CELEUS: Certifiable and Efficient LLM Evaluation via E-Processes

Zhijian Zhou[1*] **Zesheng Ye**[1*] **Zhaorun Chen**[2] **Bo Li**[2,3] **Feng Liu**[1†]
The University of Melbourne[1] University of Chicago[2]
University of Illinois Urbana-Champaign[3]
fengliu.ml@gmail.com

Abstract

Can we trust evaluation scores to capture an LLM's true real-world performance? *Certifiable* evaluation answers this question by providing a confidence interval (CI) that covers the LLM's true performance with high probability (e.g., 95%). In particular, existing methods sequentially select evaluation samples from a fixed pool and keep updating the CI until certain conditions are satisfied, e.g., the CI width reaches a target precision. However, existing methods are not generally anytime-valid: the claimed coverage (e.g., 95%) may fail when CIs are repeatedly updated and used to decide when to stop, leaving a *gap* between theoretical rigor and practice. This paper bridges this gap by proposing **CELEUS**, a **C**ertifiable framework for **E**fficient **L**LM evaluation, which leverages ***E-processes*** to build anytime-valid CIs. Concretely, we propose *a new signal* that combines two ingredients: (i) **U**ncertainty-guided sampling to select informative samples for evaluation, and (ii) **S**urrogate-assisted approximations for unevaluated samples. We prove that the new signal remains unbiased for the evaluation score conditional on the past, enabling statistically-grounded and anytime-valid e-process CIs. More importantly, the two ingredients reduce estimation variance and help reach the target precision with fewer evaluated samples. We also prove that CIs obtained by CELEUS can shrink at a near-parametric rate up to logarithmic factors and analyze the oracle variance-optimal sampling rule that motivates the empirical uncertainty-guided one. Experiments show that CELEUS reaches the target precision using 54-62% fewer evaluated samples than baselines, while preserving anytime-valid coverage. Code is available at github.com/zyecs/celeus.

1 Introduction

Large language models (LLMs) are increasingly deployed as general-purpose systems for tasks such as question answering [1, 2], coding [3], and interactive assistance [4, 5], making reliable evaluation essential for understanding their capabilities, limitations, and readiness for real-world deployment [6]. To meet this need, recent benchmark efforts [7, 8] have sought to assess LLMs systematically and at scale, rather than through a handful of narrow evaluation sets. However, scale alone is not enough for reliable deployment: simply collecting more evaluation samples and reporting an estimated performance score does not certify whether the estimate is precise enough for decision-making, leaving no statistically grounded criterion for deciding when enough samples have been evaluated.

Certifiable evaluation addresses the above issue by sequentially selecting evaluation samples from a fixed pool and updating CIs [9]. The updating process will end until certain conditions are satisfied, e.g., the CI width reaches a target precision. A common line of evaluation with CIs considers static settings, where the evaluation samples are fixed in advance and a one-shot CI is reported for model

*Equal contribution. †Corresponding author.

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

performance [10, 11]. These guarantees apply to the final interval at a pre-specified endpoint, rather than to the evolving sequence of intervals monitored during evaluation, and therefore do not directly support stopping once the CI width reaches a target precision. Another line of work provide considers a different setting in which data arrive sequentially and CIs are updated as new samples become available [12–14]. In these settings, the evaluator has no access to future samples, and these methods do not support the sample selection required by certifiable evaluation with a fixed pool.

Within the fixed pool setting, existing approaches to certifiable evaluation are not generally anytime-valid: their claimed coverage (e.g., 95%) may fail to hold when CIs are repeatedly updated and used to decide when to stop, leaving a *gap* between theoretical rigor and practice. For example, the state-of-the-art method Cer-Eval provides certification for a fixed partition, while its practical algorithm learns and updates the partition using evaluated data [9]. Using the same data to choose the partition and construct CIs introduces a dependence that is not addressed by the fixed-partition analysis.

Key Idea. This paper bridges this gap by proposing **CELEUS**, a certifiable framework for efficient LLM evaluation, which leverages e-*processes* to build anytime-valid CIs. Given a fixed data pool, CELEUS combines uncertainty-guided sampling, which prioritizes informative samples for evaluation, with surrogate-assisted approximations for unevaluated samples. These strategies reduce the number of costly evaluations needed to reach a target precision. Under such non-uniform selection, however, a naive signal based on the observed risk can be biased for the target performance. This creates the key *obstacle* for e-process construction: the signal must remain conditionally unbiased for the evaluation risk as evaluation proceeds.

Technical Contributions. To cope with this obstacle, CELEUS builds *inferential signals* for sampling without replacement from a fixed evaluation pool. At each round, the signal combines the observed losses of previously evaluated samples with surrogate scores for the remaining samples. After a new sample is selected and evaluated, it adds an inverse-probability-weighted correction for that sample's surrogate residual, using its actual selection probability. We prove that the resulting signal is conditionally unbiased for the finite-pool risk under history-dependent proposal distributions satisfying a probability floor (see Thm. 3, Sec. 4). This conditional unbiasedness enables finite-sample, anytime-valid e-process CIs through test inversion [15]. We then lift these finite-pool CIs to population-risk CIs through a correction that depends on the full pool size.

We also justify the effectivenss of CELEUS from two theoretical perspectives: 1) We characterize how surrogate scores and sample selection can reduce estimation variance and help reach the target precision with fewer evaluated samples, with an oracle variance-optimal sampling rule motivating our empirical uncertainty-guided sampling strategy (see Thms. 4 and 5); 2) We prove that CELEUS's CIs are anytime-valid, covering target risk with probability at least $1-\alpha$ at any stopping time, and that their widths shrink at a near-parametric rate up to logarithmic factors (see Thm. 6).

Empirical Contributions. Extensive experiments validate both the statistical validity and efficiency of CELEUS. In a diagnostic setting where CER-EVAL's partitioning can lose coverage, CELEUS preserves nominal anytime-valid coverage. On real LLM evaluation tasks, CELEUS reaches the target precision using 54–62% fewer evaluated samples than baselines. Moreover, compared with the no-surrogate version of CELEUS, leveraging pre-trained surrogate models with only 7–8B parameters, without any task-specific training or fine-tuning, further improves evaluation efficiency by 10–23% when evaluating larger models, including 67–72B dense models and an 8×7B MoE model.

2 Certifiable and Efficient Evaluation: Setup and Overview

In this section, we introduce the evaluation setup and give an overview of certifiable and efficient evaluation based on e-processes. We characterize a task by a joint distribution $\mathbb{P}_{\boldsymbol{x},y}$ over input-label pairs $(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in \mathcal{X}$ denotes an input and $y \in \mathcal{Y}$ denotes the corresponding ground-truth label. Given an LLM $f : \mathcal{X} \to \mathcal{Y}$ fixed before the evaluation pool is drawn and a bounded risk function $\ell : \mathcal{Y} \times \mathcal{Y} \to [L, U]$ for some constants $L, U \in \mathbb{R}$, we aim to evaluate the LLM's population risk:

$$R(f; \mathbb{P}_{\boldsymbol{x},y}) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathbb{P}_{\boldsymbol{x},y}}\big[\ell(f(\boldsymbol{x}), y)\big],$$

where, for brevity, $R = R(f; \mathbb{P}_{\boldsymbol{x},y})$.

In practice, the distribution $\mathbb{P}_{\boldsymbol{x},y}$ is unknown. Instead, we are given an evaluation pool $D_N = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^N$, whose input pool is $X_N = \{\boldsymbol{x}_j\}_{j=1}^N$. For population-risk results, these input-label

pairs are *i.i.d.* draws from $\mathbb{P}_{\boldsymbol{x},y}$.[2] The reference labels or evaluation judgments are initially withheld, and each sample has a fixed target loss throughout evaluation. *Certifiable evaluation* sequentially evaluates samples and stops once the CI reaches a target precision. Because evaluating a sample can be costly, especially when human judgments are required, *efficient evaluation* aims to reach this precision with as few evaluated samples as possible.

A natural idea is to *select informative samples for evaluation*, which has been extensively studied in active testing [18, 19]. Rather than evaluating samples uniformly at random, active testing assigns larger selection probabilities to samples expected to be more informative, with the aim of reaching the target precision using fewer evaluations. At each round t, the sampling policy selects an unevaluated sample i_t from the pool and queries the information required to determine its target loss $\ell(f(\boldsymbol{x}_{i_t}), y_{i_t})$. This query may reveal a reference label, obtain a human judgment, or invoke a model judge. We call a sample *evaluated* once its target loss becomes available.[3] For offline labeled benchmarks, we simulate this process by revealing labels only for selected samples. After round t, the evaluated data are $\{(\boldsymbol{x}_{i_m}, y_{i_m})\}_{m=1}^{t}$, and the remaining index set is

$$\mathcal{J}_t = [N] \setminus \{i_1, \dots, i_t\}, \qquad [N] = \{1, \dots, N\}, \qquad i_{1:t} = (i_1, \dots, i_t).$$

Another effective strategy is to use *surrogate-assisted approximations for unevaluated samples* [10, 11]. Specifically, a surrogate score $\tilde{\ell}_f(\boldsymbol{x}_j) \in [L, U]$ approximates the unknown target loss $\ell(f(\boldsymbol{x}_j), y_j)$ without revealing y_j. For example, the score can be produced by a surrogate model that predicts the target loss or by a label-free proxy constructed from available model outputs. The surrogate scores are computed before evaluation and kept fixed throughout the procedure. They allow us to extract information from unevaluated samples: when the scores are close to the target losses, we can obtain an accurate estimate with fewer evaluated samples.

For the analysis, let $\mathcal{P}_0$ collect the label-free information computed before evaluation, including model outputs and surrogate scores. We write

$$\mathcal{F}_t = \sigma(D_N, \mathcal{P}_0, i_{1:t}), \qquad 0 \le t \le N, \tag{1}$$

where $\sigma(\cdot)$ denotes the information determined by the listed quantities and $i_{1:0}$ is empty. Thus, $\mathcal{F}_t$ contains the complete pool, fixed preprocessing information, and the first t selections. Conditioning on the complete pool isolates the randomness from sample selection; the evaluator still sees a label only after selecting its index. The observed information and sampling law are formalized in App. C.1.

Given the selected evaluation samples and surrogate-assisted approximations, the key question is how to construct CIs that remain anytime-valid as evaluation proceeds. Standard static evaluation reports a CI for a fixed set of evaluated samples [10, 11] and can lose validity when the CI is repeatedly updated and used to decide when to stop. Sequential concentration inequalities and e-processes provide tools for inference under continuous monitoring [9, 20, 15, 21]. Applying these tools to our setting requires a signal with the correct conditional mean under non-uniform sampling without replacement. We use e-processes and recall their definition and anytime-valid implication below.

Definition 1. *[22] Let $\mathcal{H}_0$ be a null hypothesis, and let $\{\mathcal{F}_t\}_{0\le t\le N}$ describe the information used for conditioning. A nonnegative process $\{\mathcal{K}_t\}_{0\le t\le N}$ is an e-*process *for $\mathcal{H}_0$ if $\mathcal{K}_0 = 1$, each $\mathcal{K}_t$ is determined by $\mathcal{F}_t$, and $\mathbb{E}_P[\mathcal{K}_\tau] \le 1$ for every $P \in \mathcal{H}_0$ and every stopping time $\tau \le N$ with respect to $\{\mathcal{F}_t\}$. A stopping time is a decision to stop based only on information available by that round. A sufficient construction is a nonnegative supermartingale satisfying $\mathbb{E}_P[\mathcal{K}_t \mid \mathcal{F}_{t-1}] \le \mathcal{K}_{t-1}$ at each round under every $P \in \mathcal{H}_0$. Consequently, for any $\alpha \in (0, 1)$, it satisfies the anytime-valid control*

$$\mathbb{P}_P\left(\max_{0\le t\le N} \mathcal{K}_t \ge \frac{1}{\alpha}\right) \le \alpha, \qquad \forall P \in \mathcal{H}_0.$$

Thus, rejecting $\mathcal{H}_0$ whenever $\mathcal{K}_t \ge 1/\alpha$ gives an anytime-valid test: if $\mathcal{H}_0$ is true, the probability of falsely rejecting it at any round is at most α.

[2] The i.i.d. assumption is used for the population variance formula, consistency, and the correction from finite-pool to population risk. The finite-pool results condition on the complete pool and do not require i.i.d. sampling; for non-i.i.d. pools, population-risk guarantees require an appropriate distribution-dependent generalization bound [16, 17].

[3] Our efficiency objective is to reduce the number of ground-truth labels needed to reach the target precision, as acquiring these labels is often costly in practice. Our sampling rules use target-model outputs computed over the full pool, together with surrogate-model outputs when applicable. These preprocessing costs are included in the end-to-end cost decomposition in App. J, which also reports online timing with precomputed outputs and scores.

To turn e-processes into CIs, we use test inversion for the data-pool risk $R_N = \frac{1}{N}\sum_{j=1}^{N} \ell(f(\boldsymbol{x}_j), y_j)$. Conditional on $\mathcal{F}_0$, this risk is fixed, and the remaining randomness comes from sample selection. For each candidate value v, we test $\mathcal{H}_0(v) : R_N = v$ by constructing an e-process $\{\mathcal{K}_t(v)\}_{0\le t\le N}$. Intuitively, $\mathcal{K}_t(v)$ measures how strongly the information collected up to round t contradicts v. If $\mathcal{K}_t(v) \ge 1/\alpha$, then v is rejected at that round. The CI keeps the remaining candidate values:

$$\mathrm{CI}_t = \{v : \mathcal{K}_t(v) < 1/\alpha\}.$$

Since R_N corresponds to the true null hypothesis $\mathcal{H}_0(R_N)$, non-coverage can occur only if this true null is falsely rejected:

$$\mathbb{P}\left(\exists 0 \le t \le N : R_N \notin \mathrm{CI}_t\right) = \mathbb{P}\left(\exists 0 \le t \le N : \mathcal{K}_t(R_N) \ge \frac{1}{\alpha}\right) \le \alpha.$$

The resulting CIs therefore cover R_N simultaneously over all rounds with probability at least $1-\alpha$, allowing evaluation to stop once the target precision is reached.

For certifiable and efficient LLM evaluation, the key technical step is to construct valid inferential signals from the evaluated samples and surrogate approximations for unevaluated samples. These signals capture the new information obtained at each round and are designed to be conditionally unbiased for the data-pool risk, which allows us to construct an e-process satisfying the supermartingale condition above [21, Proposition 2]. Their variance also provides a natural criterion for sample selection, guiding the evaluator toward samples with informative surrogate-correction residuals that help the CI reach the target precision with fewer evaluations. We then lift this CI to the population risk R using a finite-pool-to-population correction.

3 The CELEUS Evaluation Method

We now instantiate the overview above by giving the concrete construction of CELEUS, a sequential evaluation procedure that yields *anytime-valid and shrinking CIs* for R, defined as:

Definition 2. *Fix a significance level $\alpha \in (0,1)$. For each pool size N, a sequence of random sets $\{C_{t,N}\}_{1\le t\le N}$ is called a sequence of $(1-\alpha)$ anytime-valid CIs for R if*

$$\Pr\left(\exists\, 1 \le t \le N : R \notin C_{t,N}\right) \le \alpha.$$

The family $\{C_{t,N} : 1 \le t \le N,\ N \ge 1\}$ is additionally called shrinking if, for every sequence $N \to \infty$ and $1 \le t_N \le N$ satisfying $t_N \to \infty$, $\mathrm{width}(C_{t_N,N}) \xrightarrow{\mathbb{P}} 0$.

The shrinkage condition concerns joint growth of the pool size N and the number of evaluated samples t. For a fixed finite pool, evaluating more samples does not reduce the population correction $\Delta_N(\alpha_2)$, so the pool may be exhausted before the requested precision is reached.

Constructing such CIs for R involves two conceptually different sources of randomness. First, we only observe risks for a selected subset of the evaluation pool, which creates randomness in estimating the *pool risk* R_N. Second, even if the entire pool were evaluated, R_N itself may differ from R because the pool is a finite i.i.d. sample from the population distribution. To handle these two sources of randomness separately, we split the overall error budget α into $\alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1 + \alpha_2 = \alpha$. Here, α_1 controls anytime-valid inference for R_N through test inversion of e-processes, while α_2 controls the finite-pool-to-population correction from R_N to R.

Anytime-valid CIs for R_N at level α_1 via e-process. At the beginning of round t, after the first $t-1$ evaluations, the evaluator has access to two sources of information: the observed risks on the evaluated samples, $\{\ell(f(\boldsymbol{x}_{i_m}), y_{i_m})\}_{m=1}^{t-1}$, and the surrogate risk scores on the remaining unevaluated samples, $\{\tilde{\ell}_f(\boldsymbol{x}_j)\}_{j\in\mathcal{J}_{t-1}}$. Based on this information, the sampling policy assigns selection probabilities to the remaining unevaluated samples; we denote by $q_t(i_t \mid i_{1:t-1}, X_N)$ the probability of selecting index i_t from $\mathcal{J}_{t-1}$ at round t. This notation also includes dependence on the fixed preprocessing information and the labels revealed in previous rounds, which we suppress for brevity. After selecting i_t according to this selection probability and observing its risk, we first derive the per-round *inferential signal* for constructing e-processes as

$$\hat{S}_t = \frac{\sum_{m=1}^{t-1} \ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) + \sum_{j\in\mathcal{J}_{t-1}} \tilde{\ell}_f(\boldsymbol{x}_j)}{N} + \frac{\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{N q_t(i_t \mid i_{1:t-1}, X_N)}. \tag{2}$$

The first term in $\hat{S}_t$ is a surrogate-completed estimate of the data-pool risk: it uses the observed risks for the previously evaluated samples and fills in the remaining unevaluated samples with their surrogate risk scores. The second term is an inverse-probability-weighted correction: the selected sample reveals the surrogate residual $\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})$, and the factor $1/q_t(i_t \mid i_{1:t-1}, X_N)$ turns this residual into an unbiased estimate of the normalized total residual over the remaining pool, as in a conditional Horvitz–Thompson-type estimator [23]. Hence, conditional on the complete pool, preprocessing information, and past selections, $\hat{S}_t$ is unbiased for the finite-pool risk R_N, as shown in Thm. 3.

In Rem. 1, we propose *uncertainty*-guided sampling strategies that aim to reduce the remaining statistical *uncertainty* in the per-round signal (i.e., the conditional variance of signal $\hat{S}_t$). This quantity appears explicitly in the width analysis of the e-process confidence interval: lower conditional variance yields signals that concentrate more tightly around the target risk, thereby providing stronger sequential evidence against incorrect candidate values and helping the CI reach the target precision with fewer evaluated samples. In particular, the *uncertainty*-guided sampling strategies assign larger probabilities to samples with larger discrepancies between the observed risk and the surrogate risk. Formal details are provided in Thms. 4 and 5 in Sec. 4.

We now construct CIs for R_N from the sequentially updated signals $\{\hat{S}_t\}_{1\le t\le N}$ using e-processes. Following standard e-process constructions [21], we map the signal to $[0,1]$ by setting $\bar{S}_t = (\hat{S}_t - a_t)/(b_t - a_t)$, where a_t and b_t are predictable[4] lower and upper bounds of $\hat{S}_t$ given the information available before round t. Accordingly, for each candidate value υ of R_N, we define its scaled counterpart as $\bar{\upsilon}_t = (\upsilon - a_t)/(b_t - a_t)$. Since this is an affine transformation and $\hat{S}_t$ is conditionally unbiased for R_N under the null hypothesis $\mathcal{H}_0(\upsilon) : R_N = \upsilon$, we have $\mathbb{E}[\bar{S}_t \mid \mathcal{F}_{t-1}] = \bar{\upsilon}_t$. This leads to a centered sequential signal: for each candidate υ, the sequence $\bar{S}_t - \bar{\upsilon}_t$ has conditional mean zero for all t under $\mathcal{H}_0(\upsilon)$. This ensures that, when the candidate value is correct, the signal has no systematic drift against it, making it a valid sequential signal for testing $\mathcal{H}_0(\upsilon)$ on scaled space.

To implement a sequential test for $\mathcal{H}_0(\upsilon)$ at each candidate value υ, we use a *hedged capital process* [21, Section 4.4]. For any $\theta \in (0,1)$ and candidate υ, define

$$\mathcal{K}_t^{\pm}(\upsilon) = \theta \prod_{i=1}^{t} \left(1 + \lambda_i^{+}(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right)\right) + (1-\theta) \prod_{i=1}^{t} \left(1 - \lambda_i^{-}(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right)\right). \tag{3}$$

Here, $\lambda_i^{+}(\upsilon)$ and $\lambda_i^{-}(\upsilon)$ are predictable fractions in $[0, 1/\bar{\upsilon}_i)$ and $[0, 1/(1-\bar{\upsilon}_i))$, respectively; see Remark 3 for implementation details. The resulting $(1-\alpha_1)$ confidence set for R_N at time t is[5]

$$\mathcal{M}_t^{\pm} = \bigcap_{i \le t} \left\{\upsilon \in [L,U] : \mathcal{K}_i^{\pm}(\upsilon) < 1/\alpha_1\right\}. \tag{4}$$

Finite-pool-to-population correction at level α_2. We next transfer the CI guarantee from the pool risk R_N to the population risk R by controlling the generalization gap between them at level α_2. Specifically, by Hoeffding's inequality [24], with probability at least $1-\alpha_2$,

$$|R_N - R| \le \Delta_N(\alpha_2) = (U-L)\sqrt{\frac{\log(2/\alpha_2)}{2N}}.$$

Notably, this correction depends on the full data-pool size N, rather than the number of evaluated samples, and can therefore be small or even negligible for large pools.

Considering both (i) $R_N \in \mathcal{M}_t^{\pm}$ for all $1 \le t \le N$ and (ii) $|R_N - R| \le \Delta_N(\alpha_2)$ hold, we have for all t that $R \in \mathcal{M}_t^{\pm} \oplus [-\Delta_N(\alpha_2), \Delta_N(\alpha_2)]$, where $\oplus$ denotes the Minkowski sum[6]. Therefore, we define the CI for R at time t as

$$C_t = \left(\mathcal{M}_t^{\pm} \oplus [-\Delta_N(\alpha_2), \Delta_N(\alpha_2)]\right) \cap [L,U].$$

Performance Estimation at Stopping. Given a target precision $\epsilon \in (0,1)$, the procedure stops when $\mathrm{width}(C_t) \le \epsilon$ or the pool is exhausted. With exact test inversion or the conservative numerical construction in App. M.1, and under the stated assumptions, the returned CI covers R with probability

[4] A sequence $(s_t)_{t\ge1}$ is predictable if s_t can only depend on the information available up to time $t-1$. The bounds a_t and b_t can also be chosen as constants, as long as they satisfy $a_t \le \hat{S}_t \le b_t$. See Rem. 4 for further discussion.

[5] The experiments use grid-point inversion as described in App. B. App. M.1 gives a conservative alternative that returns an interval containing the exact confidence set and preserves its coverage guarantee.

[6] For sets $A, B \subseteq \mathbb{R}$, the Minkowski sum is $A \oplus B = \{a + b : a \in A,\ b \in B\}$.

Algorithm 1 Overall procedure of CELEUS

Require: Input pool $X_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, surrogate scores $\{\tilde{\ell}_f(\boldsymbol{x}_j)\}_{j=1}^N$, error split $\alpha_1 + \alpha_2 = \alpha$, parameters β and θ, and target precision ϵ
Ensure: Stopping time τ, CI C_τ, and performance estimate $\hat{R}_\tau$
$\mathcal{J}_0 \leftarrow \{1, \ldots, N\}$ and $\tau \leftarrow N$
for $t = 1, \ldots, N$ **do**
  Compute an acquisition score for every $j \in \mathcal{J}_{t-1}$ using only information available before round t
  Convert the scores into the sampling probabilities $q_t(j \mid i_{1:t-1}, X_N) \geq \beta/(N-t+1)$ (Assumption 2)
  Draw $i_t \sim q_t(\cdot \mid i_{1:t-1}, X_N)$ over $\mathcal{J}_{t-1}$, evaluate sample i_t, and observe $\ell(f(\boldsymbol{x}_{i_t}), y_{i_t})$
  Compute $\hat{S}_t$ in Eqn. (2) and $\bar{S}_t = (\hat{S}_t - a_t)/(b_t - a_t)$
  Update $\mathcal{K}_t^{\pm}(\upsilon)$ in Eqn. (3) for every candidate $\upsilon \in [L, U]$, and $\mathcal{M}_t^{\pm}$ in Eqn. (4)
  $C_t \leftarrow \left(\mathcal{M}_t^{\pm} \oplus [-\Delta_N(\alpha_2), \Delta_N(\alpha_2)]\right) \cap [L, U]$ and $\mathcal{J}_t \leftarrow \mathcal{J}_{t-1} \setminus \{i_t\}$
  if $\mathrm{width}(C_t) \leq \epsilon$ **then**
    $\tau \leftarrow t$ and **break**
  end if
end for
return $(\tau, C_\tau, \hat{R}_\tau)$, with $\hat{R}_\tau$ in Eqn. (5)

at least $1 - \alpha$ under either stopping rule. If the pool is exhausted before the width criterion is met, the procedure reports that the target precision was not reached. Alongside the CI, CELEUS reports the following weighted surrogate-residual estimate (see Appendix A.5):

$$\hat{R}_t = \frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f\left(\boldsymbol{x}_n\right) + \frac{1}{t}\sum_{m=1}^{t} u_{m,t}\left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right), \tag{5}$$

For $1 \leq m \leq t < N$, the weight $u_{m,t}$, motivated by the LURE estimator [25], adjusts the surrogate residual for sequential, non-uniform sample selection:

$$u_{m,t} = 1 + \frac{N-t}{N-m}\left(\frac{1}{(N-m+1)q_m(i_m \mid i_{1:m-1}, X_N)} - 1\right).$$

The first term in $\hat{R}_t$ provides a low-cost approximation to the evaluation performance by averaging the surrogate risk scores $\tilde{\ell}_f(\boldsymbol{x}_n)$ over all samples. The second term corrects this approximation using the observed surrogate residuals from the evaluated samples, yielding a fixed-time unbiased estimator. At pool exhaustion, set $u_{m,N} = 1$ for all m, so that $\hat{R}_N = R_N$; the preceding weight formula is used only for $t < N$. Fixed-time unbiasedness does not imply unbiasedness at a data-dependent stopping time. The CI, rather than this scalar estimate, supplies the stopping-time coverage guarantee.

Overall procedure. Alg. 1 summarizes the procedure of CELEUS. The complete end-to-end implementation, including acquisition-score normalization, probability-floor mixing, sequential e-process updates, numerical test inversion, and precision-based stopping, is given in Alg. 2 (App. B).

4 Theoretical Analysis

Proofs and assumptions are deferred to App. C. Assumption 1 only requires bounded evaluation scores, which can usually be ensured by normalization or clipping. The population-risk results also use the i.i.d. input-label pairs and fixed target model specified in the evaluation setup. The remaining assumptions are algorithmic conditions satisfied by CELEUS.

Risk estimator. We first establish the fixed-time unbiasedness and consistency of the risk estimator.

Theorem 1. *Under Assumptions 1-2, for each fixed $1 \leq t < N$, $\hat{R}_t$ has $\mathbb{E}[\hat{R}_t] = R$ with $\mathrm{Var}[\hat{R}_t] = \mathrm{Var}(\ell(f(\boldsymbol{x}), y))/N + \mathbb{E}\left[\mathrm{Var}\left[\sum_{m=1}^{t} w_m \psi_m / t \mid \mathcal{F}_0\right]\right]$, where w_m and ψ_m are defined in Eqn. (10) and Eqn. (11). At $t = N$, $\hat{R}_N = R_N$ has expectation R and variance $\mathrm{Var}(\ell(f(\boldsymbol{x}), y))/N$.*

The next theorem further shows that the estimator is consistent.

Theorem 2. *Under Assumptions 1-2, for any sequence $\{t_N\}_{N \geq 1}$ satisfying $1 \leq t_N \leq N$, $N \to \infty$, and $(N - t_N + 1)/N \to 0$, we have $\lim_{N\to\infty} \mathbb{E}\left[\left(\hat{R}_{t_N} - R\right)^2\right] = 0$, i.e., $\hat{R}_{t_N} \to R$ in L^2.*

Conditionally unbiased signal. We next study the signals used to construct e-processes. The key property is conditional unbiasedness, which enables anytime-valid CIs under sample selection.

Theorem 3. *Under Assumption 1-2, the $\hat{S}_t$ satisfies $\mathbb{E}[\hat{S}_t \mid \mathcal{F}_{t-1}] = R_N$ with $\mathrm{Var}[\hat{S}_t \mid \mathcal{F}_{t-1}] =$*

$$\sum_{j \in \mathcal{J}_{t-1}} \frac{(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j))^2}{N^2 q_t(j \mid i_{1:t-1}, X_N)} - \left(\frac{\sum_{j \in \mathcal{J}_{t-1}} (\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j))}{N} \right)^2 .$$

Variance-driven efficiency. The variance expression above also indicates how the sampling distribution should be designed. Since the second term is fixed, q_t only affects the first term, where samples with larger discrepancies between the observed risk and the surrogate risk score contribute more to the conditional variance and should therefore be sampled with higher probability. The following theorem formalizes this intuition by characterizing the variance-optimal sampling distribution.

Theorem 4. *Under Assumptions 1-2, fix a round t. If not all $\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j), j \in \mathcal{J}_{t-1}$, are zero, the following proposal distribution minimizes the conditional variances of $\hat{R}_t$ and $\hat{S}_t$:* [7]

$$q_t^*(j \mid i_{1:t-1}, D_N) = \max \left\{ \beta/(N-t+1), \left| \ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j) \right| / \eta_t \right\} .$$

The oracle uses the complete labeled pool D_N. Here, $\beta \in (0, 1]$ controls the minimum sampling probability as specified in Assumption 2, and $\eta_t > 0$ is chosen so that $\sum_{j \in \mathcal{J}_{t-1}} q_t^*(j \mid i_{1:t-1}, D_N) = 1$. In contrast, if all $\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)$ are zero, then every proposal distribution q_t satisfying Assumption 2 is optimal.

We next translate this variance reduction into a guarantee on efficiency. The following theorem shows that the CIs provide anytime-valid coverage for R_N and shrink as more samples are evaluated.

Theorem 5. *Fix $\alpha_1 \in (0, 1)$ and $c \in (0, 1)$. Under Assumptions 1-4, for every $N \in \mathbb{Z}_+$, the sequence $\left(\mathcal{M}_t^{\pm}\right)_{1 \le t \le N}$, defined in Eqn. (4), is a $(1-\alpha_1)$ anytime-valid confidence sequence for R_N. Moreover, for any $\omega \in (0, 1)$, there exists a constant $G_6 > 0$ such that, with probability at least $1-\omega$, the following bound holds simultaneously for all $t_0 \le t \le N$:*

$$\mathrm{width}\left(\mathcal{M}_t^{\pm}\right) \le G_6 \sqrt{\frac{\log t \log\log t}{t} \left(\log\log t + \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \right)} .$$

Here, t_0 is defined in Eqn. (19). Writing $\mathcal{M}_{t,N}^{\pm}$ to make the dependence on N explicit, for any integer sequence $\{t_N\}_{N \ge 1}$ satisfying $1 \le t_N \le N$ and $t_N \to \infty$ as $N \to \infty$,

$$\mathrm{width}\left(\mathcal{M}_{t_N,N}^{\pm}\right) = O_{\mathbb{P}}\left(\log\log t_N \sqrt{\frac{\log t_N}{t_N}} \right) ,$$

and hence $\mathrm{width}\left(\mathcal{M}_{t_N,N}^{\pm}\right) \xrightarrow{\mathbb{P}} 0$ as $N \to \infty$. Thus, the confidence sequences are shrinking in the joint asymptotic regime $N \to \infty$, $t = t_N \to \infty$, and $t_N \le N$.

