TRUST THE STREAM: ONLINE CONFIDENCE CALIBRATION FOR SELECTIVE LLM JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used as automated judges, motivating selective evaluation methods that certify human agreement by accepting only sufficiently confident predictions. However, with limited calibration data, noisy estimates of the confidence distribution can destabilize the certified threshold and degrade the coverage–agreement trade-off. We introduce **O**nline **C**onfidence **C**alibration (OCC), which continuously refines the empirical confidence distribution using confidence scores observed during deployment. OCC maps confidence scores to their empirical percentiles, re-scores the fixed calibration set, and re-certifies the acceptance threshold without requiring additional human labels. We prove that the resulting threshold converges to its population counterpart, achieving finite-time exact recovery with high probability and eventual almost-sure lock-in, while preserving the human-agreement guarantee at every round. Across six benchmarks, eleven LLM judges, and four confidence functions, OCC consistently improves threshold stability and calibration quality, certifies stronger agreement at matched coverage than label-free baselines, approaches the population-CDF oracle.

1 INTRODUCTION

Large language models (LLMs) are increasingly used as automated evaluators of output quality and preference (Zheng et al., 2023; Dubois et al., 2023; Chiang & Lee, 2023), raising the question of when their judgments can be trusted (Xiong et al., 2024). A leading approach is the selective evaluation framework of Jung et al. (2025), which accepts a judge's prediction only when its confidence score exceeds a threshold certified by fixed-sequence hypothesis testing on a calibration set. Under standard exchangeability, this delivers a high-probability bound on the human-disagreement risk on accepted instances. Related conformal approaches (Yadkori et al., 2024; Mohri & Hashimoto, 2024) target adjacent problems such as hallucination abstention, not the selective-evaluation setting we consider.

Fixed-sequence certification (Bauer, 1991; Jung et al., 2025) sweeps confidence thresholds from high to low, selecting the most permissive one whose estimated risk meets the target guarantee; it relies on an approximately monotonic confidence–agreement relationship. Within this framework, the confidence distribution is estimated once from a finite calibration set. The population distribution is deterministic, but its empirical estimate varies with the sample drawn: different calibration sets of the same size yield different empirical distributions (Figure 1(a), red curves) and different certified thresholds. By contrast, if threshold selection were based directly on the population confidence distribution (Figure 1(a), dashed black), the distribution-estimation variability would be eliminated, yielding a stable, statistically valid threshold. We refer to this threshold as the population threshold. In practice, however, calibration sets are small, often a few hundred instances, and the resulting empirical distribution can deviate substantially from the population (Figure 1(b), red bars). When confidence scores are unevenly distributed, the number exceeding a threshold changes irregularly, giving high-variance agreement estimates in low-density regions. These jagged, potentially non-monotonic confidence–agreement curves (Figure 1(c), red) destabilize the fixed-sequence sweep: a noise-induced spike or dip in the estimated risk can stop it too early or too late. The certified threshold thus scatters widely around its population value, and deployment settles at a worse coverage–agreement trade-off than the population threshold itself would deliver.

Enlarging the calibration set would reduce this variability, but additional human annotations are prohibitively expensive. In contrast, deployment naturally generates an unlimited stream of unlabeled

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

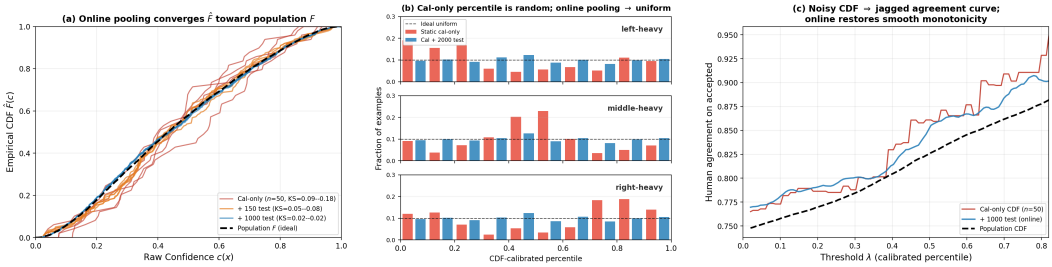


Figure 1: **Small calibration sets produce noisy confidence calibration; online pooling restores accuracy and monotonicity.** (a) Five independent draws (n=50) yield empirical CDFs (red) that deviate from the population CDF (dashed black); pooling 150 (orange) then 1000 (blue) unlabeled confidences closes the gap. (b) Held-out percentile distributions under three small draws (red) vary markedly; online pooling (blue) collapses them to near-uniform. (c) Held-out confidence–agreement curve: jagged and non-monotone under a small-cal CDF (red); online pooling (blue) smooths it toward the population curve (dashed black).

confidence scores, one for each incoming instance. Our key observation is that estimating the confidence distribution requires only these confidence values, not their corresponding labels. Because the confidence distribution determines the location of the certified threshold, we normalize confidence scores using their empirical cumulative distribution function (CDF). Unlike a fixed calibration set, this CDF can be continuously refined by pooling confidences from both the calibration data and the deployment stream, yielding a progressively sharper estimate of the underlying distribution. Crucially, this benefit is specific to the CDF space: the unlabeled observations inform only the distribution, so thresholding raw confidences directly gains nothing from them.

With each new unlabeled instance, we update the CDF, re-score the calibration set, and re-certify the threshold on it by the fixed-sequence procedure of Jung et al. (2025). As deployment progresses, the pooled CDF converges toward the population (Figure 1(b), blue bars), driving the certified threshold toward its population counterpart. It also restores a smoother, more uniform score density, so the held-out confidence–agreement curve becomes approximately monotonic (Figure 1(c), blue), a calibration-quality gain. Pooling the stream shrinks this scatter rather than shifting the threshold in any fixed direction, since the static threshold is a near-unbiased estimate of its population counterpart. We call this per-round procedure Online Confidence Calibration (OCC), illustrated in Figure 2. It touches only the fixed calibration set, so its cost scarcely grows with the deployment stream and is negligible beside a single judge evaluation. To summarize, the contributions are as follows:

- ⋆ **Problem characterization.** We find that small calibration sets make the confidence-distribution estimate noisy and can break the monotonicity the certification relies on, turning the certified threshold into a high-variance estimate of its population value.
- ⋆ **Online calibration algorithm.** We recast the estimated confidence distribution as a label-free empirical CDF, continually refined by the unlabeled deployment stream, and re-certify the threshold on the original calibration set after each update.
- ⋆ **Theoretical analyses.** We prove that, beyond a finite, explicit number $t^\star$ of unlabeled rounds, the online threshold matches the population threshold exactly at each round with high probability, and that the human-agreement guarantee of Jung et al. (2025) is preserved at every round.
- ⋆ **Empirical evidence.** Across six benchmarks (e.g., TL;DR, Chatbot Arena, HH-RLHF, AlpacaEval, HelpSteer3, Community Alignment), eleven judges (e.g., MiMo, Gemma, GPT-OSS), and four scoring functions, OCC improves calibration quality and threshold stability, certifies stronger agreement at matched coverage, and preserves the guarantee throughout.

2 Preliminary

2.1 Setting and Notation

We consider a judge model that, given a query x and m candidate responses $g_1, \ldots, g_m$, selects the one it deems best. A scoring function $\mathbb{S}$ assigns each candidate a preference score $\mathbb{S}(x, g_i)$, $i \in \{1, \ldots, m\}$; under the default setting, the judge's prediction is the highest-scoring candidate,

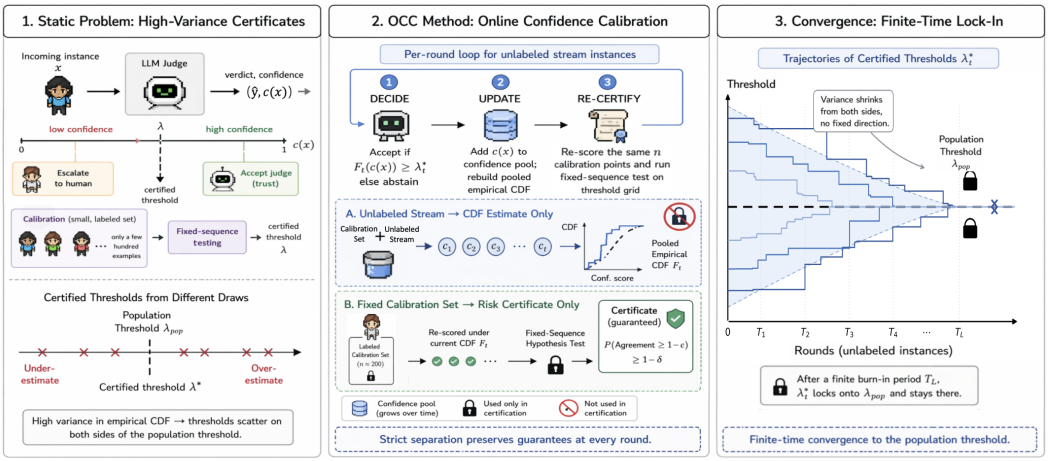


Figure 2: **OCC stabilizes the certified threshold by refining the confidence distribution with unlabeled deployment data, preserving the human-agreement guarantee at every round. (1) Static problem.** A judge's prediction is accepted only when its confidence $c(x)$ clears a threshold certified on a small labeled calibration set of size $n{\approx}200$; the small sample makes the certified threshold λ^* scatter on both sides of the population threshold λ^{pop}. **(2) Method.** Each round, OCC decides accept/abstain, adds the new confidence to the pooled empirical CDF, and re-certifies the threshold on the same n calibration points; the unlabeled stream informs only the CDF, the calibration labels only the certificate. **(3) Convergence.** As the pool grows, the scatter shrinks, and after a finite burn-in the per-round certified threshold λ_t^* locks onto λ^{pop} (Theorem 3.2).

$\hat{y}(x) = \arg\max_i \mathbb{S}(x, g_i)$. The ground-truth label $y \in \{1, \ldots, m\}$ is the candidate human annotators prefer, and the judge agrees with the human label on x when $\hat{y}(x) = y$.

We assume access to a labeled calibration set $\mathcal{D}_{\mathrm{cal}} = \{(x_i, y_i)\}_{i=1}^n$ with human-annotated preferences. During deployment, unseen (unlabeled) instances $x_{n+1}, x_{n+2}, \ldots$ arrive one at a time. At round t we observe the judge's scalar confidence $c(x_{n+t})$ in its own prediction (a measure of how certain it is, e.g. the top-1 probability $\max_i \mathbb{S}(x_{n+t}, g_i)$; Section 2.2 lists other choices), but not the human preference y_{n+t}. At each round we must decide, using only the information observed so far, whether to accept the judge's prediction on the arriving instance or to abstain and defer it to a human.

