Model-Free Assessment of Simulator Fidelity via Quantile Curves

Garud Iyengar **Yu-Shiou Willy Lin** **Kaizheng Wang**
Department of Industrial Engineering and Operations Research
Columbia University
New York, NY 10027
garud@ieor.columbia.edu yl5782@columbia.edu kaizheng.wang@columbia.edu.

Abstract

Simulation is now pervasive, arising from manufacturing to LLM-driven applications in research, education, and consumer surveys. Yet, fully characterizing the discrepancy between simulators and ground truth remains challenging. We propose a computationally tractable method to estimate the quantile function of the discrepancy between the simulated and ground-truth distributions. The approach does not impose any modeling assumptions on the simulator and it applies broadly across many parameter families: from Bernoulli and multinomial to continuous, vector-valued settings. The resulting quantile curve supports risk-aware summaries (e.g., VaR/CVaR) and comparison of simulators or prompts performance. We illustrate our framework through an application assessing LLM simulation fidelity on the OpinionQA dataset, augmented with simulations spanning seven LLMs.

1 Introduction

The adoption of simulators across operations and manufacturing, agent-based modeling in social-science research, user surveys, and education/training [20, 10, 2, 1] has accelerated with recent advances in artificial intellegence (AI). At the platform level, industry ecosystems such as NVIDIA Omniverse and Earth-2 exemplify the push toward high-fidelity, AI-enabled digital twins. Collectively, these developments fuel the sustained growth in modeling and simulation across domains.

Against this backdrop, it becomes increasingly important to understand the discrepancy between simulation outputs and the real-world. Such discrepancies are documented across many domains: the behavior of large language models (LLMs) , model discrepancy in computing systems, and sim-to-real gap in robotics [5, 4, 14, 17, 13]. To quantify this gap, a large body of work performs Bayesian inference on simulator discrepancy [9]. Their validity hinges on modeling assumptions. On the other hand, recent studies of LLMs yield a variety of model-agnostic discrepancy measures. They are numerical summaries of the overall error [15] or the severity at a given quantile level [8]. To provide flexible and comprehensive analytics, we propose a model-free procedure to approximate the entire quantile function of the discrepancy between the real-world and simulated distributions. Our work shares a similar spirit with conformal inference approaches [18, 3]. However, they typically target pointwise coverage rather than a full distributional characterization, which is a gap we aim to fill.

Ideally, one wants to recover the quantile function for a specified discrepancy measure between the real-world and the simulated distributions. The quantile function directly leads to confidence intervals for the estimators, and risk summaries, e.g. VaR/CVaR. In practice, we only observe finite samples of the real-world and simulated outcomes, and therefore, the true quantile function is not computable. We construct a model-free conservative estimate for the quantile function, and prove that our estimate comes with finite-sample guarantees for any desired α-quantile.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Our contributions are twofold:

1. Our procedure is model-free in that it does not impose any parametric assumptions on either the simulator or the ground truth; therefore, it can be applied to black-box simulators.
2. Since our procedure results in a quantile *function*, we can produce a range of different summaries – from means to tail-risk measures.

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 result, and in Section 4 we discuss an application of our methodology. In Section 5 we conclude by discussing future directions.

2 Problem Formulation and Motivating Example

In this section, we start with a concrete use case that motivates our formulation, and then formally define the quantile estimation problem. Although the terminology below is specific to the example use case, the setting can be generalized. See Appendix B for more examples.

Suppose a media research company plans to survey customers on a particular topic with multinomial outcomes (eg. “Agree”, "Neutral" ,“Disagree”). The company wants to estimate the customer opinion before committing resources for an expensive population study. The company has access to a database of past questions and human answers, and have been training an LLM-based “digital twin” for its customer base. By querying this digital twin with the new question, the company can generate an estimate for the mean response for the new question. The problem now facing the company is to characterize how close this estimate is to the true population mean. Can one provide a confidence interval for the true value? Or, better yet, estimate the quantile function for a suitable discrepancy measure between the simulator and population estimates? This work addresses precisely this problem.

To formally describe the discrepancy between simulator (LLM) and ground truth (human population), and the theoretical challenges we are facing, we consider two levels of randomness in this problem.