The bound shows that the CI width is driven by the cumulative conditional variance of the per-round signals. By Theorem 3, surrogate and acquisition quality affect this variance through different components: the surrogate score determines the residuals $\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)$, whereas the acquisition proxy determines how closely the resulting proposal distribution approaches the constrained variance-optimal proposal in Theorem 4. Smaller surrogate residuals and a smaller mismatch from the variance-optimal proposal tighten the width bound and can reduce the number of evaluations required to reach the target precision. These approximation choices affect efficiency but not anytime-valid coverage under the assumptions. A detailed theoretical and empirical analysis is provided in App. A.4.

Coverage for population risk and shrinkage rate. We finally lift the guarantee to population risk.

[7] This is stepwise optimality in round t. Optimizing the cumulative conditional variance term in Thm. 5 is a trajectory-level adaptive stochastic optimization problem; related adaptive optimization problems can be NP-hard [26].

Table 1: Empirical miscoverage at the stopping time when the CI width reaches $\epsilon = 0.05$ in the synthetic setting of App. D. Each entry reports the mean miscoverage rate over 50 independent runs, with Wilson 95% intervals shown in grey brackets. CER-EVAL exceeds the nominal miscoverage level $\alpha = 0.05$, indicating a gap between its theoretical guarantees and its proposed adaptive-partition algorithm. In contrast, all e-process-based methods remain within the nominal level.

CER-EVAL	EVALUE	W/O SURR	CELEUS	ORACLE
0.161	0.030	0.030	0.021	0.028
[0.158, 0.164]	[0.029, 0.032]	[0.029, 0.032]	[0.020, 0.023]	[0.027, 0.030]

Table 2: Median number of evaluated samples to reach $\epsilon = 0.05$, across the six surrogate-target pairs and 50 independent runs. We mark the relative savings of CELEUS vs. EVALUE, and the **top-performer** in bold. The ORACLE is included for reference only.

	SST-2 (N=68,221)	**MMLU** (N=115,700)	**AG News** (N=127,600)
EVALUE	6,889	8,144	15,336
W/O SURR	3,844 $_{(-44\%)}$	5,110 $_{(-37\%)}$	8,298 $_{(-46\%)}$
CELEUS	**3,156** $_{(-54\%)}$	**3,242** $_{(-60\%)}$	**5,840** $_{(-62\%)}$
ORACLE	2,698	1,923	4,238

Theorem 6. *Fix a significance level $\alpha \in (0,1)$ and choose any split $\alpha_1, \alpha_2 \in (0,\alpha)$ with $\alpha_1+\alpha_2 = \alpha$. Under the population sampling setup and Assumptions 1-4, for every $N \in \mathbb{Z}_+$, the sequence $(C_t)_{1\le t\le N}$ is a $(1-\alpha)$ anytime-valid confidence sequence for the population risk R. Moreover, for all $1 \le t \le N$,*

$$\operatorname{width}(C_t) \le \operatorname{width}\left(\mathcal{M}_t^{\pm}\right) + 2(U-L)\sqrt{\frac{\log(2/\alpha_2)}{2N}}.$$

Writing $C_{t,N}$ to make the dependence on N explicit, for any integer sequence $\{t_N\}_{N\ge1}$ satisfying $1 \le t_N \le N$ and $t_N \to \infty$ as $N \to \infty$,

$$\operatorname{width}(C_{t_N,N}) = O_{\mathbb{P}}\left(\log\log t_N \sqrt{\frac{\log t_N}{t_N}}\right) + O\left(\sqrt{\frac{1}{N}}\right),$$

and hence $\operatorname{width}(C_{t_N,N}) \xrightarrow{\mathbb{P}} 0$ *as $N \to \infty$. Thus, the confidence intervals are shrinking in the joint asymptotic regime $N \to \infty$, $t = t_N \to \infty$, and $t_N \le N$.*

Together, these results explain why CELEUS provides certifiable and efficient evaluation. Thms. 1-2 show that the final performance estimator targets the true evaluation risk and is consistent. Thm. 3 ensures the conditional unbiasedness of signal $\hat{S}_t$ for e-processes, while Thm. 4 shows how surrogate scores and sampling probabilities reduce variance, thereby tightening the CI in Thm. 5. Finally, we show in Thm. 6 that the resulting anytime-valid CIs shrink with t and N, yielding tight certified bounds under the stated sampling conditions; we further provide empirical validation in our experiments.

5 Experiments

In this section, we evaluate CELEUS on real LLM evaluation tasks. Coverage at the reported stopping rule is an empirical diagnostic; it does not establish simultaneous coverage over all rounds or certify the grid approximation. Sec. 5.1 describes the experimental setup, Sec. 5.2 reports the main results, and Sec. 5.3 provides additional analyses.

5.1 Setup

Datasets. Following Berrada et al. [18], we evaluate LLMs on downstream tasks. We use three benchmarks with different levels of task difficulty, label spaces, and evaluation-pool sizes: Stanford Sentiment Treebank 2 (SST-2) [27] for binary sentiment classification, AG News Topic Classification (AG News) [28] for multi-class topic classification, and Massive Multitask Language Understanding (MMLU) [1], a broad multiple-choice benchmark across a wide range of academic subjects. For each benchmark, all available samples are treated as the evaluation pool D_N. Table 5 summarizes the pool sizes and model pairs used in the main classification experiments.

Surrogate-target pairs. We instantiate *six* surrogate-target pairs across model-family and generation axes. Specifically, Llama-3-8B is used as the surrogate for four targets: Llama-3-70B (same family, same generation), DeepSeek-67B, Mixtral-8×7B, and Qwen-2.5-72B (all cross-family, same generation). We further use Llama-2-7B as the surrogate for Llama-3-70B (same family, cross-generation) and DeepSeek-67B (cross-family, cross-generation). In all pairs, the surrogate model is substantially smaller without any task-specific training or fine-tuning.

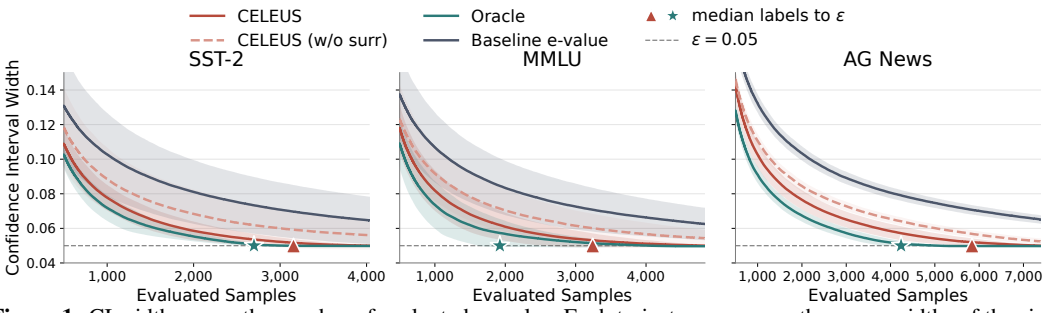

Figure 1: CI width versus the number of evaluated samples. Each trajectory averages the mean widths of the six surrogate–target pairs, with each pair averaged over 50 seeds. Shaded bands show ± 1 standard deviation across these pair means. The dashed line marks $\epsilon = 0.05$.

Baselines. We first use a diagnostic case to examine CER-EVAL [9], showing that its adaptive partitioning can fail to maintain anytime-valid coverage; and in the following comparisons, we focus on methods designed to be anytime-valid: ① EVALUE applies the standard e-processes for bounded mean estimation [21] under uniform sampling without a surrogate. ② CELEUS (W/O SURR) sets $\tilde{\ell}_f \equiv 0$ and samples uniformly without replacement from the remaining pool, with $q_t(j) = 1/(N-t+1)$. It retains the inferential signal and CI construction of CELEUS, but uses neither surrogate assistance nor uncertainty-guided sampling. ③ ORACLE uses the true surrogate-residual magnitudes as sample-selection scores, giving the oracle sampling rule in Remark 6. It provides an infeasible reference for one-step variance reduction; it is not claimed to minimize the stopping time.

Evaluation metric and sample selection. We report the evaluation results with 0-1 risk in the main text, with additional results on cross-entropy risk provided in App. G. For CELEUS, we use the uncertainty-guided sampling strategies from Rems. 1 and 2.

Hyperparameters. We set the overall significance level to $\alpha = 0.05$, which controls the miscoverage probability of the CIs, and split it evenly as $\alpha_1 = \alpha_2 = 0.025$ between anytime-valid inference for R_N and the finite-pool-to-population correction. Following previous e-process [21], we use predictable fractions in Rem. 3 with $c = 0.5$, set $\theta = 0.5$ in Eqn. (3), and use the scaling strategy in Rem. 4. For sampling, we set $\beta = 0.4$ to control the lower bound of sampling probabilities, with sensitivity analyzed in App. I. We stop the evaluation at the first time t such that $\mathrm{width}(C_t) \leq \epsilon = 0.05$.

5.2 Main Results

RQ1: Empirical coverage at stopping. For each method, we repeat the experiment 50 times; each experiment consists of 1000 Monte Carlo trials to estimate the miscoverage rate, where each trial regenerates a finite evaluation pool and runs the method (see App. D for details). As shown in Table 1, CER-EVAL has a miscoverage rate of 0.161, far above the level $\alpha = 0.05$, indicating that its adaptive-partition algorithm can lose the claimed coverage. In contrast, all e-process-based methods stay below 0.05, with miscoverage rates between 0.021 and 0.030.

RQ2: Label Efficiency. Across SST-2, MMLU, and AG News, CELEUS uses 54%, 60%, and 62% fewer labels than EVALUE (Table 2), evaluating only 4.6%, 2.8%, and 4.6% of each pool. Using the dataset-level counts in Table 2, the reductions relative to CELEUS (W/O SURR) are approximately 18%, 37%, and 30% on SST-2, MMLU, and AG News, respectively. The 7–8B surrogate models receive no task-specific training or fine-tuning. This comparison reflects the combined contribution of surrogate assistance and uncertainty-guided sampling. Figure 1 shows that CELEUS's mean CI width remains below EVALUE and CELEUS (W/O SURR) across the displayed evaluation budgets. At the budget where ORACLE first reaches ϵ, CELEUS closes 90%/93%/84% of the EVALUE-to-ORACLE width gap, compared with 53%/60%/60% for CELEUS (W/O SURR). This is consistent with the role of conditional variance in Thm. 5. The oracle uses the true residual magnitudes to reduce one-step conditional variance; this does not establish optimality for cumulative variance or stopping time. App. F reports miscoverage results, with all methods below $\alpha = 0.05$.

RQ3: What makes CELEUS work. ① *Estimator unbiasedness.* Fig. 2 (left) compares the empirical bias of $\hat{R}_t$ in Eqn. (5) with a naive unweighted estimator that removes the weight $u_{m,t}$, using R_N as the reference since the population risk R is unknown. Across all three datasets, $\hat{R}_t$ stays centered

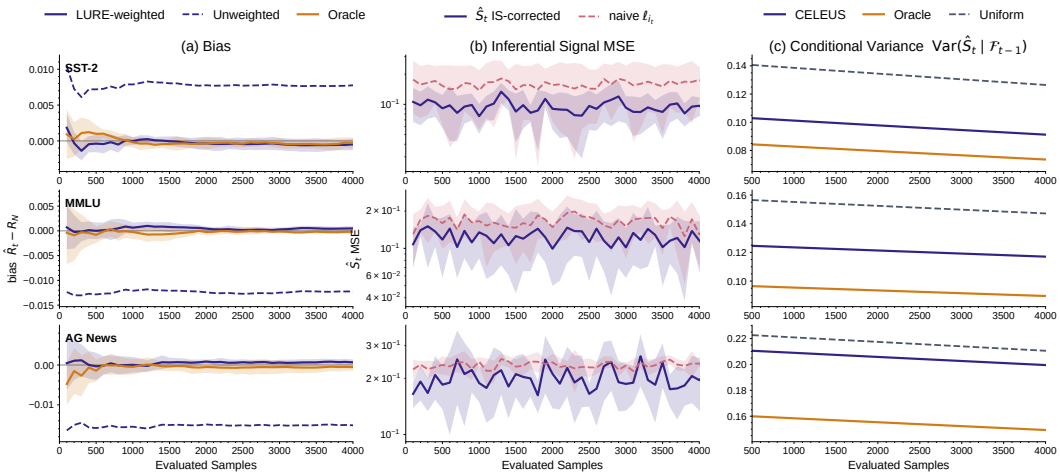

Figure 2: Mechanism analysis for CELEUS, pooled across the six surrogate-target pairs. *Left:* empirical bias of $\hat{R}_t$ in Eqn. (5) versus the unweighted estimator. *Middle:* MSE of $\hat{S}_t$ in Eqn. (2) versus the naive single-step observed risk. *Right:* $\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1})$ under uniform sampling, CELEUS's uncertainty-guided sampling, and the oracle q_t^{orc} in Remark 6.

around zero throughout the evaluation process, consistent with the unbiasedness guarantee in Thm. 1. In contrast, the naive estimator has a *persistent* bias of order 10^{-2}, about 1% on SST-2 and MMLU and 2% on AG News. This shows that the weight $u_{m,t}$ is necessary under non-uniform sample selection.

② *Inferential signal.* Fig. 2 (middle) compares $\hat{S}_t$ in Eqn. (2) with the naive single-step risk $\ell(f(\boldsymbol{x}_{i_t}), y_{i_t})$ as estimators of the finite-pool risk R_N. On SST-2 and MMLU, $\hat{S}_t$ achieves 1.5-$2\times$ lower mean squared error (MSE) across t. Unlike the naive single-step risk, $\hat{S}_t$ leverages richer information, making the signal more stable.

③ *Conditional variance.* Fig. 2 (right) reports $\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1})$ over 50 seeds under uniform sampling, CELEUS's uncertainty-guided sampling, and the oracle q_t^{orc} in Remark 6. Compared with uniform sampling, CELEUS reduces the variance by 27%, 20%, and 5% on SST-2, MMLU, and AG News, respectively, and closes $\approx 67\%$, $\approx 53\%$, and $\approx 18\%$ of the uniform-to-ORACLE variance gap. This is consistent with the variance-minimization principle in Thm. 4: sampling probabilities that better reflect surrogate-residual magnitudes reduce variance more effectively.

5.3 Additional Experiments

App. H compares the sample-selection strategies in Rem. 1. App. I studies the sensitivity of CELEUS to key hyperparameters. App. J compares the per-round runtime of CELEUS with the baselines. App. K reports detailed results for each surrogate-target pair, corresponding to the summary results in Tab. 2 and Fig. 1. App. A.2 further compares CELEUS with CER-EVAL, and App. A.4 separates the effects of surrogate and acquisition quality. App. L evaluates open-ended generation on RealToxicityPrompts and studies a pre-specified cross-entropy bound.

6 Conclusion

This paper proposed CELEUS, an e-process-based framework for certifiable and efficient LLM evaluation. CELEUS builds anytime-valid CIs from conditionally unbiased signals, uses surrogate-assisted approximation and uncertainty-guided sampling to reduce variance, and stops once the CI reaches a target precision. Theory and experiments show that CELEUS preserves coverage and reaches the target precision with fewer evaluated samples than baselines. An interesting direction for future work is to extend CELEUS to distribution-shift-aware evaluation scenarios. Additional limitations and the scope of the theoretical guarantees are discussed in App. M.

Acknowledgments and Disclosure of Funding

This research was supported by the University of Melbourne's Research Computing Services and the Petascale Campus Initiative. ZJZ is supported by the Melbourne Research Scholarship and the ARC with grant number DE240101089. FL is supported by the ARC with grant numbers DP230101540, DE240101089 and DP260103996. This work is partially supported by the National Science Foundation under grant No. 1910100, No. 2046726, NSF AI Institute ACTION No. IIS-2229876, DARPA Cooperative Agreement HR00112520013, the National Aeronautics and Space Administration (NASA) under grant No. 80NSSC20M0229, ARL Grant W911NF-23-2-0137, Alfred P. Sloan Fellowship, the research grant from eBay, AI Safety Fund, and Schmidt Science.

References

[1] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

[2] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*, 2022.

[3] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, et al. Evaluating large language models trained on code. *arXiv*, 2021.

[4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

[5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

[6] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[7] Aarohi Srivastava, Abhinav Rastogi, Abhay Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[8] Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

[9] Ganghua Wang, Zhaorun Chen, Bo Li, and Haifeng Xu. Cer-eval: Certifiable and cost-efficient evaluation framework for llms. *arXiv*, 2025.

[10] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 2023.

[11] Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. Ppi++: Efficient prediction-powered inference. *arXiv*, 2023.

[12] Tijana Zrnic and Emmanuel J Candès. Active statistical inference. In *ICML*, 2024.

[13] Daniel Csillag, Cláudio José Struchiner, and Guilherme Tegoni Goedert. Prediction-powered e-values. In *ICML*, 2025.

[14] Valentin Kilian, Stefano Cortinovis, and Francois Caron. Anytime-valid, bayes-assisted, prediction-powered inference. In *NeurIPS*, 2025.

[15] Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *NeurIPS*, 2020.

[16] Leonid Kontorovich and Kavita Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6):2126–2158, 2008.

[17] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research*, 11(2), 2010.

[18] Gabrielle Berrada, Jannik Kossen, Freddie Bickford Smith, Muhammed Razzak, Yarin Gal, and Tom Rainforth. Scaling up active testing to large language models. In *NeurIPS*, 2025.

[19] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *ICML*, 2021.

[20] Shengjia Zhao, Enze Zhou, Ashish Sabharwal, and Stefano Ermon. Adaptive concentration inequalities for sequential decision problems. In *NeurIPS*, 2016.

[21] Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

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

[23] Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

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

[25] Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. In *ICLR*, 2021.

[26] Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *Journal of Artificial Intelligence Research*, 42: 427–486, 2011.

[27] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pages 1631–1642, 2013.

[28] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, 2015.

[29] Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Thomas Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. *NeurIPS*, 2022.

[30] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *ICLR*, 2020.

[31] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv*, 2023.

[32] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[33] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[34] Sangwoo Park, Matteo Zecchin, and Osvaldo Simeone. Adaptive prediction-powered autoeval with reliability and efficiency guarantees. In *NeurIPS*, 2025.

[35] Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *NeurIPS*, 2024.

[36] Maria-Eleni Sfyraki and Jun-Kun Wang. Revisiting active sequential prediction-powered mean estimation. In *ICLR*, 2026.

[37] Daniel Hsu, Aryeh Kontorovich, David A. Levin, Yuval Peres, Csaba Szepesvari, and Geoffrey Wolfer. Mixing time estimation in reversible markov chains from a single sample path. *The Annals of Applied Probability*, 29(4):2439–2480, 2019.

[38] Jean Ville. *Etude critique de la notion de collectif*, volume 3. Gauthier-Villars Paris, 1939.

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023.

[40] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv*, 2024.

[41] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv*, 2024.

[42] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv*, 2024.

[43] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv*, 2024.

[44] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of EMNLP*, 2020.

[45] Laura Hanu, James Thewlis, and Sasha Haco. How AI is learning to identify toxic online content. *Scientific American*, 2021.

Appendix Contents

A Additional discussions **16**

A.1 Related work . . . 16

A.2 Comparison with Closely Related Methods . . . 16

A.3 Remarks on practical choices and implementation details . . . 17

A.4 Effects of surrogate and acquisition quality on CI width . . . 22

A.5 From per-round signals to performance estimation . . . 25

B End-to-end implementation of CELEUS **26**

C Detailed Proofs for Our Theoretical Results **28**

C.1 Assumptions . . . 28

C.2 Proofs of Theorem 1 . . . 29

C.3 Proofs of Theorem 2 . . . 33

C.4 Proofs of Theorem 3 . . . 35

C.5 Proofs of Theorem 4 . . . 38

C.6 Proofs of Theorem 5 . . . 42

C.7 Proofs of Theorem 6 . . . 49

D Synthetic Experimental Setting for Table 1 **51**

E Experimental Setup Details **52**

E.1 Benchmarks . . . 52

E.2 Models and surrogate-target pairs . . . 53

E.3 Baselines . . . 53

E.4 Hyperparameters . . . 54

E.5 Trials and reporting . . . 54

F Empirical coverage on real LLM evaluation under 0-1 loss **55**

G Certifiable and Label-efficient LLM evaluation on Cross-Entropy Loss **55**

H Sample-selection strategy comparison **56**

I Hyperparameter sensitivity **57**

J Computational cost and break-even analysis **59**

K Results for each surrogate-target model pair **61**

L Additional Empirical Results **61**

L.1 Open-Ended Safety Evaluation . . . 61

L.2 Sensitivity to a Pre-Specified Cross-Entropy Bound . . . 63

L.3 Computational Cost and the Break-Even Label Price 64

M Limitations and discussion **65**

M.1 Discussion: conservative numerical inversion 66

A Additional discussions

A.1 Related work

Efficient evaluation via sample selection. A growing line of work studies how to improve evaluation efficiency through active testing and adaptive sample selection. Existing approaches include cheaper proxy predictors, representative subset selection, and adaptive selection of informative examples [29, 30, 19, 18]. Active testing and related adaptive evaluation methods show that evaluation can be made more efficient by prioritizing informative examples rather than evaluating the entire test pool uniformly at random [19, 25]. These methods provide a natural framework for pool-based evaluation under limited evaluation budgets [18]. However, while they offer principled sampling strategies and, in some cases, valid estimation under adaptive sampling, their interval guarantees are typically not designed for fully sequential use. In particular, the resulting intervals or stopping criteria are generally not *anytime-valid*, and may fail to remain valid when the evaluator repeatedly monitors the current interval and stops at a data-dependent time.

Prediction-powered and semi-supervised inference for evaluation. Another closely related direction is prediction-powered inference (PPI), which combines a limited set of evaluated samples with auxiliary predictions on unevaluated samples [10, 11]. In LLM evaluation, auxiliary predictions can come from surrogate models or automatic raters [5, 31]. Classical PPI primarily provides fixed-time inference, while active statistical inference and subsequent prediction-powered confidence sequences also address sequential validity [12–14]. Our setting differs in its adaptive, unequal-probability sampling without replacement from a fixed finite pool. This sampling structure requires a signal whose conditional mean remains the same finite-pool risk as the remaining set and proposal probabilities change.

Sequential inference and anytime-valid CIs. Our work is also related to sequential inference, confidence sequences, and e-processes [32, 33, 21]. This line of work develops hypothesis tests and confidence intervals that remain valid under continuous monitoring, making them suitable for settings with adaptive stopping. In contrast to classical fixed-horizon intervals, anytime-valid procedures preserve statistical validity no matter when the analyst chooses to stop. However, many existing constructions are developed for online settings in which observations arrive sequentially from an external stochastic process [12, 34–36]. In our setting, evaluation begins with a fixed evaluation pool, and the evaluator sequentially selects which examples to evaluate without replacement. This pool-based structure induces a different dependence pattern and allows the procedure to use information from both evaluated and unevaluated samples. Existing sequential inference tools provide the foundation for anytime-valid inference, including confidence sequences for sampling without replacement [15], but do not directly address the combination of surrogate-assisted estimation, adaptive sample selection, and certifiable LLM evaluation considered here.

A.2 Comparison with Closely Related Methods

Recent prediction-powered methods are closely related to CELEUS in their use of auxiliary predictions to improve the efficiency of sequential inference. The main distinction lies in the underlying data-acquisition protocol. Csillag et al. [13] propose Prediction-Powered E-Values for an externally arriving data stream, where the procedure decides at each round whether to query the label of the current observation. Although the query probability may depend on previously observed information, the procedure does not control which observation arrives next. CELEUS instead begins with a fixed evaluation pool and adaptively selects the next item from the remaining unevaluated samples. This results in adaptive, unequal-probability sampling without replacement, which CELEUS explicitly accounts for through its conditionally unbiased signal construction.

Kilian et al. [14] develop anytime-valid, Bayes-assisted prediction-powered inference under independently sampled labelled and unlabelled data streams, with coverage established asymptotically uniformly over time. In contrast, CELEUS provides finite-sample, time-uniform coverage under adaptive sampling from a fixed evaluation pool. Therefore, although these methods share the goal of using inexpensive auxiliary predictions to reduce evaluation costs, the guarantees established in these works do not directly cover the adaptive sampling-without-replacement setting considered by CELEUS.

Table 3: Number of target labels required by CELEUS and CER-EVAL to reach the target precision on MMLU. Results are averaged over three random seeds. Label reduction is computed relative to CER-EVAL. Lower values are better.

Surrogate → Target	CELEUS	CER-EVAL	Label Reduction
Llama-3-8B → Llama-3-70B	2,373.7	3,771.0	37.1%
Llama-2-7B → Llama-3-70B	2,799.0	3,771.0	25.8%
Llama-3-8B → Qwen2.5-72B	1,478.3	2,349.0	37.1%
Llama-3-8B → Mixtral-8×7B	3,856.0	5,474.7	29.6%
Llama-2-7B → DeepSeek-67B	5,394.7	6,451.7	16.4%
Llama-3-8B → DeepSeek-67B	4,734.7	6,451.7	26.6%

CER-EVAL and adaptive partitioning. Wang et al. [9] study a closely related fixed-pool evaluation setting and propose a partition-based strategy for reducing evaluation cost. Their theoretical result considers the algorithm when supplied with a fixed partition satisfying the required conditions, whereas the practical procedure repeatedly learns and updates the partition using evaluated samples and their observed risks. Consequently, the stated guarantee does not directly cover the complete adaptive partition-learning procedure, for which an additional argument accounting for the data-dependent updates would be required. This distinction concerns the scope of the current theoretical analysis rather than the validity of adaptive partitioning in general. By comparison, CELEUS allows the sampling distribution to depend on the observed history while preserving the conditional unbiasedness of its inferential signal, thereby supporting anytime-valid inference under the adaptive sampling and stopping rules considered here.

CER-EVAL is therefore a particularly relevant empirical baseline for CELEUS. We compare the two methods under the same MMLU evaluation setting and target precision. Table 3 reports the number of queried target losses required to reach the target precision, averaged over three random seeds, 42, 43, and 44. The surrogate model listed in the first column is used only by CELEUS, whereas CER-EVAL does not use a surrogate model.

Across all six surrogate–target pairs, CELEUS requires fewer target labels than CER-EVAL to reach the same target precision. The reduction ranges from 16.4% to 37.1%, with an average reduction of 28.8%. The improvement is observed across all four target models considered, including both dense and mixture-of-experts architectures, indicating that it is not driven by a single favorable model pair.

The results also illustrate how the choice of surrogate can affect the efficiency of CELEUS. For both Llama-3-70B and DeepSeek-67B, using Llama-3-8B as the surrogate results in a larger label reduction than using Llama-2-7B. Since the CER-EVAL result remains unchanged for a fixed target, this variation arises from the CELEUS surrogate-target pairing. This observation is consistent with our theoretical analysis: when a surrogate more accurately approximates the target losses, the resulting surrogate residuals and the conditional variance of the inferential signals can be reduced, allowing the confidence interval to reach the target precision with fewer evaluated labels. Overall, these results show that the label-efficiency improvement of CELEUS over CER-EVAL persists across the surrogate–target configurations considered in this MMLU experiment.

A.3 Remarks on practical choices and implementation details

We first describe how the uncertainty-guided sampling strategies are instantiated for the two loss families used in our experiments.

Remark 1 (Surrogate scoring strategies and sample-selection scores for 0-1 loss.)**.** *We describe one oracle strategy and three practical strategies for the case of* 0-1 *loss.*

Let $p_f(\cdot \mid \boldsymbol{x})$ *and* $p_g(\cdot \mid \boldsymbol{x})$ *denote the predictive distributions of the target LLM* f *and a surrogate model* g*, respectively. Define the true* 0-1 *loss of the target model by*

$$\ell_{0\text{-}1}(f, \boldsymbol{x}, y) = \mathbb{I}\{\hat{y}_f(\boldsymbol{x}) \neq y\}, \qquad \hat{y}_f(\boldsymbol{x}) = \arg\max_{y' \in \mathcal{Y}} p_f(y' \mid \boldsymbol{x}).$$

For any model $h \in \{f, g\}$*, define the predictive dispersion score*

$$\ell_{\text{disp}}(h, \boldsymbol{x}) = 1 - \sum_{y \in \mathcal{Y}} p_h^2(y \mid \boldsymbol{x}).$$

For $h_1, h_2 \in \{f, g\}$, define the hard pseudo-loss and soft pseudo-loss as

$$\ell_{\text{hard}}(h_1, h_2, \boldsymbol{x}) = \mathbb{I}\{\hat{y}_{h_1}(\boldsymbol{x}) \neq \hat{y}_{h_2}(\boldsymbol{x})\},$$

and

$$\ell_{\text{soft}}(h_1, h_2, \boldsymbol{x}) = 1 - p_{h_1}(\hat{y}_{h_2}(\boldsymbol{x}) \mid \boldsymbol{x}), \qquad \hat{y}_h(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} p_h(y \mid \boldsymbol{x}).$$

Here, the top-1 prediction of h_2 is used as a pseudo label for assessing the prediction of h_1.

- ***Oracle strategy*** *(true 0-1 loss and surrogate score). If the true 0-1 loss of the target model were accessible, we would set*

  $$\ell(f(\boldsymbol{x}_j), y_j) = \ell_{\text{0-1}}(f, \boldsymbol{x}_j, y_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) \in \{\ell_{\text{disp}}(f, \boldsymbol{x}_j), \ell_{\text{soft}}(f, g, \boldsymbol{x}_j)\}.$$

  The sample-selection score is

  $$|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)|.$$

  This oracle strategy is infeasible in practice because it requires the true loss values of unevaluated samples. In the oracle experiments, these scores are sampled with $\beta = 0.4$ according to Remark 6.

- ***Strategy A*** *(hard pseudo-loss as target-loss proxy, predictive dispersion as surrogate score). Set*

  $$\ell(f(\boldsymbol{x}_j), y_j) \approx \ell_{\text{hard}}(f, g, \boldsymbol{x}_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) = \ell_{\text{disp}}(f, \boldsymbol{x}_j).$$

  The sample-selection score is

  $$|\ell_{\text{hard}}(f, g, \boldsymbol{x}_j) - \ell_{\text{disp}}(f, \boldsymbol{x}_j)|.$$

  This compares the hard disagreement between the target and surrogate predictions with the dispersion of the target model's predictive distribution.

- ***Strategy B*** *(predictive dispersion as target-loss proxy, soft pseudo-loss as surrogate score). Set*

  $$\ell(f(\boldsymbol{x}_j), y_j) \approx \ell_{\text{disp}}(f, \boldsymbol{x}_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) = \ell_{\text{soft}}(f, g, \boldsymbol{x}_j).$$

  The sample-selection score is

  $$|\ell_{\text{disp}}(f, \boldsymbol{x}_j) - \ell_{\text{soft}}(f, g, \boldsymbol{x}_j)|.$$

  This compares the dispersion of the target model's predictive distribution with its soft loss evaluated at the surrogate model's top-1 prediction.

