DISTRIBUTION-LEVEL CONFORMAL SCREENING FOR PROMPT-CONTROLLED LLM SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) can generate synthetic data at relatively low marginal cost, yet identifying prompts whose induced distributions are sufficiently aligned with a target population remains statistically challenging when only a small sample from that population is available. We cast this task as distribution-level set-valued inference and propose a conformal screening procedure that identifies, from a finite family of prompt-induced candidate distributions, those that are statistically compatible with the target population. Specifically, we construct a distribution-level nonconformity score using the Gaussian-smoothed empirical f-divergence between n real and N synthetic observations for each candidate. We then calibrate this score by its randomized conformal rank among m replicate scores generated from the candidate's synthetic pool, and obtain the admissible subset by inverting the resulting conformal tests. Our first theoretical contribution is an exact finite-sample, candidate-wise validity guarantee. Under the compatibility null, exchangeability of the observed and calibration scores guarantees that each compatible candidate is retained with probability at least $1 - \alpha$. Our second theoretical contribution is an asymptotic power guarantee: candidates separated from the target beyond a detection boundary of order $\mathcal{O}(r_{n,N}^{-1/2} + m^{-1/2})$, $r_{n,N} = nN/(n + N)$ are rejected with probability tending to one. Simulations and experiments on OpinionQA and MovieLens show that the retained prompts generate synthetic distributions closer to the full-data target than the empirical distribution based on the small real sample, thereby reducing downstream estimation error.

1 INTRODUCTION

High collection costs, ethical considerations, and application-specific constraints often limit the availability of real-world data for regression analysis and statistical inference. For example, studies in economics, medicine, and the social sciences often depend on human responses collected through surveys and experiments Gu et al. (2025); Wu et al. (2025); Xiao et al. (2025); Nagarajan et al. (2024); Qiao et al. (2025); Cuevas et al. (2025). In these settings, the reliability of downstream statistical analyses depends not only on sample size but also on how well the collected data represent the target population and capture variation across respondents and outcomes, i.e., quality (e.g., data coverage, diversity, label balance).

To alleviate data scarcity, data augmentation provides a practical means of expanding existing datasets and potentially increasing sample diversity. Representative techniques include input perturbations and transformations Krizhevsky et al. (2012); Bae et al. (2018); Zhong et al. (2020), sample interpolation Zhang et al. (2017), synonym replacement Wei & Zou (2019), back-translation Sennrich et al. (2016), and generative adversarial network (GAN)-based synthesis Lupión et al. (2024); Golhar et al. (2025). However, increasing dataset size alone does not guarantee more reliable statistical inference. Transformation-based methods construct variants of existing observations, whereas generative models trained on the available data seek to learn the underlying distribution. Training these models can be computationally demanding Wang et al. (2023); Liu et al. (2020), and GANs are particularly susceptible to overfitting and mode collapse when training data are scarce Wang et al. (2024); Karras et al. (2020). Consequently, the generated samples may fail to faithfully repre-

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

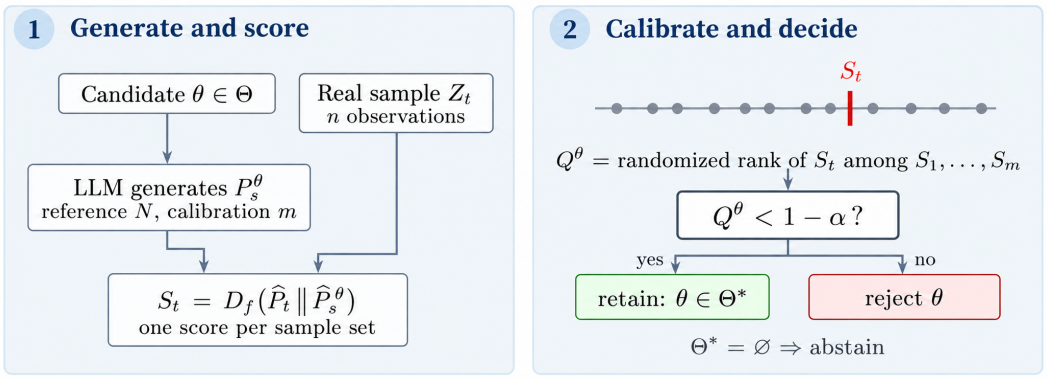


Figure 1: **Overview of conformal prompt screening. Left:** The LLM generates a synthetic reference sample together with m calibration sets and calculate the nonconformity score S_t. **Right:** The score S_t is converted into a randomized conformal rank Q^θ by comparison with the calibration scores $S_1, \cdots, S_m$. The candidate is retained when $Q^\theta < 1 - \alpha$ and rejected otherwise. Augmentation uses only Θ^*; an empty Θ^* triggers abstention.

sent the target distribution Arora et al. (2017); Van Breugel et al. (2023), potentially compromising downstream inference.

Large-scale pretraining, including multilayer perceptrons (MLP)-based architectures Tolstikhin et al. (2021); Lee-Thorp et al. (2022), GAN-based models Wang et al. (2020); Yang et al. (2023), diffusion and flow-based generative models, offers an alternative to learning solely from scarce task-specific data by leveraging knowledge acquired from large-scale datasets. LLMs extend this paradigm by enabling adaptation to downstream tasks through natural-language prompting Vaswani et al. (2017); Brown et al. (2020); Ouyang et al. (2022); Liu et al. (2022). In survey settings, this capability allows researchers to generate synthetic participant responses without task-specific fine-tuning or additional participant recruitment Dominguez-Olmedo et al. (2024); Xu et al. (2025). Yet, generating plausible individual responses does not guarantee that their aggregate distribution faithfully represents a target population. Although LLMs can mimic certain characteristics of human response distributions, empirical studies document systematic discrepancies between model-generated and human response distributions Bellemare-Pepin et al. (2024); Maier et al. (2025); Steyvers et al. (2025); Cai et al. (2024); Calderon et al. (2025); Durmus et al. (2023). Turing-test-based evaluations further identify persistent differences between human-written and model-generated text Pagan et al. (2025). More fundamentally, plausible individual responses do not guarantee that their aggregate distribution faithfully represents the population of interest. The central challenge for downstream statistical inference is therefore not merely to generate more synthetic responses, but to account for the mismatch between the model-induced and target distributions.

To address this distributional mismatch, we propose *conformal screening*, a statistical testing framework for selecting among candidate synthetic distributions rather than individual observations. Specifically, we consider a prespecified finite candidate set Θ, where each $\theta \in \Theta$ induces a synthetic distribution P_s^θ and assess each candidate's compatibility with the target distribution P_t by testing $H_0^\theta : P_t = P_s^\theta$. For each candidate, we independently draw a reference sample of size N and m calibration batches of size n from P_s^θ, with n matching the real sample size. We evaluate the real and calibration batches using the same distributional nonconformity score against the shared reference sample. Under $H_0^\theta : P_t = P_s^\theta$, the real and calibration batches are identically distributed; their scores are therefore exchangeable conditional on the reference sample. This exchangeability enables a conformal test based on the randomized rank of the real-data score relative to the calibration scores. Figure 1 summarizes the framework and our main contributions are as follows.

- **Distribution-level conformal screening.** We introduce a conformal testing framework that selects among prompt-induced synthetic distributions rather than individual observations. The framework operates on batch-level nonconformity scores, which are exchangeable under the null conditional on a shared synthetic reference sample. This construction enables distribution-level testing without imposing a parametric model on the target distribution.

- **Finite-sample calibration and sufficient separation.** We establish exact finite-sample validity under the stated sampling assumptions, pointwise in θ rather than simultaneously over Θ. We further derive a separation rate of $\mathcal{O}(r_{n,N}^{-1/2} + m^{-1/2})$, where $r_{n,N} = nN/(n+N)$. The rate makes the role of scarce real data explicit: as $N/n \to \infty$, the first term scales as $n^{-1/2}$, while increasing m reduces calibration uncertainty.
- **Evaluation from controlled simulations to real tasks.** We evaluate the framework through controlled simulations and experiments on OpinionQA and MovieLens. The results show that retained candidates more closely match the full-data distribution than the small real sample used for screening and yield lower downstream estimation error.

2 PRELIMINARIES

We first introduce the notation, then review two tools underlying our method: f-divergences for quantifying distributional discrepancies and randomized conformal ranks for finite-sample calibration under exchangeability.

Notations. Let $\Omega \subseteq \mathbb{R}^d$ be equipped with its Borel σ-algebra, and let P_t denote the target probability measure on Ω. We observe a *target sample* $\mathcal{Z}_t = \{z_{t,i}\}_{i=1}^n$, where $z_{t,1}, \ldots, z_{t,n} \overset{\text{i.i.d.}}{\sim} P_t$. Each observation $z_{t,i} = (x_i, y_i)$ consists of covariates x_i and a response y_i. We use a pretrained LLM for data synthesis, with θ indexing a candidate mixture of profile-conditioned prompt. For each fixed θ, let P_s^θ denote the induced synthetic distribution on Ω. We draw a *synthetic reference sample* $\mathcal{Z}_s^\theta = \{z_{s,i}^\theta\}_{i=1}^N$ consisting of N i.i.d. observations from P_s^θ. The empirical measures associated with $\mathcal{Z}_t$ and $\mathcal{Z}_s^\theta$ are denoted by $P_{t,n}$ and $P_{s,N}^\theta$, respectively. Section 3.1 describes the construction of candidate prompts.

Throughout, we quantify discrepancies between the target and synthetic distributions using f-divergences Ali & Silvey (1966); Csiszár (1967); Rubenstein et al. (2019); Sreekumar et al. (2023). For probability measures P_t and P_s on Ω satisfying $P_t \ll P_s$, the f-divergence is defined as

$$D_f(P_t \| P_s) = \int_\Omega f\left(\frac{dP_t}{dP_s}\right) dP_s,$$

where $f : [0, \infty) \to \mathbb{R}$ is convex with $f(1) = 0$, and dP_t/dP_s is the Radon–Nikodym derivative of P_t with respect to P_s. By Jensen's inequality, $D_f(P_t \| P_s) \geq 0$, with equality whenever $P_t = P_s$. This family includes the Kullback–Leibler divergence, total variation distance, and χ^2 divergence. In practice, empirical measures need not satisfy the required absolute continuity condition. We therefore compute a smoothed empirical f-divergence, denoted by $D_{f,n,N}(\cdot\|\cdot)$, by applying a common nondegenerate Gaussian kernel to both empirical measures. This smoothing yields strictly positive densities, ensuring a well-defined density ratio while preserving the identifiability of the underlying measures. Details are provided in Appendix C.2.

To assess distributional compatibility while accounting for sampling variability, we use a conformal transducer (Vovk, 2019) to compare a target score with a set of reference scores. Specifically, we apply the same scoring rule to m calibration batches and target batch to obtain the calibration scores $S_1, \ldots, S_m$ and the target score S_t, respectively. Given $\tau \sim \mathrm{Unif}(0, 1)$ independent of all scores, the transducer returns the randomized rank

$$Q := \frac{\sum_{i=1}^m \mathbf{1}\{S_i < S_t\} + \tau\,(1 + \sum_{i=1}^m \mathbf{1}\{S_i = S_t\})}{m+1}. \tag{1}$$

For fixed calibration scores and τ, Q is nondecreasing in S_t, with large ranks corresponding to the upper tail of the calibration scores. If $S_1, \ldots, S_m, S_t$ are exchangeable, then Q is marginally distributed as $\mathrm{Unif}(0, 1)$Vovk (2019). In next section, we specify how the calibration and target scores are constructed.

3 DISTRIBUTION-LEVEL CONFORMAL SCREENING FRAMEWORK

We first describe how candidate prompts induce synthetic distributions from an LLM, then introduce distribution-level conformal screening to select among them.

3.1 CANDIDATE CONSTRUCTION

We generate synthetic responses using a pretrained LLM G. For each application, candidate prompts share the same task instructions and question context but differ in their user-profile descriptions. Figure 4 in Appendix B.2 illustrates this prompt construction.

The prompt-control parameter θ determines a mixture-weight vector $\pi_\theta = (\pi_{\theta,1}, \ldots, \pi_{\theta,L})$, where $\pi_{\theta,\ell} \geq 0$ and $\sum_{\ell=1}^{L} \pi_{\theta,\ell} = 1$. Conditional on θ, each synthetic response is generated independently by sampling a profile and querying the LLM with the corresponding prompt:

$$\ell \sim \text{Categorical}(\pi_\theta), \qquad y \mid \ell \sim G(\cdot \mid \mathsf{p}_\ell).$$

Let $P_s^{(\ell)} := G(\cdot \mid \mathsf{p}_\ell)$ denote the response distribution induced by prompt p_ℓ. Marginalizing over the profile index yields the mixture distribution $P_s^\theta = \sum_{\ell=1}^{L} \pi_{\theta,\ell} P_s^{(\ell)}$. Thus, each θ encodes a candidate profile composition for the synthetic population and determines its synthetic distribution P_s^θ.

We consider a finite candidate set $\Theta = \{\theta_1, \ldots, \theta_k\}$, specified independently of the target sample used for screening, and test each induced distribution P_s^θ against P_t separately. The choice of Θ affects the availability of well-aligned candidates, but not the pointwise validity of each test under the stated sampling assumptions. In our LLM experiments, we construct Θ by discretizing the probability simplex of profile-mixture weights. Task-specific prompts and candidate-grid details are provided in Appendices B.2 and H.

3.2 CONFORMAL SCREENING PROCEDURE

In this section, we now introduce conformal screening for selecting among the synthetic distributions indexed by Θ. For each candidate $\theta \in \Theta$, we test the population null hypothesis

$$H_0^\theta : P_t = P_s^\theta.$$

For each candidate, we draw a synthetic reference sample $\mathcal{Z}_s^\theta$ of size N and m calibration batches $\mathcal{Z}_{s,1}^\theta, \ldots, \mathcal{Z}_{s,m}^\theta$, each of size n. All observations in these samples are drawn independently from P_s^θ. We evaluate the target and calibration batches against the shared reference sample using the nonconformity scores

$$S_t^\theta = D_{f,n,N}(P_{t,n} \| P_{s,N}^\theta), \qquad S_i^\theta = D_{f,n,N}(P_{i,n}^\theta \| P_{s,N}^\theta),$$

where $P_{t,n}$, $P_{i,n}^\theta$, and $P_{s,N}^\theta$ are the empirical measures of the target sample, the ith calibration batch, and the reference sample, respectively. Applying the conformal transducer in Eq. (1) to these scores yields the randomized rank Q^θ. At significance level α, we test each candidate separately, rejecting H_0^θ when $Q^\theta \geq 1 - \alpha$ and retaining the candidate otherwise. This yields the retained set

$$\Theta^* := \left\{\theta \in \Theta : Q^\theta < 1 - \alpha\right\}.$$

Only retained candidates are used for augmentation. If $\Theta^* = \varnothing$, the procedure abstains. Figure 1 provides an overview, and Algorithm 1 in Appendix F presents the pseudocode. The next section establishes the finite-sample validity of the corresponding upper-tail p-value, $p^\theta = 1 - Q^\theta$, for each fixed candidate under H_0^θ.

Remark 1 (Interpretation of Θ^*) *Retention in Θ^* reflects non-rejection of the equality null H_0^θ, not certification of distributional closeness. In particular, $Q^\theta < 1 - \alpha$ does not establish that $D(P_t, P_s^\theta) < \delta$ for a prespecified, practically meaningful tolerance $\delta > 0$. For a fixed candidate, such certification would instead require an equivalence-style test of*

$$H_{0,\delta}^\theta : D(P_t, P_s^\theta) \geq \delta, \qquad H_{1,\delta}^\theta : D(P_t, P_s^\theta) < \delta,$$

with the candidate retained only upon rejection of $H_{0,\delta}^\theta$. We leave this extension to future work.

4 THEORETICAL GUARANTEES

This section analyzes the validity and power of our proposed conformal screening. Section 4.1 establishes exact conditional finite-sample validity under randomized tie-breaking and quantifies the effect of deterministic tie-breaking. Section 4.2 then characterizes the procedure's asymptotic power against candidate distributions that are sufficiently separated from the target.

4.1 POINTWISE FINITE-SAMPLE VALIDITY

Theorem 1 *Under H_0^θ and the stated sampling assumptions, the scores $S_1^\theta, \dots, S_m^\theta, S_t^\theta$ are exchangeable conditional on the reference sample $\mathcal{Z}_s^\theta$. Consequently, the randomized conformal rank in Eq.* (1) *satisfies*

$$Q^\theta \mid \mathcal{Z}_s^\theta \sim \mathrm{Unif}(0,1).$$

Equivalently, $p^\theta := 1 - Q^\theta$ is a conditionally valid randomized p-value.

The proof of Theorem 1 establishes conditional exchangeability of the scores and applies the randomized-rank argument in Remark 13 of Vovk (2019). Full details can be found in Appendix D.3. The theorem yields exact conditional finite-sample Type I error control for our screening rule. Specifically, under H_0^θ,

$$\mathbb{P}\big(\theta \notin \Theta^* \,\big|\, \mathcal{Z}_s^\theta\big) = \mathbb{P}\big(Q^\theta \geqslant 1-\alpha \,\big|\, \mathcal{Z}_s^\theta\big) = \alpha. \tag{2}$$

Thus, any fixed, prespecified candidate satisfying $P_s^\theta = P_t$ is retained with conditional probability $1-\alpha$. This guarantee holds for all finite sample sizes n, N, and m, without asymptotic approximation. It is pointwise in θ and does not by itself provide simultaneous level-α error control over the candidate set Θ.

Remark 2 (Multiplicity) *For each question $j = 1, \dots, T$, let Q_j^θ denote the randomized conformal rank obtained by Eq.* (1) *to the target and calibration scores for that question. For a fixed candidate θ, simultaneous testing across T survey or rating questions can be handled by applying Bonferroni or Holm–Bonferroni to the randomized upper-tail p-values $p_j^\theta := 1 - Q_j^\theta$. Since each p_j^θ is uniform on $(0,1)$ under $H_{0,j}^\theta$, these corrections control the familywise error rate across questions at level α, without requiring independence. Details are provided in Appendix A.2.*

Exact calibration in Eq. (2) relies on the independent random tie-breaking variable $\tau_\theta \sim \mathrm{Unif}(0,1)$. Randomized tie-breaking can be reproduced by recording the random draws. Fixing the tie-breaking value deterministically, however, changes the calibration guarantee. The following corollary quantifies the resulting deviation from the nominal level in terms of the largest tie block among the calibration and target scores.

Corollary 2 *Suppose the assumptions of Theorem 1 hold. Fix a candidate $\theta \in \Theta$ and $\tau \in [0,1]$, and let Q_τ^θ denote the rank in Eq.* (1) *with the randomization variable fixed at τ. Let $\mathfrak{S}^\theta = \{S_1^\theta, \dots, S_m^\theta, S_t^\theta\}$ denote the unordered multiset of scores, with largest tie-block size $B_{\max}^\theta = \max_{v \in \mathfrak{S}^\theta} \#\big\{j \in \{1,\dots,m\} \cup \{t\} : S_j^\theta = v\big\}$. Then, under H_0^θ, for every $\alpha \in [0,1]$,*

$$\Big(1 - \alpha - \frac{\tau B_{\max}^\theta}{m+1}\Big) \vee 0 \;\leqslant\; \mathbb{P}\big(Q_\tau^\theta < 1-\alpha \,\big|\, \mathcal{Z}_s^\theta, \mathfrak{S}^\theta\big) \;\leqslant\; \Big(1 - \alpha + \frac{(1-\tau) B_{\max}^\theta}{m+1}\Big) \wedge 1.$$

The proof is provided in Appendix D.4. Setting $\tau = 0$ yields a conservative deterministic screening rule, with conditional Type I error at most α under H_0^θ. If the scores are almost surely distinct, this error equals $\lfloor \alpha(m+1) \rfloor / (m+1)$, so its gap from the nominal level is strictly less than $1/(m+1)$.

4.2 ASYMPTOTIC SCREENING POWER

In this subsection, we analyze the power of our conformal screening test against alternatives $P_t \neq P_s^\theta$. Using the Kullback–Leibler (KL) divergence as a representative discrepancy measure, we derive an upper bound on the non-rejection probability and establish consistency under the stated regularity and separation conditions.

In the following analysis, we condition on the synthetic reference sample $\mathcal{Z}_s^\theta$. Let $q_{1-\alpha}^\theta$ denote the $(1-\alpha)$-quantile of the conditional calibration-score distribution, and let $\hat{q}_{1-\alpha,m}^\theta$ denote its empirical counterpart computed from $S_1^\theta, \dots, S_m^\theta$. These quantiles serve as reference thresholds for the power analysis; the rank-based rejection criterion remains $Q^\theta \geq 1-\alpha$.

We define the population Gaussian-smoothed KL divergence as $S_t^* := D_{\mathrm{KL}}\big(P_t * \varphi_\sigma \,\big\|\, P_s^\theta * \varphi_\sigma\big)$, where φ_σ is a Gaussian kernel with fixed bandwidth $\sigma > 0$. The theorem below bounds the non-rejection probability and establishes consistency under the stated regularity and separation conditions. We first introduce the required support assumption.