2.2 Static Selective Evaluation

In this subsection, we briefly review the selective evaluation pipeline of Jung et al. (2025), which our method builds upon.

Confidence scoring. The scalar confidence $c(x)$ can be instantiated in many ways: the top-1 probability $\max_i \mathbb{S}(x, g_i)$, the margin $\mathbb{S}(x, g_{\hat{y}}) - \max_{i \neq \hat{y}} \mathbb{S}(x, g_i)$, vote-based agreement, or KL-divergence-based measures. Our framework is agnostic to this choice.

Threshold selection with risk control. We are given a tolerated disagreement rate ε and a confidence level $1{-}\delta$. A threshold λ partitions instances into an accepted set ($c(x) \geq \lambda$) and an abstained set ($c(x) < \lambda$); we call the fraction of accepted instances the *coverage*. Let $N(\lambda) = |\{i : c(x_i) \geq \lambda\}|$ be the number of calibration instances that pass the threshold, and $Y(\lambda) = \sum_{i:\, c(x_i) \geq \lambda} \mathbf{1}\{\hat{y}(x_i) \neq y_i\}$ the number among them on which the judge disagrees with the human label. Under standard exchangeability conditions, $Y(\lambda) \mid N(\lambda)$ follows a binomial distribution, and an exact $(1{-}\delta)$ upper confidence bound on the disagreement rate is

$$\widehat{R}^+(\lambda) = \sup\left\{r \in [0,1] : \mathbb{P}\big(\mathrm{Bin}(N(\lambda), r) \leq Y(\lambda)\big) \geq \delta\right\}. \tag{1}$$

Let $\Lambda \subseteq [0,1]$ be the finite set of evaluated candidate thresholds. Sweeping from high to low and stopping at the first violation of $\widehat{R}^+(\lambda) \leq \varepsilon$ (Bauer, 1991) gives

$$\lambda^* = \min\left\{\lambda \in \Lambda : \widehat{R}^+(\lambda') \leq \varepsilon \ \text{ for all } \ \lambda' \in \Lambda \cap [\lambda, 1]\right\}. \tag{2}$$

If no candidate passes, the evaluator abstains. Choosing λ^* this way yields the human-agreement guarantee of Jung et al. (2025): with probability at least $1-\delta$ over the draw of $\mathcal{D}_{\text{cal}}$, the disagreement rate among accepted instances is controlled,

$$\mathbb{P}\big(\hat{y}(x) \neq y \,\big|\, c(x) \geq \lambda^*\big) \leq \varepsilon, \tag{3}$$

i.e., the judge agrees with humans on at least a $1-\varepsilon$ fraction of accepted instances. Evaluating $Y(\lambda)$, and hence certifying λ^*, requires the human labels in $\mathcal{D}_{\text{cal}}$. A confidence score whose risk decreases with its threshold can make this ordering effective. Monotonicity is a consideration for statistical power, rather than a requirement of fixed-sequence family-wise error control itself.

The role of the confidence distribution. A uniform grid in raw-score space can allocate many candidate thresholds to sparsely populated regions and few to regions containing substantial probability mass. A percentile grid instead targets approximately uniform increments of score mass. Estimating that coordinate system from a small sample introduces variability in the corresponding raw cutoffs. Improving the CDF can reduce this variability, but it cannot replace the labels needed to assess disagreement or remove the uncertainty associated with those labels. A detailed discussion of related work is provided in Appendix A

3 Online Confidence Calibration

We introduce Online Confidence Calibration (OCC), a CDF-based confidence-calibration layer progressively refined using unlabeled deployment data. At the start of round t, OCC maintains a state $(\widehat{F}_t, \lambda_t^*)$: an empirical CDF $\widehat{F}_t$ pooled from the calibration confidences and all stream confidences seen before round t, and the threshold λ_t^* currently certified on the calibration set (both initialized from the calibration set alone). On the arriving instance x_{n+t}, OCC performs two steps (note that deployment labels are never used):

(i) **Decide:** accept x_{n+t} if its calibrated confidence satisfies $\widehat{F}_t(c(x_{n+t})) \geq \lambda_t^*$, and abstain otherwise;

(ii) **Update:** append the observed confidence $c(x_{n+t})$ to the pool to form $\widehat{F}_{t+1}$, then re-score the labeled calibration set and re-certify the threshold λ_{t+1}^* by fixed-sequence testing (Section 3.2).

3.1 CDF Calibration Layer

We first map each raw confidence score to its percentile rank via the empirical CDF. The estimate is initialized on the calibration set and updated recursively as each unlabeled stream confidence arrives:

$$\widehat{F}_1(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{c(x_i) \leq u\}, \qquad \widehat{F}_{t+1}(u) = \frac{(n+t-1)\,\widehat{F}_t(u) + \mathbf{1}\{c(x_{n+t}) \leq u\}}{n+t}, \quad t \geq 1. \tag{4}$$

Unrolling the recursion, $\widehat{F}_t$ is the empirical CDF of the pooled sample $\{c(x_1), \ldots, c(x_{n+t-1})\}$, on which the analysis of Section 3.3 operates. As t grows, this pool approaches the true deployment-time confidence distribution, so later instances are scored under progressively more accurate percentiles.

3.2 Re-Scoring and Threshold Re-Certification

With the updated $\widehat{F}_t$, we re-compute the calibrated confidence for all calibration instances:

$$s_i^{(t)} = \widehat{F}_t(c(x_i)), \quad i = 1, \ldots, n. \tag{5}$$

We then apply the standard fixed-sequence testing procedure ((1) and (2)) to $\{(s_i^{(t)}, y_i)\}_{i=1}^n$ to obtain a re-certified threshold λ_t^*. The sweep is performed over a fixed grid $G = \{0, \Delta, 2\Delta, \ldots, 1\} \subset [0,1]$ of candidate thresholds with spacing Δ (we use $\Delta = 0.005$), so the certified threshold satisfies $\lambda_t^* \in G$. The evaluated candidate path is $\Lambda \subseteq G$; Appendix C specifies its start for the deployment-frontier comparison.

The threshold is always computed on the same n labeled calibration instances; the only change from the static baseline is that the scores $s_i^{(t)}$ use a sharper CDF estimate. Because the labels, the

testing procedure, and the levels (ε, δ) are unchanged, the per-round guarantee is preserved, as we formalize in Section 3.3. Computationally, each round inserts the new confidence into the sorted pool ($\mathcal{O}(\log(n+t))$), re-scores the fixed n calibration points by binary search in that pool ($\mathcal{O}(n\log(n+t))$), and re-runs the sweep, which touches each of the $|G|$ grid points and n sorted scores once; since the grid is fixed, the per-round cost is $\mathcal{O}\big(n\log(n+t)\big)$, growing at most logarithmically in the deployment horizon and dominated by the judge's own forward pass.

Accept/abstain decision. For the arriving instance x_{n+t}, we accept the judge's prediction if its calibrated confidence $\widehat{F}_t(c(x_{n+t}))$ is at least λ_t^*, and abstain otherwise.

3.3 Convergence and Scope of Risk Control

We give two guarantees for the online procedure. First, the certified threshold does not merely converge to its population counterpart: beyond a finite, explicit number of unlabeled rounds it matches that counterpart exactly at each round with probability at least $1-\delta$, and almost surely locks onto it from some finite round onward (Theorem 3.2); CDF normalization further caps the coverage lost to grid discretization at essentially the grid resolution Δ (up to a vanishing $1/(n+t-1)$ term; Proposition 3.3). Second, the human-agreement guarantee is preserved at every round, regardless of the CDF estimate.

The convergence results concern a fixed threshold-selection rule, including its testing order and fallback convention. They establish agreement between the outputs obtained with an estimated CDF and with the population CDF. This conclusion is distinct from establishing that either output provides a valid risk certificate.

Assumption 3.1. *The calibration instances $\{x_1, \dots, x_n\}$ and stream instances $\{x_{n+1}, x_{n+2}, \dots\}$ are drawn i.i.d. from a common distribution, with confidence scores $c(x_i)$ having a continuous CDF F.*

Recall the sweep runs over the fixed grid G of spacing Δ (Section 3.2); write λ^{pop} for the threshold that fixed-sequence testing certifies on the calibration set under the population CDF F. The following theorem formalizes this first guarantee.

Theorem 3.2 (Finite-time convergence of certification)**.** *Under Assumption 3.1, let*

$$\rho \;=\; \min_{1\le i\le n}\; \min_{\lambda\in G}\; \big|F(c(x_i)) - \lambda\big|$$

be the smallest gap between a population calibration percentile $F(c(x_i))$ and a grid point. Then, for each fixed round

$$t \;\ge\; t^\star \;:=\; \left\lceil \frac{\log(2/\delta)}{2\rho^2} \right\rceil - n + 2,$$

the online certified threshold satisfies $\lambda_t^* = \lambda^{\mathrm{pop}}$ with probability at least $1-\delta$, taken over the joint draw of the calibration set and the stream (ρ and $t^\star$ are functions of the calibration draw). Moreover, since $\sup_u |\widehat{F}_t(u) - F(u)| \to 0$ almost surely (Glivenko–Cantelli), almost surely there is a finite round beyond which the certification stops changing and equals λ^{pop} at every subsequent round.

Proof. See Appendix B.1. □

Unlike a vanishing error bound, this is a finite-time exactness guarantee: at any fixed round $t \ge t^\star$ the certified threshold equals λ^{pop} with probability at least $1-\delta$, and almost surely locks onto it from some finite round onward. Here $t^\star$ scales as ρ^{-2} in the calibration-to-grid margin. This is a worst-case bound: ρ is a minimum over all n calibration percentiles, hence typically far smaller than the gap that governs the sweep near λ^{pop}, so the effective burn-in falls far short of $t^\star$ (see Section 4.2). An auxiliary fixed-round bound on certification mismatch probability is given in Appendix B.2.

The limit λ^{pop} is itself a grid quantity; how closely it matches the unconstrained (fine-grid) optimum depends on how finely the grid resolves the score distribution.

Proposition 3.3 (Grid-induced coverage loss under CDF normalization)**.** *Suppose fixed-sequence testing sweeps a grid of spacing Δ in score space, and let m_j denote the score mass in grid cell j (the fraction of instances whose score lands in the cell). The certified coverage can fall*

short of its unconstrained (fine-grid) optimum by at most $\max_j m_j$. *Under the empirical-CDF normalization* $s = \widehat{F}_t(c)$ *the scores are sample quantiles, so this loss is capped by the grid resolution:* $\max_j m_j \le \Delta + 1/(n+t-1)$.

