Model-Free Assessment of Simulator Fidelity via Quantile Curves

Yu-Shiou Willy Lin[1] **Kaizheng Wang**[1,2] **Garud Iyengar**[1,2]

[1]Department of Industrial Engineering and Operations Research
[2]Data Science Institute
Columbia University*

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 systems 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.

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 [37, 22, 2, 1]. 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 [26, 21]. 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 [11, 10, 33, 27]. 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

*Email addresses: yl5782@columbia.edu, kaizheng.wang@columbia.edu, garud@ieor.columbia.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

LLM-based simulator might generate coherent individual responses while yielding systematically biased *population-level aggregates* [36], and recent work has identified distributional alignment as a key objective [7]. 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 guarantees 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* [38], constructed from the World Values Survey [13], 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 and Oberkampf [29], we distinguish *input* uncertainty from *output* uncertainty. Work on input uncertainty [8, 4, 20] 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 [18] and simulation-based inference [9, 25], 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 methods are also sample-based and model-agnostic [12, 31]. They can provide within-scenario discrepancy measures and confidence sets through suitable test inversion. Our target, however, is the distribution of discrepancies across scenarios, where each discrepancy involves a latent real-world parameter observed only through a finite batch of outcomes. Our contribution is, therefore, by adopting a two-level procedure: construct observable one-sided proxies within each scenario, then calibrate their quantiles across scenarios, yielding finite-sample fidelity profiles and inference for latent parameters in new scenarios.

A second line of work studies LLM-based human simulation and opinion alignment, asking whether synthetic responses reproduce human responses [30, 17]. 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 [34, 5], 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 [32, 6] 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 picked 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 $Q^{sim}(\cdot \mid \psi)$. [2] 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)
$\vdots$	$\cdots$	$\vdots$	$\cdots$
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

[2] 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.

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.

Remark 2.1 (Limited real-world samples). In general, we require $n_j > 1$ to avoid the single-observation identifiability issue in the absence of structural information linking scenarios. For example, consider independent observations $y^{\mathrm{rw}}_{j,1} \mid p_j \sim \mathrm{Bernoulli}(p_j)$. Taking $p_j \equiv 1/2$ or drawing each p_j independently from $\{0, 1\}$ with equal probabilities produces identical single-observation laws, despite different latent parameter distributions. [3]

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. The resulting notion of fidelity depends on the chosen target and discrepancy L; Appendix A.3 provides discussion on these choices.

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* $\tau \in (0, 1)$,

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

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

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

   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

[3] In data-scarce settings, possible remedies include pooling comparable scenarios to assess group-level fidelity or imposing a parametric model linking p_j to scenario covariates. These approaches entail a trade-off: pooling changes the target from individual scenarios to groups, while structural models introduce additional assumptions.

$\mathcal{C}_j(\cdot)$ 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

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

The confidence set results from a variant of Chernoff Bound (Lemma B.9). 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:

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

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:

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

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 $g : \mathbb{R}_+ \to (0,1)$. Two natural criteria for asymptotic tightness are $g(n) \to 1$ and $r(n, g(n)) \to 0$, where $r(n,\gamma)$ denotes the confidence-set radius. Section 4 presents a specific choice and numerical evidence illustrating the conservativeness of fixed γ and the benefits of adaptive coverage.[4] We are now ready to state the main theorem rigorously.

[4]Our average-coverage calibration relies on coverage levels being independent of the latent discrepancy ranks (scenario difficulty), as ensured by Assumption 3.2. For difficulty-dependent sample allocation, a possible remedy is to hold $\gamma_j \equiv \gamma$ constant and use additional samples to tighten C_j, provided conditional coverage and batch independence are maintained. Allowing coverage levels themselves to depend on difficulty requires additional assumptions or a different calibration, an extension beyond the scope of this paper.

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 $\tau \in (0,1)$,*

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

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 $S_\tau := \{j : \Delta_j \ge x_\tau\}$. By definition, $\frac{|S_\tau|}{m} \approx 1-\tau$.

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

(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 B.1. The proof of Theorem 3.4 then follows as a clean-index consequence of that sharper result and is provided in Appendix B.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, $\Delta^* := L(p,q)$ 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 B.3.

Remark 3.6. The discrepancy in Theorem 3.4 measures the sim-to-real gap. We can also extend this framework to the performance of simulators by comparing two simulators' discrepancies on the same scenarios, enabling paired statistical assessment of their relative fidelity. Complementing the marginal comparisons above, we present this extension in Appendix C and the experiments in Appendix D.

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 $\hat{V}_m^{\mathrm{cal}}(\tau) := \hat{V}_m\Big(\frac{1+\tau}{2}\Big)$, 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

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

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,

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

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 [38] and constructed from the World Values Survey [13]. 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.

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: [13].

Preprocessing. After data cleaning, we retain 235 distinct questions and responses from 96,220 individuals. Each 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 discuss the relationship between our sampling assumptions and this particular application in Remark D.1.

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.[5]. 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

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

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, [6] 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://github.com/yushiouwillylin/simulator_fidelity`.

4.2 Fidelity Profiling

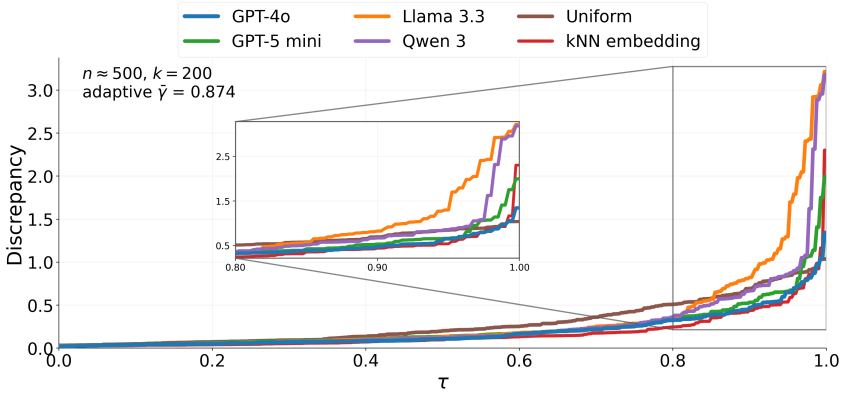


Figure 2: Calibrated discrepancy quantile curves $\hat{V}(\tau)$ across LLM simulators with zoom-in on upper 80-100 quantile.

Figure 2 plots the calibrated discrepancy quantile curves, which use the asymptotic index adjustment. For $m = 235$ and $\delta = 0.05$, the sharper finite-m calibration requires an additional index adjustment and retains a coverage correction of approximately 0.105; Appendix D.4 presents the corresponding non-asymptotic curves, and discuss their informative range and the endpoint convention. 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 and different query sample size k in Section D.5, D.6, and this pattern is stable.

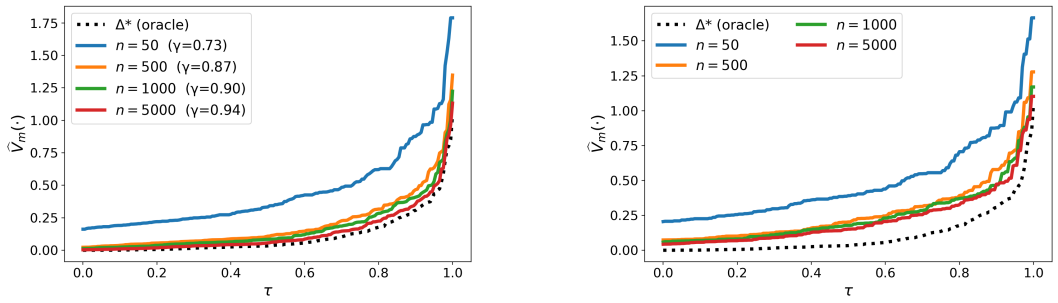
The two baselines serve as diagnostics rather than competing black-box simulators. The uniform baseline is not intended to be a poor estimator, since for near-balanced questions a uniform response distribution can have a small sim-to-real gap. Its role is to provide a structure-free benchmark without 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 generative simulators recover it.

[6] 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.7.