Assumption 1 *There exists a fixed compact set $\Omega \subseteq \mathbb{R}^d$ with $0 < \lambda_d(\Omega) < \infty$ such that $\mathrm{supp}(P_t) \subseteq \mathrm{supp}(P_s^\theta) \subseteq \Omega$, where λ_d denotes the d-dimensional Lebesgue measure.*

Assumption 2 *Recall that the definition of the conditional cumulative distribution function of the calibration scores, i.e., $P_{s,n}^\theta(s) := \mathbb{P}\big(S_i^\theta \leqslant s \mid \mathcal{Z}_s^\theta\big)$. Assume that $P_{s,n}^\theta$ is differentiable at its $(1-\alpha)$-quantile $q_{1-\alpha}^\theta$ and that there exist positive lower bounds $\zeta_{n,\alpha}$ and a constant $\zeta_0 > 0$, independent of n, such that*

$$f_{s,n}^\theta(q_{1-\alpha}^\theta) = P_{s,n}^{\theta\prime}(q_{1-\alpha}^\theta) \geqslant \zeta_{n,\alpha} \geqslant \zeta_0. \tag{3}$$

Eq. (2) states the conditional Type I error of the randomized rank-based rule, whereas the power analysis below uses score quantiles as reference thresholds without altering the rejection criterion.

Theorem 3 (Consistency of the screening test) *Suppose that $P_t \neq P_s^\theta$ and that Assumptions 1 and 2 hold. Let $r_{n,N} := nN/(n+N)$, and consider $n, N, m \to \infty$ with $N/(n+N) \to \rho \in [0,1]$. Then there exist constants $c_1, c_2, c_q > 0$ such that, for any positive sequence $\varepsilon = \varepsilon_{n,N,m}$ satisfying*

$$\Delta_{n,N}(\varepsilon) \;:=\; S_t^* - q_{1-\alpha}^\theta - \varepsilon - c_1 r_{n,N}^{-1} \;>\; 0, \quad \text{for fixed } \alpha \in (0,1), \tag{4}$$

the score-threshold test that rejects when $S_t^\theta > \widehat{q}_{1-\alpha,m}^\theta$ has Type II error bounded by

$$\mathbb{P}\big(S_t^\theta \leqslant \widehat{q}_{1-\alpha,m}^\theta\big) \;\leqslant\; 2\exp\big(-c_q\, m\, \varepsilon^2\big) \;+\; 2\exp\big(-c_2\, r_{n,N}\, \Delta_{n,N}(\varepsilon)^2\big).$$

Consequently, if $m\varepsilon^2 \to \infty$ and $r_{n,N}\Delta_{n,N}(\varepsilon)^2 \to \infty$, then $\mathbb{P}\big(S_t^\theta \leq \widehat{q}_{1-\alpha,m}^\theta\big) \to 0$ so the power converges to one. In particular, a sufficient separation condition is $S_t^* - q_{1-\alpha}^\theta \gg r_{n,N}^{-1/2} + m^{-1/2}$.

The proof combines concentration bounds for the target score and the empirical cutoff. Appendix D.5 provides the cutoff concentration result and its assumptions, while Appendix D.6 gives the full proof of the theorem. The resulting exponential bound establishes consistency of the score-threshold test whenever $S_t^* - q_{1-\alpha}^\theta \gg r_{n,N}^{-1/2} + m^{-1/2}$. The two terms capture distinct sources of uncertainty: the target score S_t^θ concentrates around the population divergence S_t^* at scale $r_{n,N}^{-1/2}$, whereas the empirical cutoff $\widehat{q}_{1-\alpha,m}^\theta$ concentrates around the conditional quantile $q_{1-\alpha}^\theta$ at scale $m^{-1/2}$. Their sum therefore defines a sufficient separation scale: when the population gap dominates both sources of uncertainty, the Type II error converges to zero.

Remark 3 *Let $S_{(1)}^\theta \leq \cdots \leq S_{(m)}^\theta$ denote the ordered calibration scores. Define the finite-sample-adjusted cutoff by $j_{\alpha,m} := \lceil (m+1)(1-\alpha) \rceil$, $\widehat{q}_{1-\alpha,m}^\theta := S_{(j_{\alpha,m})}^\theta$ for fixed $\alpha \in (0,1)$, with the convention $S_{(m+1)}^\theta := +\infty$. The score-threshold rule $S_t^\theta > \widehat{q}_{1-\alpha,m}^\theta$ is equivalent to the conservative rank rule $Q_0^\theta \geq 1-\alpha$, where Q_0^θ is the rank in Eq. (1) with $\tau_\theta = 0$. Since $Q^\theta \geq Q_0^\theta$, every rejection by the cutoff rule is also a rejection by the randomized rule. Consequently,*

$$\mathbb{P}\big(Q^\theta < 1-\alpha\big) \leq \mathbb{P}\big(S_t^\theta \leq \widehat{q}_{1-\alpha,m}^\theta\big),$$

and the same inequality holds conditional on the reference sample. Thus, any upper bound on the non-rejection probability of this cutoff-based test also applies to the randomized screening test. In particular, consistency of the former implies consistency of the latter, even when scores are tied.

Theorems 1 and 3 characterize the validity and power of our conformal screening procedure. The former guarantees exact conditional Type I error α for each fixed candidate satisfying the population null. Under the stated regularity conditions, the latter establishes vanishing Type II error whenever $S_t^* - q_{1-\alpha}^\theta \gg r_{n,N}^{-1/2} + m^{-1/2}$. The sufficient separation scale $r_{n,N}^{-1/2} + m^{-1/2}$ clarifies the distinct roles of the three sample sizes. The calibration term $m^{-1/2}$ captures uncertainty in the empirical cutoff $\widehat{q}_{1-\alpha,m}^\theta$. Increasing m improves cutoff precision without changing the empirical divergence score. The score term $r_{n,N}^{-1/2}$ reflects uncertainty from both the target and synthetic reference samples. Since $r_{n,N}^{-1/2} = (1/n + 1/N)^{1/2}$, it is governed by the smaller sample size. In particular, $r_{n,N}^{-1/2} \sim n^{-1/2}$ as $N/n \to \infty$, so enlarging the reference sample alone cannot improve the leading-order score term in this bound. Once calibration uncertainty is also sufficiently small, increasing n improves the overall separation bound, whereas further synthetic sampling yields diminishing returns.

5 EXPERIMENTS

In this section, we evaluate the proposed conformal screening procedure through simulations and real-world experiments. Our simulations use a tri-hedge distribution that combines Gaussian side modes with a non-Gaussian component. We compare interval-construction methods and examine screening over a fixed candidate grid as the target sample size n and synthetic reference sample size N vary. Appendix G.1 provides a Gaussian-mixture experiment, detailed definitions of the alternative discrepancy measures and interval-construction baselines, and additional simulation results. Our real-world experiments evaluate prompt-controlled synthetic data generated by black-box LLMs on OpinionQA and MovieLens. Using small target samples for screening, we assess both screening quality and downstream estimation performance. We also examine whether the procedure abstains when the candidate family is misspecified.

5.1 SIMULATION ON TRI-HEDGE DISTRIBUTION

We evaluate conformal screening on the tri-hedge distribution. Following the mixture parameterization in Section 3.1, we construct 93,750 candidates by combining a grid of component parameters with six admissible mixture-weight configurations. Appendix G.2 provides the distribution and full candidate-grid specifications. We compare interval coverage and length across methods and examine how the candidate acceptance ratio varies with the discrepancy measure, target sample size n, synthetic reference sample size N, and number of calibration batches m. For a fixed target sample and candidate grid, a lower acceptance ratio indicates greater selectivity, but does not alone imply better screening performance; it must be interpreted alongside empirical coverage.

5.1.1 COMPARISON OF INTERVAL-CONSTRUCTION METHODS

Table 1 compares our proposed method (Conformal-base) with the ground-truth benchmark (Oracle) and Chebyshev-, Chernoff-, and CLT-based interval-construction methods on the tri-hedge distribution. We report empirical coverage and the mean and standard deviation of interval length. Appendix G provides details of each method.

Table 1: Confidence interval performance for the tri-hedge model at $\alpha = 0.05$, $N = 2000$, and $n = 100$. Cov. denotes empirical coverage (nominal: 0.95); Avg. Len. and SD denote the mean and standard deviation of interval length, respectively. Among methods meeting the nominal coverage target, shorter and less variable intervals are preferred.

Method	$m = 20$			$m = 160$			$m = 320$		
	Cov.	**Avg. Len.**	**SD**	**Cov.**	**Avg. Len.**	**SD**	**Cov.**	**Avg. Len.**	**SD**
Oracle	0.950249	0.052710	0.010962	0.950031	0.051265	0.009737	0.950016	0.049540	0.010427
Chebyshev-base	0.985721	0.087156	0.025300	0.995197	0.089569	0.016123	0.994917	0.086002	0.015024
Chernoff-base	1.000000	3.047602	**0.014467**	1.000000	3.046348	0.011274	1.000000	3.043597	0.011534
CLT-base	0.926318	0.049722	0.015138	0.955509	0.052979	**0.009890**	0.954367	0.050822	**0.009442**
Conformal-base	**0.953383**	**0.061050**	0.020570	**0.954416**	**0.052856**	0.010105	**0.950631**	**0.049976**	0.009900

As shown in Table 1, at $m = 20$, our method achieves empirical coverage closer to the nominal level of 0.95 than the CLT-based method, although this difference narrows as m increases. Both the Chebyshev- and Chernoff-based methods are conservative, with the latter producing substantially wider intervals. For the conformal method, increasing m changes the average interval length only modestly, while the standard deviation of interval length decreases from 0.020570 to 0.009900.

5.1.2 FIXED-GRID SCREENING ACROSS DISCREPANCY MEASURES

Figure 2 shows that the fraction of retained candidates generally decreases as the target sample size n increases, indicating greater screening selectivity. The ordering across discrepancy measures is consistent for both values of m: MMD and W_2 retain the largest fractions of candidates, KL, L_2, and TV are more selective, and CLT-based KL falls between these groups. The curves for $N = 1000$ and $N = 2000$ remain close relative to the differences among discrepancy measures, indicating limited sensitivity to the reference sample size on this grid. Increasing m from 20 to 40 modestly raises

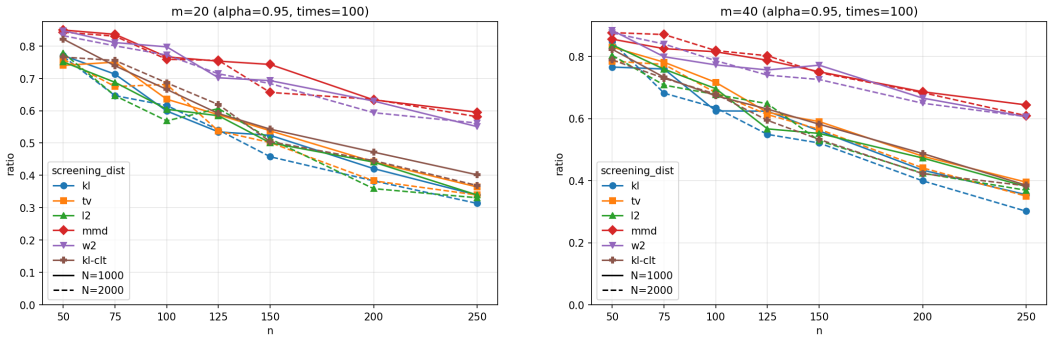


Figure 2: Average screening ratio for the tri-hedge distribution of target sample size n. Solid and dashed lines correspond to synthetic reference sample sizes $N = 1000$ and $N = 2000$. The nominal confidence level is $1 - \alpha = 0.95$.

the retention ratio without changing the ordering. The contrasting effects of increasing n and N are consistent with the sample-size analysis in Section 4.

Additional fixed-grid results for other values of N/n and α are provided in Appendix G.2. Together with the computational comparisons in Appendix G.4, these results suggest that KL divergence offers a favorable balance between screening selectivity and computational efficiency. We therefore use KL divergence in subsequent experiments.

5.2 Real-Data Experiments

In this section, we evaluate conformal screening of prompt-controlled LLM response distributions on OpinionQA and MovieLens. In both applications, the control parameter π specifies the mixture weights over profile-conditioned prompts. Appendix H provides full details of data preprocessing, prompt construction, candidate grids, generation settings, and evaluation protocols.

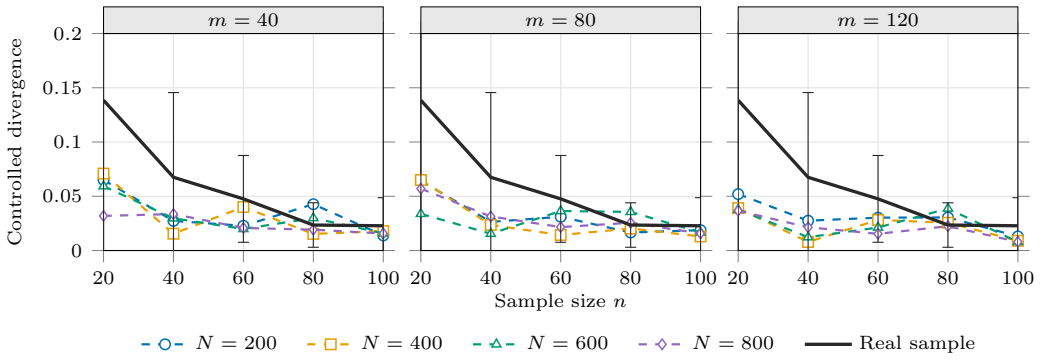
5.2.1 Experiments on OpinionQA

OpinionQA Santurkar et al. (2023) evaluates the alignment between LLM-generated survey responses and human opinions. We use Wave 26 and construct candidate synthetic distributions by varying the mixture weights over prompts conditioned on six political-ideology profiles.

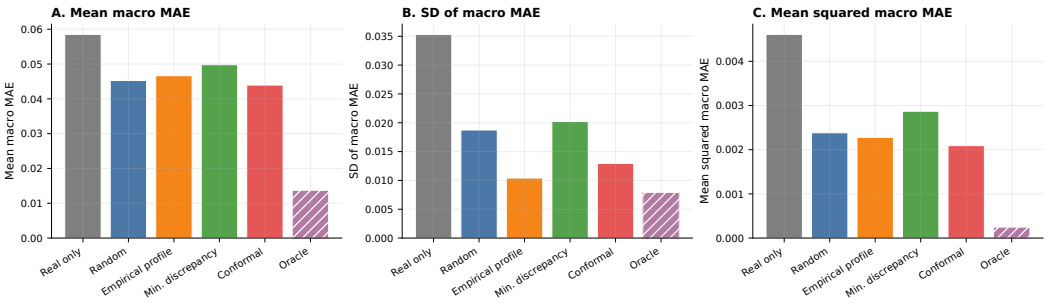
Figure 3a suggests that screening resolution is driven primarily by the target sample size n, while increasing the synthetic reference sample size N yields diminishing returns once the reference sample is sufficiently large. Appendix H.1 reports sensitivity to the number of calibration batches m and the complete counts of retained candidates. Figure 3b shows that conformal-screened augmentation achieves a mean macro MAE of 0.0440, compared with 0.0585 for the Real only baseline and 0.0453–0.0498 for the feasible synthetic baselines. These results suggest that screening can reduce response-proportion estimation error while preserving some of the variability reduction achieved by synthetic augmentation. The Oracle serves only as an infeasible benchmark. Appendix H.1 provides the complete data-processing and generation settings, metric definitions, and retained-candidate counts across (n, N, m). It also reports a comparison between Qwen3-8B and DeepSeek-V4-Flash and an additional downstream linear-regression experiment.

5.2.2 Experiments on MovieLens

On MovieLens Harper & Konstan (2015), we construct $43{,}758$ candidate controls by varying mixture weights over nine profile-conditioned prompts based on rating activity and average rating. In the primary misspecification experiment, all candidates are rejected across the displayed (m, N) configurations when $n \in \{80, 100\}$. The resulting empty retained set, $\Theta^* = \varnothing$, triggers abstention from synthetic augmentation. This result is specific to the tested prompt family and does not rule out compatible distributions under alternative prompt constructions.



(a) controlled divergence across m, N, and n. panels vary the number of calibration batches m. the solid line is the real-sample benchmark; colored dashed lines indicate synthetic reference sample sizes N.



(b) Downstream response-proportion estimation on OpinionQA at $m = 40$. panels show mean, standard deviation, and mean squared macro MAE across (n, N) configurations. Hatching marks the oracle benchmark.

Figure 3: Controlled divergence and OpinionQA downstream estimation for different m, N, and n.

We also evaluate simultaneous screening of the rating distributions for *Forrest Gump* and *Fight Club*. For each fixed candidate, we apply Bonferroni and Holm corrections to control the familywise error rate across the two movie-level hypotheses. Across the displayed configurations, Holm yields lower controlled divergence than Bonferroni while maintaining a stable set of retained candidates. Appendix H.2 provides the complete primary screening grid and abstention results, additional-movie experiments, mean-rating estimation results, retained-mixture diagnostics, and the full Bonferroni–Holm comparison.

6 Conclusion

We introduced *conformal screening*, a framework that uses a small real sample to select among prompt-controlled LLM-generated distributions rather than individual observations. Conditional exchangeability of batch-level scores under the null yields exact finite-sample validity for each fixed candidate. Under the stated regularity conditions, the test is consistent when the divergence-to-cutoff gap dominates $r_{n,N}^{-1/2} + m^{-1/2}$. This sufficient separation scale clarifies the value of scarce real data: when $N \gg n$ and calibration uncertainty is sufficiently small, the real-sample term $n^{-1/2}$ dominates, and additional synthetic sampling yields diminishing improvements to the bound. Experiments on simulated and real-world data support the theoretical analysis and demonstrate downstream estimation gains and abstention under candidate-family misspecification. Our framework assumes a prespecified finite candidate family, independent sampling, and score exchangeability under each null. They do not directly cover adaptive prompt search or simultaneous error control across candidates. Retention does not certify distributional closeness, and performance depends on how well candidates represent the target and on divergence estimation accuracy. Extensions to adaptive candidate construction, dependent sampling, and high-dimensional settings are left for future work.

AI USE STATEMENT

The authors used a general-purpose LLM assistant for language editing of author-written prose (grammar, word choice, and conciseness). The research questions, the theoretical development including all theorem statements and proofs, the experimental design, and the analysis of results were conceived and carried out by the authors. All text and code suggested by the assistant were reviewed, verified, and edited by the authors, who take full responsibility for the entire content of this paper, including any remaining errors.

REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our results. The source code, along with detailed instructions for data preparation, training and evaluation, is provided in the supplementary material and will be publicly released at an anonymous repository upon publication. The real-world datasets used in this work are publicly available and properly cited. We report the full set of hyperparameters, model architectures, and optimization configurations in Appendix G and H, including search ranges for all tuned hyperparameters. All experiments are conducted on [4 $\times$ Nvidia RTX 3090] with [CUDA Toolkit 12.0]. A random seed of 42 was used for simulation experiments, with statistics averaged over 100 repeated runs. In real-data experiments, OpinionQA used seed 42. For MovieLens, seeds 42 and 123 were used to randomly sample movies, and separate experiments were performed with each seed. For theoretical results, complete proofs of all propositions and theorems, together with the assumptions they rely on, are given in Appendix D.

REFERENCES

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966. doi: 10.1111/j.2517-6161.1966.tb00626.x.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *International conference on machine learning*, pp. 224–232. PMLR, 2017.

Hyun-Jin Bae, Chang-Wook Kim, Namju Kim, BeomHee Park, Namkug Kim, Joon Beom Seo, and Sang Min Lee. A perlin noise-based augmentation strategy for deep learning with small data samples of hrct images. *Scientific reports*, 8(1):17687, 2018.

Antoine Bellemare-Pepin, François Lespinasse, Philipp Thölke, Yann Harel, Kory Mathewson, Jay A Olson, Yoshua Bengio, and Karim Jerbi. Divergent creativity in humans and large language models. *arXiv preprint arXiv:2405.13012*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zhenguang Cai, Xufeng Duan, David Haslett, Shuqi Wang, and Martin Pickering. Do large language models resemble humans in language use? In *Proceedings of the workshop on cognitive modeling and computational linguistics*, pp. 37–56, 2024.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms. *arXiv preprint arXiv:2501.10970*, 2025.

Imre Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 2:299–318, 1967.

Alejandro Cuevas, Jennifer V Scurrell, Eva M Brown, Jason Entenmann, and Madeleine IG Daepp. Collecting qualitative data at scale with large language models: A case study. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–27, 2025.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37: 45850–45878, 2024.

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.

Mayank V. Golhar, Taylor L. Bobrow, Saowanee Ngamruengphong, and Nicholas J. Durr. Gan inversion for data augmentation to improve colonoscopy lesion classification. *IEEE Journal of Biomedical and Health Informatics*, 29(6):3864–3873, 2025. doi: 10.1109/JBHI.2024.3397611.