First, scenarios (questions) are drawn as $\psi \sim \Psi$. A real system (human) is characterized by $z \in \mathcal{Z}$ with a population distribution $\mathcal{P}$ over the latent profile state $\mathcal{Z}$. For each pair (ψ, z), the real system (human population) produces an categorical outcome Y^{gt}, with conditional distribution $Q^{\mathrm{gt}}(\cdot \mid z, \psi)$ over a space $\mathcal{Y} := \big\{y \in \mathbb{N}^d : \sum_{i=1}^d y_i = 1\big\}$. The simulator (LLM) produces an outcome Y^{sim} with conditional distribution $Q^{\mathrm{sim}}(z^{\mathrm{sim}}, \psi, r)$ over the same outcome space $\mathcal{Y}$, where the $z^{\mathrm{sim}} \in \mathcal{P}^{\mathrm{sim}}$ denotes the i.i.d. synthetic profile that is fed into the LLM, and r denotes LLM settings, including prompting strategy, hyperparameters, and other API settings[1] In our example, $Q^{\mathrm{gt}}(\cdot|z, \psi) =$ Categorical$(\Pi^{\mathrm{gt}}(z, \psi))$, where $\Pi^{\mathrm{gt}}(\psi, z) := \big(Q^{\mathrm{gt}}(\{1\} \mid z, \psi), \ldots, Q^{\mathrm{gt}}(\{d\} \mid z, \psi)\big) \in \mathcal{P}^d$ and $\mathcal{P}^d := \{u \in [0,1]^d : \sum_{i=1}^d u_i = 1\}$. For any question ψ, we can marginalize the population effect, hence denote by $Q^{\mathrm{gt}}(\cdot \mid \psi) = \mathbb{E}_{z \sim \mathcal{P}}[Q^{\mathrm{gt}}(\cdot \mid z, \psi)]$, the conditional distribution of outcome Y^{gt} given ψ. Let $p(\psi)$ be a population statistic of interest, which is a functional of the conditional distribution $Q^{\mathrm{gt}}(\cdot \mid \psi)$ and lives in a parameter space Θ, in this case $= \mathcal{P}^d$, and simulator can be defined similarly under the simulator population $\mathcal{P}^{\mathrm{sim}}$, summing up:

$$p(\psi) := \mathbb{E}_{y \sim Q^{\mathrm{gt}}}[y] = \mathbb{E}_{z \sim \mathcal{P}}\big[\Pi^{\mathrm{gt}}(\psi, z)\big] \in \Theta, \qquad q(\psi) := \mathbb{E}_{z \sim \mathcal{P}^{\mathrm{sim}}}\big[\Pi^{\mathrm{sim}}(\psi, z)\big] \in \Theta.$$

Second, we only observe finite samples per question. For each $j \in [m]$, we are given a question $\psi_j \sim \Psi$ and n_j i.i.d. profiles $z_{j,1:n_j} \sim \mathcal{P}$, with which we observe n_j ground-truth outcomes $y^{\mathrm{gt}}_{j,i} \sim Q^{\mathrm{gt}}(\cdot \mid z_{j,i}, \psi_j)$. Next, we generate k simulator outcomes $y^{\mathrm{sim}}_{j,\ell} \sim Q^{\mathrm{sim}}(\cdot \mid z^{\mathrm{sim}}, \psi_j, r)$ using a simulation pool $z^{\mathrm{sim}}_{j,1:k} \sim \mathcal{P}^{\mathrm{sim}}$ with fixed k across j to standardize simulator sampling. In addition, let $\hat{p}_j$ and $\hat{q}_j$ be estimators of $p(\psi_j)$ and $q(\psi_j)$. Concluding, the dataset is $\mathcal{D} = \big\{(\psi_j, \hat{p}_j, \hat{q}_j, n_j, k)\big\}_{j=1}^m$. Notice that we set k to be fixed across j so that $\{\hat{q}_j\}_{j=1}^m$ are identically distributed, which can be justified since sample collection for simulators are relatively inexpensive.

On a $\psi \sim \Psi$, the discrepancy between simulated and real output distributions is defined as $L(p(\psi), q(\psi))$, where $L : \Theta \times \Theta \to [0, \infty]$ is a discrepancy function. Our method is agnostic

[1] We keep r fixed during calibration so that variation in $\hat{q}_j$ reflects scenario differences rather than encoding drift or model choice; choosing a different r effectively defines a different simulator.

to the choice of L and allows practitioners to choose one that suits their use, such as the Kullback-Leibler (KL) divergence for categorical outputs. With the above setup, we can now formally state our goal:

Construct a function $V(\cdot, \mathcal{D}) : [0,1] \to \mathbb{R}$ from the data $\mathcal{D}$ such that the coverage guarantee for a new ψ

$$\mathbb{P}_{\psi\sim\Psi}\big(L(p(\psi), \hat{q}(\psi)) \le V(\alpha, \mathcal{D}) | \mathcal{D}\big) \ge 1 - \alpha - \varepsilon_m, \qquad \forall \alpha \in [0,1],$$

holds with high probability over the draw of $\mathcal{D}$. The quantity ε_m should vanish as m tends to infinity.