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*

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

where $\mathcal{C}_j(\hat{p}_j, \gamma) \subset \Theta$ *are* 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* $\tau \in (0, 1)$*,*

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

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

The formal finite-sample statement is given in Theorem B.4. Together, Theorems 3.4 and B.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 B.4 adds a lower envelope, yielding

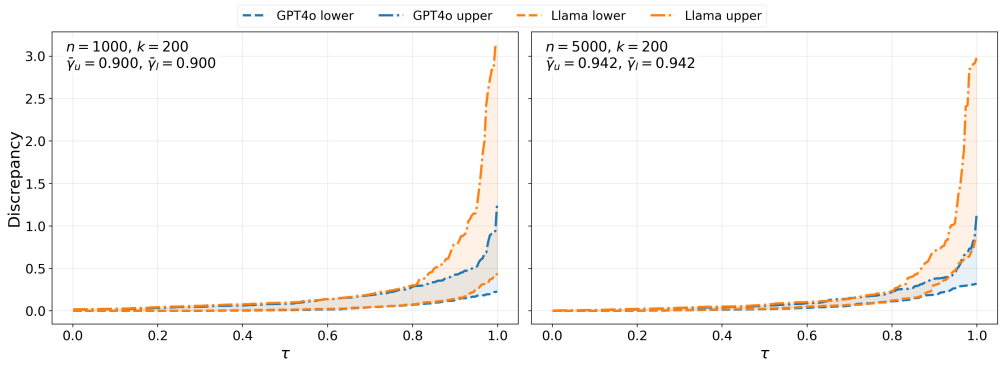


Figure 4: Confidence bands for GPT-4o and LLAMA 3.3. The left panel uses $n_j = 1000$, and $\bar{\gamma}_L = \bar{\gamma}_U = 0.9$; the right panel uses $n_j = 5000$, and $\bar{\gamma}_L = \bar{\gamma}_U = 0.94$, both with $k = 200$.

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.

Acknowledgments and Disclosure of Funding

The authors would like to thank Henry Lam for helpful discussions, as well as the anonymous reviewers for their valuable feedback. Yu-Shiou Willy Lin's research is supported by the Government Scholarship for Study Abroad from the Ministry of Education, Taiwan. Kaizheng Wang's research is supported by NSF grants DMS-2544147, DMS-2515679, and DMS-2210907. Garud Iyengar's research is supported by ONR grant N000142312374 and NSF grant EFMA-2132142. The authors declare no competing interests relevant to this work.

References

[1] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 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.

[2] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023. doi: 10.1017/pan.2023.2.

[3] Rémi Bardenet and Odalric-Ambrym Maillard. 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.

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

[5] Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6):1–34, 2021.

[6] Carlos E Budde, Arnd Hartmanns, Tobias Meggendorfer, Maximilian Weininger, and Patrick Wienhöft. Statistical model checking beyond means: Quantiles, cvar, and the dkw inequality (extended version). *arXiv preprint arXiv:2509.11859*, 2025.

[7] Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. arXiv preprint, 2025. URL https://arxiv.org/abs/2502.07068.

[8] Motong Chen, Henry Lam, and Zhenyuan Liu. Quantifying distributional input uncertainty via inflated kolmogorov-smirnov confidence band. *arXiv preprint arXiv:2403.09877*, 2024.

[9] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117.

[10] Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

[11] Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. 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.

[12] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

[13] Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, et al. World values survey: Round seven – country-pooled datafile (2017–2020), 2020. URL https://doi.org/10.14281/18241.1.

[14] Joy He-Yueya, Wanjing Anya Ma, Kanishk Gandhi, Benjamin W Domingue, Emma Brunskill, and Noah D Goodman. Psychometric alignment: Capturing human knowledge distributions via language models. *arXiv preprint arXiv:2407.15645*, 2024.

[15] Philipp Heer, Curdin Derungs, Benjamin Huber, Felix Bünning, Reto Fricker, Sascha Stoller, and Björn Niesen. 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.

[16] Wassily Hoeffding. 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.

[17] Chengpiao Huang, Yuhang Wu, and Kaizheng Wang. Uncertainty quantification for LLM-based survey simulations. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=nY1Ge2wxtP.

[18] Yongseok Jeon, Yi Chu, Raghu Pasupathy, and Sara Shashaani. Uncertainty quantification using simulation output: Batching as an inferential device, 2024. URL https://arxiv.org/abs/2311.04159.

[19] Prashanth L.A. and Sanjay P. Bhat. 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.

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

[21] Yuxuan Lu, Jing Huang, Yan Han, Bingsheng Yao, Sisong Bei, Jiri Gesi, Yaochen Xie, Qi He, Dakuo Wang, 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.

[22] C. Macal. 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.

[23] Jay Mardia, Jiantao Jiao, Ervin Tánczos, Robert D Nowak, and Tsachy Weissman. 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.

[24] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.

[25] George Papamakarios, David Sterratt, and Iain Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 837–848. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/papamakarios19a.html.

[26] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[27] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018.

[28] P. Richner, P. Heer, R. Largo, E. Marchesi, and E. Zimmermann. 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.

[29] Christopher J. Roy and William L. Oberkampf. 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.

[30] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[31] Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. Mmd aggregated two-sample test. *Journal of Machine Learning Research*, 24(194): 1–81, 2023. URL http://jmlr.org/papers/v24/21-1289.html.

[32] Jake C Snell, Thomas P Zollo, Zhun Deng, Toniann Pitassi, and Richard Zemel. Quantile risk control: A flexible framework for bounding the probability of high-loss predictions. *arXiv preprint arXiv:2212.13629*, 2022.

[33] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 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)*, pages 23–30. IEEE, 2017.

[34] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005. ISBN 0387001522.

[35] Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Jordan Zaykov, Jose Miguel Hernandez-Lobato, Richard E. Turner, Richard G. Baraniuk, Eedi Craig Barton, Simon Peyton Jones, Simon Woodhead, and Cheng Zhang. Results and insights from diagnostic questions: The neurips 2020 education challenge. In Hugo Jair Escalante and Katja Hofmann, editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 191–205. PMLR, 06–12 Dec 2021. URL `https://proceedings.mlr.press/v133/wang21a.html`.

[36] Sean J. Westwood. 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.

[37] Lin Zhang, Longfei Zhou, Lei Ren, and Yuanjun Laili. 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`.

[38] Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 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

Our target is the quantile function of $\Delta_\psi = L(p_\psi, \widehat{q}_\psi)$ across scenarios $\psi \sim \Psi$, where $\widehat{q}_\psi$ is computed using a fixed simulator query budget. Although the sample estimates $\widehat{p}_j$ and $\widehat{q}_j$ are observable, the real-world parameter p_j, and hence the discrepancy Δ_j, remains latent. For example, consider conditionally i.i.d. outcomes $y^{\mathrm{rw}}_{j,i} \mid \psi_j \sim \mathrm{Bernoulli}(p_j)$ and their sample mean $\widehat{p}_j = n_j^{-1} \sum_{i=1}^{n_j} y^{\mathrm{rw}}_{j,i}$. The outcome loss $(y^{\mathrm{rw}}_{j,i} - \widehat{q}_j)^2$ is observable, whereas the discrepancy $(p_j - \widehat{q}_j)^2$ is not. With fixed n_j and conditionally independent real and simulator batches, replacing p_j by $\widehat{p}_j$ yields

$$\mathbb{E}\left[(\widehat{p}_j - \widehat{q}_j)^2 \mid \psi_j, \widehat{q}_j, n_j\right] = (p_j - \widehat{q}_j)^2 + \frac{p_j(1-p_j)}{n_j}.$$

Thus, plug-in discrepancies incorporate sampling uncertainty that varies with n_j, and calibrating their quantiles does not directly calibrate the latent discrepancy distribution. Our two-level construction instead forms observable one-sided proxies using scenario-wise confidence sets and accounts for both within-scenario uncertainty and variation across the finitely many sampled scenarios.

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

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

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

A.3 Choice and Interpretation of the Discrepancy

Our framework does not claim to provide a universally optimal discrepancy. Instead, the target parameter and the discrepancy L determine which aspects of simulator fidelity are assessed, while the calibrated quantile curve characterizes how the resulting discrepancy $\Delta_\psi = L(p_\psi, \widehat{q}_\psi)$ varies across scenarios. The framework is therefore agnostic to the simulator's internal model, but the meaning of fidelity remains relative to the chosen target, outcome representation, and discrepancy.

Choosing a discrepancy. The choice should reflect the application-relevant errors. Comparing means is appropriate when the average outcome is the quantity of interest, whereas comparing probability vectors or full distributions can reveal differences that averaging conceals. Table 3 summarizes several choices. Each must be paired with confidence sets for the corresponding target and satisfy the stated conditions on the discrepancy.

B Theorems Proof

Theorem B.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 $[1-\tau]_m := m^{-1}\lceil m(1-\tau)\rceil$*

Table 3: Discrepancy choices and their interpretation.

Discrepancy	Primary sensitivity
Mean loss	Differences in the average outcome.
Wasserstein distance W_1	Distributional shifts measured by the distance over which probability mass is transported.
Total variation	Redistribution of probability mass across categories.
KL divergence	Outcomes to which the simulator assigns too little probability relative to the real distribution.
MMD or energy distance	Distributional differences emphasized by the chosen kernel or outcome-space distance.

