# 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]

## Abstract

Can we trust evaluation scores to capture an LLM's true real-world performance? *Certifiable* evaluation answers this question by providing guarantee for LLM evaluation. In particular, existing methods sequentially curate evaluation samples and keep updating confidence intervals (CIs) that cover the true performance with high probability (e.g., 95%) until some 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 signals that combine 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 such signals remain 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.

[*]Equal contribution [1]The University of Melbourne [2]Department of Computer Science, University of Chicago [3]Department of Computer Science, University of Illinois Urbana-Champaign. Correspondence to: Feng Liu <fengliu.ml@gmail.com>.

*ICML 2026 Workshop on Hypothesis Testing*, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

## 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 curating evaluation samples and updating confidence intervals (CIs) that cover the true performance with high probability (e.g., 95%) [9]. The updating process will end until some conditions are satisfied, e.g., the CI width reaches a target precision. A common line of evaluation with CIs considers static settings (i.e., a special case in sequential setting), where the evaluation samples are fixed in advance and a one-shot CI is reported for model 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. More recently, Wang et al. [9] proposed Cer-Eval, which moves closer to certifiable evaluation in a general sequential setting, where an evaluator is given a data pool and sequentially selects samples for evaluation.

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. Taking a state-of-the-art method as an example, Cer-Eval's certification relies on adaptive concentration inequalities for a fixed partition, whereas the practical algorithm repeatedly learns and updates the partition from the evaluated data. This induces double dipping, where the same data are used to choose the partition and quantify uncertainty within it, creating a mismatch between the fixed-partition analysis and the adaptive

procedure used in practice.

**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. A key advantage of $e$-processes is that, from the application perspective, their guarantees rely only on the simple assumption that the evaluation score is bounded [12], which can usually be ensured by normalization or clipping, making $e$-processes well suited for certified LLM evaluation in real-world settings[1]. However, directly applying $e$-processes is not enough for efficient evaluation. Practical LLM evaluation must also reduce the cost of high-quality evaluation, which grows with the number of evaluated samples, especially for open-ended generation and human-preference benchmarks [5, 13, 14]. Given a fixed data pool, efficiency strategies often select informative samples for evaluation [15] and use surrogate-assisted approximations for unevaluated samples [10]. Under such non-uniform selection, 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 sequential evidence must be built from a signal that remains conditionally unbiased for the evaluation score as the evaluation proceeds.

**Technical Contributions.** To cope with the above key obstacle, CELEUS builds inferential signals around two ingredients: (i) uncertainty-guided sampling, which prioritizes informative samples for evaluation, and (ii) surrogate-assisted approximations, which provide approximate scores for samples that have not been evaluated. Concretely, at each round, the signal first forms a surrogate-completed estimate of the finite-pool evaluation risk [2] by combining the observed risks of previously evaluated samples with surrogate scores for the remaining samples. Then, after the uncertainty-guided sampling rule selects a new sample and its risk is observed, the signal adds an inverse-probability-weighted correction for that sample's surrogate residual. Together, these two ingredients yield a conditionally unbiased signal for the finite-pool evaluation risk under sample selection (see Thm. 3, Sec. 4). This conditional unbiasedness is the central validity mechanism that enables statistically grounded, anytime-valid $e$-process CIs through test inversion [16].

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 $8 \times$ 7B MoE model.

## 2. Certifiable and Efficient Evaluation: Setup and Overview

In this section, we introduce the evaluation setup and give a high-level overview of the certifiable and efficient evaluation procedure 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 $f : \mathcal{X} \to \mathcal{Y}$ 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 on the task $\mathbb{P}_{\boldsymbol{x},y}$:

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

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 a data pool $D_N = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ consisting of $N$ *i.i.d.* samples[3] drawn from the distribution $\mathbb{P}_{\boldsymbol{x}}$ on $\mathcal{X}$. *Certifiable evaluation* sequentially evaluates the samples of $D_N$ and stops once a target precision is achieved. Because evaluating a sample can be costly, especially when human judgments are required, *efficient evaluation* aims to estimate accurately using as few evaluated samples as possible.

A natural idea is to *select informative samples for evaluation*, which has been extensively studied in the literature on active

[1] All other assumptions required for the theoretical guarantees are algorithmic conditions that can be directly satisfied through the design of the $e$-process testing procedure and impose no additional restrictions on the application scenario.

[2] CELEUS lifts the CIs of finite-pool risk constructed from the signal to the population risk via a Hoeffding-style correction that depends on the *full pool size*. This correction is small or even negligible for large pools; for small pools, CELEUS can evaluate all samples, leaving only the population correction and yielding a CI comparable to static evaluation over the full pool.

[3] The *i.i.d.* assumption is only used when applying Hoeffding's inequality to transfer the guarantee from $R_N$ to $R(f; \mathbb{P}_{\boldsymbol{x},y})$; for non-*i.i.d.* or curated pools, CELEUS still provides a finite-pool guarantee for $R_N$, while the population-level interpretation requires an appropriate distribution-dependent generalization bound [17, 18]. More discussion can be found in App. A.4.

testing [15, 19]. Rather than evaluating samples uniformly at random, active testing chooses samples that are expected to be more informative, with the aim of improving statistical efficiency and reducing evaluation cost. When evaluation proceeds sequentially, at each round $t = 1, 2, \ldots$, the sampling policy chooses one unevaluated sample from the pool $D_N$ for evaluation, and the evaluator may stop based on the current information. We index evaluated samples by their evaluation order: after round $t$, the evaluated data are $\{(\boldsymbol{x}_{i_m}, y_{i_m})\}_{m=1}^{t}$, and the remaining unevaluated pool is $\{\boldsymbol{x}_j\}_{j \in \mathcal{J}_t}$, where $\mathcal{J}_t = [N] \setminus i_{1:t}$ and $i_{1:t} = \{i_1, \ldots, i_t\}$.

Another effective strategy is to use *surrogate-assisted approximations for unevaluated samples* [10, 11], which approximates the true risk $\ell(f(\boldsymbol{x}), y)$ with a surrogate score $\tilde{\ell}_f(\boldsymbol{x}) \in [L, U]$. For example, $\tilde{\ell}_f(\boldsymbol{x})$ can be produced by a model that predicts the true risk $\ell(f(\boldsymbol{x}), y)$ directly, or by computing the risk using a proxy label $\hat{y}(\boldsymbol{x})$, namely $\tilde{\ell}_f(\boldsymbol{x}) = \ell(f(\boldsymbol{x}), \hat{y}(\boldsymbol{x}))$. The surrogate scores allow us to extract information from the unevaluated examples in the pool: if it is close to the true risk, we can obtain an accurate estimate even with few evaluated samples.

Given the selected evaluation samples and the surrogate-assisted approximations for unevaluated samples, the key question is how to construct CIs that remain anytime-valid as evaluation proceeds. Standard static evaluation, based on a fixed set of evaluated samples, reports a CI for a single evaluation run [10, 11] and can lose validity when the CI is repeatedly updated and used to decide when to stop. Alternatively, one may use adaptive concentration inequalities [9, 20], but these results typically rely on *i.i.d.* samples and are not directly compatible with the efficiency strategies for LLM evaluation[4]. These requirements are directly matched by $e$-processes [16], which provide a principled way to build anytime-valid CIs under the simple assumption that the evaluation score is bounded [12]. We recall the $e$-process, together with its anytime-valid implication.

**Definition 1.** *[21] Let $\mathcal{H}_0$ be a null hypothesis, and let $\{\mathcal{F}_t\}_{t \geq 0}$ describe the information available up to time $t$. A nonnegative process $\{\mathcal{K}_t\}_{t \geq 0}$ is called an* $e$-process *for $\mathcal{H}_0$ if $\mathcal{K}_0 = 1$, each $\mathcal{K}_t$ can be computed using only the information in $\mathcal{F}_t$, and $\mathbb{E}_P[\mathcal{K}_t \mid \mathcal{F}_{t-1}] \leq \mathcal{K}_{t-1}$, $\forall P \in \mathcal{H}_0$, $\forall 1 \leq t \leq N$. Consequently, for any $\alpha \in (0, 1)$, it satisfies the anytime-valid control*

$$\mathbb{P}_P \left( \sup_{t \geq 0} \mathcal{K}_t \geq \frac{1}{\alpha} \right) \leq \alpha, \qquad \forall\, P \in \mathcal{H}_0.$$

*Thus, rejecting $\mathcal{H}_0$ whenever $\mathcal{K}_t \geq 1/\alpha$ gives an anytime-valid test: if $\mathcal{H}_0$ is true, the probability of falsely rejecting it at any time during the sequential procedure is at most $\alpha$.*

To turn $e$-processes into CIs, we use test inversion for the data-pool risk $R_N = \frac{1}{N} \sum_{j=1}^{N} \ell(f(\boldsymbol{x}_j), y_j)$[5]. For each candidate value $\upsilon$, we test $\mathcal{H}_0(\upsilon) : R_N = \upsilon$ by constructing an $e$-process $\{\mathcal{K}_t(\upsilon)\}_{t \geq 0}$. Intuitively, $\mathcal{K}_t(\upsilon)$ measures how strongly the information collected up to round $t$ contradicts the candidate value $\upsilon$. If $\mathcal{K}_t(\upsilon) \geq 1/\alpha$, then $\upsilon$ is rejected. Therefore, the CI at round $t$ is formed by keeping all candidate values that have not been rejected: $\mathrm{CI}_t = \{\upsilon : \mathcal{K}_t(\upsilon) < 1/\alpha\}$. The $\mathrm{CI}_t$ covers the true value $R_N$ with probability at least $1 - \alpha$, since $R_N$ corresponds to the true null hypothesis $\mathcal{H}_0(R_N)$. Thus, non-coverage can occur only if this true null is falsely rejected:

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

For certifiable and efficient LLM evaluation, the key technical step is to construct valid inferential signals from the evaluated samples and the 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 $\mathcal{K}_t(\upsilon)$ satisfying $\mathbb{E}_P[\mathcal{K}_t(\upsilon) \mid \mathcal{F}_{t-1}] \leq \mathcal{K}_{t-1}(\upsilon)$ [12, 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. Finally, when CI reaches the target precision, CELEUS reports the final performance estimate by averaging the per-round signals.

## 3. The CELEUS Evalution Method

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

**Definition 2.** *Fix a significance level $\alpha \in (0, 1)$. A sequence of random sets $\{C_t\}_{t \geq 1}$ is called a sequence of* $(1 - \alpha)$ anytime-valid and shrinking CIs *for $R$ if, for every distribution $\mathbb{P}_{\boldsymbol{x},y}$,*

$$\mathbb{P}_{\boldsymbol{x},y}(\exists 1 \leq t \leq N :\ R \notin C_t) \leq \alpha,$$

*and its width* $\mathrm{width}(C_t) = \sup C_t - \inf C_t$ *satisfies* $\mathrm{width}(C_t) \to 0$ *as* $t \to \infty$.

The shrinkage condition means that, as more samples are evaluated, $C_t$ becomes increasingly informative: its width decreases toward zero, progressively reducing the uncertainty about $R$; hence, the target precision can be guaranteed to be achieved with enough samples.

[4] More discussion of related work is provided in App. A.1.

[5] Conditional on a data pool, the randomness from drawing the pool is removed, and the remaining randomness comes only from the sequential evaluation procedure. Therefore, the resulting CIs provide inference for the finite-pool risk $R_N$.

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 $\alpha$ into $\alpha_1, \alpha_2 \in (0,1)$ with $\alpha_1 + \alpha_2 = \alpha$. Here, $\alpha_1$ controls anytime-valid inference for $R_N$ through test inversion of $e$-processes, while $\alpha_2$ controls the finite-pool-to-population correction from $R_N$ to $R$.

**Anytime-valid CIs for $R_N$ at level $\alpha_1$ 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, $\{\tilde{\ell}_f(\boldsymbol{x}_j)\}_{j\in\mathcal{J}_{t-1}}$. [6] Based on this information, the sampling policy assigns selection probabilities to the remaining unevaluated samples; we denote by $q_t(i_t \mid i_{1:t-1}, D_N)$ the conditional probability of selecting index $i_t$ from $\mathcal{J}_{t-1}$ at round $t$, given the previously selected indices $i_{1:t-1}$ and the data pool $D_N$. 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

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

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 $1/q_t(i_t \mid i_{1:t-1}, D_N)$ turns this residual into an unbiased estimate of the normalized total residual over the remaining pool. Hence, conditional on the past, $\hat{S}_t$ 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 [12], 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[7] lower and upper bounds of $\hat{S}_t$ given the information available before round $t$. Accordingly, for each candidate value $\upsilon$ 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 $\mathbb{E}[\bar{S}_t \mid \mathcal{F}_{t-1}, \mathcal{H}_0(\upsilon)] = \bar{\upsilon}_t$. This leads to a centered sequential signal: for each candidate $\upsilon$, 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 $\upsilon$, we use a *hedged capital process* [12, Section 4.4]. For any $\theta \in (0,1)$ and candidate $\upsilon$, define

$$\begin{aligned}\mathcal{K}_t^{\pm}(\upsilon) &= \theta \prod_{i=1}^{t} \left(1 + \lambda_i^{+}(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right)\right) \\ &+ (1-\theta) \prod_{i=1}^{t} \left(1 - \lambda_i^{-}(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right)\right). \end{aligned} \tag{2}$$

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[8]

$$\mathcal{M}_t^{\pm} = \bigcap_{i\le t} \left\{\upsilon \in [L, U] : \mathcal{K}_i^{\pm}(\upsilon) < 1/\alpha_1\right\}. \tag{3}$$

**Finite-pool-to-population correction at level $\alpha_2$.** 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 $\alpha_2$. Specifically, by Hoeffding's

[6] CELEUS mainly improves efficiency by reducing the number of samples that require target-model evaluation, which is often the main source of labelled-data or annotation cost in certifiable evaluation. However, surrogate-assisted evaluation introduces additional cost for obtaining surrogate scores. In this paper, we use fixed 7–8B open-source models as surrogates without task-specific training or online updates, which keeps this additional cost relatively small. More discussion can be found in App. A.5.

[7] 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.

[8] In practice, we approximate the continuum of candidate values for the finite-pool risk by a finite grid on the original scale.

inequality [22], 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

$$\left\{\forall 1 \le t \le N:\ R_N \in \mathcal{M}_t^{\pm}\right\} \cap \left\{|R_N - R| \le \Delta_N(\alpha_2)\right\},$$

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. 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 at the first time $t$ such that $\mathrm{width}(C_t) \le \epsilon$, and returns $C_t$ as the final CI, which covers the true performance with probability at least $1-\alpha$. Besides the CI, CELEUS also reports a final performance estimate by aggregating the per-round inferential signals (see Appendix A.3 for details):

$$\begin{aligned}\hat{R}_t &= \frac{1}{N}\sum_{n=1}^{N} \tilde{\ell}_f(\boldsymbol{x}_n) \\ &+ \frac{1}{t}\sum_{m=1}^{t} u_{m,t}\left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right),\end{aligned} \tag{4}$$

where $u_{m,t}$, motivated by the LURE estimator [23], adjusts the surrogate residual for sequential, non-uniform sample selection up to time $t$:

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

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. Specifically, each evaluated sample $i_m$ reveals the residual $\ell(f(\boldsymbol{x}_{i_m}), y_{i_m}) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})$, and the weight $u_{m,t}$ adjusts this discrepancy for the sequential, non-uniform sample selection. Thus, $\hat{R}_t$ combines surrogate information from all samples with weighted corrections, yielding a fixed-time unbiased estimator.

Following the previous $e$-process [12], CELEUS involves several practical hyperparameters when implementing the betting strategies and constructing the confidence intervals, such as the hedge weight $\theta$, predictable betting fractions, and scaling bounds. These choices mainly affect empirical efficiency rather than the anytime-valid coverage guarantee under the stated conditions, and we directly adopt the default choices from this prior $e$-process approach whenever applicable. Due to page limitations, we defer the detailed practical choices and implementation details of CELEUS to App. A.2, with the experimental setup provided in Sec. 5.1.

## 4. Theoretical Analysis

Proofs and assumptions are deferred to App. B. Assumption 1 only requires bounded evaluation scores, which can be ensured by normalization or clipping. All other assumptions are algorithmic conditions satisfied by CELEUS and impose no additional restrictions on the application.

