Post-Selection Distributional Model Evaluation

Amirmohammad Farzaneh **Osvaldo Simeone**
Institute for Intelligent Networked Systems (INSI)
Northeastern University London
London, UK
{a.farzaneh,o.simeone}@nulondon.ac.uk

Abstract

Formal model evaluation methods typically certify that a model satisfies a prescribed target key performance indicator (KPI) level. However, in many applications, the relevant target KPI level may not be known a priori, and the user may instead wish to compare candidate models by analyzing the full trade-offs between performance and reliability achievable at test time by the models. This task, requiring the reliable estimate of the test-time KPI distributions, is made more complicated by the fact that the same data must often be used both to pre-select a subset of candidate models and to estimate their KPI distributions, causing a potential post-selection bias. In this work, we introduce post-selection distributional model evaluation (PS-DME), a general framework for statistically valid distributional model assessment after arbitrary data-dependent model pre-selection. Building on e-values, PS-DME controls post-selection false coverage rate (FCR) for the distributional KPI estimates and we establish explicit conditions under which it is provably more sample efficient than a baseline method based on sample splitting. Experiments on synthetic data, text-to-SQL decoding with large language models, and telecom network performance evaluation demonstrate that PS-DME enables reliable comparison of candidate configurations across a range of reliability levels, supporting the statistically reliable exploration of performance–reliability trade-offs.

1 Introduction

Current formal model selection and evaluation strategies, grounded in *multiple hypothesis testing* (MHT) (Angelopoulos et al., 2025; Farzaneh and Simeone, 2025), control the probability that the risk measure of interest exceeds a maximum tolerated target value. However, in practice, a suitable target risk level may not be known in advance, and the user may instead wish to explore different *trade-offs* between test-time performance and reliability. As studied in this paper, this requires estimating the full test-time distribution of the *key performance indicator* (KPI) of interest for the candidate models.

To exemplify the problem of interest, consider a large language model (LLM) agent that can be deployed on a mobile device with different configurations $k = 1, \ldots, K$. As studied in (Huang et al., 2026), these may correspond to models obtained by progressively pruning a base model. Accordingly, the available configurations range from high-quality and high-latency models to lower-quality and lower-latency models. Latency here refers to the prefill delay, which dictates the time to first token. Existing formal model selection methods (Angelopoulos et al., 2025), reviewed in (Farzaneh and Simeone, 2025), require the a priori specification of a latency target. However, as pointed out in (Huang et al., 2026), the latency target may change as the use-cases evolve, and there is a need to understand the full trade-off between performance and reliability.

For example, consider two LLM configurations – a lower-quality, simpler, model and a higher-quality, more complex, model. As illustrated in Fig. 1(a), suppose that the user knew the *cumulative*

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

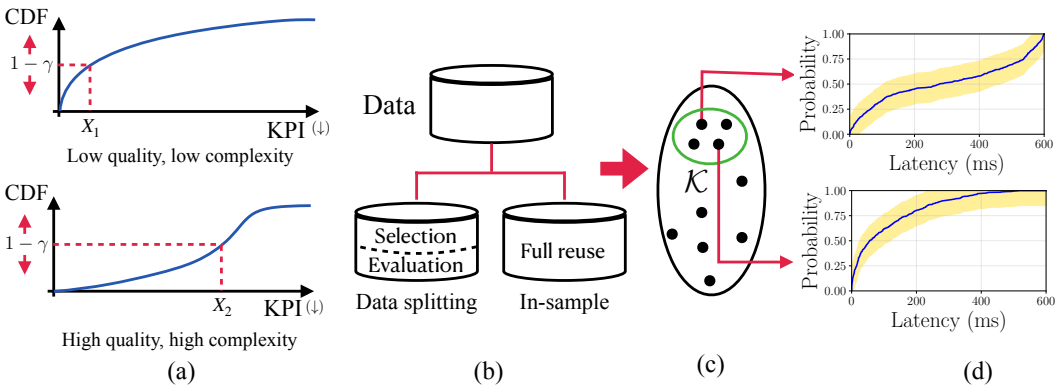


Figure 1: (a) CDF of a negatively-oriented key performance indicator (KPI), such as the prefill latency for an LLM agent to be deployed on a device (Huang et al., 2026), for two model configurations. For a test-time reliability level $1-\gamma$, knowing the CDF for each configuration allows the user to compare achievable performance across different configurations. Furthermore, it also supports the exploration of different reliability levels $1-\gamma$. (b) Two strategies for handling model pre-selection, namely data splitting and in-sample inference. (c)-(d) Illustration of the proposed post-selection distributional model evaluation (PS-DME), which applies in-sample pre-selection: (c) A user-defined selection strategy identifies a subset $\mathcal{K}$ of promising configurations among the K candidates; (d) For all the configurations in set $\mathcal{K}$, PS-DME constructs confidence bands for the CDFs of the KPI of interest with guaranteed *false coverage rate* (FCR), enabling reliable comparison of performance–reliability trade-offs across configurations.

distribution function (CDF) of the latency for the two different configurations. Note that the higher quality model, due to its higher complexity, tends to have a larger latency. If the user tolerates a maximum test-time failure probability γ, it may prefer the higher-quality model, but only as long as the benefits accrued from higher quality outweigh the larger latency ($X_2 > X_1$ in the figure). More generally, the user may also be interested to explore different values of the failure probability γ in order to obtain more favorable operating points on the trade-off between performance (latency) and test-time reliability $1-\gamma$.

The model selection strategy outlined in the previous example is challenged by the fact that the CDFs of the KPI of interest, e.g., the latency, are not known and must be estimated from data. Furthermore, the configurations being evaluated are often pre-selected from a large pool of candidates to satisfy requirements or criteria such as cost or expected accuracy. Because this pre-selection must generally rely on the same data later used for CDF estimation, naively applying classical confidence bands in this setting yields overly optimistic estimates (see, e.g., (Kuchibhotla, 2025)).

We address these challenges by proposing *post-selection distributional model evaluation* (PS-DME), a novel formal model evaluation protocol that supports the following steps:

1) **Arbitrary model pre-selection:** As illustrated in Fig. 1(c), the user can pre-select a subset $\mathcal{K} \subseteq \{1, \ldots, K\}$ of promising configurations using any arbitrary and possibly data-dependent procedure. As shown in Fig. 1(b), we study both *data splitting*, where one portion of the data is used for pre-selection and the remaining portion is reserved for distributional inference, and *in-sample* inference, in which the same data serve both for model screening and for uncertainty quantification (Chakraborty et al., 2026).
2) **Distributional model evaluation:** For each pre-selected configuration $k \in \mathcal{K}$, as illustrated in Fig. 1(d), PS-DME constructs confidence bands for the CDF of the KPI of interest. These bands quantify the uncertainty in the entire distribution of the performance metric, enabling the reliable estimation of quantities such as latency quantiles or outage probabilities (see Fig. 1(a)).
3) **Model selection based on performance-reliability trade-off with statistical guarantees:** The estimated CDF confidence bands enable the user to compare configurations by examining the trade-off between achievable performance and reliability levels. This allows the user to select a final configuration that strikes the desired balance between performance and reliability, while accounting for statistical uncertainty.

The rest of this paper is organized as follows. In Sec. 2, we formalize the post-selection distributional model evaluation problem, and introduce the FCR-based validity criterion adopted throughout the paper. Sec. 3 then presents the proposed PS-DME framework together with the sample-splitting baseline, establishes their statistical guarantees, and compares their statistical efficiency through explicit confidence band width comparisons. Sec. 4 positions our contribution with respect to prior work. Sec. 5 validates the proposed methodology on synthetic data, LLM decoding, and telecom performance evaluation. Sec. 7 concludes the paper with a discussion of the main findings and directions for future work, while the appendices collect proofs and extensions.

2 Problem Setup

For a given system, we consider K candidate configurations indexed by integer $k \in \{1, \ldots, K\}$. These may correspond to different hyperparameter settings, decoding strategies, or precision levels. Henceforth, we will refer to each configuration k as a *model*. For each model k, we observe n_k i.i.d. realizations of a scalar, negatively-oriented, *key performance indicator* (KPI),

$$\mathcal{D}_k = \{X_{k,1}, \ldots, X_{k,n_k}\}, \quad \text{with } X_{k,i} \overset{\text{i.i.d.}}{\sim} P_k, \tag{1}$$

where P_k is an unknown distribution with CDF $F_k(x) = \mathbb{P}_k(X \leq x)$. The datasets $\mathcal{D}_k$ can be arbitrarily correlated. For instance, in the case with the same number of data points for all configurations, i.e., $n_k = n$, the i-th measures $\{X_{k,i}\}_{k=1}^K$ may be a function of the same observations, with observations being i.i.d. across index $i = 1, \ldots, n$.

Let $\widehat{F}_k(\cdot)$ denote the empirical CDF formed from dataset $\mathcal{D}_k$ as

$$\widehat{F}_k(x) \;=\; \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{1}\{X_{k,i} \leq x\}. \tag{2}$$

We study an arbitrary two-step model evaluation pipeline consisting of the following phases:

1) Data-based model pre-selection: Using data $\mathcal{D} = \cup_{k=1}^K \mathcal{D}_k$ for all k configurations, an arbitrary (possibly unknown) data-dependent pre-selection rule produces a subset

$$\mathcal{K} \;=\; S(\mathcal{D}) \;\subseteq\; \{1, \ldots, K\} \tag{3}$$

of models.

2) Reliable distributional model evaluation: For each selected model $k \in \mathcal{K}$, we aim to report lower and upper bounds on the CDF $F_k(x)$, i.e.,

$$L_k(x) \;\leq\; F_k(x) \;\leq\; U_k(x), \tag{4}$$

that contain the true CDF uniformly with a user-defined probability.

Formally, for each selected model $k \in \mathcal{K}$, define the miscoverage event

$$\mathcal{O}_k \;=\; \Big\{\exists x \in \mathbb{R} :\; F_k(x) \notin [L_k(x), U_k(x)]\Big\}, \tag{5}$$

i.e., the event that there exists a KPI value x for which the interval $[L_k(x), U_k(x)]$ does not contain its true CDF value $F_k(x)$ (see Appendix A for an illustration). *Post-selection statistical validity* is measured via the *false coverage proportion* (FCP) (Xu et al., 2024). The FCP is defined as the fraction of pre-selected models in set $\mathcal{K}$ for which the miscoverage event (5) holds true, i.e.,

$$\text{FCP} \;=\; \frac{1}{\max\{|\mathcal{K}|, 1\}} \sum_{k \in \mathcal{K}} \mathbf{1}\{\mathcal{O}_k\}. \tag{6}$$

The *false coverage rate* (FCR) is the expectation of the FCP, i.e.,

$$\text{FCR} \;=\; \mathbb{E}[\text{FCP}]\,, \tag{7}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and the expectation is taken with respect to the joint distribution of data $\mathcal{D}$. Note that the selected set $\mathcal{K} = S(\mathcal{D})$ is itself data-dependent, and hence its cardinality $|\mathcal{K}|$ is a random variable. Our goal is to guarantee the condition

$$\text{FCR} \leq \delta \tag{8}$$

for a user-defined target probability $\delta \in (0, 1)$.