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* $b_m(\delta) = \sqrt{\frac{1}{2m}\log\frac{3m}{\delta}}$*.*[7] *Then, with probability at least* $1-\delta$ *over* $\mathcal{D}$*, simultaneously for all* $\tau \in (0,1)$*,*

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

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* $\tau \in (0,1)$*.*

B.1 Proof for Theorem B.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

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

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

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

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

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

[7] 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 $\big(\bar{\gamma}_m - \varepsilon_{\text{al}}(\tau,\delta)\big)[1-\tau]_m - b_m(\delta)\sqrt{[1-\tau]_m} > 0$.

Set $\varepsilon_{\mathrm{DKW}}(\delta) := \sqrt{\frac{\log(6/\delta)}{2m}}$. By the DKW inequality (Lemma B.5), the event

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

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

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

Equivalently,

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

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 $r = 1, \dots, m-1$, 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 $(\gamma_1, \dots, \gamma_m)$ is fixed independently of the exchangeable vector $(\Delta_1, \dots, \Delta_m)$. 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 [3], 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.6 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 $\mathbb{P}(T_r \geq \sum_{j\in S_r}\gamma_j - c_m\sqrt{s_r} \mid \mathcal{G}) \geq 1-\delta_{\mathrm{cnt}}/m$. A union bound over $r = 1,\dots,m-1$ yields an event $\mathcal{E}_{\mathrm{cnt}}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{cnt}}) \geq 1-\delta_{\mathrm{cnt}}$ on which, for all $r = 1,\dots,m-1$,

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

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

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

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

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

If $\beta_r = 0$, then $1-\beta_r = 1$, and the desired comparison $\hat{V}_m(1-\beta_r) \geq \bar{V}_m(1-\alpha_r)$ 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 $s \geq r$,

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

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

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

As U_r is integer,

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

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.11, with $N = \lfloor m\beta_r\rfloor + 1$,

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

Since

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

this is exactly

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

Finally, work on $\mathcal{E}_{\mathrm{DKW}} \cap \mathcal{E}_{\mathrm{al}} \cap \mathcal{E}_{\mathrm{cnt}}$, 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 $1-\alpha-\varepsilon_{\mathrm{DKW}}(\delta) - \frac{1}{m} < 0$, 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 $\alpha_{\mathrm{eff}}(\alpha)$ through the rounded grid point, $\alpha_{\mathrm{eff}}(\alpha) = \alpha_{\mathrm{eff}}(\alpha_+)$. Applying (12) at this r gives

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

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

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

By monotonicity in the threshold,

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

Define

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

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

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

where

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

This is exactly (8).

□

B.2 Proof for Theorem 3.4

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

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

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 $\alpha_{\text{eff}}(\cdot)$,

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

Thus, to guarantee $\alpha_{\text{eff}}(\alpha') \ge \bar{\gamma}_m \alpha$, 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 $\alpha + \rho_m(\delta) \ge 1$, then $1 - \alpha - \varepsilon_m(\delta) - \rho_m(\delta) \le 0$, so the desired lower bound is trivial. Hence assume $\alpha + \rho_m(\delta) < 1$, and by applying Theorem B.1 at α' gives

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

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 $\alpha_{\text{eff}}(\alpha') \ge \bar{\gamma}_m\alpha$, and hence $1 - \alpha_{\text{eff}}(\alpha') \le 1 - \bar{\gamma}_m\alpha$. Since $\hat{V}_m(\cdot)$ is nondecreasing,

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

Therefore, by monotonicity of the event in the threshold,

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

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

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

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 B.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 B.2. *Adopt the setting of Theorem B.1 and fix $\gamma \in (0,1)$. Set $\gamma_j = \gamma$ for all $j \in [m]$. Then, for every $\eta \in (0,1)$, with probability at least $1-\eta$ over $\mathcal{D}$, simultaneously for all $\alpha \in (0,1)$,*

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

where

$$\varepsilon_\gamma(\alpha,m,\eta) := \sqrt{\frac{1}{2}\log\frac{6}{\eta}} + \frac{1}{\sqrt{m}} + \frac{1}{\gamma}\sqrt{2\log\frac{3m}{\eta}}.$$

The inner probability averages over a new scenario and its k simulator samples. For fixed γ and η, $\varepsilon_\gamma(\alpha,m,\eta) = O(\sqrt{\log m})$ uniformly over $\alpha \in (0,1)$.

Proof. Apply Theorem 3.4 with $\gamma_j = \gamma$ for all $j \in [m]$, $\delta = \eta$, and $\tau = 1-\alpha$. Then $\bar{\gamma}_m = \gamma$, so the quantile index becomes

$$\bar{\gamma}_m\tau + (1-\bar{\gamma}_m) = \gamma(1-\alpha) + (1-\gamma) = 1-\gamma\alpha.$$

The corresponding coverage remainder is

$$\sqrt{\frac{\log(6/\eta)}{2m}} + \frac{1}{m} + \frac{1}{\gamma}\sqrt{\frac{2\log(3m/\eta)}{m}} = \frac{\varepsilon_\gamma(\alpha,m,\eta)}{\sqrt{m}}.$$

Since the theorem holds simultaneously for all $\tau \in (0,1)$, the substitution $\tau = 1-\alpha$ gives (13) simultaneously for all $\alpha \in (0,1)$. □

B.3 Corollary for simulator budget:

Corollary B.3. *Suppose Assumptions 3.2 and 3.3 hold, allowing fixed, possibly heterogeneous simulator sample sizes $k_j \in \mathbb{N}$. Define the intrinsic 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 $C_j^p(\hat{p}_j), C_j^q(\hat{q}_j) \subset \Theta$ are nonempty compact confidence sets constructed from the corresponding real and simulated sample batches of scenario j, and satisfy

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

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

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

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta_\psi^\star \le \hat{V}_m^\star\big(1-\tfrac{\alpha}{2}\big)\,\Big|\mathcal{D}\Big) \ge 1-\alpha-\frac{\varepsilon(\alpha,m,\eta)}{\sqrt{m}}, \tag{14}$$

where

$$\varepsilon(\alpha,m,\eta) := \varepsilon_{1/2}(\alpha,m,\eta) = \sqrt{\frac{1}{2}\log\frac{6}{\eta}} + \frac{1}{\sqrt{m}} + 2\sqrt{2\log\frac{3m}{\eta}},$$

using the remainder defined in Corollary B.2. For fixed η, $\varepsilon(\alpha,m,\eta)/\sqrt{m} = O(\sqrt{\log m/m})$ uniformly over $\alpha \in (0,1)$.

Proof. Define

$$\mathcal{G}_j^\star := \sigma(\psi_j, n_j, k_j), \qquad \mathcal{G}^\star := \sigma(\mathcal{G}_1^\star, \dots, \mathcal{G}_m^\star), \qquad Y_j^\star := \mathbf{1}\{\hat{\Delta}_j^{\star,(k)} \ge \Delta_j^\star\}.$$

Both the real and simulated sampling batches remain random conditional on $\mathcal{G}^\star$. On the joint confidence event

$$\{p_j \in C_j^p(\hat{p}_j),\ q_j \in C_j^q(\hat{q}_j)\},$$

the definition of the supremum gives $\hat{\Delta}_j^{\star,(k)} \geq \Delta_j^\star$. Consequently, the assumed joint coverage and independence across scenarios imply

$$\mathbb{P}(Y_j^\star = 1 \mid \mathcal{G}^\star) \geq \frac{1}{2},$$

and the indicators $Y_1^\star, \ldots, Y_m^\star$ are conditionally independent.