3 Main Result

We introduce our method for constructing such V and then provide theoretical guarantees. Recall that our data $\mathcal{D}$ consists of m questions; the j-th question has population parameters $p(\psi_j)$ and $q(\psi_j)$ in the real and simulated systems; $\hat{p}_j$ and $\hat{q}_j$ are their point estimates. Our procedure has two steps:

1. For each j, compute a confidence set $\mathcal{C}_j(\hat{p}_j)$ on p_j, and calculate the pseudo-discrepancy $\hat{\Delta}_j := \sup_{u\in\mathcal{C}_j(\hat{p}_j)} L\big(u, \hat{q}_j\big)$.
2. Denote $\hat{V}_m(\alpha) :=$ the empirical α-quantile of $\{\hat{\Delta}_j\}_{j=1}^m$.

The key design choice in our methodology is the construction of $\mathcal{C}_j(\cdot)$. Consistent with our motivating example, we illustrate with a multinomial setting.

Example 1 (Multinomial Confidence Set) *Under our problem formulation, we have multinomial outcomes with $\Theta = \mathcal{P}^d := \{u \in [0,1]^d : \sum_{i=1}^d u_i = 1\}$ and denote KL divergence as $\mathrm{KL}(\cdot\|\cdot)$. For any $\gamma \in (0,1)$, the set*

$$\mathcal{C}_j(\hat{p}_j) := \Big\{u \in \mathcal{P}^d : \mathrm{KL}(\hat{p}_j\|u) \le \frac{d-1}{n_j}\log(2(d-1)/\gamma)\Big\}$$

covers p_j with probability at least γ.

The validity of this confidence set is proved in Lemma A.5. In addition to multinomial outcomes, Appendix A provides the bound required for Bernoulli models and more generally, exponential family, covering additional classes of simulation outputs. With the above methodology, we can now state the main theoretical guarantee. We introduce two assumptions on which it relies.

Assumption 1 (Independent data) *Scenarios $\psi_j \in \Psi$ are drawn i.i.d. from Ψ. In addition, given $(\psi_1, \dots, \psi_m)$, the pairs $\mathcal{D} = \{(\mathcal{D}_j^{\mathrm{gt}}, \mathcal{D}_j^{sim})\}_{j=1}^m$ are independent with $\mathcal{D}_j^{\mathrm{gt}} \perp\!\!\!\perp \mathcal{D}_j^{sim}$ conditional on ψ_j, and a new $(\psi, \mathcal{D}^{sim})$ is independent of $\mathcal{D}$.*

Assumption 2 (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 $u \in \Theta$.*

We are now equipped to introduce the main theorem. A complete proof is in Appendix C.

Theorem 3.1 *Under assumption 1 and 2, define the per-scenario discrepancy (unobservable) $\Delta(\psi)$ and the pseudo-discrepancy (observable) $\hat{\Delta}(\psi)$ by*

$$\Delta(\psi) := L(p(\psi), \hat{q}(\psi))\,, \qquad \hat{\Delta}_j := \sup_{u\in\mathcal{C}_j(\hat{p}_j)} L\big(u, \hat{q}_j\big),$$

where $\mathcal{C}_j(\hat{p}_j) \subset \Theta$ are data–driven compact confidence sets satisfying $\mathbb{P}(p_j \in \mathcal{C}_j(\hat{p}_j) \,\big|\, \psi_j, n_j) \ge \frac{1}{2}$. Then, for any $\alpha \in (0,1)$, with probability at least $1-\delta$ over $\mathcal{D}$, we have the following guarantee:

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta(\psi) \le \hat{V}_m\big(1 - \tfrac{\alpha}{2}\big) \,\Big|\, \mathcal{D}\Big) \ge 1 - \alpha - \frac{C(\alpha, m, \delta)}{\sqrt{m}},$$

where $C(\alpha, m, \delta) = \sqrt{2\alpha \log \frac{2m}{\delta} + \frac{(\log \frac{2m}{\delta})^2 + 4\log\frac{2m}{\delta}}{m}} + \frac{\log \frac{2m}{\delta} + 2}{\sqrt{m}} + \sqrt{\frac{\log(4/\delta)}{2}}$, hence the remainder is $O\big(\sqrt{(\log m)/m}\big)$ as $m \to \infty$ for fixed α.