3 Post-Selection Distributional Model Evaluation

In order to ensure the condition (8), a conventional approach would adopt a sample splitting strategy, followed by the evaluation of standard uniform confidence intervals for CDFs (Dvoretzky et al., 1956; Massart, 1990). After briefly reviewing this baseline method in Sec. 3.1, we introduce PS-DME in Sec. 3.2. As introduced in Sec. 1, PS-DME leverages the entire dataset $\mathcal{D}$ for both model selection and distributional model evaluation tasks (see Fig. 1).

3.1 Sample-Splitting Distributional Model Evaluation

For each configuration $k \in \{1, \ldots, K\}$, suppose the available data $\mathcal{D}_k = \{X_{k,1}, \ldots, X_{k,n_k}\}$ in (1) are partitioned into two disjoint subsets $\mathcal{D}_k^{\text{sel}}$ and $\mathcal{D}_k^{\text{eval}}$ of sizes n_k^{sel} and n_k^{eval} respectively. In the baseline methodology reviewed here, referred to as *sample-splitting post-selection distributional model evaluation* (SS-DME), the pre-selection subset $\mathcal{D}_k^{\text{sel}}$ is used exclusively for model selection, while the model evaluation subset $\mathcal{D}_k^{\text{eval}}$ is reserved for distributional inference.

1) Model selection: The selection rule $\mathcal{K} = S\big(\mathcal{D}^{\text{sel}}\big) \subseteq \{1, \ldots, K\}$ leverages the dataset $\mathcal{D}^{\text{sel}} = \cup_{k=1}^{K} \mathcal{D}_k^{\text{sel}}$ using any arbitrary criterion. For instance, the function $S(\cdot)$ may select the $\bar{K} \leq K$ models with the smallest empirical KPIs $\hat{X}_k^{\text{sel}} = \sum_{i=1}^{n_k^{\text{sel}}} X_{k,i}/n_k^{\text{sel}}$.

2) Distributional inference: For each selected configuration $k \in \mathcal{K}$, SS-DME constructs a uniform confidence band for the true CDF $F_k(\cdot)$ using only the inference subset $\mathcal{D}_k^{\text{eval}}$. This is elaborated on in the rest of this subsection.

Let $\widehat{F}_k^{\text{eval}}(\cdot)$ denote the empirical CDF constructed from data $\mathcal{D}_k^{\text{eval}}$ using (2). By the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al., 1956), for any $\varepsilon > 0$ we have

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left|\widehat{F}_k^{\text{eval}}(x) - F_k(x)\right| > \varepsilon\right) \leq 2\exp\left(-2n_k^{\text{eval}}\varepsilon^2\right). \tag{9}$$

Using this inequality, SS-DME constructs the confidence bounds as

$$L_k(x) = \max\{0,\, \widehat{F}_k^{\text{eval}}(x) - \varepsilon_k\}, \tag{10a}$$

$$\text{and} \quad U_k(x) = \min\{1,\, \widehat{F}_k^{\text{eval}}(x) + \varepsilon_k\}, \tag{10b}$$

where

$$\varepsilon_k = \sqrt{\frac{1}{2n_k^{\text{eval}}} \log\left(\frac{2}{\delta}\right)}. \tag{11}$$

SS-DME provides the following guarantee.

Lemma 1. *For any pre-selection rule $S(\cdot)$, the confidence bands $\mathcal{C}_k = [L_k(x), U_k(x)]$ produced by SS-DME, where $L_k(x)$ and $U_k(x)$ are defined by* (10a) *and* (10b)*, respectively, satisfy the requirement* $\text{FCR} \leq \delta$.

A proof of Lemma 1 can be found in Appendix C.

3.2 In-Sample Distributional Model Evaluation

The SS-DME strategy described in Sec. 3.1 ensures validity by enforcing independence between model selection and distributional inference. However, this separation reduces the effective sample size available for constructing confidence bands, potentially leading to occasionally wide intervals. In this subsection, we introduce PS-DME, which enables full-data reuse while maintaining the requirement (8) on the FCR. To achieve this, PS-DME adopts the framework of e-values (Ramdas and Wang, 2025) to construct confidence intervals with post-selection validity (Xu et al., 2024).

E-values are statistics used to test null hypotheses that have additional properties of robustness as compared to p-values, including, notably, the support of the post-hoc selection of the significance level (Vovk and Wang, 2021; Koning, 2025). An e-value can be constructed from a p-value via the application of an e-calibrator function, which is defined as follows.

Definition 1 (e-calibrator)**.** *A function $f : [0, 1] \to [0, \infty]$ is called an e-calibrator if it satisfies the following properties: (i) f is nonincreasing and upper semicontinuous; and (ii)* $\displaystyle\int_0^1 f(u)\, du \leq 1$.

For example, a commonly used family of e-calibrators is given by the functions

$$f_\tau(p) = (1-\tau)\, p^{-\tau}, \qquad \tau \in (0,1). \tag{12}$$

An optimal e-calibrator would yield the largest e-value across all possible p-values $p \in [0,1]$. However, within the family (12), no single value of parameter τ yields the largest e-value $f_\tau(p)$ for all possible values of $p \in [0,1]$ (Vovk and Wang, 2021). Therefore, the optimal choice of τ is non-trivial, and a common selection is $\tau = 1/2$ (Yanchenko et al., 2025).

For any e-calibrator $f(\cdot)$, define its generalized inverse as $f^{-1}(t) = \sup\{u \in [0,1] : f(u) \geq t\}$.

For each configuration $k \in \{1, \ldots, K\}$, let $\widehat{F}_k(\cdot)$ denote the empirical CDF formed by (2) using the entire dataset $\mathcal{D}_k$. For an arbitrary e-calibrator $f(\cdot)$, PS-DME produces the bounds

$$L_k(x) = \max\{0,\, \widehat{F}_k(x) - \varepsilon_k\}, \tag{13a}$$

$$\text{and} \quad U_k(x) = \min\{1,\, \widehat{F}_k(x) + \varepsilon_k\}, \tag{13b}$$

with band width

$$\varepsilon_k = \sqrt{\frac{1}{2n_k} \log\left(\frac{2}{f^{-1}\left(\frac{K}{\delta|\mathcal{K}|}\right)}\right)}. \tag{14}$$

Using the properties of e-values (Vovk and Wang, 2021), PS-DME can be shown to have the following validity property, which is proved in Appendix D.

Theorem 1. *For any selection rule $S(\cdot)$ and for any e-calibrator $f(\cdot)$, the confidence bands in (13a) and (13b) produced by PS-DME satisfy the requirement (8), i.e.,* $\mathrm{FCR} \leq \delta$.

An alternative implementation of PS-DME that leverages Berk-Jones (BJ) confidence bounds (Berk and Jones, 1979) instead of the DKW inequality (9) is presented in Appendix F. BJ confidence bounds are implicitly defined based on likelihood ratios, yielding input-dependent confidence band widths. This contrasts with the sup-norm form of the DKW statistic (9), which yields efficient, symmetric, and input-independent confidence widths (14).

3.3 Comparing SS-DME and PS-DME

Thanks to its use of the full dataset for distributional inference, PS-DME can be more efficient than SS-DME. In fact, a direct comparison of the band widths (14) and (11) yields an explicit condition under which PS-DME yields a narrower band than SS-DME as follows.

Lemma 2. *For a pre-selection set size $|\mathcal{K}|$, the band width of PS-DME (14) is smaller than that of SS-DME (11) if and only if the inequality*

$$\frac{n_k^{\mathrm{eval}}}{n_k} < \frac{\log\left(\frac{2}{\delta}\right)}{\log\left(\frac{2}{f^{-1}\left(\frac{K}{\delta|\mathcal{K}|}\right)}\right)} \tag{15}$$

holds.

The condition (15) highlights the fundamental trade-off between the two methods. PS-DME benefits from using the full sample size n_k for both pre-selection and inference, but it pays an additional post-selection correction factor captured by the term $f^{-1}(K/\delta|\mathcal{K}|)$. In contrast, SS-DME avoids this correction by enforcing separation between pre-selection and inference, but it uses only the reduced sample size n_k^{eval}.

Using the e-calibrator family (12), the result in Lemma 2 can be specialized as follows (see Appendix E for a proof).

Corollary 2. *Assume that the e-calibrator belongs to the power family* (12)*. A necessary condition for the PS-DME width (14) to be smaller than the SS-DME width (11) is*

$$\frac{n_k^{\mathrm{eval}}}{n_k} < \frac{\log\left(\frac{2}{\delta}\right)}{\log 2 - W\left(-\frac{\delta|\mathcal{K}|}{eK}\right)}, \tag{16}$$

where $W(\cdot)$ denotes the lower real branch of the Lambert W function, i.e., the inverse of the function $f(w) = we^w$ on the interval $[-1/e, 0)$. Furthermore, when the candidate set size $|\mathcal{K}|$ is fixed and does not depend on the data $\mathcal{D}_k$, if condition (16) holds, there exists a value of the constant τ for which the width (14) of PS-DME is smaller than the width of SS-DME (11), namely

$$\tau = 1 + \frac{1}{W\left(-\frac{\delta|\mathcal{K}|}{eK}\right)}. \tag{17}$$

An illustration of this result can be found in Appendix A.

4 Related Work

Referring to Appendix B for further details, we highlight here connections to the literature on post-selection inference.

In the literature on inference after data-dependent selection (Lee et al., 2016; Taylor and Tibshirani, 2018), the FCR criterion was introduced in (Benjamini and Yekutieli, 2005) as an analog of the false discovery rate for confidence intervals. More recently, reference (Xu et al., 2024) developed general procedures for controlling FCR under arbitrary data-dependent selection. This methodology leverages e-values for their in-sample post-selection validity (Grünwald, 2024; Vovk and Wang, 2021). Related work has clarified the connection between e-values, post-hoc p-values, and data-dependent testing procedures (Koning, 2025; Chugg et al., 2026).

In the context of language models, prompt risk control (Zollo et al., 2024) develops upper confidence bounds for a family of scalar risk functionals of the loss distribution, including value-at-risk and conditional value-at-risk. Furthermore, (Feng et al., 2026) studies Type-I error control when the KPI is itself produced by a noisy LLM judge. This is complementary to our framework, which can be instantiated with any judge-based KPI (see Appendix G.11).

In concurrent work, leveraging the framework of e-values, reference (Koobs and Koning, 2026) developed a mechanism for obtaining confidence intervals based on data-dependent confidence levels. Framing the contribution in the context of the present work, the method in (Koobs and Koning, 2026) applies to a single model and to a fixed statistic, e.g., the average risk. Therefore, it cannot be leveraged for model selection, as well as for exploring the test-time trade-offs between performance and reliability.

5 Experiments

5.1 Benchmarks and Performance Measures

In this section, we compare three distributional model evaluation strategies by leveraging different metrics capturing reliability and efficiency for both performance evaluation and final selection.

5.1.1 Distributional Model Evaluation Strategies

We compare the following strategies:

1) Sample-splitting distributional model evaluation (SS-DME): As described in Sec. 3.1, SS-DME partitions the dataset for each configuration into independent selection and inference subsets. Model selection is carried out using the selection split, and CDF bands are constructed on the inference split via the DKW-based confidence intervals (10a) and (10b).