Xujun Gu, Hanwen Zhang, Divya Patil, Zafar Zafari, Julia Slejko, and Eberechukwu Onukwugha. Use of large language models to extract cost-effectiveness analysis data: A case study. *Value in Health*, 28(11):1637–1645, 2025. ISSN 1098-3015. doi: https://doi.org/10.1016/j.jval.2025.05.008. URL https://www.sciencedirect.com/science/article/pii/S1098301525023691.

F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015. doi: 10.1145/2827872.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the north American chapter of the Association for Computational Linguistics: human language technologies*, pp. 4296–4313, 2022.

Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In *International conference on learning representations*, 2020.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 61–68, 2022.

Marcos Lupión, Federico Cruciani, Ian Cleland, Chris Nugent, and Pilar M. Ortigosa. Data augmentation for human activity recognition with generative adversarial networks. *IEEE Journal of Biomedical and Health Informatics*, 28(4):2350–2361, 2024. doi: 10.1109/JBHI.2024.3364910.

Benjamin F Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C Luhmann, Robbie Dow, Kli Pappas, and Thomas V Wiecki. Llms reproduce human purchase intent via semantic similarity elicitation of likert ratings. *arXiv preprint arXiv:2510.08338*, 2025.

Radha Nagarajan, Midori Kondo, Franz Salas, Emre Sezgin, Yuan Yao, Vanessa Klotzman, Sandip A Godambe, Naqi Khan, Alfonso Limon, Graham Stephenson, et al. Economics and equity of large language models: health care perspective. *Journal of Medical Internet Research*, 26:e64226, 2024.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Nonparametric estimation of the likelihood ratio and divergence functionals. *IEEE Transactions on Information Theory*, 56(11): 5847–5861, 2010.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Nicolò Pagan, Petter Törnberg, Christopher A. Bail, Anikó Hannák, and Christopher Barrie. Computational turing test reveals systematic differences between human and ai language, 2025. URL https://arxiv.org/abs/2511.04195.

Tingrui Qiao, Caroline Walker, Chris Cunningham, Adam Jang-Jones, Susan Morton, Kane Meissel, and Yun Sing Koh. Longitudinal surveys are texts: Llm-enhanced analysis of school attendance in new zealand. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 310–327. Springer, 2025.

Paul Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O Tolstikhin. Practical and consistent estimation of f-divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 86–96, 2016.

Sreejith Sreekumar, Ziv Goldfeld, and Kengo Kato. Limit distribution theory for f-divergences. *IEEE Transactions on Information Theory*, 70(2):1233–1267, 2023.

Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2):221–231, 2025.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261–24272, 2021.

Boris Van Breugel, Zhaozhi Qian, and Mihaela Van Der Schaar. Synthetic data, real errors: how (not) to publish and use synthetic data. In *International Conference on Machine Learning*, pp. 34793–34808. PMLR, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vladimir Vovk. Universally consistent conformal predictive distributions. In *conformal and probabilistic prediction and applications*, pp. 105–122. PMLR, 2019.

Jian Wang, Xin Lan, Yuxin Tian, and Jiancheng Lv. MS^{3}d: A RG flow-based regularization for GAN training with limited data. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50746–50765. PMLR, 21–27 Jul 2024.

Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: effective knowledge transfer from gans to target domains with few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9332–9341, 2020.

Zhendong Wang, Yifan Jiang, Huangjie Zheng, Peihao Wang, Pengcheng He, Zhangyang Wang, Weizhu Chen, Mingyuan Zhou, et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36:72137–72154, 2023.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Jing Cynthia Wu, Jin Xi, and Shihan Xie. Llm survey framework: Coverage, reasoning, dynamics, identification. Working Paper 34308, National Bureau of Economic Research, October 2025. URL http://www.nber.org/papers/w34308.

Yuan-An Xiao, Weixuan Wang, Dong Liu, Junwei Zhou, Shengyu Cheng, and Yingfei Xiong. Predicatefix: Repairing static analysis alerts with bridging predicates. *arXiv preprint arXiv:2503.12205*, 2025.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. In *International Conference on Learning Representations*, volume 2025, pp. 76346–76382, 2025.

Jian Yang, Shuming Ma, Li Dong, Shaohan Huang, Haoyang Huang, Yuwei Yin, Dongdong Zhang, Liqun Yang, Furu Wei, and Zhoujun Li. Ganlm: Encoder-decoder pre-training with an auxiliary discriminator. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9394–9412, 2023.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *Proceedings of the AAAI conference on artificial intelligence*, 34(07):13001–13008, 2020.

A Supplementary materials for Methods

A.1 Properties of the Conformal Predictive System

Following Vovk (2019), we summarize the monotonicity, boundary behavior, and calibration properties of the conformal transducer on the score scale. Write $Q(s,\tau) := Q(S_1,\dots,S_m,s,\tau)$ for the transducer in Eq. (1), where $\tau \sim \mathrm{Unif}(0,1)$ is independent of all scores.

(P1) Monotonicity. For fixed calibration scores, $Q(s,\tau)$ is nondecreasing in s for each $\tau \in [0,1]$ and in τ for each s. The variable τ determines the rank's position within the interval associated with the test score's tie block, including the test score itself.

(P2) Boundary behavior. At the lower and upper ends of the support of S_t, the predictive distribution approaches 0 and 1, respectively. More precisely, in the randomized formulation of Vovk (2019), the lower boundary is obtained with $\tau = 0$ and the upper boundary with $\tau = 1$. Together with (P1), this ensures that Q has the monotonicity and boundary behavior required of a randomized distribution function in S_t.

(P3) Calibration. If $S_1,\dots,S_m,S_t$ are exchangeable, independent uniform randomization yields $\mathbb{P}\{Q(S_t,\tau) \le u\} = u,\ u \in [0,1]$. Hence, $Q(S_t,\tau)$ is exactly uniform, and $p := 1 - Q(S_t,\tau)$ is an exact randomized upper-tail p-value.

In our framework, this calibration also holds conditional on the shared reference sample under H_0^θ, by conditional exchangeability of the batch-level scores.

A.2 Conformal screening for Multihypothesis

When synthetic-data reliability is assessed across T survey or rating questions, screening the joint response distribution may be computationally demanding. An alternative is to combine question-wise conformal tests with standard multiplicity corrections, such as Bonferroni or Holm–Bonferroni. For a fixed candidate $\theta \in \Theta$, let $P_{t,j}$ and $P^\theta_{s,N,j}$ denote the target and synthetic response distributions for question j. We test $H^\theta_{0,j} : P_{t,j} = P^\theta_{s,N,j},\ j = 1,\dots,T$. Let Q^θ_j denote the randomized conformal rank obtained by applying Eq. (1) to the scores for question j. Under $H^\theta_{0,j}$ and the stated sampling assumptions, $p^\theta_j := 1 - Q^\theta_j$ is an exact randomized upper-tail p-value.

The Bonferroni correction assigns significance level α/T to each test. A candidate θ is retained only if no question-level null is rejected:

$$Q^\theta_j < 1 - \frac{\alpha}{T}, \qquad j = 1,\dots,T.$$

This correction controls the familywise error rate across questions at level α for each fixed candidate, without requiring independence across questions. It does not provide simultaneous error control across candidates in Θ. Moreover, these tests assess question-wise marginal compatibility rather than joint distributional compatibility.

A.3 Additional discussions

The choice of Θ affects the availability of well-aligned candidates, but not the pointwise validity of each conformal test, provided that the candidate set is constructed independently of the target sample used for screening and the stated assumptions hold. Candidate controls may specify prompt-mixture weights or index prompts sampled from a structured space. Details are provided in Appendix B.1.

For more complex prompt spaces, a structured representation can guide the sampling of a finite candidate set, enabling broader exploration without exhaustive enumeration. The resulting diversity depends on the representation and sampling scheme. Appendix B.2 provides further details.

B Candidate Prompt-Control Construction

This section details the data-splitting procedure and the construction of the candidate parameter set Θ. Figure 4 illustrates how candidate prompts are constructed for MovieLens.

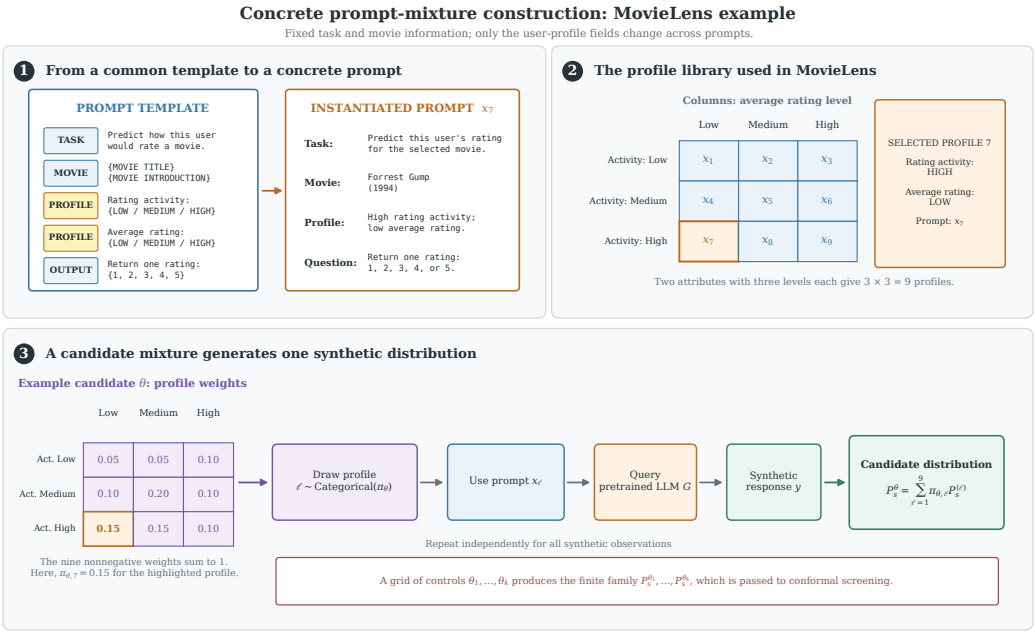


Figure 4: An concrete example to illustrate the prompt candidate construction: MovieLens example

B.1 Data Splitting

Candidate prompts may be constructed using domain knowledge, external data, expert input, or other information independent of the target sample used for screening. When collected target data inform prompt design or calibration, we partition them into disjoint development and evaluation samples, with the split chosen independently of the observed values: $\mathcal{Z}_t = \mathcal{Z}_t^{\text{dev}} \cup \mathcal{Z}_t^{\text{eval}}$, $\mathcal{Z}_t^{\text{dev}} \cap \mathcal{Z}_t^{\text{eval}} = \varnothing$. The development sample is used to construct Θ, which is fixed before screening. The evaluation sample is reserved for computing the target nonconformity score; accordingly, $n = |\mathcal{Z}_t^{\text{eval}}|$ in the subsequent analysis.

The proposed procedure does not require Θ to approximate the entire natural-language prompt space, as compatibility is tested separately for each candidate. Conditional on the development sample and the resulting candidate set, each test retains its pointwise finite-sample validity under the corresponding null and the stated assumptions. Thus, candidate construction affects the availability of well-aligned distributions, but not the validity of the individual tests.

B.2 Candidate Parameter Sampling

Notice that prompts provide a flexible means of controlling an LLM's response distribution. However, the space of possible prompts is prohibitively large for exhaustive search, even with a finite vocabulary and a fixed length limit. We therefore adopt a controlled sampling method: we define a structured prompt family indexed by a small number of control parameters and sample a finite candidate set Θ from the resulting parameter space. The objective is not to cover the entire prompt space, but to include candidates whose induced distributions approximate the target. This requires both a sufficiently expressive parameterization and a sampling scheme that can reach well-aligned candidates; increasing the sampling budget alone does not ensure their availability. The following assumptions formalize the conditions underlying this construction.

Assumption 3 *Let $(\Theta_{\text{feas}}, d_\Theta)$ be a compact metric space of feasible control parameters, and let $\mathcal{S}_\Theta$ denote the probability measure used to sample candidates. Assume that $\mathcal{S}_\Theta$ has full support on Θ_{feas}: for every $\theta \in \Theta_{\text{feas}}$ and $\epsilon > 0$,*

$$\mathcal{S}_\Theta(\{\theta' \in \Theta_{\text{feas}} : d_\Theta(\theta', \theta) < \epsilon\}) > 0.$$

Assumption 3 provides a high-probability ϵ-net guarantee for the sampled candidate set. When $\mathcal{S}_\Theta$ admits a density with respect to a reference measure having full support on Θ_{feas}, a density bounded away from zero suffices for the full-support condition. Fix $\epsilon > 0$. By compactness, Θ_{feas} admits a finite cover $B_1, \ldots, B_{N_{\epsilon/2}}$ by open balls of radius $\epsilon/2$ under d_Θ. The full-support assumption ensures that each ball has positive sampling probability, so $p_{\min} := \min_{1 \le j \le N_{\epsilon/2}} \mathcal{S}_\Theta(B_j) > 0$. Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ consist of k independent draws from $\mathcal{S}_\Theta$. By the union bound, the probability that at least one covering ball contains no sampled candidate satisfies

$$\mathbb{P}(\exists j : \Theta \cap B_j = \varnothing) \leqslant \sum_{j=1}^{N_{\epsilon/2}} (1 - \mathcal{S}_\Theta(B_j))^k \leqslant N_{\epsilon/2}(1 - p_{\min})^k,$$

which converges to zero as $k \to \infty$. Whenever every covering ball contains a sampled candidate, each feasible parameter θ shares a ball with some $\theta' \in \Theta$. The triangle inequality then gives $d_\Theta(\theta, \theta') < \epsilon$. Thus, for every fixed $\epsilon > 0$, the sampled candidate set Θ forms an ϵ-net of Θ_{feas} with probability tending to one as the sampling budget increases.

C Discrepancy Measures

C.1 Examples of f-Divergences

We consider the following special cases of the f-divergence defined in Eq. (2). Let p_t and p_s denote the densities of P_t and P_s with respect to a common dominating measure μ.

- **Kullback–Leibler (KL) divergence.** Choosing $f_{\text{KL}}(x) = x \log x$, with $f_{\text{KL}}(0) = 0$, yields
$$D_{\text{KL}}(P_t \| P_s) = \int_\Omega p_t(z) \log \frac{p_t(z)}{p_s(z)} \, d\mu(z).$$
KL divergence is generally asymmetric and measures the expected log-density ratio under P_t.
- **Total variation (TV) distance.** Choosing $f_{\text{TV}}(x) = |x - 1|/2$ yields
$$D_{\text{TV}}(P_t \| P_s) = \frac{1}{2} \int_\Omega |p_t(z) - p_s(z)| \, d\mu(z).$$
TV distance is symmetric and satisfies $0 \leqslant D_{\text{TV}}(P_t \| P_s) \leqslant 1$.
- **Pearson χ^2 divergence.** Choosing $f_{\chi^2}(x) = (x - 1)^2$ yields
$$D_{\chi^2}(P_t \| P_s) = \int_\Omega \frac{(p_t(z) - p_s(z))^2}{p_s(z)} \, d\mu(z).$$
χ^2 Divergence is generally asymmetric and weights squared density differences by the inverse synthetic density.

C.2 Empirical f-Divergences

We estimate distributional discrepancies by applying a common smoothing kernel to the target and synthetic empirical measures.

Definition 1 *Let $K : \mathbb{R}^d \to (0, \infty)$ be a strictly positive kernel satisfying $\int_{\mathbb{R}^d} K(z)\, dz = 1$. Given a target sample of size n and a synthetic reference sample of size N, define the smoothed empirical densities*

$$\widehat{p}_{t,K}(z) := \frac{1}{n} \sum_{i=1}^{n} K(z - z_{t,i}), \qquad \widehat{p}^{\,\theta}_{s,K}(z) := \frac{1}{N} \sum_{j=1}^{N} K(z - z^{\theta}_{s,j}).$$

The smoothed empirical f-divergence is

$$D_{f,n,N}\big(P_{t,n} \| P^{\theta}_{s,N}; K\big) := \int_{\mathbb{R}^d} f\left(\frac{\widehat{p}_{t,K}(z)}{\widehat{p}^{\,\theta}_{s,K}(z)}\right) \widehat{p}^{\,\theta}_{s,K}(z)\, dz. \tag{5}$$

Throughout, we use the Gaussian kernel

$$\varphi_\sigma(x) = (2\pi\sigma^2)^{-d/2} \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}^d, \tag{6}$$

where $\sigma > 0$ is a fixed bandwidth and $\|\cdot\|_2$ denotes the Euclidean norm. We use $*$ to denote convolution, with

$$(P * \varphi_\sigma)(x) = \int_\Omega \varphi_\sigma(x - z)\, dP(z), \qquad x \in \mathbb{R}^d.$$

Gaussian smoothing yields strictly positive densities, ensuring that the density ratio is well-defined. We omit the kernel argument in $D_{f,n,N}$ when the Gaussian kernel is understood.

The KL, TV, and χ^2 choices introduced above yield nonconformity scores under this common smoothing construction. For a fixed divergence and bandwidth, smaller scores indicate closer agreement between the smoothed empirical distributions. Under independent i.i.d. sampling and the compact-support condition in Assumption 1, these estimators are consistent, as $n, N \to \infty$, for the corresponding population smoothed divergence Nguyen et al. (2010). We use the smoothed empirical divergence as a distribution-level nonconformity score. Conditional exchangeability and randomized ranking ensure finite-sample calibration, whereas the choice of divergence, smoothing bandwidth, and estimation accuracy determine the screening test's power.

D Supplementary Theoretical Analysis

We use the shorthand $S_t^* := D_{\mathrm{KL}}\big(P_t \| P_s^\theta; \varphi_\sigma\big) = D_{\mathrm{KL}}\big(P_t * \varphi_\sigma \,\big\|\, P_s^\theta * \varphi_\sigma\big)$ for the Gaussian-smoothed nonconformity score, where φ_σ is the mollifier function with fixed bandwidth $\sigma > 0$.

D.1 EFD Estimation Theory

Fix a candidate $\theta \in \Theta$. For notational simplicity, we suppress the superscript θ on the synthetic reference and calibration samples, writing $\mathcal{Z}_s$ and $\mathcal{Z}_{s,i}$, respectively. Throughout this subsection, define

$$r_{n,N} := \left(\frac{1}{n} + \frac{1}{N}\right)^{-1} = \frac{nN}{n+N}, \qquad \rho_{n,N} := \frac{N}{n+N}.$$

Theorem 4 *Suppose Assumption 1 holds and the Gaussian bandwidth $\sigma > 0$ is fixed. Let the target and synthetic reference samples consist of n and N i.i.d. observations from P_t and P_s^θ, respectively, with the two samples independent. If $n, N \to \infty$ and $\rho_{n,N} \to \rho \in (0,1)$, then*

$$r_{n,N}^{1/2}\left[D_{\mathrm{KL},n,N}\big(P_{t,n} \| P_{s,N}^\theta; \varphi_\sigma\big) - S_t^*\right] \xrightarrow{d} \mathcal{N}\big(0, v_{2,\mathrm{KL}}^2(\rho, P_t, P_s^\theta, \sigma)\big), \tag{7}$$

where S_t^* is the population Gaussian-smoothed KL divergence defined above. The asymptotic variance is

$$v_{2,\mathrm{KL}}^2(\rho, P_t, P_s^\theta, \sigma) := \sum_{i,j=1}^{2} c_{i,j}(\rho) \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Sigma^{(i,j)}_{P_t, P_s^\theta, \sigma}(x,y) L_{i,\mathrm{KL}}(x) L_{j,\mathrm{KL}}(y)\, dx\, dy,$$

with coefficients $c_{1,1}(\rho) = \rho,\ c_{2,2}(\rho) = 1-\rho,\ c_{1,2}(\rho) = c_{2,1}(\rho) = \sqrt{\rho(1-\rho)}$, and derivative functions

$$L_{1,\mathrm{KL}}(x) := \log\left(\frac{(P_t * \varphi_\sigma)(x)}{(P_s^\theta * \varphi_\sigma)(x)}\right), \quad L_{2,\mathrm{KL}}(x) := -\frac{(P_t * \varphi_\sigma)(x)}{(P_s^\theta * \varphi_\sigma)(x)}.$$

The covariance function is

$$\Sigma_{P_t, P_s^\theta, \sigma}(x,y) := \begin{bmatrix} \mathrm{Cov}(\varphi_\sigma(x-Y), \varphi_\sigma(y-Y)) & 0 \\ 0 & \mathrm{Cov}(\varphi_\sigma(x-X), \varphi_\sigma(y-X)) \end{bmatrix}, \tag{8}$$

where $Y \sim P_t$ and $X \sim P_s^\theta$ are independent, and $\Sigma^{(i,j)}_{P_t, P_s^\theta, \sigma}$ denotes the (i,j)th entry.