Moreover, $\Delta_j^\star = L(p(\psi_j), q(\psi_j))$ are i.i.d., because the scenarios are i.i.d. The fixed sample sizes n_j and k_j do not enter these oracle discrepancies. Thus the oracle DKW bound and conditional counting argument in the proof of Theorem B.1 apply to $(\Delta_j^\star, \hat{\Delta}_j^{\star,(k)})$ with $\gamma_j \equiv \frac{1}{2}$. The affine-index reduction in the proof of Theorem 3.4 then gives (14), simultaneously for all $\alpha \in (0,1)$, with $\varepsilon(\alpha, m, \eta) = \varepsilon_{1/2}(\alpha, m, \eta)$. □

Theorem B.4. *Suppose the setup of Theorem B.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 $\varepsilon_{\mathrm{al}}^{\mathrm{bd}}(\alpha, \delta) := \sqrt{\frac{\log(6m/\delta)}{2m\tilde{\alpha}(\alpha)}}$, and $b_m^{\mathrm{bd}}(\delta) := \sqrt{\frac{1}{2m}\log\frac{6m}{\delta}}$. Define the effective indices

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

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

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

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

$$\hat{V}_m^-\big(\bar{\gamma}_{L,m}\tau\big) \lesssim V(\tau) \lesssim \hat{V}_m^+\big(\bar{\gamma}_{U,m}\tau + (1 - \bar{\gamma}_{U,m})\big) \tag{15}$$

B.4 Proof for Theorem B.4:

We prove the lower-band inequality with failure probability $\delta/2$. The upper-band inequality follows by applying the α-form of Theorem B.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 $\hat{\Delta}_j^- \le \Delta_j$. Therefore, by conditional coverage,

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

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

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

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

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

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

By construction, $U_r^- \ge T_r^-$. Since the coverage levels $\{\gamma_{L,j}\}_{j=1}^m$ are fixed design quantities and $(\Delta_1, \dots, \Delta_m)$ 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 $r = 1, \dots, m$, there is an event $\mathcal{E}_{\mathrm{al}}^-$ with probability at least $1 - \delta_{\mathrm{al}}$ on which, for all $r = 1, \dots, m$,

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

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 $r = 1, \dots, m$, there is an event $\mathcal{E}_{\mathrm{cnt}}^-$ with probability at least $1 - \delta_{\mathrm{cnt}}$ on which, for all $r = 1, \dots, m$,

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

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

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

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 $s \ge r$,

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

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

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

As U_r^- is integer, $U_r^- \ge \lceil m\beta_r^- \rceil$. 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,

$$\hat{V}_m^-\big(\alpha_{\mathrm{eff}}^L(\alpha_r)\big) \le \bar{V}_m(\alpha_r), \quad r = 1, \dots, m. \tag{16}$$

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 $1 - \delta_{\mathrm{DKW}}$,

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

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 (16) at this r gives

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

Therefore, by monotonicity and the strict-DKW event,

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

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

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

Equivalently,

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

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 B.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,

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

and the DKW remainder becomes

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

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

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

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

B.5 Technical Lemma

Lemma B.5. *(Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality via [24])*

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.6. *[Hoeffding (additive) via [16]]*

Let $Z_1, \dots, Z_n \in [0,1]$ 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.7. *[Chernoff–Hoeffding for one-parameter exponential family]*

Let $X_1, \dots, X_n$ 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. Write $\mu = A'(\theta)$. For $t \in A'(\Theta)$, let θ_t satisfy $A'(\theta_t) = t$ and define

$$I_\theta(t) := D(p_{\theta_t} \| p_\theta) = (\theta_t - \theta)t - A(\theta_t) + A(\theta).$$

The last equality follows directly from the exponential-family density.

For any λ such that $\theta + \lambda \in \Theta$, independence gives

$$\mathbb{E}_\theta\left[e^{\lambda n \bar{T}_n}\right] = \exp(n\{A(\theta+\lambda) - A(\theta)\}).$$

Suppose first that $t > \mu$. Since the mean map is strictly increasing, $\lambda_t := \theta_t - \theta > 0$. Markov's inequality therefore gives

$$\begin{aligned}
\mathbb{P}_\theta(\bar{T}_n \ge t) &\le e^{-n\lambda_t t}\mathbb{E}_\theta\left[e^{n\lambda_t \bar{T}_n}\right] \\
&= \exp(-n\{\lambda_t t - A(\theta+\lambda_t) + A(\theta)\}) \\
&= e^{-nI_\theta(t)}.
\end{aligned}$$

If $t < \mu$, then $\lambda_t < 0$, so the corresponding lower-tail event satisfies

$$\begin{aligned}
\mathbb{P}_\theta(\bar{T}_n \le t) &= \mathbb{P}_\theta\left(e^{n\lambda_t \bar{T}_n} \ge e^{n\lambda_t t}\right) \\
&\le e^{-n\lambda_t t}\mathbb{E}_\theta\left[e^{n\lambda_t \bar{T}_n}\right] \\
&= e^{-nI_\theta(t)}.
\end{aligned}$$

At $t = \mu$, both bounds reduce to the trivial bound 1.

Moreover,

$$I_\theta(\mu) = 0, \qquad I'_\theta(t) = \theta_t - \theta.$$

Thus I_θ is decreasing to the left of μ and increasing to the right of μ.

Fix $\varepsilon > 0$. On each side of μ where $I_\theta(t) > \varepsilon$ is nonempty, continuity and monotonicity give a threshold t_- or t_+ satisfying

$$t_- < \mu < t_+, \qquad I_\theta(t_-) = I_\theta(t_+) = \varepsilon,$$

with a threshold omitted if its corresponding tail is empty. Consequently,

$$\{I_\theta(\bar{T}_n) > \varepsilon\} \subseteq \{\bar{T}_n \le t_-\} \cup \{\bar{T}_n \ge t_+\},$$

again omitting any empty tail. Applying the appropriate one-sided bounds and the union bound yields

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

□

Lemma B.8. *(Chernoff-Hoeffding Inequality)*

Let $X_1, \dots, X_n \sim$ 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 $\varepsilon > 0$,

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

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

Proof. First suppose $\tilde{p} \in (0, 1)$. The Bernoulli family has canonical parameter

$$\theta = \log\frac{\tilde{p}}{1-\tilde{p}}, \qquad A(\theta) = \log(1 + e^\theta).$$

The one-sided bounds established in the proof of Lemma B.7 therefore give

$$\mathbb{P}(\bar{X}_n \ge t) \le e^{-nD(t\|\tilde{p})}, \qquad \tilde{p} < t < 1,$$

and

$$\mathbb{P}(\bar{X}_n \le t) \le e^{-nD(t\|\tilde{p})}, \qquad 0 < t < \tilde{p}.$$

These bounds also hold at the endpoints, since

$$\mathbb{P}(\bar{X}_n = 1) = \tilde{p}^n = e^{-nD(1\|\tilde{p})}, \qquad \mathbb{P}(\bar{X}_n = 0) = (1-\tilde{p})^n = e^{-nD(0\|\tilde{p})}.$$

The function $t \mapsto D(t\|\tilde{p})$ is continuous on $[0, 1]$, decreasing on $[0, \tilde{p}]$, and increasing on $[\tilde{p}, 1]$. Its ε-superlevel set therefore consists of at most two tails. Applying the corresponding one-sided bounds and the union bound gives

$$\mathbb{P}\big(D(\bar{X}_n\|\tilde{p}) > \varepsilon\big) \le 2e^{-n\varepsilon}.$$

If $\tilde{p} \in \{0, 1\}$, then $\bar{X}_n = \tilde{p}$ almost surely, so the divergence is zero and the conclusion is immediate. □

Lemma B.9 (Multinomial Chernoff-Hoeffding Bound via [23])**.** *For all $d \le (\frac{nC_0}{4})^{\frac{1}{3}}$ 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 $\epsilon > 0$,*

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

Lemma B.10 (Wasserstein distance bound via [19])**.** *Let X be a sub-Gaussian r.v. with parameter σ. Let F denote the CDF of X. Then, for every $n \ge 1$ and ε such that $\frac{512\sigma}{\sqrt{n}} < \varepsilon < \frac{512\sigma}{\sqrt{n}} + 16\sigma\sqrt{e}$, we have*

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

where e is Euler's number.

Lemma B.11 (Order-statistic thresholding)**.** *Let $x_1, \dots, x_m \in \mathbb{R}$ and let $x_{(1)} \le \cdots \le x_{(m)}$ be their order statistics. Fix a threshold $\tau \in \mathbb{R}$ and an integer $N \in \{1, \dots, m\}$. If at least N of the sample values are at least τ, i.e. $\big|\{j : x_j \ge \tau\}\big| \ge N$, then*

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

(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 $x_j \ge \tau$, contradicting $\big|\{j : x_j \ge \tau\}\big| \ge N$. Hence $x_{(m-N+1)} \ge \tau$.

C Paired Comparison of Simulator Fidelity

In Section 3.2, we have developed a distributional characterization of a single simulator's sim-to-real discrepancy performance. A natural next step is to ask: *Can we extend this viewpoint to compare two simulators, i.e., obtain a distributional characterization for pairwise performance?*

To that end, we mirror the construction in Section 3.2. Consider two simulators, denoted S_1, S_2 and their corresponding point estimates as $\hat{q}^{(1)}, \hat{q}^{(2)}$. For a given scenario ψ, define the performance discrepancy as

$$\delta_\psi := L(p_\psi, \hat{q}^{(1)}_\psi) - L(p_\psi, \hat{q}^{(2)}_\psi).$$

We are interested in making the inferential claim: *"S_1 is at least as good as S_2"*, where $\delta_\psi < 0$ means S_1 is closer to the ground truth than S_2 on scenario ψ.