- ***Strategy C*** *(soft pseudo-loss as target-loss proxy, predictive dispersion as surrogate score). Set*

  $$\ell(f(\boldsymbol{x}_j), y_j) \approx \ell_{\text{soft}}(f, g, \boldsymbol{x}_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) = \ell_{\text{disp}}(f, \boldsymbol{x}_j).$$

  The sample-selection score is

  $$|\ell_{\text{soft}}(f, g, \boldsymbol{x}_j) - \ell_{\text{disp}}(f, \boldsymbol{x}_j)|,$$

  which is numerically identical to that of Strategy B. However, Strategies B and C differ in which quantity plays the role of the target-loss proxy and which plays the role of the surrogate score inside the estimator $\hat{R}_t$ in Eqn. (5). This distinction changes the correction term and can affect the variance of the estimator and the resulting e-process.

The three practical strategies use the surrogate model in different ways. Strategy A uses the surrogate model's top-1 prediction to form a hard pseudo-loss, while Strategies B and C use the surrogate model's top-1 prediction to form a soft pseudo-loss. In our main 0-1 loss experiments, we use Strategy A. The other strategies are also compatible with our framework, and we compare them empirically in App. H; see Fig. 4.

Remark 2 (Surrogate scoring strategies and sample-selection scores for cross-entropy loss.)**.** *Recall that, when the residuals are not all zero, the oracle sampling distribution in Theorem 4 is*

$$q_t^*(j) = \max\left\{\frac{\beta}{N - t + 1}, \frac{|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)|}{\eta_t}\right\},$$

where $\eta_t > 0$ is chosen so that the probabilities sum to one. This distribution is unavailable in practice because it depends on the unknown label y_j. To obtain computable sample-selection scores, we replace the unknown target loss with label-free proxies when constructing the sampling distribution.

Let $p_f(\cdot \mid \boldsymbol{x})$ and $p_g(\cdot \mid \boldsymbol{x})$ denote the predictive distributions of the target LLM f and a surrogate model g, respectively. Before the sequential evaluation, the evaluator prespecifies a fixed clipping level $B > 0$. Throughout the cross-entropy evaluation, we define the target loss as

$$\ell_{\mathrm{ce}}(f, \boldsymbol{x}, y) = \min\left\{-\log p_f(y \mid \boldsymbol{x}), B\right\}.$$

For any model $h \in \{f, g\}$, define its clipped predictive entropy score as

$$H(h, \boldsymbol{x}) = \min\left\{-\sum_{y \in \mathcal{Y}} p_h(y \mid \boldsymbol{x}) \log p_h(y \mid \boldsymbol{x}), B\right\}.$$

We also define the clipped mode loss of the target model as

$$\ell_{\mathrm{mode}}(f, \boldsymbol{x}) = \min\left\{-\log p_f(\hat{y}_f(\boldsymbol{x}) \mid \boldsymbol{x}), B\right\}, \qquad \hat{y}_f(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} p_f(y \mid \boldsymbol{x}),$$

which uses the target model's own top-1 prediction as a pseudo label.

In all strategies below, the target loss used in the estimator after a sample is evaluated remains $\ell_{\mathrm{ce}}(f, \boldsymbol{x}_j, y_j)$. The label-free proxy is used only to construct the sample-selection score.

- ***Oracle strategy*** *(true cross-entropy loss and surrogate score). If the true clipped cross-entropy loss were accessible, we would set*

  $$\ell(f(\boldsymbol{x}_j), y_j) = \ell_{\mathrm{ce}}(f, \boldsymbol{x}_j, y_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) \in \{H(g, \boldsymbol{x}_j), H(f, \boldsymbol{x}_j), \ell_{\mathrm{mode}}(f, \boldsymbol{x}_j)\}.$$

  The sample-selection score is

  $$|\ell_{\mathrm{ce}}(f, \boldsymbol{x}_j, y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)|.$$

  This oracle strategy is infeasible because it requires the true loss values of unevaluated samples. In the oracle experiments, these scores are sampled with $\beta = 0.4$ according to Remark 6.

- ***Strategy A*** *(mode loss as target-loss proxy, entropy as surrogate score). For sample selection, set*

  $$\ell_{\mathrm{proxy}}(\boldsymbol{x}_j) = \ell_{\mathrm{mode}}(f, \boldsymbol{x}_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) = H(f, \boldsymbol{x}_j).$$

  The sample-selection score is

  $$|\ell_{\mathrm{mode}}(f, \boldsymbol{x}_j) - H(f, \boldsymbol{x}_j)|.$$

  This compares the loss at the target model's top-1 prediction with the entropy of its predictive distribution.

- ***Strategy B*** *(entropy as target-loss proxy, mode loss as surrogate score). For sample selection, set*

  $$\ell_{\mathrm{proxy}}(\boldsymbol{x}_j) = H(f, \boldsymbol{x}_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) = \ell_{\mathrm{mode}}(f, \boldsymbol{x}_j).$$

  The sample-selection score is

  $$|H(f, \boldsymbol{x}_j) - \ell_{\mathrm{mode}}(f, \boldsymbol{x}_j)|,$$

  which is numerically identical to that of Strategy A. However, Strategies A and B use different surrogate scores $\tilde{\ell}_f$ in the estimator $\hat{R}_t$ in Eqn. (5). This changes the residual correction and can affect the variance of the estimator and the resulting e-process.

- ***Strategy C*** *(target-surrogate entropy gap). For sample selection, set*

  $$\ell_{\mathrm{proxy}}(\boldsymbol{x}_j) = H(f, \boldsymbol{x}_j), \qquad \tilde{\ell}_f(\boldsymbol{x}_j) = H(g, \boldsymbol{x}_j).$$

  The sample-selection score is

  $$|H(f, \boldsymbol{x}_j) - H(g, \boldsymbol{x}_j)|.$$

  This compares the predictive entropies of the target and surrogate models, capturing inter-model disagreement at the distributional level.

Computing the practical acquisition scores above requires the target predictive outputs $p_f(\cdot \mid \boldsymbol{x}_j)$ over the full pool. Thus, the current implementation uses these outputs to determine which target losses should be revealed, but does not reduce the number of target forward passes. Its direct efficiency benefit is a reduction in the labels, human judgments, or model-judge evaluations required for certification. Acquisition rules based entirely on surrogate or previously available information could avoid full-pool target inference, but are not the instantiation evaluated in this paper.

All target losses, surrogate scores, and target-loss proxies defined above lie in $[0, B]$. Therefore, Assumption 1 holds with $L = 0$ and $U = B$, and every sample-selection score lies in $[0, B]$. In this setting, R_N and R denote the finite-pool and population means of the clipped target cross-entropy loss, respectively.

Strategy C is the only strategy that uses a separate surrogate model g, while Strategies A and B are self-referential and use only the target model f. In our main cross-entropy experiments, we use Strategy A. The other strategies are also compatible with our framework, and we compare them empirically in App. H; see Fig. 5.

Overall, these practical strategies are uncertainty-guided sample-selection rules: they use label-free quantities to approximate the oracle residual magnitude $|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)|$ in Theorem 4. Since this residual determines the variance-optimal sampling distribution, assigning larger probabilities to samples with larger proxy residuals aims to reduce the conditional variance of $\hat{S}_t$ and hence accelerate CI-width shrinkage. These label-free quantities follow sample-selection scores commonly used in active testing and active statistical inference, such as the expected surrogate loss, target–surrogate disagreement, and predictive uncertainty [19, 29, 18, 12]. The choice among them affects efficiency but not validity, and the most effective strategy depends on the task, the evaluation loss, and the available model outputs. Principled strategy selection remains an open problem.

Having specified the practical sample-selection strategies, we next describe the implementation choices used to construct the e-processes.

Remark 3. *Following [21], we use predictable, time-adaptive betting fractions to control the Type-I error uniformly over time, ensuring that the resulting confidence sequence remains valid under data-dependent stopping. Since the scaling bounds (a_t, b_t) may vary across rounds, each candidate value υ is transformed using the same round-specific scaling as the signal:*

$$\bar{\upsilon}_t = \frac{\upsilon - a_t}{b_t - a_t}.$$

To construct candidate-independent base fractions, define the regularized running mean and variability estimate

$$\hat{\mu}_t = \frac{1/2 + \sum_{i=1}^{t} \bar{S}_i}{t+1}, \qquad \hat{\sigma}_t^2 = \frac{1/4 + \sum_{i=1}^{t} (\bar{S}_i - \hat{\mu}_i)^2}{t+1}.$$

Using the constants $\underline{\gamma}$ and $\overline{\gamma}$ from Assumption 4, we set

$$\gamma_t = \left(\sqrt{\frac{2\log(2/\alpha_1)}{\hat{\sigma}_{t-1}^2}} \vee \underline{\gamma} \right) \wedge \overline{\gamma}, \qquad \tilde{\lambda}_t^+ = \tilde{\lambda}_t^- = \frac{\gamma_t}{\sqrt{t \log(t+1)}}.$$

For each candidate value υ, the final betting fractions are

$$\lambda_t^+(\upsilon) = \tilde{\lambda}_t^+ \wedge \frac{1}{\bar{\upsilon}_t + c}, \qquad \lambda_t^-(\upsilon) = \tilde{\lambda}_t^- \wedge \frac{1}{1 - \bar{\upsilon}_t + c},$$

where $c \in (0, 1)$ is a small constant.

Because $\hat{\sigma}_{t-1}^2$, a_t, and b_t depend only on information available before round t, the resulting fractions are predictable. Under $\mathcal{H}_0(\upsilon)$, the centered scaled signal $\bar{S}_t - \bar{\upsilon}_t$ has conditional mean zero. The truncation of γ_t ensures the envelope condition in Assumption 4, while the candidate-specific clipping keeps the final fractions within their admissible ranges.

Finally, we describe how the scaling bounds are chosen so that the per-round signal can be mapped to a bounded range before applying the e-process construction.

Remark 4. *The scaling bounds a_t and b_t need to be predictable and satisfy $a_t \le \hat{S}_t \le b_t$ almost surely. In practice, one simple choice is to use fixed bounds $a_t = a$ and $b_t = b$ for all rounds. Suppose that*

$$q_t(j \mid i_{1:t-1}, X_N) \ge q_{\min} > 0$$

for all t and all $j \in \mathcal{J}_{t-1}$. We can rewrite the signal as

$$\begin{aligned}\hat{S}_t &= \frac{\sum_{m=1}^{t-1} \ell(f(\boldsymbol{x}_{i_m}), y_{i_m}) + \sum_{j \in \mathcal{J}_t} \tilde{\ell}_f(\boldsymbol{x}_j) + \ell(f(\boldsymbol{x}_{i_t}), y_{i_t})}{N} \\ &\quad + \frac{1 - q_t(i_t \mid i_{1:t-1}, X_N)}{N q_t(i_t \mid i_{1:t-1}, X_N)} \left(\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})\right).\end{aligned}$$

The first term is an average of observed losses and surrogate loss scores, and therefore lies in $[L, U]$. The second term is bounded in absolute value by

$$\frac{(U - L)(1 - q_{\min})}{N q_{\min}}.$$

Thus, a valid fixed choice is

$$a = L - \frac{(U - L)(1 - q_{\min})}{N q_{\min}}, \qquad b = U + \frac{(U - L)(1 - q_{\min})}{N q_{\min}},$$

which ensures $\hat{S}_t \in [a, b]$ almost surely for all t. Under Assumption 2, we can take $q_{\min} = \beta/N$.

Remark 5. *For notational simplicity, we write $q_t(j)$ below, suppressing the conditioning on past selections and on X_N for practical sampling or D_N for the oracle benchmark. At round t, let $s_t(j) \ge 0$ denote the sample-selection score for $j \in \mathcal{J}_{t-1}$. For the experimental oracle benchmark, we use the true residual score*

$$s_t^{\text{oracle}}(j) = \left|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|.$$

This score is unavailable to the practical method because y_j has not yet been observed. For the practical samplers, we instead use

$$s_t^{\text{proxy}}(j) = \left|\ell_{\text{proxy}}(\boldsymbol{x}_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|,$$

where the label-free target-loss proxies and surrogate scores are defined in Remarks 1 and 2.

For the practical scores, we first define

$$p_t(j) = \begin{cases} \dfrac{s_t(j)}{\sum_{k \in \mathcal{J}_{t-1}} s_t(k)}, & \displaystyle\sum_{k \in \mathcal{J}_{t-1}} s_t(k) > 10^{-12}, \\ \dfrac{1}{N - t + 1}, & \text{otherwise.} \end{cases}$$

The threshold 10^{-12} is a numerical safeguard: when the total score is negligible, we revert to uniform sampling. The final practical sampling distribution is

$$q_t(j) = \frac{\beta}{N - t + 1} + (1 - \beta) p_t(j).$$

This construction satisfies

$$q_t(j) \ge \frac{\beta}{N - t + 1},$$

as required by Assumption 2. For practical sampling, we use $\beta = 0.4$ by default, unless otherwise specified, to balance uniform exploration and score-guided sampling. The probability floor prevents very small sampling probabilities and unstable inverse-probability corrections when the practical scores only approximate the unknown residuals.

Here, "oracle" refers to access to the true residual scores when constructing the experimental benchmark (Remark 6). The floor-mixture above is the stable sampling rule used in our experiments, while Theorem 4 separately characterizes the exact constrained variance-minimizing distribution. For the practical method, the proxy scores affect only sample selection; after a sample is selected, its observed target loss is used in the estimator and the e-process. Provided that the scores are available before selection, the probability-floor and scaling assumptions hold, and the actual sampling probabilities are used in the inverse-probability corrections, the proxy scores and positive β affect efficiency without invalidating the coverage guarantee.

Remark 6 (Oracle experiments)**.** *In the oracle experiments, the complete labeled pool D_N is available to the sampling rule. The oracle uses the true residual magnitudes as the sample-selection scores in Eqn. (8), with the same $\beta = 0.4$ as the practical method. This is an idealized benchmark.*

Write $m_t = N - t + 1$, $r_{t,j} = \ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)$, and $A_t = \sum_{j \in \mathcal{J}_{t-1}} |r_{t,j}|$. Retaining the score-normalization safeguard $\delta_{\text{num}} = 10^{-12}$, the experimental oracle uses

$$q_t^{\text{orc}}(j) = \frac{\beta}{m_t} + (1-\beta) p_t^{\text{res}}(j), \qquad p_t^{\text{res}}(j) = \begin{cases} |r_{t,j}|/A_t, & A_t > \delta_{\text{num}}, \\ 1/m_t, & A_t \le \delta_{\text{num}}. \end{cases} \tag{6}$$

Since $q_t^{\text{orc}}(j) \ge \beta/m_t$, the oracle satisfies Assumption 2. Here, q_t^{orc} is the implemented oracle-information reference based on unavailable true residuals; it is distinct from the exact constrained optimizer q_t^* in Theorem 4 unless the two constructions coincide.

A.4 Effects of surrogate and acquisition quality on CI width

We now clarify how surrogate quality and acquisition quality affect the confidence-interval width. The two quantities play different roles: the surrogate score determines the residuals appearing in the inferential signal, whereas the acquisition proxy determines how the sampling probabilities allocate evaluations across these residuals.

Theoretical analysis. Fix a round t and write

$$m_t = N - t + 1, \qquad r_{t,j} = \ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j), \qquad j \in \mathcal{J}_{t-1}.$$

If $r_{t,j} = 0$ for every $j \in \mathcal{J}_{t-1}$, then Theorem 3 gives

$$\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1}) = 0,$$

and every proposal satisfying Assumption 2 is optimal. We therefore consider the case in which the residuals are not all zero. Define

$$A_t = \sum_{j \in \mathcal{J}_{t-1}} |r_{t,j}|, \qquad \pi_t(j) = \frac{|r_{t,j}|}{A_t}.$$

Here, π_t is the residual-normalized distribution obtained without imposing the probability floor.

For a feasible proposal q, let $\mathrm{Var}_q(\hat{S}_t \mid \mathcal{F}_{t-1})$ denote the conditional variance of $\hat{S}_t$ when q is used at round t. The part of this variance that depends on the proposal is

$$\Phi_t(q) = \frac{1}{N^2} \sum_{j \in \mathcal{J}_{t-1}} \frac{r_{t,j}^2}{q(j)}.$$

Since Assumption 2 ensures $q(j) > 0$, we have the exact decomposition

$$\begin{aligned} \Phi_t(q) &= \frac{A_t^2}{N^2} \sum_{j \in \mathcal{J}_{t-1}} \frac{\pi_t(j)^2}{q(j)} \\ &= \frac{A_t^2}{N^2} \left(1 + \chi^2(\pi_t \| q)\right), \end{aligned}$$

where

$$\chi^2(\pi_t \| q) = \sum_{j \in \mathcal{J}_{t-1}} \frac{(\pi_t(j) - q(j))^2}{q(j)}.$$

Consequently,

$$\mathrm{Var}_q(\hat{S}_t \mid \mathcal{F}_{t-1}) = \frac{A_t^2}{N^2} \left(1 + \chi^2(\pi_t \| q)\right) - \left(\frac{1}{N} \sum_{j \in \mathcal{J}_{t-1}} r_{t,j} \right)^2.$$

To account for the probability floor, define

$$\mathcal{Q}_{t,\beta} = \left\{ q : \sum_{j \in \mathcal{J}_{t-1}} q(j) = 1, \quad q(j) \ge \frac{\beta}{m_t} \right\}.$$

Because the second term in the preceding variance expression is independent of q, the constrained optimizer in Theorem 4 equivalently satisfies

$$q_t^* = \arg\min_{q \in \mathcal{Q}_{t,\beta}} \chi^2(\pi_t \| q).$$

Thus, for every feasible proxy-induced proposal q_t,

$$\begin{aligned}\mathrm{Var}_{q_t}(\hat{S}_t \mid \mathcal{F}_{t-1}) - \mathrm{Var}_{q_t^*}(\hat{S}_t \mid \mathcal{F}_{t-1}) &= \frac{A_t^2}{N^2}\left(\chi^2(\pi_t \| q_t) - \chi^2(\pi_t \| q_t^*)\right) \\ &\geq 0.\end{aligned}$$

If π_t itself satisfies the probability floor, then $q_t^* = \pi_t$. Otherwise, q_t^* is the water-filling proposal characterized in Theorem 4.

This decomposition separates the two approximation effects. First, the surrogate score $\tilde{\ell}_f$ determines the residuals $r_{t,j}$. In particular, Assumption 2 gives

$$\begin{aligned}\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1}) &\leq \frac{1}{N^2} \sum_{j \in \mathcal{J}_{t-1}} \frac{r_{t,j}^2}{q_t(j)} \\ &\leq \frac{m_t}{\beta N^2} \sum_{j \in \mathcal{J}_{t-1}} r_{t,j}^2 \\ &= \frac{m_t^2}{\beta N^2} \mathrm{MSE}_t,\end{aligned}$$

where

$$\mathrm{MSE}_t = \frac{1}{m_t} \sum_{j \in \mathcal{J}_{t-1}} \left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2.$$

Hence, a more accurate surrogate, in the sense of smaller residual mean squared error, tightens this upper bound on the conditional variance.

Second, the acquisition proxy is used to approximate the unknown residual magnitudes $|r_{t,j}|$ and thereby construct q_t. Holding the surrogate residuals fixed, its effect is measured exactly by

$$\chi^2(\pi_t \| q_t) - \chi^2(\pi_t \| q_t^*).$$

A proposal with a smaller divergence gap incurs less variance relative to the constrained oracle. Through the cumulative conditional-variance term in Theorem 5, improvements in either surrogate accuracy or acquisition quality can therefore tighten the confidence-width bound.

In Remarks 1 and 2, the target-loss proxy is used only to construct the sample-selection score, whereas $\tilde{\ell}_f$ is the surrogate score used in $\hat{S}_t$ and $\hat{R}_t$. Therefore, two strategies may produce the same acquisition score and hence the same sampling probabilities, while still producing different residuals and conditional variances because they use different surrogate scores.

These approximation choices affect efficiency rather than validity. No quantitative approximation-quality assumption is required for conditional unbiasedness. Provided that the surrogate scores are fixed before evaluation, the sampling probabilities depend only on observed information, the boundedness assumptions hold, the actual proposal probabilities satisfy Assumption 2 and are used in the inverse-probability correction, and the observed target loss is used after selection, the coverage guarantees in Theorems 5 and 6 remain unchanged. The population guarantee additionally uses the i.i.d. input-label pairs specified in the setup. Less accurate approximations may increase the conditional variance and delay stopping, but they do not invalidate the confidence sequence.

Controlled semi-synthetic analysis. We further conduct a controlled semi-synthetic analysis that varies surrogate and acquisition quality independently. Let

$$\ell_j = \ell(f(\boldsymbol{x}_j), y_j), \qquad \mathrm{MSE}(h, \ell) = \frac{1}{N} \sum_{j=1}^{N} \left(h(\boldsymbol{x}_j) - \ell_j\right)^2.$$

We measure surrogate quality using the normalized MSE skill relative to a fixed baseline predictor $\tilde{\ell}^{\mathrm{base}}$:

$$Q_{\mathrm{surr}} = 1 - \frac{\mathrm{MSE}(\tilde{\ell}_f, \ell)}{\mathrm{MSE}(\tilde{\ell}^{\mathrm{base}}, \ell)}.$$

Table 4: Number of evaluated samples required to reach CI width $\epsilon = 0.05$, reported as median [interquartile range] across 50 trajectories. Lower is better.

Task	Acquisition quality	$Q_{\text{surr}} = 0.8$	$Q_{\text{surr}} = 0.4$	$Q_{\text{surr}} = 0.0$
SST-2	$Q_{\text{acq}} = 1.0$	1,830 [1,801–1,863]	4,542 [4,490–4,629]	8,060 [7,983–8,172]
SST-2	$Q_{\text{acq}} = 0.5$	2,133 [2,060–2,168]	6,062 [5,890–6,148]	11,040 [10,881–11,244]
SST-2	$Q_{\text{acq}} = 0.0$	2,718 [2,681–2,776]	8,842 [8,586–9,079]	16,854 [16,529–17,287]
MMLU	$Q_{\text{acq}} = 1.0$	1,339 [1,328–1,346]	1,741 [1,712–1,766]	2,434 [2,404–2,510]
MMLU	$Q_{\text{acq}} = 0.5$	1,387 [1,377–1,399]	2,236 [2,176–2,275]	3,532 [3,410–3,636]
MMLU	$Q_{\text{acq}} = 0.0$	1,678 [1,628–1,714]	4,054 [3,902–4,270]	7,348 [7,148–7,664]

A larger value of Q_{surr} therefore corresponds to a smaller surrogate residual MSE. We consider

$$Q_{\text{surr}} \in \{0.8, 0.4, 0.0\}.$$

To control acquisition quality, we first construct an infeasible oracle-information reference proposal using the true residual magnitudes. Specifically, define

$$p_t^{\text{res}}(j) = \begin{cases} \dfrac{|r_{t,j}|}{\sum_{k \in \mathcal{J}_{t-1}} |r_{t,k}|}, & \displaystyle\sum_{k \in \mathcal{J}_{t-1}} |r_{t,k}| > 10^{-12}, \\ \dfrac{1}{m_t}, & \text{otherwise}, \end{cases}$$

and

$$q_t^{\text{orc}}(j) = \frac{\beta}{m_t} + (1 - \beta) p_t^{\text{res}}(j).$$

We use $\beta = 0.4$, as in the practical experiments. Here, q_t^{orc} is the implemented oracle-information reference based on unavailable true residuals; it is distinct from the exact constrained optimizer q_t^* in Theorem 4 unless the two constructions coincide.

Let

$$q_t^{\text{unif}}(j) = \frac{1}{m_t}.$$

For $\kappa \in [0, 1]$, we then define

$$q_t^{(\kappa)} = \kappa q_t^{\text{orc}} + (1 - \kappa) q_t^{\text{unif}}.$$

Both component distributions satisfy the probability floor, so every $q_t^{(\kappa)}$ also satisfies Assumption 2. We measure acquisition quality by its normalized total-variation proximity to the oracle-information reference:

$$Q_{\text{acq}} = 1 - \frac{\text{TV}\left(q_t^{(\kappa)}, q_t^{\text{orc}}\right)}{\text{TV}\left(q_t^{\text{unif}}, q_t^{\text{orc}}\right)} = \kappa,$$

where

$$\text{TV}(p, q) = \frac{1}{2} \sum_{j \in \mathcal{J}_{t-1}} |p(j) - q(j)|.$$

The equality holds whenever $q_t^{\text{orc}} \neq q_t^{\text{unif}}$; when the two distributions coincide, there is no acquisition-quality gap to control. We consider

$$Q_{\text{acq}} \in \{1.0, 0.5, 0.0\},$$

representing oracle-information sampling, an equal oracle–uniform mixture, and uniform sampling, respectively.

For each combination of surrogate and acquisition quality, we generate five surrogate constructions and run ten sampling repetitions, giving 50 trajectories per cell. Table 4 reports the number of evaluated samples required to reach CI width $\epsilon = 0.05$ as the median and interquartile range.

Holding surrogate quality fixed, improving acquisition quality from $Q_{\text{acq}} = 0$ to $Q_{\text{acq}} = 1$ reduces the median sample requirement by 32.7%–52.2% on SST-2 and 20.2%–66.9% on MMLU. Holding

acquisition quality fixed, improving surrogate quality from $Q_{\text{surr}} = 0$ to $Q_{\text{surr}} = 0.8$ reduces the median sample requirement by 77.3%–83.9% on SST-2 and 45.0%–77.2% on MMLU.

These results are consistent with the theoretical decomposition above. Better surrogate scores reduce the residual-error component, whereas better acquisition scores reduce the variance inflation caused by allocating sampling probability differently from the oracle-information reference. Both improvements therefore allow the confidence interval to reach the target width using fewer evaluated samples.

A.5 From per-round signals to performance estimation

In this appendix, we explain how the final performance estimator in Eqn. (5) is obtained by aggregating the correction information exposed by the inferential signal in Eqn. (2). The inferential signal is designed for e-process construction, while the final estimator uses the same surrogate-correction structure to report a scalar performance estimate at stopping.

Recall that the per-round inferential signal is

$$\hat{S}_t = \frac{\sum_{m=1}^{t-1} \ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) + \sum_{j \in \mathcal{J}_{t-1}} \tilde{\ell}_f(\boldsymbol{x}_j)}{N} + \frac{\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{N q_t(i_t \mid i_{1:t-1}, X_N)}.$$

Equivalently, it can be written as

$$\hat{S}_t = \frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f(\boldsymbol{x}_n) + \frac{1}{N}\sum_{m=1}^{t-1} \left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right) + \frac{\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{N q_t(i_t \mid i_{1:t-1}, X_N)}.$$

This form shows that each round contributes a surrogate correction term

$$\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_t}),$$

which measures the discrepancy between the observed risk and the surrogate risk score on the selected sample. The signal uses this correction in an inverse-probability-weighted way so that it remains conditionally unbiased for R_N at each round.

To obtain the final estimator at horizon t, we aggregate the correction terms revealed over the first t rounds. Before observing the risk at round t, the horizon-t estimator has the predictable component

$$B_{t-1}^{(t)} = \frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f(\boldsymbol{x}_n) + \frac{1}{t}\sum_{m=1}^{t-1} u_{m,t} \left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right).$$

Relative to this predictable component, the inferential signal can be decomposed as

$$\hat{S}_t = B_{t-1}^{(t)} + \hat{Z}_t,$$

where

$$\hat{Z}_t = \left[\frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f(\boldsymbol{x}_n) + \frac{1}{N}\sum_{m=1}^{t-1} \left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right) - B_{t-1}^{(t)}\right] + \frac{\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{N q_t(i_t \mid i_{1:t-1}, X_N)}.$$

Here, $\hat{Z}_t$ is the one-step correction used to turn the predictable component into a conditionally unbiased signal for e-process construction.

For performance estimation, we use the same predictable component but aggregate the observed correction terms with horizon-dependent weights. After observing the risk at round t, the final estimator is

$$\hat{R}_t = B_{t-1}^{(t)} + \frac{1}{t} u_{t,t} \left(\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})\right).$$

Substituting the definition of $B_{t-1}^{(t)}$ yields

$$\hat{R}_t = \frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f(\boldsymbol{x}_n) + \frac{1}{t}\sum_{m=1}^{t} u_{m,t} \left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right),$$

where

$$u_{m,t} = 1 + \frac{N-t}{N-m}\left(\frac{1}{(N-m+1)q_m(i_m \mid i_{1:m-1}, X_N)} - 1\right).$$

Thus, the estimator in Eqn. (5) is obtained by aggregating the same surrogate correction terms that appear in the inferential signal. The signal $\hat{S}_t$ uses a one-step correction for valid e-process updates, whereas $\hat{R}_t$ aggregates the correction information to produce the final performance estimate. This construction also clarifies the connection to Horvitz–Thompson estimation [23]. Conditional on $\mathcal{F}_{t-1}$, the inverse-probability-weighted correction in $\hat{S}_t$ is an unbiased estimate of the normalized total residual over the remaining pool, so $\hat{S}_t$ adds a conditional Horvitz–Thompson-type correction to the surrogate-completed estimate. The estimator $\hat{R}_t$ also uses inverse-probability weighting, but it is not the standard Horvitz–Thompson estimator: its horizon-dependent weights $u_{m,t}$ follow the LURE construction for sequential, non-uniform sample selection [25].

B End-to-end implementation of CELEUS

This section presents the complete implementation of CELEUS, connecting the practical sample-selection strategies in Remarks 1 and 2 to the proposal probabilities, inferential signals, e-processes, confidence intervals, stopping rule, and final performance estimate. At each round, the acquisition scores, proposal distribution, scaling bounds, and betting fractions are determined before the newly selected target loss is observed. The oracle experiments use the true residual magnitudes as the sample-selection scores (Remark 6); Alg. 2 describes practical sampling with label-free acquisition scores.

At round t, let $s_t(j) \geq 0$ denote the label-free acquisition score assigned to each remaining sample $j \in \mathcal{J}_{t-1}$. For the residual-based strategies considered in this paper, the score takes the form

$$s_t(j) = \left|\ell_{\text{proxy},t}(\boldsymbol{x}_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|,$$

where $\ell_{\text{proxy},t}(\boldsymbol{x}_j)$ is a label-free proxy for the unknown target loss. The score may depend on information available before round t, including previously observed losses and surrogate information, but not on the unrevealed label y_j. Define

$$m_t = |\mathcal{J}_{t-1}| = N - t + 1, \qquad Z_t^{\text{score}} = \sum_{k \in \mathcal{J}_{t-1}} s_t(k).$$

We first normalize the acquisition scores according to

$$p_t(j) = \begin{cases} \dfrac{s_t(j)}{Z_t^{\text{score}}}, & Z_t^{\text{score}} > \delta_{\text{num}}, \\ \dfrac{1}{m_t}, & Z_t^{\text{score}} \leq \delta_{\text{num}}, \end{cases} \qquad j \in \mathcal{J}_{t-1}, \tag{7}$$

where $\delta_{\text{num}} > 0$ is a numerical tolerance, set to 10^{-12} in our implementation. Thus, if all scores are zero, or their sum is numerically negligible, the normalized distribution falls back to uniform sampling. The actual proposal distribution used to draw the next sample is

$$q_t(j \mid i_{1:t-1}, X_N) = \frac{\beta}{m_t} + (1-\beta)p_t(j), \qquad j \in \mathcal{J}_{t-1}. \tag{8}$$

This construction defines a probability distribution satisfying

$$q_t(j \mid i_{1:t-1}, X_N) \geq \frac{\beta}{N-t+1},$$

as required by Assumption 2. If all scores are zero, both p_t and q_t are uniform over the remaining pool. If only some scores are zero, the probability floor ensures that the corresponding samples remain selectable. The proposal in Eqn. (8) is the practical score-based sampling rule motivated by Theorem 4; it is not, in general, identical to the constrained oracle proposal q_t^*.