Proof. See Appendix B.3. □

By contrast, raw scoring ($s = c$) carries no such cap: where confidences concentrate, a single width-Δ interval can hold a constant fraction of the mass, so its grid loss need not shrink with Δ and can force a suboptimal threshold. CDF normalization avoids this by spreading the scores near-uniformly, as Figure 1(b) illustrates.

Remark. At every round t, the threshold λ_t^* is certified by fixed-sequence testing on the fixed calibration set $\{(s_i^{(t)}, y_i)\}_{i=1}^n$. Under Assumption 3.1, the binomial UCB in (1) is an exact $(1-\delta)$ upper confidence bound on the conditional disagreement rate for any scoring map that does not use the calibration labels: conditioning on the confidences fixes the accepted set, leaving $Y(\lambda) \mid N(\lambda)$ binomial. Since $\widehat{F}_t$ is computed from confidence values only (calibration and stream), it meets this condition, so the guarantee $\mathbb{P}\big(\hat{y}(x) \neq y \mid \widehat{F}_t(c(x)) \ge \lambda_t^*\big) \le \varepsilon$ holds with probability at least $1-\delta$ at every round, regardless of the CDF estimation quality. The convergence result above therefore quantifies efficiency (how close to the population threshold), not validity (which holds unconditionally).

4 Experiments

Building on the online CDF update and certification analysis, we evaluate whether unlabeled stream data improve threshold tracking and deployment performance. In the experiments, we mainly ask four questions: **(Q1)** Does tracking error decrease and the threshold stabilize (Section 4.2)? **(Q2)** How do signed errors vary across calibration splits, and do thresholds move in a common direction (Section 4.3)? **(Q3)** Does online updating allow higher agreement targets at the same coverage (Section 4.4)? **(Q4)** How reliable is deployment, and how sensitive is tracking to the calibration budget and scoring function (Section 4.5)?

4.1 Experimental Setup

Benchmarks and judges. We evaluate six pairwise preference benchmarks with human labels: **TL;DR** (Stiennon et al., 2020), **Chatbot Arena** (Zheng et al., 2023), **HH-RLHF** (Bai et al., 2022), **AlpacaEval** (Dubois et al., 2023), **HelpSteer3** (Wang et al., 2026), and **Community Alignment** (Zhang et al., 2026). We test eleven open-weight judges (7B–120B), highlighting **MiMo-7B** (Xiaomi et al., 2025), **Gemma-3-27B** (Team et al., 2025), and **GPT-OSS-120B** (Agarwal et al., 2025); full results appear in Appendix E. The default score is top-1 probability under the Simulated Annotators protocol (Jung et al., 2025); Appendix D gives the elicitation details and score definitions.

Online protocol. Each default split comprises three disjoint subsets: n=200 labeled calibration instances, a randomly ordered stream of T=1,250 instances, and N_{ref}=1,550 reference instances. OCC follows the decide-then-update protocol of Section 3; deployment labels are used only for post-hoc evaluation, except by ACI-Label. Threshold and CDF diagnostics are measured after t completed updates, with $t = 0$ denoting initialization. Unless a target sweep is specified, $\varepsilon = 0.20$ and $\delta = 0.10$. Trajectory figures use 100 splits by default; aggregate tables and distribution figures use 1,000. Bands show ± 1 cross-split standard deviation unless otherwise stated; captions specify exceptions. Appendix C details the indexing and calibration-size variants, including the 3,800-instance pool for Figure 3(c).

Population reference. For each split, we transform the calibration confidences using the reference CDF $\widehat{F}^{\text{ref}}$ and rerun the same certification on the same calibration labels. The resulting threshold λ_r^{ref} approximates λ^{pop} for that split. The reference set is disjoint from calibration and deployment data, remains fixed within the split, and never enters the online pool. Threshold errors are measured against λ_r^{ref} without subtracting reference uncertainty (Appendix C).

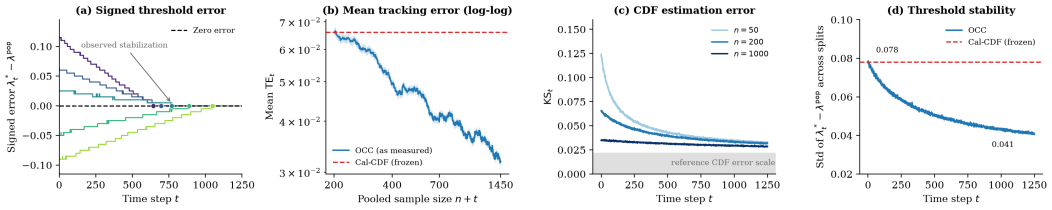


Figure 3: **Threshold tracking and stability under online CDF updates** (TL;DR, Gemma-3-27B; n=200 unless noted; (b,d): 1,000 splits; (c): 100). **(a)** Signed errors relative to each split's finite-reference threshold for five example draws; dots mark the last change observed by T=1,250. **(b)** Mean absolute tracking error versus pool size $n + t$ (log-log axes); band: ± 1 s.e. **(c)** Mean KS distance to the fixed reference CDF (N_{ref}=1,550) for $n \in \{50, 200, 1000\}$. Gray marks the approximate reference-CDF error scale (0.022), not an error floor. **(d)** Cross-split standard deviation of signed reference errors. Red dashed curves in (b,d): frozen Cal-CDF.

Compared methods. Certification-based methods share the calibration labels and testing rule. **Static** (Jung et al., 2025) freezes a threshold certified from raw confidence; **Cal-CDF** freezes the calibration-set CDF and its certified threshold, serving as the primary control for online updating. **W-ECDF** (Hu et al., 2026) re-certifies using a CDF of the latest 100 stream confidences, excluding calibration scores from its CDF pool; **OCC** re-certifies using calibration and all past stream confidences. We also include two privileged references. **ACI-Label** (Gibbs & Candes, 2021) adapts using deployment labels, without the selective-risk certificate considered here. **Oracle** certifies under the held-out reference CDF. Both use information unavailable to OCC. Implementation and tuning details are in Appendix C.

Metrics. **Threshold tracking error (TE)** is the mean absolute difference between the certified threshold and each split's reference threshold. **KS distance** compares the online CDF with the reference CDF. **Coverage** is the fraction of stream instances accepted over deployment. **Cumulative empirical selective risk (FCP)** measures disagreement among accepted instances. **Empirical GSR** is the fraction of runs with at least one acceptance whose empirical agreement meets $1 - \varepsilon$. GSR is a finite-evaluation statistic, distinct from the population-risk guarantee at nominal level $1 - \delta$. Appendix C defines the aggregation and the auxiliary grid-based monotonicity diagnostic.

4.2 Convergence to the Population Threshold (Q1)

Motivated by Theorem 3.2, we ask whether threshold trajectories stabilize and whether their distance from the reference decreases over $T = 1{,}250$ updates. Frozen Cal-CDF provides the control for online updating.

The threshold stabilizes and tracking error decreases. The five example trajectories approach their references from both sides and remain unchanged over the rest of the observed deployment (Figure 3(a)). OCC and Cal-CDF share the same initialization, but only OCC incorporates stream confidences. Its mean TE falls from 0.066 to 0.032 on TL;DR; across all six benchmarks, the endpoint reduction is 2.0–2.1-fold (Figure 3(b), Table 7). Mean KS distance also decreases, and the gaps between initial calibration sizes narrow as the pool grows (Figure 3(c)), supporting improved CDF estimation. We also check whether empirical agreement is nondecreasing as the percentile threshold increases. On the evaluated 20-point grid, the fraction of threshold pairs violating this ordering falls from $0.16-0.22$ at initialization to $0.03-0.06$ at the final checkpoint (Table 7). Section 4.3 examines the corresponding change in signed-error dispersion.

4.3 Signed-Error Dispersion and Net Threshold Movement (Q2)

We next examine how signed reference errors vary across calibration splits and whether thresholds move in a common direction. Here, net movement means the change from the initial to the final percentile threshold. Across 1,000 splits, the initial and final mean signed errors are both approximately zero (Figure 5(a)), while their standard deviation falls from 0.078 to 0.041 (Figure 3(d)). Final errors concentrate near zero for both signs of the initial error (Figure 5(b)). Of these 1,000 splits, 50.2%

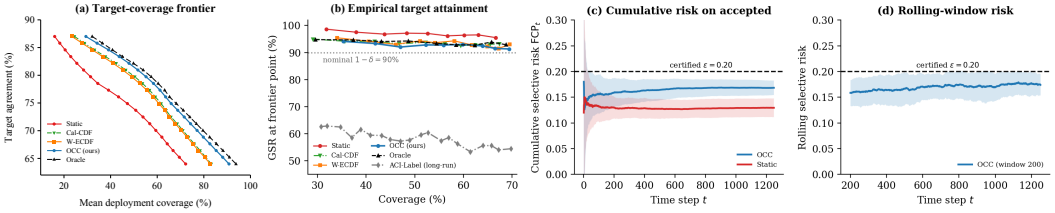


Figure 4: **Deployment frontiers and empirical reliability** (TL;DR, Gemma-3-27B, n=200; (a,b): 1,000 evaluated splits per target; (c,d): 100 splits). **(a)** Target agreement $1-\varepsilon$ versus mean deployment coverage; 20 risk targets $\varepsilon \in [0.130, 0.360]$ (exact readings: Table 4). **(b)** Empirical GSR for a separate 20-target sweep, $\varepsilon \in [0.05, 0.35]$, using nonempty runs only (undefined if none). Each point uses its own target; 90% is a nominal reference. ACI-Label uses label feedback. **(c,d)** Mean cumulative and rolling-window (200) selective risk at $\varepsilon = 0.20$; we set the bands with ± 1 cross-split standard deviation.

Table 1: **Target agreement (%) at 55% mean deployment coverage across three judges and six benchmarks** (n=200, δ=0.10, 1,000 splits). Within each dataset: St. = Static, W = W-ECDF, and **O** = OCC. Bold entries are the highest among these three methods. The TL;DR comparison including Cal-CDF and Oracle is shown in Figure 4(a), with exact readings in Table 4.