4 Application: LLM Fidelity Profiling

We apply the methodology from Section 3 to real data. Our primary dataset is OpinionQA [16], 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 [8], 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 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 [8] 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 [8]. We also provide the same procedure onto a Bernoulli setting using the EEDI dataset, details can be found in Appendix D.

We apply our methodology to produce a fidelity profile for each candidate LLM ℓ. We use total variation as the discrepancy measure, $L(p, q) = \frac{1}{2} \|p - q\|_1$, and set $\delta = 0.05$. In addition, we set the simulation budget $k = 100$, the result is presented in Figure 1.

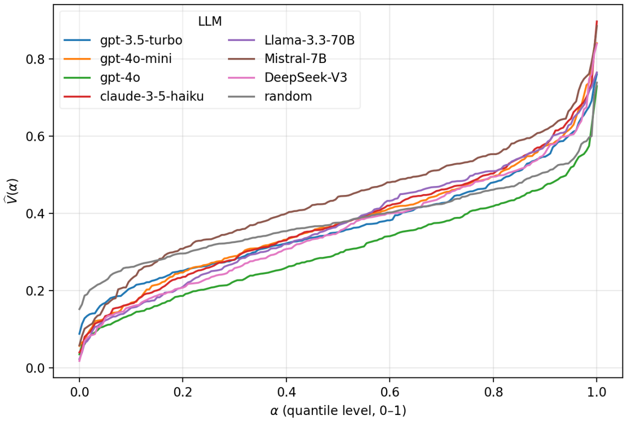

Figure 1: Quantile fidelity profiles $\hat{V}(\alpha)$ across LLMs.

Figure 1 compares models by how tightly their synthetic outcomes track the human distribution across items. We plot $\hat{V}_\ell(\alpha)$ against α, 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, with Mistral 7B clearly performing worse. 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.

5 Discussion

We present a model-free estimator of the quantile function that makes no parametric assumptions on either the simulator or the ground truth, delivers finite-sample guarantees for any desired α-quantile, and supports flexible summaries from means to tail-risk measures. Several promising directions remain: first, our proof relies on DKW-type concentration—often conservative for small m—and a grid-uniform step that further loosens constants; tightening these bounds is an immediate target. Second, many applications involve temporally dependent, dynamic simulation processes; extending our static framework to dynamic settings would broaden applicability. Third, our analysis assumes i.i.d. scenarios, whereas covariate shift or endogenous sampling may invalidate marginal guarantees; addressing such distribution shifts is an important avenue for future work.

References

[1] G. Aher, R. I. Arriaga, and A. T. 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] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

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

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

[5] Y. Gao, D. Lee, G. Burtch, and S. Fazelpour. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025.

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

[7] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[8] C. Huang, Y. Wu, and K. Wang. Uncertainty quantification for LLM-based survey simulations. In *Forty-second International Conference on Machine Learning*, 2025.

[9] M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

[10] C. Macal. Everything you need to know about agent-based modelling and simulation. *Journal of Simulation*, 10:144–156, 05 2016.

[11] J. Mardia, J. Jiao, E. Tánczos, R. D. Nowak, and T. 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.

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

[13] X. B. Peng, M. Andrychowicz, W. Zaremba, and P. 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.

[14] C. J. Roy and W. 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.

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

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

[17] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. 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.

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

[19] Z. Wang, A. Lamb, E. Saveliev, P. Cameron, J. Zaykov, J. M. Hernandez-Lobato, R. E. Turner, R. G. Baraniuk, E. Craig Barton, S. Peyton Jones, S. Woodhead, and C. Zhang. Results and insights from diagnostic questions: The neurips 2020 education challenge. In H. J. Escalante and K. 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.

[20] L. Zhang, L. Zhou, L. Ren, and Y. Laili. Modeling and simulation in intelligent manufacturing. *Computers in Industry*, 112:103123, 2019.