For numerical test inversion, let

$$\mathcal{G} = \{v_1, \ldots, v_K\} \subset [L, U]$$

be an equally spaced candidate grid containing L and U. For compactness in Alg. 2, we write

$$q_t(j) = q_t(j \mid i_{1:t-1}, X_N), \qquad z_t = \ell(f(\boldsymbol{x}_{i_t}), y_{i_t}), \qquad \tilde{\ell}_j = \tilde{\ell}_f(\boldsymbol{x}_j).$$

The predictable scaling bounds and betting fractions are selected according to Remarks 4 and 3, respectively. Algorithm 2 summarizes the complete sequential procedure.

Algorithm 2. End-to-end implementation of CELEUS

Require: Input pool $X_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, surrogate scores $\{\tilde{\ell}_j\}_{j=1}^N$, and acquisition-score rule s_t
Require: Loss bounds $[L, U]$, grid $\mathcal{G}$, error split $\alpha_1 + \alpha_2 = \alpha$, parameters β, θ, c, δ_{num}, and target width ϵ
Ensure: Stopping time τ, grid-based population interval $\widehat{C}_\tau$, performance estimate $\hat{R}_\tau$, and stopping status `reached`

$\mathcal{J}_0 \leftarrow \{1, \ldots, N\}$, $\mathcal{A}_0 \leftarrow \mathcal{G}$, $\tau \leftarrow N$, and `reached` $\leftarrow$ false
$\Delta_N \leftarrow (U - L)\sqrt{\dfrac{\log(2/\alpha_2)}{2N}}$
for each $v \in \mathcal{G}$ **do**
  $\mathcal{K}_{0,+}(v) \leftarrow 1$ and $\mathcal{K}_{0,-}(v) \leftarrow 1$
end for
for $t = 1, \ldots, N$ **do**
  $m_t \leftarrow N - t + 1$
  for each $j \in \mathcal{J}_{t-1}$ **do**
    Compute $s_t(j) \geq 0$ using only information available before round t
  end for
  $Z_t^{\text{score}} \leftarrow \sum_{k \in \mathcal{J}_{t-1}} s_t(k)$
  if $Z_t^{\text{score}} > \delta_{\text{num}}$ **then**
    $p_t(j) \leftarrow s_t(j)/Z_t^{\text{score}}$ for every $j \in \mathcal{J}_{t-1}$
  else
    $p_t(j) \leftarrow 1/m_t$ for every $j \in \mathcal{J}_{t-1}$
  end if
  $q_t(j) \leftarrow \beta/m_t + (1 - \beta)p_t(j)$ for every $j \in \mathcal{J}_{t-1}$
  Before observing z_t, choose predictable scaling bounds a_t, b_t according to Remark 4
  for each $v \in \mathcal{G}$ **do**
    $\bar{v}_t \leftarrow (v - a_t)/(b_t - a_t)$
    Compute predictable $\lambda_t^+(v)$ and $\lambda_t^-(v)$ according to Remark 3
  end for
  Draw $i_t \sim q_t(\cdot)$ over $\mathcal{J}_{t-1}$ without replacement
  Evaluate sample i_t and observe $z_t = \ell(f(\boldsymbol{x}_{i_t}), y_{i_t})$
  $q_t^{\text{obs}} \leftarrow q_t(i_t \mid i_{1:t-1}, X_N)$
  $\hat{S}_t \leftarrow \dfrac{1}{N}\left(\sum_{m=1}^{t-1} z_m + \sum_{j \in \mathcal{J}_{t-1}} \tilde{\ell}_j\right) + \dfrac{z_t - \tilde{\ell}_{i_t}}{N q_t^{\text{obs}}}$
  $\bar{S}_t \leftarrow (\hat{S}_t - a_t)/(b_t - a_t)$
  $\mathcal{J}_t \leftarrow \mathcal{J}_{t-1} \setminus \{i_t\}$
  for each $v \in \mathcal{G}$ **do**
    $d_t(v) \leftarrow \bar{S}_t - \bar{v}_t$
    $\mathcal{K}_{t,+}(v) \leftarrow \mathcal{K}_{t-1,+}(v)\left[1 + \lambda_t^+(v) d_t(v)\right]$
    $\mathcal{K}_{t,-}(v) \leftarrow \mathcal{K}_{t-1,-}(v)\left[1 - \lambda_t^-(v) d_t(v)\right]$
    $\mathcal{K}_t^{\pm}(v) \leftarrow \theta \mathcal{K}_{t,+}(v) + (1 - \theta)\mathcal{K}_{t,-}(v)$
  end for
  $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1} \cap \left\{v \in \mathcal{G} : \mathcal{K}_t^{\pm}(v) < 1/\alpha_1\right\}$
  if $\mathcal{A}_t \neq \varnothing$ **then**
    $\widehat{\mathcal{M}}_{t,\mathcal{G}}^{\pm} \leftarrow [\min \mathcal{A}_t, \max \mathcal{A}_t]$
    $\widehat{C}_t \leftarrow \left(\widehat{\mathcal{M}}_{t,\mathcal{G}}^{\pm} \oplus [-\Delta_N, \Delta_N]\right) \cap [L, U]$
  else
    $\widehat{\mathcal{M}}_{t,\mathcal{G}}^{\pm} \leftarrow [L, U]$ and $\widehat{C}_t \leftarrow [L, U]$
  end if
  if $\mathcal{A}_t \neq \varnothing$ and $\text{width}(\widehat{C}_t) \leq \epsilon$ **then**
    $\tau \leftarrow t$ and `reached` $\leftarrow$ true
    break
  end if
end for

if $\tau < N$ **then**
  for $m = 1, \ldots, \tau$ **do**
    $u_{m,\tau} \leftarrow 1 + \dfrac{N-\tau}{N-m}\left(\dfrac{1}{(N-m+1)q_m^{\mathrm{obs}}} - 1\right)$
  end for
  $\hat{R}_\tau \leftarrow \dfrac{1}{N}\sum_{n=1}^{N}\tilde{\ell}_n + \dfrac{1}{\tau}\sum_{m=1}^{\tau} u_{m,\tau}(z_m - \tilde{\ell}_{i_m})$
else
  $\hat{R}_N \leftarrow \dfrac{1}{N}\sum_{m=1}^{N} z_m = R_N$
end if
return $(\tau, \widehat{C}_\tau, \hat{R}_\tau, \texttt{reached})$

The distinction between p_t and q_t is essential. The distribution p_t only normalizes the acquisition scores, whereas q_t in Eqn. (8) is the distribution actually used to sample i_t. Consequently, the realized probability q_t^{obs} is used both in the inverse-probability correction in Eqn. (2) and in the weight $u_{m,\tau}$ in Eqn. (5). Using p_t, the unnormalized scores, or the oracle proposal in either correction would not correspond to the implemented sampling distribution.

For each grid candidate, the algorithm recursively updates the two one-sided capital processes and their hedged combination in Eqn. (3). The set $\mathcal{A}_t$ retains only the candidates that have remained unrejected at every round up to t, thereby implementing the running intersection in Eqn. (4) on the grid. The numerical interval $\widehat{\mathcal{M}}^{\pm}_{t,\mathcal{G}}$ spans the smallest and largest surviving grid points, after which $\widehat{C}_t$ applies the finite-pool-to-population correction. If no grid point survives, the full interval $[L, U]$ is used as a numerical safeguard, so an empty grid is not interpreted as attaining zero width. Theorems 5 and 6 concern exact inversion over the continuum $[L, U]$, while the grid provides its numerical implementation.

The procedure stops at the first round for which the reported interval reaches the target width ϵ. If this does not occur by round N, the procedure terminates when the pool is exhausted and reports the resulting interval with $\texttt{reached} = \mathrm{false}$ rather than declaring that the target precision has been attained. At $t = N$, only one sample remains and Eqn. (8) assigns it probability one, regardless of the acquisition scores. Consequently, $\hat{S}_N = R_N$. We also set $\hat{R}_N = R_N$ directly, avoiding the undefined expression obtained by substituting $m = t = N$ into $u_{m,t}$. For an earlier stopping time $\tau < N$, the final estimator uses the actual proposal probabilities stored at the corresponding sampling rounds. The confidence interval is the certification output of the procedure, while $\hat{R}_\tau$ is reported as an accompanying scalar performance estimate.

C Detailed Proofs for Our Theoretical Results

C.1 Assumptions

In this appendix, we collect the standing assumptions used in the theoretical analysis. The first assumption is an application-level boundedness condition on the evaluation scores. In most evaluation tasks, this condition is mild and can be satisfied by normalizing the metric or clipping the scores to a fixed range. The population-risk results additionally use the i.i.d. input-label pairs and fixed target model specified in the evaluation setup. The remaining assumptions are algorithmic conditions: they concern how the procedure selects samples, scales the inferential signals, and chooses the predictable fractions used in the e-process construction.

We distinguish the information available to the evaluator from the complete-pool information used in the proofs. The observed information after round t is

$$\mathcal{O}_t = \sigma\left(X_N, \mathcal{P}_0, (i_m, y_{i_m})_{m=1}^{t}\right).$$

The sequences $(\mathcal{O}_t)$ and $(\mathcal{F}_t)$ are filtrations, meaning that each records information accumulated up to a given round. Here, $\mathcal{F}_t$ is defined in Eqn. (1), and $\mathcal{O}_t \subseteq \mathcal{F}_t$. Practical sampling probabilities, scaling bounds, and betting fractions are predictable with respect to $(\mathcal{O}_t)$: their round-t values are determined using only $\mathcal{O}_{t-1}$. The surrogate scores in $\mathcal{P}_0$ remain fixed throughout evaluation.

Each index is drawn using fresh randomization independent of $\mathcal{F}_{t-1}$, including the complete pool, preprocessing information, and past selections. Consequently, for every $j \in \mathcal{J}_{t-1}$,

$$\mathbb{P}(i_t = j \mid \mathcal{F}_{t-1}) = \mathbb{P}(i_t = j \mid \mathcal{O}_{t-1}) = q_t(j \mid i_{1:t-1}, X_N). \tag{9}$$

Thus, conditioning on the complete pool fixes its losses without changing the probabilities used to draw the next index. Conditional expectations and variances given $\mathcal{F}_0$ concern only sample-selection randomness, whereas unconditional quantities also average over the pool and preprocessing information.

Assumption 1. *The evaluation risk score and the surrogate risk score are uniformly bounded: there exist constants $L, U \in \mathbb{R}$ such that*

$$L \le \ell(f(\boldsymbol{x}), y) \le U, \qquad L \le \tilde{\ell}_f(\boldsymbol{x}) \le U.$$

Assumption 2. *There exists a constant $\beta \in (0, 1]$[8] such that*

$$\min_{j \in \mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, X_N) \ge \frac{\beta}{N - t + 1}, \qquad \forall N \in \mathbb{Z}_+,\ \forall 1 \le t \le N.$$

Assumption 2 ensures that every remaining sample has a nonzero chance of being selected. This prevents inverse-probability weights from becoming ill-defined or arbitrarily large.

Assumption 3. *For every $N \in \mathbb{Z}_+$ and $1 \le t \le N$, the scaling bounds a_t and b_t are $\mathcal{O}_{t-1}$-measurable and hence $\mathcal{F}_{t-1}$-measurable. Moreover, there exist deterministic constants $0 < \underline{\Delta} \le \overline{\Delta} < \infty$, independent of N and t, such that, almost surely, $a_t \le \hat{S}_t \le b_t$, $[L, U] \subseteq [a_t, b_t]$, $\underline{\Delta} \le b_t - a_t \le \overline{\Delta}$.*

Assumption 3 is an algorithmic condition on the scaling step. Predictability preserves the conditional mean property after scaling. The containment conditions ensure that both the inferential signal and every candidate value in $[L, U]$ are mapped to $[0, 1]$, while the uniform width bounds make the transformation non-degenerate and provide the scale control required for the confidence-width bound.

Assumption 4. *Fix a constant $c \in (0, 1)$. There exist deterministic constants $0 < \underline{\gamma} \le \overline{\gamma} < \infty$, independent of N and t, such that, for every $N \in \mathbb{Z}_+$, the evaluator chooses real-valued sequences predictable with respect to $(\mathcal{O}_t)$ $\big(\tilde{\lambda}_t^+\big)_{1 \le t \le N}$ and $\big(\tilde{\lambda}_t^-\big)_{1 \le t \le N}$ that do not depend on the candidate value υ and satisfy, almost surely,*

$$\frac{\underline{\gamma}}{\sqrt{t \log(t+1)}} \le |\tilde{\lambda}_t^{\pm}| \le \frac{\overline{\gamma}}{\sqrt{t \log(t+1)}}, \qquad \forall 1 \le t \le N.$$

For each $\upsilon \in [L, U]$, define

$$\lambda_t^+(\upsilon) = |\tilde{\lambda}_t^+| \wedge \frac{1}{\bar{\upsilon}_t + c}, \qquad \lambda_t^-(\upsilon) = |\tilde{\lambda}_t^-| \wedge \frac{1}{1 - \bar{\upsilon}_t + c}.$$

Assumption 4 specifies the predictable fractions used in the e-process. Predictability and clipping ensure the validity of the capital processes, while the candidate-independent envelope condition provides the uniform control over $\upsilon \in [L, U]$ required for the confidence-width bound. The constants are uniform in N and t, as required for the joint asymptotic result.

C.2 Proofs of Theorem 1

Theorem 1′. *Under the population sampling setup and Assumptions 1-2, for each fixed $1 \le t < N$, the estimator $\hat{R}_t$ as defined in Eqn. (5) satisfies $\mathbb{E}[\hat{R}_t] = R$ with*

$$\mathrm{Var}[\hat{R}_t] = \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} w_m \psi_m \mid \mathcal{F}_0\right]\right],$$

[8] More generally, the sampling floor may use a time-varying sequence $\{\beta_t\}_{t \ge 1}$ with $\beta_t \in (0, 1]$. If β_t is predictable and there exists a deterministic constant $\underline{\beta} > 0$ such that $\beta_t \ge \underline{\beta}$ for all t, all theoretical guarantees continue to hold. The round-t variance-optimal proposal uses β_t, while the uniform variance and confidence-width bounds use $\underline{\beta}$. This changes only the constants in the bounds, not the validity or convergence rates. We use a time-independent β for notational simplicity.

where

$$w_m = \frac{N(N-t)}{(N-m)(N-m+1)}, \tag{10}$$

and

$$\psi_m = \frac{\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})}{N q_m(i_m \mid i_{1:m-1}, X_N)} + \frac{1}{N}\sum_{s=1}^{m-1}\left(\ell\left(f(\boldsymbol{x}_{i_s}), y_{i_s}\right) - \tilde{\ell}_f\left(\boldsymbol{x}_{i_s}\right)\right). \tag{11}$$

At $t = N$, $\hat{R}_N = R_N$ has expectation R and variance $\mathrm{Var}(\ell(f(\boldsymbol{x}), y))/N$.

We present the detailed proof of Theorem 1 as follows.

Proof. At $t = N$, the convention $\hat{R}_N = R_N$ gives the stated expectation and variance directly under the population sampling setup. We therefore prove the remaining case $1 \le t < N$. Under Assumptions 1–2, we first prove unbiasedness and then compute the variance, starting from Eqn. (5):

$$\mathbb{E}[\hat{R}_t] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right] + \mathbb{E}\Bigg[\underbrace{\frac{1}{t}\sum_{m=1}^{t} u_{m,t}\left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right)}_{\text{(a)}}\Bigg], \tag{12}$$

where

$$u_{m,t} = 1 + \frac{N-t}{N-m}\left(\frac{1}{(N-m+1)q_m(i_m \mid i_{1:m-1}, X_N)} - 1\right).$$

By Assumption 2, the proposal probabilities are strictly positive on the remaining pool, so the inverse-probability terms above are well defined.

The key step is to show that the residual-correction term (a) exactly matches, in expectation, the average residual over the entire pool. To this end, we invoke the unbiasedness results for adaptive sampling without replacement from Farquhar et al. [25], which imply that (a) can be rewritten as

$$\text{(a)} = \frac{1}{t}\sum_{m=1}^{t} w_m \psi_m, \qquad w_m = \frac{N(N-t)}{(N-m)(N-m+1)},$$

and

$$\psi_m = \frac{\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})}{N q_m(i_m \mid i_{1:m-1}, X_N)} + \frac{1}{N}\sum_{s=1}^{m-1}\left(\ell\left(f(\boldsymbol{x}_{i_s}), y_{i_s}\right) - \tilde{\ell}_f\left(\boldsymbol{x}_{i_s}\right)\right).$$

We now compute $\mathbb{E}[\psi_m]$ via the tower property, which averages a conditional expectation over the information being conditioned on. Conditioning on $\mathcal{F}_{m-1}$ fixes the complete pool, preprocessing information, and the first $m-1$ selections, giving

$$\begin{aligned}
\mathbb{E}[\psi_m] &= \mathbb{E}_{D_N, \mathcal{P}_0, i_{1:m-1}}\left[\mathbb{E}_m\left[\psi_m \mid \mathcal{F}_{m-1}\right]\right] \\
&= \underbrace{\mathbb{E}_{D_N, \mathcal{P}_0, i_{1:m-1}}\left[\mathbb{E}_m\left[\frac{\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})}{N q_m(i_m \mid i_{1:m-1}, X_N)} \mid \mathcal{F}_{m-1}\right]\right]}_{\text{(b)}} \\
&\quad + \mathbb{E}_{D_N, \mathcal{P}_0, i_{1:m-1}}\left[\frac{1}{N}\sum_{s=1}^{m-1}\left(\ell\left(f(\boldsymbol{x}_{i_s}), y_{i_s}\right) - \tilde{\ell}_f\left(\boldsymbol{x}_{i_s}\right)\right)\right].
\end{aligned}$$

For (b), the conditional expectation over the m-th draw expands as a finite sum over the remaining unlabeled indices. Specifically, given $\mathcal{F}_{m-1}$ and the sampling law in Eqn. (9), the next selected index takes values $j \in \mathcal{J}_{m-1}$ with conditional probability $q_m(j \mid i_{1:m-1}, X_N)$, hence

$$\begin{aligned}
\text{(b)} &= \mathbb{E}_{D_N, \mathcal{P}_0, i_{1:m-1}}\left[\sum_{j \in \mathcal{J}_{m-1}} q_m(j \mid i_{1:m-1}, X_N)\frac{\ell\left(f(\boldsymbol{x}_j), y_j\right) - \tilde{\ell}_f\left(\boldsymbol{x}_j\right)}{N q_m(j \mid i_{1:m-1}, X_N)}\right] \\
&= \mathbb{E}_{D_N, \mathcal{P}_0, i_{1:m-1}}\left[\sum_{j \in \mathcal{J}_{m-1}} \frac{1}{N}\left(\ell\left(f(\boldsymbol{x}_j), y_j\right) - \tilde{\ell}_f\left(\boldsymbol{x}_j\right)\right)\right].
\end{aligned}$$

Combining the two terms yields

$$\begin{aligned}\mathbb{E}[\psi_m] &= \mathbb{E}_{D_N,\mathcal{P}_0,i_{1:m-1}}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n),y_n\right)-\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right]\\ &= \mathbb{E}_{D_N,\mathcal{P}_0}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n),y_n\right)-\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right],\end{aligned}$$

where the last equality holds because the expression inside the brackets no longer depends on the selection history $i_{1:m-1}$.

Substituting this back into the representation of ⓐ, we have

$$\begin{aligned}\mathbb{E}[\text{ⓐ}] &= \frac{1}{t}\sum_{m=1}^{t} w_m\mathbb{E}[\psi_m]\\ &= \frac{1}{t}\sum_{m=1}^{t} w_m\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n),y_n\right)-\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right]\\ &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n),y_n\right)-\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right]\times\frac{1}{t}\sum_{m=1}^{t} w_m\\ &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n),y_n\right)-\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right].\end{aligned}$$

The last equality holds since

$$\begin{aligned}\sum_{m=1}^{t} w_m &= N(N-t)\sum_{m=1}^{t}\frac{1}{(N-m)(N-m+1)}\\ &= N(N-t)\sum_{m=1}^{t}\left(\frac{1}{N-m}-\frac{1}{N-m+1}\right)=t.\end{aligned}$$

Substituting $\mathbb{E}[\text{ⓐ}]$ back into Eqn. (12), we obtain

$$\begin{aligned}\mathbb{E}[\hat{R}_t] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right]+\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n),y_n\right)-\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right]\\ &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\ell\left(f(\boldsymbol{x}_n),y_n\right)\right]\\ &= R,\end{aligned}$$

which proves the unbiasedness of $\hat{R}_t$.

We next investigate the variance of $\hat{R}_t$ under Assumption 1. By the law of total variance, we have

$$\mathrm{Var}[\hat{R}_t]=\mathbb{E}\Big[\underbrace{\mathrm{Var}\left[\hat{R}_t\mid\mathcal{F}_0\right]}_{\text{ⓒ}}\Big]+\mathrm{Var}\Big[\underbrace{\mathbb{E}\left[\hat{R}_t\mid\mathcal{F}_0\right]}_{\text{ⓓ}}\Big]. \tag{13}$$

We first investigate the term ⓓ. The same unbiasedness calculation holds with the complete pool and preprocessing information fixed, so conditional on $\mathcal{F}_0$ we have

$$\mathbb{E}\left[\hat{R}_t\mid\mathcal{F}_0\right]=\frac{1}{N}\sum_{n=1}^{N}\ell\left(f(\boldsymbol{x}_n),y_n\right).$$

Since the input-label pairs are i.i.d. and f is fixed, the pool losses are i.i.d. Therefore,

$$\mathrm{Var}[\text{ⓓ}]=\mathrm{Var}\left[\frac{1}{N}\sum_{n=1}^{N}\ell\left(f(\boldsymbol{x}_n),y_n\right)\right]=\frac{\mathrm{Var}(\ell(f(\boldsymbol{x}),y))}{N}.$$

We now consider the term ⓒ. Since the first term in $\hat{R}_t$ is measurable with respect to $\mathcal{F}_0$, it does not contribute to the conditional variance. Hence,

$$\text{ⓒ} = \text{Var}\left[\underbrace{\frac{1}{t}\sum_{m=1}^{t} w_m \psi_m}_{\text{ⓐ}} \mid \mathcal{F}_0\right].$$

Write

$$\mathbb{E}\left[\text{ⓐ} \mid \mathcal{F}_0\right] = \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)$$

for the average residual over the pool. Then

$$\begin{aligned}
\text{ⓒ} &= \mathbb{E}\left[\text{ⓐ}^2 \mid \mathcal{F}_0\right] - \left(\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)^2 \\
&= \mathbb{E}\left[\left(\frac{1}{t}\sum_{m=1}^{t} w_m\psi_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)^2 \mid \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\left(\frac{1}{t}\sum_{m=1}^{t} w_m\left(\psi_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)\right)^2 \mid \mathcal{F}_0\right] \\
&= \frac{1}{t^2}\sum_{m=1}^{t}\sum_{s=1}^{t} w_m w_s \mathbb{E}\left[\underbrace{\left(\psi_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)}_{\text{ⓔ}}\right. \\
&\qquad \left.\cdot\underbrace{\left(\psi_s - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)}_{\text{ⓕ}} \mid \mathcal{F}_0\right].
\end{aligned}$$

For $s < m$, since ψ_s is measurable with respect to $\mathcal{F}_{m-1}$, we have

$$\begin{aligned}
\mathbb{E}\left[\psi_m \cdot \text{ⓕ} \mid \mathcal{F}_0\right] &= \mathbb{E}\left[\mathbb{E}\left[\psi_m \cdot \text{ⓕ} \mid \mathcal{F}_{m-1}\right] \mid \mathcal{F}_0\right] \\
&= \mathbb{E}\left[\text{ⓕ} \cdot \mathbb{E}\left[\psi_m \mid \mathcal{F}_{m-1}\right] \mid \mathcal{F}_0\right].
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}\left[\psi_m \mid \mathcal{F}_{m-1}\right] &= \sum_{j \in \mathcal{J}_{m-1}} q_m(j \mid i_{1:m-1}, X_N)\frac{\ell\left(f(\boldsymbol{x}_j), y_j\right) - \tilde{\ell}_f\left(\boldsymbol{x}_j\right)}{N q_m(j \mid i_{1:m-1}, X_N)} \\
&\quad + \frac{1}{N}\sum_{k=1}^{m-1}\left(\ell\left(f(\boldsymbol{x}_{i_k}), y_{i_k}\right) - \tilde{\ell}_f\left(\boldsymbol{x}_{i_k}\right)\right) \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right).
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\psi_m \cdot \text{ⓕ} \mid \mathcal{F}_0\right] = \mathbb{E}\left[\text{ⓕ} \cdot \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right) \mid \mathcal{F}_0\right].$$

Hence, when $s < m$, the conditional expectation of the cross term ⓔ × ⓕ given $\mathcal{F}_0$ is equal to zero. By symmetry, the same conclusion holds when $m < s$. Therefore, only the diagonal terms remain, and

$$\text{ⓒ} \quad = \quad \frac{1}{t^2}\sum_{m=1}^{t} w_m^2 \mathbb{E}\left[\left(\psi_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)^2 \mid \mathcal{F}_0\right].$$

Using the tower property once more, we can rewrite the remaining term as

$$\begin{aligned} & \mathbb{E}\left[\left(\psi_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)^2 \mid \mathcal{F}_0\right] \\ = \quad & \mathbb{E}\Big[\mathrm{Var}\left[\psi_m \mid \mathcal{F}_{m-1}\right] \mid \mathcal{F}_0\Big]. \end{aligned}$$

Therefore,

$$\mathrm{Var}\left[\hat{R}_t \mid \mathcal{F}_0\right] = \frac{1}{t^2}\sum_{m=1}^{t} w_m^2 \mathbb{E}\Big[\mathrm{Var}\left[\psi_m \mid \mathcal{F}_{m-1}\right] \mid \mathcal{F}_0\Big].$$

Substituting this and the expression for Var[ⓓ] back into Eqn. (13), we obtain

$$\begin{aligned} \mathrm{Var}[\hat{R}_t] \quad &= \quad \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\frac{1}{t^2}\sum_{m=1}^{t} w_m^2 \mathbb{E}\Big[\mathrm{Var}\left[\psi_m \mid \mathcal{F}_{m-1}\right] \mid \mathcal{F}_0\Big]\right] \\ &= \quad \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} w_m \psi_m \mid \mathcal{F}_0\right]\right], \end{aligned}$$

which completes the proof. □

C.3 Proofs of Theorem 2

We present the detailed proof of Theorem 2 as follows.

Proof. Under Assumption 1-2, Theorem 1 establishes the unbiasedness of the estimator $\hat{R}_t$ at each round t. Hence

$$\mathbb{E}\left[\left(\hat{R}_t - R\right)^2\right] = \mathrm{Var}[\hat{R}_t].$$

To establish convergence, it suffices to show that the variance vanishes along the sequence $\{t_N\}_{N\geq 1}$.

Step 1: Control the variance of the residual-correction term. Recall from Theorem 1 that, for $t < N$,

$$\mathrm{Var}[\hat{R}_t] = \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} w_m \psi_m \mid \mathcal{F}_0\right]\right],$$

where

$$w_m = \frac{N(N-t)}{(N-m)(N-m+1)},$$

and

$$\psi_m = \frac{\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})}{N q_m(i_m \mid i_{1:m-1}, X_N)} + \frac{1}{N}\sum_{s=1}^{m-1}\left(\ell\left(f(\boldsymbol{x}_{i_s}), y_{i_s}\right) - \tilde{\ell}_f\left(\boldsymbol{x}_{i_s}\right)\right).$$

Moreover, by the variance analysis in the proof of Theorem 1,

$$\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} w_m \psi_m \mid \mathcal{F}_0\right] = \frac{1}{t^2}\sum_{m=1}^{t} w_m^2 \mathbb{E}\Big[\mathrm{Var}\left[\psi_m \mid \mathcal{F}_{m-1}\right] \mid \mathcal{F}_0\Big].$$

We now bound the conditional variance of ψ_m. Since the second term in ψ_m is measurable with respect to $\mathcal{F}_{m-1}$, it does not contribute to the conditional variance. Hence,

$$\begin{aligned}
&\operatorname{Var}\left[\psi_m \mid \mathcal{F}_{m-1}\right] \\
&= \operatorname{Var}\left[\frac{\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})}{N q_m(i_m \mid i_{1:m-1}, X_N)} \mid \mathcal{F}_{m-1}\right] \\
&\leq \mathbb{E}\left[\left(\frac{\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})}{N q_m(i_m \mid i_{1:m-1}, X_N)}\right)^2 \mid \mathcal{F}_{m-1}\right] \\
&= \sum_{j \in \mathcal{J}_{m-1}} q_m(j \mid i_{1:m-1}, X_N) \frac{\left(\ell\left(f(\boldsymbol{x}_j), y_j\right) - \tilde{\ell}_f\left(\boldsymbol{x}_j\right)\right)^2}{N^2 q_m^2(j \mid i_{1:m-1}, X_N)} \\
&= \sum_{j \in \mathcal{J}_{m-1}} \frac{\left(\ell\left(f(\boldsymbol{x}_j), y_j\right) - \tilde{\ell}_f\left(\boldsymbol{x}_j\right)\right)^2}{N^2 q_m(j \mid i_{1:m-1}, X_N)}.
\end{aligned}$$

Under Assumption 1, $L \leq \ell, \tilde{\ell}_f \leq U$, so

$$\left|\ell\left(f(\boldsymbol{x}_j), y_j\right) - \tilde{\ell}_f\left(\boldsymbol{x}_j\right)\right| \leq U - L, \qquad j \in \mathcal{J}_{m-1}.$$

Using the positivity assumption (Assumption 2)

$$q_m(j \mid i_{1:m-1}, X_N) \geq \beta/(N - m + 1), \qquad j \in \mathcal{J}_{m-1},$$

we obtain

$$\frac{1}{q_m(j \mid i_{1:m-1}, X_N)} \leq \frac{N - m + 1}{\beta}.$$

Therefore,

$$\begin{aligned}
\operatorname{Var}\left[\psi_m \mid \mathcal{F}_{m-1}\right] &\leq \sum_{j \in \mathcal{J}_{m-1}} \frac{(U-L)^2}{N^2 q_m(j \mid i_{1:m-1}, X_N)} \\
&\leq \frac{(U-L)^2}{N^2} \sum_{j \in \mathcal{J}_{m-1}} \frac{N - m + 1}{\beta} \\
&\leq \frac{(U-L)^2}{\beta} \frac{|\mathcal{J}_{m-1}|}{N} \\
&= \frac{(U-L)^2}{\beta} \frac{N - m + 1}{N}.
\end{aligned}$$

Substituting this into the above display yields

$$\mathbb{E}\left[\operatorname{Var}\left[\frac{1}{t}\sum_{m=1}^{t} w_m \psi_m \mid \mathcal{F}_0\right]\right] \leq \frac{(U-L)^2}{\beta} \frac{1}{t^2} \sum_{m=1}^{t} w_m^2 \frac{N - m + 1}{N}. \tag{14}$$

We next bound the coefficient sum. For $t < N$,

$$\begin{aligned}
&\frac{1}{t^2}\sum_{m=1}^{t} w_m^2 \frac{N-m+1}{N} \\
&= \frac{1}{t^2}\sum_{m=1}^{t}\left(\frac{N(N-t)}{(N-m)(N-m+1)}\right)^2 \frac{N-m+1}{N} \\
&= \frac{N(N-t)^2}{t^2}\sum_{m=1}^{t}\frac{1}{(N-m)^2(N-m+1)} \\
&\le \frac{N(N-t)^2}{t^2}\sum_{m=1}^{t}\frac{1}{(N-t)(N-m)(N-m+1)} \\
&= \frac{N(N-t)}{t^2}\sum_{m=1}^{t}\frac{1}{(N-m)(N-m+1)} \\
&= \frac{N(N-t)}{t^2}\sum_{m=1}^{t}\left(\frac{1}{N-m}-\frac{1}{N-m+1}\right) \\
&= \frac{N(N-t)}{t^2}\left(\frac{1}{N-t}-\frac{1}{N}\right) \\
&= \frac{1}{t}.
\end{aligned}$$

Combining this with Eqn. (14), we obtain

$$\mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} w_m\psi_m \mid \mathcal{F}_0\right]\right] \le \frac{(U-L)^2}{\beta t}. \tag{15}$$

Step 2: Conclude the L^2 convergence. If $t < N$, combining Theorem 1 with Eqn. (15) gives

$$\mathbb{E}\left[\left(\hat{R}_t - R\right)^2\right] = \mathrm{Var}[\hat{R}_t] \le \frac{\mathrm{Var}\left(\ell(f(\boldsymbol{x}), y)\right)}{N} + \frac{(U-L)^2}{\beta t}.$$

If $t = N$, then $\hat{R}_N = R_N$ pathwise, so

$$\mathbb{E}\left[\left(\hat{R}_N - R\right)^2\right] = \mathbb{E}\left[(R_N - R)^2\right] = \frac{\mathrm{Var}\left(\ell(f(\boldsymbol{x}), y)\right)}{N}.$$

Therefore, for all $1 \le t \le N$,

$$\mathbb{E}\left[\left(\hat{R}_t - R\right)^2\right] \le \frac{\mathrm{Var}\left(\ell(f(\boldsymbol{x}), y)\right)}{N} + \frac{(U-L)^2}{\beta t}.$$

Finally, if $t = t_N \le N$ is any sequence such that $N \to \infty$ and

$$\frac{N - t_N + 1}{N} \to 0,$$

then $t_N/N \to 1$, and hence $t_N \to \infty$. Therefore both terms on the right-hand side vanish, and we obtain

$$\mathbb{E}\left[\left(\hat{R}_{t_N} - R\right)^2\right] \to 0.$$

Hence $\hat{R}_{t_N} \to R$ in L^2, and therefore in probability. □

C.4 Proofs of Theorem 3

We present the detailed proof of Theorem 3 as follows.