Proof of the Theorem 4: We adapt the two-sample alternative argument in Proposition 2(iv) of Sreekumar et al. (2023) to independent samples of unequal sizes. The proof of Eq. (7) proceeds in three steps. First, we verify the density-ratio and integrability conditions required by Theorem 2(iv) of the same reference. Second, we apply their Theorem 6 to establish weak convergence of the two empirical processes in the appropriate weighted L^2 spaces; independence then yields joint convergence. Finally, we apply the functional delta method and identify the asymptotic variance.

We begin by verifying that the population and empirical Gaussian-smoothed KL divergences are well-defined and finite. Strict positivity of the Gaussian kernel ensures that all smoothed measures have strictly positive densities on $\mathbb{R}^d$ and are therefore mutually absolutely continuous. To establish finiteness, note that Assumption 1 places the supports of P_t, P_s^θ, and their empirical counterparts within the fixed compact set Ω, almost surely. Applying Jensen's inequality to the jointly convex KL functional yields:

$$\begin{aligned}\mathbb{E}\big[D_{\mathrm{KL},n,N}\big(P_{t,n}\|P_{s,N}^\theta,\varphi_\sigma\big)\big] &= \mathbb{E}\Big[D_{\mathrm{KL}}\Big(\mathbb{E}_{Y\sim P_{t,n}}\big[\mathcal{N}(Y,\sigma^2 I_d)\big],\mathbb{E}_{X\sim P_{s,N}^\theta}\big[\mathcal{N}(X,\sigma^2 I_d)\big]\Big)\Big]\\ &\leqslant \mathbb{E}_{X\sim P_s^\theta, Y\sim P_t}\Big[D_{\mathrm{KL}}\Big(\mathcal{N}(Y,\sigma^2 I_d),\mathcal{N}(X,\sigma^2 I_d)\Big)\Big]\\ &= \frac{1}{2\sigma^2}\mathbb{E}_{X\sim P_s^\theta, Y\sim P_t}\big[\|X-Y\|^2\big] < \infty,\end{aligned}$$

Furthermore, based on that there exists a finite constant $R<\infty$ such that, $\|X_j\|\leqslant R,\quad \|Y_i\|\leqslant R$ almost surely, we obtain

$$\left|\frac{P_{s,N}^\theta * \varphi_\sigma(x)}{P_{t,n} * \varphi_\sigma(x)}\right| = \frac{\frac{1}{N}\sum_{i=1}^N e^{-\frac{\|x-X_i\|^2}{2\sigma^2}}}{\frac{1}{n}\sum_{i=1}^n e^{-\frac{\|x-Y_i\|^2}{2\sigma^2}}} \leqslant \exp\left(\frac{R^2}{2\sigma^2}+\frac{2R}{\sigma^2}\|x\|\right)\leqslant c_1 e^{c_2\|x\|}, \tag{9}$$

for some constants $c_1, c_2 > 0$ that depend only on σ and $\mathrm{supp}(P_{s,N}^\theta)$.

Moreover, Jensen's inequality together with the expansion used above yields the two bounds needed below. Write $C_\sigma = (2\pi\sigma^2)^{-d/2}$ and let R be the radius of a ball containing $\mathrm{supp}(P_t)\cup\mathrm{supp}(P_s^\theta)$. Expanding $\|x-y\|^2 = \|x\|^2 - 2\langle x,y\rangle + \|y\|^2$ and bounding $|\langle x,y\rangle|\leqslant R\|x\|$, we have

$$P_s^\theta * \varphi_\sigma(x) \;\leqslant\; C_\sigma\, e^{-\frac{\|x\|^2}{2\sigma^2}+\frac{R\|x\|}{\sigma^2}},\qquad P_t * \varphi_\sigma(x) \;\geqslant\; C_\sigma\, e^{-\frac{\|x\|^2}{2\sigma^2}-\frac{R\|x\|}{\sigma^2}-\frac{R^2}{2\sigma^2}}.$$

Dividing the two cancels the Gaussian factor and leaves a linear exponent. Applying Jensen's inequality to the numerator before dividing, so that $\big(\int e^{\frac{2\langle x,y\rangle-\|y\|^2}{2\sigma^2}}\,dP_s^\theta\big)^2 \;\leqslant\; \int e^{\frac{2\langle x,y\rangle-\|y\|^2}{\sigma^2}}\,dP_s^\theta$, retains one factor $e^{-\|x\|^2/(2\sigma^2)}$, which dominates any linear exponent. Together,

$$\frac{P_s^\theta * \varphi_\sigma(x)}{P_t * \varphi_\sigma(x)} \;\leqslant\; c_5\, e^{c_6\|x\|},\qquad \frac{\big(P_s^\theta * \varphi_\sigma(x)\big)^2}{P_t * \varphi_\sigma(x)} \;\leqslant\; c_3\, e^{-\frac{\|x\|^2}{2\sigma^2}+c_4\|x\|} \;\leqslant\; \tilde{c},$$

for constants $c_3, c_4, c_5, c_6, \tilde{c} > 0$ that depend only on σ and on the supports of P_t and P_s^θ. Hence $\big\|(P_s^\theta * \varphi_\sigma)^2/(P_t * \varphi_\sigma)\big\|_\infty < \infty$, and both $P_t * \varphi_\sigma$ and $\log\big(P_t * \varphi_\sigma / P_s^\theta * \varphi_\sigma\big)$ lie in $L^2(P_s^\theta * \varphi_\sigma)$. The same inequalities hold with P_t and P_s^θ interchanged.

The purpose of the next construction is to specify the weighted L^2 spaces in which the two centered smoothed empirical-density processes must converge before the functional delta method can be applied. These spaces are inherited directly from the two-sample alternative result, Theorem 2(iv) of Sreekumar et al.Sreekumar et al. (2023). In their Theorem 2(iv), η_1 and η_2 have relative densities $1 + 1/p_\mu$ and $1 + p_\mu + q_{\mathrm{env}}$, respectively, with respect to the reference measure. Here in our scenario, the reference measure has Lebesgue density $P_s^\theta * \varphi_\sigma$, $p_\mu(x) = P_t * \varphi_\sigma(x)/P_s^\theta * \varphi_\sigma(x)$, and equation (9) provides $q_{\mathrm{env}}(x) = c_1 e^{c_2\|x\|}$. Therefore, the derivatives of the two measures become

$$\begin{aligned}\frac{d\eta_1}{dx}(x) &= P_s^\theta * \varphi_\sigma(x) + \frac{\big(P_s^\theta * \varphi_\sigma(x)\big)^2}{P_t * \varphi_\sigma(x)},\\ \frac{d\eta_2}{dx}(x) &= P_s^\theta * \varphi_\sigma(x) + c' P_s^\theta * \varphi_\sigma(x) e^{c\|x\|} + P_t * \varphi_\sigma(x),\end{aligned}$$

and hence are finite measures based on the inequalities above. Then, with $r_{n,N} = (1/n + 1/N)^{-1}$, we claim that

$$\begin{aligned} & r_{n,N}^{\frac{1}{2}} \left(D_{\mathrm{KL},n,N} \left(P_{t,n} \| P_{s,N}^{\theta}, \varphi_\sigma \right) - D_{\mathrm{KL}}^{*} \left(P_t \| P_s^{\theta}, \varphi_\sigma \right) \right) \\ & \xrightarrow{d} \sqrt{\rho} \int_{\mathbb{R}^d} G_{P_t,\sigma}(x) \log \left(\frac{P_t * \varphi_\sigma(x)}{P_s^{\theta} * \varphi_\sigma(x)} \right) dx - \sqrt{1-\rho} \int_{\mathbb{R}^d} G_{P_s^{\theta},\sigma}(x) \frac{P_t * \varphi_\sigma(x)}{P_s^{\theta} * \varphi_\sigma(x)} dx. \end{aligned} \tag{10}$$

By the functional delta method applied to the Kullback–Leibler functional at $\left(P_t * \varphi_\sigma,\, P_s^{\theta} * \varphi_\sigma\right)$, the claim (10) reduces to the two weak-convergence statements

$$r_{n,N}^{1/2} \left(\frac{P_{t,n} * \varphi_\sigma}{P_s^{\theta} * \varphi_\sigma} - \frac{P_t * \varphi_\sigma}{P_s^{\theta} * \varphi_\sigma} \right) \xrightarrow{w} \sqrt{\rho}\, \frac{G_{P_t,\sigma}}{P_s^{\theta} * \varphi_\sigma} \quad \text{in } L^2(\eta_1), \tag{11}$$

$$r_{n,N}^{1/2} \left(\frac{P_{s,N}^{\theta} * \varphi_\sigma}{P_s^{\theta} * \varphi_\sigma} - 1 \right) \xrightarrow{w} \sqrt{1-\rho}\, \frac{G_{P_s^{\theta},\sigma}}{P_s^{\theta} * \varphi_\sigma} \quad \text{in } L^2(\eta_2). \tag{12}$$

It remains to verify the conditions under which (11) and (12) hold. By the L^2-valued CLT in Theorem 6 of Sreekumar et al., it suffices to verify the following second-moment integrability conditions,

$$\int_{\mathbb{R}^d} \frac{\mathrm{Var}_{P_t} \left(\varphi_\sigma(x - \cdot) \right)}{P_s^{\theta} * \varphi_\sigma(x)} dx + \int_{\mathbb{R}^d} \frac{\mathrm{Var}_{P_t} \left(\varphi_\sigma(x - \cdot) \right)}{P_t * \varphi_\sigma(x)} dx < \infty, \tag{13}$$

and

$$\int_{\mathbb{R}^d} \frac{\mathrm{Var}_{P_s^{\theta}} \left(\varphi_\sigma(x - \cdot) \right)}{P_s^{\theta} * \varphi_\sigma(x)} \left(1 + e^{c\|x\|} \right) dx + \int_{\mathbb{R}^d} \frac{\mathrm{Var}_{P_s^{\theta}} \left(\varphi_\sigma(x - \cdot) \right) P_t * \varphi_\sigma(x)}{\left(P_s^{\theta} * \varphi_\sigma(x) \right)^2} dx < \infty, \tag{14}$$

To see this, note that the second term in (13) and the first term in (14) are both controlled by

$$\int_{\mathbb{R}^d} \frac{\mathrm{Var}_P \left\{ \varphi_\sigma(x - \cdot) \right\}}{P * \varphi_\sigma(x)} \left(1 + e^{c\|x\|} \right) dx < \infty, \quad P \in \left\{ P_t, P_s^{\theta} \right\}. \tag{15}$$

Neither the argument nor the constants depend on which of the two measures is taken, so we verify (15) for a generic P supported in a ball of radius R. Writing $C_\sigma = (2\pi\sigma^2)^{-d/2}$ and bounding the variance by the second moment,

$$\begin{aligned} \frac{\mathrm{Var}_P \left(\varphi_\sigma(x - \cdot) \right)}{P * \varphi_\sigma(x)} \;\leqslant\; \frac{\mathbb{E}_P \left[\varphi_\sigma^2(x - \cdot) \right]}{P * \varphi_\sigma(x)} \;&=\; C_\sigma \, \frac{\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2}{\sigma^2}} \, dP(y)}{\int_{\mathbb{R}^d} e^{-\frac{\|x-y\|^2}{2\sigma^2}} \, dP(y)} \\ &=\; C_\sigma \, e^{-\frac{\|x\|^2}{2\sigma^2}} \, \frac{\int_{\mathbb{R}^d} e^{\frac{2\langle x,y \rangle - \|y\|^2}{\sigma^2}} \, dP(y)}{\int_{\mathbb{R}^d} e^{\frac{2\langle x,y \rangle - \|y\|^2}{2\sigma^2}} \, dP(y)} \;\leqslant\; C_\sigma \, e^{-\frac{\|x\|^2}{2\sigma^2}} \, \frac{e^{\frac{2R\|x\|}{\sigma^2}}}{e^{-\frac{2R\|x\|+R^2}{2\sigma^2}}} \\ &=\; C_\sigma \, e^{\frac{R^2}{2\sigma^2}} \, e^{-\frac{\|x\|^2}{2\sigma^2} + \frac{3R\|x\|}{\sigma^2}}. \end{aligned}$$

Multiplying by $1 + e^{c\|x\|} \leqslant 2e^{c\|x\|}$ and integrating,

$$\int_{\mathbb{R}^d} \frac{\mathrm{Var}_P \left\{ \varphi_\sigma(x - \cdot) \right\}}{P * \varphi_\sigma(x)} \left(1 + e^{c\|x\|} \right) dx \;\leqslant\; 2\, C_\sigma \, e^{\frac{R^2}{2\sigma^2}} \int_{\mathbb{R}^d} e^{-\frac{\|x\|^2}{2\sigma^2} + \bar{c}\,\|x\|} \, dx \;<\; \infty, \qquad \bar{c} := c + \frac{3R}{\sigma^2},$$

The last integral is finite because the Gaussian factor dominates any linear exponent.

For the remaining terms, we have

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{\mathrm{Var}_{P_s^{\theta}} \left(\varphi_\sigma(x - \cdot) \right) P_t * \varphi_\sigma(x)}{\left(P_s^{\theta} * \varphi_\sigma(x) \right)^2} dx &\leqslant c' \int_{\mathbb{R}^d} e^{-\frac{\|x\|^2}{2\sigma^2}} e^{c\|x\|} \frac{P_t * \varphi_\sigma(x)}{P_s^{\theta} * \varphi_\sigma(x)} dx \\ &\leqslant c'^2 \int_{\mathbb{R}^d} e^{-\frac{\|x\|^2}{2\sigma^2}} e^{2c\|x\|} dx < \infty, \end{aligned}$$

and

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{\mathrm{Var}_{P_t} \left(\varphi_\sigma(x - Y) \right)}{P_s^{\theta} * \varphi_\sigma(x)} dx &\leqslant \int_{\mathbb{R}^d} \frac{\mathbb{E}_{P_t} \left[\varphi_\sigma^2(x - Y) \right]}{P_s^{\theta} * \varphi_\sigma(x)} dx \\ &\leqslant c' \int_{\mathbb{R}^d} e^{-\frac{\|x\|^2}{2\sigma^2}} e^{c\|x\|} dx < \infty. \end{aligned}$$

Hence all four integrability requirements in (13) and (14) are satisfied, and Theorem 6 yields the weak convergences in (11) and (12). Finally, the proof of the claim is completed by noting a zero-mean Gaussian random variable with variance

$$v_{2,KL}^2(\rho, \mathcal{Z}_s, \mathcal{Z}_t, \sigma) := \sum_{1\leqslant i,j\leqslant 2} \int\int c_{i,j}(\rho)\Sigma^{(i,j)}_{P_t,P_s^\theta,\sigma}(x,y)L_{i,KL}(x)L_{j,KL}(y)dxdy.$$
$$= \sum_{1\leqslant i,j\leqslant 2} \int\int c_{i,j}(\rho)\Sigma^{(i,j)}_{P_t,P_s^\theta,\sigma}(x,y)\tilde{L}_{i,KL}(x)\tilde{L}_{j,KL}(y)dxdy,$$

where

$$L_{1,f}(x) := L_{1,f,P_t,P_s^\theta,\sigma}(x) := f'\left(\frac{P_t * \varphi_\sigma(x)}{P_s^\theta * \varphi_\sigma(x)}\right),$$
$$L_{2,f}(x) := L_{2,f,P_t,P_s^\theta,\sigma}(x) := f\left(\frac{P_t * \varphi_\sigma(x)}{P_s^\theta * \varphi_\sigma(x)}\right) - \frac{P_t * \varphi_\sigma(x)}{P_s^\theta * \varphi_\sigma(x)} f'\left(\frac{P_t * \varphi_\sigma(x)}{P_s^\theta * \varphi_\sigma(x)}\right).$$

By taking $f(x) = x\log(x)$, we have $\tilde{L}_{1,KL} := \log\left(P_t * \varphi_\sigma / P_s^\theta * \varphi_\sigma\right)$ and $\tilde{L}_{2,KL} := -P_t * \varphi_\sigma / P_s^\theta * \varphi_\sigma$. We leave out the constant part of the $L_{1,KL}$ to $\tilde{L}_{1,KL}$, as

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \Sigma^{(1,1)}_{P_t,P_s^\theta,\sigma}(x,y)dxdy = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \operatorname{cov}\left(\varphi_\sigma(x-Y), \varphi_\sigma(y-Y)\right)dxdy$$
$$= \operatorname{Var}\left(\int_{\mathbb{R}^d} \varphi_\sigma(x-Y)\,dx\right) = 0.$$

As defined in equation ((8)), $\Sigma^{(i,j)}_{P_t,P_s^\theta,\sigma}$ is the covariance matrix of the 2-dimensional centered Gaussian process $(G_{P_t,\sigma}, G_{P_s^\theta,\sigma})$. Since the target sample and the synthetic sample are independent,

$$\Sigma^{(1,2)}_{P_t,P_s^\theta,\sigma}(x,y) = \Sigma^{(2,1)}_{P_t,P_s^\theta,\sigma}(x,y) = 0.$$

Therefore, the asymptotic variance equivalently reduces to

$$v_{2,KL}^2(\rho, P_t, P_s^\theta, \sigma) = \rho \iint \Sigma^{(1,1)}_{P_t,P_s^\theta,\sigma}(x,y)L_{1,KL}(x)L_{1,KL}(y)\,dx\,dy$$
$$+ (1-\rho)\iint \Sigma^{(2,2)}_{P_t,P_s^\theta,\sigma}(x,y)L_{2,KL}(x)L_{2,KL}(y)\,dx\,dy.$$

Meanwhile from the convergence rate in equation (11) and equation (12), we obtain the corresponding coefficient $c_{i,j}(\rho)$ for each term, $c_{1,1} = \rho, c_{1,2} = c_{2,1} = \sqrt{\rho(1-\rho)}, c_{2,2} = 1-\rho$. □

Theorem 4 provides a direct characterization of the convergence rate of the smoothed KL estimator. The convergence rate is given by $r_{n,N} = (1/n + 1/N)^{-1}$, which depends on n and N. As the number of synthetic samples N becomes substantially larger than that of the target samples n, ρ approaches 1. Under this condition, the efficiency of the divergence estimator obtained by the EFD approach is dominated by the term $c_{1,1}$, while the convergence rate is determined by the target sample size n.

Moreover, by imposing stronger regularity conditions, a more precise characterization of the tail probabilities of the estimation bias exceeding ϵ associated with the smoothed KL estimator's convergence can be derived.

Condition 1 (Uniform density-ratio bounds) *There exist constants* $0 < \underline{r} \leq \overline{r} < \infty$ *such that*

$$\underline{r} \leq \frac{(P_t * \varphi_\sigma)(z)}{(P_s^\theta * \varphi_\sigma)(z)} \leq \overline{r}, \qquad \forall z \in \Omega.$$

Lemma 5 *Under Condition 1, Let* $r_{n,N} = (1/n + 1/N)^{-1} = nN/(n+N)$*, and let*

$$b_{n,N} = \left|\mathbb{E}(D_{f,n,N}(P_t\|P_s^\theta, \varphi_\sigma)) - D_f^*(P_t\|P_s^\theta, \varphi_\sigma)\right|.$$

If $f \in C^1(I)$ and denote

$$\kappa_f = \sup_{r\in I} \max\{|f'(r)|,\ |f(r) - rf'(r)|\} < \infty.$$

Then, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|D_{f,n,N}(P_t\|P_s^\theta, \varphi_\sigma) - D_f^*(P_t\|P_s^\theta, \varphi_\sigma)\right| > b_{n,N} + \varepsilon\right) \leqslant 2\exp\left(-cr_{n,N}\varepsilon^2\right).$$

Proof of Lemma 5: Consider the f-divergence integrand:

$$\Phi(a, b) = f(\frac{a}{b})b,$$

where a stands for some $p_t * \varphi_\sigma$ and b stands for some $P_s^\theta * \varphi_\sigma$. Denote $r = a/b$, direct differentiation gives

$$\frac{\partial\Phi}{\partial a}(a, b) = f'(r), \quad \frac{\partial\Phi}{\partial b}(a, b) = f(r) - rf'(r).$$

Let $I := [\underline{r}, \overline{r}]$, (a, b) and (a', b') be two admissible pairs, where $a/b, a'/b' \in I$. For $\lambda \in [0, 1]$,

$$a_\lambda = \lambda a + (1 - \lambda)a',$$
$$b_\lambda = \lambda b + (1 - \lambda)b'.$$

Then

$$\frac{a_\lambda}{b_\lambda} = \omega_\lambda \frac{a}{b} + (1 - \omega_\lambda)\frac{a'}{b'}, \quad \omega_\lambda = \frac{\lambda b}{\lambda b + (1 - \lambda)b'} \in [0, 1].$$

It follows that $a_\lambda/b_\lambda \in I$, and by the mean-value formula

$$\begin{aligned} &|\Phi(a, b) - \Phi(a', b')| \\ &\leqslant \int_0^1 \left| f'\left(\frac{a_\lambda}{b_\lambda}\right)(a - a') + \left[f\left(\frac{a_\lambda}{b_\lambda}\right) - \frac{a_\lambda}{b_\lambda}f'\left(\frac{a_\lambda}{b_\lambda}\right)\right](b - b')\right| d\lambda \\ &\leqslant \kappa_f\left(|a - a'| + |b - b'|\right). \end{aligned} \tag{16}$$

We next verify the bounded-difference condition. If one target sample observation Y_i is replaced by another admissible value Y_i', then $p_{t,n} * \varphi_\sigma$ changes. Denote the resulting empirical density by $p_{t,n}^{(i)} * \varphi_\sigma$. Since φ_σ is nonnegative and integrates to one,

$$\begin{aligned} \left\|p_{t,n} * \varphi_\sigma - p_{t,n}^{(i)} * \varphi_\sigma\right\|_{L^1(\mathcal{Z})} &\leqslant \frac{1}{n}\int_\Omega |\varphi_\sigma(z - Y_i) - \varphi_\sigma(z - Y_i')|\, dz \\ &\leqslant \frac{2}{n}. \end{aligned} \tag{17}$$

Combining (16) and (17), replacing the target sample Y_i changes $D_{f,n,N}(P_t\|P_s^\theta, \varphi_\sigma)$ by at most $2\kappa_f/n$. Similarly to the denotes in (17), if one synthetic sample X_j is replaced, then

$$\left\|P_{s,N}^\theta * \varphi_\sigma - P_{s,N}^{\theta,(j)} * \varphi_\sigma\right\|_{L^1(\mathcal{Z})} \leqslant \frac{2}{N},$$

and replacing one synthetic observation changes $D_{f,n,N}(P_t, P_s^\theta, \varphi_\sigma)$ by at most $2\kappa_f/N$.