Judge	TL;DR			Arena			HH-RLHF			AlpacaE.			HelpSt.			Comm.A.		
	St.	W	**O**	St.	W	**O**	St.	W	**O**	St.	W	**O**	St.	W	**O**	St.	W	**O**
MiMo-7B	68.2	73.6	**76.0**	66.1	71.8	**74.0**	70.4	75.9	**77.6**	68.0	73.1	**75.7**	66.9	72.9	**74.8**	66.6	72.1	**74.5**
Gemma-3-27B	72.0	77.5	**79.8**	70.1	75.4	**78.2**	74.1	79.5	**81.2**	71.3	76.6	**79.5**	71.0	76.7	**78.5**	70.6	75.8	**78.4**
GPT-OSS-120B	75.4	81.0	**82.8**	73.3	78.8	**81.5**	77.4	82.1	**84.7**	74.7	80.5	**82.4**	74.4	79.5	**82.0**	73.9	79.6	**81.3**

show a net decrease, 2.4% no net change, and 47.4% a net increase (Figure 5(c)). These observations support reduced error dispersion across splits rather than a systematic shift of all thresholds in one direction. A lower percentile threshold need not imply a lower raw-confidence cutoff when the CDF also changes. Net movement alone therefore does not determine realized coverage. We assess the coverage–agreement trade-off directly in Section 4.4.

4.4 Deployment Frontiers at Matched Coverage (Q3)

Following the threshold-tracking analysis, we examine whether online CDF updates improve the trade-off between deployment coverage and the requested agreement target. Frozen Cal-CDF is the primary control for assessing the benefit of unlabeled stream updates. For the TL;DR comparison with Gemma-3-27B, we evaluate 20 equally spaced targets $\varepsilon \in [0.130, 0.360]$. At each target, we compute coverage over the full $T = 1{,}250$ deployment rounds and average it across all 1,000 splits, including zero-acceptance runs. Each method thus produces a curve relating mean coverage to target agreement $1 - \varepsilon$. We use piecewise-linear interpolation to compare target agreement at a common mean coverage. The separate empirical-reliability analyses complement this target–coverage comparison (Geifman & El-Yaniv, 2017; Gibbs & Candès, 2024). Figure 4(a) shows the complete TL;DR frontier, while Table 1 compares Static, W-ECDF, and OCC across three judges and six benchmarks at 55% mean coverage. We provide the details of exact TL;DR readings, paired intervals, and effective denominators in Appendix C.

Online updates improve the deployment frontier. On TL;DR with Gemma-3-27B, OCC improves on frozen Cal-CDF by 1.5–2.4 percentage points across the four reported coverages. At 55% coverage, the interpolated target increases from 78.0% to 79.8%. It also exceeds W-ECDF by 1.8–2.8 points and remains 0.4–0.9 points below the finite-reference Oracle; Tables 4 and 5 give the exact readings and pointwise 95% paired bootstrap intervals. At 55% mean coverage, Table 1 shows higher reported targets for OCC than for Static and W-ECDF in all 18 judge–benchmark combinations, extending the comparison beyond the representative TL;DR case.

Table 3: **Empirical GSR (%) across benchmarks and deployment checkpoints.** Mean over the three representative judges (n=200, ε=0.20, 1,000 evaluated splits per judge). GSR uses cumulative agreement among runs with at least one acceptance by the checkpoint; 90% is a nominal reference.

Checkpoint	TL;DR	Arena	HH-RLHF	AlpacaEval	HelpSteer3	Comm. Align.
t=250	92.8	91.9	93.4	92.2	92.9	91.5
t=500	92.3	91.4	93.7	91.8	93.1	90.9
t=750	92.6	91.7	93.0	92.1	92.5	91.3
t=1250	92.4	91.2	93.2	91.6	92.7	90.8

4.5 Empirical Reliability and Sensitivity (Q4)

Following the matched-coverage comparison, we examine empirical risk and target attainment during deployment, then vary the calibration budget and scoring function. These diagnostics complement the theory; they do not establish simultaneous population-risk control along the stream.

Empirical risk and target attainment. At $\varepsilon = 0.20$, the mean cumulative empirical risk of OCC approaches 0.17, and its mean rolling-window risk remains below the target over the displayed horizon (Figure 4(c, d)). Static has lower plotted mean risk (0.13), but lower risk alone does not establish a better deployment trade-off. We therefore read the risk curves alongside the coverage comparison in Section 4.4. Figure 4(b) reports empirical GSR across its separate target sweep. To assess reliability beyond TL;DR, Table 3 reports the default-target checkpoint GSRs across all six benchmarks, averaged over the three representative judges. Values range from 90.8% to 93.7%, exceeding the nominal 90% reference at every reported checkpoint.

Tracking improves across calibration budgets. We fix the final pool size at $n + T = 1{,}450$ and vary the initial calibration budget. The reported tracking time is the start of the first 50-round interval during which mean TE remains at most 0.035. Larger calibration sets reach this criterion earlier, while final mean TE varies little, remaining within $(3.1\text{–}3.3) \times 10^{-2}$ (Table 9). Appendix F gives the timing details and empirical budget fit. Calibration labels remain necessary for risk certification.

Tracking improves across scoring functions. Final TE is lower than initial TE for all four scores and all three judges, with initial-to-final ratios of 1.4–2.2 (Table 2). In pairwise judging, EMP margin is a strictly increasing transformation of top-1 probability. The two therefore yield identical CDF-normalized scores, explaining the matching rows and providing an invariance check. Vote agreement shows smaller relative gains. Its discrete support falls outside the continuous-CDF assumption; Appendix H discusses how to interpret this empirical comparison. Additional robustness analyses, including a contaminated stream and non-uniform arrival orders, are provided in Appendix J.

Table 2: **Tracking across scores.** TL;DR, n=200, 1,000 splits. Mean $\mathrm{TE}_0 \rightarrow \mathrm{TE}_{1250}$ ($\times 10^{-2}$).

Score	MiMo 7B	Gemma-3 27B	GPT-OSS 120B
Top-1	$7.7 \rightarrow 3.5$	$6.6 \rightarrow 3.2$	$5.6 \rightarrow 2.9$
EMP margin	$7.7 \rightarrow 3.5$	$6.6 \rightarrow 3.2$	$5.6 \rightarrow 2.9$
KL margin	$7.3 \rightarrow 3.3$	$6.3 \rightarrow 3.0$	$5.3 \rightarrow 2.8$
Vote	$7.0 \rightarrow 4.4$	$5.8 \rightarrow 3.9$	$5.0 \rightarrow 3.6$

5 Conclusion

In this work, we introduced Online Confidence Calibration to address threshold instability caused by estimating confidence distributions from small calibration sets. OCC incorporates unlabeled deployment scores into an evolving empirical CDF, re-scores the original calibration set, and re-certifies the acceptance threshold without requiring additional human annotations. Under i.i.d. sampling and a continuous confidence distribution, our analysis establishes that the selected threshold almost surely stabilizes at the threshold obtained using the population CDF on the same calibration set. Experiments across six benchmarks, eleven LLM judges, and four confidence functions demonstrate more accurate confidence-distribution estimates, reduced threshold variability, and higher agreement targets at matched coverage, with performance approaching the reference-CDF oracle. These results highlight the value of unlabeled deployment data for improving the coverage–agreement trade-off in selective LLM evaluation under a fixed annotation budget.

AI USE STATEMENT

We used generative AI tools to assist in debugging parts of the experiment code and to improve the language and typesetting of the manuscript. They did not generate any data or experimental results. AI-assisted code was reviewed and tested by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve the reliability of selective LLM evaluation through online confidence calibration. Our experiments use existing preference benchmarks and annotations. These resources and the judge models may reflect annotator, social, or cultural biases. Our theoretical results rely on the assumptions stated in the paper, and empirical agreement is measured against the reference annotations provided by the benchmarks. Responsible deployment should consider the representativeness of these preferences, potential distribution shifts, and appropriate human oversight.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

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

Sher Badshah, Ali Emami, and Hassan Sajjad. Scope: Selective conformal optimized pairwise llm judging. *arXiv preprint arXiv:2602.13110*, 2026.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. Cap: A general algorithm for online selective conformal prediction with fcr control. *arXiv preprint arXiv:2403.07728*, 2024.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Peter Bauer. Multiple testing in clinical trials. *Statistics in medicine*, 10(6):871–890, 1991.

Aadyot Bhatnagar, Huan Wang, Caiming Xiong, and Yu Bai. Improved online conformal prediction via strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 2337–2363. PMLR, 2023.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631, 2023.

Chao Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069, 2023.

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

Ran El-Yaniv et al. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5), 2010.

Shai Feldman, Liran Ringel, Stephen Bates, and Yaniv Romano. Achieving risk control in online learning settings. *arXiv preprint arXiv:2205.09095*, 2022.

Haosen Ge, Hamsa Bastani, and Osbert Bastani. Stochastic online conformal prediction with semi-bandit feedback. *arXiv preprint arXiv:2405.13268*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

Isaac Gibbs and Emmanuel J Candès. Conformal inference for online prediction with arbitrary distribution shifts. *Journal of Machine Learning Research*, 25(162):1–36, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Dongjian Hu, Junxi Wu, Shu-Tao Xia, and Changliang Zou. Distribution-informed online conformal prediction. In *International Conference on Learning Representations*, volume 2026, pp. 58230–58255, 2026.

Pierre Humbert, Ulysse Gazin, Ruth Heller, and Etienne Roquain. Online selective conformal inference: adaptive scores, convergence rate and optimality. *arXiv preprint arXiv:2508.10336*, 2025.

Gaojie Jin, Yong Tao, Lijia Yu, and Tianjin Huang. Margin-adaptive confidence ranking for reliable llm judgement. *arXiv preprint arXiv:2605.15416*, 2026.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, volume 2025, pp. 3101–3125, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Xinyu Li, Yi Zhou, Guanqun Cao, Zeyu Fu, Tianjin Huang, and Gaojie Jin. Localize-then-decide guarantees for llm judgments. *arXiv preprint arXiv:2608.25824*, 2026.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

Tuo Liu, Edgar Dobriban, and Francesco Orabona. Online conformal prediction via universal portfolio algorithms. *arXiv preprint arXiv:2602.03168*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Sangwoo Park, Matteo Zecchin, and Osvaldo Simeone. Adaptive prediction-powered autoeval with reliability and efficiency guarantees. *Advances in Neural Information Processing Systems*, 38: 158758–158784, 2026.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8352–8370, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

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

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *Advances in Neural Information Processing Systems*, 38, 2026.

LLM Xiaomi, Bingquan Xia, Bowen Shen, Dawei Zhu, Di Zhang, Gang Wang, Hailin Zhang, Huaqiu Liu, Jiebao Xiao, Jinhao Dong, et al. Mimo: Unlocking the reasoning potential of language model–from pretraining to posttraining. *arXiv preprint arXiv:2505.07608*, 2025.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1040–1051, 2021.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International conference on machine learning*, pp. 25834–25866. PMLR, 2022.