Proof. Recall the signal in Eqn. (2):

$$\hat{S}_t = \frac{\sum_{m=1}^{t-1} \ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) + \sum_{j \in \mathcal{J}_{t-1}} \tilde{\ell}_f\left(\boldsymbol{x}_j\right)}{N} + \frac{\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f\left(\boldsymbol{x}_{i_t}\right)}{N q_t(i_t \mid i_{1:t-1}, X_N)} .$$

By Assumption 2, the proposal probabilities are strictly positive on the remaining pool, so the inverse-probability terms above are well defined.

We first rewrite $\hat{S}_t$ around the finite-pool risk

$$R_N = \frac{1}{N} \sum_{j=1}^{N} \ell(f(\boldsymbol{x}_j), y_j) = \frac{\sum_{m=1}^{t-1} \ell(f(\boldsymbol{x}_{i_m}), y_{i_m}) + \sum_{j \in \mathcal{J}_{t-1}} \ell(f(\boldsymbol{x}_j), y_j)}{N} ,$$

subtracting this identity from Eqn. (2) yields

$$\hat{S}_t = R_N + \underbrace{\frac{1}{N} \left(\frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{q_t(i_t \mid i_{1:t-1}, X_N)} - \sum_{j \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j) \Big) \right)}_{\textcircled{a}} . \tag{16}$$

Under Assumption 1-2, we first prove the conditional unbiasedness. Since $\mathcal{F}_{t-1}$ includes the complete pool D_N, R_N is measurable with respect to $\mathcal{F}_{t-1}$. Thus, taking conditional expectation on both sides of Eqn. (16) gives

$$\mathbb{E}[\hat{S}_t \mid \mathcal{F}_{t-1}] = R_N + \mathbb{E}[\textcircled{a} \mid \mathcal{F}_{t-1}].$$

We now compute $\mathbb{E}[\textcircled{a} \mid \mathcal{F}_{t-1}]$. Conditional on $\mathcal{F}_{t-1}$, the next queried index i_t takes values in $\mathcal{J}_{t-1}$ with probabilities $q_t(j \mid i_{1:t-1}, X_N)$. Therefore,

$$\begin{aligned}
\mathbb{E}[\textcircled{a} \mid \mathcal{F}_{t-1}] &= \frac{1}{N} \sum_{j \in \mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, X_N) \frac{\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)}{q_t(j \mid i_{1:t-1}, X_N)} \\
&\quad - \frac{1}{N} \sum_{j \in \mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, X_N) \sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k) \Big) \\
&= \frac{1}{N} \sum_{j \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j) \Big) \\
&\quad - \frac{1}{N} \left(\sum_{j \in \mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, X_N) \right) \left(\sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k) \Big) \right) \\
&= 0,
\end{aligned}$$

where the last equality uses

$$\sum_{j \in \mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, X_N) = 1.$$

Hence

$$\mathbb{E}[\hat{S}_t \mid \mathcal{F}_{t-1}] = R_N,$$

which proves the conditional unbiasedness.

We next investigate the conditional variance of $\hat{S}_t$. Since R_N is $\mathcal{F}_{t-1}$-measurable, by Eqn. (16),

$$\mathrm{Var}[\hat{S}_t \mid \mathcal{F}_{t-1}] = \mathrm{Var}[\textcircled{a} \mid \mathcal{F}_{t-1}].$$

Moreover, since $\mathbb{E}[\textcircled{a} \mid \mathcal{F}_{t-1}] = 0$,

$$\mathrm{Var}[\textcircled{a} \mid \mathcal{F}_{t-1}] = \mathbb{E}[\textcircled{a}^2 \mid \mathcal{F}_{t-1}].$$

Expanding ⓐ^{2} from Eqn. (16), we obtain

$$
\begin{aligned}
\text{ⓐ}^2 \;=\; & \underbrace{\frac{\Big(\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})\Big)^2}{N^2 q_t^2(i_t \mid i_{1:t-1}, X_N)}}_{\text{ⓑ}} \\
& \underbrace{- \frac{2}{N^2} \cdot \frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{q_t(i_t \mid i_{1:t-1}, X_N)} \left(\sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k)\Big) \right)}_{\text{ⓒ}} \\
& + \underbrace{\frac{1}{N^2} \left(\sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k)\Big) \right)^2}_{\text{ⓓ}}.
\end{aligned}
$$

Then, we have

$$
\mathbb{E}[\text{ⓐ}^2 \mid \mathcal{F}_{t-1}] = \mathbb{E}[\text{ⓑ} \mid \mathcal{F}_{t-1}] - \mathbb{E}[\text{ⓒ} \mid \mathcal{F}_{t-1}] + \mathbb{E}[\text{ⓓ} \mid \mathcal{F}_{t-1}].
$$

For the first term,

$$
\begin{aligned}
& \mathbb{E}[\text{ⓑ} \mid \mathcal{F}_{t-1}] \\
& \quad = \sum_{j \in \mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, X_N) \frac{\Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\Big)^2}{N^2 q_t^2(j \mid i_{1:t-1}, X_N)} \\
& \quad = \sum_{j \in \mathcal{J}_{t-1}} \frac{\Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\Big)^2}{N^2 q_t(j \mid i_{1:t-1}, X_N)}.
\end{aligned}
$$

For the second term, in a similar manner, we have

$$
\begin{aligned}
& \mathbb{E}[\text{ⓒ} \mid \mathcal{F}_{t-1}] \\
& \quad = \sum_{j \in \mathcal{J}_{t-1}} \frac{2\Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\Big)\Big(\sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k)\Big)\Big)}{q_t^{-1}(j \mid i_{1:t-1}, X_N) N^2 q_t(j \mid i_{1:t-1}, X_N)} \\
& \quad = \sum_{j \in \mathcal{J}_{t-1}} \frac{2\Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\Big)\Big(\sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k)\Big)\Big)}{N^2} \\
& \quad = 2\left(\frac{\sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k)\Big)}{N} \right)^2.
\end{aligned}
$$

Finally, since ⓓ is $\mathcal{F}_{t-1}$-measurable, we have

$$
\mathbb{E}[\text{ⓓ} \mid \mathcal{F}_{t-1}] = \left(\frac{\sum_{k \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_k), y_k) - \tilde{\ell}_f(\boldsymbol{x}_k)\Big)}{N} \right)^2.
$$

Therefore,

$$\mathrm{Var}[\hat{S}_t \mid \mathcal{F}_{t-1}] = \sum_{j \in \mathcal{J}_{t-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{N^2 q_t(j \mid i_{1:t-1}, X_N)} - \left(\frac{\sum_{j \in \mathcal{J}_{t-1}} \left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)}{N} \right)^2,$$

which completes the proof. □

C.5 Proofs of Theorem 4

We present the detailed proof of Theorem 4 as follows.

Proof. Fix t. Throughout this proof, we condition on $\mathcal{F}_{t-1}$ from Eqn. (1), which fixes the complete pool, preprocessing information, and previous selections. The oracle proposal may use the unobserved pool losses; it is a mathematical benchmark and need not be determined by the observed information $\mathcal{O}_{t-1}$. By Assumption 1, both $\ell(f(\boldsymbol{x}_j), y_j)$ and $\tilde{\ell}_f(\boldsymbol{x}_j)$ are uniformly bounded, hence all quantities

$$\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j), \qquad j \in \mathcal{J}_{t-1},$$

are finite. Under the conditioning above, the pool losses are fixed, and $\tilde{\ell}_f$ is measurable with respect to $\mathcal{F}_{t-1}$. Therefore, these residuals are fixed conditional on $\mathcal{F}_{t-1}$.

If $t = N$, there is only one remaining sample, so the only feasible proposal assigns it probability one. With the terminal convention $\hat{R}_N = R_N$, both $\hat{R}_N$ and $\hat{S}_N$ equal R_N, and the claimed optimality follows immediately. We therefore consider $t < N$ below.

We first consider the inferential signal $\hat{S}_t$. Recall that

$$\hat{S}_t = \frac{\sum_{m=1}^{t-1} \ell(f(\boldsymbol{x}_{i_m}), y_{i_m}) + \sum_{j \in \mathcal{J}_{t-1}} \tilde{\ell}_f(\boldsymbol{x}_j)}{N} + \frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{N q_t(i_t \mid i_{1:t-1}, X_N)}.$$

By Assumption 2, the proposal probabilities are strictly positive on the remaining pool, so the inverse-probability terms above are well defined.

We also recall that

$$R_N = \frac{\sum_{m=1}^{t-1} \ell(f(\boldsymbol{x}_{i_m}), y_{i_m}) + \sum_{j \in \mathcal{J}_{t-1}} \ell(f(\boldsymbol{x}_j), y_j)}{N}.$$

Hence

$$\hat{S}_t - R_N = \frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{N q_t(i_t \mid i_{1:t-1}, X_N)} - \frac{1}{N} \sum_{j \in \mathcal{J}_{t-1}} \left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right). \tag{17}$$

By Theorem 3,

$$\mathbb{E}[\hat{S}_t \mid \mathcal{F}_{t-1}] = R_N,$$

and

$$\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1}) = \sum_{j \in \mathcal{J}_{t-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{N^2 q_t(j \mid i_{1:t-1}, X_N)} - \left(\frac{\sum_{j \in \mathcal{J}_{t-1}} \left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)}{N} \right)^2.$$

The second term does not depend on the proposal distribution, so minimizing $\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1})$ is equivalent to minimizing

$$\sum_{j \in \mathcal{J}_{t-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{q_t(j \mid i_{1:t-1}, X_N)}$$

over proposal distributions satisfying

$$\sum_{j \in \mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, X_N) = 1, \qquad q_t(j \mid i_{1:t-1}, X_N) \geq \frac{\beta}{N-t+1}.$$

Suppose first that the residuals are not all zero. For this optimization, write

$$M = N - t + 1, \qquad r_j = \ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j), \qquad q_j = q_t(j \mid i_{1:t-1}, X_N).$$

If $\beta = 1$, the only feasible proposal is uniform:

$$q_j = \frac{1}{M}, \qquad j \in \mathcal{J}_{t-1}.$$

This agrees with the claimed formula for any

$$\eta_t \geq M \max_{j \in \mathcal{J}_{t-1}} |r_j|.$$

Now consider $0 < \beta < 1$. The objective is convex, the feasible set is compact, and the uniform proposal strictly satisfies the lower-bound constraints. Thus, the Karush–Kuhn–Tucker conditions are necessary and sufficient for optimality.

The Lagrangian is

$$\mathcal{L}(q, \zeta, \mu) = \sum_{j \in \mathcal{J}_{t-1}} \frac{r_j^2}{q_j} + \zeta \left(\sum_{j \in \mathcal{J}_{t-1}} q_j - 1 \right) + \sum_{j \in \mathcal{J}_{t-1}} \mu_j \left(\frac{\beta}{M} - q_j \right),$$

where $\zeta \in \mathbb{R}$ and $\mu_j \geq 0$. Stationarity and complementary slackness give

$$-\frac{r_j^2}{q_j^2} + \zeta - \mu_j = 0, \qquad \mu_j \left(q_j - \frac{\beta}{M} \right) = 0.$$

Since at least one residual is nonzero, there exists some j such that $r_j \neq 0$. For this j, stationarity gives

$$\zeta = \frac{r_j^2}{q_j^2} + \mu_j > 0.$$

Therefore, $\sqrt{\zeta}$ is well defined and positive.

If $q_j > \beta/M$, complementary slackness gives $\mu_j = 0$. The stationarity condition then becomes

$$\zeta = \frac{r_j^2}{q_j^2}.$$

Since $q_j > 0$, taking the positive square root gives

$$\sqrt{\zeta} = \frac{|r_j|}{q_j},$$

or equivalently,

$$q_j = \frac{|r_j|}{\sqrt{\zeta}}.$$

Thus, the common normalizing denominator of the residual-proportional probabilities is $\sqrt{\zeta}$. We denote this denominator by

$$\eta_t = \sqrt{\zeta}.$$

Consequently,

$$q_j = \frac{|r_j|}{\eta_t}$$

whenever $q_j > \beta/M$.

If $q_j = \beta/M$, stationarity and $\mu_j \geq 0$ imply

$$\zeta - \frac{r_j^2}{q_j^2} \geq 0,$$

and hence

$$\frac{|r_j|}{\eta_t} \leq \frac{\beta}{M}.$$

Therefore, the optimal proposal must satisfy

$$q_t^*(j \mid i_{1:t-1}, D_N) = \max\left\{\frac{\beta}{M}, \frac{|r_j|}{\eta_t}\right\}.$$

We next derive η_t explicitly. Let the residual magnitudes be ordered as

$$|r|_{(1)} \geq |r|_{(2)} \geq \cdots \geq |r|_{(M)},$$

where $\{|r|_{(1)}, \ldots, |r|_{(M)}\} = \{|r_j| : j \in \mathcal{J}_{t-1}\}$, and set $|r|_{(M+1)} = 0$. For each $k \in \{1, \ldots, M\}$, define

$$\eta_{t,k} = \frac{\sum_{s=1}^{k} |r|_{(s)}}{1 - (M-k)\beta/M}.$$

Let

$$k_t = \max\left\{k \in \{1, \ldots, M\} : |r|_{(k)} > \frac{\beta}{M}\eta_{t,k}\right\}.$$

This set is nonempty because $|r|_{(1)} > 0$ and $\beta < 1$. The normalizing constant is

$$\eta_t = \eta_{t,k_t} = \frac{\sum_{s=1}^{k_t} |r|_{(s)}}{1 - (M-k_t)\beta/M}.$$

By the ordering of the residual magnitudes and the maximality of k_t,

$$|r|_{(s)} > \frac{\beta}{M}\eta_t \quad \text{for } s \leq k_t, \qquad |r|_{(s)} \leq \frac{\beta}{M}\eta_t \quad \text{for } s > k_t.$$

Indeed, if $k_t < M$, the maximality of k_t gives $|r|_{(k_t+1)} \leq \beta\eta_{t,k_t+1}/M$, which, by the definition of $\eta_{t,k}$, is equivalent to $|r|_{(k_t+1)} \leq \beta\eta_{t,k_t}/M = \beta\eta_t/M$. Thus, the first k_t samples receive probabilities proportional to their residual magnitudes, while the remaining samples receive the lower-bound probability β/M. It follows that

$$\begin{aligned}
\sum_{j \in \mathcal{J}_{t-1}} q_t^*(j \mid i_{1:t-1}, D_N) &= \sum_{s=1}^{k_t} \frac{|r|_{(s)}}{\eta_t} + (M-k_t)\frac{\beta}{M} \\
&= \frac{\sum_{s=1}^{k_t} |r|_{(s)}}{\eta_t} + (M-k_t)\frac{\beta}{M} \\
&= 1.
\end{aligned}$$

Hence this explicit value of η_t normalizes the proposal distribution.

The resulting proposal satisfies the KKT conditions with

$$\zeta = \eta_t^2, \qquad \mu_j = \eta_t^2 - \frac{r_j^2}{(q_j^*)^2}.$$

If $q_j^* > \beta/M$, then $q_j^* = |r_j|/\eta_t$, so $\mu_j = 0$. If $q_j^* = \beta/M$, then $|r_j|/\eta_t \leq \beta/M$, so $\mu_j \geq 0$. Thus, stationarity, primal feasibility, dual feasibility, and complementary slackness all hold, and q_t^* is optimal.

Since the objective is strictly convex in every coordinate with $r_j \neq 0$, and every coordinate with $r_j = 0$ must receive the lower-bound probability β/M, the optimal proposal distribution is unique.

We next consider the estimator $\hat{R}_t$. Recall that

$$\hat{R}_t = \frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f(\boldsymbol{x}_n) + \frac{1}{t}\sum_{m=1}^{t} u_{m,t}\left(\ell(f(\boldsymbol{x}_{i_m}), y_{i_m}) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right).$$

Keeping all previous sampling decisions and probabilities fixed, only the $m = t$ term depends on the current proposal. Therefore, we can write

$$\begin{aligned}\hat{R}_t \quad = \quad & \underbrace{\frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f(\boldsymbol{x}_n) + \frac{1}{t}\sum_{m=1}^{t-1} u_{m,t}\left(\ell(f(\boldsymbol{x}_{i_m}), y_{i_m}) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right)}_{\text{ⓐ}} \\ & + \frac{1}{t} u_{t,t}\left(\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})\right).\end{aligned}$$

Since $t < N$, the definition of $u_{m,t}$ gives

$$u_{t,t} = 1 + \frac{N-t}{N-t}\left(\frac{1}{(N-t+1)q_t(i_t \mid i_{1:t-1}, X_N)} - 1\right) = \frac{1}{(N-t+1)q_t(i_t \mid i_{1:t-1}, X_N)}.$$

Hence

$$\hat{R}_t = \text{ⓐ} + \frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{t(N-t+1)q_t(i_t \mid i_{1:t-1}, X_N)}. \tag{18}$$

Since ⓐ is measurable with respect to $\mathcal{F}_{t-1}$, conditioning on $\mathcal{F}_{t-1}$ gives

$$\begin{aligned}\mathrm{Var}(\hat{R}_t \mid \mathcal{F}_{t-1}) = & \sum_{j \in \mathcal{J}_{t-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{t^2(N-t+1)^2 q_t(j \mid i_{1:t-1}, X_N)} \\ & - \left(\frac{\sum_{j \in \mathcal{J}_{t-1}}\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)}{t(N-t+1)}\right)^2.\end{aligned}$$

Again, the second term does not depend on the proposal distribution, so minimizing $\mathrm{Var}(\hat{R}_t \mid \mathcal{F}_{t-1})$ is equivalent to minimizing

$$\sum_{j \in \mathcal{J}_{t-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{q_t(j \mid i_{1:t-1}, X_N)}$$

subject to the same probability constraints. This is exactly the optimization problem above. Therefore, the same proposal distribution

$$q_t^*(j \mid i_{1:t-1}, D_N) = \max\left\{\frac{\beta}{N-t+1}, \frac{\left|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|}{\eta_t}\right\}$$

also minimizes $\mathrm{Var}(\hat{R}_t \mid \mathcal{F}_{t-1})$.

Furthermore, with all sampling policies before round t held fixed, the law of total variance gives

$$\mathrm{Var}(\hat{R}_t) = \mathrm{Var}\big(\mathbb{E}[\hat{R}_t \mid \mathcal{F}_{t-1}]\big) + \mathbb{E}\big[\mathrm{Var}(\hat{R}_t \mid \mathcal{F}_{t-1})\big].$$

From Eqn. (18),

$$\mathbb{E}[\hat{R}_t \mid \mathcal{F}_{t-1}] = \text{ⓐ} + \frac{1}{t(N-t+1)}\sum_{j \in \mathcal{J}_{t-1}}\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right),$$

which does not depend on the current proposal. Consequently, applying the conditional minimizer at each history also minimizes $\mathrm{Var}(\hat{R}_t)$ with respect to the round-t proposal, while the earlier policies remain fixed.

In contrast, suppose that

$$\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j) = 0, \qquad \forall j \in \mathcal{J}_{t-1}.$$

For the signal $\hat{S}_t$, Eqn. (17) immediately gives

$$\hat{S}_t - R_N = 0,$$

and hence

$$\hat{S}_t = R_N.$$

Therefore,

$$\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1}) = 0,$$

so the conditional variance is minimized by every proposal distribution satisfying Assumption 2.

For the estimator $\hat{R}_t$, since $i_t \in \mathcal{J}_{t-1}$, Eqn. (18) yields

$$\hat{R}_t = ⓐ + \frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{t(N-t+1)q_t(i_t \mid i_{1:t-1}, X_N)} = ⓐ,$$

which is $\mathcal{F}_{t-1}$-measurable and does not depend on the current proposal. Hence

$$\mathrm{Var}(\hat{R}_t \mid \mathcal{F}_{t-1}) = 0.$$

Thus, this conditional variance is minimized by every proposal distribution satisfying Assumption 2.

In summary, when the residuals are not all zero, q_t^* uniquely minimizes both conditional variances among proposal distributions satisfying Assumption 2. When all residuals are zero, both conditional variances vanish, and every feasible proposal distribution is optimal. This completes the proof. □

C.6 Proofs of Theorem 5

Before presenting the detailed proof of Theorem 5, we first recall the notion of a test martingale and a time-uniform concentration inequality for bounded martingale differences.

Definition 3. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ be a filtered probability space, where $(\mathcal{F}_t)_{t\geq 0}$ is a filtration satisfying $\mathcal{F}_t \subseteq \mathcal{F}$ for all t. A nonnegative, $(\mathcal{F}_t)$-adapted process $(Y_t)_{t\geq 0}$ is called a* test martingale *if $Y_0 = 1$ and, for all $1 \leq t \leq N$,*

$$\mathbb{E}_{\mathbb{P}}[Y_t \mid \mathcal{F}_{t-1}] = Y_{t-1} \quad a.s.$$

Theorem 7. *[37, Corollary 8.2] Let $(Y_t)_{1\leq t\leq N}$ be a bounded martingale difference sequence with respect to a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$. Assume that there exists $b > 0$ such that $|Y_t| \leq b$ almost surely for all $t \in \mathbb{N}$. For each $t \in \mathbb{N}$, define*

$$V_t = \sum_{i=1}^{t} \mathbb{E}[Y_i^2 \mid \mathcal{F}_{i-1}], \qquad W_t = \sum_{i=1}^{t} Y_i.$$

Then, for any $T \geq 1$, $\delta > 0$, and $\rho > 1$,

$$\Pr\left(\exists t \in [T] : W_t > \sqrt{2\rho V_t \delta} + \frac{4b\delta}{3}\right) \leq \left(1 + \left\lceil \log_\rho\left(\frac{2T}{\delta}\right)\right\rceil_+\right) \exp(-\delta).$$

We further introduce the following elementary inequality.

Theorem 8. *For any $r > 0$ and any $x \geq -1/(1+r)$,*

$$\log(1+x) \geq x - \frac{x^2}{2(1 - 1/(1+r))}.$$

Proof of Theorem 5. Fix $N \in \mathbb{Z}_+$. The proof consists of two steps.

- First, we establish the anytime-valid coverage of the inverted hedged capital process, showing that $\left(\mathcal{M}_t^{\pm}\right)_{1\le t\le N}$ is a $(1-\alpha_1)$ confidence sequence for R_N.
- Second, we derive a time-uniform width bound for $\mathcal{M}_t^{\pm}$ and use it to establish shrinkage in the joint asymptotic regime.

Step 1: Anytime-validity via test martingales and inversion.

We use $\mathcal{F}_t$ from Eqn. (1). First condition on $\mathcal{F}_0$, so the complete pool and preprocessing information are fixed and only sample-selection randomness remains. By Theorem 3,

$$\mathbb{E}[\hat{S}_t \mid \mathcal{F}_{t-1}] = R_N.$$

Recall that we use the predictable scaling

$$\bar{S}_t = \frac{\hat{S}_t - a_t}{b_t - a_t}, \qquad \bar{\upsilon}_t = \frac{\upsilon - a_t}{b_t - a_t}, \qquad \bar{R}_{N,t} = \frac{R_N - a_t}{b_t - a_t},$$

where a_t and b_t are $\mathcal{F}_{t-1}$-measurable. By Assumption 1, $R_N \in [L, U]$. Moreover, Assumption 3 gives $b_t - a_t > 0$, $\hat{S}_t \in [a_t, b_t]$ almost surely, and $[L, U] \subseteq [a_t, b_t]$. Hence

$$\bar{S}_t, \bar{\upsilon}_t, \bar{R}_{N,t} \in [0, 1].$$

Since the scaling is affine,

$$\mathbb{E}[\bar{S}_t \mid \mathcal{F}_{t-1}] = \bar{R}_{N,t}.$$

In particular, under the null hypothesis $\mathcal{H}_0(\upsilon) : R_N = \upsilon$,

$$\mathbb{E}[\bar{S}_t \mid \mathcal{F}_{t-1}] = \bar{\upsilon}_t.$$

Recall the one-sided capital processes

$$\mathcal{K}_t^{+}(\upsilon) = \prod_{i=1}^{t} \left(1 + \lambda_i^{+}(\upsilon)(\bar{S}_i - \bar{\upsilon}_i)\right),$$

and

$$\mathcal{K}_t^{-}(\upsilon) = \prod_{i=1}^{t} \left(1 - \lambda_i^{-}(\upsilon)(\bar{S}_i - \bar{\upsilon}_i)\right).$$

By Assumption 4,

$$\lambda_t^{+}(\upsilon) = |\tilde{\lambda}_t^{+}| \wedge \frac{1}{\bar{\upsilon}_t + c}, \qquad \lambda_t^{-}(\upsilon) = |\tilde{\lambda}_t^{-}| \wedge \frac{1}{1 - \bar{\upsilon}_t + c}.$$

Since $\bar{S}_t, \bar{\upsilon}_t \in [0, 1]$,

$$1 + \lambda_t^{+}(\upsilon)(\bar{S}_t - \bar{\upsilon}_t) \ge 1 - \lambda_t^{+}(\upsilon)\bar{\upsilon}_t \ge \frac{c}{\bar{\upsilon}_t + c} > 0,$$

and similarly,

$$1 - \lambda_t^{-}(\upsilon)(\bar{S}_t - \bar{\upsilon}_t) \ge 1 - \lambda_t^{-}(\upsilon)(1 - \bar{\upsilon}_t) \ge \frac{c}{1 - \bar{\upsilon}_t + c} > 0.$$

Thus both capital processes are nonnegative.

Because $\lambda_t^{+}(\upsilon)$ is $\mathcal{F}_{t-1}$-measurable, under $\mathcal{H}_0(\upsilon)$,

$$\begin{aligned}\mathbb{E}\left[\mathcal{K}_t^{+}(\upsilon) \mid \mathcal{F}_{t-1}\right] &= \mathcal{K}_{t-1}^{+}(\upsilon)\mathbb{E}\left[1 + \lambda_t^{+}(\upsilon)(\bar{S}_t - \bar{\upsilon}_t) \mid \mathcal{F}_{t-1}\right] \\ &= \mathcal{K}_{t-1}^{+}(\upsilon).\end{aligned}$$

Therefore, $\{\mathcal{K}_t^{+}(\upsilon)\}_{1\le t\le N}$ is a nonnegative martingale under $\mathcal{H}_0(\upsilon)$. The same argument gives

$$\mathbb{E}\left[\mathcal{K}_t^{-}(\upsilon) \mid \mathcal{F}_{t-1}\right] = \mathcal{K}_{t-1}^{-}(\upsilon),$$

so $\{\mathcal{K}_t^{-}(\upsilon)\}_{1\le t\le N}$ is also a nonnegative martingale under $\mathcal{H}_0(\upsilon)$.

Hence, for $\theta \in (0,1)$,

$$\mathcal{K}_t^{\pm}(\upsilon) = \theta \mathcal{K}_t^{+}(\upsilon) + (1-\theta)\mathcal{K}_t^{-}(\upsilon)$$

is a nonnegative martingale under $\mathcal{H}_0(\upsilon)$ with $\mathcal{K}_0^{\pm}(\upsilon) = 1$.

For each fixed complete pool and preprocessing information, take $\upsilon = R_N$. By Ville's inequality [38],

$$\Pr\left(\exists 1 \le t \le N : \mathcal{K}_t^{\pm}(R_N) \ge \frac{1}{\alpha_1} \mid \mathcal{F}_0\right) \le \alpha_1.$$

Therefore, defining

$$\mathcal{B}_t^{\pm} = \left\{\upsilon \in [L,U] : \mathcal{K}_t^{\pm}(\upsilon) < \frac{1}{\alpha_1}\right\},$$

we have

$$\Pr\left(\exists 1 \le t \le N : R_N \notin \mathcal{B}_t^{\pm} \mid \mathcal{F}_0\right) \le \alpha_1.$$

Recall that

$$\mathcal{M}_t^{\pm} = \bigcap_{i \le t} \mathcal{B}_i^{\pm}.$$

For every t,

$$R_N \notin \mathcal{M}_t^{\pm} \iff \exists i \le t : R_N \notin \mathcal{B}_i^{\pm}.$$

Consequently,

$$\left\{\exists 1 \le t \le N : R_N \notin \mathcal{M}_t^{\pm}\right\} = \left\{\exists 1 \le t \le N : R_N \notin \mathcal{B}_t^{\pm}\right\}.$$

It follows that

$$\Pr\left(\exists 1 \le t \le N : R_N \notin \mathcal{M}_t^{\pm} \mid \mathcal{F}_0\right) \le \alpha_1.$$

Taking expectations over $\mathcal{F}_0$ gives the same bound unconditionally. Thus, $\left(\mathcal{M}_t^{\pm}\right)_{1 \le t \le N}$ is a $(1-\alpha_1)$ anytime-valid confidence sequence for R_N.

Step 2: Shrinkage and width bound via a time-uniform martingale concentration argument.