Hence, the variation of all the target samples and all synthetic samples is limited by

$$\sum_{i=1}^n (\frac{2\kappa_f}{n})^2 + \sum_{j=1}^N (\frac{2\kappa_f}{N})^2 = 4\kappa_f^2\left(\frac{1}{n} + \frac{1}{N}\right) = \frac{4\kappa_f^2}{r_{n,N}}.$$

McDiarmid's bounded-difference inequality gives, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|D_{f,n,N}(P_t\|P_s^\theta, \varphi_\sigma) - \mathbb{E}(D_{f,n,N}(P_t\|P_s^\theta, \varphi_\sigma))\right| > \varepsilon\right) \leqslant 2\exp\left(-\frac{r_{n,N}\varepsilon^2}{2\kappa_f^2}\right).$$

Finally, by the definition of $b_{n,N}$, we have finished the proof. □

Lemma 6 *Suppose the assumptions of Lemma 5 hold. In addition, If $f \in C^2(I)$ and*

$$M_f = \sup_{r \in I} |f''(r)| < \infty. \tag{18}$$

Assume also that the denominator smoothed densities are uniformly bounded away from zero: there exists $\ell_s > 0$ such that

$$P_s^\theta * \varphi_\sigma(z) \geqslant \ell_s, \qquad P_{s,N}^\theta * \varphi_\sigma(z) \geqslant \ell_s, \qquad \forall z \in \mathcal{Z}.$$

Then

$$b_{n,N} \leqslant c \left(\frac{1}{n} + \frac{1}{N} \right),$$

Consequently, combining with the lemma 5 for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left| D_{f,n,N}(P_t \| P_s^\theta, \varphi_\sigma) - D_f^*(P_t, P_s^\theta, \varphi_\sigma) \right| > c r_{n,N}^{-1} + \varepsilon \right) \leq 2 \exp\left(-c r_{n,N} \varepsilon^2 \right). \tag{19}$$

Proof of Lemma 6: Let $\Delta_t(z) := p_{t,n} * \varphi_\sigma(z) - p_t * \varphi_\sigma(z), \quad \Delta_s(z) := P_{s,N}^\theta * \varphi_\sigma(z) - P_s^\theta * \varphi_\sigma(z)$. Recall the notation in proof of lemma 5, $\Phi(a,b) = f \circ (a/b)b$. Its Hessian is

$$\nabla^2 \Phi(a,b) = \frac{f''(a/b)}{b} \begin{pmatrix} 1 & -a/b \\ -a/b & (a/b)^2 \end{pmatrix}.$$

By the ratio bound $a/b \in I$, the lower bound $b \geqslant \ell_s$, and the equation (18), The first-order Taylor terms have zero expectation,

$$\mathbb{E}\Delta_t(z) = 0, \quad \mathbb{E}\Delta_s(z) = 0.$$

The quadratic Taylor remainder satisfies

$$|R(z)| \leqslant \frac{M_f(1+\overline{r}^2)}{2\ell_s} \left[\Delta_t^2(z) + \Delta_s^2(z) \right].$$

Therefore,

$$\begin{aligned} b_{n,N} &= \left| \mathbb{E}(D_{f,n,N}(P_t \| P_s^\theta, \varphi_\sigma)) - D_f^*(P_t, P_s^\theta, \varphi_\sigma) \right| \\ &\leqslant \frac{M_f(1+\overline{r}^2)}{2\ell_s} \mathbb{E}(\int_\Omega \left[\Delta_t^2(z) + \Delta_s^2(z) \right] dz. \end{aligned}$$

For the target density estimator,

$$\begin{aligned} \mathbb{E} \int_{\mathcal{Z}} \Delta_t^2(z) dz &= \int_\Omega \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^n \varphi_\sigma(z - Y_i) \right) dz \\ &= \frac{1}{n} \int_\Omega \mathrm{Var}\left(\varphi_\sigma(z - Y_1) \right) dz \\ &\leqslant \frac{1}{n} \int_{\mathbb{R}^d} \mathbb{E}\varphi_\sigma^2(z - Y_1) dz = \frac{C_{\sigma,d}}{n}, \end{aligned} \tag{20}$$

where $C_{\sigma,d} = (4\pi\sigma^2)^{-d/2}$.

Similarly,

$$\mathbb{E} \int_{\mathcal{Z}} \Delta_s^2(z) dz \leqslant \frac{C_{\sigma,d}}{N}. \tag{21}$$

Combining (20) and (21) gives

$$b_{n,N} \leqslant \frac{M_f(1+\overline{r}^2)}{2\ell_s} C_{\sigma,d} \left(\frac{1}{n} + \frac{1}{N} \right),$$

which proves lemma 6. Inequality (19) follows immediately from Lemma 5.

Corollary 7 (Synthetic-rich resolution) *Suppose Assumption 1 and the conditions of Lemmas 5 and 6 hold. If $n, N \to \infty$ and $N/n \to \infty$, then*

$$r_{n,N} = \frac{nN}{n+N} = n\{1 + o(1)\},$$

More precisely, the stochastic component is $O_{\mathbb{P}}(n^{-1/2})$ and the bias is $O(n^{-1})$.

It follows from the concentration and bias bounds, which remain valid without requiring $\rho_{n,N}$ to stay away from one. Thus, in the synthetic-rich regime, additional synthetic observations yield diminishing improvements because the target sample determines the leading estimation scale.

With assumption 1, Ω is compact. All observations lie in Ω, and consider the Gaussian kernel φ_σ in equation (6). Then there exist constants $0 < \varphi_{\min} \leq \varphi_{\max} < \infty$ such that

$$\varphi_{\min} \leqslant \varphi_\sigma(z - z') \leqslant \varphi_{\max}, \quad \forall z, z' \in \mathcal{Z}.$$

Consequently, all smoothed empirical densities are bounded between $\varphi_{\min}$ and $\varphi_{\max}$ on Ω, and one may take

$$\underline{r} = \frac{\varphi_{\min}}{\varphi_{\max}}, \quad \overline{r} = \frac{\varphi_{\max}}{\varphi_{\min}}.$$

we have established Condition 1 using Assumption 1. integration domain is $\mathbb{R}^d$ rather than a compact Ω, compact support alone does not generally imply such a uniform empirical ratio bound in the Gaussian tails; in that case, the ratio bound 1 must be imposed explicitly or the EFD should be evaluated on a compact truncation region.

For the KL case $f(r) = r \log r$, one has

$$f'(r) = 1 + \log r, \quad f''(r) = \frac{1}{r}.$$

Hence, under the ratio bound condition 1,

$$\kappa_{\mathrm{KL}} = \max\left\{ \sup_{r \in [\underline{r}, \overline{r}]} \|1 + \log r\|, \overline{r} \right\}, \quad M_{\mathrm{KL}} = \frac{1}{\underline{r}}.$$

Therefore, we provide a more refined characterization of inequality (19) as

$$\mathbb{P}\left(\left| D_{\mathrm{KL},n,N}(P_{t,n} \| P^\theta_{s,N}, \varphi_\sigma) - D^*_{\mathrm{KL}}(P_t \| P^\theta_s, \varphi_\sigma) \right| > c_1 r^{-1}_{n,N} + \varepsilon \right) \leq 2 \exp\left(-c_2 r_{n,N} \varepsilon^2 \right). \quad (22)$$

Lastly, we observe that the estimation efficiency of EFD relies primarily on the functions L_1 and L_2 for different f-divergence functions $f(x)$. The first-order derivative components of the f-divergence functional are $L_{1,f}(r) = f'(r)$ and $L_{2,f}(r) = f(r) - rf'(r)$. A bounded second derivative f'' is additionally required for the second-order bias analysis. Then we can construct the asymptotic variance of the EFD estimator, while the convergence rate is preserved as $r_{n,N}$. Notably, for the TV distance, the corresponding f-function is not twice differentiable, the first-order derivation is $f'_{TV}(x) = 1/2 * sgn(x - 1)$, which is not differentiable at $x = 1$, additional nonsmooth empirical-process arguments would be required. In contrast, for the χ^2 distance, the substitution using $L_{1,\chi^2} = 2(x - 1)$ and $L_{2,\chi^2} = 1 - x^2$ is applicable.

D.2 Additional Risk-Control Discussion

In this section, we will discuss the main issue of synthetic data, which is the probability risk that the generated synthetic data does not match the distribution of the target data. We analyze the risk level of the control parameters obtained by the LLM-Based Synthetic Data Parameter Selection Algorithm during data synthesis, and prove that the risk of the synthetic data generated by the selected control parameters can be effectively controlled below α.

D.3 Proof of Pointwise Finite-Sample Validity

Proof of Theorem 1: To prove Theorem 1, we build on the results of Vovk (2019). Specifically, we adopt Lemma 11 from that work and restate the definition of the function $A(\cdot)$ ($A(\cdot)$ is defined in Definition 7 in Vovk (2019)) as

$$A\left(\left(\mathcal{Z}_{s,1}, D\left(P^\theta_{s,1} \| P^\theta_{s,N} \right) \right), \ldots, \left(\mathcal{Z}_{s,m}, D\left(P^\theta_{s,m} \| P^\theta_{s,N} \right) \right), \left(\mathcal{Z}_t, D\left(P_{t,n} \| P^\theta_{s,N} \right) \right) \right) = \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\left(D\left(P^\theta_{s,i} \| P^\theta_{s,N} \right) < D\left(P_{t,n} \| P^\theta_{s,N} \right) \right).$$

To simplify notation, we use the previously defined S_i^θ and S_t^θ in place of $D(P_{s,i}^\theta \| P_{s,N}^\theta)$ and $D(P_t \| P_{s,N}^\theta)$. From the definition, we can check that the function A has the permutation invariance. For a permutation $\varrho(i) = j$, where $i, j \in 1, 2, \cdots, m$,

$$\begin{aligned}
&A_t\left(\left(\mathcal{Z}_{s,1}, S_1^\theta\right), \ldots, \left(\mathcal{Z}_{s,m}, S_m^\theta\right), \left(\mathcal{Z}_t, S_t^\theta\right)\right) \\
=&A_t\left(\left(\mathcal{Z}_{s,\varrho(1)}, S_{\varrho(1)}^\theta\right), \ldots, \left(\mathcal{Z}_{s,\varrho(m)}, S_{\varrho(m)}^\theta\right), \left(\mathcal{Z}_t, S_t^\theta\right)\right) \\
=&\frac{1}{m}\left(\sum_{i=1}^{m} \mathbf{1}\left(S_i^\theta < S_t^\theta\right)\right)
\end{aligned}$$

Corresponding to $A(\cdot)$, the function that exchanges the i-th variable and the final variable is defined as A_i

$$\begin{aligned}
A_i\left(\left(\mathcal{Z}_{s,1}, S_1^\theta\right), \ldots, \left(\mathcal{Z}_t, S_t^\theta\right), \cdots, \left(\mathcal{Z}_{s,m}, S_m^\theta\right), \left(\mathcal{Z}_{s,i}, S_i^\theta\right)\right) = \\
\frac{1}{m}\left(\sum_{j=1, j\neq i}^{m} \mathbf{1}\left(S_j^\theta < S_i^\theta\right) + \mathbf{1}\left(S_t^\theta < S_i^\theta\right)\right)
\end{aligned}$$

Subsequently, we may verify the monotonicity of the function $A(\cdot)$ with respect to S_i^θ and S_t^θ.

$$A\left(\left(\mathcal{Z}_{s,1}, S_1^\theta\right), \ldots, \left(\mathcal{Z}_{s,m}, S_m^\theta\right), \left(\mathcal{Z}_t, S_t^\theta\right)\right)$$

monotonically increase in S_t^θ and monotonically decrease in S_i^θ.

Furthermore, we verify whether $A(\cdot)$ satisfies the following conditions. Suppose the monotonic $A(\cdot)$ satisfies the following three conditions under the null hypothesis:

- for any synthetic sample $\mathcal{Z}_s$, calibration samples $(\mathcal{Z}_{s,1}, \ldots, \mathcal{Z}_{s,n})$, and target sample $\mathcal{Z}_t$, for all $i \in \{1, 2, \cdots, m\}$,
$$\begin{aligned}
\inf_{\mathcal{Z}_{s,i}} S_i^\theta &= \inf_{\mathcal{Z}_t} S_t^\theta, \\
\sup_{\mathcal{Z}_{s,i}} S_i^\theta &= \sup_{\mathcal{Z}_t} S_t^\theta;
\end{aligned} \tag{23}$$
- the $\inf_{\mathcal{Z}_{s,i}}$ and $\inf_{\mathcal{Z}_t}$ in Eq.(23) is either attained for all $(\mathcal{Z}_{s,1}, \ldots, \mathcal{Z}_{s,n})$ and $\mathcal{Z}_t$ or not attained;
- the $\sup_{\mathcal{Z}_{s,i}}$ and $\sup_{\mathcal{Z}_t}$ in Eq.(23) is either attained for all $(\mathcal{Z}_{s,1}, \ldots, \mathcal{Z}_{s,n})$ and $\mathcal{Z}_t$ or not attained;

The conditions of the lemma 11 can be verified readily. Under the null hypothesis, the distribution of the calibration sample from P_s^θ of size n, is identical to the sample-set from P_t, the distribution of the target sample set, which implies that the range of S_i^θ and S_t^θ coincide. Then the conformal transducer corresponding to A have the properties in Theorem 1, where the conformal transducer of A is defined as

$$\begin{aligned}
Q^\theta(\mathcal{Z}_s, \mathcal{Z}_{s,1}, \ldots, \mathcal{Z}_{s,m}, \mathcal{Z}_t, \tau) := &\frac{1}{m+1}\mathbf{1}\left(\{i = 1, \ldots, m+1 \mid A_i < A_t\}\right) \\
&+ \frac{\tau}{m+1}\mathbf{1}\left(\{i = 1, \ldots, m+1 \mid A_i = A_t\}\right).
\end{aligned}$$

From the definitions of A_i and A, $A_i < A$ if and only if $S_i^\theta < S_t^\theta$, so that $\mathbf{1}(A_i < A)$ can be replaced by $\mathbf{1}(S_i^\theta < S_t^\theta)$. Consequently, the transducer corresponding to A reduces to the form in Equation (1). □

D.4 Fixed-τ Tie-Block Bounds

For our test, we assess the validity of synthetic samples $\mathcal{Z}_s$ by using them as a proxy for the synthetic distribution P_s^θ, denoted as $P_{s,N}^\theta$. We test whether target samples come from P_s^θ, under which null hypothesis, calibration and target samples are exchangeable.

The proof of Theorem 1 follows from the calibration samples and their exchangeability with the target sample set under the null hypothesis. With the conformal predictive system following a uniform distribution, Corollary 2 further gives the upper and lower bounds for risk control.

Proof of Corollary 2: Conditional on $\mathcal{Z}_s$, under the null hypothesis, the scores $S_1^\theta, \ldots, S_m^\theta, S_t^\theta$ are exchangeable by Theorem 1. Let k denote the total number of score positions belonging to the tie blocks for which

$$Q^\theta(\mathcal{Z}_s, \mathcal{Z}_{s,1}, \ldots, \mathcal{Z}_{s,m}, \mathcal{Z}_t, \tau) \leqslant 1-\alpha.$$

By Theorem 1, the target score is uniformly distributed over the $m+1$ score positions. Therefore,

$$\mathbb{P}\left(Q^\theta \leqslant 1-\alpha \mid \mathcal{Z}_s, \{S_1^\theta, \ldots, S_m^\theta, S_t^\theta\}\right) = \frac{k}{m+1}.$$

For the last tie block satisfying $Q^\theta \leqslant 1-\alpha$, its size is at most B_{max}^θ. Hence,

$$\frac{k-(1-\tau)B_{max}^\theta}{m+1} \leqslant 1-\alpha,$$

which gives

$$\frac{k}{m+1} \leqslant 1-\alpha+\frac{(1-\tau)B_{max}^\theta}{m+1}.$$

Similarly, if not all tie blocks satisfy $Q^\theta \leqslant 1-\alpha$, the first tie block satisfying $Q^\theta > 1-\alpha$ has size at most B_{max}^θ. Therefore,

$$\frac{k+\tau B_{max}^\theta}{m+1} > 1-\alpha,$$

and thus

$$\frac{k}{m+1} > 1-\alpha-\frac{\tau B_{max}^\theta}{m+1}.$$

If all tie blocks are accepted, the lower bound holds automatically. This completes the proof. □

Corollary 2 controls the probability of incorrectly rejecting a parameter satisfying the population null hypothesis. As for the upper bound, continuity of the conformal measure (e.g., the f-divergence discussed in the previous part) implies that ties in the nonconformity scores $\{S_i^\theta = S_t^\theta\}$ occur with probability zero. Accordingly, the upper bound of the risk control bias is $1/(m+1)$, confirming the stability of the conformal predictive system without excessive conservatism. Conversely, under a discrete conformal measure, the ties in $\{S_i^\theta = S_t^\theta\}$ may be a larger value, the conformal predictive system controls the risk but tends to be conservative.

It follows from the upper bound in Corollary 2 that as $m \to \infty$, the upper bounds converge to the lower bound at the rate of $O(1/(m+1))$, leading to an asymptotic coverage level of $1-\alpha$. Moreover, aside from the support set condition on the synthetic distribution, Corollary 2 only relies on the independently generated calibration subsets, which does not require assumptions on the generative model for synthetic data. This guarantees that the conformal method provides valid and robust risk control for synthetic data generated by LLM.

D.5 Empirical Cutoff Concentration

Theorem 8 *Fix a significance level $\alpha \in (0,1)$. The conformal tail rejection region is $(\widehat{q}_{1-\alpha,m}^\theta, \infty)$. When the Assumption 2 Then there exist constants $c > 0$, such that for every small $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\widehat{q}_{1-\alpha,m}^\theta - q_{1-\alpha}^\theta\right| > \varepsilon | \mathcal{Z}_s\right) \leqslant 2\exp\left(-cm\varepsilon^2\right). \tag{24}$$

Then, for every $0 < \varepsilon \leqslant \delta_0$,

$$\mathbb{P}\left(\left|\widehat{q}_{1-\alpha,m}^\theta - q_{1-\alpha}^\theta\right| > \varepsilon \,\middle|\, Z_s\right) \leqslant 2\exp\left(-c_q m\varepsilon^2\right),$$

where $c_q > 0$ is independent of n and m. Consequently, $\widehat{q}_{1-\alpha,m}^\theta - q_{1-\alpha}^\theta = O_p(m^{-1/2})$ uniformly over n.

This theorem can be viewed as a stability result for the empirical conformal cutoff. The assumption that $P^{\theta}_{s,n}$ is differentiable at $q^{\theta}_{1-\alpha}$ and that $f^{\theta}_{s,n}(q^{\theta}_{1-\alpha}) \geqslant \zeta_0 > 0$ is a local regularity around the rejection boundary. Intuitively, it rules out the case where the distribution of calibration nonconformity scores is nearly flat, has a large gap, or has too much discreteness near the critical quantile. The rate $\widehat{q}^{\theta}_{1-\alpha,m} - q^{\theta}_{1-\alpha} = O_{\mathrm{P}}(m^{-1/2})$ follows immediately from this exponential bound.