Lily Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, et al. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *International Conference on Learning Representations*, volume 2026, pp. 147076–147215, 2026.

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. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025.

A Related Work

We situate OCC within selective LLM evaluation, confidence estimation, and online conformal inference, emphasizing the available feedback and the guarantees studied.

LLM judges and selective evaluation. LLM evaluation uses benchmarks such as MT-Bench and Arena-Hard (Zheng et al., 2023; Li et al., 2024), with improvements from structured scoring and specialized judges (Liu et al., 2023; Kim et al., 2024; Zhu et al., 2025), task decomposition (Saha et al., 2024), and panels of diverse models (Verga et al., 2024). Bias studies document limitations of empirical judge quality (Wang et al., 2024; Panickssery et al., 2024; Xiong et al., 2024). Alongside these improvements, Jung et al. (2025) introduce selective evaluation with fixed-sequence testing on labeled calibration data, and extend it to cascades that escalate uncertain judgments to stronger models. SCOPE couples a bias-aware uncertainty signal with calibrated acceptance thresholds (Badshah et al., 2026); Localize-Then-Decide addresses multiple candidates through conformal shortlisting followed by selective choice (Li et al., 2026). R-AutoEval+ instead uses prediction-powered inference to improve the reliability and efficiency of aggregate estimates of model performance (Park et al., 2026), a different target from the disagreement risk of accepted individual judgments.

Confidence signals and percentile normalization. Selective evaluation relies on scalar confidence signals derived from token probabilities, verbalized confidence, self-evaluation, or agreement across generations (Kadavath et al., 2022; Lin et al., 2022; Tian et al., 2023; Wang et al., 2022; Kuhn et al., 2023; Jin et al., 2026). Probability-calibration methods such as temperature scaling aim to align predicted probabilities with observed correctness (Guo et al., 2017). Percentile normalization represents an existing score by its estimated CDF value. A common CDF transform is nondecreasing and does not reverse raw-confidence ordering, although it can introduce ties; its output is a percentile, not an estimated probability of correctness. This distinction separates estimation of the confidence distribution from calibration of the confidence–correctness relationship.

Selective prediction and online risk control. Selective prediction trades coverage for risk through abstention (Chow, 1970; El-Yaniv et al., 2010; Geifman & El-Yaniv, 2017; 2019), with extensions to language-model prediction and factuality (Xin et al., 2021; Yadkori et al., 2024; Mohri & Hashimoto, 2024). Conformal risk control handles monotone loss families, while Learn-then-Test selects risk-controlling rules by testing hypotheses (Angelopoulos et al., 2024; 2025). Weighted conformal methods address non-exchangeable data (Barber et al., 2023). For sequential prediction, ACI uses outcome feedback to target long-run coverage (Gibbs & Candes, 2021); expert aggregation, strongly adaptive learning, and parameter-free updates improve online adaptation (Zaffran et al., 2022; Gibbs & Candès, 2024; Bhatnagar et al., 2023; Liu et al., 2026), while online risk-control frameworks study broader losses (Feldman et al., 2022). Conformal Optimistic Prediction (COP) incorporates an estimated CDF of nonconformity scores into its update and observes deployment outcomes (Hu et al., 2026). CAP studies selection-conditional coverage and FCR control (Bao et al., 2024); OnlineSCI studies adaptive selective inference, including classification with abstention (Humbert et al., 2025). Related work also considers semi-bandit feedback (Ge et al., 2024).

Positioning of our work. OCC extends the selective-evaluation pipeline of Jung et al. (2025) by refining confidence percentiles with unlabeled deployment observations. The calibration labels and testing rule remain fixed, but each updated CDF triggers re-scoring and re-certification, so the selected threshold can change. W-ECDF provides a windowed confidence-distribution estimate under the same feedback restriction. We study the resulting threshold tracking, stability, and deployment frontiers, complementing work that changes the judge's confidence signal, localizes candidate responses, or adapts using newly observed labels.

B PROOFS

B.1 PROOF OF THEOREM 3.2

Proof. Fix a round $t \geq t^\star$. By the DKW inequality (Dvoretzky et al., 1956; Massart, 1990), with probability at least $1-\delta$,

$$\sup_u |\widehat{F}_t(u) - F(u)| \;\leq\; \eta_t := \sqrt{\tfrac{\log(2/\delta)}{2(n+t-1)}},$$

and $t \geq t^\star$ gives $n+t-1 > \log(2/\delta)/(2\rho^2)$, hence $\eta_t < \rho$ strictly. On this event, for every calibration instance i and grid point $\lambda \in G$,

$$\big|\widehat{F}_t(c(x_i)) - F(c(x_i))\big| \;\leq\; \eta_t \;<\; \rho \;\leq\; \big|F(c(x_i)) - \lambda\big|,$$

so $\widehat{F}_t(c(x_i))$ and $F(c(x_i))$ fall on the same side of every λ. Hence the accepted calibration subset is identical under $\widehat{F}_t$ and F at each grid point; since fixed-sequence testing (Equation (2)) depends on the scores only through these subsets and the labels $\{y_i\}$, it returns the same threshold, $\lambda_t^* = \lambda^{\mathrm{pop}}$. For the almost-sure claim, Glivenko–Cantelli gives $\sup_u |\widehat{F}_t(u) - F(u)| \to 0$ a.s., hence a.s. $\sup_u |\widehat{F}_t(u) - F(u)| < \rho$ for all large t, and the same argument then yields $\lambda_t^* = \lambda^{\mathrm{pop}}$ from that round onward. □

B.2 AN AUXILIARY CERTIFICATION DEVIATION BOUND

The following result complements Theorem 3.2 by bounding certification mismatch probability at a fixed round. Here t follows the decision-time indexing of Section 3, so the CDF contains $n+t-1$ samples.

Corollary B.1 (Fixed-round certification deviation bound)**.** *Under Assumption 3.1, for each fixed deterministic round t,*

$$\mathbb{P}\big(\lambda_t^* \neq \lambda^{\mathrm{pop}}\big) \leq \min\{1,\; \delta + 2n|G|\eta_t\}\,, \qquad \eta_t = \sqrt{\frac{\log(2/\delta)}{2(n+t-1)}}.$$

Proof. Fix a deterministic round t and work on the DKW event $\mathcal{E}_t = \{\sup_u |\widehat{F}_t(u) - F(u)| \leq \eta_t\}$, which has probability at least $1-\delta$. As in the proof of Theorem 3.2, if every calibration percentile $F(c(x_i))$ lies farther than η_t from every grid point, then $\widehat{F}_t(c(x_i))$ and $F(c(x_i))$ fall on the same side of each $\lambda \in G$, the accepted subsets coincide at every grid point, and the sweep certifies the same threshold. On $\mathcal{E}_t$ a deviation therefore requires some percentile to lie within η_t of some grid point:

$$\{\lambda_t^* \neq \lambda^{\mathrm{pop}}\} \cap \mathcal{E}_t \;\subseteq\; \bigcup_{i=1}^{n} \Big\{ \min_{\lambda \in G} \big|F(c(x_i)) - \lambda\big| \leq \eta_t \Big\}.$$

Since F is continuous, the probability integral transform makes each $F(c(x_i))$ uniform on $[0,1]$. The union of the $|G|$ intervals of half-width η_t centred at the grid points has Lebesgue measure at most $2|G|\eta_t$, so each event above has probability at most $2|G|\eta_t$. A union bound over the n percentiles, together with $\mathbb{P}(\mathcal{E}_t^c) \leq \delta$, gives $\mathbb{P}(\lambda_t^* \neq \lambda^{\mathrm{pop}}) \leq \delta + 2n|G|\eta_t$. Taking the minimum with the trivial upper bound 1 gives the stated result. □

For fixed n, G, and δ, the second term decreases at the CDF estimation rate, but the bound tends to δ and can be vacuous at the experimental sample sizes. This is a coarse bound on the probability of a certification mismatch; it does not identify an asymptotic exponent for the measured mean absolute error against a finite reference in Figure 3(b).

B.3 PROOF OF PROPOSITION 3.3

Proof. Since s is monotone in c, each grid point accepts a top-by-confidence subset, and moving from one grid point to the next changes the coverage by the intervening cell mass m_j. The grid thus realizes only coverages on this discrete set, and rounding the optimum to the nearest realizable value costs at most $\max_j m_j$. Under $s = \widehat{F}_t(c)$ the scores are sample quantiles, so a width-Δ interval holds at most $\Delta(n+t-1)+1$ of them, giving $\max_j m_j \leq \Delta + 1/(n+t-1)$. □

C Implementation Details of Compared Methods

This appendix makes the procedure in Section 3 explicit in Algorithm 1, then specifies the diagnostics and baseline implementations underlying Section 4.1. The candidate-path rule $\mathcal{A}$ uses calibration scores, not calibration errors; the deployment-frontier sweep below gives its experimental instantiation. The empirical CDF counts repeated confidence values with their multiplicities.

Algorithm 1 OCC: decide, update the CDF, and re-certify on fixed calibration labels

Require: Calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n}$; fixed judge $\hat{y}$ and confidence c;
risk target $\varepsilon \in (0,1)$, $\delta \in (0,1)$, grid G, and score-only path rule $\mathcal{A}$;
unlabeled stream $x_{n+1}, \ldots, x_{n+T}$.

Cache $c_i \leftarrow c(x_i)$ and $\ell_i \leftarrow \mathbf{1}\{\hat{y}(x_i) \neq y_i\}$ for $i = 1, \ldots, n$.
$\mathcal{P} \leftarrow (c_1, \ldots, c_n)$; $\widehat{F}_1 \leftarrow \text{ECDF}(\mathcal{P})$.
$\lambda_1^* \leftarrow \textsc{Certify}(\widehat{F}_1)$.
for $t = 1, \ldots, T$ **do**
  Observe $q_t \leftarrow c(x_{n+t})$.
  if $\lambda_t^* \neq \perp$ **and** $\widehat{F}_t(q_t) \geq \lambda_t^*$ **then**
    Accept the judge's prediction $\hat{y}(x_{n+t})$.
  else
    Abstain on x_{n+t}.
  end if
  Append q_t to $\mathcal{P}$, including when the decision is to abstain.
  $\widehat{F}_{t+1} \leftarrow \text{ECDF}(\mathcal{P})$.
  $\lambda_{t+1}^* \leftarrow \textsc{Certify}(\widehat{F}_{t+1})$.
end for

function $\textsc{Certify}(H)$ ▷ Uses the cached n calibration errors only.
  $s_i \leftarrow H(c_i)$ for $i = 1, \ldots, n$; $\Lambda \leftarrow \mathcal{A}((s_i)_{i=1}^{n}; G) \subseteq G$.
  $\lambda_{\text{best}} \leftarrow \perp$.
  for $\lambda \in \Lambda$ in descending order **do**
    $N \leftarrow \sum_{i=1}^{n} \mathbf{1}\{s_i \geq \lambda\}$; $Y \leftarrow \sum_{i=1}^{n} \ell_i \mathbf{1}\{s_i \geq \lambda\}$.
    $U \leftarrow$ the binomial upper bound in Equation (1) using (N, Y); set $U = 1$ if $N = 0$.
    if $U > \varepsilon$ **then**
      return λ_{best}. ▷ Stop at the first failure.
    end if
    $\lambda_{\text{best}} \leftarrow \lambda$.
  end for
  return λ_{best}.
end function

Here $\perp$ denotes no passing candidate, not a numerical threshold. The state used to decide round t contains exactly $n + t - 1$ confidences; its update uses no stream labels or reference samples. An abstention does not terminate deployment.