The challenge, as in Section 3.2, is that p_ψ is unobserved and, with heterogeneous n_j, the direct plug-in discrepancies $L(\hat{p}_\psi, \hat{q}^{(1)}_\psi) - L(\hat{p}_\psi, \hat{q}^{(2)}_\psi)$ are non-i.i.d., making it harder to establish uniform claims across quantile levels. We therefore follow the same idea and work with a conservative, data-driven surrogate that constructs a confidence set, enabling a valid distributional comparison between S_1 and S_2. The methodology is as follows:

1. For each j, compute a confidence set $\mathcal{C}_j(\hat{p}_j)$ on p_j, and calculate the pseudo-performance discrepancy $\hat{\delta}_j := \sup_{u \in \mathcal{C}_j(\hat{p}_j)} \{L(u, \hat{q}^{(1)}_j) - L(u, \hat{q}^{(2)}_j)\}$.
2. Let $\hat{U}_m(v) = \hat{\delta}_{(\lceil mv \rceil)}$ for $v \in (0, 1)$, where $\hat{\delta}_{(r)}$ denotes the rth order statistic of $\{\hat{\delta}_j\}_{j=1}^m$.

In addition, the calibration data are $\mathcal{D} = \{(\psi_j, \hat{p}_j, \hat{q}^{(1)}_j, \hat{q}^{(2)}_j, n_j, k)\}_{j=1}^m$. With these, we present the following theorem:

Theorem C.1. *Suppose Assumptions 3.2 and 3.3 hold, with the two simulators' sample batches treated jointly as the simulator batch in Assumption 3.2. For any simulation sample size $k \in \mathbb{N}$, denote $\hat{q}^{(1)}_\psi, \hat{q}^{(2)}_\psi$ as point estimates of the two simulators for scenario ψ. Define the per-scenario performance discrepancy δ_j and the pseudo-performance discrepancy $\hat{\delta}_j$ by*

$$\delta_j := L\big(p_j, \hat{q}^{(1)}_j\big) - L\big(p_j, \hat{q}^{(2)}_j\big)$$
$$\hat{\delta}_j := \sup_{u \in C_j(\hat{p}_j, \gamma)} \Big[L\big(u, \hat{q}^{(1)}_j\big) - L\big(u, \hat{q}^{(2)}_j\big)\Big],$$

where confidence set $\mathcal{C}_j(\hat{p}_j, \gamma)$ satisfies $\mathbb{P}\big(p_j \in \mathcal{C}_j(\hat{p}_j, \gamma) \mid \psi_j, n_j\big) \ge \frac{1}{2}$. Then, for every $\eta \in (0, 1)$, with probability at least $1 - \eta$ over the calibration data $\mathcal{D}$, simultaneously for all $\alpha \in (0, 1)$,

$$\mathbb{P}_{\psi \sim \Psi}\Big(\delta_\psi \le \widehat{U}_m(1 - \tfrac{\alpha}{2}) \Big| \mathcal{D}\Big) \ge 1 - \alpha - \frac{\varepsilon(\alpha, m, \eta)}{\sqrt{m}},$$

where

$$\varepsilon(\alpha, m, \eta) := \sqrt{2\alpha \log \frac{2m}{\eta} + \frac{\big(\log \frac{2m}{\eta}\big)^2 + 4 \log \frac{2m}{\eta}}{m}} + \frac{\log \frac{2m}{\eta} + 2}{\sqrt{m}} + \sqrt{\frac{\log(4/\eta)}{2}}.$$

The inner probability averages over a new scenario and the corresponding samples from both simulators. In particular, the remainder is $O\big(\sqrt{(\log m)/m}\big)$, which vanishes as $m \to \infty$ for any α.

The proof of Theorem C.1 mirrors the argument for Theorem B.1, where we construct analogous pseudo-discrepancies via confidence sets, and then apply the same quantile-bounding machinery. Intuitively speaking, Theorem C.1 says that: If the data-driven performance discrepancy quantile $\widehat{U}_m(1 - \frac{\bar{\alpha}}{2}) \leq 0$ for some $\bar{\alpha}$, then – up to an $O\big(\sqrt{(\log m)/m}\big)$ adjustment term – S_1 is at least as good as S_2 on at least a $(1 - \bar{\alpha})$ fraction of scenarios with high probability. Finally, for simplicity, we state the result for fixed coverage $\gamma_j \equiv 1/2$. The extension to sample-size-adaptive coverage $\gamma_j = g(n_j)$ follows directly under the same assumptions: construct the paired pseudo-discrepancies using $C_j(\hat{p}_j, \gamma_j)$ and replace the index $1 - \alpha/2$ by $1 - \bar{\gamma}_m\alpha$, with the same finite-sample remainder as in Theorem 3.4.

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.

Remark D.1. We document some practical sampling issues common in statistical applications: the retained questions come from a curated survey rather than an explicit i.i.d. sampling design over future questions, and shared respondents may induce cross-question dependence, which violates Assumption 3.2. The curves nevertheless provide empirical fidelity profiles for the evaluated questions and a proxy for population-level fidelity, while their formal coverage guarantees rely on the stated sampling assumptions, commonly imposed in predictive inference. A practical remedy is to define the target population within prespecified question categories and respondent strata and apply the calibration separately within each stratum under a conditional i.i.d. design satisfying the stated sampling and confidence-set assumptions.

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.[8]

[8]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.

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.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 4.

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.

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 Finite-m Calibration and Extreme-Tail Behavior

The fidelity profiles in this paper use the asymptotic index adjustment $\tau \mapsto \bar{\gamma}_m \tau + (1 - \bar{\gamma}_m)$ for readability. Here, we present the sharper finite-m calibration, keeping the samples, confidence sets, and pseudo-discrepancies unchanged to show finite-m's effect on index adjustment. Notice that the

Table 4: 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[9]
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)

sharper guarantee still includes a coverage correction: under the stated assumptions, the threshold at plotted level τ has population coverage of at least $\tau - \varepsilon_m(\delta)$, with probability at least $1 - \delta$ over the calibration data. At the present settings, $\varepsilon_m(\delta) \approx 0.1052$, so $\tau = 0.80$ certifies coverage of at least approximately 69.5%, rather than 80%. With this qualification understood, we leave the coverage correction implicit in the following discussion and focus on how the finite-m index adjustment changes the plotted curves.

Finite-sample interpretation. We follow the same plotting convention as the main text, changing only the quantile-index mapping from $\bar{\gamma}_m\tau + (1-\bar{\gamma}_m)$ in (4) to $\tau_{\text{eff}}(\tau)$ in Theorem B.1. For $m = 235$, $\delta = 0.05$, and $\bar{\gamma}_m \approx 0.874$, the two mappings give:

Quantile level τ	Asymptotic index	Finite-m index
0.50	0.5630	0.7632
0.80	0.8252	0.9527

For example, at $\tau = 0.80$, the asymptotic curve evaluates $\widehat{V}_m$ at 0.8252, whereas the finite-m curve evaluates it at 0.9527. The latter therefore selects a higher empirical quantile and produces a more conservative discrepancy threshold. Both characterize the same population discrepancy distribution, up to their respective finite-sample remainders. The sharper calibration reduces the coverage correction from approximately 0.4314 in Theorem 3.4 to 0.1052 in Theorem B.1, whose guarantee is coverage of at least $\tau - 0.1052$ with 95% confidence under the stated assumptions.

Informative range. At these settings, the finite-m mapping reaches its endpoint at approximately $\tau = 0.8936$. Together with the coverage correction, this gives positive, data-dependent coverage bounds over approximately $0.1052 < \tau < 0.8936$. Beyond the upper cutoff, we use the deterministic squared-loss bound 4, as explained below. Increasing the number of scenarios reduces the finite-m correction and extends this range; additional human observations can tighten the scenario-wise confidence sets but do not replace the need for more scenarios.

Endpoint convention. When the effective tail mass vanishes, Theorem B.1 uses the extended endpoint $+\infty$. Here, both means lie in $[-1, 1]$, so squared loss is bounded by 4, which we use instead. The resulting plateau is a conservative fallback, not evidence of identical tail performance. Without a known bound, the largest observed discrepancy alone cannot certify the population maximum.