Proof of theorem 8: Since $P^{\theta}_{s,n}$ is differentiable at $q^{\theta}_{1-\alpha}$ and $f^{\theta}_{s,n}(q^{\theta}_{1-\alpha}) > 0$, there exists $\delta_0 > 0$ such that, for every $0 < \varepsilon \leqslant \delta_0$,

$$P^{\theta}_{s,n}(q^{\theta}_{1-\alpha} + \varepsilon) - (1-\alpha) \geqslant \frac{1}{2} f^{\theta}_{s,n}(q^{\theta}_{1-\alpha})\varepsilon,$$

and

$$(1-\alpha) - P^{\theta}_{s,n}(q^{\theta}_{1-\alpha} - \varepsilon) \geqslant \frac{1}{2} f^{\theta}_{s,n}(q^{\theta}_{1-\alpha})\varepsilon.$$

We first control the upper tail of $\widehat{q}^{\theta}_{1-\alpha,m}$. If $\widehat{q}^{\theta}_{1-\alpha,m} > q^{\theta}_{1-\alpha} + \varepsilon$, then necessarily

$$\widehat{P}^{\theta}_{s,n,m}(q^{\theta}_{1-\alpha} + \varepsilon) < 1 - \alpha.$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{q}^{\theta}_{1-\alpha,m} > q^{\theta}_{1-\alpha} + \varepsilon | \mathcal{Z}_s\right) \leqslant \mathbb{P}\left(\widehat{P}^{\theta}_{s,n,m}(q^{\theta}_{1-\alpha} + \varepsilon) < 1 - \alpha | \mathcal{Z}_s\right) \\
&= \mathbb{P}\left(\widehat{P}^{\theta}_{s,n,m}(q^{\theta}_{1-\alpha} + \varepsilon) - P^{\theta}_{s,n}(q^{\theta}_{1-\alpha} + \varepsilon) < -\left(P^{\theta}_{s,n}(q^{\theta}_{1-\alpha} + \varepsilon) - (1-\alpha)\right) | \mathcal{Z}_s\right).
\end{aligned}$$

Since

$$m\widehat{P}^{\theta}_{s,n,m}(q^{\theta}_{1-\alpha} + \varepsilon) \sim \mathrm{Binomial}\left(m, P^{\theta}_{s,n}(q^{\theta}_{1-\alpha} + \varepsilon)\right).$$

Conditional on Z_s, Hoeffding's inequality gives

$$\begin{aligned}
\mathbb{P}\left(\widehat{q}^{\theta}_{1-\alpha,m} > q^{\theta}_{1-\alpha} + \varepsilon | \mathcal{Z}_s\right) &\leqslant \exp\left(-2m\left(P^{\theta}_{s,n}(q^{\theta}_{1-\alpha} + \varepsilon) - (1-\alpha)\right)^2\right) \\
&\leqslant \exp\left(-\frac{1}{2} m f^{\theta 2}_{s,n}(q^{\theta}_{1-\alpha})\varepsilon^2\right).
\end{aligned}$$

With condition (3), we have:

$$\mathbb{P}\left(\widehat{q}_{1-\alpha,m} > q^{\theta}_{1-\alpha} + \varepsilon | \mathcal{Z}_s\right) \leqslant \exp\left(-\frac{1}{2} m\zeta_0^2\varepsilon^2\right).$$

Similarly, we have the other side

$$\mathbb{P}\left(\widehat{q}^{\theta}_{1-\alpha,m} < q^{\theta}_{1-\alpha} - \varepsilon | \mathcal{Z}_s\right) \leqslant \exp\left(-\frac{1}{2} m\zeta_0^2\varepsilon^2\right).$$

Combining the two bounds yields

$$\mathbb{P}\left(\left|\widehat{q}^{\theta}_{1-\alpha,m} - q^{\theta}_{1-\alpha}\right| > \varepsilon | \mathcal{Z}_s\right) \leqslant 2\exp\left(-\frac{1}{2} m\zeta_0^2\varepsilon^2\right),$$

which proves the stated concentration inequality. □

D.6 Proof of Asymptotic Screening Power

Proof of theorem 3: By Proposition 9, Assumption 1 implies the uniform density-ratio and denominator bounds required in Lemmas 5 and 6. Therefore,

$$\mathbb{P}\left(|S^{\theta}_t - S^*_t| > c_1 r^{-1}_{n,N} + \varepsilon\right) \leqslant 2\exp(-c_2 r_{n,N}\varepsilon^2).$$

Building on the results of inequality (22)

$$\mathbb{P}\left(|S^{\theta}_t - S^*_t| > c_1 r^{-1}_{n,N} + \varepsilon\right) \leqslant 2\exp\{-c_2 r_{n,N}\varepsilon^2\},$$

and Theorem 8

$$\mathbb{P}\left(\left|\widehat{q}^{\theta}_{1-\alpha,m} - q^{\theta}_{1-\alpha}\right| > \varepsilon | Z_s\right) \leqslant 2\exp\left(-c_q m\varepsilon^2\right),$$

for some constants $c_1, c_2, c_q > 0$.

We first decompose the non-rejection region $\{S^{\theta}_t < \widehat{q}^{\theta}_{1-\alpha,m}\}$. For any $\varepsilon_1 > 0$,

$$\{S^{\theta}_t < \widehat{q}^{\theta}_{1-\alpha,m}\} \subseteq \{\widehat{q}^{\theta}_{1-\alpha,m} \geqslant q^{\theta}_{1-\alpha} + \varepsilon_1\} \cup \{S^{\theta}_t \leqslant q^{\theta}_{1-\alpha} + \varepsilon_1\}.$$

Indeed, on the complement of the right-hand side, we have simultaneously

$$\widehat{q}^{\theta}_{1-\alpha,m} < q^{\theta}_{1-\alpha} + \varepsilon_1 \quad \text{and} \quad S^{\theta}_t > q^{\theta}_{1-\alpha} + \varepsilon_1,$$

which implies $S^{\theta}_t > \widehat{q}^{\theta}_{1-\alpha,m}$ and hence contradicts $S^{\theta}_t < \widehat{q}^{\theta}_{1-\alpha,m}$.

Therefore, by the union bound,

$$\mathbb{P}(S^{\theta}_t < \widehat{q}^{\theta}_{1-\alpha,m}) \leqslant \mathbb{P}(\widehat{q}^{\theta}_{1-\alpha,m} \geqslant q + \varepsilon_1) + \mathbb{P}(S^{\theta}_t \leqslant q + \varepsilon_1).$$

The first term is controlled by the inequality (24):

It remains to control the second term. Since condition (4), the event $\{S^{\theta}_t \leqslant q^{\theta}_{1-\alpha} + \varepsilon_1\}$ implies

$$S^*_t - S^{\theta}_t \geqslant S^*_t - q_{1-\alpha} - \varepsilon_1 = c_1 r^{-1}_{n,N} + \Delta_{n,N}(\varepsilon_1).$$

Thus,

$$\{S^{\theta}_t \leqslant q_{1-\alpha} + \varepsilon_1\} \subseteq \left\{|S^{\theta}_t - S^*_t| \geqslant c_1 r^{-1}_{n,N} + \Delta_{n,N}(\varepsilon_1)\right\}.$$

Applying the KL-EFD concentration inequality (22) with $\varepsilon_2 = \Delta_{n,N}(\varepsilon_1)$ gives

$$\begin{aligned}\mathbb{P}(S^{\theta}_t \leqslant q^{\theta}_{1-\alpha} + \varepsilon_1) &\leqslant \mathbb{P}\left(|S^{\theta}_t - S^*_t| \geqslant c_1 r^{-1}_{n,N} + \Delta_{n,N}(\varepsilon_1)\right) \\ &\leqslant 2\exp\left(-c_2 r_{n,N}\Delta_{n,N}(\varepsilon_1)^2\right).\end{aligned}$$

Combining the two bounds proves

$$\mathbb{P}\left(S^{\theta}_t < \widehat{q}^{\theta}_{1-\alpha,m}\right) \leqslant 2\exp\left(-c_q m\varepsilon_1^2\right) + 2\exp\left(-c_2 r_{n,N}\left(S^*_t - q^{\theta}_{1-\alpha} - \varepsilon_1 - c_1 r^{-1}_{n,N}\right)^2\right).$$

Finally, if $\varepsilon_1 \gg m^{-1/2}$, then $m\varepsilon_1^2 \to \infty$, so the first exponential term converges to zero. If

$$S^*_t - q^{\theta}_{1-\alpha} - \varepsilon_1 - c_1 r^{-1}_{n,N} \gg r^{-\frac{1}{2}}_{n,N},$$

then

$$r_{n,N}\left(S^*_t - q^{\theta}_{1-\alpha} - \varepsilon_1 - c_1 r^{-1}_{n,N}\right)^2 \to \infty,$$

so the second exponential term also converges to zero. Hence $\mathbb{P}\left(S^{\theta}_t < \widehat{q}^{\theta}_{1-\alpha,m}\right) \to 0$. The sufficient condition

$$S^*_t - q^{\theta}_{1-\alpha} - c_1 r^{-1}_{n,N} \gg r^{-\frac{1}{2}}_{n,N} + m^{-\frac{1}{2}}$$

follows by choosing ε_1 such that

$$m^{-\frac{1}{2}} \ll \varepsilon_1 \ll S^*_t - q^{\theta}_{1-\alpha} - c_1 r^{-1}_{n,N}. \qquad \square$$

Proposition 9 *Under Assumption 1, Condition 1 holds. Moreover, the same uniform bounds hold for the empirical smoothed densities, their one-observation replacements, and the denominator lower bound required in Lemma 6.*

Proof of Proposition 9: By Assumption 1, $\mathrm{supp}(P_t) \subseteq \Omega$, and both $\mathrm{supp}(P^{\theta}_s)$ and Ω are compact.

Since the Gaussian kernel φ_σ is continuous and strictly positive, there exist constants $0 < \varphi_{\min} \leqslant \varphi_{\max} < \infty$ such that

$$\varphi_{\min} \leqslant \varphi_\sigma(z - x) \leqslant \varphi_{\max}, \qquad \forall z \in \Omega,\ x \in \Omega.$$

Consequently,

$$\varphi_{\min} \leqslant (P_t * \varphi_\sigma)(z),\ (P_s^\theta * \varphi_\sigma)(z) \leqslant \varphi_{\max}, \qquad \forall z \in \Omega.$$

Therefore,

$$\frac{\varphi_{\min}}{\varphi_{\max}} \leqslant \frac{P_t * \varphi_\sigma(z)}{P_s^\theta * \varphi_\sigma(z)} \leqslant \frac{\varphi_{\max}}{\varphi_{\min}},$$

so Condition 1 holds with

$$\underline{r} = \frac{\varphi_{\min}}{\varphi_{\max}}, \qquad \overline{r} = \frac{\varphi_{\max}}{\varphi_{\min}}.$$

The same argument applies to every empirical smoothed density, because each empirical density is an average of Gaussian kernel values lying in $[\varphi_{\min}, \varphi_{\max}]$. It also applies after replacing one observation and to every convex interpolation used in the proofs of Lemmas 5 and 6.

In particular,

$$P_s^\theta * \varphi_\sigma(z) \geqslant \varphi_{\min}, \qquad P_{s,N}^\theta * \varphi_\sigma(z) \geqslant \varphi_{\min},$$

so the denominator condition in Lemma 6 holds with $\ell_s = \varphi_{\min}$. □

E Additional Theoretical Results

E.1 Asymptotic Normality of the Empirical Quantile

Theorem 10 *Let $S_1^\theta, S_2^\theta, \ldots, S_m^\theta$ be i.i.d. random samples from a population with cumulative distribution function $\widehat{P}_{s,n,m}^\theta$. For a fixed quantile level $\alpha \in (0,1)$, denote*

$$\hat{q}_{1-\alpha,m}^\theta = \widehat{P}_{s,n,m}^{\theta-1}(1-\alpha) = \inf\left\{x \in \mathbb{R} : \widehat{P}_{s,n,m}^\theta(x) \geqslant 1-\alpha\right\}.$$

Assuming the distribution function F is differentiable at $q_{1-\alpha}^\theta$, and its probability density function $f_{s,n}^\theta(x)$ satisfies $f_{s,n}^\theta(q_{1-\alpha}^\theta) > 0$.

Then, the empirical quantile estimator $\hat{q}_{1-\alpha,m}^\theta$ is asymptotically normal:

$$\sqrt{m}\left(\hat{q}_{1-\alpha,m}^\theta - q_{1-\alpha}^\theta\right) \xrightarrow{d} N\left(0,\ \frac{\alpha(1-\alpha)}{f_{s,n,m}^\theta(q_{1-\alpha}^\theta)^2}\right),$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes the normal distribution with mean μ and variance σ^2.

Proof of Theorem 10: This theorem can be verified via the following steps. Note that $m\widehat{P}_{s,n,m}^\theta(q_{1-\alpha,m}^\theta) \sim \text{Binomial}(m, 1-\alpha)$. By the Central Limit Theorem,

$$\sqrt{m}\left(\widehat{P}_{s,n,m}^\theta(\hat{q}_{1-\alpha,m}^\theta) - (1-\alpha)\right) \xrightarrow{d} N\left(0,\ \alpha(1-\alpha)\right). \tag{25}$$

By the inverse function theorem and $f_{s,n}^\theta(q_{1-\alpha}^\theta) > 0$, $P^{-1}(\cdot)$ is differentiable with derivative $P^{-1\prime}(1-\alpha) = \frac{1}{f(q_{1-\alpha})}$.

Treat $\hat{q}_{1-\alpha,m}^\theta = \widehat{P}_{s,n,m}^{\theta-1}(1-\alpha)$ as a functional of $\widehat{P}_{s,n,m}^\theta$. The Delta method gives

$$\sqrt{m}\left(\widehat{P}_{s,n,m}^{\theta-1}(1-\alpha) - P_{s,n}^{\theta-1}(1-\alpha)\right) = \frac{1}{f_{s,n}^\theta(q_{1-\alpha}^\theta)} \cdot \sqrt{m}\left(1-\alpha - \widehat{P}_{s,n,m}^\theta(q_{1-\alpha}^\theta)\right) + o_p(1). \tag{26}$$

Denote the first item in equation (26) as X_m. From equation (25) and the continuous mapping theorem,

$$X_m \xrightarrow{d} N\left(0,\ \frac{\alpha(1-\alpha)}{f_{s,n}^\theta(q_{1-\alpha}^\theta)^2}\right).$$

Meanwhile, $Y_m = o_p(1)$ implies $Y_n \xrightarrow{p} 0$. By Slutsky's Theorem, we have

$$X_m + Y_m \xrightarrow{d} N\left(0,\ \frac{\alpha(1-\alpha)}{f_{s,n}^\theta(q_{1-\alpha}^\theta)^2}\right),$$

which establishes the asymptotic normality of $\sqrt{m}(\hat{q}_{1-\alpha,m}^\theta - q_{1-\alpha}^\theta)$. □

E.2 MULTIPLE-OUTPUT COROLLARY

Corollary 11 *Fix a candidate control parameter $\theta \in \Theta$. For $j = 1, \ldots, T$, consider the question-level null hypotheses*

$$H_{0,j}^{\theta} : P_{t,j} = P_{s,j}^{\theta},$$

and let Q_j^{θ} be the corresponding randomized conformal predictive system values. Suppose that $H_{0,j}^{\theta}$ holds for all $j = 1, \ldots, T$. Under the Bonferroni rule, define

$$\theta \in \Theta^* \quad \Longleftrightarrow \quad Q_j^{\theta} < 1 - \frac{\alpha}{T} \quad \textit{for every } j = 1, \ldots, T.$$

Then

$$\mathbb{P}(\theta \notin \Theta^*) \leq \alpha.$$

With the same assumptions as in Corollary 2, Algorithm 1 is applied to a multiple hypothesis testing problem with the Bonferroni correction.

Proof of Corollary 11: For each $j = 1, \ldots, T$, the randomized validity result in Theorem 1 gives

$$\mathbb{P}\Big(Q_j^{\theta} \geq 1 - \frac{\alpha}{T}\Big) \leqslant \frac{\alpha}{T}.$$

Therefore,

$$\begin{aligned}
\mathbb{P}(\theta \notin \Theta^*) &= \mathbb{P}\left(\bigcup_{j=1}^{T} \Big\{Q_j^{\theta} \geq 1 - \frac{\alpha}{T}\Big\}\right) \\
&\leq \sum_{j=1}^{T} \mathbb{P}\Big(Q_j^{\theta} \geq 1 - \frac{\alpha}{T}\Big) \\
&\leq \sum_{j=1}^{T} \frac{\alpha}{T} = \alpha.
\end{aligned}$$

The argument does not require independence among the Multi-tests. □

The above simultaneous guarantee should be interpreted as a simultaneous non-exclusion guarantee. It ensures that, with probability at least $1-\alpha$, the screening procedure does not remove any candidate parameter whose synthetic distribution is exchangeable with the target distribution. It does not imply that every parameter retained in Θ^* is exactly correct.

F ALGORITHMS AND COMPLEXITY

Computational complexity. Let $k = |\Theta|$. The total computational cost can be written as

$$O(k\,[C_G(N + mn) + (m + 1)C_D(N, n)]),$$

where C_G is the cost of generating one synthetic observation and $C_D(N, n)$ is the cost of one distributional nonconformity evaluation. The calibration sampling term mn is usually negligible compared with LLM generation and divergence evaluation. For empirical f-divergence with kernel smoothing, the naive divergence cost can scale as $O(Nn)$. The parameters therefore affect computation and statistical resolution differently: N should be a moderate multiple of n, and m should be large enough to stabilize the cutoff without dominating the computational budget.

G ADDITIONAL SIMULATION RESULTS

In addition to the f-divergences introduced in Section 2, we consider the L_2 distance, 2-Wasserstein distance, and maximum mean discrepancy (MMD) as alternative distribution-level nonconformity scores for conformal screening. We also compare conformal interval construction with alternatives based on concentration inequalities and the central limit theorem (CLT), assessing both interval quality and effectiveness in screening candidate control parameters.

Algorithm 1 LLM-Based Synthetic Data Parameter Selection Algorithm

Require: Significance level α, number of calibration samples m, target dataset size n, synthetic dataset size N, $\tau_\theta \sim U[0,1]$.
Ensure: Retained parameter set Θ^*.

1: Prepare the target dataset $\mathcal{Z}_t$ with sample size n.
2: Construct the candidate set of LLM control parameters, yielding a candidate set of control parameters Θ.
3: **For** $\theta \in \Theta$ **do**
4: Generate a synthetic dataset $\mathcal{Z}_s^\theta$ with sample size N using the LLM under parameter θ.
5: genertae m calibration samples $\mathcal{Z}_{s,i}^\theta$ $(i = 1, 2, \ldots, m)$ with sample size n under θ.
6: Calculate the EFD nonconformity scores S_i^θ and S_t^θ for the calibration and target samples, respectively.
7: Compute the conformal predictive system value Q^θ of candidate θ:
8: If $Q^\theta \geqslant 1-\alpha$, reject the parameter θ;
9: Otherwise, add θ to the retained set Θ^*.
10: **End For**
11: **Return** the retained compatibility set Θ^*

The L_p norm between distribution P and R is defined as

$$D_{L_p}(P\|R) = \left(\int \|p(x) - r(x)\|^p dx \right)^{\frac{1}{p}}$$

where $p(x)$ and $r(x)$ denote the density function of distribution P and R. In particular, when $p = 1$, we introduce the total variation as a substitute for the L_1 norm. Including $p = 1$, we consider the case where $p = 2$.

The 2-Wasserstein distance between distribution P and R is defined as

$$W_2(P,R) = \left(\inf_{\gamma \in \Pi(P,R)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|^2 \, d\gamma(x,y) \right)^{\frac{1}{2}},$$

where $\Pi(P,R)$ denotes the set of all joint distributions with marginals P and R.

The squared Maximum Mean Discrepancy (MMD) between distributions P and R is defined as

$$\mathrm{MMD}^2(P,R) = \left\| \mathbb{E}_{x\sim P}[\varphi(x)] - \mathbb{E}_{y\sim R}[\varphi(y)] \right\|_{\mathcal{H}}^2,$$

Given i.i.d. samples $\mathcal{Z}_s = \{z_{s,i}\}$ and $\mathcal{Z}_t = \{z_{t,j}\}$, The empirical MMD is computed by

$$\widehat{\mathrm{MMD}}^2(\mathcal{Z}_s, \mathcal{Z}_t) = \frac{1}{N^2} \sum_{i,j=1}^{N} K(z_{s,i}, z_{s,j}) - \frac{2}{nN} \sum_{i=1}^{n} \sum_{j=1}^{N} K(z_{t,i}, z_{s,j}) + \frac{1}{n^2} \sum_{i,j=1}^{n} K(z_{t,i}, z_{t,j}).$$

where $K(\cdot,\cdot)$ is a kernel function.

which corresponds to the one-sided interval $[0, u_{\mathrm{or}}]$. This approach uses the true sampling distribution and is included solely for reference.