Update indexing and repeated splits. Default experiments use approximately 3,000 instances per benchmark, randomly partitioned into calibration, deployment, and reference subsets. For threshold and CDF diagnostics, t counts completed stream updates: the experimental state at t corresponds to $(\widehat{F}_{t+1}, \lambda_{t+1}^*)$ in Section 3, with $n + t$ pooled confidences. Initialization is $t = 0$. Trajectory figures use 100 splits unless stated otherwise; aggregate tables and distribution figures use 1,000. In particular, Figure 3(b,d) uses 1,000 splits, and the band in (b) is a pointwise standard error. Other trajectory bands show ± 1 cross-split standard deviation by default.

Finite-reference construction and uncertainty. Following online conformal prediction (Gibbs & Candès, 2024; Humbert et al., 2025), we report convergence against an oracle reference: λ^{pop} runs the identical certification of Section 3.2 on $\mathcal{D}_{\text{cal}}$, but with scores transformed by the population CDF F; being conditional on the calibration draw, it is recomputed for every split, and all threshold

comparisons below are made within a split. We approximate F by the reference empirical CDF $\widehat{F}^{\mathrm{ref}}$, computed from a set disjoint from both the calibration set and the stream and held fixed throughout each split. Reference samples are never added to the online pool. The finite reference introduces sampling uncertainty. Its CDF error is of order $N_{\mathrm{ref}}^{-1/2}$ (approximately $0.87/\sqrt{N_{\mathrm{ref}}} \approx 0.022$ as a descriptive scale for $N_{\mathrm{ref}} = 1{,}550$), not a deterministic lower bound on the distance between two empirical CDFs. This scale does not directly quantify threshold error because certification is a discontinuous functional of the CDF. We report TE as measured without quadrature correction: reference error can increase or decrease an absolute deviation, so it does not provide a one-sided bound on the underlying tracking error. By construction λ^{pop} uses the same calibration labels and certification rule; its experimental approximation additionally depends on the finite reference set.

Aggregation of threshold tracking errors. For split r and update count t, let $e_{r,t} = |\lambda^*_{r,t} - \lambda^{\mathrm{ref}}_r|$, where λ^{ref}_r is the finite-reference approximation to that split's population threshold and is held fixed throughout deployment. Figure 3(b) reports $\bar{e}_t = R^{-1}\sum_{r=1}^{R} e_{r,t}$ for $R = 1{,}000$, with pointwise standard error $s_t/\sqrt{R}$, where s_t is the sample standard deviation of the absolute errors $e_{r,t}$ across splits. Figure 3(d) reports $\mathrm{SD}_r(d_{r,t})$ at each t, where $d_{r,t} = \lambda^*_{r,t} - \lambda^{\mathrm{ref}}_r$, over the same R splits. This measures signed-error dispersion, not the standard error in (b) or the dispersion of absolute thresholds, since reference thresholds can differ across splits. Cal-CDF retains its initial threshold in each split, so both its mean absolute error and signed-error dispersion remain constant in t.

CDF and grid diagnostics. For split r, the KS distance is $\mathrm{KS}_{r,t} = \sup_u |\widehat{F}_{r,t}(u) - \widehat{F}^{\mathrm{ref}}_r(u)|$; reported curves average these distances across splits. The auxiliary grid-based monotonicity violation rate $\mathcal{V}_{\mathrm{grid},t}$ is the fraction of threshold pairs on the evaluated 20-point percentile grid that violate empirical agreement monotonicity on the reference set. It describes the evaluated grid; CDF normalization does not reverse the ordering of raw confidence scores.

CDF comparison across initial calibration sizes. For Figure 3(c), each run selects three disjoint subsets from 3,800 TL;DR instances: 1,550 reference instances, 1,250 stream instances, and $n \in \{50, 200, 1000\}$ calibration instances. Any remaining instances are unused in that run. The reference CDF is fixed throughout deployment, and its sample size is the same for every n. Each curve averages the per-split KS distances over 100 splits. This comparison holds T fixed, whereas Table 9 holds $n + T = 1{,}450$ fixed.

Deployment-frontier sweep. For the TL;DR comparison in Section 4.4, all methods use the same $n = 200$ calibration labels and $T = 1{,}250$ stream instances in each of $R = 1{,}000$ splits. Let $s^{(m,t)}_{r,(30)}$ be the 30th-largest calibration score in method m's score coordinate. The start is

$$\lambda^{(m)}_{\mathrm{start},r,t} = \Delta \left\lfloor s^{(m,t)}_{r,(30)}/\Delta \right\rfloor, \qquad \Delta = 0.005.$$

We evaluate $\Lambda = G \cap [0, \lambda_{\mathrm{start}}]$ in descending order, stop at the first failure of Equation (1), and retain the last passing candidate. If no candidate passes, the round explicitly abstains. This start ensures at least 30 accepted calibration instances; discretization and ties may increase the count. It uses scores rather than calibration errors. Static, Cal-CDF, and Oracle freeze their certified thresholds; OCC and W-ECDF re-certify as their CDFs change.

Deployment-frontier aggregation. Let $A_{m,r,t}(\varepsilon)$ indicate actual acceptance at risk target ε. The mean coverage is

$$\bar{C}_m(\varepsilon) = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{T}\sum_{t=1}^{T} A_{m,r,t}(\varepsilon)\right).$$

We sweep 20 equally spaced ε values in $[0.130, 0.360]$ and interpolate $1 - \varepsilon$ as a piecewise-linear function of $\bar{C}_m(\varepsilon)$. All four coverages in Table 4 lie within the common observed range, 32.1%–72.2%; we do not extrapolate. Zero-acceptance runs contribute zero coverage. Empirical-risk and GSR averages exclude them, so their effective denominator is $N_{\mathrm{valid},m}(\varepsilon) = \#\{r : \sum_t A_{m,r,t}(\varepsilon) > 0\}$; these conditional statistics are undefined when the count is zero. Table 6 reports counts and conditional coverage at selected targets. The same definition applies to the separate GSR sweeps in Figure 4(b) and Figure 6: a target with no nonempty runs has zero coverage and undefined GSR.

The detailed TL;DR values in Table 4 include the frozen Cal-CDF and finite-reference Oracle comparisons underlying Figure 4(a).

Table 4: **Matched-coverage target agreement (%) for the TL;DR frontier in Figure 4(a)** (Gemma-3-27B, n=200, T=1,250, 1,000 splits). The 20 targets span $\varepsilon \in [0.130, 0.360]$. Values are descriptive interpolations, not additional risk certificates. Table 5 reports paired differences and intervals.

Method	Stream use	Interpolated target agreement (%)			
		35%	45%	55%	65%
Static (Jung et al., 2025)	none (raw)	78.4	75.6	72.0	67.3
Cal-CDF	none (cal CDF)	84.1	81.4	78.0	73.0
W-ECDF (Hu et al., 2026)	windowed scores	83.8	81.0	77.5	72.6
OCC (ours)	pooled scores	**85.6**	**83.0**	**79.8**	**75.4**
ACI-Label (Gibbs & Candes, 2021)	stream labels[†]	no selective-risk certificate			
Oracle	reference CDF[†]	86.2	83.4	80.6	76.3

[†] Privileged information unavailable in our setting; shown as reference, not as a competitor.

Paired uncertainty. We use 2,000 bootstrap resamples of split IDs, retaining each split's complete target vector and sharing the resampling indices across methods. Each resample recomputes mean coverage, interpolation, and paired differences. Table 5 reports pointwise 95% percentile intervals for variation across random splits of the fixed benchmark. Comparisons remain within support in all resamples. Differences are computed before rounding and can therefore differ by 0.1 point from differences of the rounded values in Table 4.

Table 5: **Paired differences in interpolated target agreement** (percentage points; TL;DR, Gemma-3-27B). Entries are point estimates with pointwise 95% percentile intervals from 2,000 paired bootstrap resamples of 1,000 splits.

Coverage	OCC − Cal-CDF	Oracle − OCC
35%	1.502 [1.446, 1.567]	0.610 [0.512, 0.703]
45%	1.603 [1.503, 1.702]	0.403 [0.288, 0.519]
55%	1.804 [1.661, 1.943]	0.785 [0.639, 0.944]
65%	2.407 [2.213, 2.609]	0.879 [0.694, 1.081]

Table 6: **Nonempty deployment runs and conditional coverage** (TL;DR, Gemma-3-27B, R = 1,000, T = 1,250). Each entry is the nonempty-run count, followed by mean coverage among those runs. Column targets are rounded for display.

Method	$\varepsilon = 0.1300$	0.1784	0.2147	0.2511
Static	884 (18.1%)	970 (26.3%)	993 (34.7%)	1000 (47.2%)
Cal-CDF	907 (27.0%)	979 (43.3%)	996 (53.9%)	1000 (61.7%)
W-ECDF	926 (25.3%)	991 (41.5%)	999 (52.5%)	1000 (60.7%)
OCC	908 (32.4%)	980 (48.9%)	996 (58.6%)	1000 (66.0%)
Oracle	897 (35.8%)	974 (51.1%)	996 (60.6%)	1000 (68.2%)

Window timing of W-ECDF. W-ECDF is a window-CDF baseline inspired by Hu et al. (2026) and uses our fixed-label certification. At round t, W-ECDF uses the $m_t = \min(100, t-1)$ most recent stream confidences, deciding before the current confidence is added. It abstains at $t = 1$, when no window CDF is available; the full window is available from $t = 101$. Unlabeled confidences are retained after abstention.

Adaptation of ACI-Label to abstention rounds. Following the convention of online selective inference (Humbert et al., 2025), ACI-Label updates its threshold only on rounds where the instance is accepted, using the realized error indicator of that round; abstained rounds leave the threshold unchanged.