**Risk estimator.** We first establish the fixed-time unbiasedness and consistency of the risk estimator. We emphasize that the unbiasedness result is a fixed-time property and does not, by itself, imply unbiasedness after data-dependent stopping. In CELEUS, the stopping rule and certification rely on the anytime-valid confidence interval, while the point estimate reported at stopping is used as a summary statistic.

**Theorem 1.** *Under Assumption 1-2, the estimator $\hat{R}_t$ has $\mathbb{E}[\hat{R}_t] = R$ with $\mathrm{Var}[\hat{R}_t] = \mathrm{Var}(\ell(f(\boldsymbol{x}), y))/N + \mathbb{E}\left[\mathrm{Var}\left[\sum_{m=1}^{t} c_m a_m/t \mid D_N, \tilde{\ell}_f\right]\right]$, where $c_m$ and $a_m$ are defined in Eqn. (5) and Eqn. (6).*

The next theorem further shows that estimator is consistent.

**Theorem 2.** *Under Assumption 1-2, for any sequence $\{t_N\}_{N\ge1}$ satisfying $1 \le t_N \le N$, $N \to \infty$, 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.

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

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

**Variance-driven efficiency.** The variance above 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 and

the surrogate risk scores 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 variance of $\hat{R}_t$ and $\hat{S}_t$:* [9]

$$q_t^*(j \mid i_{1:t-1}, D_N) = \frac{\left|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|}{\sum_{s \in \mathcal{J}_{t-1}} \left|\ell(f(\boldsymbol{x}_s), y_s) - \tilde{\ell}_f(\boldsymbol{x}_s)\right|}.$$

*In contrast, if all $\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)$ are zero, then every proposal distribution $q_t$ 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, the sequence $\left(\mathcal{M}_t^{\pm}\right)_{1 \le t \le N}$, defined in Eqn. (3), is a sequence of $(1-\alpha_1)$ anytime-valid, shrinking CIs for $R_N$. Moreover, for $t \ge t_0$, define*

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

*Then, for some constant $G_6 > 0$ and all $t \ge t_0$,*

$$\mathrm{width}\left(\mathcal{M}_t^{\pm}\right) \le G_6 \sqrt{\frac{\log t \log\log t}{t} B_t},$$

*where $t_0$ is defined in Eqn. (31) (App. B.6). It follows that* $\mathrm{width}\left(\mathcal{M}_t^{\pm}\right) = O\left(\log\log t \sqrt{\log t / t}\right)$.

The bound shows that the CI width is driven by the cumulative conditional variance of the per-round signals. By Thm. 3, this variance is mainly affected by the surrogate residuals $\ell(f(\boldsymbol{x}), y) - \tilde{\ell}_f(\boldsymbol{x})$ and the sampling distribution $q_t$. Without surrogate information, e.g., $\tilde{\ell}_f(\boldsymbol{x}) = 0$, residuals can be large; under uniform sampling, the procedure cannot prioritize high-variance samples. Accurate surrogate scores and suitable non-uniform sampling reduce these residual and sampling effects, thereby shrinking the CI faster and reaching the target precision with fewer evaluated samples.

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

**Theorem 6.** *Fix a level $\alpha \in (0,1)$ and choose any split $\alpha_1, \alpha_2 \in (0, \alpha)$ with $\alpha_1 + \alpha_2 = \alpha$. Under Assumptions 1-4, the sequence $(C_t)_{1 \le t \le N}$ is a sequence of $(1-\alpha)$ anytime-valid, shrinking CIs for the population risk $R$ with width* $\mathrm{width}\,(C_t) = O\left(\log\log t \sqrt{\log t / t}\right) + O\left(\sqrt{1/N}\right)$.

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 also provide validation in our experiments.

## 5. Experiments

In this section, we evaluate CELEUS on real LLM evaluation tasks. Specifically, we sequentially curate evaluation samples, update the CI after each round, and stop once the CI width reaches the target precision $\epsilon = 0.05$. Under this setup, Sec. 5.2 tests the anytime-valid guarantee: with $\alpha = 0.05$, the CI should cover the true performance with probability at least $1-\alpha$ at the stopping time when the CI width reaches the target precision. Beyond theoretical rigor, we evaluate the efficiency of CELEUS by measuring how many samples are evaluated before reaching the target precision, and also study the contribution of CELEUS's key design components. Sec. 5.1 describes the detailed experimental setup. Sec. 5.3 provides additional analyses. Our code is available at: https://github.com/zyecs/celeus.

### 5.1. Setup

**Datasets.** Following Berrada et al. [15], 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) [25] for binary sentiment classification, AG News Topic Classification (AG News) [26] for multi-class topic classification, and Massive Multitask Language Understanding (MMLU) [1], a broad multiple-choice benchmark that tests LLMs across a wide range of academic subjects, including mathematics, law, history, and other domains requiring factual knowledge and reasoning. For each benchmark, all available samples are treated as the evaluation pool $D_N$.

**Surrogate-target pairs.** The target model is the model whose performance we aim to evaluate, and the surrogate model is used to provide surrogate scores for unevaluated samples. We evaluate multiple target models, including Llama-3-70B, DeepSeek-67B, Mixtral-8×7B, and Qwen-2.5-72B, covering different model families and generations. We instantiate *six* surrogate-target pairs across model-family and generation axes. Specifically, Llama-3-8B is used as the

[9] 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 [24].

**Table 1.** Empirical miscoverage at the stopping time when the CI width reaches $\epsilon = 0.05$ in the synthetic setting of App. C. 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 |

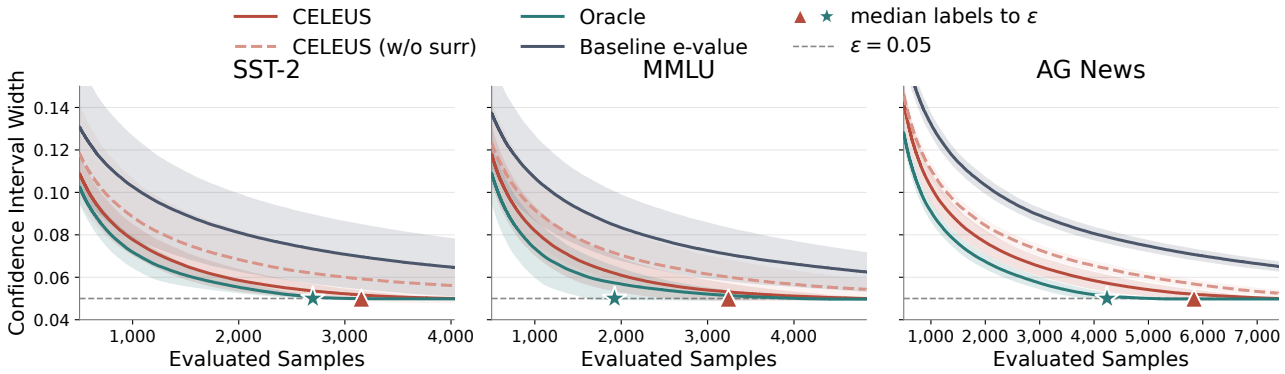


**Figure 1.** CI width versus the number of evaluated samples. Each trajectory is averaged over 50 seeds and surrogate–target pairs, with shaded bands showing $\pm 1$ standard deviation. The dashed line marks $\epsilon = 0.05$. At the budget where ORACLE reaches $\epsilon$, CELEUS reduces the CI-width gap between EVALUE and ORACLE by 90%/93%/84% on SST-2/MMLU/AG News, compared with 53%/60%/60% for CELEUS (W/O SURR).

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.

**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 [12] under uniform sampling without a surrogate. ② CELEUS (W/O SURR) ablates the surrogate component of CELEUS by setting $\tilde{\ell}_f \equiv 0$ while retaining the uncertainty-guided sampling rule, isolating the contribution of surrogate-assisted approximation. ③ ORACLE uses the variance-optimal sampling rule $q_t^\star$ in Thm. 4, which requires the per-sample residuals and is unknown in practice; it serves as an oracle reference for the best sampling efficiency attainable by CELEUS.

**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. F. For CELEUS, we use the uncertainty-guided sampling strategies from Rems. 1 and 2. For the no-surrogate baselines, i.e., EVALUE and CELEUS (W/O SURR), we use uniform sampling over the unevaluated pool. Detailed results for different sample-selection strategies are provided in App. G.

**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 [12], we use predictable fractions in Rem. 3 with $c = 0.5$, set $\theta = 0.5$ in Eqn. (2), 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. D.4. We stop the evaluation at the first time $t$ such that $\mathrm{width}(C_t) \leq \epsilon = 0.05$.

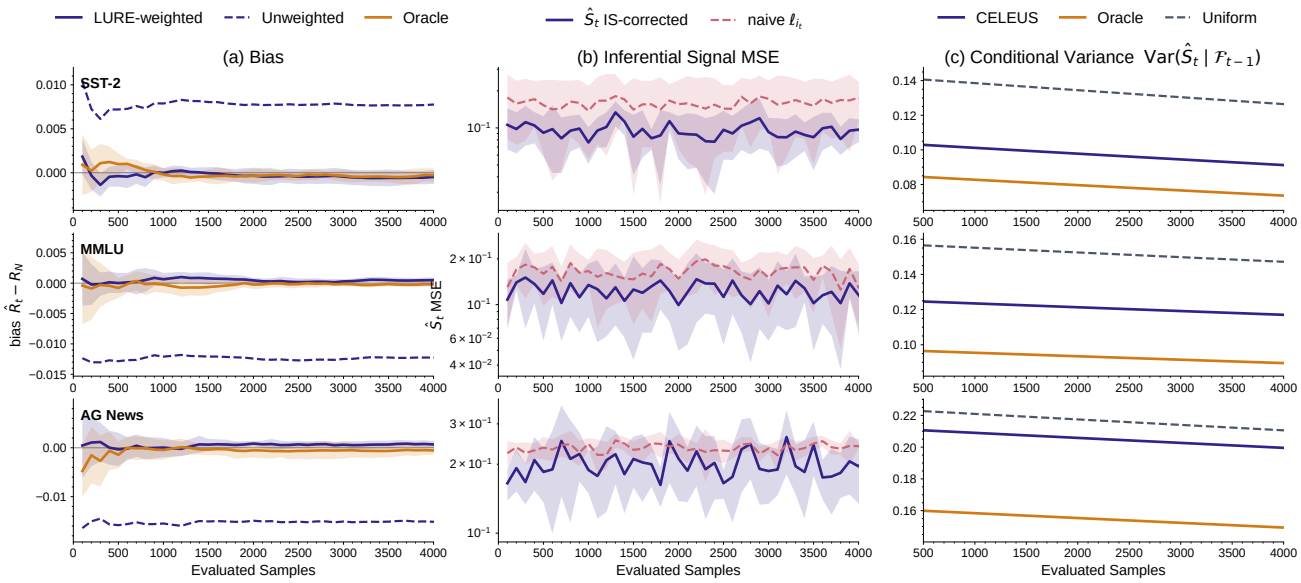


**Figure 2.** Mechanism analysis for CELEUS, pooled across the six surrogate-target pairs. *Left:* empirical bias of the final estimator $\hat{R}_t$ in Eqn. (4) versus the unweighted estimator that removes $u_{m,t}$. *Middle:* MSE of the proposed signal $\hat{S}_t$ in Eqn. (1) versus the naive single-step observed risk of the newly evaluated sample. *Right:* $\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1})$ under uniform sampling, CELEUS's uncertainty-guided sampling, and the oracle $q_t^\star$ in Thm. 4.

### 5.2. Main Results

We now evaluate three claims about CELEUS under the setup of Sec. 5.1: its CIs retain anytime-valid coverage rate at the stopping time when the CI width reaches the target precision $\epsilon = 0.05$ **(RQ1)**; it reaches precision $\epsilon$ with fewer evaluated samples than compared methods **(RQ2)**; and the final performance estimator Eqn. (4), the inferential signal Eqn. (1), and the uncertainty-guided sampling strategy motivated by Thm. 4 each contribute to this efficiency gain **(RQ3)**.

**RQ1: Anytime-valid coverage.** 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. C 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: Efficiency.** We compare CELEUS by CI width and the number of evaluated samples needed to reach $\epsilon = 0.05$. 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. Compared with CELEUS (W/O SURR), it reduces the required labels by 10-23% using 7-8B surrogate models without task-specific training or fine-tuning. Figure 1 shows that this gain holds anytime: CELEUS's CI width remains below EVALUE and CELEUS (W/O SURR) at every budget. At the budget where ORACLE first reaches $\epsilon$, CELEUS closes 90%/93%/84% of the EVALUE-to-ORACLE width gap, compared with 53%/60%/60% for CELEUS (W/O SURR). This matches Thm. 5, as the CI width is controlled by the cumulative conditional variance of the signals, which is minimized by the oracle variance-optimal strategy. App. E reports miscoverage results, with all methods below $\alpha = 0.05$.

**RQ3: What makes CELEUS work.** We isolate three key designs of CELEUS: the weighted final performance estimator, the surrogate-corrected inferential signal, and the uncertainty-guided sampling rule, and examine how they contribute to certifiable and efficiency LLM evaluation.

① *Estimator unbiasedness.* Fig. 2 (left) compares the empirical bias of $\hat{R}_t$ in Eqn. (4) 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 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. (1) 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× lower mean squared error (MSE) across $t$. This improvement comes from the structure of $\hat{S}_t$: it combines

previously observed risks, surrogate scores for unevaluated samples, and an inverse-probability-weighted correction from the newly evaluated sample. Thus, 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 $q_t^\star$. Compared with uniform sampling, CELEUS reduces the variance by 27%, 20%, and 5% on SST-2, MMLU, and AG News, and closes $\approx$ 67%, $\approx$ 53%, and $\approx$ 18% of the uniform-to-ORACLE variance gap. This matches Thm. 4: sampling probabilities that better reflect surrogate-residual magnitudes reduce variance more effectively. Moreover, the reduction in $\mathrm{Var}(\hat{S}_t \mid \mathcal{F}_{t-1})$ is exactly the variance that enters the width bound of Thm. 5, explaining the efficiency gain.

Taken together, the three sub-claims under RQ3 close the loop: LURE weights reduce bias, surrogate scores improve signal quality, and uncertainty-guided sampling reduces conditional variance. Combining RQ1 and RQ2, they complete empirical confirmation of theoretical claims of CELEUS.

### 5.3. Additional Experiments

In addition to the three RQs above, we provide further analyses of sample-selection strategies, hyperparameter sensitivity, and practical overhead. Due to the page limit, we summarize their scope here and defer the full results to the appendix. App. G compares the sample-selection strategies in Rem. 1. App. D.4 studies the sensitivity of CELEUS to key hyperparameters. App. I compares the per-round runtime of CELEUS with the baselines. App. J reports detailed results for each surrogate-target pair, corresponding to the summary results in Tab. 2 and Fig. 1.

## 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.

## Acknowledgements

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 number DE240101089, LP240100101, DP230101540 and the NSF&CSIRO Responsible AI program with grant number 2303037.

## Impact Statement

This paper aims to improve the reliability and efficiency of LLM evaluation by providing anytime-valid confidence intervals for evaluation scores. CELEUS can reduce the number of expensive evaluated labelled samples needed for certification while maintaining valid uncertainty quantification, which is especially useful when evaluation relies on human annotation or strong LLM judges. By encouraging evaluation with confidence intervals rather than only point estimates, this work may support more transparent and statistically grounded model comparison. However, CELEUS does not guarantee that a benchmark is representative of real-world deployment conditions, and surrogate-assisted evaluation introduces additional computational cost. More broadly, our method addresses statistical uncertainty in evaluation, but does not by itself resolve other risks of LLM deployment, such as safety, bias, privacy, or misuse.

## 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 eval-

uation 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] 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.

[13] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. *GitHub repository*, 2023.

[14] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *ACL*, 2023.

[15] 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.

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

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

[18] Mehryar Mohri and Afshin Rostamizadeh. Stability bounds for stationary $\varphi$-mixing and $\beta$-mixing processes. *Journal of Machine Learning Research*, 11(2), 2010.

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

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

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

[24] 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.

[25] 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.