For confidence interval construction procedures that rely on concentration inequalities, we employ the Chebyshev bound and the Chernoff bound. A finite-sample confidence interval for ξ with confidence level $1-\alpha$ is given by solving

$$\frac{2}{m} \sum_{i=1}^{m} \exp(\xi_i - \epsilon) = \alpha,$$

and

$$\frac{\sum_{i=1}^{m} (\xi_i - \bar{\xi})^2}{(m-1)\epsilon} = \alpha.$$

The corresponding Chebyshev-based confidence interval and Chernoff based confidence interval are $\hat{\xi}_n \pm \epsilon$.

A CLT-based asymptotic $1-\alpha$ confidence interval for ξ is given by

$$\hat{\xi}_n \pm z_{1-\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}},$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution, and $\hat{\sigma}$ is a consistent estimator of the asymptotic standard deviation σ.

G.1 Gaussian-Mixture Simulation and Additional Method Details

This section presents the Gaussian-mixture simulation and provides detailed definitions of the alternative discrepancy measures and interval-construction baselines omitted from the main text. We specify the target distribution as an equally weighted mixture of two Gaussian components, $G_0 = \frac{1}{2}\mathcal{N}(0,1) + \frac{1}{2}\mathcal{N}(5,1)$, and use a two-component Gaussian mixture as the synthetic data generator. We apply conformal screening to candidate values of the mean parameter μ, varying the target sample size n, synthetic reference sample size N, and number of calibration batches m. We begin by comparing the widths of confidence intervals constructed by the competing methods under the true distribution.

Table 2: Confidence intervals for the Gaussian mixture model at $\alpha = 0.05$, $N = 2000$, $n = 100$, and varying m.

m	Metric	Oracle	Conformal-base	Chebyshev-base	Chernoff-base	CLT-base
20	Empirical Coverage	0.950000	**0.948458**	0.991891	1.000000	0.920647
	Avg. Length	0.056362	**0.060491**	0.086367	3.065647	0.043857
	StDev.	0.009998	0.014289	0.018336	0.017566	**0.010426**
80	Empirical Coverage	0.950000	**0.949588**	0.997004	1.000000	0.947091
	Avg. Length	0.055415	**0.056137**	0.085910	3.063859	0.044915
	StDev.	0.009698	0.010007	0.013412	0.016623	**0.006345**
320	Empirical Coverage	0.950000	**0.951737**	0.998394	1.000000	0.954464
	Avg. Length	0.056207	**0.056765**	0.087951	3.065650	0.046331
	StDev.	0.011051	0.011387	0.015783	0.017490	**0.007257**

In the table 2, we construct $\alpha = 0.05$ confidence intervals for the KL divergence between m calibration samples and samples of size N with various methods, and repeat 100 times, comparing the average real risk control level and interval length. The results demonstrate that the confidence intervals derived from the conformal method possess a risk level closer to the nominal level and shorter lengths relative to competing methods. Moreover, the CLT method is less stable in interval construction for small m but achieves comparable performance with an increasing number of calibration samples m. The concentration inequality approaches tend to overestimate the risk control level. However, as predicted by Theorem 1, we observe that m has almost no effect on the width of the confidence interval but exerts a considerable influence on the stability of the prediction interval.

From figure 5 and figure 6, it is noted that f-divergence yields consistently good performance in our experiments. Among alternative conformal measures, the Wasserstein-2 distance performs poorly, MMD performs favorably, and the KL divergence attains the optimal performance. The conformal method consistently surpasses the CLT-based method in confidence interval construction, which is both based on the KL divergence. Additionally, the screening efficacy of control parameters is markedly enhanced with increasing n. A significant performance boost occurs when N becomes comparable to n, whereas for sufficiently large N, the behavior is dominated by n in light of Theorem 3, yielding only marginal improvement.

G.2 Additional Tri-Hedge Screening Results

The Laplace-center tri-hedge distribution is defined as

$$\bar{p}(z) = 0.5 \cdot \frac{1}{2}\exp\left(-|z|\right) + 0.3 \cdot \phi(z; 3, 1) + 0.2 \cdot \phi(z; -3, 1),$$

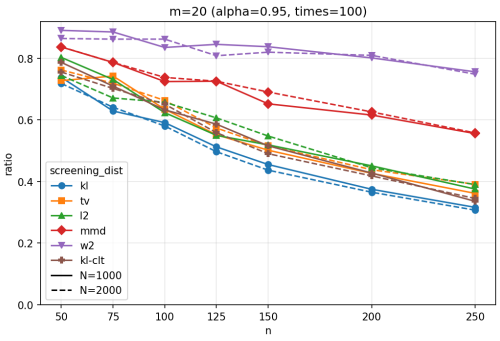


Figure 5: Acceptance ratio of candidate parameters versus n, $m = 20$, different methods and N values.

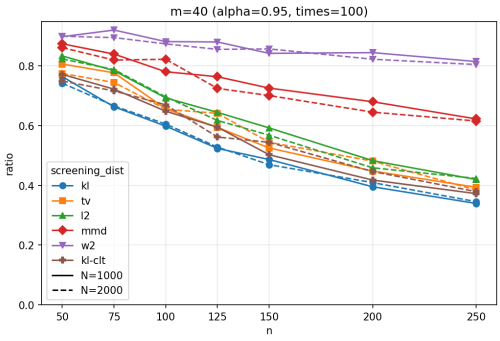
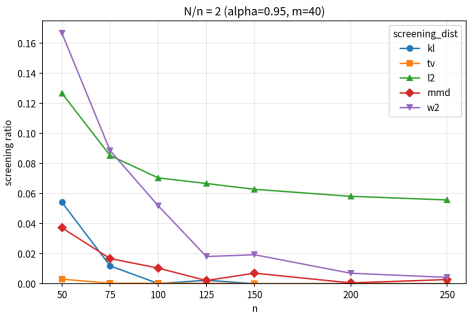


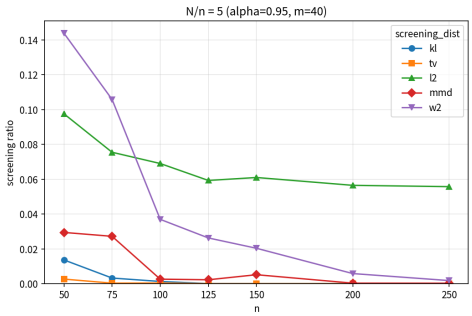
Figure 6: Acceptance ratio of candidate parameters versus n, $m = 40$, different methods and N values.

where $\frac{1}{2}\exp(-|z|)$ is the density of the $\mathrm{Laplace}(0, 1)$ distribution under the location-scale parameterization, and $\phi(z; \mu, \sigma^2)$ denotes the Gaussian density with mean μ and variance σ^2. This distribution combines two Gaussian side modes with a non-Gaussian Laplace central component and provides a challenging multimodal distribution.

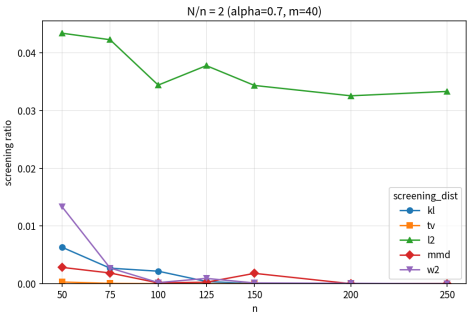
In algorithm initialization, we construct the parameter grid as follows: 5 mean values are selected with a step size of 2 over the interval $(-4, 4)$, and 5 variance values are selected with a step size of 0.2 over the interval $(0, 1)$. Taking the component-wise Cartesian product across the three dimensions, combined with 6 sets of mixture weights satisfying the unit-sum constraint, the total number of parameter configurations equals 93750.



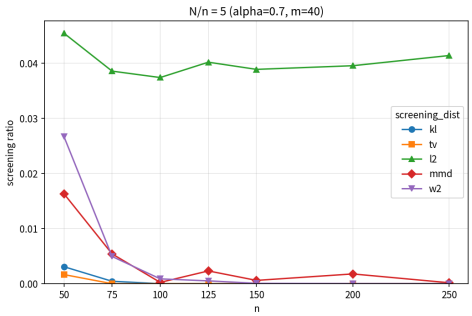
(a) $\alpha = 0.05$, $N/n = 2$



(b) $\alpha = 0.05$, $N/n = 5$



(c) $\alpha = 0.3$, $N/n = 2$



(d) $\alpha = 0.3$, $N/n = 5$

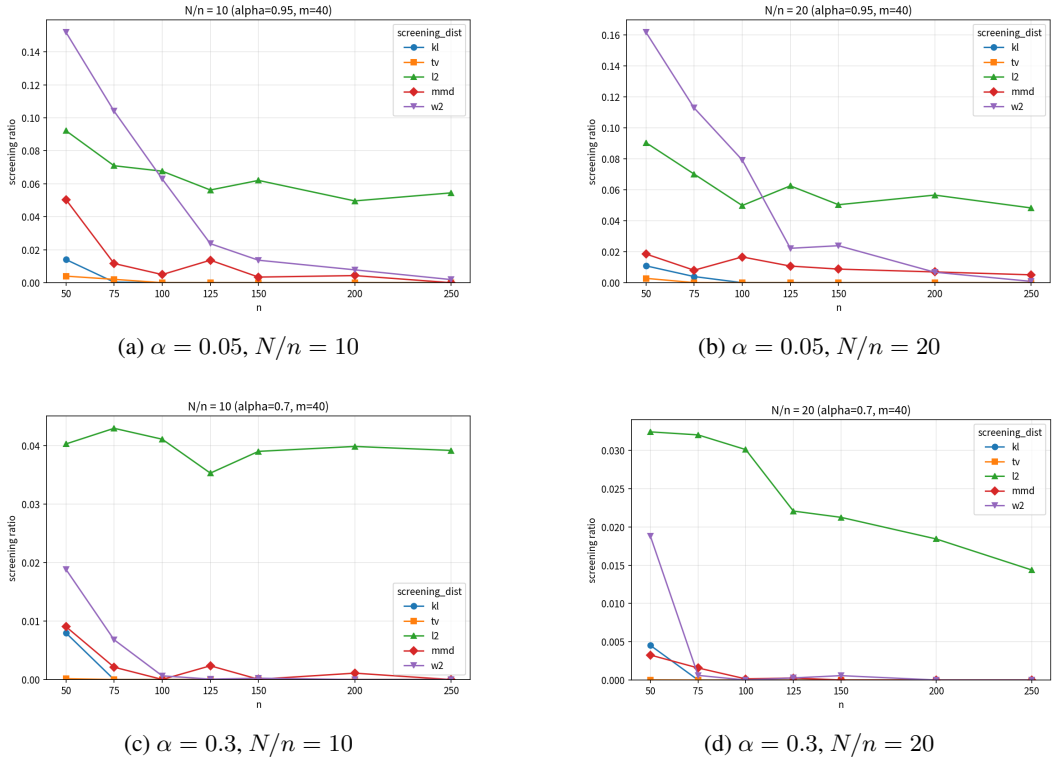
Figure 7: Screening ratio versus real sample size n on the tri-hedge distribution ($m = 40$). The columns correspond to $N/n = 2$ and $N/n = 5$, while the rows correspond to $\alpha = 0.05$ and $\alpha = 0.3$.

Figure 7 reports the fixed-grid screening results for $N/n \in \{2, 5\}$ at $\alpha \in \{0.05, 0.3\}$. All panels use the same candidate grid and $m = 40$.

Overall, we have the following observations. First, as n increases, the number of accepted candidate parameters generally decreases across different methods, indicating that a larger real sample size improves the precision and stability of the screening procedure. Second, among the distance measures, the L_2 and W_2 distances tend to retain more candidate parameters. In contrast, the KL, TV, and MMD distances are more effective in reducing the candidate set, demonstrating stronger screening behavior under the same fixed grid. Third, comparing $N/n = 2$ with $N/n = 5$, increasing the number of generated samples improves the screening results to some extent; however, this improvement is less pronounced than that obtained by increasing the real sample size n. Fourth, increasing α from 0.05 to 0.3 produces a clear reduction in the number of accepted candidate parameters, while the remaining qualitative patterns are largely unchanged. Since a larger α lowers the rejection threshold $1 - \alpha$, more generated distributions are rejected, resulting in fewer retained parameters.

Moreover, the calibration repetition m primarily controls the stability of parameter selection with negligible effect on interval width, while the quantity of accepted candidate parameters is strongly determined by the real sample size n, with the synthetic sample size N also playing a role. These results support the theoretical findings presented in Section 4.

Figure 8 further reports the larger synthetic-to-real sample ratios $N/n = 10$ and $N/n = 20$ under $\alpha = 0.05$ and $\alpha = 0.3$.



(a) $\alpha = 0.05$, $N/n = 10$

(b) $\alpha = 0.05$, $N/n = 20$

(c) $\alpha = 0.3$, $N/n = 10$

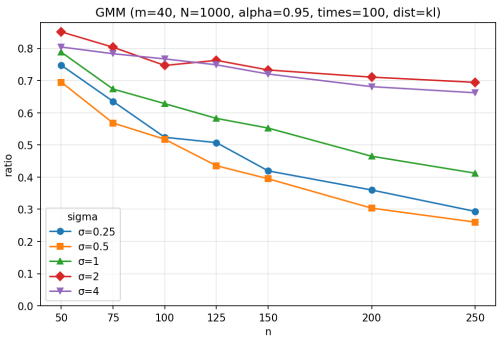
(d) $\alpha = 0.3$, $N/n = 20$

Figure 8: Additional fixed-grid screening results on the tri-hedge distribution for $N/n \in \{10, 20\}$ and $\alpha \in \{0.05, 0.3\}$ ($m = 40$).

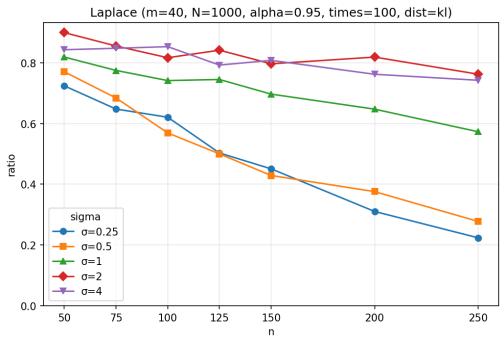
G.3 Effect of Gaussian Kernel Bandwidth on Screening Ratio

In conformal screening, the KL divergence is constructed via one-dimensional Gaussian kernel density estimation (KDE). The kernel bandwidth σ controls the smoothness of the estimated density. In previous experiments, we used Silverman's rule to determine the value of σ. In this section, we fix the calibration sample size $m = 40$ and the reference sample size $N = 1000$, and sweep the KDE bandwidth grid $\sigma \in \{0.25, 0.5, 1.0, 2.0, 4.0\}$ across observation sample sizes

$n \in \{50, 75, 100, 125, 150, 200, 250\}$. The screening ratio is defined as the number of candidate parameters passing the conformal test divided by the size of the candidate grid, and is averaged over 100 independent replications. The true distributions are specified as the Gaussian mixture model and the Tri-hedge model, respectively. The candidate grids are consistent with those in previous experiments.



(a) Effect of KL–KDE bandwidth σ on the screening ratio with fixed $m = 40$ and $N = 1000$ on GMM



(b) Effect of KL–KDE bandwidth σ on the screening ratio with fixed $m = 40$ and $N = 1000$ on Tri-Hedge model

Figure 9a reports the results under the GMM setting. Figure 9b shows the same pattern under the Laplace-TH setting. Three conclusions can be drawn from both figures. First, Curves for small bandwidths decrease monotonically with n, larger bandwidths are notably more robust. Second, increasing n mainly reduces the screening ratio under small bandwidths, while curves for large bandwidths are insensitive to changes in sample size. Third, the relative performance ranking between GMM and Tri-Hedge is consistent, indicating that this bandwidth effect is not limited to a specific true distribution.

G.4 Computational Cost of Discrepancy Measures

Table 3: Single-computation runtime (10^{-5} seconds) across different distance metrics ($\alpha = 0.05$, $m = 40$).

n	N	KL	TV	L_2	MMD	W_2
50	100	1.2	11	1.8	1.9	227
75	150	2.4	17	3.0	3.8	305
100	200	4.1	22	4.4	7.7	304
125	250	7	28	6.9	12	281
150	300	11	34	11	17	223
200	400	21	42	21	30	349
250	500	27	54	28	36	210

Under the setting $\alpha = 0.05$ and $m = 40$, Table 3 reports the single-computation runtime of different discrepancy measures. The runtime of the EFD-based estimation procedure increases as both n and N grow. For instance, when (n, N) increases from $(50, 100)$ to $(250, 500)$, both dimensions increase by a factor of 5. If the leading computational cost scales with nN, the total computational burden is expected to increase by approximately 25 times. This is consistent with the KL runtime, which increases from 1.2×10^{-5} to 27×10^{-5} seconds, approximately 22.5 times larger. Overall, KL divergence and the L_2 distance are the most computationally efficient, with runtimes of the same order across all considered values of (n, N). In contrast, the W_2 distance is consistently the most time-consuming measure, which is expected since its computation typically involves solving an optimal transport problem.

H Additional Real-Data Experiments

H.1 OpinionQA Data, Evaluation Protocol, and Screening Counts

OpinionQA Santurkar et al. (2023) is a survey-based benchmark derived from the Pew Research Center's nationally representative American Trends Panel and is designed to evaluate how well language models reflect diverse human opinions. We use Wave 26, which contains multiple-choice questions concerning firearms and public policy. After the filtering procedure described in Appendix H.1, the dataset contains 916 respondents. Political ideology is used as the individual-level prompt feature, and the control parameter $\pi = (\pi_1, \ldots, \pi_6)$ specifies the mixing proportions of its six profile categories, including Refused.

OpinionQA is constructed from 14–15 waves of the Pew Research Center's American Trends Panel, a nationally representative survey of U.S. adults. We use Wave 26, which focuses on firearms-related topics and contains multiple-choice questions covering gun ownership rights, public safety concerns, and gun-control proposals. After retaining questions with at least five response options, dropping questions whose per-question Refused rate exceeds 10%, and keeping the full respondent pool without profile-based subsampling, the resulting real dataset contains 916 respondents. Each observation corresponds to a respondent, and each selected question is a multiple-choice survey item.

The LLM prompt contains the survey instruction, respondent characteristics, the questionnaire, and optionally a short description of the answer preference used to modulate the response tendencies of LLMs. We use political ideology as the individual-level prompt feature and construct prompts with varying proportional representations of its six categories, including Refused. The control parameter is $\theta = (\pi_1, \ldots, \pi_6)$ with $\sum_i \pi_i = 1$. A simplex grid with spacing 0.1 gives 3003 candidate control parameters and a resolution in Euclidean distance no larger than $0.1\sqrt{6}$. The primary generator is Qwen3-8B via the SiliconFlow API, with temperature 0.4, max tokens 2048, and no top-p sampling.

Downstream response-proportion evaluation. The acceptance ratio is

$$\mathbf{a}_{\mathrm{acc}} = \frac{|\Theta^*|}{|\Theta|},$$

where Θ is the candidate set and Θ^* is the retained compatibility set.

For a nonempty retained set, the controlled divergence is

$$D_{\mathrm{ctrl}} = \frac{1}{|\Theta^*|} \sum_{\theta \in \Theta^*} D\left(\widehat{P}^{\theta}_{s,N}, \widehat{P}^{\mathrm{full}}_{t}\right),$$

where $\widehat{P}^{\mathrm{full}}_{t}$ is the full-data empirical distribution, which refers to the target empirical distribution derived from the entire dataset, rather than from a sampled subset of size n. The evaluation distribution is the retrospective complement of the fixed real reference subsample. Let $\widehat{\boldsymbol{p}}_{\mathrm{real}}$, $\widehat{\boldsymbol{p}}_{\mathrm{syn}}$, and $\widehat{\boldsymbol{p}}_{\mathrm{eval}}$ denote the response-proportion vectors from the reference sample, the selected synthetic distribution, and the retrospective evaluation sample. The augmented estimator is

$$\widehat{\boldsymbol{p}}_{\mathrm{aug}} = \frac{n\widehat{\boldsymbol{p}}_{\mathrm{real}} + N\widehat{\boldsymbol{p}}_{\mathrm{syn}}}{n+N},$$

and its run-level loss over J response categories is

$$\mathrm{MAE}_{\mathrm{macro}} = \frac{1}{J} \sum_{j=1}^{J} \left|\widehat{p}_{\mathrm{aug},j} - \widehat{p}_{\mathrm{eval},j}\right|.$$

We summarize this loss by its mean, standard deviation, and mean square across the evaluated configurations. The comparison includes Real only, Random control, Empirical-profile control, Minimum-discrepancy control, Conformal screening, and an infeasible Oracle selected using the evaluation distribution.

Table 4: Retrospective OpinionQA holdout KL after replacing Qwen3-8B with DeepSeek-V4-Flash under the matched configuration $(n, N, m) = (60, 600, 40)$. Oracle selects on the evaluation distribution and is infeasible.

Model	Random	Empirical profile	Min. discrepancy	Conformal	Oracle
Qwen3-8B	0.022609	0.065756	0.014290	0.019396	0.000559
DeepSeek-V4-Flash	0.012437	0.021574	0.014912	0.010655	0.000293

Performance comparison after increasing model capability. We next compare Qwen3-8B with the higher-performing DeepSeek-V4-Flash generator while applying the same screening procedure.