Hyperparameter grids. ACI-Label uses $\gamma_t = c\,t^{-3/4}$. Its step-size coefficient is grid-searched over $c \in \{0.5, 1, 2, 5\}$ per benchmark, selected post hoc to minimize the end-of-stream tracking error; this oracle tuning favors the baseline and is consistent with its role as a privileged reference. The fixed-sequence threshold grid has spacing 0.005 in the method's score coordinate: raw confidence for Static and CDF percentiles for Cal-CDF, W-ECDF, OCC, and Oracle.

D Simulated Annotators Protocol and Scoring Function Definitions

This appendix details the confidence-elicitation protocol used throughout the paper and formally defines the four scoring functions compared in Section 4.5.

Simulated Annotators protocol. We follow the Simulated Annotators framework of Jung et al. (2025). For each instance x with $m{=}2$ candidates, $N{=}15$ simulated annotators independently evaluate the instance. Each annotator receives a distinct prompt consisting of (i) a task instruction describing the evaluation criterion, (ii) $K{=}15$ few-shot demonstrations randomly sampled from the calibration pool (excluding the current instance), each containing a query, its candidate responses, and the human-preferred label, and (iii) the target query with its two candidate responses. The judge LLM is queried with greedy decoding (temperature 0), and the top-token log-probabilities over the candidate labels ("A", "B") are extracted from the first generated token, exponentiated, and normalized into a probability vector $\mathbf{p}_j = (p_{j,1}, p_{j,2})$ for annotator j. Varying the few-shot subset across annotators introduces diversity analogous to inter-annotator variation in human labeling.

Scoring functions. Let $\bar{p}_i = \frac{1}{N}\sum_{j=1}^{N} p_{j,i}$ denote the ensemble mean probability (EMP) of candidate g_i, and let $\hat{y} = \arg\max_i \bar{p}_i$ be the judge's prediction. The four scalar confidences $c(x)$ used in Table 2 are: **Top-1 probability**, the EMP of the predicted winner, $c(x) = \max_i \bar{p}_i$; **EMP margin**, the gap between the EMP scores, $c(x) = \bar{p}_{\hat{y}} - \max_{i \neq \hat{y}} \bar{p}_i$ (for $m{=}2$ this equals $2\max_i \bar{p}_i - 1$, a strictly monotone transform of the top-1 probability, so both yield identical normalized scores under CDF normalization); **KL margin**, the same gap computed on the KL-divergence-weighted voting score of Jung et al. (2025), which scores each candidate by aggregating only the annotators who rank that candidate first, weighting annotator j's contribution to candidate g_i by $p_{j,i}$ (derived from $\mathrm{KL}(e_i \,\|\, \mathbf{p}_j) = -\log p_{j,i}$ via exponential transform) and normalizing by voter consensus; **Vote agreement**, the fraction of annotators voting for the majority choice, $c(x) = \max_i \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}[\arg\max_l p_{j,l} = i]$, supported on the discrete grid $\{\lceil N/2 \rceil / N, \ldots, 1\}$. Whichever scalar is chosen, the online procedure is unchanged: $c(x)$ is passed through the pooled CDF $\widehat{F}_t$ and the certification of Section 3.2 runs on the normalized scores.

E Full Results Across Judges and Benchmarks

This appendix reports convergence results for all eleven judge LLMs spanning seven families across the six benchmarks, complementing the main-text deployment-frontier comparison in Table 1. We first give checkpoint metrics for the representative judge across all six benchmarks (Table 7), then initialization and end-of-deployment tracking errors for all eleven judges (Table 8). TE and KS are reported against the finite reference, without subtracting a presumed threshold-error floor. Their residual values do not identify a separate population-tracking component.

Table 7: **OCC threshold tracking, CDF estimation, and grid-based empirical monotonicity during deployment** (Gemma-3-27B, n=200, mean over 1,000 splits; t=0 denotes initialization). TE and KS use the split-specific finite reference. $\mathcal{V}_{\mathrm{grid},t}$ is the empirical monotonicity violation rate on the evaluated 20-point percentile grid (Appendix C).

Dataset	TE_t ($\times 10^{-2}$, ↓)				KS_t ($\times 10^{-2}$, ↓)				$\mathcal{V}_{\mathrm{grid},t}$ (↓)	
	t=0	t=250	t=500	t=1250	t=0	t=250	t=500	t=1250	t=0	t=1250
TL;DR	6.6	4.7	4.0	3.2	6.4	4.8	3.9	3.2	0.18	0.04
Chatbot Arena	7.1	5.2	4.3	3.4	6.6	4.6	4.0	3.1	0.21	0.05
HH-RLHF	6.0	4.1	3.6	3.0	6.5	4.7	3.9	3.2	0.16	0.03
AlpacaEval	6.9	5.0	4.2	3.3	6.7	4.7	4.0	3.2	0.19	0.05
HelpSteer3	6.4	4.4	3.8	3.1	6.4	4.6	4.0	3.1	0.17	0.04
Comm. Align.	7.4	5.4	4.5	3.5	6.6	4.8	4.1	3.2	0.22	0.06

Table 8: **Initial and end-of-deployment threshold tracking error of OCC** ($\times 10^{-2}$, ↓) across all eleven judges and six benchmarks (n=200, 1,000 splits). The table reports both initialization and end-of-deployment errors under the same finite-reference convention.

Judge	TL;DR		Arena		HH-RLHF		AlpacaE.		HelpSt.		Comm.A.	
	t=0	t=T	t=0	t=T	t=0	t=T	t=0	t=T	t=0	t=T	t=0	t=T
MiMo-7B	7.7	3.5	8.2	3.7	7.1	3.3	8.0	3.6	7.3	3.4	8.6	3.7
Qwen3-8B	7.4	3.4	8.2	3.6	6.7	3.2	7.5	3.4	7.2	3.3	8.1	3.6
Llama-3-8B	7.5	3.4	8.0	3.5	6.9	3.3	7.8	3.6	7.1	3.4	8.4	3.7
Ministral-14B	7.1	3.3	7.8	3.5	6.4	3.1	7.2	3.4	6.9	3.2	7.8	3.4
DeepSeek-16B	7.0	3.3	7.5	3.4	6.5	3.1	7.3	3.3	6.6	3.2	7.9	3.5
Gemma-3-27B	6.6	3.2	7.1	3.4	6.0	3.0	6.9	3.3	6.4	3.1	7.4	3.5
Qwen2.5-32B	6.4	3.1	6.8	3.2	5.9	3.0	6.6	3.2	6.1	3.0	7.3	3.3
DeepSeek-67B	6.1	3.0	7.0	3.2	5.5	2.9	6.3	3.1	6.0	3.0	6.8	3.2
Llama-3-70B	5.9	3.0	6.5	3.1	5.4	2.8	6.3	3.0	5.6	2.9	6.8	3.2
Qwen2.5-72B	5.8	2.9	6.2	3.0	5.4	2.8	6.0	3.0	5.8	2.9	6.5	3.1
GPT-OSS-120B	5.6	2.9	6.1	3.0	5.1	2.8	5.9	2.9	5.3	2.8	6.4	3.1

F Calibration Budget and Observed Tracking Times

This appendix examines the calibration-budget comparison in Section 4.5. We vary n while fixing the final pool size $n + T = 1{,}450$ and reference size $N_{\mathrm{ref}} = 1{,}550$. Table 9 reports mean tracking errors and two complementary timing summaries.

Timing definitions. Let $e_{r,t} = |\lambda^*_{r,t} - \lambda^{\mathrm{ref}}_r|$ and $\bar{e}_t = R^{-1}\sum_{r=1}^{R} e_{r,t}$, with $R = 1{,}000$. This is the mean-TE convention of Figure 3(b), applied separately at each budget. Define

$$t_{0.035} = \min\left\{t \in \{0, \ldots, T-49\} : \bar{e}_{t+j} \le 0.035 \text{ for all } j \in \{0, \ldots, 49\}\right\}.$$

All five budgets attain this condition within deployment. The statistic is read from the mean curve, without selecting or excluding individual splits by their attainment times. For each split, the last-change round is the largest $t \in \{1, \ldots, T\}$ with $\lambda^*_{r,t} \neq \lambda^*_{r,t-1}$, or 0 if no change occurs. A change at T is recorded as T. The table reports the median of these R observed last-change rounds as t_{lock}; it does not infer stability beyond T.

Observed budget dependence. Across the five budgets, $n + t_{0.035}$ ranges from 1,034 to 1,051. A least-squares fit with slope fixed at -1 gives $K = 1{,}043.4$ in $t_{0.035} \approx K - n$; the table rounds this intercept to $K = 1{,}040$ for an illustrative fit. This is an empirical summary. Theorem 3.2 does not predict a common intercept or slope across budgets, since its margin ρ changes with the calibration draw and n. The error target 0.035 is distinct from this margin, which satisfies $\rho \le \Delta/2 = 0.0025$ on the grid used here. At the fixed final pool size, mean TE lies between 3.1 and 3.3×10^{-2}. Unlabeled confidences help estimate the confidence distribution, while changing n also changes the labeled sample available for certification.

Table 9: **Calibration budget and observed tracking times** (TL;DR, Gemma-3-27B, 1,000 splits; $n + T = 1{,}450$, $N_{\text{ref}} = 1{,}550$). TE columns report the split mean $\bar{e}_t$; $t = 0$ uses Cal-CDF initialization. $t_{0.035}$ starts the first 50-round window with $\bar{e}_t \leq 0.035$, while t_{lock} is the median last-change round within deployment. The illustrative-fit column uses the rounded intercept $K = 1{,}040$; timing definitions and boundary conventions are given above.

n	T	Mean TE ($\times 10^{-2}$, ↓)		Observed timing and fit		
		t=0 (Cal-CDF)	t=T	rounded fit ($K-n$)	$t_{0.035}$	median t_{lock}
50	1400	11.8	3.3	990	984	1027
100	1350	8.9	3.2	940	947	981
200	1250	6.6	3.2	840	838	882
500	950	4.6	3.2	540	551	587
1,000	450	3.6	3.1	40	47	84

G Initial-to-Reference Deviation and Net Threshold Movement

This appendix expands the direction analysis in Section 4.3. For each split, we compare the initial Cal-CDF certification, the final online certification, and the reference certification obtained from the same calibration labels under the reference CDF. Holding the reference fixed within a split makes the end-to-end change in signed error equal to the net change in the certified percentile threshold. Define $\Delta\lambda_r = \lambda^*_{r,T} - \lambda^*_{r,0}$: negative, zero, and positive values indicate a net decrease, no net change, and a net increase, respectively. Zero net change means equal endpoint grid thresholds; intermediate thresholds may still vary. The three categories contain 502, 24, and 474 splits, respectively. Each percentage uses all $R = 1{,}000$ splits as its denominator, giving 50.2%, 2.4%, and 47.4%. Splits with no net change are retained as a separate category.