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

[27] 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.

[28] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. Safetybench: Evaluating the safety of large language models. In *ACL*, 2024.

[29] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[30] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, et al. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *ACM LAMPS*, 2023.

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

[32] 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.

[33] 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.

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

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

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

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

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

[39] 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.

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

## Appendix Contents

**A Additional discussions** **13**

A.1 Related work . . . 13

A.2 Remarks on practical choices and implementation details . . . 13

A.3 From per-round signals to performance estimation . . . 16

A.4 Population Interpretation for Non-i.i.d. Evaluation Pools . . . 18

A.5 Limitations . . . 18

**B Detailed Proofs for Our Theoretical Results** **19**

B.1 Assumptions . . . 19

B.2 Proofs of Theorem 1 . . . 20

B.3 Proofs of Theorem 2 . . . 23

B.4 Proofs of Theorem 3 . . . 26

B.5 Proofs of Theorem 4 . . . 28

B.6 Proofs of Theorem 5 . . . 31

B.7 Proofs of Theorem 6 . . . 43

**C Synthetic Experimental Setting for Table 1** **45**

**D Experimental Setup Details** **47**

D.1 Benchmarks . . . 47

D.2 Models and surrogate-target pairs . . . 47

D.3 Baselines . . . 48

D.4 Hyperparameters . . . 48

D.5 Trials and reporting . . . 48

**E Empirical coverage on real LLM evaluation under 0-1 loss** **49**

**F Certifiable and Label-efficient LLM evaluation on Cross-Entropy Loss** **49**

**G Sample-selection strategy comparison** **49**

**H Hyperparameter sensitivity** **51**

**I Computational cost** **52**

**J Results for each surrogate-target model pair** **53**

# A. Additional discussions

## A.1. Related work

**LLM evaluation.** A large body of work on LLM evaluation focuses on benchmark construction, capability coverage, and evaluation criteria [7, 8, 5]. Existing studies have developed diverse datasets and protocols for assessing model performance across reasoning, knowledge, instruction following, safety, and robustness [1, 2, 27–30]. These works substantially improve our understanding of what aspects of LLM behavior should be measured. However, this literature is primarily concerned with *what* to evaluate, rather than with the statistical reliability of the evaluation procedure itself. In common practice, evaluation often reports an average score on a benchmark under a chosen metric, with limited attention to whether the resulting estimate is sufficiently precise for decision-making.

**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 [31, 32, 19, 15]. 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, 23]. These methods provide a natural framework for pool-based evaluation under limited evaluation budgets [15]. 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) and, more broadly, semi-supervised inference with auxiliary predictors or automatic raters [10, 11, 5]. These methods improve statistical efficiency by combining evaluated samples with auxiliary predictions on unevaluated samples. This paradigm is particularly relevant to LLM evaluation, where automatic raters or surrogate models can provide useful information on unevaluated examples and reduce the number of human evaluations needed [27]. However, existing PPI-style guarantees are typically derived under *i.i.d.* sampling assumptions and rely on asymptotic normality or related large-sample approximations. As a result, they do not directly apply to the adaptive evaluation setting considered here, where evaluated examples are selected sequentially from a finite pool. More importantly, they do not by themselves provide anytime-valid guarantees under continuous monitoring and adaptive stopping.

**Sequential inference and anytime-valid CIs.** Our work is also related to sequential inference, confidence sequences, and $e$-processes [33, 34, 12]. 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 [35–38]. 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 [16], but do not directly address the combination of surrogate-assisted estimation, adaptive sample selection, and certifiable LLM evaluation considered here.

## A.2. 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*

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

*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.*

- ***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. (4). 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. G; see Fig. 4.*

**Remark 2** (Surrogate scoring strategies and sample-selection scores for cross-entropy loss.)**.** *Recall that the oracle sampling distribution in Theorem 4 is $q_t^\star(j) \propto |\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)|$, which is unavailable in practice because it depends on the unknown label $y_j$. To obtain a computable sample-selection score, we replace the true cross-entropy loss with label-free proxies.*

*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. The true cross-entropy loss of the target model is*

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

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

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

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

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

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

- ***Oracle strategy*** *(true cross-entropy loss and surrogate score). If the true 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*

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

  *This oracle strategy is infeasible because it requires the true loss values of unevaluated samples.*

- ***Strategy A*** *(mode loss as target-loss proxy, entropy as surrogate score). Set*

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

  *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). Set*

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

  *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 differ in which quantity is used as the target-loss proxy and which is used as the surrogate score inside the estimator $\hat{R}_t^f$ in Eqn. (4). This changes the correction term and can affect the variance of the estimator and the resulting e-process.*

- ***Strategy C*** *(target-surrogate entropy gap). Set*

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

  *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.*

*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. G; 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.

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

**Remark 3.** *Following [12], we use predictable, time-adaptive fractions to obtain Type-I error control uniformly over time, so 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 $\upsilon$ is also scaled at round $t$ as $\bar{\upsilon}_t = (\upsilon - a_t)/(b_t - a_t)$. Thus, the fractions are chosen based on the centered scaled signal $\bar{S}_i - \bar{\upsilon}_i$, which has conditional mean zero under $\mathcal{H}_0(\upsilon)$. Specifically, we set*

$$\lambda_t^+(\upsilon) = \sqrt{\frac{2\log(2/\alpha_1)}{\hat{\sigma}_{t-1}^2(\upsilon)\, t\log(t+1)}} \wedge \frac{1}{\bar{\upsilon}_t + c}, \qquad \lambda_t^-(\upsilon) = \sqrt{\frac{2\log(2/\alpha_1)}{\hat{\sigma}_{t-1}^2(\upsilon)\, t\log(t+1)}} \wedge \frac{1}{1 - \bar{\upsilon}_t + c},$$

*where $c > 0$ is a small constant, and*

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

*The clipping terms ensure that $\lambda_t^+(\upsilon)$ and $\lambda_t^-(\upsilon)$ stay within their admissible ranges, while $\hat{\sigma}_{t-1}^2(\upsilon)$ adapts the fractions to the past variability of the centered scaled signals.*

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 $b_t$ only 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*

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

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

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

*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 $q_{\min} = \beta/N$.*

### A.3. From per-round signals to performance estimation

In this appendix, we explain how the final performance estimator in Eqn. (4) is obtained by aggregating the correction information exposed by the inferential signal in Eqn. (1). 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

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

Equivalently, it can be written as

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

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

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

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

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

Thus, the estimator in Eqn. (4) 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.

### A.4. Population Interpretation for Non-i.i.d. Evaluation Pools

In this appendix, we clarify the role of the *i.i.d.* assumption in our population-level guarantee. This assumption is not needed for the e-process confidence interval constructed for the finite-pool risk. It is only used in the finite-pool-to-population correction that transfers the finite-pool guarantee to the population risk.

Conditional on the given evaluation pool, CELEUS first constructs anytime-valid confidence intervals for the finite-pool risk

$$R_N = \frac{1}{N}\sum_{j=1}^{N} \ell(f(x_j), y_j).$$

This finite-pool e-process component is the core sequential guarantee of CELEUS. It follows from the construction of conditionally unbiased signals and remains valid even when the evaluation pool is fixed in advance or not sampled *i.i.d.* from a population distribution.

The population-level guarantee targets

$$R(f; P_{x,y}) = \mathrm{E}_{(x,y)\sim P_{x,y}}\left[\ell(f(x), y)\right].$$

To obtain this guarantee, CELEUS adds a correction term that controls the gap between the finite-pool risk $R_N$ and the population risk $R(f; P_{x,y})$. In the main text, this correction is instantiated using Hoeffding's inequality [22], which is appropriate when the evaluation pool consists of independent samples drawn from the target distribution. This yields

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

Thus, the Hoeffding correction should be viewed as the finite-pool-to-population component of CELEUS under the *i.i.d.* pool assumption, rather than as a requirement for the finite-pool e-process component itself.

For non-*i.i.d.* evaluation pools, the finite-pool e-process component remains unchanged, but the correction from $R_N$ to $R(f; P_{x,y})$ should be modified according to the pool-generation assumption. When the dependence structure is known, Hoeffding's inequality can be replaced by a concentration or generalization bound adapted to that structure. For example, Kontorovich and Ramanan [17] provide concentration inequalities for functions of dependent random variables. Their results assume that dependence can be controlled by mixing coefficients, which quantify how strongly future variables depend on the past. Under this condition, they obtain a Hoeffding-like deviation bound with an additional dependence factor determined by these coefficients. Mohri and Rostamizadeh [18] study a related learning-theoretic setting for stationary $\phi$-mixing and $\beta$-mixing processes. Their bounds assume stationarity, bounded losses, and dependence that decays over time according to the corresponding mixing coefficients. For algorithm-dependent generalization bounds, they additionally require stability of the learning algorithm. Under these conditions, they show that generalization bounds can be extended beyond the *i.i.d.* setting, with additional terms depending on the mixing rate. These results illustrate how the finite-pool-to-population correction in CELEUS can be replaced when the evaluation pool has a known weak-dependence structure.

In summary, CELEUS always provides an anytime-valid guarantee for the finite-pool risk $R_N$ conditional on the evaluation pool. Interpreting the resulting interval as a population-level guarantee additionally requires an assumption connecting the evaluation pool to the target population. Under the *i.i.d.* assumption, CELEUS uses Hoeffding's inequality for this correction. Under suitable non-*i.i.d.* assumptions, one can instead use dependent-data concentration or generalization bounds. Without such assumptions, the conservative interpretation is that CELEUS certifies performance on the given evaluation pool, while broader population or deployment claims depend on how representative that pool is.

### A.5. Limitations

CELEUS mainly improves efficiency by reducing the number of samples that require target-model evaluation, which is often the main source of labelled-data or annotation cost in certifiable evaluation. In this paper, efficiency therefore mainly refers to reducing the number of evaluated labelled samples needed to reach a target confidence interval width, i.e., reducing the cost of obtaining target evaluation scores. As shown in Table 2 and Figure 1, using surrogate scores can substantially reduce the required number of target-model evaluated samples compared with EVALUE and the no-surrogate variant of CELEUS. However, this label efficiency does not necessarily mean that CELEUS has lower computational cost in every setting. Surrogate-assisted evaluation introduces additional cost for obtaining surrogate scores, as well as additional per-round

computation for surrogate correction, sample selection, and confidence interval updates. Thus, the per-round runtime of CELEUS can be higher than the standard EVALUE, as illustrated by the per-round computation costs in Table 6.

The practical computational cost of surrogate scoring depends on how the surrogate is obtained and used. Some prior evaluation pipelines use commercial auto-raters [13] or strong LLM judges [5] as surrogate evaluators, whose monetary and computational costs can be substantial. The computational cost may also increase if the surrogate model is updated or fine-tuned during the evaluation process. In contrast, following Berrada et al. [15], we use fixed smaller open-source models as surrogates without task-specific training or online updates. Specifically, our experiments use 7–8B surrogate models to evaluate much larger target models, including 67–72B dense models and an 8×7B MoE model. This design keeps the surrogate computational cost relatively small, avoids extra monetary costs associated with commercial surrogate evaluators, and still provides clear reductions in the number of evaluated labelled samples.

## B. Detailed Proofs for Our Theoretical Results

### B.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 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. In CELEUS, these algorithmic conditions are designed to be satisfied and impose no additional restrictions on the application scenario.

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

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

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

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

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.** *The scaling intervals are non-degenerate and their widths are uniformly bounded above and below: there exist constants $0 < \underline{\Delta} \leq \overline{\Delta} < \infty$ such that*

$$\underline{\Delta} \leq b_t - a_t \leq \overline{\Delta}, \qquad \forall 1 \leq t \leq N.$$

Assumption 3 is an algorithmic condition on the scaling step. It ensures that the affine transformation used to map the inferential signal to the scaled space is well-defined and stable.

**Assumption 4.** *Fix a constant $c > 0$. For each candidate value $\upsilon$, choose*

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

*where $\big(\tilde{\lambda}_t^+(\upsilon)\big)_{1 \leq t \leq N}$ and $\big(\tilde{\lambda}_t^-(\upsilon)\big)_{1 \leq t \leq N}$ are real-valued predictable sequences. The clipping ensures that the resulting fractions remain in the admissible ranges required for the $e$-process construction. We assume that there exist constants $0 < \underline{\gamma} \leq \overline{\gamma} < \infty$ such that, for every candidate value $\upsilon$ and all $1 \leq t \leq N$,*

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

Assumption 4 specifies the predictable fractions used in the $e$-process. This is also an algorithmic condition: the fractions are chosen by the evaluator before observing the risk at each round and can therefore be designed to satisfy the stated bounds.

### B.2. Proofs of Theorem 1

**Theorem 1′.** *Under Assumption 1-2, the estimator $\hat{R}_t$ as defined in Eqn. (4) satisfies $\mathbb{E}[\hat{R}_t] = R$ with*

$$\mathrm{Var}[\hat{R}_t] = \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} c_m a_m \mid D_N, \tilde{\ell}_f\right]\right],$$

*where*

$$c_m = \frac{N(N-t)}{(N-m)(N-m+1)}, \tag{5}$$

*and*

$$a_m = \frac{\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})}{N q_m(i_m \mid i_{1:m-1}, D_N)} + \frac{1}{N}\sum_{s=1}^{m-1}\left(\ell\left(f(\boldsymbol{x}_{i_s}), y_{i_s}\right) - \tilde{\ell}_f\left(\boldsymbol{x}_{i_s}\right)\right). \tag{6}$$

We present the detailed proof of Theorem 1 as follows.

*Proof.* Under Assumption 1-2, we first prove the unbiasedness of $\hat{R}_t$, and then investigate its variance. Recall the estimator in Eqn. (4):

$$\mathbb{E}[\hat{R}_t] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right] + \mathbb{E}\Bigg[\underbrace{\frac{1}{t}\sum_{m=1}^{t} u_{m,t}\left(\ell\left(f(\boldsymbol{x}_{i_m}), y_{i_m}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_m})\right)}_{ⓐ}\Bigg], \tag{7}$$

where

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

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 ⓐ 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. [23], which imply that ⓐ can be rewritten as

$$ⓐ = \frac{1}{t}\sum_{m=1}^{t} c_m a_m, \qquad c_m = \frac{N(N-t)}{(N-m)(N-m+1)},$$

and

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

We now compute $\mathbb{E}[a_m]$ via the tower property. Conditioning on the pool $D_N$, the surrogate predictor $\tilde{\ell}_f$, and the first $m-1$ selections, we obtain

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

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

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

Combining the two terms yields

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

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

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

The last equality holds since

$$\sum_{m=1}^{t} c_m = N(N-t) \sum_{m=1}^{t} \frac{1}{(N-m)(N-m+1)} = N(N-t) \sum_{m=1}^{t} \left( \frac{1}{N-m} - \frac{1}{N-m+1} \right) = t.$$

Substituting $\mathbb{E}[\text{ⓐ}]$ back into Eqn. (7), 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

$$\mathrm{Var}[\hat{R}_t] = \mathbb{E} \Bigg[ \underbrace{\mathrm{Var}\left[ \hat{R}_t \mid D_N, \tilde{\ell}_f \right]}_{\text{ⓒ}} \Bigg] + \mathrm{Var} \Bigg[ \underbrace{\mathbb{E}\left[ \hat{R}_t \mid D_N, \tilde{\ell}_f \right]}_{\text{ⓓ}} \Bigg]. \tag{8}$$

We first investigate the term ⓓ. From the unbiasedness argument above, conditional on $D_N$ and $\tilde{\ell}_f$, we have

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

Therefore,

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

We now consider the term ⓒ. Since the first term in $\hat{R}_t$ is measurable with respect to $(D_N, \tilde{\ell}_f)$, it does not contribute to the conditional variance. Hence,

$$ⓒ = \mathrm{Var}\left[\underbrace{\frac{1}{t}\sum_{m=1}^{t} c_m a_m}_{ⓐ} \mid D_N, \tilde{\ell}_f\right].$$