2) Naive in-sample distributional model evaluation (N-PS-DME): This benchmark reuses the full dataset $\mathcal{D}$ for both model selection and distributional inference, applying the same construction (10a)-(10b) as in SS-DME with n_k in place of n_k^{eval}. Due to selection bias, N-PS-DME cannot generally guarantee the FCR requirement (8).

3) In-sample distributional model evaluation (PS-DME): The proposed PS-DME, detailed in Sec. 3.2, reuses the full dataset for both selection and inference, leveraging an e-value-based methodology to construct valid post-hoc CDF bounds (13a) and (13b) that satisfy the FCR condition (8). Unless otherwise stated, we use the power e-calibrator (12) with the optimal parameter τ calculated as discussed in Sec. 3.3.

5.1.2 Performance Measures

We adopt the following metrics to assess both reliability and efficiency of different strategies.

1) False coverage proportion (FCP): For each simulated trial, we compute the realized FCP as defined in (6), namely the fraction of selected configurations $k \in \mathcal{K}$ for which the miscoverage event $\mathcal{O}_k$ in (5) occurs. Averaging across repetitions yields an empirical estimate of the FCR (7).

2) Confidence band width: We quantify statistical efficiency via the average uniform band radius. For SS-DME and N-PS-DME this corresponds to ε_k in (11), and for PS-DME to ε_k in (14). We report averages across selected configurations and experimental repetitions.

3) Post-selection guaranteed KPI: With reference to the example in Fig. 1(a), suppose that the system under study has a maximum allocated failure probability equal to γ. Using the lower bound $L_k(x)$, each configuration k can guarantee a KPI no worse than

$$X_k^* = \inf\{x : L_k(x) \geq 1 - \gamma\}, \tag{18}$$

while satisfying an error rate (FCR) no larger than δ. We provide an illustration of this definition in Fig. 3, which is further discussed in the next section. Upon optimization over the configuration $k \in \mathcal{K}$, the *post-selection best guaranteed KPI* is thus given by $X^* = \min_{k\in\mathcal{K}} X_k^*$.

5.2 Synthetic Data

We begin by validating the proposed PS-DME in a fully controlled synthetic setting, where the true data-generating distribution is known exactly. This allows us to directly assess both coverage properties and the tightness of the resulting CDF bands.

Setup: We consider a standard linear-Gaussian model with covariates drawn as $U \sim \mathcal{N}(0, I_{10})$, and responses generated according to the linear model $V = U^\top \beta + \varepsilon$, with noise $\varepsilon \sim \mathcal{N}(0, 1)$ and coefficient vector $\beta \sim \mathcal{N}(0, I_{10})$ whose entries are held fixed for all data points.

We construct $K = 2000$ candidate predictors $\widehat{g}_k(\cdot)$ by training ridge regression models over a logarithmically spaced grid of regularization parameters $\lambda_1, \ldots, \lambda_K$, where the grid spans the interval $[10^{-4}, 10^2]$. Each model is trained on an independent training dataset of size 600. For each model k, the KPI is defined as the squared prediction error $X_{k,i} = (V_i - \widehat{g}_k(U_i))^2$, evaluated on the same independent calibration dataset $\{(U_i, V_i)\}_{i=1}^n$ of size n. Thus, for each model k, we have the KPI samples $\mathcal{D}_k = \{X_{k,i}\}_{i=1}^n$ that are i.i.d. draws from an unknown distribution P_k, and samples $\{X_{k,i}\}_{k=1}^K$ are correlated across different hyperparameters. Model pre-selection chooses the $M = 1000$ configurations with the smallest empirical mean KPI. Throughout all experiments, we target an FCR level $\delta = 0.1$ in the requirement (8).

Results: To start, we illustrate the per-configuration CDF estimates produced by different schemes for a given candidate hyperparameter, selected here as $\lambda_k = 0.0231$. Fig. 2 plots the corresponding true risk CDF $F_k(x)$, the empirical CDF $\widehat{F}_k(x)$, and the corresponding post-selection confidence band $[L_k(x), U_k(x)]$ provided by SS-DME, N-PS-DME, and PS-DME using a calibration dataset of size $n = 20$. For SS-DME, we use 50% of the available calibration dataset for pre-selection, and the remainder 50% for inference. In this example, while SS-DME and PS-DME provide miscoverage control, N-PS-DME exhibits a miscoverage event. This illustrates the general theoretical result that SS-DME and PS-DME offer the FCR requirement (8), while N-PS-DME does not provide error control guarantees due to selection bias.

Fig. 2 also confirms the comparison between the widths of SS-DME and PS-DME formalized by Lemma 2. For the given split, the condition (15) is satisfied, and accordingly, the band width achieved by PS-DME is markedly narrower than that of SS-DME. As detailed in Sec. 3.2, this can be attributed to the fact that PS-DME uses the entire calibration dataset for forming the bands, while SS-DME is forced to use only a subset of the dataset for inference. As highlighted in the figure, this improved efficiency allows PS-DME to select a KPI X_k^* in (18) that is smaller, and thus preferable, than SS-DME.

We now turn to further analyzing the post-selection guaranteed KPI. Fig. 3 reports the best post-selection guaranteed KPI as a function of the target probability γ for SS-DME and for PS-DME. N-PS-DME does not guarantee FCR control, and thus is not included in this analysis. PS-DME is seen to achieve uniformly lower KPI (quadratic error) across all values of target probability $1 - \gamma$.

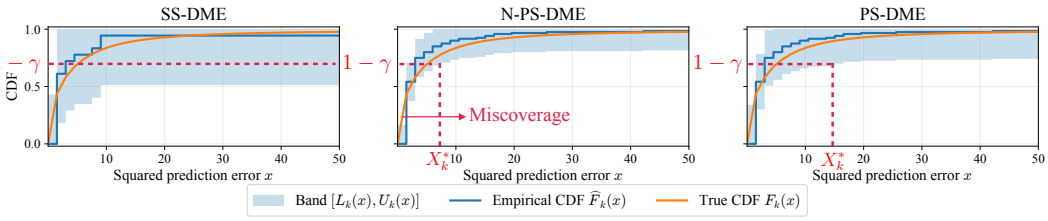
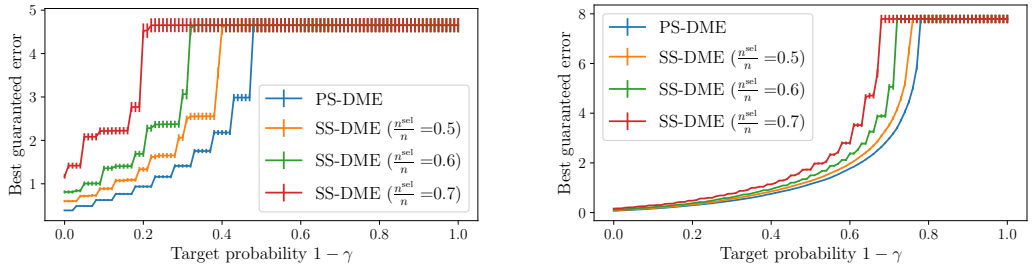


Figure 2: For a candidate hyperparameter $\lambda_k = 0.0231$, we show the true CDF $F_k(x)$, the empirical CDF $\widehat{F}_k(x)$, and the post-selection confidence band $[L_k(x), U_k(x)]$ for the data splitting, naive in-sample, and post-hoc in-sample benchmarks.



(a) Calibration size $n = 20$.

(b) Calibration size $n = 100$.

Figure 3: Best guaranteed KPI for the synthetic data experiment as a function of the target probability $1-\gamma$ for SS-DME and PS-DME. For SS-DME, a fraction $n^{\text{sel}}/n \in \{0.5, 0.6, 0.7\}$ of the available data is used for model pre-selection, while the remaining samples are used to construct the CDF confidence bands. (a) Results with 20 calibration samples. (b) Results with 100 calibration samples. The vertical error bars show the standard error of the mean across 100 random splits of the dataset.

Highest benefits are observed in the regimes of low test-time failure rates γ and when calibration data are scarce, i.e., for $n = 20$. In fact, in a low-sample regime, SS-DME suffers from a severe reduction in effective inference sample size, since only a fraction of the already limited data is available for band construction. Empirical FCR results are reported in Appendix G.1.

5.3 Evaluating and Selecting LLM Decoding Configurations

We next instantiate the proposed PS-DME in a text-to-SQL task performed by an LLM, where the KPI is a score accounting for the quality of the generated SQL query.

Task and dataset: We consider the text-to-SQL Spider benchmark (Yu et al., 2018), in which the LLM receives as input (*i*) a natural-language question, (*ii*) a database identifier specifying which database the question refers to, and (*iii*) the corresponding database schema, including table names, columns, and their relationships. The goal of the model is to generate an SQL query that, when executed on the specified database, correctly answers the natural-language question. Performance is evaluated by comparing the execution result of the generated query with that of the gold (ground-truth) SQL query provided in the dataset. In our experiments, we restrict attention to a fixed subset of $n = 200$ (randomly selected) validation examples from Spider (Yu et al., 2018).

We adopt the `Qwen/Qwen2.5-7B-Instruct` LLM (Yang et al., 2024) with nucleus (top-p) sampling (Holtzman et al., 2019), with parameters given by `top_p` $\in \{0.90, 0.95\}$ and temperature $\tau \in \{0.3, 0.7, 1.0\}$. The maximum output length is 160 tokens.

The hyperparameters correspond to decoding strategies that trade off computational budget against solution quality. In total, we consider $K = 43$ candidate configurations, organized into three families:

(i) Single decode: We include one greedy (deterministic) decoding together with the six stochastic single-pass decoding strategies obtained by varying `top_p` $\in \{0.90, 0.95\}$ and temperature $\tau \in \{0.3, 0.7, 1.0\}$.

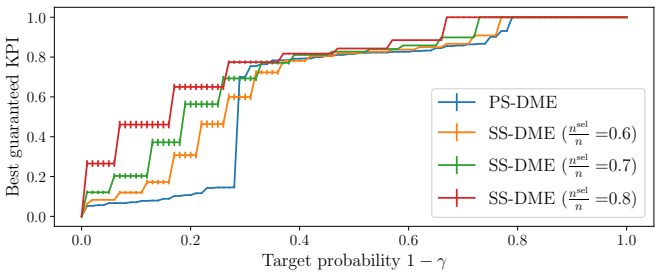


Figure 4: Best guaranteed KPI (loss) for the Spider text-to-SQL experiment as a function of the target probability $1-\gamma$ for SS-DME and PS-DME. For SS-DME, a fraction $n^{\text{sel}}/n \in \{0.6, 0.7, 0.8\}$ of the available data is used for decoding strategy pre-selection, while the remaining samples are used to construct the CDF confidence bands. Results are shown for $n = 50$ calibration samples. The vertical error bars show the standard error of the mean across 100 random splits of the dataset.

(ii) Best-of-n (BoN) decoding: Best-of-n sampling (Brown et al., 2020; Wang et al., 2022) generates $n \in \{2, 4, 6\}$ independent SQL candidates for the same input, and selects the one achieving the highest reward. For each value of parameter n, we vary temperature $\tau \in \{0.3, 0.7, 1.0\}$ and `top_p` $\in \{0.90, 0.95\}$, yielding 18 BoN configurations.