A Useful Lemmas

Lemma A.1 *(Dvoretzky–Kiefer–Wolfowitz (DKW) Inequality via [12])*

Let $X_1, X_2, \dots, X_n$ 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 A.2 *[Hoeffding (additive) via [7]]*

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 A.3 *[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 $\varepsilon > 0$,

$$\mathbb{P}\big(D(p_{\theta_{\bar{T}_n}} \| p_\theta) > \varepsilon\big) \le 2e^{-n\varepsilon}, \quad \textit{or equivalently, } \mathbb{P}\big(D(p_{\hat{\theta}_n} \| p_\theta) > \varepsilon\big) \le 2e^{-n\varepsilon}.$$

Proof:

Define the shifted log-MGF under p_θ,

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

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

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

$$\mathbb{P}_\theta(\bar{T}_n \geq m) = \mathbb{P}\big(e^{\lambda n \bar{T}_n} \geq e^{\lambda n m}\big) \leq e^{-\lambda n m}\, \mathbb{E}_\theta\big[e^{\lambda n \bar{T}_n}\big] = \exp\big(-n(\lambda m - \psi_\theta(\lambda))\big).$$

Optimizing over λ gives the Chernoff bound

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hence

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

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

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

which proves the lemma.

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

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

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

Then for any $\varepsilon > 0$,

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

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

Proof: Via Lemma A.3

Lemma A.5 (Multinomial Chernoff-Hoeffding Bound via [11]) *For all $d \leq (\frac{nC_0}{4})^{\frac{1}{3}}$ and $P \in \mathcal{M}_k$, the following holds with the universal constants $C_0 = \frac{e^3}{2\pi} \approx\ 3.1967$, for any $\epsilon > 0$,*

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

Lemma A.6 (Order–statistic thresholding) *Let $x_1, \dots, x_m \in \mathbb{R}$ and let $x_{(1)} \leq \cdots \leq x_{(m)}$ be their order statistics. Fix a threshold $\tau \in \mathbb{R}$ and an integer $N \in \{1, \dots, m\}$. If at least N of the sample values are at least τ, i.e. $\big|\{j : x_j \geq \tau\}\big| \geq N$, then*

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

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

Proof:

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

B 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 + shift plan.
- Profiles: $\mathcal{Z}$ = machine/operator states, shift team, lot sizes; $\mathcal{P}$ = plant variability.
- Laws: $Q^{\text{gt}}(\cdot \mid z, \psi)$ = empirical cycle-time distribution on the floor; $Q^{\text{sim}}(\cdot \mid a(z), \psi, r)$ = 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: $\mathcal{X} = \{1, \dots, K\}$ 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: $Q^{\text{gt}}(\cdot \mid z, \psi)$ denotes the empirical technology-choice distribution, and $Q^{\text{sim}}(\cdot \mid a(z), \psi, r)$ 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 $\sum_{c=1}^{K} p_c \log\big(p_c/q_c\big)$.
- Sampling: n_j human records per scenario and k synthetic replications per scenario.

C Proof to Theorem 3.1

We first prove the fixed level of confidence $\bar{\alpha}$ case.

As a setup, we work conditionally on $\mathcal{G}_j := \sigma(\psi_j, \hat{q}_j, n_j), \mathcal{G} := \sigma\big(\{\mathcal{G}_j\}_{j=1}^m\big)$. For brevity, we abuse notation and write $p(\psi_j), q(\psi_j)$ as p_j, q_j, and similarly $\hat{p}, \hat{q}$. In addition, for any quantities $\{\Delta_j\}_{j=1}^m$, we denote the sorted version as $\{\Delta_{(i)}\}_{i=1}^m$, ie. $\Delta_{(1)} \leq \cdots \leq \Delta_{(m)}$. For any sequence $\{\Delta_j\}_{j=1}^m$, let $\Delta_{(1)} \leq \cdots \leq \Delta_{(m)}$ denote its order statistics. Throughout the proof, $\alpha \in (0,1)$ denotes a generic quantile index (a function argument) and is distinct from the target coverage level $\bar{\alpha}$.

We seek the quantile function of $\Delta_j = L(p_j, q_j)$, but only observe the estimators $(\hat{p}_j, \hat{q}_j)$. To preserve an i.i.d. structure, we instead work with the proxy $\bar{\Delta}_j := L(p_j, \hat{q}_j)$, for which the sequence $\{\bar{\Delta}_j\}_{j=1}^m$ is i.i.d., since we fix the simulator budget k and each scenarios $\{\psi_j\}$ are i.i.d. as assumed in Section 2.