Write

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

for the average residual over the pool. Then

$$\begin{aligned}
ⓒ &= \mathbb{E}\left[ⓐ^2 \mid D_N, \tilde{\ell}_f\right] - \left(\frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)^2 \\
&= \mathbb{E}\left[\left(\frac{1}{t}\sum_{m=1}^{t} c_m a_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)^2 \mid D_N, \tilde{\ell}_f\right] \\
&= \mathbb{E}\left[\left(\frac{1}{t}\sum_{m=1}^{t} c_m\left(a_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)\right)^2 \mid D_N, \tilde{\ell}_f\right] \\
&= \frac{1}{t^2}\sum_{m=1}^{t}\sum_{s=1}^{t} c_m c_s \mathbb{E}\left[\underbrace{\left(a_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)}_{ⓔ}\right. \\
&\qquad \left.\cdot \underbrace{\left(a_s - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)}_{ⓕ} \mid D_N, \tilde{\ell}_f\right].
\end{aligned}$$

For $s < m$, since $a_s$ is measurable with respect to $(D_N, \tilde{\ell}_f, i_{1:m-1})$, we have

$$\begin{aligned}
\mathbb{E}\left[a_m \cdot ⓕ \mid D_N, \tilde{\ell}_f\right] &= \mathbb{E}\left[\mathbb{E}\left[a_m \cdot ⓕ \mid D_N, \tilde{\ell}_f, i_{1:m-1}\right] \mid D_N, \tilde{\ell}_f\right] \\
&= \mathbb{E}\left[ⓕ \cdot \mathbb{E}\left[a_m \mid D_N, \tilde{\ell}_f, i_{1:m-1}\right] \mid D_N, \tilde{\ell}_f\right].
\end{aligned}$$

Moreover,

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

Therefore,

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

Hence, when $s < m$, the cross term ⓔ $\times$ ⓕ is equal to zero. By symmetry, the same conclusion holds when $m < s$. Therefore, only the diagonal terms remain, and

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

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

$$\begin{aligned}&\mathbb{E}\left[\left(a_m - \frac{1}{N}\sum_{n=1}^{N}\left(\ell\left(f(\boldsymbol{x}_n), y_n\right) - \tilde{\ell}_f\left(\boldsymbol{x}_n\right)\right)\right)^2 \mid D_N, \tilde{\ell}_f\right] \\ =\ &\mathbb{E}\left[\mathrm{Var}\left[a_m \mid D_N, \tilde{\ell}_f, i_{1:m-1}\right] \mid D_N, \tilde{\ell}_f\right].\end{aligned}$$

Therefore,

$$\mathrm{Var}\left[\hat{R}_t \mid D_N, \tilde{\ell}_f\right] = \frac{1}{t^2}\sum_{m=1}^{t} c_m^2 \mathbb{E}\left[\mathrm{Var}\left[a_m \mid D_N, \tilde{\ell}_f, i_{1:m-1}\right] \mid D_N, \tilde{\ell}_f\right].$$

Substituting this and the expression for Var[ⓓ] back into Eqn. (8), we obtain

$$\begin{aligned}\mathrm{Var}[\hat{R}_t] &= \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\frac{1}{t^2}\sum_{m=1}^{t} c_m^2 \mathbb{E}\left[\mathrm{Var}\left[a_m \mid D_N, \tilde{\ell}_f, i_{1:m-1}\right] \mid D_N, \tilde{\ell}_f\right]\right] \\ &= \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} c_m a_m \mid D_N, \tilde{\ell}_f\right]\right],\end{aligned}$$

which completes the proof. □

### B.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 $t < N$,

$$\mathrm{Var}[\hat{R}_t] = \frac{\mathrm{Var}(\ell(f(\boldsymbol{x}), y))}{N} + \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} c_m a_m \mid D_N, \tilde{\ell}_f\right]\right],$$

where

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

and

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

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

$$\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} c_m a_m \mid D_N, \tilde{\ell}_f\right] = \frac{1}{t^2}\sum_{m=1}^{t} c_m^2 \mathbb{E}\left[\mathrm{Var}\left[a_m \mid D_N, \tilde{\ell}_f, i_{1:m-1}\right] \mid D_N, \tilde{\ell}_f\right].$$

We now bound the conditional variance of $a_m$. Since the second term in $a_m$ is measurable with respect to $(D_N, \tilde{\ell}_f, i_{1:m-1})$, it does not contribute to the conditional variance. Hence,

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

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)

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

we obtain

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

Therefore,

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

Substituting this into the above display yields

$$\mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} c_m a_m \mid D_N, \tilde{\ell}_f\right]\right] \leq \frac{(U-L)^2}{\beta}\frac{1}{t^2}\sum_{m=1}^{t} c_m^2 \frac{N-m+1}{N}. \tag{9}$$

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

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

Combining this with Eqn. (9), we obtain

$$\mathbb{E}\left[\mathrm{Var}\left[\frac{1}{t}\sum_{m=1}^{t} c_m a_m \mid D_N, \tilde{\ell}_f\right]\right] \leq \frac{(U-L)^2}{\beta t}. \tag{10}$$

**Step 2: Conclude the $L^2$ convergence.** If $t < N$, combining Theorem 1 with Eqn. (10) gives

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

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 $1 \leq t \leq N$,

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

Finally, if $t = t_N \leq N$ 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. □

### B.4. Proofs of Theorem 3

We present the detailed proof of Theorem 3 as follows.

*Proof.* Recall the signal in Eqn. (1):

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

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. (1) yields

$$\hat{S}_t = R_N + \underbrace{\frac{1}{N}\left(\frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{q_t(i_t \mid i_{1:t-1}, D_N)} - \sum_{j\in\mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\Big)\right)}_{ⓐ}. \tag{11}$$

Under Assumption 1-2, we first prove the conditional unbiasedness. Since $R_N$ is measurable with respect to $\mathcal{F}_{t-1}$, taking conditional expectation on both sides of Eqn. (11) gives

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

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

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

where the last equality uses

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

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. (11),

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

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

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

Expanding $\text{ⓐ}^2$ from Eqn. (11), we obtain

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

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,

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

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

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

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

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

Therefore,

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

which completes the proof. □

### B.5. Proofs of Theorem 4

We present the detailed proof of Theorem 4 as follows.

*Proof.* Fix $t$, and let

$$\mathcal{F}_{t-1} = \sigma(D_N, i_{1:t-1}, \tilde{\ell}_f).$$

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. Since $\tilde{\ell}_f$ is constructed using only the information available up to round $t-1$, these quantities are fixed conditional on $\mathcal{F}_{t-1}$.

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

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

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

$$\hat{S}_t - R_N = \frac{\ell(f(\boldsymbol{x}_{i_t}), y_{i_t}) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{N q_t(i_t \mid i_{1:t-1}, D_N)} - \frac{1}{N} \sum_{j \in \mathcal{J}_{t-1}} \Big(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\Big) . \tag{12}$$

By Theorem 3,

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

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

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

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

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

over all proposal distributions on $\mathcal{J}_{t-1}$.

By Cauchy-Schwarz,

$$\begin{aligned}
&\left(\sum_{j\in\mathcal{J}_{t-1}} \left|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|\right)^2 \\
&= \left(\sum_{j\in\mathcal{J}_{t-1}} \frac{\left|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|}{\sqrt{q_t(j \mid i_{1:t-1}, D_N)}} \sqrt{q_t(j \mid i_{1:t-1}, D_N)}\right)^2 \\
&\leq \left(\sum_{j\in\mathcal{J}_{t-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{q_t(j \mid i_{1:t-1}, D_N)}\right)\left(\sum_{j\in\mathcal{J}_{t-1}} q_t(j \mid i_{1:t-1}, D_N)\right) \\
&= \sum_{j\in\mathcal{J}_{t-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{q_t(j \mid i_{1:t-1}, D_N)}.
\end{aligned}$$

Equality holds if and only if

$$q_t(j \mid i_{1:t-1}, D_N) \propto \left|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|, \qquad j \in \mathcal{J}_{t-1}.$$

Hence the minimizer is

$$q_t^*(j \mid i_{1:t-1}, D_N) = \frac{\left|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right|}{\sum_{s\in\mathcal{J}_{t-1}} \left|\ell(f(\boldsymbol{x}_s), y_s) - \tilde{\ell}_f(\boldsymbol{x}_s)\right|}.$$

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

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

Since only the $m = t$ term depends on $q_t$, we can write

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

By the definition of $u_{m,t}$,

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

Hence

$$\hat{R}_t = ⓐ + \frac{\ell\left(f(\boldsymbol{x}_{i_t}), y_{i_t}\right) - \tilde{\ell}_f(\boldsymbol{x}_{i_t})}{t(N-t+1)q_t(i_t \mid i_{1:t-1}, D_N)}. \tag{13}$$

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

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

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

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

which is exactly the same objective as above. Therefore, the same proposal distribution

$$q_t^*(j \mid i_{1:t-1}, D_N) = \frac{\Big|\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\Big|}{\sum_{s \in \mathcal{J}_{t-1}} \Big|\ell(f(\boldsymbol{x}_s), y_s) - \tilde{\ell}_f(\boldsymbol{x}_s)\Big|}$$

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

Finally, by the law of total variance,

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

From Eqn. (13),

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

which does not depend on the proposal distribution. Therefore, the same proposal distribution $q_t^*$ also minimizes $\mathrm{Var}(\hat{R}_t)$.

**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. (12) immediately gives

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

and hence

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

Therefore,

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

so the variance is already minimized, and the choice of $q_t$ is immaterial and every proposal distribution is optimal.

For the estimator $\hat{R}_t$, since $i_t \in \mathcal{J}_{t-1}$, Eqn. (13) yields

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

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

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

By the law of total variance,

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

which again does not depend on the proposal distribution. Therefore, in this case the variance is also minimized, and the choice of $q_t$ is immaterial and every proposal distribution is optimal.

In summary, when the quantities $\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)$, $j \in \mathcal{J}_{t-1}$, are not all zero, the proposal distribution $q_t^*$ uniquely minimizes the variance. In contrast, when all these quantities are zero, the variance is already minimized, so the choice of $q_t$ is immaterial and every proposal distribution is optimal. This completes the proof.

□

### B.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 $1 \leq t \leq N$,*

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

**Theorem 7.** *[39, 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 elementary inequality as follows

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

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

*Proof.* The proof structure consists of two steps.

- First, we establish the anytime-valid coverage of the inverted hedged capital process under the scaling, showing that $\left(\mathcal{M}_t^{\pm}\right)_{1\leq t\leq N}$ is a $(1-\alpha_1)$ confidence sequence for $R_N$.
- Second, we derive an explicit contraction rate for $\mathcal{M}_t^{\pm}$ under the choice $\lambda_t^+(\upsilon) = |\tilde{\lambda}_t^+| \wedge 1/(\bar{\upsilon}_t + c)$ and $\lambda_t^-(\upsilon) = |\tilde{\lambda}_t^-| \wedge 1/(1 - \bar{\upsilon}_t + c)$ together with the envelope condition $|\tilde{\lambda}_t^{\pm}| \asymp 1/\sqrt{t\log(t+1)}$.

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

Let $\mathcal{F}_{t-1} = \sigma(D_N, i_{1:t-1}, \tilde{\ell}_f)$. 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 and $\hat{S}_t \in [a_t, b_t]$ almost surely. By Assumption 3, the scaling intervals are non-degenerate, so $b_t - a_t > 0$ for all $1 \leq t \leq N$. Hence the above normalization is well defined. Since this is an affine transformation,

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

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

$$\mathbb{E}[\bar{S}_t \mid \mathcal{F}_{t-1}, \mathcal{H}_0(\upsilon)] = \bar{\upsilon}_t.$$

Moreover, by construction $\bar{S}_t \in [0,1]$. We now verify directly that the one-sided capital processes

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

are test martingales under $\mathcal{H}_0(\upsilon)$.

Recall Assumption 4 that we choose

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

where $(\tilde{\lambda}_t^+)_{1\le t\le N}$ and $(\tilde{\lambda}_t^-)_{1\le t\le N}$ are real-valued predictable sequences not depending on $\upsilon$. First, since $\bar{S}_t \in [0,1]$ and $\lambda_t^+(\upsilon) \le 1/(\bar{\upsilon}_t + c)$ by Assumption 4, we have

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

Hence each multiplicative factor in $\mathcal{K}_t^+(\upsilon)$ is nonnegative.

Similarly, since $\bar{S}_t \in [0,1]$ and $\lambda_t^-(\upsilon) \le 1/(1 - \bar{\upsilon}_t + c)$ by Assumption 4, we have

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

Hence each multiplicative factor in $\mathcal{K}_t^-(\upsilon)$ is also nonnegative.

Next, since $\lambda_t^+(\upsilon)$ is $\mathcal{F}_{t-1}$-measurable by Assumption 4, under $\mathcal{H}_0(\upsilon)$ we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{K}_t^+(\upsilon) \mid \mathcal{F}_{t-1}\right] &= \mathcal{K}_{t-1}^+(\upsilon)\,\mathbb{E}\left[1 + \lambda_t^+(\upsilon)\left(\bar{S}_t - \bar{\upsilon}_t\right) \mid \mathcal{F}_{t-1}, \mathcal{H}_0(\upsilon)\right] \\
&= \mathcal{K}_{t-1}^+(\upsilon)\left(1 + \lambda_t^+(\upsilon)\mathbb{E}\left[\bar{S}_t - \bar{\upsilon}_t \mid \mathcal{F}_{t-1}, \mathcal{H}_0(\upsilon)\right]\right) \\
&= \mathcal{K}_{t-1}^+(\upsilon).
\end{aligned}$$

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

In the same way, since $\lambda_t^-(\upsilon)$ is also $\mathcal{F}_{t-1}$-measurable by Assumption 4,

$$\begin{aligned}
\mathbb{E}\left[\mathcal{K}_t^-(\upsilon) \mid \mathcal{F}_{t-1}\right] &= \mathcal{K}_{t-1}^-(\upsilon)\,\mathbb{E}\left[1 - \lambda_t^-(\upsilon)\left(\bar{S}_t - \bar{\upsilon}_t\right) \mid \mathcal{F}_{t-1}, \mathcal{H}_0(\upsilon)\right] \\
&= \mathcal{K}_{t-1}^-(\upsilon)\left(1 - \lambda_t^-(\upsilon)\mathbb{E}\left[\bar{S}_t - \bar{\upsilon}_t \mid \mathcal{F}_{t-1}, \mathcal{H}_0(\upsilon)\right]\right) \\
&= \mathcal{K}_{t-1}^-(\upsilon).
\end{aligned}$$

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

Hence their hedged combination

$$\mathcal{K}_t^{\pm}(\upsilon) = \theta\mathcal{K}_t^+(\upsilon) + (1-\theta)\mathcal{K}_t^-(\upsilon), \qquad \theta \in [0,1],$$

is again a nonnegative martingale under $\mathcal{H}_0(\upsilon)$, and therefore a test martingale.

By Ville's inequality [40],

$$\Pr_{R_N=\upsilon}\left(\exists 1 \le t \le N:\ \mathcal{K}_t^{\pm}(\upsilon) \ge \frac{1}{\alpha_1}\right) \le \alpha_1.$$

Therefore, the inverted set with $\upsilon \in [L, U]$ (Assumption 1)

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

is a $(1-\alpha_1)$ confidence sequence for $R_N$.