Empirical comparison. The finite-m adjustment raises the thresholds and moves the sharp upper-tail increases to lower plotted levels. Before the common plateau, Llama 3.3 and Qwen 3 retain

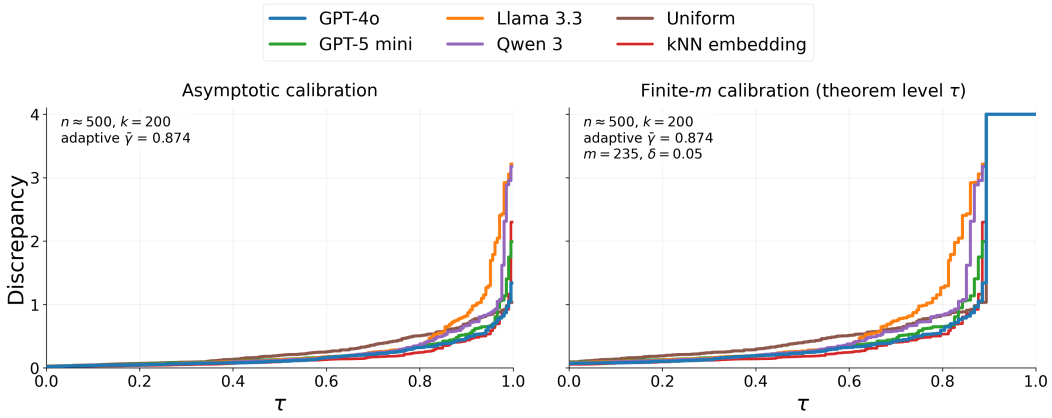


Figure 9: Asymptotic and finite-m calibration of squared-mean-loss fidelity profiles. The left panel uses $\widehat{V}_m(\bar{\gamma}_m\tau + 1 - \bar{\gamma}_m)$. The right panel uses $\widehat{V}_m(\tau_{\text{eff}}(\tau))$ with $m = 235$ and $\delta = 0.05$ per curve. When the effective tail mass vanishes, the curves use the common deterministic bound 4.

substantially larger upper-tail values than GPT-4o and GPT-5-mini. Thus, the broad differences remain visible, although the bounds no longer distinguish simulators once they reach 4.

Implications for tail summaries. Finite-sample upper bounds on the mean discrepancy and CVaR are obtained by the empirical curve $\hat{V}_m(\tau)$, including its support-bound fallback. Without a known upper bound, an unresolved tail assigned $+\infty$ makes these integrated bounds infinite.

D.5 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 10, which shows performance of the LLMs for $n_j = \{50, 500, 5000, 10000\}$. The results indicate that the dominance relation between LLMs remains is robust.

D.6 Simulator Query-Budget Sensitivity

We examine whether the fidelity profiles are sensitive to the simulator query budget by varying $k \in \{25, 50, 100, 200, 400\}$. We hold the human batch and question set fixed and use nested subsets of retained simulator responses to recompute $\hat{q}_j^{(k)}$ and the corresponding squared-mean-loss fidelity profiles. Figure 11 shows that the curves largely overlap across the tested budgets, and the overall profile shapes remain stable, which suggests that the broad fidelity patterns are not driven by the particular choice $k = 200$, although individual extreme-tail estimates remain sensitive to the retained simulator samples. As discussed in Remark 3.5, this analysis concerns the finite-query discrepancy $L(p_j, \hat{q}_j^{(k)})$, rather than the intrinsic discrepancy $L(p_j, q_j)$.

D.7 Adaptive coverage rate γ_j choice

For bounded outcomes in $[a, b]$, the Hoeffding confidence-set radius is

$$r(n, \gamma) = (b - a)\sqrt{\log(2/(1 - \gamma))/(2n)},$$

so $r(n, g(n)) \to 0$ whenever $\log(1/(1 - g(n))) = o(n)$. Therefore, as mentioned in the main text, we focus on the generic functional form

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

This polynomial schedule ensures the radius approach 0 and $g(n) \to 1$ for every fixed $\beta > 0$. In contrast, other functional forms such as constant coverage γ leaves persistent quantile-index inflation and $g(n) = 1 - e^{-cn}$ with $c > 0$ leaves a nonvanishing radius.

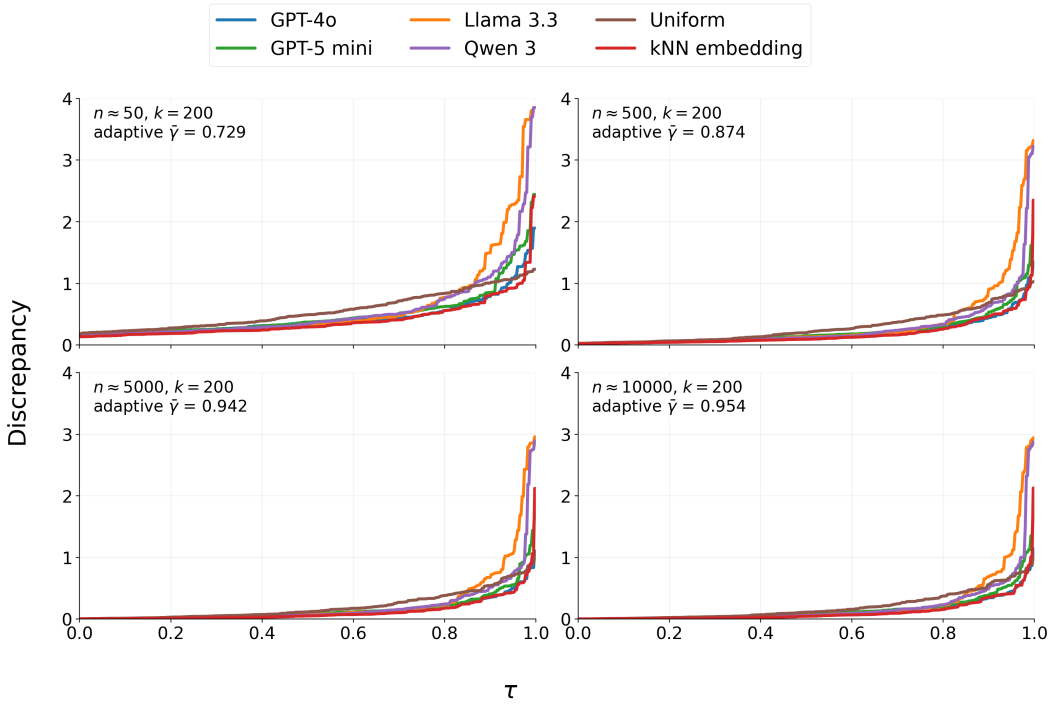


Figure 10: Robustness check of simulator performance under different human sample sizes. Panels correspond to $n_j = 50$ (top left), $n_j = 500$ (top right), $n_j = 5000$ (bottom left), and $n_j = 10000$ (bottom right), with $k = 200$ fixed. Colors match Figure 2.

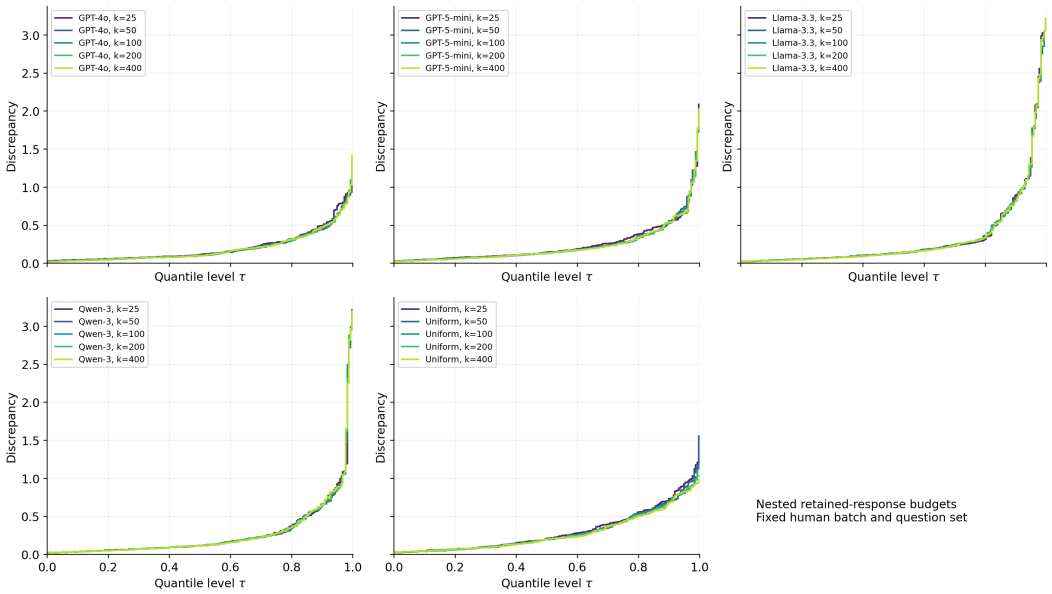


Figure 11: Sensitivity of squared-mean-loss fidelity profiles to the simulator query budget. Each panel corresponds to an LLM simulator or the uniform baseline, with colors indicating $k \in \{25, 50, 100, 200, 400\}$ retained responses per question. The simulator response subsets are nested, while the human batch and question set are held fixed. Profiles largely overlap across budgets, with the most visible differences near the extreme upper tail.