By defining $\bar{V}_m(\alpha)$ as an empirical α quantile of $\{\bar{\Delta}_j\}_{j=1}^m$, formally $\bar{V}_m(\alpha) = \inf\{t : \bar{F}_m(t) \geq \alpha\} = \bar{\Delta}_{(\lceil m\alpha \rceil)}$ for the ordered $\bar{\Delta}_{(i)}$, where $\bar{F}_m(t) = \frac{1}{m}\sum_{j=1}^m \mathbb{I}(\bar{\Delta}_j \leq x)$. If we have access to $\bar{V}_m(\alpha)$, then we can claim, by Lemma A.1, that with probability $1-\delta$:

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

and this would deliver our desired evelope bound.

However, we do not have access to p_j, so we need to estimate it with $\hat{p}_j$. If we are in the case of $n_j = n$, then this can be solved by creating an uncertainty set across all j. Specifically, we take $\sup L$ out of an uncertainty set over $\hat{p}_j$ that ensures with high probability p_j lies inside, and because n_js are identical this quantity is still i.i.d. , hence the above procedure is still valid. This leads to another problem, which is that we do not have the same sample size n_j for each ψ_j, so $L(\hat{p}_j, \hat{q}_j)$ will not be i.i.d. , so we need to try to control for this heterogeneity.

We instead construct a randomized pseudo-divergences with respect to $\bar{\Delta}_j$. Specifically, we allow a randomized level of coverage $\gamma_j \sim Unif(0,1), \forall j \in [m]$, and define the two divergence terms as

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

where $\mathcal{C}_j(\hat{p}_j, \gamma_j) \subset \Theta$ are data–driven confidence sets satisfying $\mathbb{P}(p_j \in \mathcal{C}_j(\hat{p}_j, \gamma_j) \,|\, \psi_j, n_j) \;\geq\; \gamma_j$. By Assumption 2 and the compactness of the confidence set $C_j(\hat{p}_j, \gamma_j) \subset \Theta$, Berge's maximum theorem guarantees that the supremum in the definition of $\hat{\Delta}_j$ is attained, hence $\hat{\Delta}_j$ is well defined. Therefore, by the coverage property of C_j, we have

$$\mathbb{P}(\hat{\Delta}_j \geq \bar{\Delta}_j | \mathcal{G}_j \cup \sigma(\gamma_j)) \geq \gamma_j,$$

and if we marginalize over γ_j, by $\gamma_j \perp\!\!\!\perp \mathcal{G}_j$ we get $\forall j$

$$\begin{aligned} \mathbb{P}(\hat{\Delta}_j \geq \bar{\Delta}_j | \mathcal{G}_j) &= \mathbb{E}[\mathbb{P}(\hat{\Delta}_j \geq \bar{\Delta}_j | \mathcal{G}_j \cup \sigma(\gamma_j) | \mathcal{G}_j] \\ &\geq \mathbb{E}[\gamma_j | \mathcal{G}_j] = \frac{1}{2}. \end{aligned} \tag{2}$$

Furthermore, by the tower property and independence of γ_j from $\mathcal{G}_j$,

$$\mathbb{P}\big(\hat{\Delta}_j \geq \bar{\Delta}_j\big) = \mathbb{E}\big[\mathbb{P}(\hat{\Delta}_j \geq \bar{\Delta}_j \,|\, \mathcal{G}_j)\big] \geq \tfrac{1}{2},$$

and, conditionally on $\mathcal{G}$, the indicators $Y_j = \mathbf{1}\{\hat{\Delta}_j \geq \bar{\Delta}_j\}$ are independent across j.

We will now use $\{\hat{\Delta}_j\}_{j=1}^m$ to create an upper bound of $\bar{V}_m(1-\alpha)$, which along side (1) will give us our desired envelope. Intuitively, we want to find a larger quantile of $\hat{\Delta}_j$ and with (2), we can claim an upper bound of $\bar{V}_m(1-\alpha)$ with high probability.