We further define the running intersection

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

Then $\mathcal{M}_t^{\pm} \subseteq \mathcal{B}_t^{\pm}$ for each $t$, so $\{\mathcal{M}_t^{\pm}\}_{1\le t\le N}$ is nested and has non-increasing width. It also remains $(1-\alpha_1)$ anytime-valid because the corresponding ever-miss events coincide:

$$\upsilon \notin \mathcal{M}_t^{\pm} \iff \exists i \le t \text{ such that } \upsilon \notin \mathcal{B}_i^{\pm},$$

and therefore

$$\left\{\exists 1 \le t \le N : \upsilon \notin \mathcal{M}_t^{\pm}\right\} = \bigcup_{1\le t\le N} \bigcup_{i\le t} \left\{\upsilon \notin \mathcal{B}_i^{\pm}\right\} = \bigcup_{i\ge 1} \left\{\upsilon \notin \mathcal{B}_i^{\pm}\right\} = \left\{\exists 1 \le t \le N : \upsilon \notin \mathcal{B}_t^{\pm}\right\}.$$

Hence

$$\Pr\left(\exists 1 \le t \le N : \upsilon \notin \mathcal{M}_t^{\pm}\right) = \Pr\left(\exists 1 \le t \le N : \upsilon \notin \mathcal{B}_t^{\pm}\right) \le \alpha_1,$$

which shows that $\{\mathcal{M}_t^{\pm}\}_{1\le t\le N}$ is also 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 $c \in (0,1)$. Recall that, under Assumption 4, we choose $\lambda_t^+(\upsilon) = |\tilde{\lambda}_t^+| \wedge 1/(\bar{\upsilon}_t + c)$ and $\lambda_t^-(\upsilon) = |\tilde{\lambda}_t^-| \wedge 1/(1-\bar{\upsilon}_t + c)$, where $(\tilde{\lambda}_t^+)_{1\le t\le N}$ and $(\tilde{\lambda}_t^-)_{1\le t\le N}$ are real-valued predictable sequences not depending on $\upsilon$, and there exist constants $0 < \underline{\gamma} \le \overline{\gamma} < \infty$ such that for all $1 \le t \le N$,

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

**We start with $\upsilon < R_N$ and analyze the one-sided capital process $\mathcal{K}_t^+(\upsilon)$.** For each $i \ge 1$ with $\bar{S}_i \in [0,1]$ and $\bar{\upsilon}_i \in [0,1]$, we have

$$\lambda_i^+(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right) \ge \frac{1}{\bar{\upsilon}_i + c}(-\bar{\upsilon}_i) \ge -\frac{1}{1+c}.$$

Applying the elementary inequality (Theorem 8) termwise with $u = \lambda_i^+(\upsilon)(\bar{S}_i - \bar{\upsilon}_i)$ gives

$$\begin{aligned}
\log \mathcal{K}_t^+(\upsilon) &= \sum_{i=1}^{t} \log\left(1 + \lambda_i^+(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right)\right) \\
&\ge \sum_{i=1}^{t} \lambda_i^+(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right) - \frac{1}{2(1-1/(1+c))} \sum_{i=1}^{t} \left(\lambda_i^+(\upsilon)\right)^2 \left(\bar{S}_i - \bar{\upsilon}_i\right)^2.
\end{aligned}$$

Using $\bar{S}_i - \bar{\upsilon}_i = (\bar{S}_i - \bar{R}_{N,i}) + (\bar{R}_{N,i} - \bar{\upsilon}_i)$ and $(\bar{S}_i - \bar{\upsilon}_i)^2 \le 1$, we obtain

$$\begin{aligned}
\log \mathcal{K}_t^+(\upsilon) &\ge \sum_{i=1}^{t} \lambda_i^+(\upsilon)\left(\bar{R}_{N,i} - \bar{\upsilon}_i\right) + \sum_{i=1}^{t} \lambda_i^+(\upsilon)(\bar{S}_i - \bar{R}_{N,i}) - \frac{\sum_{i=1}^{t}\left(\lambda_i^+(\upsilon)\right)^2\left(\bar{S}_i - \bar{\upsilon}_i\right)^2}{2(1-1/(1+c))} \\
&\ge (R_N - \upsilon)\sum_{i=1}^{t} \frac{\lambda_i^+(\upsilon)}{b_i - a_i} + \underbrace{\sum_{i=1}^{t} \lambda_i^+(\upsilon)(\bar{S}_i - \bar{R}_{N,i})}_{ⓒ} - \frac{\sum_{i=1}^{t}\left(\lambda_i^+(\upsilon)\right)^2}{2(1-1/(1+c))}.
\end{aligned} \tag{14}$$

Define the martingale-difference increments

$$Y_t = \lambda_t^+(\upsilon)\left(\bar{S}_t - \bar{R}_{N,t}\right), \qquad W_t = \sum_{i=1}^{t} Y_i = \sum_{i=1}^{t} \lambda_i^+(\upsilon)\left(\bar{S}_i - \bar{R}_{N,i}\right) = ⓒ.$$

Since $\mathbb{E}[\bar{S}_t \mid \mathcal{F}_{t-1}] = \bar{R}_{N,t}$ and $\lambda_t^+(\upsilon)$ is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = \lambda_t^+(\upsilon)\mathbb{E}[\bar{S}_t - \bar{R}_{N,t} \mid \mathcal{F}_{t-1}] = 0,$$

so $(Y_t)_{1\le t\le N}$ is a martingale difference sequence. Moreover,

$$|Y_t| \le \lambda_t^+(\upsilon)|\bar{S}_t - \bar{R}_{N,t}| \le \frac{1}{c} \quad \text{almost surely.}$$

Finally, its predictable quadratic variation is

$$\begin{aligned}
V_t^+(\upsilon) &= \sum_{i=1}^{t} \mathbb{E}[Y_i^2 \mid \mathcal{F}_{i-1}] \\
&= \sum_{i=1}^{t} \mathbb{E}\left[\left(\lambda_i^+(\upsilon)\right)^2 \left(\bar{S}_i - \bar{R}_{N,i}\right)^2 \mid \mathcal{F}_{i-1}\right] \\
&= \sum_{i=1}^{t} \left(\lambda_i^+(\upsilon)\right)^2 \mathbb{E}\left[\left(\bar{S}_i - \bar{R}_{N,i}\right)^2 \mid \mathcal{F}_{i-1}\right] \\
&= \sum_{i=1}^{t} \left(\lambda_i^+(\upsilon)\right)^2 \mathrm{Var}\left(\bar{S}_i \mid \mathcal{F}_{i-1}\right) \\
&= \sum_{i=1}^{t} \frac{\left(\lambda_i^+(\upsilon)\right)^2}{(b_i - a_i)^2} \mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) \\
&\overset{(a)}{\leq} \sum_{i=1}^{t} \frac{\left(\lambda_i^+(\upsilon)\right)^2}{\underline{\Delta}^2} \mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) \\
&\overset{(b)}{\leq} \sum_{i=1}^{t} \frac{\overline{\gamma}^2}{i \log(i+1) \underline{\Delta}^2} \mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) \\
&= \frac{\overline{\gamma}^2}{\underline{\Delta}^2} \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \,, \qquad (15)
\end{aligned}$$

where $(a)$ follows from Assumption 3, and $(b)$ follows from Assumption 4.

We now upgrade the finite-horizon bound in Theorem 7 to an anytime-valid bound via stitching. Fix a target error probability $\omega \in (0,1)$ and any $\rho > 1$. For $j = 0, 1, 2, \ldots$, define the dyadic horizons

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

Let $-W_t = \sum_{i=1}^{t}(-Y_i)$. Then $(-Y_t)_{1 \leq t \leq N}$ is also a bounded martingale difference sequence with the same predictable quadratic variation $V_t^+(\upsilon)$ and bound $|-Y_t| \leq 1/c$. For each $j$, consider the event

$$\begin{aligned}
\mathcal{E}_j &= \left\{\forall t \leq T_j : W_t \geq -\sqrt{2\rho V_t^+(\upsilon)\delta_j} - \frac{4\delta_j}{3c}\right\} \\
&= \left\{\forall t \leq T_j : -W_t \leq \sqrt{2\rho V_t^+(\upsilon)\delta_j} + \frac{4\delta_j}{3c}\right\} .
\end{aligned}$$

Theorem 7 applied to $-W_t$ yields, for each $j$,

$$\begin{aligned}
\Pr(\mathcal{E}_j^c) &= \Pr\left(\left\{\exists t \leq T_j : -W_t > \sqrt{2\rho V_t^+(\upsilon)\delta_j} + \frac{4\delta_j}{3c}\right\}\right) \\
&\leq \left(1 + \left\lceil \log_\rho\left(\frac{2T_j}{\delta_j}\right)\right\rceil_+\right) \exp(-\delta_j) \,,
\end{aligned}$$

where $\mathcal{E}_j^c$ denotes the complement event of $\mathcal{E}_j$.

Since $\delta_j \geq 1$, we have $2T_j/\delta_j \leq 2T_j$, hence $\lceil \log_\rho(2T_j/\delta_j) \rceil_+ \leq \lceil \log_\rho(2T_j) \rceil_+$, and therefore

$$\Pr(\mathcal{E}_j^c) \leq \left(1 + \left\lceil \log_\rho(2T_j) \right\rceil_+\right) \exp(-\delta_j) \leq \omega_j,$$

where the last inequality holds by the definition of $\delta_j$. By a union bound,

$$\Pr\Big(\bigcap_{j\geq 0}\mathcal{E}_j\Big) \geq 1-\sum_{j\geq 0}\Pr(\mathcal{E}_j^c) \geq 1-\sum_{j\geq 0}\omega_j = 1-\omega/2.$$

On the event $\bigcap_{j\geq 0}\mathcal{E}_j$, all bounds $\mathcal{E}_j$ hold simultaneously. Fix any $1\leq t\leq N$ and let $j(t)=\lceil \log_2 t\rceil$. Then $t\leq 2^{j(t)}=T_{j(t)}$. Since $\mathcal{E}_{j(t)}$ ensures that the stated inequality holds for all times up to $T_{j(t)}$, it in particular holds at time $t$, that is,

$$W_t \geq -\sqrt{2\rho V_t^+(\upsilon)\delta_{\lceil \log_2 t\rceil}} - \frac{4\delta_{\lceil \log_2 t\rceil}}{3c}. \tag{16}$$

Combining Eqn. (16) with Eqn. (14), we obtain that, with probability at least $1-\omega/2$, simultaneously for all $1\leq t\leq N$,

$$\begin{aligned}\log \mathcal{K}_t^+(\upsilon) \quad \geq \quad & (R_N-\upsilon)\sum_{i=1}^t \frac{\lambda_i^+(\upsilon)}{b_i-a_i} - \frac{1}{2(1-1/(1+c))}\sum_{i=1}^t \big(\lambda_i^+(\upsilon)\big)^2 \\ & -\sqrt{2\rho V_t^+(\upsilon)\delta_{\lceil \log_2 t\rceil}} - \frac{4}{3c}\delta_{\lceil \log_2 t\rceil}.\end{aligned}$$

Substituting Eqn. (15) into the display above further gives

$$\begin{aligned}\log \mathcal{K}_t^+(\upsilon) \quad \geq \quad & (R_N-\upsilon)\underbrace{\sum_{i=1}^t \frac{\lambda_i^+(\upsilon)}{b_i-a_i}}_{\text{ⓓ}} - \frac{1}{2(1-1/(1+c))}\underbrace{\sum_{i=1}^t \big(\lambda_i^+(\upsilon)\big)^2}_{\text{ⓔ}} \\ & -\sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2}\sum_{i=1}^t \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\delta_{\lceil \log_2 t\rceil}} - \frac{4}{3c}\delta_{\lceil \log_2 t\rceil}.\end{aligned} \tag{17}$$

We now bound the two deterministic aggregates

$$\text{ⓓ} = \sum_{i=1}^t \frac{\lambda_i^+(\upsilon)}{b_i-a_i}, \qquad \text{ⓔ} = \sum_{i=1}^t \big(\lambda_i^+(\upsilon)\big)^2,$$

under the standing assumption that there exist constants $0<\underline{\gamma}\leq\overline{\gamma}<\infty$ such that for all $1\leq t\leq N$,

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

Since $\lambda_i^+(\upsilon)\leq|\tilde{\lambda}_i^+|$,

$$\text{ⓔ} \leq \sum_{i=1}^t |\tilde{\lambda}_i^+|^2 \leq \overline{\gamma}^2\sum_{i=1}^t \frac{1}{i\log(i+1)}.$$

Since $\log(i+1)\geq \log i$ for $i\geq 2$, we have

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

Therefore, there exists an absolute constant $C_0$ such that for all $t\geq 2$,

$$\text{ⓔ} \leq \overline{\gamma}^2\big(C_0+\log\log(t+1)\big). \tag{18}$$

Since every $\bar{\upsilon}_i \in [0,1]$, we have $\bar{\upsilon}_i + c \le 1 + c$ and thus $1/(\bar{\upsilon}_i + c) \ge 1/(1+c)$. Choose a deterministic integer $t_c \ge 1$ such that

$$\frac{\overline{\gamma}}{\sqrt{t_c \log(t_c+1)}} \le \frac{1}{1+c}.$$

Such a finite $t_c$ exists because

$$\frac{\overline{\gamma}}{\sqrt{t \log(t+1)}} \to 0 \qquad \text{as } t \to \infty.$$

For any $t \ge 2t_c$ and any $i \in \{\lfloor t/2 \rfloor, \dots, t\}$, we have $i \ge t_c$ and hence

$$|\tilde{\lambda}_i^+| \le \frac{1}{1+c} \le \frac{1}{\bar{\upsilon}_i + c},$$

so $\lambda_i^+(\upsilon) = |\tilde{\lambda}_i^+|$. Moreover, by Assumption 3,

$$b_i - a_i \le \overline{\Delta},$$

and therefore for all such $i$,

$$\frac{\lambda_i^+(\upsilon)}{b_i - a_i} = \frac{|\tilde{\lambda}_i^+|}{b_i - a_i} \ge \frac{1}{\overline{\Delta}} \cdot \frac{\underline{\gamma}}{\sqrt{i \log(i+1)}} \ge \frac{1}{\overline{\Delta}} \cdot \frac{\underline{\gamma}}{\sqrt{t \log(t+1)}}.$$

Therefore, for all $t \ge 2t_c$,

$$\text{ⓓ} \ge \sum_{i=\lfloor t/2 \rfloor}^{t} \frac{\lambda_i^+(\upsilon)}{b_i - a_i} \ge \frac{t}{2} \cdot \frac{\underline{\gamma}}{\overline{\Delta}\sqrt{t \log(t+1)}} = \frac{\underline{\gamma}\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}}. \tag{19}$$

By substituting Eqn. (18) and Eqn. (19) into Eqn. (17), we obtain that, with probability at least $1 - \omega/2$, simultaneously for all $t \ge 2t_c$,

$$\begin{aligned} \log \mathcal{K}_t^+(\upsilon) \quad \ge \quad & (R_N - \upsilon)\frac{\underline{\gamma}\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}} - \frac{\overline{\gamma}^2\big(C_0 + \log\log(t+1)\big)}{2(1 - 1/(1+c))} \\ & - \sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2} \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \delta_{\lceil \log_2 t \rceil}} - \frac{4}{3c}\delta_{\lceil \log_2 t \rceil}. \end{aligned} \tag{20}$$

Recall that $\mathcal{B}_t^{\pm}$ is defined by

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

Therefore, to show that a candidate value $\upsilon$ does not belong to $\mathcal{B}_t^{\pm}$, it suffices to verify the converse inequality

$$\mathcal{K}_t^{\pm}(\upsilon) \ge \frac{1}{\alpha_1},$$

since this directly implies $\upsilon \notin \mathcal{B}_t^{\pm}$.

On the event where Eqn. (20) holds, any $\upsilon < R_N$ satisfying

$$R_N - \upsilon \ge \underbrace{\frac{\log\left(\frac{1}{\theta\alpha_1}\right) + \frac{\overline{\gamma}^2\left(C_0 + \log\log(t+1)\right)}{2(1-1/(1+c))} + \sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2} \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)} \delta_{\lceil \log_2 t \rceil}} + \frac{4}{3c}\delta_{\lceil \log_2 t \rceil}}{\underline{\gamma}\sqrt{t}/(2\overline{\Delta}\sqrt{\log(t+1)})}}_{\text{ⓕ}} \tag{21}$$

ensures that $\log \mathcal{K}_t^+(\upsilon) \geq \log(1/(\theta\alpha_1))$, and hence $\mathcal{K}_t^+(\upsilon) \geq 1/(\theta\alpha_1)$. Using $\mathcal{K}_t^\pm(\upsilon) \geq \theta\,\mathcal{K}_t^+(\upsilon)$, we further obtain $\mathcal{K}_t^\pm(\upsilon) \geq 1/\alpha_1$. Consequently, any such $\upsilon$ is excluded from $\mathcal{B}_t^\pm$.