Fix an auxiliary failure probability $\omega \in (0,1)$ and any $\rho > 1$. Under Assumption 4,

$$\frac{\underline{\gamma}}{\sqrt{t\log(t+1)}} \le |\tilde{\lambda}_t^{\pm}| \le \frac{\overline{\gamma}}{\sqrt{t\log(t+1)}}.$$

Define

$$t_c = \min\left\{t \in \mathbb{Z}_+ : t\log(t+1) \ge ((1+c)\overline{\gamma})^2\right\}, \qquad t_0 = \max\{2t_c, 8\}. \tag{19}$$

For every $i \ge t_c$,

$$|\tilde{\lambda}_i^{\pm}| \le \frac{1}{1+c}.$$

Since $\bar{v}_i \in [0,1]$,

$$\frac{1}{\bar{v}_i + c} \ge \frac{1}{1+c}, \qquad \frac{1}{1-\bar{v}_i + c} \ge \frac{1}{1+c}.$$

Therefore, simultaneously for all $\upsilon \in [L,U]$ and $i \ge t_c$,

$$\lambda_i^{+}(\upsilon) = |\tilde{\lambda}_i^{+}|, \qquad \lambda_i^{-}(\upsilon) = |\tilde{\lambda}_i^{-}|. \tag{20}$$

Let

$$Z_i = \bar{S}_i - \bar{R}_{N,i}.$$

Then

$$\mathbb{E}[Z_i \mid \mathcal{F}_{i-1}] = 0, \qquad |Z_i| \le 1.$$

Define

$$Y_i^{+} = \begin{cases} 0, & i < t_c, \\ |\tilde{\lambda}_i^{+}| Z_i, & i \ge t_c, \end{cases} \qquad Y_i^{-} = \begin{cases} 0, & i < t_c, \\ |\tilde{\lambda}_i^{-}| Z_i, & i \ge t_c, \end{cases}$$

and

$$W_t^{+} = \sum_{i=1}^{t} Y_i^{+}, \qquad W_t^{-} = \sum_{i=1}^{t} Y_i^{-}.$$

Because $|\tilde{\lambda}_i^{\pm}|$ is predictable and does not depend on v, $(Y_i^+)_{1\le i\le N}$ and $(Y_i^-)_{1\le i\le N}$ are candidate-independent martingale difference sequences. Moreover,

$$|Y_i^{\pm}| \le \frac{1}{c} \quad \text{almost surely.}$$

Their predictable quadratic variations satisfy

$$\begin{aligned} V_t^{\pm} &= \sum_{i=1}^{t} \mathbb{E}\left[(Y_i^{\pm})^2 \mid \mathcal{F}_{i-1}\right] \\ &\le \sum_{i=1}^{t} \frac{\overline{\gamma}^2}{i\log(i+1)} \mathrm{Var}\left(\bar{S}_i \mid \mathcal{F}_{i-1}\right) \\ &= \sum_{i=1}^{t} \frac{\overline{\gamma}^2}{i\log(i+1)(b_i-a_i)^2} \mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) \\ &\le \frac{\overline{\gamma}^2}{\underline{\Delta}^2} \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}. \end{aligned} \tag{21}$$

We now apply Theorem 7 using a stitching argument. For $j = 0, 1, \ldots, \lceil \log_2 N \rceil$, define

$$T_j = \min\{2^j, N\}, \qquad \omega_j = \frac{3\omega}{\pi^2(j+1)^2},$$

and

$$\delta_j = \max\left\{1, \log\left(\frac{\left(1 + \lceil \log_\rho(2^{j+1})\rceil_+\right)\pi^2(j+1)^2}{3\omega}\right)\right\}.$$

Define

$$\mathcal{E}_j^+ = \left\{\forall t \le T_j : W_t^+ \ge -\sqrt{2\rho V_t^+ \delta_j} - \frac{4\delta_j}{3c}\right\},$$

and

$$\mathcal{E}_j^- = \left\{\forall t \le T_j : W_t^- \le \sqrt{2\rho V_t^- \delta_j} + \frac{4\delta_j}{3c}\right\}.$$

Applying Theorem 7 to $-W_t^+$ and W_t^- gives

$$\Pr\left((\mathcal{E}_j^+)^c\right) \le \omega_j, \qquad \Pr\left((\mathcal{E}_j^-)^c\right) \le \omega_j.$$

Indeed, $\delta_j \ge 1$ and $T_j \le 2^j$ imply

$$\left\lceil \log_\rho\left(\frac{2T_j}{\delta_j}\right)\right\rceil_+ \le \left\lceil \log_\rho(2^{j+1})\right\rceil_+,$$

so the conclusion follows from the definition of δ_j.

Let

$$\mathcal{E}^+ = \bigcap_{j=0}^{\lceil \log_2 N\rceil} \mathcal{E}_j^+, \qquad \mathcal{E}^- = \bigcap_{j=0}^{\lceil \log_2 N\rceil} \mathcal{E}_j^-.$$

Since

$$\sum_{j=0}^{\infty} \omega_j = \frac{\omega}{2},$$

a union bound gives

$$\Pr(\mathcal{E}^+) \ge 1 - \frac{\omega}{2}, \qquad \Pr(\mathcal{E}^-) \ge 1 - \frac{\omega}{2},$$

and hence

$$\Pr(\mathcal{E}^+ \cap \mathcal{E}^-) \ge 1 - \omega.$$

On $\mathcal{E}^+$,

$$W_t^+ \geq -\sqrt{2\rho V_t^+ \delta_{\lceil \log_2 t \rceil}} - \frac{4\delta_{\lceil \log_2 t \rceil}}{3c},$$

and on $\mathcal{E}^-$,

$$W_t^- \leq \sqrt{2\rho V_t^- \delta_{\lceil \log_2 t \rceil}} + \frac{4\delta_{\lceil \log_2 t \rceil}}{3c},$$

simultaneously for all $1 \leq t \leq N$. Since these events are constructed from candidate-independent martingales, they hold simultaneously for every $\upsilon \in [L, U]$.

We first consider $\upsilon < R_N$. For every i,

$$\bar{S}_i - \bar{v}_i = Z_i + \frac{R_N - \upsilon}{b_i - a_i}, \qquad \left(\bar{S}_i - \bar{v}_i\right)^2 \leq 1.$$

Moreover,

$$\lambda_i^+(\upsilon)(\bar{S}_i - \bar{v}_i) \geq -\frac{1}{1+c},$$

so Theorem 8 gives

$$\begin{aligned}\log \mathcal{K}_t^+(\upsilon) \geq (R_N - \upsilon) \sum_{i=1}^t \frac{\lambda_i^+(\upsilon)}{b_i - a_i} + \sum_{i=1}^t \lambda_i^+(\upsilon) Z_i \\ - \frac{1}{2(1 - 1/(1+c))} \sum_{i=1}^t \left(\lambda_i^+(\upsilon)\right)^2. \end{aligned} \tag{22}$$

Let

$$G_c = \frac{t_c - 1}{c}.$$

For $i < t_c$, $\lambda_i^+(\upsilon) \leq 1/c$ and $Z_i \geq -1$. Together with Eqn. (20), this gives

$$\sum_{i=1}^t \lambda_i^+(\upsilon) Z_i \geq -G_c + W_t^+.$$

Moreover, for every $2t_c \leq t \leq N$,

$$\begin{aligned}\sum_{i=1}^t \frac{\lambda_i^+(\upsilon)}{b_i - a_i} &\geq \sum_{i=\lfloor t/2 \rfloor}^t \frac{|\tilde{\lambda}_i^+|}{b_i - a_i} \\ &\geq \frac{\underline{\gamma}\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}}. \end{aligned} \tag{23}$$

Also,

$$\begin{aligned}\sum_{i=1}^t \left(\lambda_i^+(\upsilon)\right)^2 &\leq \overline{\gamma}^2 \sum_{i=1}^t \frac{1}{i \log(i+1)} \\ &\leq \overline{\gamma}^2 \left(G_0 + \log\log(t+1)\right) \end{aligned} \tag{24}$$

for some absolute constant $G_0 > 0$. Here we use

$$\sum_{i=2}^t \frac{1}{i \log i} \leq \frac{1}{2\log 2} + \int_2^t \frac{dx}{x \log x}.$$

Combining Eqn. (21)–Eqn. (24), we obtain on $\mathcal{E}^+$, simultaneously for all $\upsilon < R_N$ and $2t_c \leq t \leq N$,

$$\begin{aligned}\log \mathcal{K}_t^+(\upsilon) \geq (R_N - \upsilon) \frac{\underline{\gamma}\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}} - G_c \\ - \frac{\overline{\gamma}^2 \left(G_0 + \log\log(t+1)\right)}{2(1 - 1/(1+c))} \\ - \sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2} \sum_{i=1}^t \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \delta_{\lceil \log_2 t \rceil}} - \frac{4\delta_{\lceil \log_2 t \rceil}}{3c}. \end{aligned} \tag{25}$$

The stitching construction gives

$$\delta_{\lceil \log_2 t \rceil} = O\left(\log(1/\omega) + \log\log t\right).$$

To make the resulting bound explicit, let

$$Q_t = \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}.$$

For the fixed ω and all $t \geq t_0$, there exists a constant $G_1 > 0$ such that

$$\delta_{\lceil \log_2 t \rceil} \leq G_1 \log\log(t+1).$$

Since $t_0 \geq 8$, $\log\log(t+1)$ is bounded away from zero, and hence, after enlarging G_1 if necessary,

$$1 + \log\log(t+1) + \sqrt{Q_t \log\log(t+1)} \leq G_1\sqrt{\log\log(t+1)\left(\log\log(t+1) + Q_t\right)}.$$

Thus, the fixed terms, the quadratic term, and the Freedman term in Eqn. (25), together with the target threshold $\log(1/(\theta\alpha_1))$, are dominated by a constant multiple of $\sqrt{\log\log(t+1)(\log\log(t+1)+Q_t)}$. Therefore, there exists a constant $G_2 > 0$ such that every $\upsilon < R_N$ satisfying

$$R_N - \upsilon \geq G_2\sqrt{\frac{\log(t+1)\log\log(t+1)}{t}\left(\log\log(t+1) + \sum_{i=1}^{t}\frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}$$

also satisfies

$$\mathcal{K}_t^+(\upsilon) \geq \frac{1}{\theta\alpha_1}.$$

Since

$$\mathcal{K}_t^{\pm}(\upsilon) \geq \theta\mathcal{K}_t^+(\upsilon),$$

every such candidate is excluded from $\mathcal{B}_t^{\pm}$. Hence, on $\mathcal{E}^+$,

$$\mathcal{B}_t^{\pm} \subseteq \left[R_N - G_2\sqrt{\frac{\log(t+1)\log\log(t+1)}{t}\left(\log\log(t+1) + \sum_{i=1}^{t}\frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}, U\right]$$

simultaneously for all $t_0 \leq t \leq N$.

We next consider $\upsilon > R_N$. Using the same decomposition and applying the elementary inequality to $-\lambda_i^-(\upsilon)(\bar{S}_i - \bar{\upsilon}_i)$ gives

$$\begin{aligned}\log\mathcal{K}_t^-(\upsilon) \geq (\upsilon - R_N)\sum_{i=1}^{t}\frac{\lambda_i^-(\upsilon)}{b_i - a_i} - \sum_{i=1}^{t}\lambda_i^-(\upsilon)Z_i \\ - \frac{1}{2(1-1/(1+c))}\sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2. \end{aligned} \tag{26}$$

For $i < t_c$, $\lambda_i^-(\upsilon) \leq 1/c$ and $-Z_i \geq -1$. Hence, as above,

$$-\sum_{i=1}^{t}\lambda_i^-(\upsilon)Z_i \geq -G_c - W_t^-,$$

and, for every $2t_c \leq t \leq N$,

$$\sum_{i=1}^{t}\frac{\lambda_i^-(\upsilon)}{b_i - a_i} \geq \frac{\underline{\gamma}\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}},$$

while

$$\sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2 \leq \overline{\gamma}^2\left(G_0 + \log\log(t+1)\right).$$

It follows that, on $\mathcal{E}^-$, simultaneously for all $\upsilon > R_N$ and $t_0 \le t \le N$,

$$\begin{aligned}\log \mathcal{K}_t^-(\upsilon) \ge (\upsilon - R_N)\frac{\gamma\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}} - G_c \\ - \frac{\overline{\gamma}^2\left(G_0 + \log\log(t+1)\right)}{2(1-1/(1+c))} \\ - \sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2}\sum_{i=1}^{t}\frac{\operatorname{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\delta_{\lceil \log_2 t\rceil}} - \frac{4\delta_{\lceil \log_2 t\rceil}}{3c}.\end{aligned} \tag{27}$$

Therefore, there exists a constant $G_4 > 0$ such that every $\upsilon > R_N$ satisfying

$$\upsilon - R_N \ge G_4\sqrt{\frac{\log(t+1)\log\log(t+1)}{t}\left(\log\log(t+1) + \sum_{i=1}^{t}\frac{\operatorname{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}$$

also satisfies

$$\mathcal{K}_t^-(\upsilon) \ge \frac{1}{(1-\theta)\alpha_1}.$$

Since

$$\mathcal{K}_t^{\pm}(\upsilon) \ge (1-\theta)\mathcal{K}_t^-(\upsilon),$$

every such candidate is excluded from $\mathcal{B}_t^{\pm}$. Hence, on $\mathcal{E}^-$,

$$\mathcal{B}_t^{\pm} \subseteq \left[L, R_N + G_4\sqrt{\frac{\log(t+1)\log\log(t+1)}{t}\left(\log\log(t+1) + \sum_{i=1}^{t}\frac{\operatorname{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}\right]$$

simultaneously for all $t_0 \le t \le N$.

On $\mathcal{E}^+ \cap \mathcal{E}^-$, which has probability at least $1-\omega$, the preceding two bounds hold simultaneously. Assigning width zero to the empty set, there exists a constant $G_5 > 0$ such that

$$\operatorname{width}\left(\mathcal{B}_t^{\pm}\right) \le G_5\sqrt{\frac{\log(t+1)\log\log(t+1)}{t}\left(\log\log(t+1) + \sum_{i=1}^{t}\frac{\operatorname{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}$$

simultaneously for all $t_0 \le t \le N$. Since

$$\mathcal{M}_t^{\pm} = \bigcap_{i \le t}\mathcal{B}_i^{\pm} \subseteq \mathcal{B}_t^{\pm},$$

the same width bound holds for $\mathcal{M}_t^{\pm}$.

For every $t \ge t_0$,

$$\log(t+1) \le 2\log t, \qquad \log\log(t+1) \le 2\log\log t.$$

Consequently, after enlarging the constant if necessary, with probability at least $1-\omega$,

$$\operatorname{width}\left(\mathcal{M}_t^{\pm}\right) \le G_6\sqrt{\frac{\log t\log\log t}{t}\left(\log\log t + \sum_{i=1}^{t}\frac{\operatorname{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)},$$

simultaneously for all $t_0 \le t \le N$.

The constant G_6 can be chosen independently of N and t.

Finally, Theorem 3 and Assumptions 1 and 2 imply

$$\begin{aligned}\operatorname{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) &\le \frac{1}{N^2}\sum_{j \in \mathcal{J}_{i-1}}\frac{(U-L)^2}{q_i(j \mid i_{1:i-1}, X_N)} \\ &\le \frac{(U-L)^2(N-i+1)^2}{\beta N^2} \\ &\le \frac{(U-L)^2}{\beta}.\end{aligned}$$

Therefore,

$$\sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} = O(\log\log t), \tag{28}$$

where the implied constant is independent of N.

Writing $\mathcal{M}_{t,N}^{\pm}$ to make the dependence on N explicit, consider any integer sequence $\{t_N\}_{N\geq 1}$ satisfying $1 \leq t_N \leq N$ and $t_N \to \infty$ as $N \to \infty$. Then $t_N \geq t_0$ for all sufficiently large N. Since $\omega \in (0,1)$ was arbitrary, the preceding time-uniform bound and Eqn. (28) imply that, for some constant $G_7 > 0$ independent of N,

$$\Pr\left(\mathrm{width}\left(\mathcal{M}_{t_N,N}^{\pm}\right) > G_7 \log\log t_N \sqrt{\frac{\log t_N}{t_N}}\right) \leq \omega$$

for all sufficiently large N. Consequently,

$$\mathrm{width}\left(\mathcal{M}_{t_N,N}^{\pm}\right) = O_{\mathbb{P}}\left(\log\log t_N \sqrt{\frac{\log t_N}{t_N}}\right).$$

Since

$$\log\log t_N \sqrt{\frac{\log t_N}{t_N}} \to 0,$$

we conclude that

$$\mathrm{width}\left(\mathcal{M}_{t_N,N}^{\pm}\right) \xrightarrow{\mathbb{P}} 0 \qquad \text{as } N \to \infty.$$

This completes the proof. □

C.7 Proofs of Theorem 6

We present the detailed proof of Theorem 6 as follows.

Proof. Fix $N \in \mathbb{Z}_+$. We decompose the proof into two high-probability events: one for the sequential uncertainty in estimating R_N from adaptive labels, and one for the finite-pool approximation error $|R_N - R|$.

Let

$$\mathcal{E}_1 = \left\{\forall 1 \leq t \leq N,\ R_N \in \mathcal{M}_t^{\pm}\right\}, \qquad \mathcal{E}_2 = \left\{|R_N - R| \leq \Delta_N(\alpha_2)\right\}.$$

By Theorem 5, $\left(\mathcal{M}_t^{\pm}\right)_{1\leq t\leq N}$ is a $(1-\alpha_1)$ anytime-valid confidence sequence for R_N, hence

$$\Pr\left(\mathcal{E}_1\right) = \Pr\left(\forall 1 \leq t \leq N,\ R_N \in \mathcal{M}_t^{\pm}\right) \geq 1 - \alpha_1.$$

Next, by Assumption 1, we have $\ell\left(f(\boldsymbol{x}), y\right) \in [L, U]$. Since the input-label pairs are i.i.d. and the target model is fixed, the pool losses are i.i.d. Hence, Hoeffding's inequality gives, for any $r > 0$,

$$\Pr\left(|R_N - R| > r\right) \leq 2\exp\left(-\frac{2Nr^2}{(U-L)^2}\right).$$

Choosing

$$\Delta_N(\alpha_2) = (U-L)\sqrt{\frac{\log\left(2/\alpha_2\right)}{2N}}$$

yields

$$\Pr\left(\mathcal{E}_2\right) \geq 1 - \alpha_2.$$

Assume $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Fix any $1 \leq t \leq N$. From $\mathcal{E}_1$, we have

$$R_N \in \mathcal{M}_t^{\pm}.$$

From $\mathcal{E}_2$, we have

$$R \in \left[R_N - \Delta_N(\alpha_2), R_N + \Delta_N(\alpha_2)\right].$$

Combining these two statements implies

$$R = R_N + (R - R_N) \in \mathcal{M}_t^{\pm} \oplus [-\Delta_N(\alpha_2), \Delta_N(\alpha_2)]\,.$$

Assumption 1 also implies $R \in [L, U]$. Therefore, by the definition of C_t, it follows that $R \in C_t$. Since $1 \le t \le N$ was arbitrary, we conclude

$$\mathcal{E}_1 \cap \mathcal{E}_2 \subseteq \{\forall 1 \le t \le N,\ R \in C_t\}\,.$$

Taking complements, we obtain

$$\{\exists 1 \le t \le N :\ R \notin C_t\} \subseteq \mathcal{E}_1^c \cup \mathcal{E}_2^c.$$

Therefore, by a union bound,

$$\Pr\left(\exists 1 \le t \le N :\ R \notin C_t\right) \le \Pr\left(\mathcal{E}_1^c\right) + \Pr\left(\mathcal{E}_2^c\right) \le \alpha_1 + \alpha_2 = \alpha.$$

Thus, for every $N \in \mathbb{Z}_+$, $(C_t)_{1 \le t \le N}$ is a $(1-\alpha)$ anytime-valid confidence sequence for R.

For any nonempty bounded set $A \subset \mathbb{R}$ and any $r \ge 0$, the Minkowski sum satisfies

$$\sup\left(A \oplus [-r, r]\right) = \sup A + r, \qquad \inf\left(A \oplus [-r, r]\right) = \inf A - r.$$

Hence

$$\operatorname{width}\left(A \oplus [-r, r]\right) = (\sup A + r) - (\inf A - r) = \operatorname{width}(A) + 2r.$$

If $\mathcal{M}_t^{\pm} = \varnothing$, then $C_t = \varnothing$, and the width bound below holds trivially, assigning width zero to the empty set. Otherwise, applying this with $A = \mathcal{M}_t^{\pm}$ and $r = \Delta_N(\alpha_2)$ gives

$$\operatorname{width}\left(\mathcal{M}_t^{\pm} \oplus [-\Delta_N(\alpha_2), \Delta_N(\alpha_2)]\right) = \operatorname{width}\left(\mathcal{M}_t^{\pm}\right) + 2\Delta_N(\alpha_2).$$

Since

$$C_t = \left(\mathcal{M}_t^{\pm} \oplus [-\Delta_N(\alpha_2), \Delta_N(\alpha_2)]\right) \cap [L, U],$$

and intersecting with $[L, U]$ can only reduce the width, for every $1 \le t \le N$ we have

$$\begin{aligned}\operatorname{width}(C_t) &\le \operatorname{width}\left(\mathcal{M}_t^{\pm}\right) + 2\Delta_N(\alpha_2) \\ &= \operatorname{width}\left(\mathcal{M}_t^{\pm}\right) + 2(U - L)\sqrt{\frac{\log(2/\alpha_2)}{2N}}.\end{aligned}$$

Writing $C_{t,N}$ and $\mathcal{M}_{t,N}^{\pm}$ to make their dependence on N explicit, let $\{t_N\}_{N \ge 1}$ be any integer sequence satisfying

$$1 \le t_N \le N, \qquad t_N \to \infty \quad \text{as } N \to \infty.$$

The preceding width inequality gives

$$\operatorname{width}\left(C_{t_N,N}\right) \le \operatorname{width}\left(\mathcal{M}_{t_N,N}^{\pm}\right) + 2\Delta_N(\alpha_2).$$

By Theorem 5,

$$\operatorname{width}\left(\mathcal{M}_{t_N,N}^{\pm}\right) = O_{\mathbb{P}}\left(\log\log t_N \sqrt{\frac{\log t_N}{t_N}}\right).$$

Moreover, since α_2 is fixed,

$$2\Delta_N(\alpha_2) = 2(U - L)\sqrt{\frac{\log(2/\alpha_2)}{2N}} = O\left(\sqrt{\frac{1}{N}}\right).$$

Consequently,

$$\operatorname{width}\left(C_{t_N,N}\right) = O_{\mathbb{P}}\left(\log\log t_N \sqrt{\frac{\log t_N}{t_N}}\right) + O\left(\sqrt{\frac{1}{N}}\right).$$

Since $t_N \to \infty$ and $N \to \infty$, both terms converge to zero. Therefore,

$$\operatorname{width}\left(C_{t_N,N}\right) \xrightarrow{\mathbb{P}} 0 \qquad \text{as } N \to \infty.$$

Since the sequence $\{t_N\}_{N \ge 1}$ was arbitrary, the confidence intervals are shrinking in the joint asymptotic regime $N \to \infty$, $t = t_N \to \infty$, and $t_N \le N$. This completes the proof. □

D Synthetic Experimental Setting for Table 1

CER-EVAL [9] adaptively selects evaluation samples and constructs CIs by partitioning the evaluation pool and applying adaptive concentration inequalities within each partition [20]. This approach can be effective when the partition is fixed and aligned with the true risk structure, since a low-variance partition can yield tighter CIs with fewer evaluated samples. However, its certification analysis relies on concentration bounds for a fixed or well-behaved partition, whereas the algorithm repeatedly learns and updates the partition from the evaluated samples, their embeddings, and their observed risk values. Consequently, the stated guarantee does not directly cover the complete adaptive partition-learning procedure; this distinction concerns the scope of the current theoretical analysis rather than the validity of adaptive partitioning in general. In this procedure, the same evaluated data are used both to choose the partition and to quantify the within-partition variation. When the embedding used for partitioning is uninformative about the target risk, the adaptive partition can overfit the evaluated samples and underestimate the variation among unevaluated samples, leading to overly narrow CIs.

We present a synthetic setting to further illustrate this distinction between the fixed-partition guarantee and the adaptive procedure of CER-EVAL [9]. Specifically, we construct a setting where the evaluation pool contains genuine risk heterogeneity, but this heterogeneity is independent of the embeddings used for partitioning. Therefore, any embedding-based partition cannot reliably transfer the risk structure observed on evaluated samples to the remaining unevaluated samples.

Data-generating process. We generate a finite input pool $X_N = \{\boldsymbol{x}_j\}_{j=1}^N$. For each sample $\boldsymbol{x}_j$, we first draw a latent group $\Gamma_j \in \{1, \dots, \Gamma\}$. The latent group determines the risk level of the sample, but it is not encoded in the embedding used for partitioning. Specifically, each group k is assigned a risk probability p_k, and we choose these Γ probabilities to be evenly spaced between 0.1 and 0.9. Given $\Gamma_j = k$, the realized per-sample risk value is generated as

$$\ell_j \mid \Gamma_j = k \sim \text{Bernoulli}(p_k),$$

so samples from different latent groups have different risk distributions. The finite-pool risk is

$$R_N = \frac{1}{N} \sum_{j=1}^{N} \ell_j.$$

The crucial design choice is that the representation used for partitioning is independent of the latent risk structure:

$$\boldsymbol{x}_j \perp (\Gamma_j, \ell_j).$$

Thus, the pool contains genuine risk heterogeneity through the latent groups, but this heterogeneity is invisible from the embeddings. In other words, two samples may have very different risk probabilities because they belong to different latent groups, but their embeddings provide no information about this difference. Therefore, an embedding-based partition cannot reliably infer the risk behavior of unevaluated samples from the evaluated samples.

This setting is not adversarial in the sense of manipulating the evaluation order or using unbounded risks. The samples are generated i.i.d., and the realized risk values remain bounded in $[0, 1]$. The only difficulty is that the representation used for partitioning is statistically unrelated to the target risk. This isolates a realistic failure mode: an embedding may capture semantic similarity, while being weakly related or unrelated to the particular evaluation risk being certified.

Why this setting is challenging for CER-EVAL. This setting is challenging for CER-EVAL because the partition is learned from the evaluated samples, while the embeddings used to construct the partition are independent of the target risk values. Hence, any apparent risk structure discovered from the evaluated samples cannot be reliably transferred to the unevaluated samples. The key issue is that the observed risk values are used both to learn the partition and to estimate the within-partition variation. As a result, the adaptive partition may look highly informative on the evaluated samples, for example by separating samples with observed risk value 0 from those with observed risk value 1, even though the embeddings themselves contain no information about the risk.

This creates a mismatch between the fixed-partition analysis and the adaptive procedure used in practice. CER-EVAL constructs its CI as if the learned partition were fixed or well behaved, so a small empirical within-partition variation on the evaluated samples leads to a small reported CI

width. However, in this setting, the learned partition provides no predictive information about the risk values of unevaluated samples, because the embeddings used to assign samples to partition cells are independent of ℓ_j. Therefore, the small within-partition variation reflects overfitting to the evaluated risks rather than genuine risk reduction on the remaining pool. Consequently, CER-EVAL can produce overly narrow CIs and suffer undercoverage, especially when the CI is repeatedly updated and used to stop once the target precision is reached. This undercoverage arises because the adaptive procedure lies outside the scope of the fixed-partition guarantee; it does not indicate that the theoretical analysis of CER-EVAL is incorrect under its stated assumptions.

Setups for CELEUS on synthetic data. We compare three variants of CELEUS in this synthetic setting. The first is the standard CELEUS, for which we generate simple noisy surrogate risk scores as auxiliary information:

$$\tilde{\ell}_f(\boldsymbol{x}_j) = \Pi_{[0,1]}(p_{\Gamma_j} + \sigma_s \xi_j), \quad \xi_j \sim \mathcal{N}(0,1),$$

where $\Pi_{[0,1]}$ denotes clipping to $[0,1]$. In this variant, the sampling probabilities are computed using only the surrogate risk scores, since the true risk values ℓ_j are unobserved before evaluation. Specifically, we use a pre-evaluation uncertainty proxy proportional to $\sqrt{\tilde{\ell}_f(\boldsymbol{x}_j)(1-\tilde{\ell}_f(\boldsymbol{x}_j))}$, which prioritizes samples whose surrogate risk scores are closer to $1/2$.

The second variant is CELEUS without surrogate assistance, where we set $\tilde{\ell}_f(\boldsymbol{x}_j) = 0$ for all samples and use uniform sampling. This variant illustrates that the certification remains valid even without informative surrogate risk scores. The third variant is Oracle CELEUS, which uses the same surrogate risk scores but samples according to the infeasible oracle proposal in Eqn. (6), with the same $\beta = 0.4$ as the practical method:

$$q_t^{\mathrm{orc}}(j) = \frac{\beta}{m_t} + (1-\beta) p_t^{\mathrm{res}}(j), \qquad p_t^{\mathrm{res}}(j) \propto |\ell_j - \tilde{\ell}_f(\boldsymbol{x}_j)|.$$

This oracle rule uses the unobserved realized risk values ℓ_j before evaluation, and is therefore included only as an idealized benchmark.

E Experimental Setup Details

This appendix expands the main-text setup of Section 5.1. Throughout, $D_N = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^N$ denotes the complete evaluation pool containing all available input-label pairs from the benchmark split used for evaluation, and $X_N = \{\boldsymbol{x}_j\}_{j=1}^N$ denotes its inputs.

E.1 Benchmarks

We evaluate on three publicly available classification benchmarks chosen to span label-space size, task difficulty, and pool size.

SST-2 [27][9]. Binary sentiment classification on movie-review sentences and phrases from the Stanford Sentiment Treebank. We use all 67,349 training examples and 872 validation examples as the evaluation pool, giving $N_{\mathrm{SST2}} = 68{,}221$. The label space $\mathcal{Y} = \{\text{negative}, \text{positive}\}$ is approximately balanced. Under 0-1 loss, the true loss is either 0 or 1, while the surrogate score may vary continuously over $[0,1]$. The residual $\ell - \widetilde{\ell}_f$ therefore lies in $[-1,1]$; its distribution and the resulting proposal q_t^* depend on the surrogate scores as well as the labels.

MMLU [1][10]. Four-way multiple-choice question answering across 57 academic subjects distributed over four broad strata: STEM (e.g., college mathematics, physics, chemistry, computer science), humanities (e.g., philosophy, formal logic, world history, moral disputes), social sciences (e.g., econometrics, sociology, public relations), and applied/professional domains (e.g., professional medicine, professional law, jurisprudence). Every question is presented with four answer options labelled $\{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$ and a single correct choice, giving label-space cardinality $|\mathcal{Y}| = 4$. Following the benchmark's evaluation convention, we obtain a single evaluation pool with $N_{\mathrm{MMLU}} = 115{,}700$.

[9] https://huggingface.co/datasets/stanfordnlp/sst2
[10] https://huggingface.co/datasets/cais/mmlu

Table 5: Datasets and surrogate-target pairs in the main classification experiments. Each of the six pairs is evaluated on each dataset. Additional studies state their own subsets and preprocessing.