First, we define an imaginary set $S_\alpha = \{j \in [m] : \bar{\Delta}_j \geq \bar{V}_m(1-\alpha)\}$, and denote $|S_\alpha| = s$. By definition of empirical $1-\alpha$ quantile, $\inf k = \lceil m(1-\alpha) \rceil$, and hence

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

We next define $Y_j = \mathbb{I}(\hat{\Delta}_j \geq \bar{\Delta}_j)$. By (2): $\mathbb{P}(Y_j = 1|\mathcal{G}) \geq \frac{1}{2}$. We now define another imaginary set $T_\alpha = \{j \in S_\alpha : \hat{\Delta}_j \geq \bar{\Delta}_j\} = \{j \in S_\alpha : Y_j = 1\} = \{j : \hat{\Delta}_j \geq \bar{\Delta}_j \geq \bar{V}_m(1-\alpha)\}$, with which we use to choose our upper bound. First we calculate $\mathbb{E}[T_\alpha]$:

$$\mathbb{E}[T_\alpha|\mathcal{G}] = \mathbb{E}[\sum_{j\in S_\alpha} Y_j|\mathcal{G}] = \sum_{j\in S_\alpha} \mathbb{P}(Y_j|\mathcal{G}) \geq \frac{1}{2}s.$$

Lemma A.2 implies that for any $\delta \in [0, \frac{1}{2}s]$, we have:

$$\mathbb{P}(|T_\alpha| \leq \frac{1}{2}s - t|\mathcal{G}) \leq \mathbb{P}(|T_\alpha| \leq \mathbb{E}[T_\alpha] - t|\mathcal{G}) \leq \exp(-\frac{2t^2}{s}),$$

where we applied $\mathbb{E}[T_\alpha|\mathcal{G}] \geq \frac{1}{2}s$ in the first inequality. By setting $t = c\sqrt{s}$:

$$\mathbb{P}(|T_\alpha| \leq \frac{1}{2}s - t|\mathcal{G}) \leq \exp(-2c^2) \tag{3}$$

With this bound, we can link the actual set of indices we have interest, ie. $U_\alpha = \{j : \hat{\Delta}_j \geq \bar{V}_m(1-\alpha)\}$ to T_α. By construction, $T_\alpha \subseteq U_\alpha$, hence by (3), with probability greather than $1 - \exp(-2c^2)$:

$$|U_\alpha| \geq |T_\alpha| \geq \frac{1}{2}s - c\sqrt{s},$$

which implies at least $\frac{1}{2}s - c\sqrt{s}$ of the $\hat{\Delta}_j$'s are larger than $\bar{V}_m(1-\alpha)$ with high probability.

We now analyze: for a fixed α, what coverage guarantee can we get for the inner probability via order statistics. Set $N := \lfloor \frac{1}{2}s - c\sqrt{s} \rfloor$. If at least N sample values exceed $\bar{V}_m(1-\alpha)$, then by order-statistics calculus (Lemma A.6)

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

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

A sufficient condition is

$$\tfrac{1}{2}m\alpha_{\text{eff}} - c\sqrt{m\alpha_{\text{eff}}} - 1 \;\geq\; m\alpha.$$

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

$$y^* \;=\; \frac{2c/\sqrt{m} + \sqrt{4c^2/m + 8\alpha + 8/m}}{2}, \qquad \alpha_{\text{eff}}(\alpha, c, m) \;=\; (y^*)^2.$$

Thus we obtain the comparison event

$$\mathcal{E}_\alpha := \Big\{\; \hat{V}_m(1-\alpha) \;\geq\; \bar{V}_m\big(1 - \alpha_{\text{eff}}(\alpha, c, m)\big) \Big\}, \qquad \mathbb{P}(\mathcal{E}_\alpha) \;\geq\; 1 - e^{-2c^2}. \tag{4}$$

Apply (1) at level $1 - \alpha_{\text{eff}}(\alpha, c, m)$:

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

On $\mathcal{E}_\alpha$ in (4), $\bar{V}_m(1-\alpha_{\text{eff}}) \leq \hat{V}_m(1-\alpha)$. Hence,

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

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

The exact algebraic form is

$$\alpha_{\text{eff}}(\alpha, c, m) = \frac{\left(\frac{2c}{\sqrt{m}} + \sqrt{\frac{4c^2}{m} + 8\alpha + \frac{8}{m}}\right)^2}{4} = 2\alpha + \frac{c}{\sqrt{m}}\sqrt{8\alpha + \frac{4c^2+8}{m}} + \frac{2c^2+2}{m},$$

and, as $m \to \infty$, $\alpha_{\text{eff}}(\alpha, c, m) = 2\alpha + c\sqrt{8\alpha/m} + O(m^{-1})$.