To turn the exclusion condition Eqn. (21) into an explicit radius, we upper bound ⓕ. There exists a constant $G_1 > 0$, depending only on $\underline{\Delta}, \overline{\Delta}, c, \rho, \underline{\gamma}, \overline{\gamma}, \alpha_1, \theta, \omega$, such that for all $t \geq 2t_c$,

$$
\begin{aligned}
\text{ⓕ} \;\leq\; & G_1\sqrt{\frac{\log(t+1)}{t}}\times \\
& \left[1+\log\log(t+1)+\sqrt{\sum_{i=1}^{t}\frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\delta_{\lceil\log_2 t\rceil}}+\delta_{\lceil\log_2 t\rceil}\right].
\end{aligned}
$$

Moreover, by Theorem 3 together with Assumptions 1 and 2, for each $i \geq 1$ we have

$$
\begin{aligned}
\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) &= \frac{1}{N^2}\sum_{j\in\mathcal{J}_{i-1}}\frac{\left(\ell\left(f(\boldsymbol{x}_j),y_j\right)-\tilde{\ell}_f\left(\boldsymbol{x}_j\right)\right)^2}{q_i(j \mid i_{1:i-1}, D_N)} \\
&\quad -\left(\frac{1}{N}\sum_{j\in\mathcal{J}_{i-1}}\left(\ell\left(f(\boldsymbol{x}_j),y_j\right)-\tilde{\ell}_f\left(\boldsymbol{x}_j\right)\right)\right)^2 \\
&\leq \frac{1}{N^2}\sum_{j\in\mathcal{J}_{i-1}}\frac{(U-L)^2}{q_i(j \mid i_{1:i-1}, D_N)} \\
&\leq \frac{1}{N^2}\sum_{j\in\mathcal{J}_{i-1}}(U-L)^2\frac{N-i+1}{\beta} \\
&= \frac{(U-L)^2(N-i+1)^2}{\beta N^2} \\
&\leq \frac{(U-L)^2}{\beta}.
\end{aligned}
$$

Hence $\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)$ is uniformly bounded, and therefore the same comparison used in Eqn. (18) yields

$$
\begin{aligned}
\sum_{i=1}^{t}\frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)} &\leq \frac{(U-L)^2}{\beta}\sum_{i=1}^{t}\frac{1}{i\log(i+1)} \\
&\leq \frac{(U-L)^2}{\beta}\left(C_0+\log\log(t+1)\right) \\
&= O(\log\log t)\,. \qquad (22)
\end{aligned}
$$

Finally, the stitching construction yields $\delta_{\lceil\log_2 t\rceil} = O(\log(1/\omega)+\log\log t)$. Substituting these scalings into the previous display shows that the dominant contribution is of order $\log\log t$, and therefore

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

for a constant $G_2$ depending only on $\underline{\Delta}, \overline{\Delta}, c, \rho, \underline{\gamma}, \overline{\gamma}, \alpha_1, \theta, \omega$.

Putting the pieces together, on the event where Eqn. (20) holds, every $\upsilon < R_N$ with

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

is excluded from $\mathcal{B}_t^{\pm}$. Equivalently, $\mathcal{B}_t^{\pm}$ cannot contain any point smaller than

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

and hence the left endpoint satisfies, with probability at least $1-\omega/2$, simultaneously for all $t \geq 2t_c$,

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

**We now turn to the upper endpoint and analyze the one-sided capital process $\mathcal{K}_t^-(\upsilon)$ for $\upsilon > R_N$.** For each $1 \leq t \leq N$ with $\bar{S}_t \in [0,1]$ and $\bar{\upsilon}_t \in [0,1]$, we have

$$-\lambda_i^-(\upsilon)\left(\bar{S}_i - \bar{\upsilon}_i\right) \geq -\frac{1}{1-\bar{\upsilon}_i + c}(1-\bar{\upsilon}_i) \geq -\frac{1}{1+c}.$$

Applying the elementary inequality (Theorem 8) termwise with $u = -\lambda_i^-(\upsilon)(\bar{S}_i - \bar{\upsilon}_i)$ gives

$$\begin{aligned}\log \mathcal{K}_t^-(\upsilon) &= \sum_{i=1}^{t}\log\left(1-\lambda_i^-(\upsilon)\left(\bar{S}_i-\bar{\upsilon}_i\right)\right)\\ &\geq -\sum_{i=1}^{t}\lambda_i^-(\upsilon)\left(\bar{S}_i-\bar{\upsilon}_i\right) - \frac{1}{2(1-1/(1+c))}\sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2\left(\bar{S}_i-\bar{\upsilon}_i\right)^2.\end{aligned}$$

Using $\bar{S}_i - \bar{\upsilon}_i = (\bar{S}_i - \bar{R}_{N,i}) + (\bar{R}_{N,i} - \bar{\upsilon}_i)$ and $(\bar{S}_i - \bar{\upsilon}_i)^2 \leq 1$, we obtain

$$\begin{aligned}\log \mathcal{K}_t^-(\upsilon) &\geq \sum_{i=1}^{t}\lambda_i^-(\upsilon)\left(\bar{\upsilon}_i - \bar{R}_{N,i}\right) - \sum_{i=1}^{t}\lambda_i^-(\upsilon)(\bar{S}_i - \bar{R}_{N,i}) - \frac{\sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2\left(\bar{S}_i-\bar{\upsilon}_i\right)^2}{2(1-1/(1+c))}\\ &\geq (\upsilon - R_N)\sum_{i=1}^{t}\frac{\lambda_i^-(\upsilon)}{b_i-a_i} - \underbrace{\sum_{i=1}^{t}\lambda_i^-(\upsilon)(\bar{S}_i - \bar{R}_{N,i})}_{ⓒ^-} - \frac{\sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2}{2(1-1/(1+c))}.\end{aligned} \tag{23}$$

Define the martingale-difference increments

$$Y_t^- = \lambda_t^-(\upsilon)\left(\bar{S}_t - \bar{R}_{N,t}\right), \qquad W_t^- = \sum_{i=1}^{t}Y_i^- = \sum_{i=1}^{t}\lambda_i^-(\upsilon)\left(\bar{S}_i - \bar{R}_{N,i}\right) = ⓒ^-.$$

Since $\mathbb{E}[\bar{S}_t \mid \mathcal{F}_{t-1}] = \bar{R}_{N,t}$ and $\lambda_t^-(\upsilon)$ is $\mathcal{F}_{t-1}$-measurable,

$$\mathbb{E}[Y_t^- \mid \mathcal{F}_{t-1}] = \lambda_t^-(\upsilon)\mathbb{E}[\bar{S}_t - \bar{R}_{N,t} \mid \mathcal{F}_{t-1}] = 0,$$

so $(Y_t^-)_{1\leq t\leq N}$ is a martingale difference sequence. Moreover,

$$|Y_t^-| \leq \lambda_t^-(\upsilon)|\bar{S}_t - \bar{R}_{N,t}| \leq \frac{1}{c} \quad \text{almost surely.}$$

Finally, its predictable quadratic variation is

$$\begin{aligned}
V_t^-(\upsilon) &= \sum_{i=1}^{t} \mathbb{E}[(Y_i^-)^2 \mid \mathcal{F}_{i-1}] \\
&= \sum_{i=1}^{t} \mathbb{E}\left[\left(\lambda_i^-(\upsilon)\right)^2 \left(\bar{S}_i - \bar{R}_{N,i}\right)^2 \mid \mathcal{F}_{i-1}\right] \\
&= \sum_{i=1}^{t} \left(\lambda_i^-(\upsilon)\right)^2 \mathbb{E}\left[\left(\bar{S}_i - \bar{R}_{N,i}\right)^2 \mid \mathcal{F}_{i-1}\right] \\
&= \sum_{i=1}^{t} \left(\lambda_i^-(\upsilon)\right)^2 \mathrm{Var}\left(\bar{S}_i \mid \mathcal{F}_{i-1}\right) \\
&= \sum_{i=1}^{t} \frac{\left(\lambda_i^-(\upsilon)\right)^2}{(b_i - a_i)^2} \mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) \\
&\overset{(a)}{\leq} \sum_{i=1}^{t} \frac{\left(\lambda_i^-(\upsilon)\right)^2}{\underline{\Delta}^2} \mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) \\
&\overset{(b)}{\leq} \sum_{i=1}^{t} \frac{\overline{\gamma}^2}{i \log(i+1) \underline{\Delta}^2} \mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) \\
&= \frac{\overline{\gamma}^2}{\underline{\Delta}^2} \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)}, \qquad (24)
\end{aligned}$$

where $(a)$ follows from Assumption 3, and $(b)$ follows from Assumption 4.

We now upgrade the finite-horizon bound in Theorem 7 to an anytime-valid bound via stitching. Fix a target error probability $\omega \in (0,1)$ and any $\rho > 1$. For $j = 0, 1, 2, \ldots$, define the dyadic horizons

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

For each $j$, consider the event

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

Theorem 7 applied to $W_t^-$ yields, for each $j$,

$$\begin{aligned}
\Pr(\mathcal{E}_j^c) &= \Pr\left(\left\{\exists t \leq T_j : W_t^- > \sqrt{2\rho V_t^-(\upsilon)\delta_j} + \frac{4\delta_j}{3c}\right\}\right) \\
&\leq \left(1 + \left\lceil \log_\rho\left(\frac{2T_j}{\delta_j}\right)\right\rceil_+\right) \exp(-\delta_j).
\end{aligned}$$

Since $\delta_j \geq 1$, we have $2T_j/\delta_j \leq 2T_j$, hence $\lceil \log_\rho(2T_j/\delta_j) \rceil_+ \leq \lceil \log_\rho(2T_j) \rceil_+$, and therefore

$$\Pr(\mathcal{E}_j^c) \leq \left(1 + \left\lceil \log_\rho(2T_j) \right\rceil_+\right) \exp(-\delta_j) \leq \omega_j,$$

where the last inequality holds by the definition of $\delta_j$. By a union bound,

$$\Pr\Big(\bigcap_{j \geq 0} \mathcal{E}_j\Big) \geq 1 - \sum_{j \geq 0} \Pr(\mathcal{E}_j^c) \geq 1 - \sum_{j \geq 0} \omega_j = 1 - \omega/2.$$

On the event $\bigcap_{j\geq 0}\mathcal{E}_j$, all bounds $\mathcal{E}_j$ hold simultaneously. Fix any $1 \leq t \leq N$ and let $j(t) = \lceil \log_2 t \rceil$. Then $t \leq 2^{j(t)} = T_{j(t)}$. Since $\mathcal{E}_{j(t)}$ ensures that the stated inequality holds for all times up to $T_{j(t)}$, it in particular holds at time $t$, that is,

$$W_t^- \leq \sqrt{2\rho V_t^-(\upsilon)\delta_{\lceil \log_2 t\rceil}} + \frac{4\delta_{\lceil \log_2 t\rceil}}{3c}. \tag{25}$$

Combining Eqn. (25) with Eqn. (23) yields that, with probability at least $1-\omega/2$, simultaneously for all $1 \leq t \leq N$,

$$\begin{aligned}\log \mathcal{K}_t^-(\upsilon) \quad \geq \quad & (\upsilon - R_N)\sum_{i=1}^{t}\frac{\lambda_i^-(\upsilon)}{b_i - a_i} - \frac{1}{2(1-1/(1+c))}\sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2 \\ & -\sqrt{2\rho V_t^-(\upsilon)\delta_{\lceil \log_2 t\rceil}} - \frac{4}{3c}\delta_{\lceil \log_2 t\rceil}.\end{aligned}$$

Substituting the bound in Eqn. (24) into the display above further gives

$$\begin{aligned}\log \mathcal{K}_t^-(\upsilon) \quad \geq \quad & (\upsilon - R_N)\underbrace{\sum_{i=1}^{t}\frac{\lambda_i^-(\upsilon)}{b_i - a_i}}_{\text{ⓓ}^-} - \frac{1}{2(1-1/(1+c))}\underbrace{\sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2}_{\text{ⓔ}^-} \\ & -\sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2}\sum_{i=1}^{t}\frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\delta_{\lceil \log_2 t\rceil}} - \frac{4}{3c}\delta_{\lceil \log_2 t\rceil}.\end{aligned} \tag{26}$$

We now bound the two deterministic aggregates

$$\text{ⓓ}^- = \sum_{i=1}^{t}\frac{\lambda_i^-(\upsilon)}{b_i - a_i}, \qquad \text{ⓔ}^- = \sum_{i=1}^{t}\left(\lambda_i^-(\upsilon)\right)^2,$$

under the standing assumption that there exist constants $0 < \underline{\gamma} \leq \overline{\gamma} < \infty$ such that for all $1 \leq t \leq N$,

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

Since $\lambda_i^-(\upsilon) \leq |\tilde{\lambda}_i^-|$,

$$\text{ⓔ}^- \leq \sum_{i=1}^{t}|\tilde{\lambda}_i^-|^2 \leq \overline{\gamma}^2\sum_{i=1}^{t}\frac{1}{i\log(i+1)}.$$

Since $\log(i+1) \geq \log i$ for $i \geq 2$, we have

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

Therefore, there exists an absolute constant $C_0$ such that for all $t \geq 2$,

$$\text{ⓔ}^- \leq \overline{\gamma}^2\left(C_0 + \log\log(t+1)\right). \tag{27}$$

Since every $\bar{\upsilon}_i \in [0,1]$, we have $1 - \bar{\upsilon}_i + c \leq 1 + c$ and thus $1/(1-\bar{\upsilon}_i + c) \geq 1/(1+c)$. Choose a deterministic integer $t_c \geq 1$ such that

$$\frac{\overline{\gamma}}{\sqrt{t_c\log(t_c+1)}} \leq \frac{1}{1+c}.$$

Such a finite $t_c$ exists because

$$\frac{\overline{\gamma}}{\sqrt{t\log(t+1)}} \to 0 \qquad \text{as } t \to \infty.$$

For any $t \geq 2t_c$ and any $i \in \{\lfloor t/2 \rfloor, \ldots, t\}$, we have $i \geq t_c$ and hence

$$|\tilde{\lambda}_i^-| \leq \frac{1}{1+c} \leq \frac{1}{1-\bar{v}_i+c},$$

so $\lambda_i^-(v) = |\tilde{\lambda}_i^-|$. Moreover, by Assumption 3,

$$b_i - a_i \leq \overline{\Delta},$$

and therefore for all such $i$,

$$\frac{\lambda_i^-(v)}{b_i - a_i} = \frac{|\tilde{\lambda}_i^-|}{b_i - a_i} \geq \frac{1}{\overline{\Delta}} \cdot \frac{\underline{\gamma}}{\sqrt{i \log(i+1)}} \geq \frac{1}{\overline{\Delta}} \cdot \frac{\underline{\gamma}}{\sqrt{t \log(t+1)}}.$$

Therefore, for all $t \geq 2t_c$,

$$ⓓ^- \geq \sum_{i=\lfloor t/2 \rfloor}^{t} \frac{\lambda_i^-(v)}{b_i - a_i} \geq \frac{t}{2} \cdot \frac{\underline{\gamma}}{\overline{\Delta}\sqrt{t \log(t+1)}} = \frac{\underline{\gamma}\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}}. \tag{28}$$

By substituting Eqn. (27) and Eqn. (28) into Eqn. (26), we have that, with probability at least $1 - \omega/2$, simultaneously for all $t \geq 2t_c$,

$$\begin{aligned} \log \mathcal{K}_t^-(v) \quad \geq \quad & (v - R_N)\frac{\underline{\gamma}\sqrt{t}}{2\overline{\Delta}\sqrt{\log(t+1)}} - \frac{\overline{\gamma}^2\big(C_0 + \log\log(t+1)\big)}{2(1-1/(1+c))} \\ & - \sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2} \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)} \delta_{\lceil \log_2 t \rceil}} - \frac{4}{3c}\delta_{\lceil \log_2 t \rceil}. \end{aligned} \tag{29}$$

Recall that $\mathcal{B}_t^{\pm}$ is defined by

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

Therefore, to show that a candidate value $v$ does not belong to $\mathcal{B}_t^{\pm}$, it suffices to verify the converse inequality

$$\mathcal{K}_t^{\pm}(v) \geq \frac{1}{\alpha_1},$$

since this directly implies $v \notin \mathcal{B}_t^{\pm}$.