As shown in Table 4, the conformal-screened holdout KL decreases from 0.0194 with Qwen3-8B to 0.0107 with DeepSeek-V4-Flash, a reduction of approximately 45% in this controlled comparison. Random and empirical-profile controls exhibit the same direction of improvement, while the minimum-discrepancy result remains similar. These results provide an empirical trend that improving the underlying generator can strengthen the alignment of the screened synthetic distribution under the same screening procedure.

Screening counts and ratios. Table 5 reports the accepted-candidate counts. We observed that the resolution of control-parameter screening is primarily influenced by the real sample size n. The synthetic size N controls the accuracy of the generated distribution, while the calibration repetition m stabilizes the conformal threshold. Increasing n tends to improve prompt discrimination, increasing N tends to help until it is sufficiently larger than n, and increasing m mainly stabilizes the cutoff. In particular, the saved configurations at $n = 20$ exhibit substantially smaller acceptance ratios in several settings; this pattern becomes less pronounced when the reference sample increases to $n = 40$. The corresponding acceptance ratio is each entry divided by the 3003 candidates in the OpinionQA grid.

Table 5: Accepted candidate counts under different values of m, N, and n in OpinionQA.

m	N	n				
		20	40	60	80	100
40	200	370	2940	2968	2951	2446
	400	496	3002	2847	3001	2932
	600	521	2941	3003	3002	2906
	800	120	2989	3001	3003	2965
80	200	602	2999	2921	2957	2742
	400	642	2995	3002	3003	2933
	600	134	3003	2966	2999	2982
	800	512	2996	2994	3003	2965
120	200	536	2883	2947	2958	2734
	400	292	3002	2981	3003	2823
	600	91	3001	3002	3001	2922
	800	150	3003	3003	3003	2974

Note. Divide each count by 3003 to obtain the acceptance ratio.

Linear Regression Downstream Task. To further assess whether screened synthetic samples improve a concrete downstream estimator, we consider a linear regression of the survey option index y on political ideology. We compare a real-only fit on the n observed units with mixed fits that concatenate those units with N synthetic draws generated from the minimum-discrepancy mixing weight π and the corresponding vertex-conditional response distributions $P(y \mid \text{POLIDEOLOGY})$.

Figure 10 shows that the benefit of augmentation is largest when the real sample is scarce. At n=20, the real-only MSE is 0.597; mixing with $N \in \{200, 400, 600\}$ reduces it modestly to about 0.56, whereas N=800 lowers it to 0.387, a relative reduction of about 35%, and already matches the

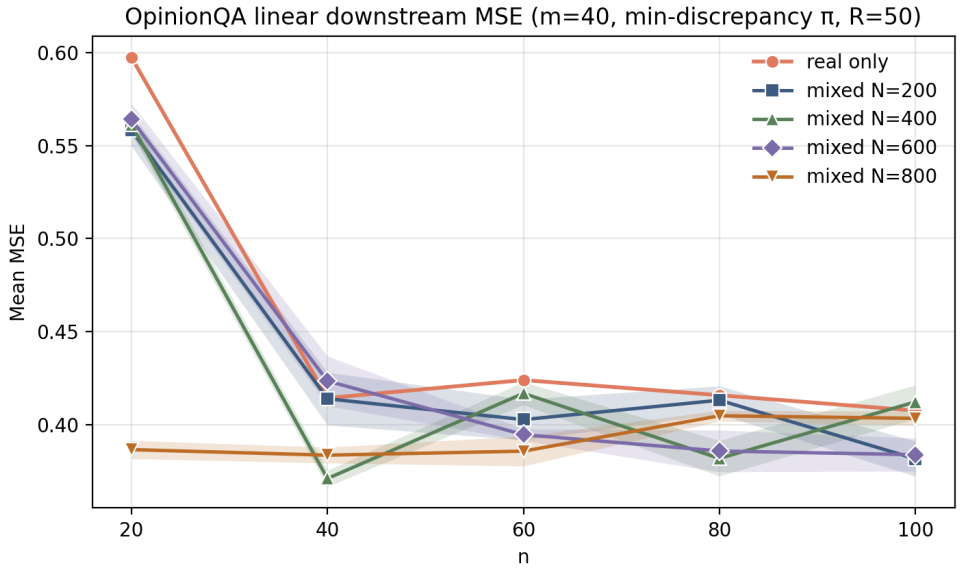


Figure 10: Holdout MSE of the OpinionQA linear model versus real sample size n (m=40). The coral curve is the real-only estimator; the remaining curves augment the same n units with N synthetic samples drawn from the minimum-discrepancy π. Mixed points are means over 50 synthetic repeats; bands are ± 1 SD.

large-n regime. Across the grid, larger N also yields tighter synthetic-repeat bands, especially at N=800, whose MSE stays in a narrow window of about 0.38–0.40 for all n. Overall, screened synthetic data are most useful as a stabilizer of the linear downstream estimator under limited real samples, rather than as a substitute once n is sufficient.

H.2 MovieLens Screening and Distribution Shift

MovieLens Harper & Konstan (2015) is a widely used benchmark for recommendation systems containing anonymous user ratings and movie metadata. It is collected from the MovieLens online movie recommendation service, containing millions of anonymous user ratings for thousands of films, along with detailed movie metadata and user information. We use the MovieLens ml-latest-small corpus: the top 30 most-rated movies, users with at least four rated movies among them, and user-level missing rates above 10% excluded. This yields a full real sample of 319 users.

Following the same procedure as for OpinionQA, all prompts follow a three-part structure: survey question background, selected movie introduction, and user characteristic description. We selected rating activity and average rating as user features. This dataset tests method performance when LLM control parameters fail to precisely align the generative distribution with the true distribution. The LLM is again Qwen3-8B via SiliconFlow, with temperature = 0.4 for probability prediction, max tokens = 2,048, and no top-p sampling.

The two three-level user attributes define nine joint profile cells. We therefore write the prompt-control vector as $\theta = (\pi_1, \ldots, \pi_9)$, where $\sum_{i=1}^{9} \pi_i = 1$, and use a grid of 43,758 candidate mixtures. Each candidate specifies how the profile-conditioned synthetic pools are combined before the resulting rating distribution is compared with the real reference sample.

However, it is observed that the prompt design space for large language models has an extremely high dimensionality, resulting in an effectively infinite number of candidate prompt options. However, sufficient and reliable supporting information must be provided in effective prompts to enable the generation of valid synthetic data by LLMs. In the experiments, a combinatorial prompt strategy is adopted, in which the control parameters are transferred to the mixing proportions π of different features; the general construction is described in Section 3.1, with task-specific details provided

in Appendix B.2. The dimensionality of the parameter space to be screened is thereby effectively reduced, and the efficiency of the parameter screening procedure is improved.

Moreover, OpinionQA uses political ideology as the main profile variable, while MovieLens uses two user rating behavior features with a three-level rating scale. It is important to note that the number of candidate control parameters to be screened exhibited a dramatic increase from 3,003 to 43,758. Strikingly, this exponential growth in parameter space was achieved by incorporating merely one additional user-level attribute into our model. This observation highlights the critical need for more computationally efficient prompt optimization techniques, which would substantially decrease the runtime of the screening procedure.

Table 6 reports the number of candidates that remain compatible with the reference distribution for every (m, N, n) configuration. Dividing a cell by 43,758 gives its acceptance ratio. Thus, a small positive count means that the real data identify a narrow subset of usable prompt mixtures, whereas a zero means that the compatibility set is empty and the procedure abstains from synthetic augmentation. It is observed that no valid control parameters π can be obtained through screening when the real sample sizes are $n = 80$ and $n = 100$, indicating that valid synthetic data cannot be generated by the LLM for any mixing proportion. When the real reference sample becomes sufficiently informative, the conformal test rejects all candidate prompt-controlled LLM distributions. This indicates that the proposed method can abstain from synthetic augmentation under prompt misspecification.

Table 6: Accepted candidate counts under different values of m, N, and n in MovieLens.

m	N	n				
		20	40	60	80	100
20	200	32234	1812	116	0	0
	400	36884	2063	110	0	0
	600	39831	1245	27	0	0
	800	40168	1072	21	0	0
40	200	37609	3094	221	0	0
	400	41382	3532	166	0	0
	600	42989	2837	10	0	0
	800	43661	1824	20	0	0
60	200	39849	3757	285	0	0
	400	43360	3515	81	0	0
	600	43701	2798	23	0	0
	800	43501	3120	16	0	0

Note. Divide each count by 43,758 to obtain the acceptance ratio. A zero entry is an empty compatibility set.

The contraction of the compatibility set is already pronounced before it becomes empty. At $n = 20$, the procedure retains 32,234–43,701 candidates (73.7%–99.9% of the grid); at $n = 40$, it retains only 1,072–3,757 candidates (2.45%–8.59%); and at $n = 60$, only 10–285 candidates remain (0.023%–0.651%). The changes across N and m are not strictly monotone in every row, but they are much smaller than the change caused by increasing n. The table should therefore be read as evidence that the amount of real information is the dominant source of prompt discrimination in this experiment, not as evidence that a particular accepted-candidate count is itself a downstream accuracy measure.

We observe that the resolution of control parameters for synthetic data is primarily influenced by the real sample size n. The synthetic size N controls the accuracy of the generated distribution, and the calibration repetition m stabilizes the conformal threshold. The empirical results should be interpreted together with the theoretical and simulation findings: increasing n should improve prompt discrimination, increasing N should help until it is sufficiently larger than n, and increasing m should mainly stabilize the cutoff.

Practical guidelines for parameter selection in real-world applications are derived from these results. The real sample size n should be as large as possible. The synthetic sample size N should be

controlled as a constant multiple of the real sample size n. The parameter m should be selected to achieve stable interval construction. For a confidence level of $\alpha = 0.05$, m is typically chosen as a multiple of 20, and a value of 40 was found to provide satisfactory performance in the experiments conducted.

H.3 Additional Movies, Mean-Rating Estimation, and Profile Mixtures

To further examine the validity of the proposed prompt-screening procedure after prompt adjustment, we additionally include the movie "Fight Club" in the synthetic rating experiment.

The results are reported in Table 7. The table compares *Forrest Gump* (Movie ID 356) and *Fight Club* (Movie ID 2959) at fixed $m = 40$ and $\alpha = 0.05$, while varying the synthetic sample size N and real reference size n. This separation makes it possible to distinguish a movie-specific change in compatibility from a change caused by the screening configuration itself.

Table 7: Accepted candidate counts for the additional MovieLens experiments at $m = 40$ and $\alpha = 0.05$.

Movie ID	N	n				
		20	40	60	80	100
356	200	26492	3793	2041	3981	1965
	400	27644	2998	1344	3279	1331
	600	28220	2669	1003	3007	1038
2959	200	23361	4115	1692	1345	614
	400	24219	3676	1351	993	386
	600	24613	3368	1267	857	308

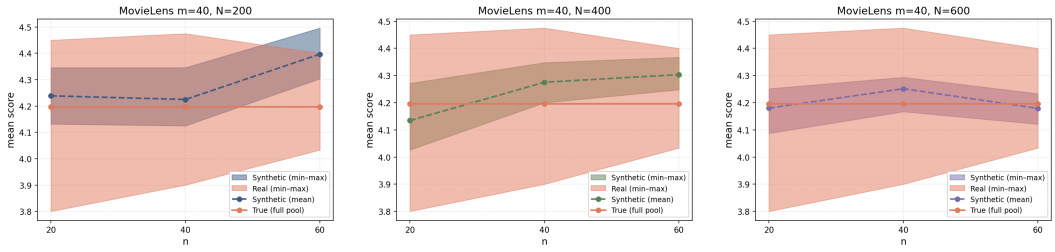
Note. Movie IDs 356 and 2959 correspond to *Forrest Gump* and *Fight Club*. Divide each count by 43,758 to obtain the acceptance ratio.

Overall, the number of accepted candidates decreases substantially as the sample size n increases, indicating that larger real-sample information leads to a stricter screening of invalid synthetic-generation parameters. Moreover, for most configurations with moderate or large n, increasing N also reduces the number of accepted candidates, suggesting that a larger synthetic sample size helps expose distributional discrepancies more clearly.

These results further confirm that validating prompt selection becomes more challenging when the LLM is required to generate rating distributions under more refined movie-specific constraints. The broad pattern is clearest for *Fight Club*: for every displayed value of N, the count falls continuously as n increases, reaching 308–614 accepted candidates at $n = 100$. For *Forrest Gump*, the count also contracts sharply from $n = 20$ to $n = 60$, but rises again at $n = 80$ before falling at $n = 100$. This local rebound shows why the trend should not be interpreted as a deterministic monotonicity claim: the retained set also depends on the realized reference sample and on how the candidate grid overlaps the movie-specific target distribution. Candidate counts indicate how selective the screen is; they do not, by themselves, rank the predictive quality of the accepted candidates.

To further evaluate the usefulness of synthetic data in downstream tasks, we generated synthetic samples using the screened control parameters and examined their performance in mean-rating prediction under different choices of N and n. We repeated the data generation procedure 40 times to assess the stability of the predicted mean scores, and the results of the movie "Fight Club" are shown in Figure 11.

The introduction of synthetic data substantially reduces the fluctuation of mean-rating prediction, indicating improved prediction stability. Although synthetic data may introduce some bias when the real sample size n becomes larger, it still effectively suppresses prediction variability and therefore improves the overall reliability of the downstream prediction task. Moreover, as the synthetic sample size N increases, the synthetic-data-induced bias is better controlled, while the variance-reduction effect becomes more pronounced. These results suggest that, with properly screened control pa-



(a) (b) (c)

Figure 11: Mean-Rating Prediction Stability with Screened Synthetic Data. The solid and dashed lines denote the mean predicted scores obtained from real and synthetic data, respectively, while the shaded regions represent the prediction variability over 40 repeated generations.

rameters, synthetic data can provide meaningful support for downstream prediction, especially by stabilizing estimates when the available real sample size is limited.

In Figure 11, the width of a shaded band reflects variability across the 40 repetitions, whereas the vertical separation between the solid real-data curve and the dashed synthetic-data curve reflects bias. The two features must be evaluated jointly: a narrower band supports a variance-reduction claim, but it does not establish accuracy if the dashed curve remains systematically displaced. The screening results above provide the accompanying distributional control by limiting augmentation to prompt mixtures that pass the compatibility test.

We next examine whether the composition of a retained prompt mixture resembles the empirical user-profile composition in the full MovieLens data. If accepted prompts concentrate around the empirical π, this suggests that the demographic or user-profile mixture is important for matching the target distribution. If accepted prompts deviate from the empirical π, this may indicate that the LLM already carries prior knowledge that shifts the generated distribution, and the prompt mixture acts as a correction mechanism.

Figure 12 reports the distribution of the L_2 gap over the candidates retained at $m = 40$; it does not report the size of the retained set. A small gap means that a compatible prompt uses a profile mixture close to the empirical mixture π^*, while a larger gap is consistent with the prompt mixture compensating for a distribution already encoded by the LLM. Because multiple mechanisms can produce the same rating distribution, this comparison is a diagnostic of how screening uses the profile mixture rather than a formal identification of the LLM's internal prior.

H.4 Multiple-Output Screening

To further evaluate the difficulty of prompt screening under joint distributional constraints, we conduct a multiple-testing-based screening experiment by combining the rating distributions of two movies, "Forrest Gump" and "Fight Club".

In this setting, the LLM is required to simultaneously approximate the rating distributions of both movies, which substantially increases the difficulty of distributional control. Consequently, it becomes challenging to identify candidate prompts that can generate validate synthetic data. To address this issue, we apply Bonferroni and Holm corrections to split the confidence level across multiple testing procedures and then screen the candidate parameters accordingly. The results are presented in Figure 13.

Overall, the Holm procedure leads to more effective confidence allocation and achieves better control of the synthetic rating distributions. Compared with the Bonferroni correction, Holm retains a more stable set of candidate parameters while consistently reducing the divergence between the synthetic and real distributions. This suggests that the stepwise Holm correction is less conservative and more suitable for prompt screening in the joint movie-rating synthesis task. For example, in the MovieLens experiment, when considering the screened rating-control parameters for two movies with $N = 600$, $n = 20$, and $m = 40$, both the Bonferroni and Holm–Bonferroni corrections fur-

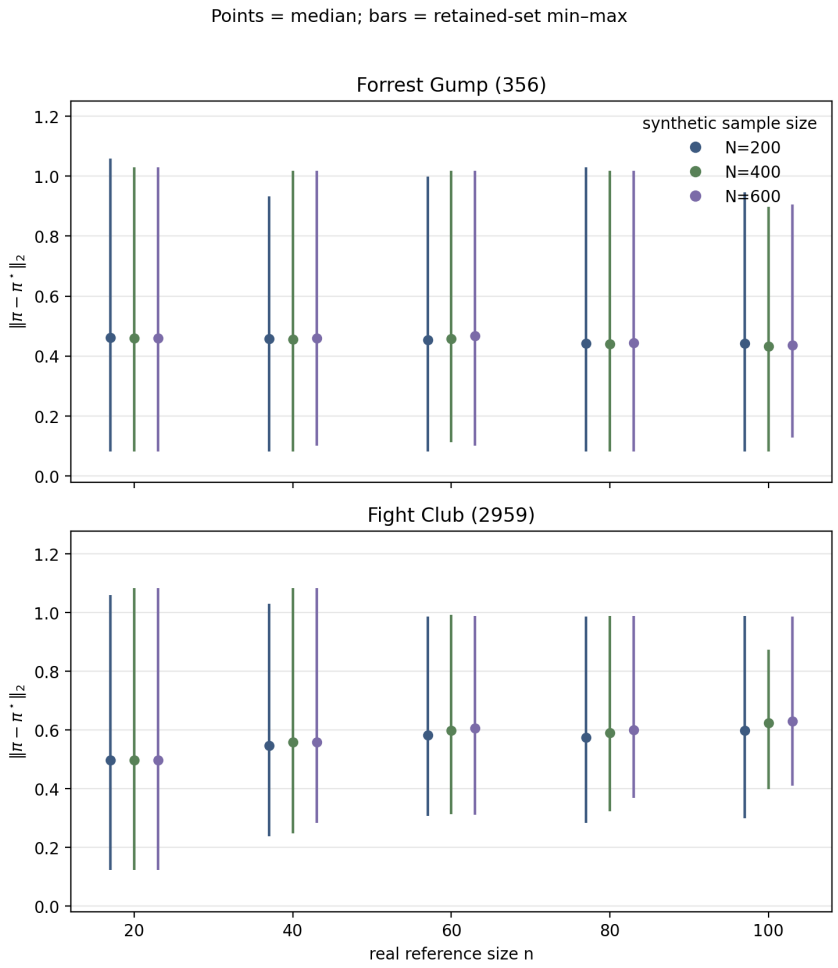


Figure 12: L_2 norm of the difference between the accepted π and ground-truth π^* at $m = 40$ under varying N and n.

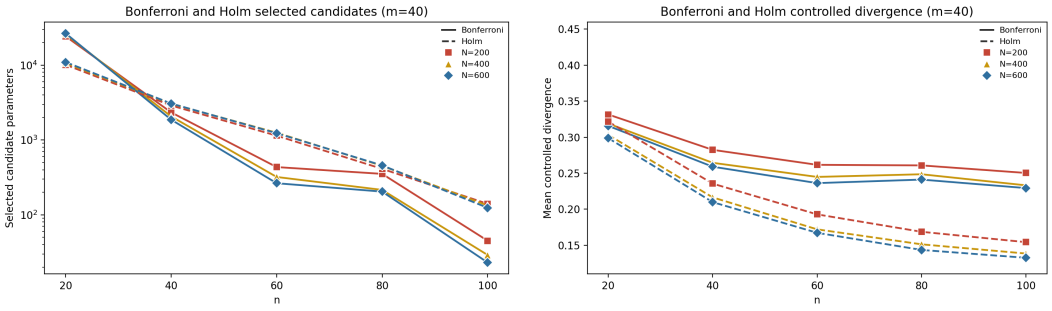


Figure 13: Prompt screening results using the Bonferroni and Holm procedures for the joint rating prediction task on *Forrest Gump* and *Fight Club*, with $m = 40$. The left panel shows the number of selected candidate parameters on a logarithmic scale, and the right panel shows the mean controlled divergence between the synthetic and real rating distributions, both as functions of the sample size n.

ther select two control parameters under the overall level $\alpha = 0.05$, with an average divergence of 0.0535.

This multiple-testing strategy avoids imposing an independence assumption between the two distributional comparisons. Although it may lead to fewer selected parameters due to its more conservative nature, it effectively reduces the complexity of control-parameter screening while preserving valid error control. Retaining more candidates is not automatically preferable if their divergence is larger. Conversely, the equality of the two procedures at the specific $(N, n, m) = (600, 20, 40)$ configuration does not contradict their broader difference across the displayed settings; it identifies one configuration in which both corrections lead to the same two-candidate decision.