Dataset	Task	Pool size N
SST-2	Binary sentiment classification	68,221
MMLU	Four-way question answering	115,700
AG News	Four-way topic classification	127,600
Pair	**Surrogate**	**Target**
1	Llama-3-8B	Llama-3-70B
2	Llama-3-8B	DeepSeek-67B
3	Llama-3-8B	Mixtral-8×7B
4	Llama-3-8B	Qwen-2.5-72B
5	Llama-2-7B	Llama-3-70B
6	Llama-2-7B	DeepSeek-67B

AG News [28][11]. Four-way topic classification from the headline and lead paragraph of news articles in the AG corpus, with label space $\mathcal{Y} = \{\text{World}, \text{Sports}, \text{Business}, \text{Sci/Tech}\}$. The standard release combines 120,000 training and 7,600 test examples for a total pool of $N_{\text{AGNews}} = 127{,}600$, with classes balanced exactly by construction. AG News probes a third regime distinct from SST-2 and MMLU. Topic classification on lead paragraphs relies on dispersed lexical cues that smaller surrogate models capture less reliably than they do sentiment polarity or short-form factual recall.

The three benchmarks vary in label-space size, task content, and pool size. Their label-space cardinalities are $|\mathcal{Y}| \in \{2, 4, 4\}$. The variance analysis depends on the resulting surrogate residuals and sampling probabilities, rather than on label-space size alone. In every case the full pool plays the role of D_N, so the finite-pool risk R_N is the average loss over the entire benchmark. The pool size varies and controls the magnitude of the population correction $\Delta_N(\alpha_2)$.

E.2 Models and surrogate-target pairs

The target model is the LLM whose population risk we certify; the surrogate model produces the auxiliary score $\widetilde{\ell}_f$ for unevaluated samples. We instantiate *six* surrogate-target pairs that vary along the model-family and model-generation axes. Each surrogate is smaller than its target; the computational benefit depends on the inference setup and the cost of obtaining target losses.

1. **Llama-3-8B** → **Llama-3-70B**: same family, same generation.
2. **Llama-3-8B** → **DeepSeek-67B**: cross-family, same generation.
3. **Llama-3-8B** → **Mixtral-8×7B**: cross-family, same generation.
4. **Llama-3-8B** → **Qwen-2.5-72B**: cross-family, same generation.
5. **Llama-2-7B** → **Llama-3-70B**: same family, cross-generation.
6. **Llama-2-7B** → **DeepSeek-67B**: cross-family, cross-generation.

For all open-weights models, Llama-2-7B [39], Llama-3-8B and Llama-3-70B [40], DeepSeek-67B [41], Mixtral-8×7B [42], and Qwen-2.5-72B [43], we use the publicly released checkpoints without further fine-tuning. All predictions are obtained at temperature 0. Following Berrada et al. [18], for the multi-choice tasks (MMLU, AG News), we score each candidate label by its next-token log-probability and predict $\arg\max$; SST-2 follows the same scoring convention restricted to the two label tokens.

E.3 Baselines

EVALUE [21]. Applies the hedged-capital confidence sequence of Eqn. (3) to uniform sampling without a surrogate ($\widetilde{\ell} \equiv 0$). The betting fractions, scaling, and α-split follow the common experimental settings.

[11] https://huggingface.co/datasets/fancyzhx/ag_news

CELEUS (W/O SURR). Sets $\tilde{\ell}_f \equiv 0$ and samples uniformly without replacement over the remaining pool, with $q_t(j) = 1/(N-t+1)$. It retains the inferential signal and CI construction of CELEUS. Both CELEUS (W/O SURR) and EVALUE use uniform sampling, so their comparison does not isolate active sample selection. Relative to full CELEUS, CELEUS (W/O SURR) removes both surrogate assistance and uncertainty-guided sampling. The controlled study in App. A.4 examines the two contributions separately.

Oracle. Uses the true per-sample residuals as sample-selection scores with $\beta = 0.4$, sampling according to q_t^{orc} in Remark 6. Computing these probabilities requires the losses of unevaluated samples, so ORACLE is not deployable in practice. It provides an idealized variance reference; comparisons with the practical method reflect unavailable residual information.

CER-EVAL [9]. At each round, CER-EVAL partitions the evaluation pool using the evaluated samples, computes within-partition statistics and CI radii through Bernstein-style or adaptive concentration bounds, selects the partition that contributes most to the current uncertainty, and evaluates a new sample from that partition. This strategy can be effective when the partition is fixed or well aligned with the true risk structure, since a low-variance partition can reduce the number of evaluated samples. However, its certification relies on concentration arguments for a fixed or well-behaved partition, whereas the algorithm repeatedly learns and updates the partition from the evaluated samples, their embeddings, and their observed risks. Consequently, the stated guarantee does not directly cover the complete adaptive partition-learning procedure; this distinction concerns the scope of the current theoretical analysis rather than the validity of adaptive partitioning in general. In this procedure, the same evaluated data are used both to choose the partition and to quantify the within-partition variation. As shown in our diagnostic coverage study Tab. 1, such adaptive partitioning can fail to maintain anytime-valid coverage under repeated monitoring and stopping. Therefore, we report CER-EVAL separately to illustrate this distinction, and focus the main efficiency comparisons on methods designed to be anytime-valid.

E.4 Hyperparameters

Significance level and split. We set the total significance level to $\alpha = 0.05$, which controls the miscoverage probability of the final CI. We split it evenly as $\alpha_1 = \alpha_2 = 0.025$: α_1 is used for the anytime-valid e-process inference of the finite-pool risk R_N, while α_2 is used for the finite-pool-to-population correction via Hoeffding's inequality. App. I studies the sensitivity of the number of evaluated samples required to reach ϵ to this split.

e-process construction. Following the standard e-process construction of Waudby-Smith and Ramdas [21], we use the predictable fractions in Rem. 3 with constant $c = 0.5$. We set the hedge weight to $\theta = 0.5$ in the hedged capital process in Eqn. (3), treating the upper and lower directions symmetrically.

Scaling. We use the global scaling strategy in Rem. 4. Specifically, we choose fixed predictable bounds a and b that satisfy $a \le \hat{S}_t \le b$ for all rounds, and scale the signal as $\bar{S}_t = (\hat{S}_t - a)/(b-a)$ before constructing the e-process.

Grid for numerical test inversion. The candidate grid $\mathcal{G}$ in App. B consists of $K = \max\{2000, \lceil 2000(U-L)\rceil\}$ equally spaced interior points $v_i = L + i(U-L)/(K+1)$, $i = 1, \ldots, K$, with K capped at 200,000; the same grid is used for CELEUS and EVALUE. Thus $K = 2000$ for the 0-1 loss and for the toxicity score in App. L.1, while under the cross-entropy loss K increases with the range $U - L$.

Sampling lower bound and stopping rule. For sampling, we set $\beta = 0.4$ to control the lower bound of the sampling probabilities, as required by Assumption 2. The sensitivity to this choice is also studied in App. I. We stop the evaluation at the first time t such that $\mathrm{width}(C_t) \le \epsilon$, with $\epsilon = 0.05$ in all main experiments.

E.5 Trials and reporting

Each configuration in the main classification experiments (a surrogate-target pair on a dataset) is run for 50 random seeds. Additional studies report their own numbers of seeds. Pooled trajectories in Figs. 1 and 3 stack per-seed widths across independent runs and report mean width with a $\pm 1\,\sigma$ band across the per-pair means.

F Empirical coverage on real LLM evaluation under 0-1 loss

In addition to the coverage investigation with synthetic data reported in Tab. 1, we also report the empirical coverage results for the main real LLM evaluation experiments under 0-1 loss (Tab. 2 and Fig. 1). The evaluation follows the same sequential protocol as in Sec. 5: after each round, the CI is updated, and the procedure stops at the time when the CI width reaches the target precision $\epsilon = 0.05$. Table 6 reports empirical miscoverage at this stopping time. The observed rates for CELEUS and EVALUE are below $\alpha = 0.05$ on all datasets. These results describe coverage at the tested stopping rule; they do not establish simultaneous coverage at every round or a formal guarantee for the numerical grid approximation.

Table 6: Empirical miscoverage rate at the stopping time when the CI width reaches $\epsilon = 0.05$, pooled across the surrogate-target pairs within each dataset. Brackets give Wilson 95% intervals. Both CELEUS and EVALUE remain below the nominal level $\alpha = 0.05$ on all datasets.

Method	SST-2	MMLU	AG NEWS
EVALUE	0.013 [0.005, 0.034]	0.007 [0.002, 0.024]	0.007 [0.002, 0.024]
CELEUS (W/O SURR)	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]
CELEUS	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]	0.013 [0.005, 0.034]
ORACLE	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]

G Certifiable and Label-efficient LLM evaluation on Cross-Entropy Loss

We follow the deployment-style protocol in Sec. 5, replacing the 0–1 risk with the cross-entropy risk $\ell_{\text{ce}}(f, \boldsymbol{x}, y) = -\log p_f(y \mid \boldsymbol{x})$. The reported experiment uses the maximum cross-entropy observed over the pool as its upper bound U. This bounds the realized pool losses, but uses information from all labels and does not supply a bound fixed before observing pool losses as required by Assumption 1. We therefore report these results as an empirical comparison, rather than as verification of the population-risk guarantee. The sample-selection rule follows Strategy A in Rem. 2, with the other hyperparameters as in App. E. Fig. 3 shows the resulting CI-width trajectories.

Table 7: Empirical miscoverage at the stopping time when the CI width reaches the target precision $\epsilon = 0.05$ under the **cross-entropy** loss, pooled across the six surrogate-target pairs within each dataset (50 seeds per pair). Point estimates are accompanied by Wilson 95% intervals (in gray). The observed miscoverage rates are below the nominal $\alpha = 0.05$ on every dataset. These rates concern the empirical pool-bound protocol described above and do not establish the population-risk guarantee of Thm. 6.

Method	SST-2	MMLU	AG NEWS
EVALUE	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]	0.007 [0.002, 0.024]
CELEUS (W/O SURR)	0.000 [0.000, 0.013]	0.007 [0.002, 0.024]	0.000 [0.000, 0.013]
CELEUS	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]
ORACLE	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]	0.000 [0.000, 0.013]

Miscoverage rate at the stopping time. Tab. 7 reports empirical miscoverage below $\alpha = 0.05$ for all displayed cross-entropy configurations. The theoretical construction applies to different loss functions when the boundedness, conditional-mean, sampling, and betting assumptions hold. Population inference additionally requires the stated population sampling assumptions. These conditions and exact test inversion are distinct from the empirical pool-bound and numerical-grid choices used here.

Efficiency. The relative ordering of methods at the stopping time matches the 0-1 case: in Fig. 3, CELEUS has smaller CI width than CELEUS (W/O SURR), which in turn improves over EVALUE, across the evaluation process. Two observations are worth noting. First, the required numbers of evaluated samples are larger, because cross-entropy risk has a wider range and the surrogate residual correspondingly contributes more to $\mathrm{Var}(\widehat{S}_t \mid \mathcal{F}_{t-1})$. Second, as in the 0–1 setting, the gain of full

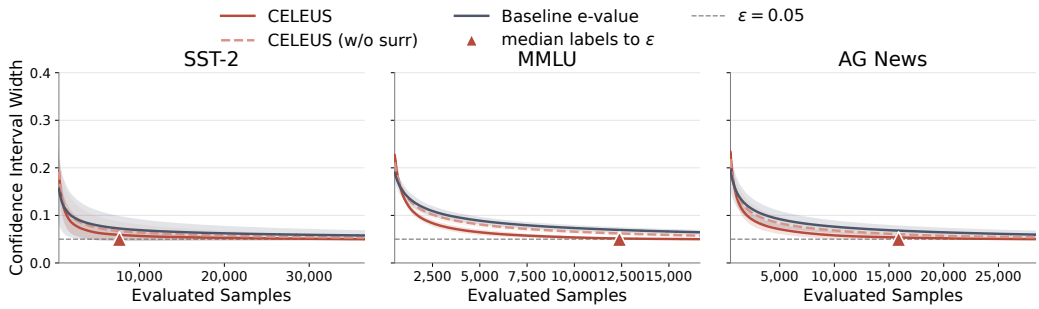

Figure 3: Confidence interval width versus number of evaluated samples under cross-entropy loss on SST-2, MMLU, and AG News. Each curve is the mean width at each label budget, aggregated over 50 seeds and the six surrogate-target pairs of App. E.2; the shaded band is $\pm 1\sigma$ across the per-pair means. The dashed line marks the target precision $\epsilon = 0.05$, and triangles mark the median samples-to-ϵ for CELEUS, which reaches ϵ first on every benchmark.

CELEUS over uniform CELEUS (W/O SURR) is largest on SST-2 and MMLU, where the entropy proxy is more aligned with the target cross-entropy risk, and smallest on AG News.

H Sample-selection strategy comparison

The deployed configurations in Sec. 5 and Sec. G use one strategy from each of the two label-free families: Strategy A of Rem. 1 for the 0-1 loss and Strategy A of Rem. 2 for cross-entropy. This appendix compares the alternatives within each family while retaining the same e-process construction and stopping rule. The strategies vary the acquisition proxy and the surrogate score used in the residual correction, so this comparison does not isolate the effect of q_t alone.

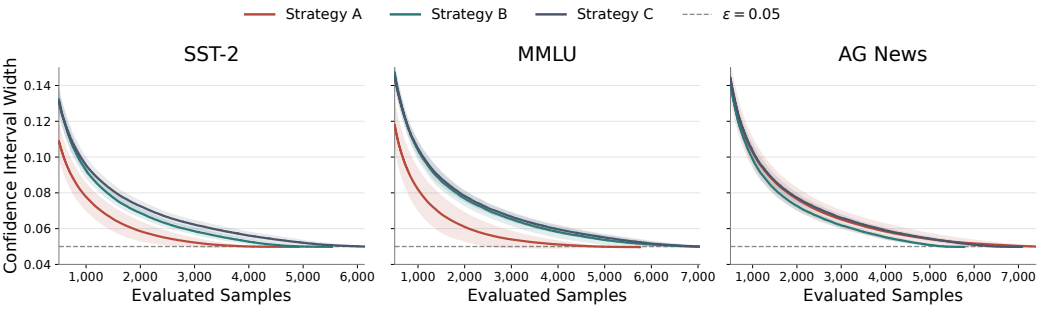

Figure 4: Sample-selection strategy comparison under 0-1 risk. We compare the sample-selection strategies in Rem. 1 by plotting CI width versus the number of evaluated samples on SST-2, MMLU, and AG News. Each curve is averaged over 50 seeds, and the dashed line marks the target precision $\epsilon = 0.05$.

0-1 **loss family.** The strategies in Rem. 1 differ in how they approximate the oracle residual $|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)|$ using label-free quantities. Strategy A uses the hard pseudo-loss $\ell_{\text{hard}}(f, g, \boldsymbol{x}_j) = \mathbf{1}\{\hat{y}_f(\boldsymbol{x}_j) \neq \hat{y}_g(\boldsymbol{x}_j)\}$ as the target-loss proxy and pairs it with the predictive dispersion score $\ell_{\text{disp}}(f, \boldsymbol{x}_j)$ as the surrogate score. Strategies B and C both use $\ell_{\text{disp}}(f, \boldsymbol{x}_j)$ and the soft pseudo-loss $\ell_{\text{soft}}(f, g, \boldsymbol{x}_j) = 1 - p_f(\hat{y}_g(\boldsymbol{x}_j) \mid \boldsymbol{x}_j)$, and therefore produce numerically identical sample-selection rankings. They differ in the assignment of the target-loss proxy and surrogate score: Strategy B uses ℓ_{disp} as the proxy and ℓ_{soft} as the surrogate, whereas Strategy C reverses them. The proxy determines acquisition scores; the estimator in Eqn. (5) uses the observed target loss and the assigned surrogate score. The strategies give the same proposal for the same remaining pool, but different surrogate corrections. Their realized sampling orders coincide only if the sampling randomness is also coupled. Fig. 4 compares the three strategies on SST-2, MMLU, and AG News. Strategy A yields smaller CI widths than Strategies B and C across evaluation budgets on all three datasets, while remaining within the prescribed miscoverage level $\alpha = 0.05$. This is intuitive for 0-1

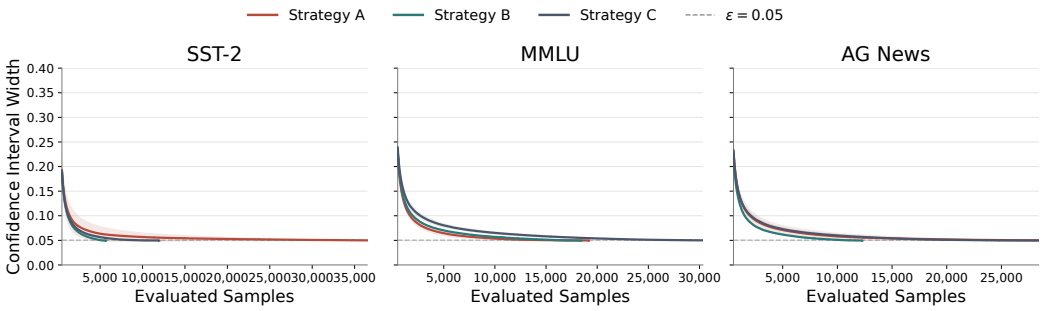

Figure 5: Sample-selection strategy comparison under cross-entropy risk. We compare the sample-selection strategies in Rem. 2 by plotting CI width versus the number of evaluated samples on SST-2, MMLU, and AG News. Each curve is averaged over 50 seeds, and the dashed line marks the target precision $\epsilon = 0.05$.

loss: the hard pseudo-loss preserves the discrete agreement structure of the target loss more directly, whereas the soft pseudo-loss spreads this information over $[0, 1]$. As a result, Strategy A more effectively prioritizes samples with large proxy residuals, aligning more closely with the variance-optimal sampling rule in Thm. 4.

Cross-entropy loss family. The three strategies in Rem. 2 differ in the same structural way, with predictive entropy and mode loss used as label-free proxies for the unobserved cross-entropy loss. Strategy A treats the mode loss $\ell_{\text{mode}}(f, \boldsymbol{x}_j) = -\log p_f(\hat{y}_f(\boldsymbol{x}_j) \mid \boldsymbol{x}_j)$ as the target-loss proxy and uses the predictive entropy $H(f, \boldsymbol{x}_j)$ as the surrogate score. Strategy B reverses this assignment, treating $H(f, \boldsymbol{x}_j)$ as the target-loss proxy and $\ell_{\text{mode}}(f, \boldsymbol{x}_j)$ as the surrogate score; hence it has the same sample-selection ranking as Strategy A but a different correction term in Eqn. (5). Strategy C is the only strategy that uses a separate surrogate model: it compares the target predictive entropy $H(f, \boldsymbol{x}_j)$ with the surrogate predictive entropy $H(g, \boldsymbol{x}_j)$, prioritizing samples where the two predictive distributions differ most in entropy. Fig. 5 compares the three strategies. The differences among strategies become smaller at large t, because their effect enters the CI width through the cumulative conditional-variance term in Thm. 5; as more samples are evaluated, this cumulative averaging reduces the impact of early sample-selection differences.

I Hyperparameter sensitivity

CELEUS has four configuration parameters that may affect the number of evaluated samples needed to reach the target precision, while preserving the exact-inversion coverage guarantee when the stated assumptions remain satisfied: the split of α between the e-process and the population correction, the lower bound β on sampling probabilities, the hedge weight θ, and the stabilizer c in the predictable fractions. The default values

$$\alpha_1 = \alpha_2 = 0.025, \qquad \beta = 0.4, \qquad \theta = 0.5, \qquad c = 0.5$$

are fixed in advance and are not tuned on evaluation risks. We vary one parameter at a time while keeping the others at their default values. For β, we sweep over $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. Tab. 8 reports the number of evaluated samples needed to reach ϵ on four representative dataset-model configurations.

Sensitivity to the α-split. Tightening α_1 from 0.025 to 0.005 increases the required number of evaluated samples by 40%, 33%, 39%, and 39% on configurations 1-4, respectively. Loosening α_1 to 0.045 reduces the required number by 14%-19%. This trend is consistent with the e-process construction: a smaller α_1 raises the rejection threshold $1/\alpha_1$, so more evidence is needed before the CI can shrink. Although a larger α_1 leaves a smaller α_2 for the finite-pool-to-population correction, this correction remains small for the large pools used in our experiments. We keep the symmetric split as default because it is conservative and keeps the population correction stable for smaller pools.

Sensitivity to the hedge weight θ. Changing θ from 0.5 to 0.25 or 0.75 changes the number of evaluated samples by less than 6% on every configuration, and both perturbations slightly increase the sample budget. This suggests that the symmetric choice $\theta = 0.5$ is near-optimal in our experiments, as it balances the upper and lower hedged capital processes in Eqn. (3).

Table 8: Hyperparameter sensitivity. We report the median number of evaluated samples required to reach $\epsilon = 0.05$ over 50 seeds. Each column corresponds to one dataset and surrogate-target pair: ① Llama-2-7B → Mixtral-8×7B on SST-2; ② Llama-2-7B → Qwen-2.5-72B on MMLU; ③ Llama-3-8B → DeepSeek-67B on AG News; and ④ Llama-3-8B → Llama-3-70B on SST-2. The pairs in columns ① and ② are not among the six pairs in Table 5. The default setting is $(\alpha_1 = 0.025,\ \theta = 0.5,\ c = 0.5)$, chosen without using prior information about the target LLMs or the evaluation datasets.

Config	SST-2/①	MMLU/②	AG News/③	SST-2/④
α-split, with $\alpha_1 + \alpha_2 = 0.05$				
baseline	2,426	2,116	5,436	3,997
$\alpha_1 = 0.005$	3,398	2,823	7,544	5,552
$\alpha_1 = 0.045$	**2,086**	**1,733**	**4,606**	**3,494**
θ, with $0 < \theta < 1$				
baseline	**2,426**	**2,116**	**5,436**	**3,997**
$\theta = 0.25$	2,531	2,199	5,673	4,188
$\theta = 0.75$	2,570	2,218	5,685	4,250
c, with $0 < c < 1$				
baseline	2,426	2,116	5,436	3,997
$c = 0.3$	2,533	2,141	5,709	4,266
$c = 0.7$	**2,415**	**2,112**	**5,213**	**3,892**

Sensitivity to the constant c. Changing c to 0.3 or 0.7 has a mild effect. The choice $c = 0.3$ increases the number of evaluated samples by 4%-7%, while $c = 0.7$ slightly reduces it on some configurations. Since c controls the scale of the predictable fractions in Rem. 3, this result indicates that the default $c = 0.5$ provides a stable trade-off across datasets.

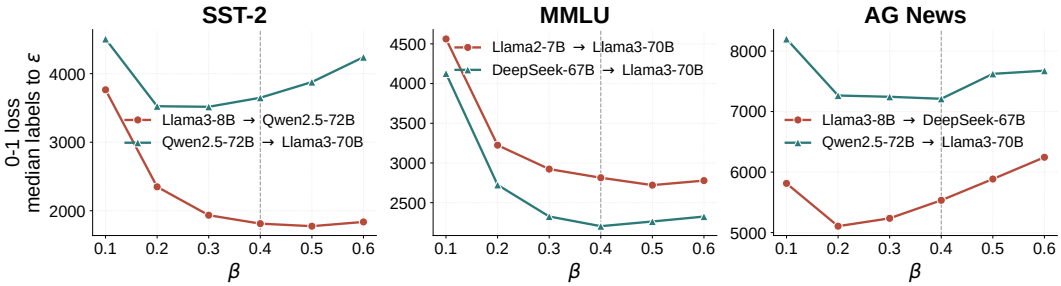

Figure 6: Sensitivity of median samples-to-ϵ to the sampling-probability floor β, on three benchmarks (SST-2, MMLU, AG News) and two surrogate-target pairs per benchmark. Each curve is the median number of evaluated samples used, over 50 seeds at $\epsilon = 0.05$ under 0-1 loss. Three curves use pairs from Table 5 (Llama-3-8B → Qwen-2.5-72B on SST-2, Llama-2-7B → Llama-3-70B on MMLU, and Llama-3-8B → DeepSeek-67B on AG News); the other three curves use two additional pairs (Qwen-2.5-72B → Llama-3-70B on SST-2 and AG News, and DeepSeek-67B → Llama-3-70B on MMLU). The vertical dashed line marks the deployed value $\beta = 0.4$. Curves are L-shaped on the left and either flat or mildly increasing on the right; the deployed value is within 5% of the per-experimental configuration minimum on every curve.

Sensitivity to β. The parameter β controls the lower bound on the sampling probabilities and hence prevents the sampling rule from assigning too little probability to any remaining sample. Fig. 6 reports the results on six dataset-model configurations. The default value $\beta = 0.4$ is at, or within 5% of, the best value on every curve. When β is too small, some sampling probabilities can become very small, which slows down the accumulation of evidence and leads to wider CIs. When β is too large, the sampling rule is forced closer to uniform sampling, which can reduce the benefit of uncertainty-guided sampling. The best values mostly lie in $\beta \in [0.3, 0.5]$, showing that the default choice is stable across benchmarks.

Overall, the sensitivity analysis shows that the performance of CELEUS changes smoothly under these hyperparameter perturbations. The default choices provide stable performance across datasets without tuning on evaluation risks.

Table 9: Per-round wall-clock cost in seconds, averaged over 50 seeds. Each column indexes a dataset-pair: ①: Llama-2-7B → Mixtral-8×7B on SST-2, ②: Llama-2-7B → Qwen-2.5-72B on MMLU, ③: Llama-3-8B → DeepSeek-67B on AG News, ④: Llama-3-8B → Llama-3-70B on SST-2. The pairs in columns ① and ② are not among the six pairs in Table 5.

Method	SST-2/①	MMLU/②	AG News/③	SST-2/④
CELEUS	0.0388	0.5101	0.0963	0.0251
EVALUE	0.0064	0.0305	0.0066	0.0018

J Computational cost and break-even analysis

Tab. 9 reports the per-round wall-clock cost of CELEUS and EVALUE baseline on the same four (dataset, pair) experimental configurations used in App. I. The wall-clock figures exclude target-model inference. The shared-output protocol computes target predictions over the full pool before sequential sampling, so this cost is common to the compared methods. The per-round measurements describe the remaining work reported for each method; they should not be interpreted as end-to-end evaluation times.

The surrogate scores are fixed before sequential evaluation. Constructing the proposal uses the scores of the remaining samples and does not require rerunning the surrogate model at every round. On the slowest listed configuration, mmlu/2, CELEUS takes about 0.51 seconds per round, approximately 0.48 seconds more than EVALUE. Whether the 54–62% reduction in target-loss queries in Table 2 offsets this overhead depends on the cost of obtaining each loss, as quantified below.

Break-even analysis. Our primary empirical efficiency measure is target-loss-query efficiency, defined as the number of target losses that must be revealed before the prescribed precision is reached. On the labeled benchmarks considered here, each query corresponds to revealing the reference label needed to compute the target loss after the target-model output is available. In other applications, it may instead correspond to obtaining a human judgment or invoking a model judge. This quantity is distinct from the number of target-model forward passes and from end-to-end computational cost.

The practical acquisition rules evaluated in this paper use target predictive outputs over the full evaluation pool. In our offline timing protocol, these outputs are computed once and reused across the compared methods. We therefore treat the corresponding full-pool target-inference time, denoted by T_{target}, as a common cost. Let n_C and n_E denote the numbers of target-loss queries required by CELEUS and EVALUE to reach the target CI width, respectively. Let T_C and T_E denote their remaining method-specific computation times accumulated up to their stopping times. For an end-to-end comparison, T_C must include any method-specific full-pool surrogate inference as well as online computation, and T_E must include the corresponding costs of EVALUE. The numerical timing results below instead condition on precomputed outputs and surrogate scores, as detailed in Sec. L.3; in that online-only comparison, T_C and T_E contain only the measured online runtimes. Finally, let p_L denote the time-equivalent marginal cost of revealing one target loss after the corresponding target output is available. The total times under this shared-output protocol are

$$T_C^{\text{tot}}(p_L) = T_{\text{target}} + T_C + n_C p_L, \qquad T_E^{\text{tot}}(p_L) = T_{\text{target}} + T_E + n_E p_L.$$

Because T_{target} is common, it does not affect the comparison. We therefore report the method-specific incremental times

$$\widetilde{T}_C(p_L) = T_C + n_C p_L, \qquad \widetilde{T}_E(p_L) = T_E + n_E p_L.$$

In all settings considered below, $n_E > n_C$ and $T_C > T_E$. The conditional break-even query cost is consequently

$$p_L^\star = \frac{T_C - T_E}{n_E - n_C},$$

and

$$\widetilde{T}_C(p_L) < \widetilde{T}_E(p_L) \quad \Longleftrightarrow \quad p_L > p_L^\star.$$

Thus, under the shared-output protocol, the query savings of CELEUS offset its additional surrogate and online computation when the marginal cost of revealing each target loss exceeds $p_L^\star$. More generally, if $n_E > n_C$ and $T_C \le T_E$, CELEUS has no greater incremental cost for every $p_L \ge 0$, with a strict improvement whenever $T_C < T_E$ or $p_L > 0$.

Table 10: Conditional break-even target-loss-query costs and online evaluation times required to reach width $\epsilon = 0.05$, with target outputs and surrogate scores precomputed. Break-even costs are point estimates [95% intervals] in milliseconds per query. Times include measured online runtime and the stated query cost, exclude preprocessing, and are reported in seconds as CELEUS / EVALUE.

Dataset	$p_L^\star$ (ms/query)	$p_L = 1$ ms/query	$p_L = 10$ ms/query
SST-2	0.92 [0.63, 1.81]	10.83 / 11.38	41.28 / 107.55
MMLU	2.24 [1.89, 2.77]	16.12 / 9.29	45.12 / 87.86
AG News	3.23 [2.37, 4.37]	36.71 / 16.66	96.78 / 157.82

The online timing study in Sec. L.3 covers SST-2, MMLU, and AG News using six random seeds per model pair. Each method stops when its numerical CI width first reaches $\epsilon = 0.05$, and all reported runs reach this width before pool exhaustion. Table 10 reports the break-even query costs and incremental times with outputs and surrogate scores precomputed. Adding a common preprocessing cost to both methods leaves their ordering unchanged; adding a method-specific surrogate-inference cost can change the break-even point.

At $p_L = 1$ ms per query, CELEUS has lower incremental time on SST-2, whereas EVALUE has lower incremental time on MMLU and AG News. At $p_L = 10$ ms per query, CELEUS has lower incremental time on all three datasets. Therefore, under the shared-output protocol, the comparison depends on the application-specific query cost: the reduction in queried target losses outweighs the additional computation when such queries are sufficiently costly, whereas EVALUE can be less expensive when they are inexpensive.

When target outputs are not already available or shared, their method-specific inference costs should also be included. Let T^C_{target} and T^E_{target} denote the target-inference costs of CELEUS and EVALUE, respectively. The corresponding total costs are

$$T^{\text{tot}}_C(p_L) = T^C_{\text{target}} + T_C + n_C p_L, \qquad T^{\text{tot}}_E(p_L) = T^E_{\text{target}} + T_E + n_E p_L,$$

which yield the general break-even point

$$p^\star_{L,\text{general}} = \frac{T_C - T_E + T^C_{\text{target}} - T^E_{\text{target}}}{n_E - n_C},$$

whenever the numerator is positive. For example, if the practical acquisition rule of CELEUS requires target inference on all N samples, while EVALUE performs target inference only on its n_E queried samples, and T_f denotes the average per-sample target-inference time, then

$$T^C_{\text{target}} = NT_f, \qquad T^E_{\text{target}} = n_E T_f,$$

and hence

$$p^\star_{L,\text{general}} = \frac{T_C - T_E + (N - n_E)T_f}{n_E - n_C}.$$

This expression separates the shared-output setting reported in Table 10 from deployment settings in which full-pool target inference constitutes an additional method-specific cost.