On the event where Eqn. (29) holds, any $v > R_N$ satisfying

$$v - R_N \geq \underbrace{\frac{\log\left(\frac{1}{(1-\theta)\alpha_1}\right) + \frac{\overline{\gamma}^2\left(C_0 + \log\log(t+1)\right)}{2(1-1/(1+c))} + \sqrt{\frac{2\rho\overline{\gamma}^2}{\underline{\Delta}^2}\sum_{i=1}^{t}\frac{\mathrm{Var}(\hat{S}_i|\mathcal{F}_{i-1})}{i\log(i+1)}\delta_{\lceil \log_2 t \rceil}} + \frac{4}{3c}\delta_{\lceil \log_2 t \rceil}}{\underline{\gamma}\sqrt{t}/(2\overline{\Delta}\sqrt{\log(t+1)})}}_{ⓕ^-} \tag{30}$$

ensures that $\log \mathcal{K}_t^-(v) \geq \log(1/((1-\theta)\alpha_1))$, and hence $\mathcal{K}_t^-(v) \geq 1/((1-\theta)\alpha_1)$. Using $\mathcal{K}_t^{\pm}(v) \geq (1-\theta)\,\mathcal{K}_t^-(v)$, we further obtain $\mathcal{K}_t^{\pm}(v) \geq 1/\alpha_1$. Consequently, any such $v$ is excluded from $\mathcal{B}_t^{\pm}$.

To turn the exclusion condition Eqn. (30) into an explicit radius, we upper bound $ⓕ^-$. There exists a constant $G_3 > 0$, depending only on $\underline{\Delta}, \overline{\Delta}, c, \rho, \underline{\gamma}, \overline{\gamma}, \alpha_1, \theta, \omega$, such that for all $t \geq 2t_c$,

$$\begin{aligned} ⓕ^- \;\leq\; & G_3\sqrt{\frac{\log(t+1)}{t}} \times \\ & \left[1 + \log\log(t+1) + \sqrt{\sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\delta_{\lceil \log_2 t \rceil}} + \delta_{\lceil \log_2 t \rceil}\right]. \end{aligned}$$

Using Eqn. (22), which gives

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

together with

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

we obtain

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

for a constant $G_4$ depending only on $\underline{\Delta}, \overline{\Delta}, c, \rho, \underline{\gamma}, \overline{\gamma}, \alpha_1, \theta, \omega$.

Putting the pieces together, on the event where Eqn. (29) holds, every $\upsilon > R_N$ with

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

is excluded from $\mathcal{B}_t^{\pm}$. Equivalently, $\mathcal{B}_t^{\pm}$ cannot contain any point bigger than

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

and hence the upper endpoint satisfies, with probability at least $1 - \omega/2$, simultaneously for all $t \geq 2t_c$,

$$\sup \mathcal{B}_t^{\pm} \leq R_N + G_4 \sqrt{\frac{\log(t+1) \log \log(t+1)}{t} \left( \log \log(t+1) + \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \right)}.$$

**Two-sided width bound and shrinkage.** On the intersection of the corresponding high-probability events for $\mathcal{K}_t^{+}$ and $\mathcal{K}_t^{-}$, which holds with probability at least $1 - \omega$ by a union bound, we have simultaneously for all $t \geq 2t_c$,

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

$$\sup \mathcal{B}_t^{\pm} \leq R_N + G_4 \sqrt{\frac{\log(t+1) \log \log(t+1)}{t} \left( \log \log(t+1) + \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \right)}.$$

Since both radii are of order

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

there exists a constant $G_5 > 0$ such that for all $t \geq 2t_c$,

$$\mathcal{B}_t^{\pm} \subseteq \left[ R_N - G_5 \sqrt{\frac{\log(t+1) \log \log(t+1)}{t} \left( \log \log(t+1) + \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \right)}, \right.$$
$$\left. R_N + G_5 \sqrt{\frac{\log(t+1) \log \log(t+1)}{t} \left( \log \log(t+1) + \sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i \log(i+1)} \right)} \right].$$

In particular, on this event,

$$\text{width}\left(\mathcal{B}_t^{\pm}\right) \leq 2G_5\sqrt{\frac{\log(t+1)\log\log(t+1)}{t}\left(\log\log(t+1)+\sum_{i=1}^{t}\frac{\text{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}.$$

Finally, recall that $\mathcal{M}_t^{\pm} = \bigcap_{i\leq t}\mathcal{B}_i^{\pm} \subseteq \mathcal{B}_t^{\pm}$ for each $t$. Therefore the same width bound holds for $\mathcal{M}_t^{\pm}$ as follows

$$\text{width}\left(\mathcal{M}_t^{\pm}\right) \quad \leq \quad 2G_5\sqrt{\frac{\log(t+1)\log\log(t+1)}{t}\left(\log\log(t+1)+\sum_{i=1}^{t}\frac{\text{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}.$$

Define

$$t_c = \min\left\{t \in \mathbb{Z}_+ : t\log(t+1) \geq ((1+c)\overline{\gamma})^2\right\}, \qquad t_0 = \max\{2t_c, 8\}. \tag{31}$$

For all $t \geq t_0$, we have

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

Indeed, $t \geq 2$ implies $\log(t+1) \leq 2\log t$, and $t \geq 8$ implies $\log\log(t+1) \leq 2\log\log t$.

Hence, after slightly enlarging the constant $G_5$ to $G_6$ if necessary, we obtain

$$\text{width}\left(\mathcal{M}_t^{\pm}\right) \leq G_6\sqrt{\frac{\log t\log\log t}{t}\left(\log\log t+\sum_{i=1}^{t}\frac{\text{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)}\right)}, \qquad \forall t \geq t_0.$$

This completes the proof. □

### B.7. Proofs of Theorem 6

We present the detailed proof of Theorem 6 as follows.

*Proof.* 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 pool examples are i.i.d., Hoeffding's inequality gives, for any $\beta > 0$,

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

Choosing

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

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

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 $R \in [R_N - \Delta_N(\alpha_2),\ R_N + \Delta_N(\alpha_2)]$. Combining these two statements implies

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

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,

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

This proves that $(C_t)_{1 \le t \le N}$ is $(1-\alpha)$ anytime-valid for $R$.

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

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

hence

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

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, we have

$$\operatorname{width}(C_t) \le \operatorname{width}\left(\mathcal{M}_t^{\pm}\right) + 2\Delta_N(\alpha_2).$$

By Theorem 5, for all sufficiently large $t$,

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

Moreover, by Theorem 3 together with Assumptions 1 and 2, for each $i \ge 1$ we have

$$\begin{aligned}
\operatorname{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right) &= \frac{1}{N^2} \sum_{j \in \mathcal{J}_{i-1}} \frac{\left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)^2}{q_i(j \mid i_{1:i-1}, D_N)} \\
&\quad - \left(\frac{1}{N} \sum_{j \in \mathcal{J}_{i-1}} \left(\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)\right)\right)^2 \\
&\le \frac{1}{N^2} \sum_{j \in \mathcal{J}_{i-1}} \frac{(U-L)^2}{q_i(j \mid i_{1:i-1}, D_N)} \\
&\le \frac{1}{N^2} \sum_{j \in \mathcal{J}_{i-1}} (U-L)^2 \frac{N-i+1}{\beta} \\
&= \frac{(U-L)^2 (N-i+1)^2}{\beta N^2} \\
&\le \frac{(U-L)^2}{\beta}.
\end{aligned}$$

Since $\log(i+1) \geq \log i$ for $i \geq 2$, we have

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

Therefore, there exists an absolute constant $C_0$ such that for all $t \geq 2$,

$$\sum_{i=1}^{t} \frac{1}{i\log(i+1)} \leq C_0 + \log\log(t+1)\ .$$

Hence

$$\begin{aligned}\sum_{i=1}^{t} \frac{\mathrm{Var}\left(\hat{S}_i \mid \mathcal{F}_{i-1}\right)}{i\log(i+1)} &\leq \frac{(U-L)^2}{\beta}\sum_{i=1}^{t}\frac{1}{i\log(i+1)} \\ &\leq \frac{(U-L)^2}{\beta}\big(C_0 + \log\log(t+1)\big) \\ &= O(\log\log t)\ .\end{aligned}$$

Substituting this into the bound above yields

$$\mathrm{width}\left(\mathcal{M}_t^{\pm}\right) = O\left(\log\log t\sqrt{\frac{\log t}{t}}\right).$$

On the other hand, by the definition of $\Delta_N(\alpha_2)$,

$$\Delta_N(\alpha_2) = O\left(\sqrt{\frac{1}{N}}\right) \quad \text{for fixed } \alpha_2.$$

Combining these bounds yields

$$\mathrm{width}\left(C_t\right) = O\left(\log\log t\sqrt{\frac{\log t}{t}}\right) + O\left(\sqrt{\frac{1}{N}}\right).$$

This completes the proof. □

## C. 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, there is a gap between its fixed-partition certification and its proposed algorithm: the 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. This creates a double-dipping issue, where 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 the gap between the theoretical guarantees and the proposed algorithm 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 evaluation pool $D_N = \{\boldsymbol{x}_j\}_{j=1}^{N}$. For each sample $\boldsymbol{x}_j$, we first draw a latent group $G_j \in \{1, \ldots, G\}$. The latent group determines the risk level of the sample, but it is not encoded in the embedding

used for partitioning. Specifically, each group $g$ is assigned a risk probability $p_g$, and we choose these $G$ probabilities to be evenly spaced between $0.1$ and $0.9$. Given $G_j = g$, the realized per-sample risk value is generated as

$$Z_j \mid G_j = g \sim \text{Bernoulli}(p_g),$$

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

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

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

$$\boldsymbol{x}_j \perp (G_j, Z_j).$$

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 a double-dipping effect: 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 $Z_j$. 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.

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

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

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 $Z_j$ 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 proportional to the true residual magnitude:

$$q_t^\star(j) \propto |Z_j - \tilde{\ell}_f(\boldsymbol{x}_j)| \,.$$

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

# D. Experimental Setup Details

This appendix expands the main-text setup of Section 5.1. Throughout, $\mathcal{D}_N = \{x_j\}_{j=1}^N$ denotes the evaluation pool containing all available samples from the benchmark split used for evaluation.

## D.1. Benchmarks

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

**SST-2** [25][10]. Single-sentence binary sentiment classification, derived from the Stanford Sentiment Treebank by restricting the phrase-level annotations on the Pang–Lee movie review corpus to the polarity-only sentence labels. We use the validation split as the evaluation pool, giving $N_{\text{SST2}} = 67{,}349$. The label space $\mathcal{Y} = \{\text{negative}, \text{positive}\}$ is approximately balanced. SST-2 instantiates the $|\mathcal{Y}| = 2$ regime: under 0-1 loss the residual $\ell - \widetilde{\ell}_f$ takes values in a discrete two-modal set whenever the surrogate is reasonably calibrated, which produces a sharply separated acquisition score and a highly non-uniform variance-optimal proposal $q_t^\star$ in Thm. 4.

**MMLU** [1][11]. 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 $\{\text{A}, \text{B}, \text{C}, \text{D}\}$ 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 $N_{\text{MMLU}} = 115{,}700$.

**AG News** [26][12]. 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.

Together the three benchmarks span the three axes whose effect on CELEUSś behaviour is captured by the analysis. The label-space cardinality varies as $|\mathcal{Y}| \in \{2, 4, 4\}$, controlling the granularity of the surrogate prediction and therefore the reachable correlation between $\ell$ and $\widetilde{\ell}_f$. The within-pool difficulty distribution varies from approximately homogeneous (SST-2, AG News) to substantially heterogeneous across 57 subjects (MMLU), which controls the tail of the residual distribution and the leverage available to $q_t^\star$. In every case the full pool plays the role of $\mathcal{D}_N$, 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 $\Delta_N(\alpha_2)$.

## D.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, holding the surrogate substantially smaller than the target so that surrogate scoring is strictly cheaper than target scoring.

1. **Llama-3-8B** $\rightarrow$ **Llama-3-70B**: same family, same generation.
2. **Llama-3-8B** $\rightarrow$ **DeepSeek-67B**: cross-family, same generation.
3. **Llama-3-8B** $\rightarrow$ **Mixtral-8$\times$7B**: cross-family, same generation[13].
4. **Llama-3-8B** $\rightarrow$ **Qwen-2.5-72B**: cross-family, same generation.
5. **Llama-2-7B** $\rightarrow$ **Llama-3-70B**: same family, cross-generation.

[10] https://huggingface.co/datasets/stanfordnlp/sst2
[11] https://huggingface.co/datasets/cais/mmlu
[12] https://huggingface.co/datasets/fancyzhx/ag_news
[13] Mixtral parameter counts are reported as active parameters per token.

6. **Llama-2-7B → DeepSeek-67B**: cross-family, cross-generation.

For all open-weights models we use the publicly released checkpoints without further fine-tuning. All predictions are obtained at temperature 0. Following Berrada et al. [15], 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.

### D.3. Baselines

**EVALUE** [12]. Applies the hedged-capital confidence sequence of Eqn. (2) to uniform sampling without a surrogate ($\tilde{\ell} \equiv 0$). All other components (prediction fractions, scaling, $\alpha$-split) match CELEUS for fairness; the only differences are the absence of the surrogate and the choice of sampler.

**CELEUS (W/O SURR).** Ablates the surrogate-assisted approximation by setting $\tilde{\ell} \equiv 0$, while retaining the uncertainty-guided sampling strategy. Comparison against EVALUE isolates the contribution of active sampling at $\tilde{\ell} \equiv 0$; comparison against CELEUS isolates the contribution of the surrogate at the same uncertainty-guided sampling strategy.

**Oracle.** Uses the variance-minimizing proposal $q_t^\star(j) \propto |\ell(f(\boldsymbol{x}_j), y_j) - \tilde{\ell}_f(\boldsymbol{x}_j)|$ from Thm. 4. Computing $q_t^\star$ requires the true per-sample residuals, which are unknown for unevaluated samples, so ORACLE is not deployable in practice. We include it as an oracle reference for the best variance reduction attainable under the inferential signal in Eqn. (1); the gap between CELEUS and ORACLE therefore measures the suboptimality of the practical sample-selection score.

**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, there is a gap between its theoretical certification and its proposed algorithm: the 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. This creates a double-dipping issue, where 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 gap, and focus the main efficiency comparisons on methods designed to be anytime-valid.

### D.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 $\alpha_1 = \alpha_2 = 0.025$: $\alpha_1$ is used for the anytime-valid $e$-process inference of the finite-pool risk $R_N$, while $\alpha_2$ is used for the finite-pool-to-population correction via Hoeffding's inequality. App. D.4 studies the sensitivity of the number of evaluated samples required to reach $\epsilon$ to this split.

**$e$-process construction.** Following the standard $e$-process construction of Waudby-Smith and Ramdas [12], 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. (2), 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 $a \leq \hat{S}_t \leq b$ for all rounds, and scale the signal as $\bar{S}_t = (\hat{S}_t - a)/(b - a)$ before constructing the $e$-process.

**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. D.4. We stop the evaluation at the first time $t$ such that $\mathrm{width}(C_t) \leq \epsilon$, with $\epsilon = 0.05$ in all main experiments.

### D.5. Trials and reporting

Each experimental configuration (i.e., a surrogate-target pair on a particular dataset) is run for 50 random 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.

## E. 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 3 reports the empirical miscoverage rate at this stopping time. Both CELEUS and EVALUE remain below the target level $\alpha = 0.05$ on all datasets, supporting the anytime-valid coverage guarantee under adaptive stopping.

**Table 3.** 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] |

## F. 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 $\ell_{\mathrm{ce}}(f, \boldsymbol{x}, y) = -\log p_f(y \mid \boldsymbol{x})$, which is bounded in $[L, U]$ by clipping at the maximum cross-entropy value observed over the pool. The sample-selection rule follows Strategy A in Rem. 2, and all other hyperparameters are kept the same as in App. D. Fig. 3 shows the resulting CI-width trajectories.

**Table 4.** 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). All $e$-process-based methods stay within the nominal $\alpha = 0.05$ on every dataset, consistent with the by-construction anytime-validity of $\mathcal{M}_t^{\pm}$ in Eqn. (3) and of $C_t$ via 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. 4 shows that the empirical miscoverage at the stopping time remains below the prescribed level $\alpha = 0.05$ for all cross-entropy configurations, consistent with the 0-1 risk results in Tabs. 1 and 3. This follows from the $e$-process construction: the anytime-valid guarantee does not depend on the specific risk form, but only requires the boundedness condition of evaluation score in Assumption 1.

