further fine-tuning to achieve more uniform discrepancy levels across this set of questions.

F.3. UMAR Dataset

We consider the UMAR unit of NEST (Next Evolution in Sustainable Building Technologies), a modular living-lab platform for building research operated by Empa (Richner et al., 2018), which is released by (Heer et al., 2024). The published dataset contains four years of one-minute measurements, from July 1, 2019 to June 30, 2023, and includes room temperatures, HVAC setpoints, valve and window states, ambient temperature, solar irradiation, and building-energy variables. In our reproduction, the UMAR portion is stored as four yearly wide-format CSV files together with a metadata table for the UMAR measurement channels.

Following the fixed experimental design used in our UMAR reproduction, we focus on room `temp_275` and aggregate the one-minute measurements to 30-minute means. We then construct a prediction table using the current room temperature, two lagged temperatures, hour of day, weekend indicator, month, ambient temperature, and, when available, room setpoint and solar irradiation as covariates. To avoid clear sensor artifacts and outliers, we remove rows for which `room_temp_t`, `lag1`, `lag2`, or the future response exceeds 30°C. This yields 66,922 cleaned 30-minute observations.

Unlike EEDI and OpinionQA, UMAR is not a multinomial-response setting. Here the unit of analysis is an operational scenario rather than a question. We define the scenario label as

$$\psi = (h, w, t_q),$$

where $h \in \{0, \ldots, 23\}$ is the hour of day, $w \in \{0, 1\}$ indicates whether the timestamp falls on a weekend, and $t_q \in \{0, 1, 2, 3\}$ is the quartile of the ambient temperature. This gives $m = 192$ scenarios in total. For each scenario ψ_j, the real-side parameter p_j is the variance of the next-step 30-minute room temperature among held-out UMAR observations in that scenario, while the simulator-side parameter q_j is the corresponding variance induced by emulator draws. Since this is a continuous-outcome problem, we construct the confidence set $\mathcal{C}_j$ for p_j using a bootstrap confidence interval for the scenario-level variance, with confidence level $\gamma_j = 1 - n_j^{-1/3}$, where n_j is the number of real observations in scenario j.

Since this is a continuous-outcome problem, we estimate the real-side and simulator-side scenario variances by the unbiased sample variances $\hat{p}_j$ and $\hat{q}_j$, and use the squared-loss discrepancy $L(p_j, q_j) = (p_j - q_j)^2$. For the real-side variance parameter p_j, we construct $\mathcal{C}_j = [\ell_j, u_j]$ as a percentile bootstrap confidence interval based on $B = 400$ bootstrap resamples of the real observations in scenario j. Writing $\hat{p}_j^{*(1)}, \ldots, \hat{p}_j^{*(B)}$ for the bootstrap sample variances and setting $\alpha_j = (1 - \gamma_j)/2$ with $\gamma_j = 1 - n_j^{-1/3}$, we define $\mathcal{C}_j = [\ell_j, u_j] = \left[Q_{\alpha_j}\left(\{\hat{p}_j^{*(b)}\}_{b=1}^B\right), Q_{1-\alpha_j}\left(\{\hat{p}_j^{*(b)}\}_{b=1}^B\right)\right]$. The scenario-level upper pseudo-discrepancy used in Figure 13 is then

$$\widehat{\Delta}_j^+ = \sup_{u \in \mathcal{C}_j} L(u, \hat{q}_j) = \max\{(\ell_j - \hat{q}_j)^2, (u_j - \hat{q}_j)^2\}.$$

We use this bootstrap interval as a practical approximation for variance uncertainty in the continuous-outcome UMAR setting, rather than as a finite-sample-exact confidence set.

We compare three emulator families using five-fold cross-fitting: linear regression model, decision tree, and multilayer perceptron. For each emulator and each scenario, we generate a balanced simulator budget of $k = 200$ draws. We then apply our quantile-fidelity procedure to the resulting scenario-level variance discrepancies.

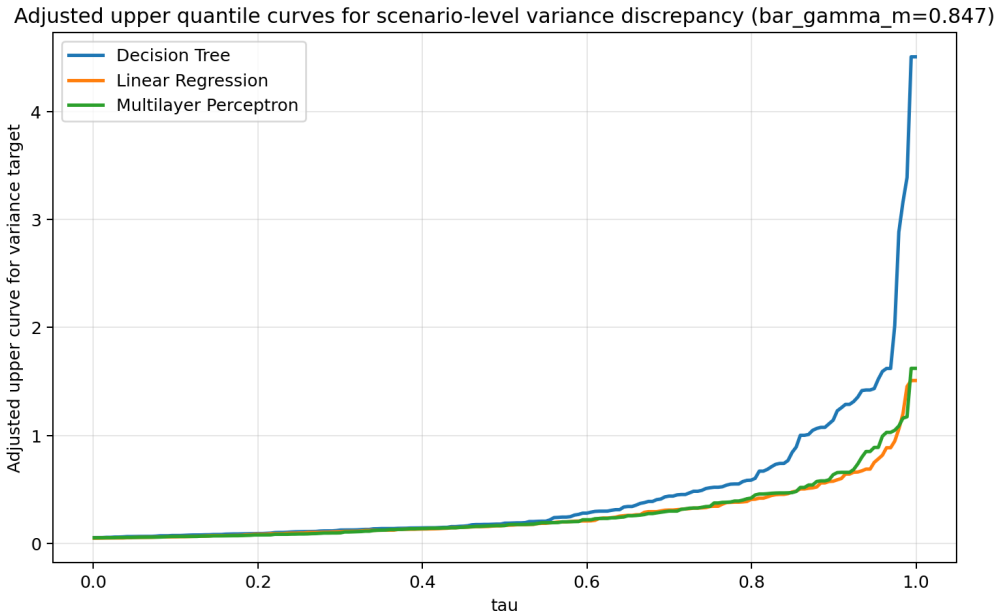


Figure 13. Adjusted upper quantile curves for scenario-level variance discrepancy across UMAR emulator families. Lower curves indicate closer agreement between emulator-side and real-side temperature variability.

Figure 13 reports the adjusted upper quantile curves for the UMAR experiment. Lower and flatter curves indicate that the emulator more faithfully reproduces the scenario-level variability observed in the real building data. In this reproduction, the linear regression model attains the lowest overall upper quantile curve, with the multilayer perceptron close behind, while the decision-tree baseline exhibits substantially larger upper-tail discrepancies. This illustrates that predictive accuracy alone is not the main object of interest here: the fidelity target is the distribution of scenario-level variability, not only pointwise temperature prediction error.

[1]Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, USA [2]**AUTHORERR: Missing \icmlaffiliation.** . Correspondence to: Yu-Shiou(Willy) Lin <yl5782@columbia.edu>.

Accepted to the Workshop on Statistical Frameworks for Uncertainty in Agentic Systems at ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).