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 12. 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.

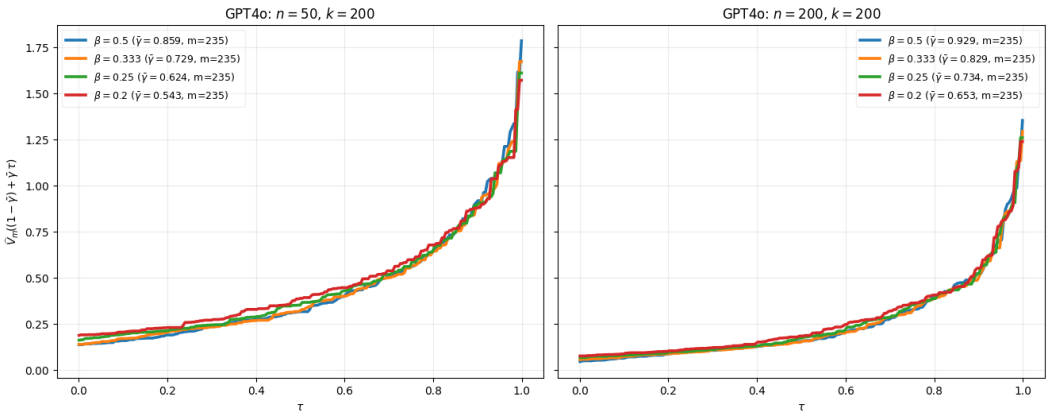


Figure 12: β sensitivtiy analysis under $n = \{50, 200\}$.

D.8 Inference for Unseen Survey Questions

We illustrate how a calibrated fidelity curve can be used to infer human population means for questions not used in calibration, as we mentioned in Section 3.3. We split the 235 retained survey questions into 200 calibration questions and 35 held-out questions, using the same split for all four simulators. On the calibration questions, we follow the main experimental setup: a human subsampling budget of 500 respondents per question, a simulator budget of $k = 200$, and coverage levels $\gamma_j = 1 - n_j^{-1/3}$ based on the effective human sample sizes.

Interval construction. For each simulator, we construct the empirical pseudo-discrepancy quantile curve using only the $m_{\text{cal}} = 200$ calibration questions. We apply the finite-m calibration with outer failure probability $\delta = 0.05$, recomputing the average coverage over this calibration subset. At target coverage τ, the upper discrepancy threshold is

$$U_{\text{cal}}(\tau) = \widehat{V}_{m_{\text{cal}}}\left(\tau_{\text{eff}}\left(\tau + \varepsilon_{m_{\text{cal}}}(\delta)\right)\right),$$

where the effective index and coverage remainder are evaluated using the calibration subset. When $\tau + \varepsilon_{m_{\text{cal}}}(\delta) \geq 1$ or the effective tail mass vanishes, we use $U_{\text{cal}}(\tau) = 4$, the deterministic upper bound on squared loss between means in $[-1, 1]$.

For a held-out question h, its simulator estimate $\widehat{q}_h$ then gives the interval

$$S_h(\tau) = \left\{p \in [-1,1] : (p - \widehat{q}_h)^2 \leq U_{\text{cal}}(\tau)\right\} = \left[\widehat{q}_h - \sqrt{U_{\text{cal}}(\tau)}, \widehat{q}_h + \sqrt{U_{\text{cal}}(\tau)}\right] \cap [-1,1].$$

Thus, one calibrated curve per simulator supplies intervals for all held-out questions and coverage levels; no additional human observations are required to construct an interval for a new question.

Evaluation. For each held-out question, we compute an oracle mean p_h^{ref} using all available human responses, following the evaluation in the tightness analysis. These estimates are used only to evaluate whether $p_h^{\text{ref}} \in S_h(\tau)$; they are not treated as directly observed population parameters. Table 5 reports the number and percentage of covered reference means and the mean length of the clipped intervals.

Results and interpretation. Reference coverage exceeds the target at every evaluated level, consistent with the one-sided nature and conservativeness of the method. For example, at $\tau = 0.50$, observed reference coverage ranges from 82.9% to 91.4%; the target specifies a coverage lower bound, not an exact coverage rate. Across the evaluated levels from 0.50 through 0.75, GPT-4o produces the smallest interval widths and Llama-3.3 the largest. At $\tau = 0.75$, all four simulators cover all 35 reference means, while their mean widths range from 1.571 to 1.967. The width comparison therefore distinguishes the precision of the intervals even when their observed coverage is identical.

At $\tau \in \{0.80, 0.90, 0.95\}$, the finite-sample calibration uses the support bound, yielding width-2 intervals and automatic coverage of every reference mean. These rows reflect the full-range fallback rather than additional evidence of simulator accuracy. Overall, the example demonstrates the reuse

Table 5: Inference for 35 held-out survey questions. Coverage reports the number of covered reference means with percentages in parentheses; width is the mean interval length on the response scale $[-1, 1]$.

	GPT-4o		GPT-5-mini		Llama-3.3		Qwen-3	
Target τ	Coverage	Width	Coverage	Width	Coverage	Width	Coverage	Width
0.50	32 (91.4)	1.009	32 (91.4)	1.080	29 (82.9)	1.220	31 (88.6)	1.130
0.60	32 (91.4)	1.163	34 (97.1)	1.231	32 (91.4)	1.498	32 (91.4)	1.327
0.70	33 (94.3)	1.324	34 (97.1)	1.452	34 (97.1)	1.875	32 (91.4)	1.473
0.75	35 (100.0)	1.571	35 (100.0)	1.817	35 (100.0)	1.967	35 (100.0)	1.940
$\{0.80, 0.90, 0.95\}^\dagger$	35 (100.0)	2.000	35 (100.0)	2.000	35 (100.0)	2.000	35 (100.0)	2.000

All coverage counts have denominator 35. †The final row reports identical results at each of the three listed levels, not pooled results: the deterministic support fallback gives the full interval $[-1, 1]$ for every held-out question.

of a single calibrated curve for inference across unseen questions and confidence levels, while the reported widths distinguish informative intervals from the full parameter range.

D.9 Paired LLM Simulator Fidelity

To complement the marginal fidelity profiles, we apply the paired comparison framework in Appendix C to compare LLM simulators on the same survey questions. For each pair S_1, S_2, we consider the signed difference $\delta_j = (p_j - \hat{q}_j^{(1)})^2 - (p_j - \hat{q}_j^{(2)})^2$, where negative values favor S_1 and positive values favor S_2. Using a common human-side confidence set for each question, we construct lower and upper paired pseudo-discrepancies and calibrate their quantiles across questions. This characterizes the distribution of within-question performance differences, rather than comparing the simulators' marginal quantile curves separately.

We compare GPT-4o (S_1) and Llama-3.3 (S_2) across the 235 WorldValueBench questions, using the same response encoding in $[-1, 1]$ and $k = 200$ simulator responses per model and question. We consider human subsample sizes $n = 1{,}000$ and $n = 5{,}000$, constructing the human-side confidence sets with coverage $\gamma_j = 1 - n_j^{-1/3}$. Figure 13 displays the paired curves using the asymptotic quantile-index adjustments. In addition, although we did not include the confidence band for the paired case in Appendix C, it is straightforward to make this extension, hence we directly plot the confidence band. We also present the paired square-loss mean difference of all available human responses as the dotted curve, which serves as a oracle ground-truth curve as we did in the tightness analysis.

The paired profiles reveal heterogeneous relative performance. At low quantile levels, both curves lie below zero, indicating questions on which GPT-4o has smaller squared loss; at high quantile levels, both lie above zero, indicating questions favoring Llama-3.3. The full-human reference exhibits a much larger negative tail than positive tail, suggesting that the largest observed advantages of GPT-4o are substantially greater in magnitude than those of Llama-3.3. Around the median, however, the differences are close to zero and the interval between the curves includes zero. Increasing the human sample size narrows this interval while preserving the qualitative pattern. Thus, the paired analysis suggests question-dependent advantages rather than uniform superiority of either simulator. These interpretations concern the signed-difference distribution and the displayed asymptotic curves, not a finite-sample dominance claim.

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.