(iii) Iterative draft–refine (IAD) decoding: Iterative self-refinement methods (Madaan et al., 2023) improve outputs by repeatedly conditioning on previous drafts. Starting from an initial draft SQL generated from the prompt, the model produces revised SQL queries conditioned on the previous draft. This process is repeated for up to $t \in \{2, 4, 6\}$ iterations, after which the final SQL query is returned. For each value of t, we vary temperature $\tau \in \{0.3, 0.7, 1.0\}$ and `top_p` $\in \{0.90, 0.95\}$, yielding 18 IAD configurations.

To evaluate the quality of a generated SQL query, we construct a performance metric that captures both structural similarity at the token level and semantic correctness relative to the reference query. The token-level KPI measures the overlap between the SQL tokens appearing in the predicted and reference queries using the F1 score, i.e., the harmonic mean of precision and recall. The semantic correctness measure determines whether the generated query retrieves the correct information from the database leveraging the F1 score (Yu et al., 2018). The final KPI is given by $1 - wF1_{\text{token}} - (1-w)F1_{\text{exec}}$, where $F1_{\text{token}}$ and $F1_{\text{exec}}$ are the discussed F1 scores. We set $w = 0.3$, assigning greater importance to execution-level correctness, since the primary objective of a text-to-SQL system is to retrieve the correct database result. Model pre-selection chooses the $M = 30$ configurations with the smallest empirical mean KPI.

Results: Fig. 4 compares the post-selection guaranteed KPI achieved by SS-DME, which uses a fraction $n^{\text{sel}}/n \in \{0.6, 0.7, 0.8\}$ of the dataset for pre-selection, with that obtained by the proposed PS-DME. As shown in the figure, PS-DME consistently guarantees lower KPI values, corresponding to higher performance for the selected decoding strategy.

6 Limitations

The validity guarantees of PS-DME in Theorem 1 rests on three assumptions that we restate below.

Independent samples: For each configuration k, the observations in dataset $\mathcal{D}_k$ are assumed to be i.i.d., and this PS-DME is not directly applicable to time-series data.

Scalar KPI: PS-DME evaluates one scalar metric at a time. Extending the framework to jointly evaluate several metrics simultaneously is left to future work.

Choice of e-calibrator: As stated in Theorem 1, the FCR guarantee holds for any valid e-calibrator, but the width of the resulting confidence bands does depend on the parameter τ. Corollary 2 gives the optimal τ when the number of selected configurations $|\mathcal{K}|$ is known in advance, and Appendix G.7 provides a sensitivity analysis on the choice of τ.

7 Conclusion

We introduced PS-DME, a framework for statistically valid distributional evaluation of pre-selected models under full data reuse. By leveraging e-values, the proposed method constructs post-selection confidence bands for entire CDFs with guaranteed false coverage rate control, addressing the bias induced by data-dependent model selection.

Unlike conventional target-based approaches, PS-DME enables uncertainty-aware assessment of the full KPI distribution, supporting reliable exploration of performance–reliability trade-offs across operating points such as latency quantiles and outage levels. Compared to sample splitting, PS-DME can yield narrower confidence bands when the efficiency gains from full data reuse outweigh the cost of post-selection correction, as formalized through explicit width comparisons.

Experiments on synthetic data, LLM decoding, and telecom evaluation demonstrate that PS-DME achieves valid post-selection inference while often providing tighter bands and stronger guarantees than sample-splitting baselines.

Future work includes extending the framework to more general and multivariate performance measures, developing sharper confidence band constructions, and integrating distributional guarantees into downstream decision-making for reliable model deployment.

Acknowledgments

This work was supported by the European Research Council (ERC) under the European Union's Horizon Europe Programme (grant agreement No. 101198347). The work of O. Simeone was also supported by an EPSRC Open Fellowship (EP/W024101/1) and by the EPSRC project (EP/X011852/1).

References

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Luiz André Barroso, Urs Hölzle, and Parthasarathy Ranganathan. *The datacenter as a computer: Designing warehouse-scale machines*. Springer Nature, 2019.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

Robert H Berk and Douglas H Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1):47–59, 1979.

Leonardo Bonati, Salvatore D'Oro, Michele Polese, Stefano Basagni, and Tommaso Melodia. Intelligence and learning in O-RAN for data-driven nextg cellular networks. *IEEE Communications Magazine*, 59(10):21–27, 2021.

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.

Abhinav Chakraborty, Junu Lee, and Eugene Katsevich. Comparing three learn-then-test paradigms in a multivariate normal means problem. *arXiv preprint arXiv:2601.07764*, 2026.

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. Certifying counterfactual bias in LLMs. In *International Conference on Learning Representations (ICLR)*, 2025.

Ben Chugg, Etienne Gauthier, Michael I Jordan, Aaditya Ramdas, and Ian Waudby-Smith. Post-hoc large-sample statistical inference. *arXiv preprint arXiv:2603.08002*, 2026.

Jeffrey Dean and Luiz André Barroso. The tail at scale. *Communications of the ACM*, 56(2):74–80, 2013.

Lutz Dümbgen and Jon A Wellner. A new approach to tests and confidence bands for distribution functions. *The Annals of Statistics*, 51(1):260–289, 2023.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pages 642–669, 1956.

Amirmohammad Farzaneh and Osvaldo Simeone. Ensuring reliability via hyperparameter selection: Review and advances. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pages 1163–1167. IEEE, 2025.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations (ICLR)*, 2026.

Peter D Grünwald. Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121, 2024.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Hanxian Huang, Igor Fedorov, Andrey Gromov, Bernard Beckerman, Naveen Suda, David Eriksson, Maximilian Balandat, Rylan Conway, Patrick Huber, Chinnadhurai Sankar, Ayushi Dalmia, Zechun Liu, Lemeng Wu, Tarek Elgamal, Adithya Sagar, Vikas Chandra, and Raghuraman Krishnamoorthi. Scale. 2026. URL `https://arxiv.org/abs/2603.15954`.

Leah Jager and Jon A. Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018 – 2053, 2007. doi: 10.1214/0009053607000000244. URL `https://doi.org/10.1214/0009053607000000244`.

Kaggle Contributors. Http server logs dataset. `https://www.kaggle.com/`, 2020. Accessed: 2026.

Nick W. Koning. Post-hoc α hypothesis testing and the post-hoc p-value, 2025. URL `https://arxiv.org/abs/2312.08040`.

Stan Koobs and Nick W. Koning. Equivalence testing with data-dependent and post-hoc equivalence margins, 2026. URL `https://arxiv.org/abs/2603.16213`.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

Arun Kumar Kuchibhotla. Post-selection inference. In *International Encyclopedia of Statistical Science*, pages 1920–1924. Springer, 2025.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. Can LLM already serve as a database interface? A BIg bench for large-scale database grounded text-to-SQLs. *Advances in Neural Information Processing Systems*, 36, 2023.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

Marc Noé. The calculation of distributions of two-sided Kolmogorov-Smirnov type statistics. *The Annals of mathematical statistics*, pages 58–64, 1972.

Art B Owen. Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association*, 90(430):516–521, 1995.

Michele Polese, Leonardo Bonati, Salvatore D'oro, Stefano Basagni, and Tommaso Melodia. Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges. *IEEE Communications Surveys & Tutorials*, 25(2):1376–1411, 2023.

Petar Popovski, Kasper Fløe Trillingsgaard, Osvaldo Simeone, and Giuseppe Durisi. 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. *IEEE Access*, 6:55765–55779, 2018.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1(1-2):1–390, 2025.

Galen R Shorack and Jon A Wellner. *Empirical processes with applications to statistics*. SIAM, 2009.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv: 2107.07511*, 2021.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Larry Wasserman. *All of nonparametric statistics*. Springer, 2006.

Ziyu Xu, Ruodu Wang, and Aaditya Ramdas. Post-selection inference for e-value based confidence intervals. *Electronic Journal of Statistics*, 18(1):2292–2338, 2024.

Eric Yanchenko, Jonathan P Williams, and Ryan Martin. Hypothesis testing for community structure in temporal networks using e-values. *arXiv preprint arXiv:2507.23034*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL `https://arxiv.org/abs/2407.10671`.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. *arXiv preprint arXiv:1809.08887*, 2018.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

A Additional Illustrations

In this section, we provide additional illustrations for concepts introduced in the main text.

Fig. 5 illustrates the definition of a miscoverage event $\mathcal{O}_k$, as defined in (5).

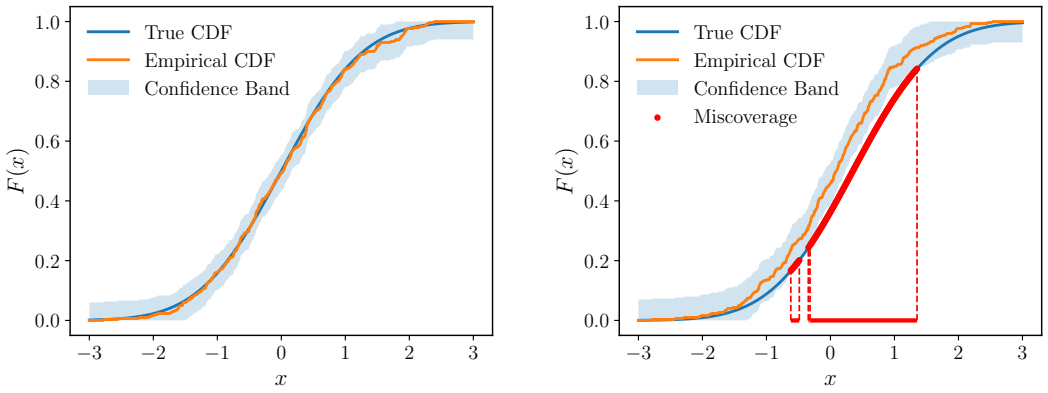


Figure 5: Illustration of the miscoverage event $\mathcal{O}_k$ defined in (5). (a) A scenario in which the miscoverage event $\mathcal{O}_k$ in (5) does not occur, as the true CDF lies entirely within the reported confidence band. (b) A scenario in which the miscoverage event $\mathcal{O}_k$ in (5) occurs, since the true CDF exits the band over the region of input x marked in red.

Fig. 6 illustrates the band width (14) achieved by PS-DME when using the e-calibrators in (12) with selected values of parameters τ, as well as with the optimal calibrator parameter τ in (17), together with the band width (11) obtained by SS-DME, as a function of the dataset split ratio n_k^{eval}/n_k for fixed values $\delta = 0.1$ and $\delta|\mathcal{K}|/K = 0.01$. The intersection of the curves corresponding to PS-DME and SS-DME occurs at the threshold (15) identified in Lemma 2 beyond which SS-DME yields a smaller width. It is observed that the choice of τ in (17) yields the larger range of values of the ratio n_k^{eval}/n_k over which PS-DME is advantageous over SS-DME.