**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 benefit of surrogate-assisted approximation 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.

## G. Sample-selection strategy comparison

The deployed configurations in Sec. 5 and Sec. F 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, holding the trajectory machinery, $e$-process, and stopping rule fixed; only the formula that produces $q_t$ changes

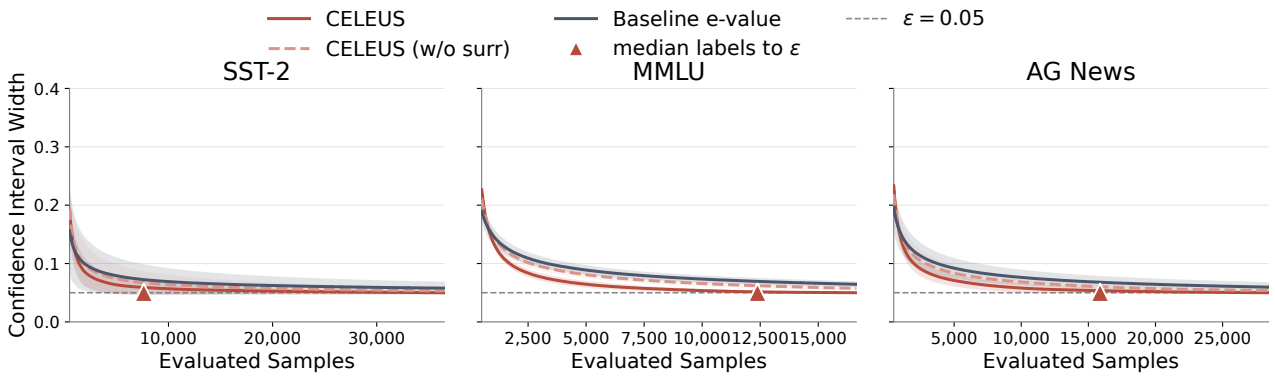


**Figure 3.** Confidence interval width versus number of evaluated samples under cross-entropy loss on SST-2, MMLU, and AG News. Each curve is the mean width at each label budget, aggregated over 50 seeds and the four surrogate-target pairs of App. D.2; the shaded band is $\pm 1\sigma$ across the per-pair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

across runs.

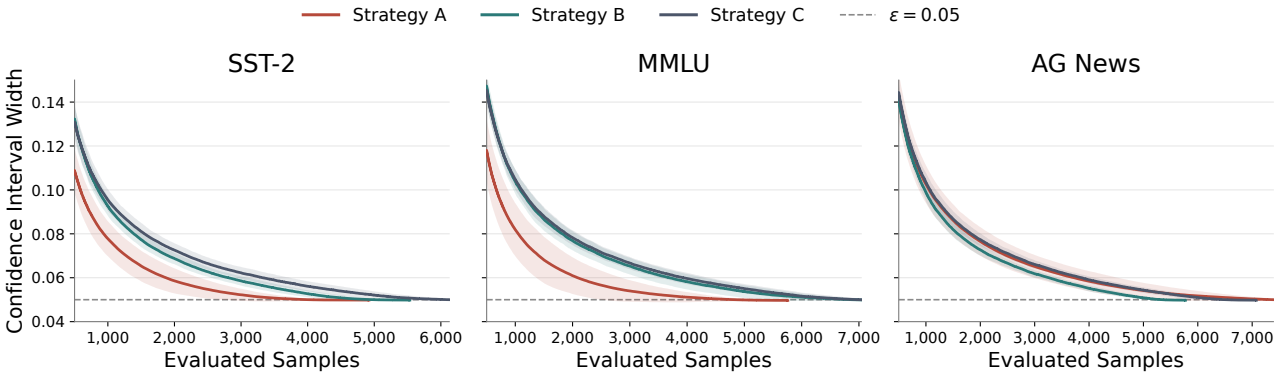


**Figure 4.** Sample-selection strategy comparison under 0-1 risk. We compare the sample-selection strategies in Rem. 1 by plotting CI width versus the number of evaluated samples on SST-2, MMLU, and AG News. Each curve is averaged over 50 seeds, and the dashed line marks the target precision $\epsilon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

**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 only in which quantity plays the role of the target-loss proxy inside the estimator in Eqn. (4): Strategy B treats $\ell_{\text{disp}}$ as the target-loss proxy and $\ell_{\text{soft}}$ as the surrogate score, whereas Strategy C reverses this assignment. This leaves the sampling order unchanged but changes the correction term, which can affect the conditional variance of $\widehat{S}_t$ and the cumulative-variance term in Thm. 5. 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 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

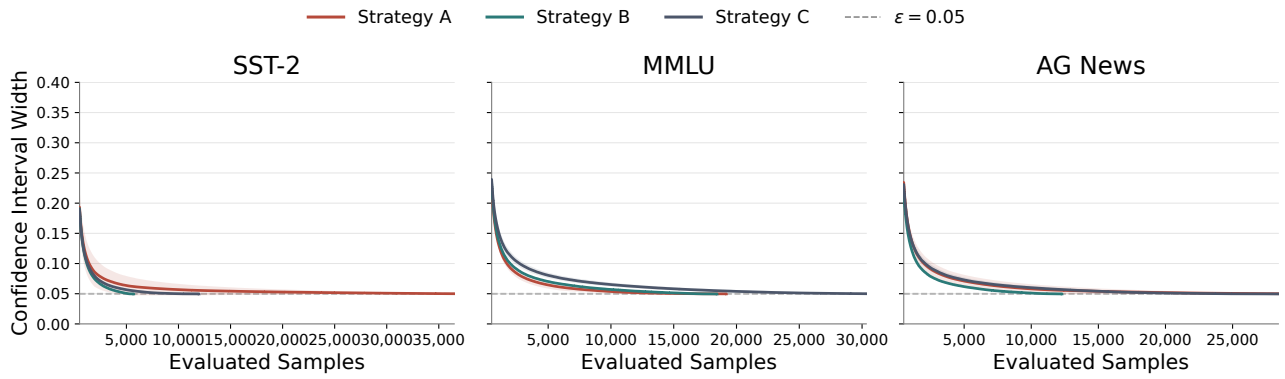


**Figure 5.** Sample-selection strategy comparison under cross-entropy risk. We compare the sample-selection strategies in Rem. 2 by plotting CI width versus the number of evaluated samples on SST-2, MMLU, and AG News. Each curve is averaged over 50 seeds, and the dashed line marks the target precision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

same sample-selection ranking as Strategy A but a different correction term in Eqn. (4). 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.

## H. Hyperparameter sensitivity

CELEUS has four configuration parameters that may affect the number of evaluated samples needed to reach the target precision, while leaving the anytime-valid coverage guarantee unchanged: the split of $\alpha$ between the $e$-process and the population correction, the lower bound $\beta$ on sampling probabilities, the hedge weight $\theta$, 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 $\beta$, we sweep over $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$. Tab. 5 reports the number of evaluated samples needed to reach $\epsilon$ on four representative dataset-model configurations.

**Sensitivity to the $\alpha$-split.** Tightening $\alpha_1$ 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 $\alpha_1$ to $0.045$ reduces the required number by $14\%$-$19\%$. This trend is consistent with the $e$-process construction: a smaller $\alpha_1$ raises the rejection threshold $1/\alpha_1$, so more evidence is needed before the CI can shrink. Although a larger $\alpha_1$ leaves a smaller $\alpha_2$ 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 $\theta$.** Changing $\theta$ 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. (2).

**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.

**Sensitivity to $\beta$.** The parameter $\beta$ 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 $\beta$ is too small, some sampling probabilities can become very small, which slows down the accumulation of evidence and leads to wider CIs. When $\beta$ is too

**Table 5.** 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 $\rightarrow$ Mixtral-8×7B on SST-2; ② Llama-2-7B $\rightarrow$ Qwen-2.5-72B on MMLU; ③ Llama-3-8B $\rightarrow$ DeepSeek-67B on AG News; and ④ Llama-3-8B $\rightarrow$ Llama-3-70B on SST-2. 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** |
|---|---|---|---|---|
| $\alpha$-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** |
| $\theta$, 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** |

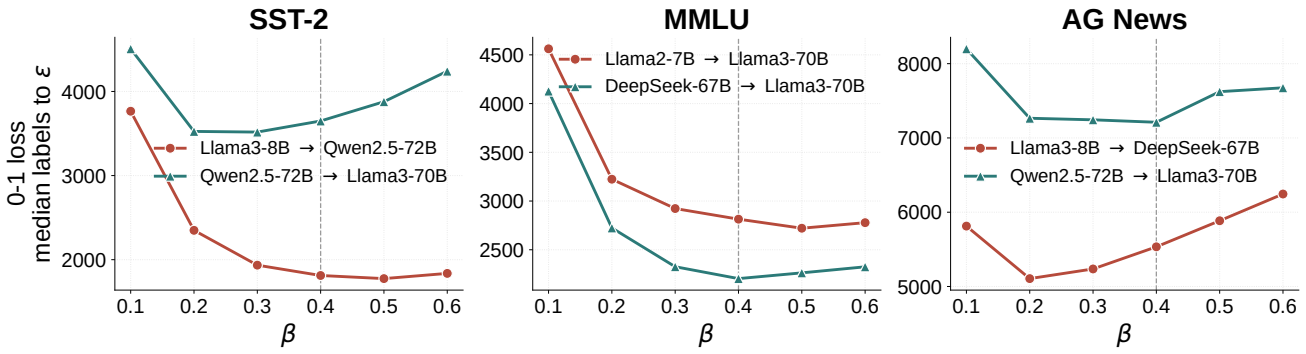


**Figure 6.** Sensitivity of median samples-to-$\epsilon$ to the betting-magnitude floor $\beta$, on three benchmarks (SST-2, MMLU, AG News) and two surrogate-target pairs per benchmark. Each curve is the median number of evaluate samples used, over 50 seeds at $\epsilon = 0.05$ under 0-1 loss. The vertical 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.

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.

## I. Computational cost

Tab. 6 reports the per-round wall-clock cost of CELEUS and EVALUE baseline on the same four (dataset, pair) experimental configurations used in App. H. The wall-clock figures exclude target-model inference, which is identical across methods at fixed acquisition order. The reported numbers therefore isolate the extra cost that CELEUS pays for the e-process update, the LURE-weighted estimator of Eqn. (4), and the surrogate-driven proposal.

Most of the per-round cost of CELEUS comes from running the surrogate model on the remaining unevaluated samples, which takes time proportional to $N - t$. Because the surrogate is much smaller than the target, this extra work is cheap in absolute terms. Even on the slowest experimental configuration, mmlu/2, CELEUS adds about 0.51 s per round. Combined with the 54–62 % reduction in samples-to-$\epsilon$ reported in Tab. 1, the time saved by labelling fewer samples is much larger than the time added by the per-round overhead.

**Table 6.** Per-round wall-clock cost in seconds, averaged over 50 seeds. Each column indexes a dataset-pair: ①: llama2-7b → Mixtral-8×7B on SST-2, ②: llama2-7b → Qwen2.5-72B on MMLU, ③: llama3-8b → DeepSeek-67B on AG News, ④: llama3-8b → Llama-3-70B on SST-2.

| 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. Results for each surrogate-target model pair

In addition to the dataset-level averages, we report the CI-width trajectories and evaluated samples-to-$\epsilon$ 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 acts as a lower envelope, 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 role of surrogate scoring. 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 benefit of surrogate assistance is smaller but the anytime-valid guarantee is still preserved. This pattern is consistent with the variance bound in Thm. 5: informative surrogate scores reduce the residual $\ell - \widetilde{\ell}_f$, which lowers the cumulative conditional variance and accelerates CI shrinkage. 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 → Mixtral-8×7B row, the median samples-to-$\epsilon$ 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 what the oracle proposal in Thm. 4 optimizes. The oracle proposal minimizes 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.

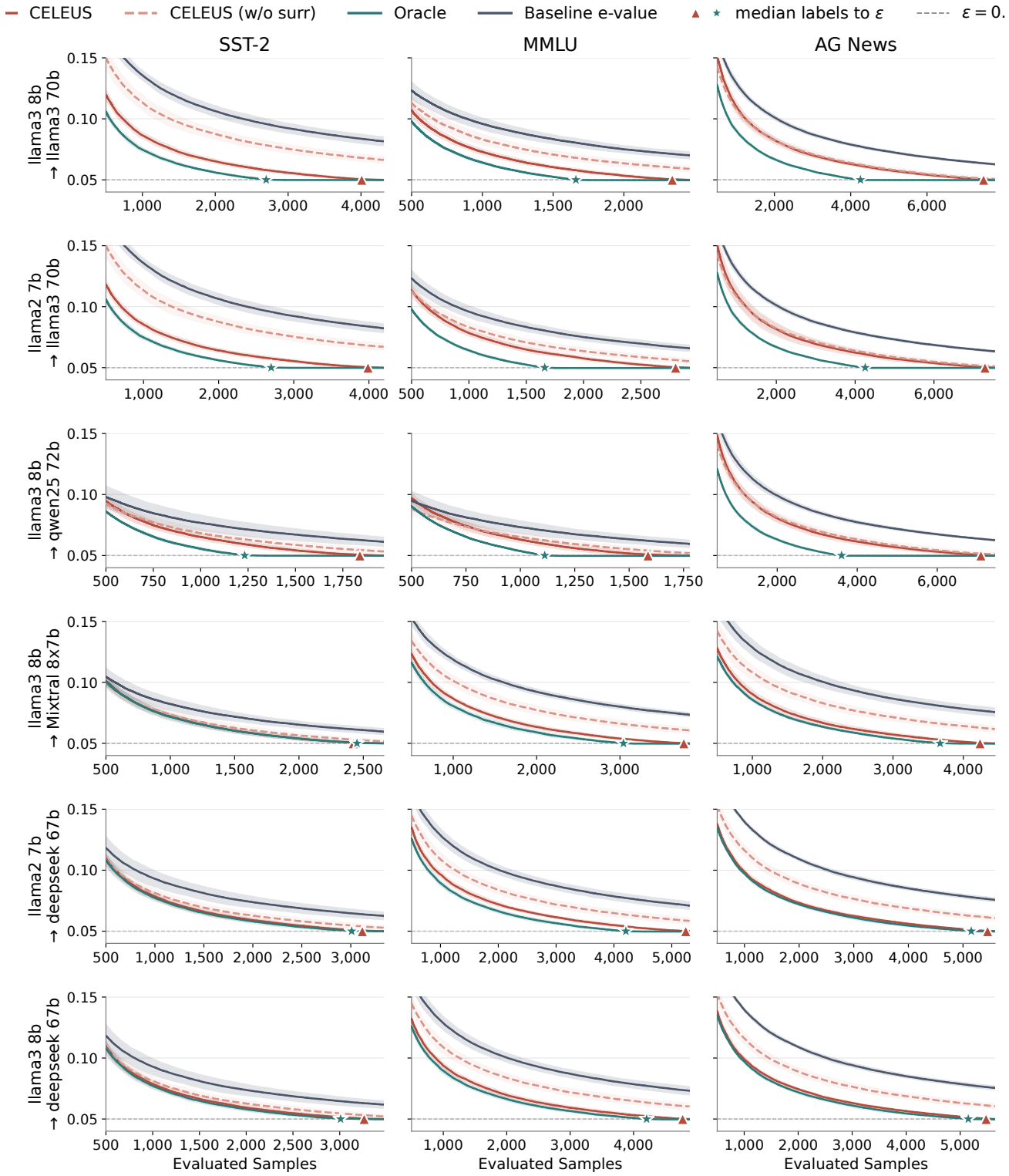


**Figure 7.** Confidence interval width versus evaluated samples on each (*dataset*, *surrogate*, *target*) experimental configuration. Rows index the six surrogate-target pairs of App. D.2, and columns index the three benchmarks. Each curve is the mean width over 50 seeds. The dashed line marks the target precision $\epsilon = 0.05$, and triangles mark the median samples-to-$\epsilon$. The ordering CELEUS $\leq$ CELEUS (W/O SURR) $\leq$ EVALUE holds on every panel, and the gap between CELEUS and CELEUS (W/O SURR) tracks the role of surrogate scoring.