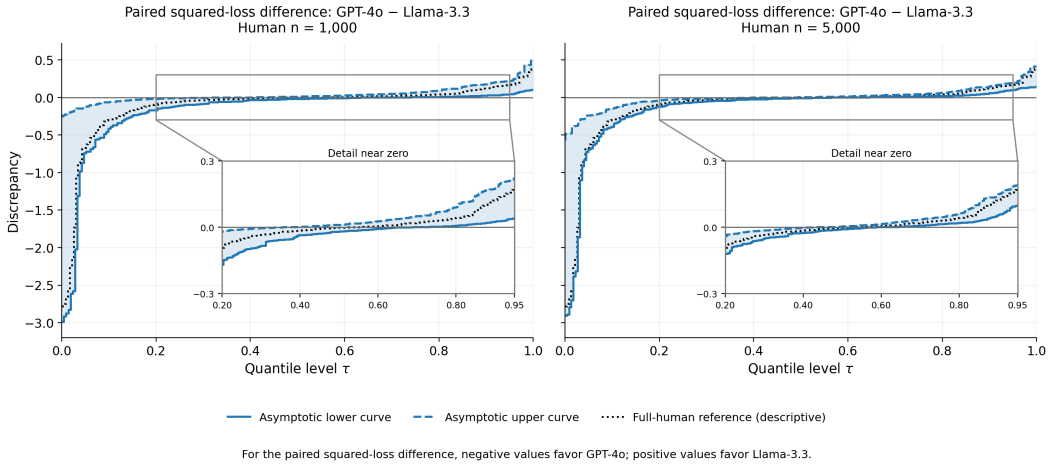


Figure 13: Paired squared-loss differences between GPT-4o and Llama-3.3, with human subsample sizes $n = 1{,}000$ (left) and $n = 5{,}000$ (right). Solid and dashed blue lines denote the asymptotically calibrated lower and upper quantile curves; shading marks the interval between them. The dotted black line is the descriptive full-human reference. Negative values favor GPT-4o, while positive values favor Llama-3.3. Insets show differences near zero.

- Profiles: $\mathcal{Z}$ = machine/operator states, shift team, lot sizes; $\mathcal{P}$ = plant variability.
- 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. and Bhat [19].

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 [14], built on the NeurIPS 2020 Education Challenge [35], 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.8. We adopt the preprocessed version curated by [17], 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 [17] 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 [17].

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.

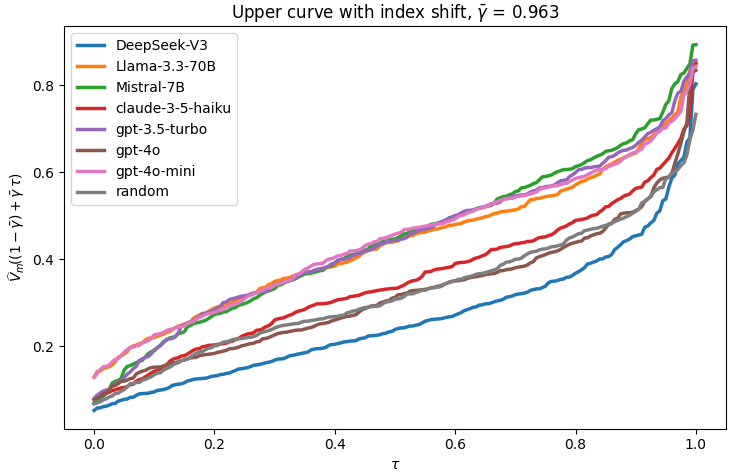


Figure 14: Quantile fidelity profiles $\hat{V}(\tau)$ across LLMs (Discrepancy: Absolute loss, $k = 50$).

Figure 14 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 [30], 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 [17], 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 [17] compute $\{\hat{p}_j, \hat{q}_j\}_{j=1}^{385}$ 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 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 [17].

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 15.

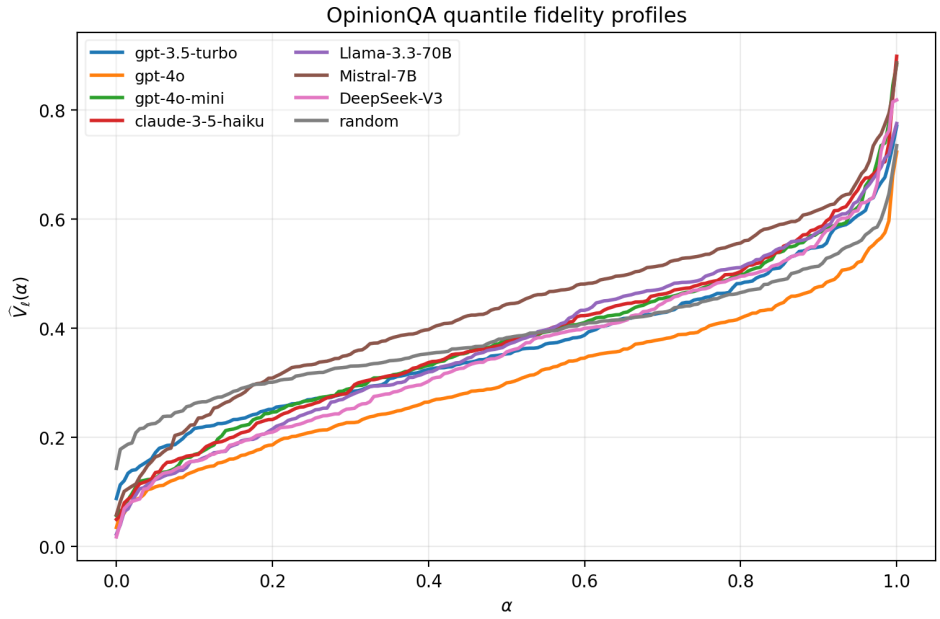


Figure 15: Quantile fidelity profiles $\hat{V}(\tau)$ across LLMs.

Figure 15 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 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 [28], which is released by [15]. 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 16 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.

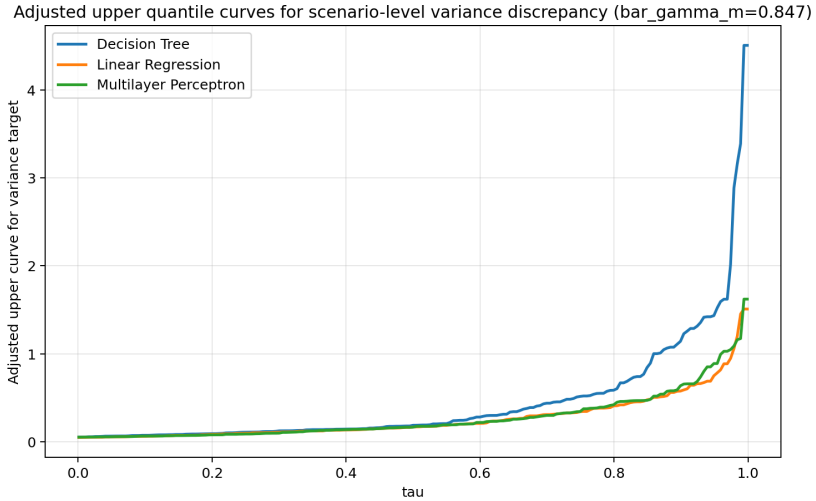


Figure 16: Adjusted upper quantile curves for scenario-level variance discrepancy across UMAR emulator families. Lower curves indicate closer agreement between emulator-side and real-side temperature variability.

Figure 16 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, all the claims we made are backed by proofs and experiment implementation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, we discuss our limitation in the Discussion section at the end of the paper.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Yes, the full proof and technical lemma required are in Appendix, while proof sketch is in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we provide all such details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we provide code through anonymous github and zip file.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, some in the main text and most in the Appendix. We also include supplementary materials for rerun of our method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The main empirical objects are calibrated quantile curves with finite-sample guarantees. We also report data-driven confidence bands and robustness checks over human sample sizes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix and the anonymized repository describe the API model calls, simulator budgets, prompts, decoding parameters, and scripts used to reproduce the figures. The post-processing and quantile-curve construction run on CPU; the main computational cost is the external LLM/API generation step.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we have thoruoghly checked alignment with Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss limitations and potential risks in the Final Section. Our method can help detect simulator failures before deployment, but fidelity profiles could also be misused to over-legitimize LLM-based human simulation in sensitive social settings if the assumptions, preprocessing choices, or tail failures are ignored.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We did not use any scraped dataset and only adopted public survey dataset, which we do not know of any risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cited the World Value Survey and the curated version World Value Bench, and respected their license terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release anonymized code, processed data files, and reproduction scripts. Documentation, dependency information, and license/access notes are provided in the anonymized repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The human survey data we used is public dataset, where the collection procedure is documented in the dataset's website and we explicitly cited the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The human survey data is public dataset, where the collection procedure is documented in the dataset's website and we explicitly cited the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are the simulators evaluated in our paper, and their prompts and decoding settings are described in Appendix and anonymous github. We also use an LLM-assisted procedure for response-category scaling, described in Appendix. LLM use for copy-editing only is not part of the methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.