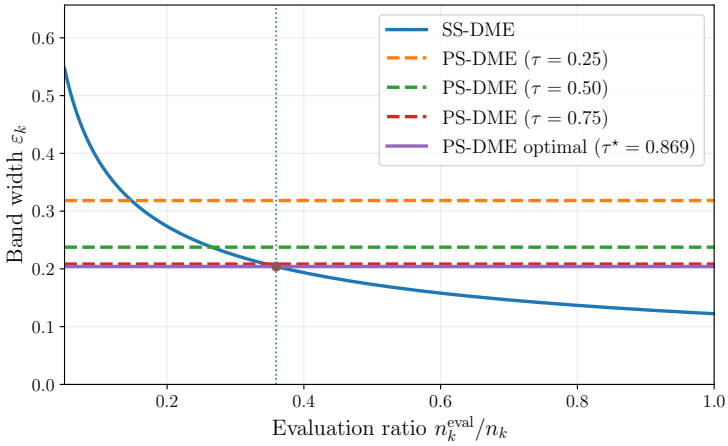


Figure 6: Comparison of the band widths produced by SS-DME and by PS-DME with the e-calibrator family (12). The parameter $\delta|\mathcal{K}|/K$ is fixed to 0.01.

B Additional Related Work

Reliable hyperparameter selection: Recent work has shown that hyperparameter selection can be cast as an MHT problem, enabling finite-sample guarantees for the selected model (Angelopoulos

et al., 2025). This framework has been extended to a variety of performance criteria and engineering applications (Stephen et al., 2021; Bates et al., 2023; Angelopoulos et al., 2022); see (Farzaneh and Simeone, 2025) for a recent review. A key limitation of this literature is that it focuses on scalar performance metrics, e.g., average risk, relative to a pre-specified target level. In contrast, this work aims at *distributional* evaluation, enabling comparison of models across all test-time reliability levels without committing to a single KPI threshold.

Confidence bands for distribution functions: Classical nonparametric methods provide confidence bands for a single distribution function, notably via the DKW inequality (Dvoretzky et al., 1956; Massart, 1990). Likelihood-ratio and goodness-of-fit approaches, such as Berk–Jones and empirical likelihood methods (see Appendix F), can yield tighter or more adaptive bands (Berk and Jones, 1979; Owen, 1995; Dümbgen and Wellner, 2023). However, these methods are designed for marginal inference and do not account for model selection. In contrast, our setting requires validity after arbitrary data-dependent screening.

C Proof of Lemma 1

For each fixed k, the DKW inequality (9) guarantees the inequality

$$\mathbb{P}(F_k \notin \mathcal{C}_k) \leq \delta. \tag{19}$$

Therefore, we have

$$\begin{aligned}
\text{FCR} &= \mathbb{E}_{\mathcal{K}}\left[\mathbb{E}\left[\frac{1}{\max\{|\mathcal{K}|,1\}}\sum_{k\in\mathcal{K}}\mathbf{1}\{F_k \notin \mathcal{C}_k\}\middle|\mathcal{K}\right]\right] && \text{(20a)}\\
&= \mathbb{E}_{\mathcal{K}}\left[\frac{1}{\max\{|\mathcal{K}|,1\}}\mathbb{E}\left[\sum_{k\in\mathcal{K}}\mathbf{1}\{F_k \notin \mathcal{C}_k\}\middle|\mathcal{K}\right]\right] && \text{(20b)}\\
&\leq \mathbb{E}_{\mathcal{K}}\left[\frac{1}{\max\{|\mathcal{K}|,1\}}\sum_{k\in\mathcal{K}}\delta\right] && \text{(20c)}\\
&\leq \delta && \text{(20d)}
\end{aligned}$$

where the outer expectation $\mathbb{E}_{\mathcal{K}}[\cdot]$ is with respect to the pre-selected subset $\mathcal{K}$, which establishes the guarantee.

D Proof of Theorem 1

For any candidate CDF F, consider the null hypothesis

$$\mathcal{H}_{0,k}(F): \quad F_k = F \tag{21}$$

that the candidate distribution F coincides with the true distribution F_k. Define the Kolmogorov distance (Wasserman, 2006)

$$T_k(F) = \sup_{x\in\mathbb{R}}\left|\widehat{F}_k(x) - F(x)\right|. \tag{22}$$

By the DKW inequality (Dvoretzky et al., 1956), under the null hypothesis (21), the inequality

$$\mathbb{P}(T_k(F) > \varepsilon \mid \mathcal{H}_{0,k}(F)) \leq 2\exp\left(-2n_k\varepsilon^2\right) \tag{23}$$

holds. As shown next, this implies that the following statistic

$$p_k(F) = \min\left\{1,\, 2\exp\left(-2n_k(T_k(F))^2\right)\right\} \tag{24}$$

is a valid p-value for the null hypothesis (21).

Lemma 3. *For the null hypothesis $\mathcal{H}_{0,k}(F)$ in* (21)*, the statistic $p_k(F)$ in* (24) *is a valid p-value, i.e., for all $\alpha \in (0,1)$,*

$$\mathbb{P}(p_k(F) \leq \alpha \,|\, \mathcal{H}_{0,k}(F)) \leq \alpha. \tag{25}$$

Proof. Fix $\alpha \in (0,1)$. The event $\{p_k(F) \leq \alpha\}$ is equivalent to the inequality

$$T_k(F) \geq \sqrt{\frac{1}{2n_k} \log\left(\frac{2}{\alpha}\right)}. \tag{26}$$

Under hypothesis $\mathcal{H}_{0,k}(F)$, the DKW inequality (Dvoretzky et al., 1956) implies

$$\mathbb{P}\left(T_k(F) \geq \sqrt{\frac{1}{2n_k} \log\left(\frac{2}{\alpha}\right)} \middle| \mathcal{H}_{0,k}(F)\right) \leq 2\exp\left(-2n_k \cdot \frac{1}{2n_k} \log\frac{2}{\alpha}\right) = \alpha, \tag{27}$$

which establishes super-uniformity and concludes the proof. □

Given the p-value $p_k(F)$ in (24), we construct the statistic

$$E_k(F) = f(p_k(F)), \tag{28}$$

which is a valid e-value, i.e.,

$$\mathbb{E}[E_k(F) \mid \mathcal{H}_{0,k}(F)] \leq 1. \tag{29}$$

The confidence bound $C_k = \{F : L_k(x) \leq F(x) \leq U_k(x) \text{ for all } x \in \mathbb{R}\}$ with (13a) and (13b) produced by PS-DME is obtained by inverting the binary test of the null hypothesis $\mathcal{H}_{0,k}(F)$ using the e-value $E_k(F)$ as

$$\mathcal{C}_k = \left\{F : \ E_k(F) < \frac{K}{\delta|\mathcal{K}|}\right\} = \left\{F : p_k(F) > f^{-1}\left(\frac{K}{\delta|\mathcal{K}|}\right)\right\}. \tag{30}$$

Let $\mathcal{O}_k$ denote the miscoverage event (5). By construction of set (30), we have

$$\mathbf{1}\{\mathcal{O}_k\} = \mathbf{1}\left\{E_k(F_k) \geq \frac{K}{\delta|\mathcal{K}|}\right\} \leq \frac{\delta|\mathcal{K}|}{K} E_k(F_k), \tag{31}$$

where F_k represents the true CDF, and the inequality follows from the condition $\mathbf{1}\{x \geq a\} \leq x/a$, which holds for all $x \geq 0$ and $a > 0$. Therefore, taking the expectation with respect to the true joint distribution of the KPIs, the FCR in (7) is evaluated as

$$\begin{aligned}
\mathrm{FCR} &= \mathbb{E}\left[\frac{1}{\max\{|\mathcal{K}|, 1\}} \sum_{k\in\mathcal{K}} \mathbf{1}\{\mathcal{O}_k\}\right] && \text{(32a)}\\
&\leq \mathbb{E}\left[\frac{1}{\max\{|\mathcal{K}|, 1\}} \sum_{k\in\mathcal{K}} \frac{\delta|\mathcal{K}|}{K} E_k(F_k)\right] && \text{(32b)}\\
&\leq \frac{\delta}{K} \sum_{k=1}^{K} \mathbb{E}[E_k(F_k)] && \text{(32c)}\\
&= \frac{\delta}{K} \sum_{k=1}^{K} \mathbb{E}\left[E_k(F_k)|\mathcal{H}_{0,k}(F_k)\right] \leq \delta && \text{(32d)}
\end{aligned}$$

E Proof of Corollary 2

The Vovk–Sellke bound (Vovk and Wang, 2021) states that for the power family of calibrators (12) we have

$$f_{\mathrm{VS}}(p) = \max_{\tau\in(0,1)} f_\tau(p) = \begin{cases} \dfrac{-\exp(-1)}{p\log p}, & p \leq e^{-1}, \\ 1, & p > e^{-1}, \end{cases} \tag{33}$$

where

$$\arg\max_{\tau\in(0,1)} f_\tau(p) = -\frac{1}{\log p} \tag{34}$$

Hence, for all $p \in (0,1)$ and $\tau \in (0,1)$, the inequality

$$f_\tau(p) \leq f_{\mathrm{VS}}(p) \tag{35}$$

holds. Note, however, that the function $f_{\mathrm{VS}}(p)$ is not a valid e-calibrator.

Since function $f_\tau(\cdot)$ is decreasing, the corresponding inverse satisfies the inequality $f_\tau^{-1}(y) \le f_{\mathrm{VS}}^{-1}(y)$ for all $y \ge 1$. Furthermore, the inverse of the Vovk-Sellke bound is given by

$$f_{\mathrm{VS}}^{-1}(y) = \exp\left(W\left(-\frac{1}{ey}\right)\right) \tag{36}$$

for $p \le e^{-1}$. Substituting $y = K/(\delta|\mathcal{K}|)$ in (36) and plugging the result into the denominator of (15), with $\log(2/\delta)$ in the numerator, yields the condition (16). Given (35), the inequality (15) for any τ implies (16), proving the necessity of condition (16). Furthermore, using (34), we obtain that the maximizing choice of the power-calibrator parameter at level $y = K/(\delta|\mathcal{K}|)$ is (17). Therefore, when $|\mathcal{K}|$ is fixed, condition (16) ensures that this choice of τ makes (15) hold, completing the proof.

F Alternative Implementation of PS-DME via Berk–Jones CDF Bands

The proof of Theorem 1 in Appendix D shows that PS-DME applies to any valid test of the null hypothesis $\mathcal{H}_{0,k}(F)$ in (21). While Sec. 3 uses the Kolmogorov statistic $T_k(F)$ in (22) together with the DKW inequality (9), one may instead use likelihood-ratio-type goodness-of-fit statistics such as the Berk–Jones (BJ) statistic (Berk and Jones, 1979).

Define the BJ statistic for a candidate CDF F as

$$S_{n_k}(F) = \sup_{x\in\mathbb{R}} K\big(\widehat{F}_k(x),\, F(x)\big), \tag{37}$$

where $K(\cdot,\cdot)$ is the Bernoulli KL divergence defined as

$$K(a,b) = a\log\left(\frac{a}{b}\right) + (1-a)\log\left(\frac{1-a}{1-b}\right), \qquad a,b \in (0,1). \tag{38}$$

Under the null hypothesis $\mathcal{H}_{0,k}(F)$ and assuming that F is continuous, the probability integral transform implies that the transformed variables $U_{k,i} = F(X_{k,i})$ are i.i.d. $\mathrm{Unif}[0,1]$ (Shorack and Wellner, 2009). Let $\widehat{G}$ denote the empirical CDF of $\{U_{k,i}\}_{i=1}^{n_k}$. Then, under the null, the statistic $S_{n_k}(F)$ in (37) has the same distribution as

$$S_{n_k}^{(0)} = \sup_{u\in[0,1]} K\big(\widehat{G}(u),\, u\big), \tag{39}$$

i.e., the BJ statistic computed from an i.i.d. sample drawn from $\mathrm{Unif}[0,1]$.