The surrogate models are 7–8B open-weight models used without task-specific fine-tuning. Their inference cost must be added when it is incurred for the evaluation, since the reported online timing starts from precomputed scores. The uniform surrogate-free variant CELEUS (W/O SURR) also requires fewer target-loss queries than EVALUE in the main experiments and does not require target outputs to select the next sample. These results concern query efficiency; end-to-end cost depends on the preprocessing and loss-acquisition costs of the deployment.

A precise monetary comparison would require application-specific costs for target inference, surrogate inference, labels or judgments, and computational hardware. Accordingly, our primary empirical efficiency claim concerns the reduction in target-loss queries required to reach the prescribed precision. Table 10 characterizes when these savings also reduce incremental cost conditional on precomputed target outputs and surrogate scores, while the general expression above covers settings with method-specific target-inference costs.

K Results for each surrogate-target model pair

In addition to the dataset-level averages, we report the CI-width trajectories and evaluated samples-to-ϵ for each $(\textit{dataset}, \textit{surrogate}, \textit{target})$ configuration separately in Fig. 7. The results show that CELEUS reaches the target precision with fewer evaluated samples than EVALUE across all configurations, not only after averaging over surrogate-target pairs. The improvement is especially clear on AG News, where the gap between CELEUS and EVALUE is large for all six surrogate-target pairs, and CELEUS typically reaches $\epsilon = 0.05$ at roughly half of the budget required by EVALUE. Across the full grid, ORACLE provides an idealized reference, while CELEUS consistently moves the CI-width trajectory closer to this oracle curve than the surrogate-free baselines.

The comparison between CELEUS and CELEUS (W/O SURR) shows the combined effects of surrogate assistance and uncertainty-guided sampling. When the surrogate scores are more informative for the target risk, the gap between the two methods becomes larger; for example, in several Llama-3-8B surrogate configurations, CELEUS shrinks the CI visibly faster than CELEUS (W/O SURR). In configurations where the surrogate-target relationship is weaker, the two curves are closer, indicating that the gain from the full procedure is smaller but the anytime-valid guarantee is still preserved. This pattern is consistent with Thm. 5, whose variance term depends on both the residuals and the sampling probabilities. These curves do not isolate either contribution; the controlled study in App. A.4 varies the two factors separately. Thus, the per-configuration results support the same conclusion as the dataset-level summary: CELEUS improves efficiency through both uncertainty-guided sampling and surrogate-assisted approximation, while retaining certifiable coverage.

A corner case. On the SST-2 evaluation of the Llama-3-8B $\rightarrow$ Mixtral-8$\times$7B row, the median samples-to-ϵ of CELEUS (2,422) is slightly smaller than that of ORACLE (2,451). The difference is about 30 evaluated samples and lies within the seed-to-seed spread of either method; it does not imply that CELEUS is better than ORACLE. However, this case is still informative about the one-step objective of the oracle in Remark 6. The oracle aims to reduce the conditional variance of the one-step signal $\widehat{S}_t$ given the past, but the reported metric is the stopping time determined by the entire CI trajectory. These are closely related but not identical objectives.

On this pair, the target risk is small, with $R_N \approx 0.104$, so the surrogate residuals are highly sparse and unevenly distributed across the pool. In such a finite-sample regime, a proposal that is optimal for the next-step conditional variance need not produce a uniformly smaller stopping time on every random run, because the CI width also depends on the accumulated trajectory of signals, the scaling bounds, and the realized sampling path. By contrast, CELEUS's practical proposal is smoother and may occasionally produce a slightly shorter trajectory to the target precision on this specific configuration. Therefore, this small reversal should be viewed as finite-sample variability rather than a violation of the oracle optimality result. It also suggests that designing a label-free proposal that targets the full CI trajectory, rather than only the one-step variance, is an interesting direction for future work.

L Additional Empirical Results

We extend evaluation to open-ended generation. We then study the choice of a cross-entropy loss bound and quantify when label savings translate into lower evaluation cost. Throughout, target precision refers to confidence-interval width: an evaluation stops at the first round at which the interval width is at most ϵ.

L.1 Open-Ended Safety Evaluation

We evaluate free-form generations on RealToxicityPrompts [44][12], using a pool of $N = 99{,}442$ prompts and toxicity scores in $[0, 1]$. The target and surrogate are Qwen2.5-72B-Instruct and Qwen2.5-7B-Instruct [43], respectively; each generates one continuation per prompt with greedy decoding and at most 64 new tokens. ToxicBERT [45][13] scores each prompt concatenated with a generated continuation: the risk value ℓ_j is the score with the target continuation, and the surrogate

[12] https://huggingface.co/datasets/allenai/real-toxicity-prompts
[13] https://huggingface.co/unitary/toxic-bert

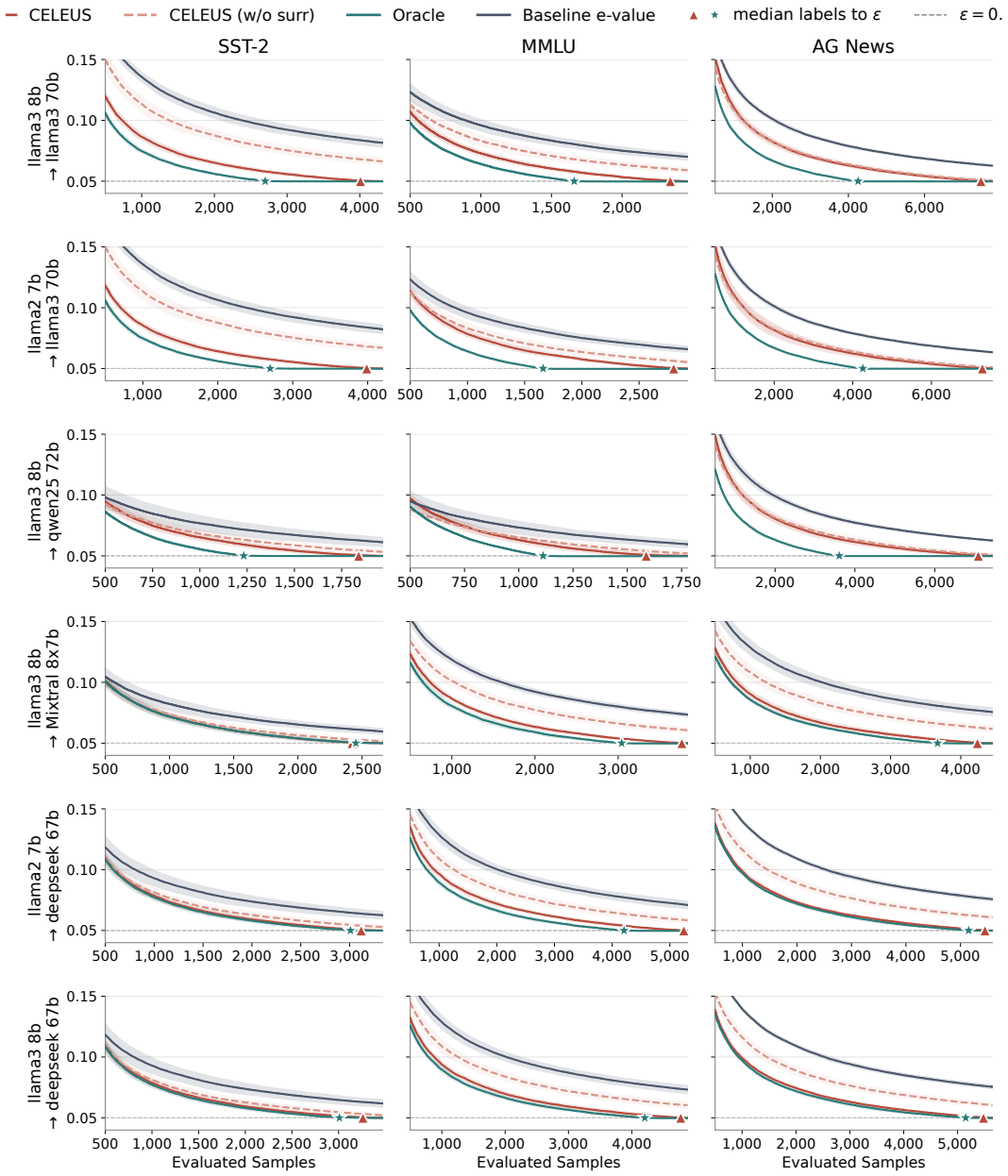

Figure 7: Confidence interval width versus evaluated samples on each (*dataset*, *surrogate*, *target*) experimental configuration. Rows index the six surrogate-target pairs of App. E.2, and columns index the three benchmarks. Each curve is the mean width over 50 seeds. The dashed line marks the target precision $\epsilon = 0.05$, and triangles mark the median samples-to-ϵ. The ordering CELEUS $\leq$ CELEUS (W/O SURR) $\leq$ EVALUE holds on every panel, and the gap between CELEUS and CELEUS (W/O SURR) reflects the combined change in surrogate assistance and sampling.

risk score $\tilde{\ell}_f(\boldsymbol{x}_j)$ is the score with the surrogate continuation. CELEUS samples with the score $s_t(j) = \tilde{\ell}_f(\boldsymbol{x}_j)(1 - \tilde{\ell}_f(\boldsymbol{x}_j))$, which favors prompts whose surrogate score is close to $1/2$. The full-pool mean toxicity is $R_N = 0.04667$. We use $\alpha = 0.05$ with $\alpha_1 = \alpha_2 = 0.025$, $[L, U] = [0, 1]$, and $K = 2000$ grid points. We compare CELEUS, its no-surrogate variant, and EVALUE over ten matched seeds (42–51), with a maximum budget of 5,000 evaluations. CELEUS uses $\beta = 0.8$. This value was selected after running CELEUS with $\beta \in \{0.1, 0.2, 0.4, 0.6, 0.8\}$ on the same pool and seeds; among these values, it required the fewest evaluations to reach $\epsilon = 0.05$. Unlike the default $\beta = 0.4$ in App. I, it is therefore not fixed in advance, and Table 11 may favor CELEUS. The no-surrogate variant sets surrogate predictions to zero and samples uniformly; EVALUE applies the baseline e-process to uniformly sampled losses.

Table 11: Open-ended safety evaluation on RealToxicityPrompts. Entries are median evaluations to reach width $\epsilon = 0.05$ over ten matched seeds.

Method	Median evaluations
CELEUS	**529.0**
CELEUS (W/O SURR)	765.0
EVALUE	745.5

CELEUS requires 29.0% fewer evaluations than EVALUE and 30.8% fewer than its no-surrogate variant to reach the same precision (Table 11). Its median estimate at stopping is 0.04708, close to the full-pool mean of 0.04667. The label savings therefore extend to evaluation of open-ended generations with a continuous safety score, in addition to classification and multiple-choice tasks.

L.2 Sensitivity to a Pre-Specified Cross-Entropy Bound

A valid bound for sequential evaluation need not equal the largest realized loss. We test this by separating dataset preprocessing from the bound supplied to the evaluation algorithm. Preprocessing removes examples with cross-entropy exceeding a pre-specified cutoff $C = 3$ and freezes the retained pool. The retained losses are unchanged. We then set $U \in \{3.0, 3.5, 4.0\}$ before sequential evaluation. Every tested U bounds the losses in this fixed pool because $U \geq C$; increasing U changes the confidence-sequence bound without changing the evaluation examples or their losses. The immediate evaluation target is the mean loss over the retained pool. Filtering changes the target relative to the original benchmark and requires inspecting every loss before sequential evaluation. The reported query counts cover the subsequent sequential phase; they do not include this preprocessing. A population interpretation additionally requires assumptions on how the retained examples represent the filtered population.

The study covers SST-2, MMLU, and AG News with four surrogate–target pairs. Two pairs evaluate Llama-3-70B using Llama-3-8B and Llama-2-7B as surrogates; the other two use Llama-3-8B to evaluate Qwen2.5-72B and Mixtral-8×7B. We compare CELEUS and EVALUE over 50 seeds per setting at $\epsilon \in \{0.05, 0.07, 0.10\}$, using nominal miscoverage $\alpha = 0.05$ with equal allocations to the finite-pool interval and population correction.

CELEUS requires fewer labels in all 27 dataset–bound–precision summaries in Table 12. At $U = 3$ and $\epsilon = 0.05$, the mean pair-median label counts are 27,245.6 versus 46,158.5 on SST-2, 10,823.6 versus 48,332.4 on MMLU, and 24,986.6 versus 84,662.4 on AG News. All trajectories reach each target before exhausting the retained pool. Larger bounds increase the evaluation budget: for example, increasing U from 3 to 4 raises the SST-2 count at $\epsilon = 0.05$ from 27,245.6 to 55,629.0 for CELEUS and from 46,158.5 to 67,760.5 for EVALUE. Thus, a tighter valid bound improves efficiency, while CELEUS retains a label advantage under the more conservative bounds.

The aggregate advantage coexists with variation across model pairs: CELEUS uses fewer labels in 102 of the 108 pair–bound–precision comparisons. The six reversals occur for Llama-3-8B→Mixtral-8×7B on SST-2 at $\epsilon = 0.07$ and 0.10, across the three values of U. The complete grid therefore shows both the broad efficiency gain and the settings in which the baseline stops earlier.

Table 12: Labels to target precision under pre-specified CE bounds. For each model pair we take the median over 50 seeds, then average these medians equally across the four pairs. Brackets report the corresponding average percentage of the retained pool, computed within each pair before averaging. The preprocessing cutoff is fixed at $C = 3$ throughout.

Dataset	U	Method	$\epsilon = 0.05$	$\epsilon = 0.07$	$\epsilon = 0.10$
SST-2	3.0	CELEUS	27,245.6 [40.2%]	6,579.0 [9.7%]	2,101.9 [3.1%]
		EVALUE	46,158.5 [68.1%]	24,675.5 [36.4%]	5,715.9 [8.4%]
SST-2	3.5	CELEUS	42,063.8 [62.0%]	9,140.8 [13.5%]	2,580.6 [3.8%]
		EVALUE	53,982.3 [79.6%]	33,275.4 [49.1%]	6,375.4 [9.4%]
SST-2	4.0	CELEUS	55,629.0 [82.0%]	13,013.0 [19.2%]	3,142.1 [4.6%]
		EVALUE	67,760.5 [99.8%]	37,380.4 [55.1%]	7,523.6 [11.1%]
MMLU	3.0	CELEUS	10,823.6 [9.5%]	3,498.8 [3.1%]	1,514.3 [1.3%]
		EVALUE	48,332.4 [42.5%]	11,148.9 [9.8%]	3,114.3 [2.7%]
MMLU	3.5	CELEUS	16,180.0 [14.2%]	4,373.5 [3.8%]	1,778.9 [1.6%]
		EVALUE	73,613.0 [64.7%]	13,391.8 [11.8%]	3,425.4 [3.0%]
MMLU	4.0	CELEUS	25,915.9 [22.8%]	5,459.1 [4.8%]	2,076.0 [1.8%]
		EVALUE	99,151.9 [87.1%]	16,797.5 [14.8%]	3,776.8 [3.3%]
AG News	3.0	CELEUS	24,986.6 [23.7%]	6,956.0 [6.6%]	2,363.4 [2.3%]
		EVALUE	84,662.4 [80.6%]	20,540.4 [19.5%]	5,583.9 [5.3%]
AG News	3.5	CELEUS	36,903.0 [35.1%]	8,623.3 [8.2%]	2,700.6 [2.6%]
		EVALUE	89,845.1 [85.9%]	24,497.6 [23.2%]	5,930.0 [5.6%]
AG News	4.0	CELEUS	57,291.9 [54.7%]	10,971.8 [10.4%]	3,123.8 [3.0%]
		EVALUE	102,529.1 [99.0%]	30,459.1 [28.9%]	6,548.6 [6.2%]

L.3 Computational Cost and the Break-Even Label Price

Label efficiency translates into computational savings when the cost of obtaining the saved labels exceeds the additional algorithmic work. We measure each method in isolation until its first crossing of width $\epsilon = 0.05$, covering three datasets, ten ordered surrogate-target pairs, and six seeds per pair. The ten pairs are the six pairs in Table 5 and four additional pairs: Llama-2-7B $\rightarrow$ Mixtral-8$\times$7B, Llama-2-7B $\rightarrow$ Qwen-2.5-72B, DeepSeek-67B $\rightarrow$ Llama-3-70B, and Qwen-2.5-72B $\rightarrow$ Llama-3-70B. All 360 runs reach the target before pool exhaustion. Timing begins with precomputed losses and surrogate scores and covers the online sampling, estimation, and confidence-sequence updates through stopping.

Let n_k be the number of labels and T_k the measured online runtime for method $k \in \{\mathrm{C}, \mathrm{E}\}$, denoting CELEUS and EVALUE. If one label costs p_L seconds-equivalent, the online evaluation cost is

$$\widetilde{T}_k(p_L) = T_k + n_k p_L. \tag{29}$$

For equal-weight means across model pairs and seeds, the break-even price is

$$p_L^\star = \frac{\bar{T}_\mathrm{C} - \bar{T}_\mathrm{E}}{\bar{n}_\mathrm{E} - \bar{n}_\mathrm{C}}, \qquad \bar{n}_\mathrm{E} > \bar{n}_\mathrm{C}. \tag{30}$$

CELEUS has lower mean online cost when $p_L > p_L^\star$. Table 13 reports this threshold and evaluates both methods at label costs of 1 and 10 milliseconds.

Table 13: Online cost to width $\epsilon = 0.05$. Break-even prices are in milliseconds per label; brackets give 95% hierarchical bootstrap intervals (20,000 replicates, resampling model pairs and then paired seeds). Total costs are in seconds and include measured algorithm time plus the indicated label-cost term. Each total is an equal-weight mean across ten pairs and six seeds.

	Break-even price	Total: CELEUS / EVALUE (s)	
Dataset	$p_L^\star$ [95% interval]	$p_L = 1$ ms	$p_L = 10$ ms
SST-2	0.92 [0.63, 1.81]	10.83 / 11.38	41.28 / 107.55
MMLU	2.24 [1.89, 2.77]	16.12 / 9.29	45.12 / 87.86
AG News	3.23 [2.37, 4.37]	36.71 / 16.66	96.78 / 157.82

The break-even prices range from 0.92 to 3.23 milliseconds per label. At 1 millisecond per label, CELEUS is already cheaper on SST-2, while EVALUE remains cheaper on MMLU and AG News.

At 10 milliseconds per label, CELEUS has lower cost on all three datasets, reducing the mean totals from 107.55 to 41.28 seconds on SST-2, from 87.86 to 45.12 seconds on MMLU, and from 157.82 to 96.78 seconds on AG News. The same ordering holds for the six-seed mean of each of the 30 dataset–pair settings at this price. These results quantify the regime in which the saved evaluations outweigh the online computational overhead.

When deployment also requires method-specific preparation, let T_k^{setup} denote its one-time cost, including surrogate inference where applicable. The complete cost becomes $T_k^{\mathrm{setup}} + T_k + n_k p_L$, and the corresponding mean break-even price is

$$p^{\star}_{L,\mathrm{setup}} = \frac{\bar{T}_{\mathrm{C}}^{\mathrm{setup}} - \bar{T}_{\mathrm{E}}^{\mathrm{setup}} + \bar{T}_{\mathrm{C}} - \bar{T}_{\mathrm{E}}}{\bar{n}_{\mathrm{E}} - \bar{n}_{\mathrm{C}}}. \tag{31}$$

The timing study measures T_k; T_k^{setup} captures deployment-dependent preparation. This separates the measured online trade-off from the additional cost of obtaining surrogate scores, and makes the benefit of reusing existing scores explicit.

M Limitations and discussion

We discuss the scope of the theoretical guarantees and several practical considerations for applying CELEUS. The confidence sets in Eqn. (4) are defined by testing every candidate value in the continuous interval $[L, U]$. In practice, we evaluate the e-processes on a finite grid and return the interval between the smallest and largest unrejected grid points. Theorems 5 and 6 apply to the exact confidence sets obtained from continuous test inversion, while the finite grid provides a computationally tractable approximation. Increasing the grid density or refining the grid near the estimated boundaries provides finer numerical resolution. App. M.1 provides a conservative alternative: it rejects whole grid cells using valid lower bounds and returns an outer enclosure of the exact confidence set. This enclosure inherits the coverage guarantee when the stated rounding rules are followed. This alternative is separate from the grid-point implementation used in the experiments. For grid-point inversion alone, a narrow unrejected region may fall between grid points and be missed, while the interval spanning the extreme unrejected grid points may include candidate values that would be rejected by exact inversion. Thus, the grid-based interval may be either narrower or wider than the exact confidence set. Assumption 1 also requires deterministic bounds L and U to be specified before observing the pool losses. This requirement is immediate for naturally bounded metrics such as 0–1 loss. For cross-entropy loss, a clipping level fixed before observing pool losses, as described in Remark 2, makes the guarantees concern the mean clipped loss. The empirical pool-maximum bound in App. G does not establish this population-level boundedness condition. The cutoff study in Sec. L.2 instead evaluates a retained pool after loss-based filtering and therefore changes the evaluation target. A larger clipping level retains more information from unusually large losses but increases the bounded range, whereas a smaller clipping level may yield tighter intervals but modifies the quantity being estimated. The clipping level should therefore be selected according to the evaluation objective before observing the pool losses.

The finite-pool confidence sequence targets R_N, the average risk over the given evaluation pool, and remains valid under adaptive sampling without replacement. Extending this guarantee to the population risk R requires the evaluation pool to consist of independent and identically distributed input-label pairs from the intended target population, with the target model fixed before the pool is drawn. Under this condition, the Hoeffding correction $\Delta_N(\alpha_2)$ controls the difference between R_N and R. If a benchmark is instead treated as a fixed or deliberately curated collection, the finite-pool guarantee for R_N remains valid and provides the natural interpretation of the result. A population-level guarantee would then require a correction suited to the process used to construct the benchmark. Because $\Delta_N(\alpha_2)$ depends on the full pool size N, rather than the number of evaluated samples t, evaluating more samples from the same pool does not reduce this correction. It becomes smaller only when the evaluation pool itself becomes larger and may therefore remain an appreciable part of the final interval for a small pool.

The validity of CELEUS does not require the surrogate or acquisition proxy to achieve any prescribed level of accuracy. The stated finite-pool theorems for practical sampling require that the surrogate scores are fixed before evaluation, acquisition scores depend only on observed information, the boundedness conditions hold, the proposal probabilities satisfy Assumption 2, and the actual sampling

probabilities are used in the inverse-probability correction. Nevertheless, their quality affects efficiency. More accurate surrogate scores can produce smaller residuals in the inferential signal, while acquisition scores that better approximate the residual magnitudes can direct evaluations toward more informative samples, as analyzed in App. A.4. The proposal in Theorem 4 is a positive-floor oracle benchmark that minimizes the conditional variance at a given round, conditional on the complete pool, preprocessing information, and previous selections. Because the true residuals are unknown before evaluation, practical strategies replace them with label-free uncertainty or disagreement scores, whose effectiveness may vary across tasks. The theorem therefore motivates these strategies but does not claim that they minimize the total number of evaluations over the entire sampling process or at the final stopping time. Surrogate scoring and online inference also incur computational costs. Reductions in target evaluations should therefore primarily be interpreted as improvements in label efficiency. Whether they also reduce total cost depends on the relative costs of surrogate inference, online computation, and target labels, as quantified by the break-even analysis in App. J.

The empirical evaluation covers classification and multiple-choice tasks and one open-ended generation task with a continuous safety score (App. L.1). Evaluation settings based on human-preference judgments are not tested in this paper.

Finally, several design choices affect the tightness of the resulting CIs. Time-uniform inference allows the intervals to be monitored repeatedly and the evaluation to be stopped based on the observed data without invalidating the coverage guarantee. The probability floor prevents inverse-probability corrections from becoming excessively large, while predictable scaling bounds and betting fractions ensure that each e-process update remains valid. Splitting the error budget between α_1 and α_2 separately controls uncertainty from evaluating only part of the pool and uncertainty from using a finite pool to represent the population. These safeguards provide finite-sample guarantees under broad conditions, but they may produce wider intervals than methods designed for a prespecified evaluation time or methods that impose stronger distributional assumptions. The width result in Theorem 5 should therefore be understood as a general upper bound on how the uncertainty decreases, rather than as an exact prediction of the number of evaluations required to reach the stopping criterion. The numerical error introduced by the finite grid is separate from this theoretical width bound. In particular, the observed miscoverage below α in Tabs. 1 and 6 reflects two sources of conservativeness that require no additional distributional assumptions: the time-uniform e-process construction [21] and the finite-pool-to-population correction based on Hoeffding's inequality [24]. Compared with the $O(t^{-1/2})$ width typical of a CI designed for a prespecified evaluation time, the additional logarithmic factors in Thm. 5 are the price of validity at every round and under data-dependent stopping. Tighter constructions that retain finite-sample, anytime-valid guarantees are left for future work.

M.1 Discussion: conservative numerical inversion

This subsection discusses an optional numerical alternative to the grid-point inversion in Alg. 2. It is not part of that algorithm, and we do not claim experimental results for this alternative. The test statistic remains $\mathcal{K}_t^{\pm}(\upsilon)$ in Eqn. (3), and the continuous confidence set remains Eqn. (4). The quantity $\underline{\mathcal{K}}_{t,j}^{\mathrm{lo}}$ below is only a computational bound used to check a range of candidate values.

Let $L = \upsilon_1 < \cdots < \upsilon_K = U$ be a fixed equally spaced grid with spacing $(U-L)/(K-1)$. A grid cell is the range $[\upsilon_j, \upsilon_{j+1}]$ between adjacent grid points. Write $\mathcal{K}_{t,+}$ and $\mathcal{K}_{t,-}$ for the two products in Eqn. (3), and define

$$\underline{\mathcal{K}}_{t,j} = \theta \mathcal{K}_{t,+}(\upsilon_{j+1}) + (1-\theta)\mathcal{K}_{t,-}(\upsilon_j). \tag{32}$$

With the fractions in Remark 3, $\mathcal{K}_{t,+}(\upsilon)$ cannot increase as υ increases, whereas $\mathcal{K}_{t,-}(\upsilon)$ cannot decrease. To see this, note that before clipping, the factors are $1 + \tilde{\lambda}_t^{+}(\bar{S}_t - \bar{\upsilon}_t)$ and $1 - \tilde{\lambda}_t^{-}(\bar{S}_t - \bar{\upsilon}_t)$; after clipping, they are $(\bar{S}_t + c)/(\bar{\upsilon}_t + c)$ and $(1 - \bar{S}_t + c)/(1 - \bar{\upsilon}_t + c)$, respectively. Each positive factor has the stated direction of change, and its branches agree at the clipping point. The products therefore have the same property. Consequently,

$$\underline{\mathcal{K}}_{t,j}^{\mathrm{lo}} \le \underline{\mathcal{K}}_{t,j} \le \mathcal{K}_t^{\pm}(\upsilon) \text{ for every } \upsilon \in [\upsilon_j, \upsilon_{j+1}],$$

where $\underline{\mathcal{K}}_{t,j}^{\mathrm{lo}}$ is a computed lower bound guaranteed to remain valid despite rounding. If this lower bound is at least $1/\alpha_1$, the original test rejects every candidate in the cell. Otherwise, the cell is retained; it may contain values rejected by the original test, since the bound is conservative.

For this optional implementation, initialize $\mathcal{A}_0^{\mathrm{cell}} = \{1, \ldots, K-1\}$ and retain

$$\mathcal{A}_t^{\mathrm{cell}} = \{j \in \mathcal{A}_{t-1}^{\mathrm{cell}} : \underline{K}_{t,j}^{\mathrm{lo}} < T^{\mathrm{up}}\},$$

where T^{up} is an upper bound on $1/\alpha_1$ that accounts for rounding. When cells remain, report their outside edges:

$$I_t = [v_{\min \mathcal{A}_t^{\mathrm{cell}}}, v_{1+\max \mathcal{A}_t^{\mathrm{cell}}}].$$

Every candidate in $\mathcal{M}_t^{\pm}$ lies in a cell whose lower bound stayed below the rejection threshold at every preceding round. Thus,

$$\mathcal{M}_t^{\pm} \subseteq I_t. \tag{33}$$

The interval I_t is an outer enclosure, meaning an interval containing the whole continuous confidence set. If no cell remains, use $I_t = [L, U]$ and do not declare the target width attained. Applying the original population correction to I_t preserves coverage, but the wider numerical interval need not satisfy the exact-inversion width bound.

For finite-precision calculations, interval arithmetic propagates lower and upper bounds through each calculation. These bounds must include accumulated errors in the signal, scaling, betting fractions, and capital updates. Retain a cell whenever rounding prevents certified rejection. Use an upper bound on Δ_N, round reported endpoints outward, and stop only when an upper bound on the full reported population-interval width is at most ϵ. The score-normalization tolerance δ_{num} does not control these inversion errors.

The cell check reuses the existing endpoint capitals. For fixed arithmetic precision, the grid computation per round and storage remain $O(K)$, with an additional constant-factor cost for interval arithmetic. The grid size and arithmetic precision should be recorded. A coarse grid or loose numerical bounds may delay stopping or prevent the target width from being reached. Stopping, coverage, and timing results would need to be evaluated for this optional implementation separately.

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 problem, the proposed e-process-based framework, the main theoretical guarantees, and the empirical claims. The claims are matched by the theoretical results and experiments reported in the paper.

   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: App. M discusses the limitations and the scope of the theoretical guarantees, including numerical test inversion on a finite grid, loss bounds fixed before observing the pool losses, the i.i.d. condition for population-risk guarantees, the computational costs of surrogate scoring and online inference, and the empirical scope of the experiments. The conclusion also outlines distribution-shift-aware evaluation as a future direction.

   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 state their assumptions explicitly, and full proofs are provided in the appendix. The results are cross-referenced with the corresponding definitions, assumptions, and proof details.

   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: The paper specifies the datasets, surrogate-target model pairs, baselines, hyperparameters, stopping rule, evaluation metrics, and Monte Carlo protocols needed to reproduce the main experimental results.

   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: We provide code and instructions in https://github.com/zyecs/celeus.

   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: The paper reports the datasets, evaluation pools, surrogate-target pairs, baselines, loss functions, hyperparameters, stopping criterion, and sample-selection strategies in the experimental setup and appendix.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The experiments report results over multiple independent seeds or Monte Carlo trials, including Wilson confidence intervals, mean trajectories, and variability bands where appropriate. The caption and text specify how the reported intervals or bands are computed.

   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: The experiments are conducted on a high-performance computing cluster equipped with multiple NVIDIA A100 GPUs using the PyTorch framework. Each GPU has 80 GB of memory, and the available disk storage exceeds 4 TB.

   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 research uses public datasets and pre-trained models for evaluation, does not involve collecting sensitive personal data, and is intended to improve reliable and statistically certified LLM evaluation.

   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 discusses that certifiable LLM evaluation can improve reliable deployment decisions by providing anytime-valid confidence intervals, without any negative societal impacts.

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 high-risk models, scraped datasets, or generative systems. It only evaluates existing models using 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: The paper cites the public datasets and pre-trained models used in the experiments. Their usage follows the corresponding public access terms and licenses.

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 or release a new dataset, model, or benchmark asset.

    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 conduct new crowdsourcing experiments or human-subject studies. It uses existing public datasets and benchmarks.

    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 new human-subject data collection or intervention, 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: [Yes]

Justification: LLMs are central to the research setting and experiments. The paper evaluates target LLMs and uses smaller pre-trained surrogate LLMs to provide surrogate scores for unevaluated samples. These roles are described in the experimental setup, including the surrogate-target model pairs.

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.