Finally, $\mathbb{P}_{\psi\sim\Psi}(\cdot \,|\, \mathcal{D})$ is over a fresh test scenario given the realized calibration data $\mathcal{D}$. High-probability qualifiers are taken over $(\mathcal{D}, \{\gamma_j\})$. We first condition on $\mathcal{G}$ (leaving $\{\hat{p}_j\}, \{\gamma_j\}$ random), derive conditional bounds, and then remove the conditioning via the tower property; both key events (1) and (4) are measurable in $(\mathcal{D}, \{\gamma_j\})$.

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

$$\mathcal{E}_{\mathrm{DKW}} := \Big\{ \sup_x \big|\hat{F}_m(x) - F^*(x)\big| \le \varepsilon_m(\delta) \Big\}, \qquad \mathcal{E}_r := \Big\{ \text{(3.1) holds with } \alpha = \alpha_r \Big\}.$$

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

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

Work on the event $\mathcal{E}_{\mathrm{DKW}} \cap \bigcap_{r=1}^m \mathcal{E}_r$. For any $\alpha \in (0,1)$ let $r = \lceil m\alpha \rceil$ and denote $\alpha_+ := \alpha_r = r/m \in [\alpha, \alpha + 1/m]$. Since the empirical quantile is piecewise constant on the m–grid,

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

Applying (3.1) at level α_+ yields

$$\mathbb{P}_{\psi\sim\Psi}\Big(\Delta(\psi) \le \widehat{V}_m(1-\alpha) \;\Big|\; \mathcal{D}\Big) \;\ge\; 1 - 2\alpha_+ - \frac{c}{\sqrt{m}}\sqrt{8\alpha_+ + \frac{4c^2+8}{m}} - \frac{2c^2+2}{m} - \varepsilon_m(\delta).$$

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

$$2\alpha_+ \le 2\alpha + \tfrac{2}{m}, \qquad \sqrt{8\alpha_+ + \tfrac{4c^2+8}{m}} \;\le\; \sqrt{8\alpha + \tfrac{4c^2+16}{m}}.$$

Therefore, with probability at least $1 - \delta - me^{-2c^2}$ over $\mathcal{D}$, the guarantee (3.1) (at α replaced by α_+) holds *uniformly* for all $\alpha \in (0,1)$. The form in the main theoerem is a mere simplification with respect to c and δ.

D Additional Applications: EEDI Dataset

We apply the methodology from Section 3 to real data. Our primary dataset is EEDI [6], built on the NeurIPS 2020 Education Challenge [19], 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 students or simulators choice, consistent with Lemma A.4. We adopt the preprocessed version curated by [8], 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 [8] 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 [8].

We apply our methodology to produce a fidelity profile for each candidate LLM ℓ. We use absolute error as the loss, $L(p,q) = |p - q|$. We set $\gamma = 0.5$ uniformly and the DKW failure probability to $\delta = 0.1$, which determines the curve's effective width at $\alpha \to 1$ in Figure 2. In addition, we set the simulation budget $k = 50$.

Figure 2 compares models by how tightly their synthetic outcomes track the human distribution across items. We plot $\hat{V}_\ell(\alpha)$ against α, 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.

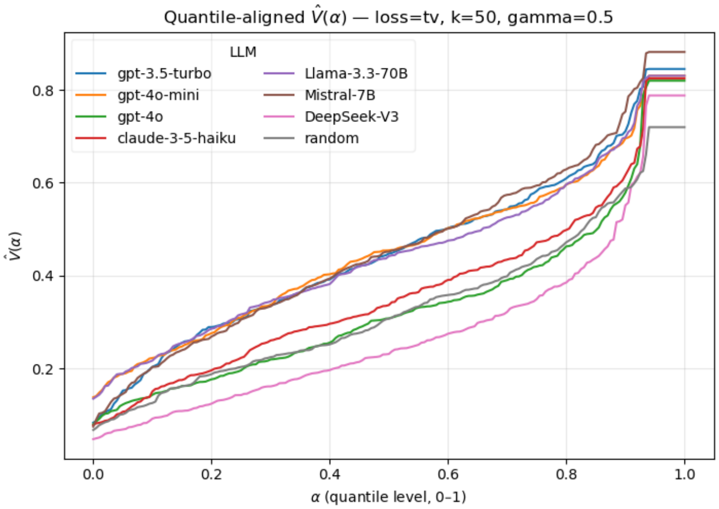

Figure 2: Quantile fidelity profiles $\hat{V}(\alpha)$ across LLMs (Discrepancy: Absolute loss, $k = 50$, $\beta = 0.5$, $\delta = 0.1$).