Let $q_{n_k}(1-\alpha)$ denote the $(1-\alpha)$ quantile of the null distribution of random variable $S_{n_k}^{(0)}$ in (39). Then, by test inversion (Owen, 1995; Jager and Wellner, 2007), the acceptance region

$$S_{n_k}(F) \le q_{n_k}(1-\alpha) \tag{40}$$

defines a valid $(1-\alpha)$ confidence set for the true CDF F_k. This condition is equivalent to requiring that, for all $x \in \mathbb{R}$, the value $F(x)$ lies in the set of all $u \in [0,1]$ satisfying

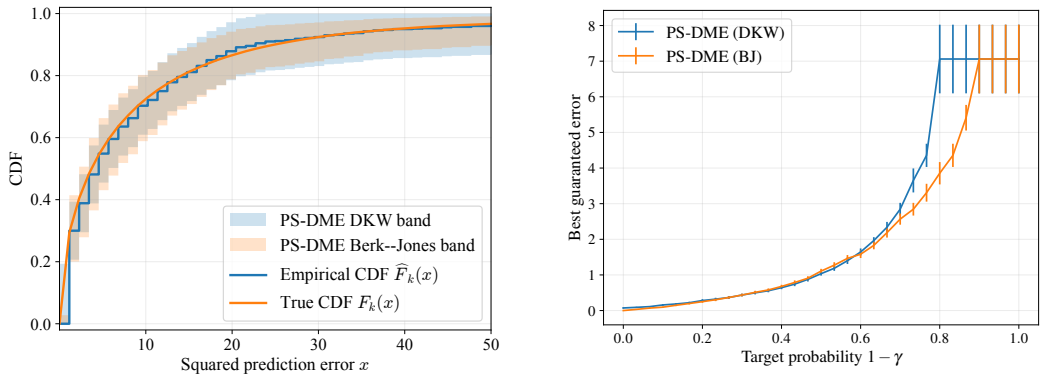
$$K\big(\widehat{F}_k(x),\, u\big) \le q_{n_k}(1-\alpha). \tag{41}$$

Accordingly, the lower and upper bounds at each x are obtained by solving this inequality for u, yielding confidence bands

$$L_k(x) \le F_k(x) \le U_k(x), \qquad \forall x \in \mathbb{R}, \tag{42}$$

which depend on the local value $\widehat{F}_k(x)$ and are generally asymmetric. To guarantee post-selection FCR control, the nominal level is set to $\alpha = f^{-1}(K/(\delta|\mathcal{K}|))$, consistent with the e-value inversion in (30).

Compared to the DKW-based bands in (13a)–(13b), whose width is the input-independent quantity ε_k in (14), the BJ bands are input-dependent. In particular, they are typically tighter in the tails and wider around the center of the CDF, reflecting the geometry of the KL divergence: deviations are penalized more strongly when $\widehat{F}_k(x)$ is close to 0 or 1, and more weakly near $\widehat{F}_k(x) \approx 1/2$ (Jager and Wellner,



(a) (b)

Figure 7: Comparison between the DKW-based and Berk–Jones-based implementations of PS-DME. (a) CDF confidence bands for a representative hyperparameter; (b) Best guaranteed KPI as a function of the outage probability γ.

2007). This improved adaptivity comes at the cost of computing the critical value $q_{n_k}(1-\alpha)$, e.g., via Noé's recursion (Noé, 1972), and solving a one-dimensional root-finding problem for each value of $\widehat{F}_k(x)$.

We finally compare the default DKW-based implementation of PS-DME in (13a)–(13b) with the BJ-based implementation described above. Fig. 7(a) shows the confidence bands obtained by the two methods for a representative hyperparameter. The BJ-based band is tighter in the tails of the CDF and wider around the center, in agreement with the discussion above.

The impact of this difference on model evaluation is illustrated in Fig. 7(b), which reports the post-selection best guaranteed KPI as a function of the target outage probability γ. The BJ-based implementation provides a stronger guarantee for most outage levels, particularly in the tail regimes. In contrast, for intermediate outage probabilities, the DKW-based implementation can yield slightly tighter guarantees, reflecting its uniform band width in (14).

G Additional Experiments

G.1 Empirical FCR for the Synthetic Data Experiment

Table 1 reports the empirical FCR for SS-DME and PS-DME on the synthetic experiment of Sec. 5.2, averaged over 100 random replications at target level $\delta = 0.1$. SS-DME uses a split ratio $n^{\text{sel}}/n = 0.5$. Both methods satisfy the FCR requirement (8).

Table 1: Empirical FCR for the synthetic experiment at $\delta = 0.1$, averaged over 100 replications.

n	Method	Empirical FCR
20	SS-DME	0.0401
	PS-DME	0.0116
100	SS-DME	0.0923
	PS-DME	0

G.2 Evaluating and Selecting Telecom Network Configurations

We evaluate the proposed PS-DME on a wireless networking dataset derived from an Open Radio Access Network (O-RAN) deployment (Bonati et al., 2021).

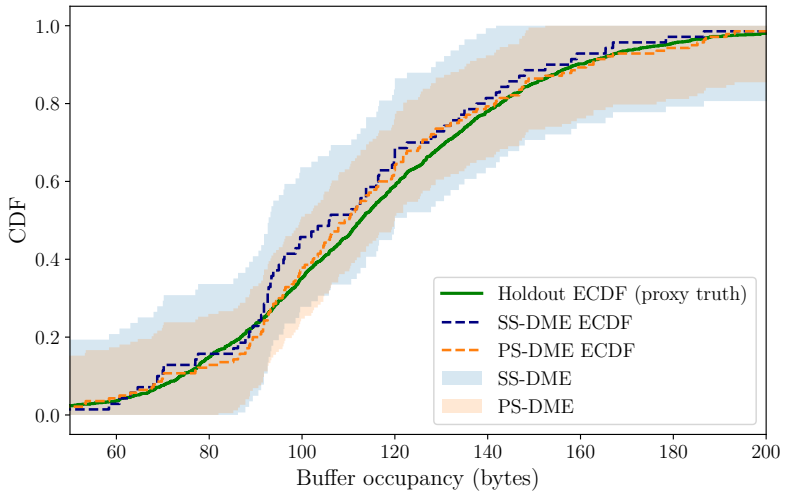


Figure 8: Representative post-selection CDF bands for the O-RAN buffer occupancy experiment for a selected configuration for PS-DME, and SS-DME with dataset split ratio $|\mathcal{D}^{\text{sel}}|/|\mathcal{D}| = 0.5$.

G.2.1 Dataset and system setup

We use the dataset introduced in (Bonati et al., 2021), which was collected on the Colosseum wireless network emulator at Northeastern University (Polese et al., 2023). The dataset contains measurements from an O-RAN deployment consisting of four base stations operating with a channel bandwidth of 3 MHz, consisting of 15 physical resource blocks (PRBs), and serving 40 user equipments (UEs). Traffic is organized into three slices representing distinct service classes, namely enhanced mobile broadband (eMBB), massive machine-type communications (MTC), and ultra-reliable low-latency communications (URLLC) (Popovski et al., 2018). The dataset reports periodic slice-level measurements, including the downlink buffer occupancy and downlink transmission rate for each slice, across different network configurations, which specify (*i*) the scheduling policy used for each slice; and (*ii*) the initial allocation of radio resources, i.e., PRBs, across slices. Specifically, the dataset specifies 18 predefined configurations, each corresponding to one particular combination of one of three schedulers and of one of six initial allocations for the three slices. We adopt the downlink buffer occupancy for the URLLC slice as the (negatively oriented) KPI $X_{k,i}$. The ground truth CDF of this KPI is estimated for each configuration by using 90% of the available data.

Model pre-selection is performed by selecting the six configurations with the smallest empirical mean buffer occupancy for the URLLC slice.

G.2.2 Results

Fig. 8 illustrates representative post-selection CDF bands for a selected O-RAN configuration for SS-DME and PS-DME, with SS-DME using half of the calibration data for pre-selection. The horizontal axis corresponds to the buffer occupancy KPI, while the vertical axis shows the corresponding CDF. As in the previous experiments, SS-DME yields wider bands compared to PS-DME. Additionally, because SS-DME uses fewer calibration samples to estimate the empirical CDF, its empirical CDF exhibits larger deviations from the proxy true CDF than that of PS-DME. Note also that, since SS-DME uses a subset of the calibration dataset for evaluating the empirical CDF, the empirical CDFs for SS-DME and PS-DME are different.

We then analyze the post-selection guaranteed KPI. Fig. 9 reports the best guaranteed buffer occupancy as a function of the target probability $1-\gamma$ for PS-DME, as well as SS-DME with dataset split ratios n^{sel}/n from the set $\{0.2, 0.3, 0.4\}$. The proposed PS-DME consistently yields stronger guarantees than the SS-DME variants.

To further investigate the trade-off induced by sample splitting, we fix two target probabilities, $1-\gamma \in \{0.7, 0.8\}$, and vary the split ratio n^{sel}/n used by SS-DME. Fig. 10 reports the resulting best guaranteed buffer occupancy as a function of the split ratio. The two panels correspond to different

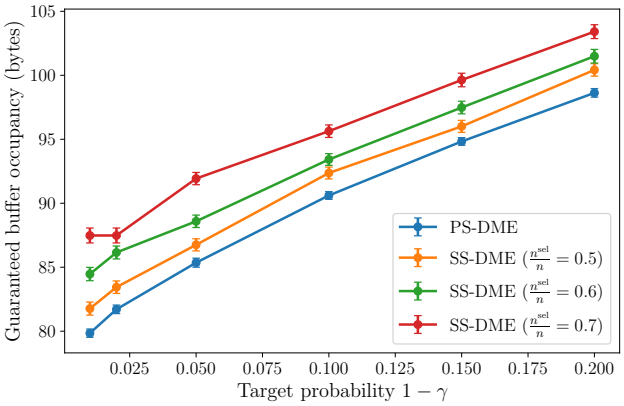
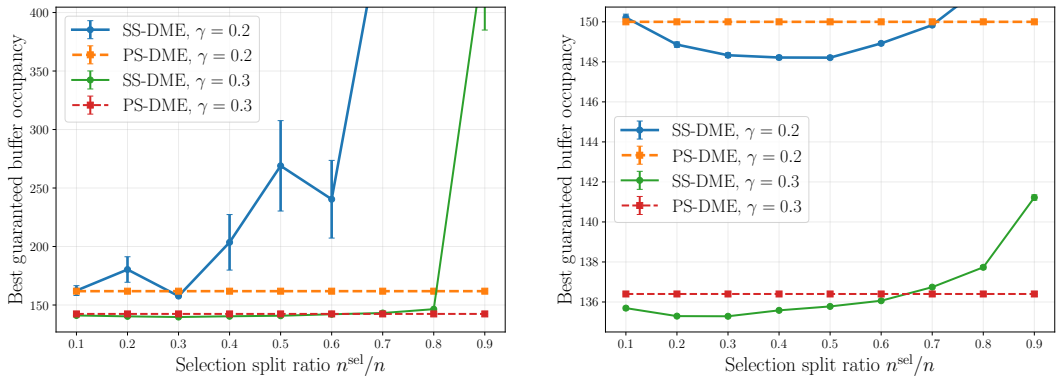


Figure 9: Best guaranteed buffer occupancy for the O-RAN experiment as a function of the target probability $1-\gamma$ for PS-DME and SS-DME with dataset split ratio $n^{\text{sel}}/n \in \{0.2, 0.3, 0.4\}$. The vertical error bars show the standard error of the mean across 100 random splits of the dataset.



(a) (b)

Figure 10: Effect of the SS-DME split ratio n^{sel}/n on the best guaranteed buffer occupancy in the O-RAN experiment, for two target probabilities $1-\gamma \in \{0.7, 0.8\}$. Each panel corresponds to a different calibration regime, with (a) 30% and (b) 100% of the available data used for calibration. The vertical error bars show the standard error of the mean across 1,000 random splits of the calibration dataset into pre-selection and evaluation datasets.

calibration regimes: Fig. 10(a) uses 30% of the available data for calibration, while Fig. 10(b) uses the entire available dataset. In each panel, we report results for both target probabilities $1-\gamma = 0.7$ and $1-\gamma = 0.8$.

In the low-calibration regime shown in Fig. 10(a), SS-DME consistently yields a larger best guaranteed buffer occupancy than PS-DME across all split ratios and both target probabilities. This reflects the fact that splitting an already limited calibration dataset reduces the effective sample size available for both model selection and inference, leading to more conservative guarantees and hence worse (larger) guaranteed KPI values. In contrast, in the high-calibration regime shown in Fig. 10(b), the trade-off becomes more nuanced. For intermediate values of the split ratio, SS-DME can outperform PS-DME for the less stringent target $1-\gamma = 0.7$, achieving a smaller guaranteed buffer occupancy. This improvement arises because allocating a moderate portion of the data to selection can lead to better configuration choices, while the remaining data are still sufficient to yield tight confidence bands. However, this advantage does not persist uniformly across all split ratios or for the more stringent target $1-\gamma = 0.8$, where the need for precise distributional estimation favors PS-DME, which leverages the full dataset.

G.3 Post-Selection CDF Inference for Server Latency

We next evaluate the proposed PS-DME confidence bands on response-time distributions extracted from HTTP server access logs. Latency is a canonical KPI in web services and distributed systems, where configurations are routinely compared and selected based on empirical performance measured on the same data (Dean and Barroso, 2013; Barroso et al., 2019). This setting provides a natural testbed for post-selection CDF inference, as distributional information, rather than a single scalar metric, is often of primary interest.

G.3.1 Setup

We consider a collection of HTTP server access logs obtained from a publicly available server log dataset released on Kaggle and commonly used for benchmarking log analysis and anomaly detection methods (Kaggle Contributors, 2020). Each log entry records, among other fields, the HTTP request method, the request path, and the response time measured at the server. In this experiment, the hyperparameters correspond to discrete request-handling configurations defined by ordered pairs $(\text{HTTP method}, \text{request path})$. The HTTP method takes values in $\{\texttt{GET}, \texttt{POST}, \texttt{PUT}, \texttt{DELETE}\}$, which comprise all methods observed in the dataset. The request path is defined as the URL path component of the request line, with query parameters removed, and takes values in the finite set of distinct paths observed across the log file.

To construct a finite hyperparameter grid, we enumerate all method–path pairs appearing in the dataset and compute their total frequency. We retain only those pairs that appear at least $N_{\min} = 200$ times, ensuring that each candidate configuration admits sufficiently many samples for stable empirical CDF estimation. Among the remaining pairs, we select the $K = 30$ most frequent configurations to form a fixed hyperparameter grid, which is determined once and held constant across all experimental repetitions. Each configuration k in this grid induces an unknown latency distribution P_k corresponding to the response times of requests matching the associated method–path pair.

Within each experimental repetition, the data corresponding to the fixed hyperparameter grid are randomly split into a calibration set and a holdout set. Specifically, 30% of the data are allocated to calibration and the remaining 70% to holdout. Using the calibration data, we compute the empirical mean latency for each configuration and select a subset $\mathcal{K} \subseteq \{1, \dots, K\}$ consisting of the $M = 12$ configurations with the smallest empirical mean latency. This selection rule is fully data-dependent and reuses the same samples later employed to construct confidence bands.

For each selected configuration $k \in \mathcal{K}$, we construct two-sided uniform CDF confidence bands using the methodology described in Section 3. To study the effect of the e-calibrator on both inferential tightness and post-selection reliability, we compare three members of the power-family of e-calibrators (12), and consider three parameters $\tau \in \{1/3,\ 1/2,\ 2/3\}$. For each calibrator, the corresponding band width is computed according to (14), and the same selected set $\mathcal{K}$ is used across all calibrators within each repetition to enable a fair comparison.

G.3.2 Results

Because the true latency distributions are unknown, we assess post-selection validity using the empirical CDF computed on the holdout set as a proxy for the true CDF. A band is declared to fail if the holdout empirical CDF exits the proposed interval at any point. We repeat the entire procedure over 300 independent random splits and compute the FCP for each repetition.

Fig. 11 compares post-selection CDF bands constructed using three different e-calibrators for a representative selected configuration. All curves share the same empirical CDF, while the shaded regions correspond to the confidence bands induced by the different values of τ. The resulting band widths vary with the choice of calibrator. In our experiment, the tightest band among the tested calibrators is obtained for $\tau = 1/3$, while both larger and smaller values of τ produce slightly wider bands. This behavior reflects the trade-off introduced by the e-calibrator in the inversion of the e-value tests: different calibrators allocate evidence differently across the range of p-values, which in turn affects the resulting confidence radius in (14). These results illustrate that the choice of calibrator can have a noticeable effect on inferential tightness, and that intermediate values of τ may provide the most informative bands in practice.

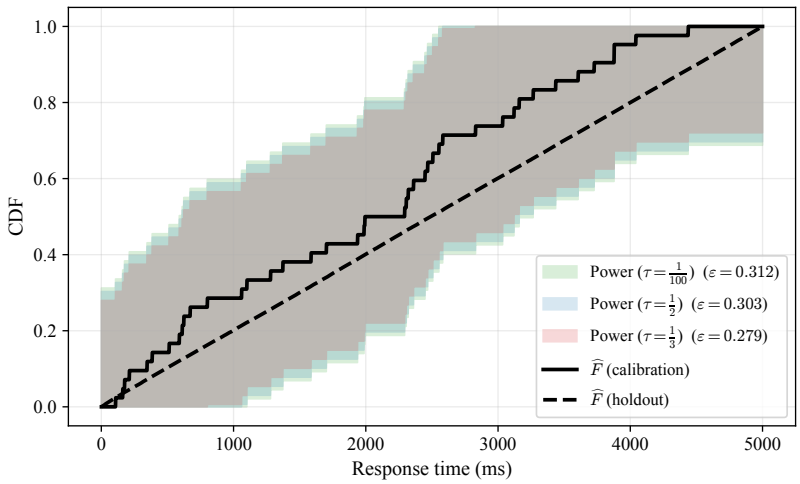


Figure 11: Representative post-selection CDF bands for a selected server configuration using three different e-calibrators. The empirical CDF is common across all curves, while the shaded regions correspond to different values of parameter τ in (12).

G.4 Empirical FCR on Real-World Benchmarks

To complement the empirical FCR table of Appendix G.1, we report the empirical FCR of PS-DME on the two real-world benchmarks of Sec. 5.3 and Appendix G.2. For the Spider text-to-SQL experiment, we use $n_{\text{cal}} = 100$ calibration samples and a held-out set of 100 samples as a proxy for the true distribution; results are averaged over 500 independent random splits. For the O-RAN experiment, we use the same split protocol as in Appendix G.2 and average over 300 random splits. In both cases, the target level is $\delta = 0.1$.

Table 2: Empirical FCR of PS-DME on the Spider text-to-SQL and O-RAN telecom benchmarks at target $\delta = 0.1$. Mean $\pm$ one standard error of the mean across random splits.

Benchmark	Setting	Empirical FCR
Spider (text-to-SQL)	$M = 5$ selected	0.031 ± 0.007
	$M = 10$ selected	0.046 ± 0.007
	$M = 20$ selected	0.070 ± 0.008
O-RAN (telecom)	$M = 6$ selected	0.007 ± 0.001

As shown in Table 2, PS-DME satisfies the FCR requirement (8) in all settings, even on the O-RAN benchmark where the i.i.d. assumption is only approximate (see Sec. 6). On the O-RAN dataset, 92% of random repetitions exhibit zero miscoverage, indicating substantial slack with respect to the nominal target.

G.5 Selection-Ratio Sweep

The experiments in Sec. 5 use a fixed selected-set size M. In this appendix, we vary M across its full operating range to characterize the dependence of the band width and the guaranteed KPI on the selection ratio M/K. We use the synthetic data generating process of Sec. 5.2 with $K = 31$ configurations, $n = 200$, $\delta = 0.1$, and 300 random repetitions. For each value of M, SS-DME is given the number of evaluation samples n^{eval} equal to 70% of the Corollary 2 crossover threshold, i.e., a regime in which Corollary 2 guarantees that PS-DME is narrower by design. We report the band width ε and the best guaranteed KPI $X^\star$ at target reliability $1 - \gamma = 0.75$ computed via the lower band and (18).

Table 3 shows that PS-DME dominates SS-DME on both the band width and the best guaranteed KPI across the entire range of M. As expected from Corollary 2, the width gap is largest when the

Table 3: Band width ε and best guaranteed KPI $X^\star$ at $1-\gamma = 0.75$ for PS-DME and SS-DME on synthetic data, as a function of the selected-set size M (with $K = 31$). Lower is better for both ε and $X^\star$.

M	M/K	$\varepsilon_{\mathrm{PS}}$	$\varepsilon_{\mathrm{SS}}$	$X^\star_{\mathrm{PS}}$	$X^\star_{\mathrm{SS}}$
2	0.06	0.149	0.179	0.793	0.814
5	0.16	0.139	0.168	0.791	0.807
10	0.32	0.132	0.158	0.790	0.804
15	0.48	0.127	0.153	0.789	0.800
20	0.65	0.124	0.148	0.789	0.799
25	0.81	0.121	0.145	0.789	0.796
30	0.97	0.119	0.142	0.787	0.798

selection is most aggressive (small M), since in this regime a sample-split allocation of n^{eval} samples to inference is especially wasteful.

G.6 Data-Dependent Selection

Theorem 1 covers arbitrary data-dependent selection rules $S(\cdot)$, including those for which $|\mathcal{K}|$ is itself random. To stress-test this property, we evaluate PS-DME under a selection rule that fixes a *target* selected set size M but keeps the realized $|\mathcal{K}|$ data-dependent: for each repetition, we rank configurations by the empirical mean of their KPIs and retain all configurations whose mean lies at or below the M-th order statistic. Ties and ex aequo values therefore induce randomness in $|\mathcal{K}|$. We use the synthetic setup of Appendix G.5 with $n_{\mathrm{cal}} = 100$ and 500 repetitions at $\delta = 0.1$.

Table 4: PS-DME under data-dependent selection with target size M. Mean realized $|\mathcal{K}|$, empirical FCR, and PS-DME band width $\varepsilon_{\mathrm{PS}}$ across 500 repetitions.

Target M	Mean $\lvert\mathcal{K}\rvert$	Empirical FCR	$\varepsilon_{\mathrm{PS}}$
5	5.0 ± 4.1	0.004 ± 0.003	0.203
10	9.2 ± 4.1	0.018 ± 0.006	0.191
20	21.3 ± 6.1	0.078 ± 0.012	0.175

Table 4 confirms that PS-DME controls FCR well below the target $\delta = 0.1$ for all target selection sizes, in agreement with Theorem 1.

G.7 Sensitivity to the Choice of e-Calibrator

We study the sensitivity of PS-DME to the e-calibrator parameter τ in the power family (12) and compare the optimal $\tau^\star$ of Corollary 2 to the common default $\tau = 1/2$ (the square-root e-calibrator). We use the synthetic setup of Appendix G.5 with 300 repetitions.

Table 5: Band width ε and best guaranteed KPI $X^\star$ at $1-\gamma = 0.75$ for the power e-calibrator with $\tau = \tau^\star$ from Corollary 2 and with $\tau = 1/2$ (square-root e-calibrator).

M	$\tau^\star$	$\varepsilon(\tau^\star)$	$\varepsilon(\tau = 1/2)$	$X^\star(\tau^\star)$	$X^\star(\tau = 1/2)$
5	0.859	0.139	0.161	0.791	0.808
10	0.841	0.132	0.150	0.790	0.793
20	0.816	0.124	0.138	0.789	0.791

Across all values of M, the optimal $\tau^\star$ yields band widths that are 11–15% tighter than the default $\tau = 1/2$ and consistently produces a stronger (smaller) best guaranteed KPI, confirming the sharpness of Corollary 2. Band widths remain within 5% of the oracle value for all $\tau \in [0.75, 0.95]$, suggesting that the choice is robust in a neighborhood of the optimum.

G.8 Decomposition of PS-DME Conservatism

To quantify the sources of conservatism in PS-DME, we compare it to two reference procedures: (*i*) a Bonferroni-style baseline that applies the DKW inequality with the inflated level δ/K to each configuration independently (ignoring post-selection), and (*ii*) a single-DKW baseline that applies (9) at level δ with no multiplicity correction. We use the synthetic setup with $K = 2000$, target $M = 1000$, $\delta = 0.1$, and 200 repetitions.

Table 6: Band widths and empirical FCRs for PS-DME, Bonferroni, and single-DKW baselines on synthetic data with $K = 2000$, $M = 1000$, $\delta = 0.1$.

n	$\varepsilon_{\mathrm{PS}}$	$\varepsilon_{\mathrm{Bonf}}$	$\varepsilon_{\mathrm{DKW1}}$	$\mathrm{FCR}_{\mathrm{PS}}$	$\mathrm{FCR}_{\mathrm{Bonf}}$	$\mathrm{FCR}_{\mathrm{DKW1}}$
20	0.401	0.498	0.274	0.005	0.000	0.102
100	0.179	0.223	0.122	0.000	0.000	0.080
200	0.127	0.157	0.087	0.000	0.000	0.106
500	0.080	0.100	0.055	0.000	0.000	0.078

Table 6 shows that PS-DME is uniformly tighter than Bonferroni by 20–25% while still controlling FCR, confirming the benefit of the e-value route over a union bound. The single-DKW baseline is roughly 30–35% tighter than PS-DME but violates the target FCR, highlighting that the gap between PS-DME and single-DKW is the price paid for post-selection validity. At $n = 100$ and $n = 200$, the empirical FCR of PS-DME is zero across all repetitions, indicating that the DKW tail bound and the e-calibration jointly contribute substantial slack in small samples.

G.9 Berk–Jones vs. DKW on LLM Data

Appendix F introduced a BJ implementation of PS-DME on synthetic data. Here we evaluate the same implementation on the Spider text-to-SQL benchmark of Sec. 5.3 with $K = 31$ (after removing degenerate configurations), $n_{\mathrm{cal}} = 100$, $\delta = 0.1$, and 500 random repetitions.

Table 7: Empirical FCR and best guaranteed KPI at $1-\gamma = 0.9$ for the DKW and BJ implementations of PS-DME on Spider text-to-SQL.

M	$\mathrm{FCR}_{\mathrm{DKW}}$	$\mathrm{FCR}_{\mathrm{BJ}}$	$X^{\star}_{\mathrm{DKW}}$ (p=0.9)	$X^{\star}_{\mathrm{BJ}}$ (p=0.9)
5	0.038 ± 0.008	0.043 ± 0.008	0.66	0.74
10	0.039 ± 0.007	0.055 ± 0.008	0.69	0.75
20	0.080 ± 0.009	0.096 ± 0.009	0.72	0.75

Both implementations control FCR at the target level $\delta = 0.1$. For this benchmark, the DKW variant yields a smaller (better) best guaranteed KPI at $1-\gamma = 0.9$, because the KPI distribution concentrates in the mid-range where the BJ bands are wider; this is consistent with the geometry discussion in Appendix F. Taken together, the results support the choice of DKW as the default calibrator for the main experiments while confirming that BJ remains a valid alternative when tail quantiles are of primary interest.

G.10 BIRD Benchmark

To extend the evaluation beyond Spider, we instantiate PS-DME on the BIRD text-to-SQL benchmark (Li et al., 2023). We restrict attention to $K = 10$ candidate configurations spanning the same single-decode, best-of-n, and iterative-draft-refine families as in Sec. 5.3, with $n = 200$ validation examples, $\delta = 0.1$, and $M = 5$. Results are averaged over 300 random repetitions.

The results in Table 8 confirm the findings of Sec. 5.3: PS-DME satisfies the FCR requirement, while the naive in-sample baseline fails by a factor of nearly four. The gap in guaranteed KPI between the two methods is modest in absolute terms, highlighting that the FCR correction factor $K/(\delta|\mathcal{K}|)$ in (14) is a small price to pay for post-selection validity. A head-to-head comparison with SS-DME on the same benchmark (with SS-DME given n^{eval} equal to 70% of the Corollary 2 crossover threshold, so that PS-DME is provably narrower by design) yields $\mathrm{FCR}_{\mathrm{PS}} = 0.035 \pm 0.004$ versus $\mathrm{FCR}_{\mathrm{SS}} = 0.088 \pm 0.006$, both below δ.

Table 8: Empirical FCR and best guaranteed KPI for PS-DME and the naive in-sample baseline (N-PS-DME) on the BIRD text-to-SQL benchmark.

Method	Empirical FCR	$X^\star$ (p=0.90)	$X^\star$ (p=0.99)
PS-DME	0.087 ± 0.011	0.751	0.772
N-PS-DME	0.372 ± 0.017	0.762	0.781

G.11 LLM-as-Judge Evaluation

The experiments in the main text use execution-based KPIs. Here we evaluate PS-DME when the KPI is produced by an LLM-as-judge, following the recent trend in LLM evaluation and the methodology of Feng et al. (2026). We use the Zephyr-7B-β model from the Mistral family as the judge. To avoid the near-saturated scores that arise when a judge is prompted with the candidate output alone, we provide the judge with both the candidate SQL and the gold reference SQL and ask it to rate semantic equivalence on a scale of 0–10.

We use $K = 5$ decoding strategies, $M = 2$, $n_{\text{cal}} = 100$, $n_{\text{hold}} = 50$, $\delta = 0.1$, and 300 repetitions. Mean judge scores vary across strategies from 8.00 (worst) to 8.68 (best), with a Pearson correlation of $r \in [0.20, 0.28]$ between judge scores and execution-based KPIs, confirming that the judge is informative.

Table 9: PS-DME and SS-DME under an LLM-as-judge KPI on Spider. SS-DME uses $n^{\text{eval}} = 31$ samples (set to 70% of the Corollary 2 crossover threshold).

Method	Empirical FCR	ε	$X^\star$ (p=0.75)
PS-DME	0.003	0.183	4.92 / 10
SS-DME	–	0.220	0.93 / 10

As shown in Table 9, PS-DME controls FCR well below $\delta = 0.1$ and yields a judge-based guaranteed KPI of $4.92/10$ at reliability $1 - \gamma = 0.75$. The SS-DME band at the same reliability level collapses to the trivial bound $0.93/10$ because only 31 samples are available after splitting, illustrating that PS-DME is especially beneficial when the calibration budget is tight.

H Compute Resources

The synthetic data experiments (Sec. 5.2), the O-RAN telecom experiments (Appendix G.2), and the HTTP server log experiment (Appendix G.3) were run on a 2021 MacBook Pro with an Apple M1 Pro chip and 16GB of RAM. LLM inference for the text-to-SQL experiment (Sec. 5.3) was performed on a single NVIDIA A100 GPU (80GB VRAM).

I Code Availability

The code to reproduce all experiments in this paper is available at `https://github.com/amirfar76/PS-DME`.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the three main contributions: the PS-DME framework for post-selection distributional model evaluation, FCR control guarantees via e-values, and sample efficiency improvements over sample splitting. These are all formally proved in Theorem 1 and Lemma 2, and validated experimentally in Sections 5.2 and 5.3.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6, and future directions are outlined in Section 7.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: All theoretical results (Lemma 1, Theorem 1, Lemma 2, Corollary 2) are accompanied by complete proofs in Appendices C, D, and E. Assumptions (i.i.d. samples, arbitrary pre-selection rule, valid e-calibrator) are clearly stated in Section 2 and in each result.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: All experimental setups are described in detail in Section 5, including dataset sizes, model configurations, hyperparameter grids, pre-selection rules, FCR target level, number of repetitions, and split ratios. The LLM model and decoding parameters are fully specified in Section 5.3. The synthetic data generative model is described in closed form in Section 5.2.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: The code to reproduce all experiments is publicly available, as described in Appendix I. All datasets used are publicly available and properly cited. Full experimental details are provided in Section 5 and the appendices.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Full experimental details are provided in Section 5, including data split ratios, hyperparameter grids, the number of candidate configurations K, pre-selection sizes M, calibration dataset sizes n, and the FCR target $\delta = 0.1$. The e-calibrator choice and optimal τ computation are described in Sections 3.2 and 3.3.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: All figures reporting averaged quantities (Figures 3, 4, 9, and 10) include vertical error bars representing the standard error of the mean random dataset splits, as stated in the respective captions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Compute resources are described in Appendix H.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and confirm that the research conforms with it in every respect.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The paper proposes a foundational statistical framework for model evaluation and has no direct path to negative societal impacts. The methodology is designed to improve the reliability of AI and engineering system evaluation, which is broadly beneficial.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release new models, datasets, or other assets with high risk for misuse. The methodology is a statistical evaluation framework, and the datasets used are pre-existing public benchmarks.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All datasets and models used are properly cited. All are publicly available resources used in accordance with their intended research purposes.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not introduce new datasets, models, or other releasable assets.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments use pre-existing datasets or are fully synthetic.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The paper does not involve human subjects research, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs were not used in the development of the core methodology of this paper.

Guidelines:

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