At ξ=0.01, the mean fraction of rounds with a threshold more than 0.01 below its split-specific reference is 31% (Table 13). This signed-deviation diagnostic is distinct from the endpoint movement shares above; risk is assessed separately in Section 4.5.

Error dispersion and direction. Figure 5(a,b) shows that final signed errors concentrate near zero for both positive and negative initial errors. Their standard deviation decreases from 0.078 to 0.041, and the initial and final mean signed errors are both approximately zero. Panel (c) summarizes the direction of net percentile-threshold movement; it does not measure a change in realized coverage. Together, these diagnostics describe a reduction in cross-split error dispersion without a substantial mean shift.

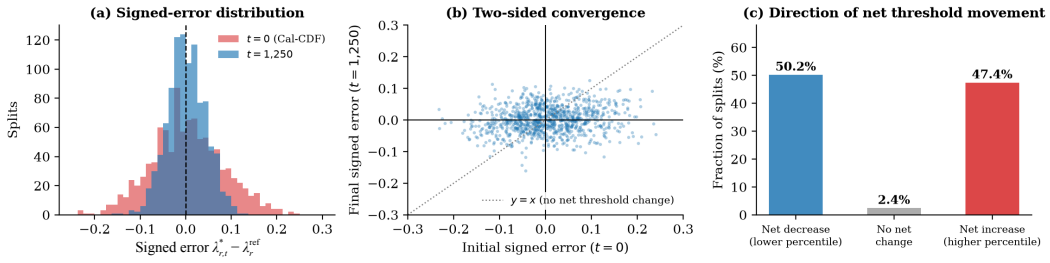


Figure 5: **Online updating reduces the dispersion of signed threshold errors** (TL;DR, Gemma-3-27B, n=200, 1,000 splits; the direction analysis of Section 4.3). **(a)** Distribution of the signed error $\lambda^*_{r,t} - \lambda^{\text{ref}}_r$ at t=0 (red: Cal-CDF initialization, mean ≈ 0, std 0.078) and t=1,250 (blue: concentrated, mean ≈ 0, std 0.041); the final distribution is more concentrated around zero. **(b)** Final signed error against initial signed error per split: final errors concentrate near zero for both signs of the initial error. The diagonal marks equal initial and final thresholds (no net change), since each split's reference is fixed; intermediate thresholds may still vary. **(c)** Net changes in percentile threshold from initialization to t=1,250: 502 net decreases (50.2%), 24 with no net change (2.4%), and 474 net increases (47.4%). All percentages use the full 1,000 splits. These endpoint changes do not directly specify changes in coverage.

H SCORING-FUNCTION SENSITIVITY

The main-text scoring comparison in Table 2 uses the four confidence functions defined in Appendix D. This appendix explains the matching top-1 and EMP-margin results and the interpretation of the discrete-score comparison. For pairwise comparison, EMP margin is a strictly monotone transform of top-1 probability, so their identical CDF-normalized results provide an invariance check. Vote agreement has discrete support and falls outside the continuous-CDF assumption of Theorem 3.1; this comparison tests empirical applicability beyond that assumption.

I PERFORMANCE UNDER DISTRIBUTION SHIFT

Section 3.3 assumes i.i.d. calibration and stream data (Theorem 3.1). This appendix evaluates a complementary empirical stress test: how do CDF updates affect threshold tracking, deployment coverage, and target attainment when the stream distribution drifts? The i.i.d. guarantees do not by themselves establish risk control for the shifted stream.

Gradual covariate-shift protocol. The calibration set is drawn from the base distribution, while the stream undergoes a slow covariate shift: the confidence density is translated by $\beta\,(t/T)$ at round t, for a total location shift of $\beta = 0.15$ over $T = 1{,}250$ rounds (e.g. progressively harder queries that lower the judge's confidence). The confidence-to-agreement map $\Pr(\hat{y} = y \mid c)$ is held fixed, so only the confidence density drifts and the population threshold moves solely through that density. We compare two frozen and two updating methods: **Static** (threshold frozen at t=0); **Cal-CDF** (percentiles computed once from $\mathcal{D}_{\text{cal}}$ and never updated by the stream); **W-ECDF** (percentiles from a sliding window of the last w=100 stream confidences); and **OCC** (calibration and all past stream confidences pooled). All methods use the same fixed calibration labels for their binomial certification step. Static and Cal-CDF perform it once; W-ECDF and OCC repeat it as their CDF estimates change.

Empirical risk and target attainment under shift. Table 10 reports end-of-stream behavior at the requested risk target $\varepsilon = 0.20$. The reported empirical selective risks are 0.27 for Static and 0.24 for Cal-CDF, with empirical GSRs of 72.0% and 78.3%. W-ECDF and OCC both have reported risk 0.19, with GSRs of 90.4% and 91.2%, respectively. W-ECDF has lower tracking error than OCC (4.1 versus 5.6, in units of 10^{-2}), whereas OCC has higher coverage (0.46 versus 0.39). These observations are consistent with a trade-off between tracking the recent distribution and retaining a larger confidence pool; they compare empirical outcomes at the same requested target, not matched statistical guarantees.

Table 10: **End-of-stream behavior under a slow covariate shift** (TL;DR, Gemma-3-27B, n=200, T=1,250, total drift β=0.15, 1,000 splits; target ε=0.20, nominal $1-\delta$=90%; TE_{1250} is measured against the current, drifted population threshold).

Method	Adaptive	Coverage	Realized risk	GSR (%)	TE_{1250} ($\times 10^{-2}$, $\downarrow$)
Static	no	0.58	0.27 ($> \varepsilon$)	72.0	12.4
Cal-CDF	no	0.51	0.24 ($> \varepsilon$)	78.3	9.8
W-ECDF	yes	0.39	0.19 ($\leq \varepsilon$)	90.4	4.1
OCC (ours)	yes	0.46	0.19 ($\leq \varepsilon$)	91.2	5.6

Empirical GSR across coverage levels. Figure 6 traces coverage and empirical GSR as the requested risk target varies over $[0.05, 0.35]$. Over the displayed coverage range, OCC has higher plotted GSR than the frozen baselines, while W-ECDF remains near the nominal 90% reference line. Each operating point is assessed against its own requested target. These curves describe empirical target attainment under the evaluated shift; the main-text comparison in Section 4.4 instead reports requested target levels at matched coverage.

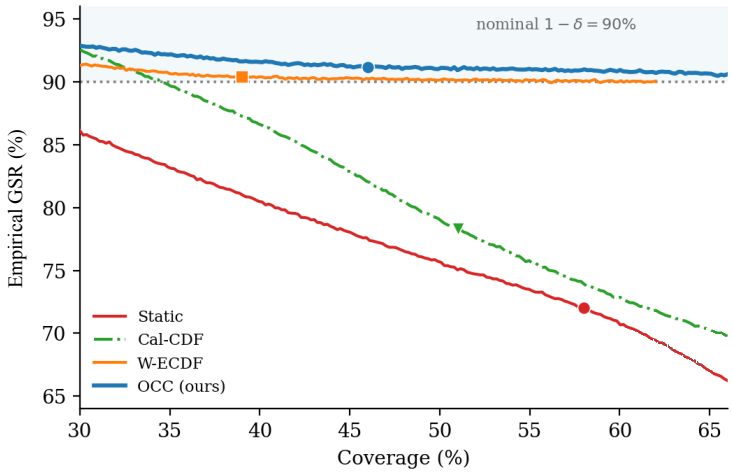


Figure 6: **Empirical target attainment under a slow covariate shift** (TL;DR, Gemma-3-27B, n=200, β=0.15; ε swept over $[0.05, 0.35]$). GSR is evaluated against each operating point's requested target, conditional on a nonempty accepted set, and is undefined if no such run remains. Large markers identify $\varepsilon = 0.20$ and match Table 10. The dashed 90% line is a nominal reference. The curves show empirical outcomes beyond the i.i.d. setting of Theorem 3.1.

Scope and limitation. These results concern the evaluated gradual-shift scenario and do not establish risk control under arbitrary drift. Pooling retains older confidence observations and can lag a changing distribution; windowing uses fewer observations but can respond more quickly to recent changes. Appendix J provides a complementary transient-contamination test.

J Robustness Analyses

We complement the stationary-stream experiments with transient contamination and sorted arrival orders, reporting empirical behavior beyond the i.i.d. setting of Section 3.3. We also vary the below-reference diagnostic tolerance introduced in Appendix G.

Contaminated stream. Following the corruption protocol of Hu et al. (2026), we simulate a distribution shift in the first 200 stream rounds by sampling confidences from a shifted distribution (+0.05 location shift), then reverting to the original distribution for the remaining rounds. Table 11 reports the recovery behavior. The difference in mean TE between contaminated and clean streams narrows from 0.017 at $t = 200$ to 0.003 at $t = 500$ and 0.001 at $t = 1{,}250$. Their end-of-deployment empirical GSRs are 91.4% and 92.1%, respectively.

Table 11: **Recovery from a contaminated stream** (TL;DR, Gemma-3-27B, n=200, 1,000 splits; contamination in rounds 1–200).

	TE_t ($\times 10^{-2}$)				GSR
Method	t=200	t=500	t=750	t=1250	(end)
OCC (clean stream)	5.1	4.0	3.5	3.2	92.1%
OCC (contaminated)	6.8	4.3	3.6	3.3	91.4%
Static	6.6	6.6	6.6	6.6	92.5%

Non-uniform arrival orders. Within each of 1,000 shared splits, we compare random, easy-first (descending confidence), and hard-first (ascending confidence) orders of the same stream, fixing calibration and reference sets, scores, and certification rules. Hard-first raises mean TE_{500} from 0.040 to 0.045 and delays $t_{0.035}$ by about 15% (963 versus 838; Table 12). After all T updates, identical confidence multisets yield identical CDFs and certified thresholds: equal endpoint TE checks this

invariance, while intermediate tracking and deployment-long GSR remain order-dependent. The contamination test changes the multiset, so its endpoints need not coincide.

Table 12: **Sensitivity to arrival order** (TL;DR, Gemma-3-27B, n=200, T=1,250, 1,000 shared splits). TE columns are split means ($\times 10^{-2}$); endpoint TE follows all T updates. $t_{0.035}$ uses the mean-curve definition in Appendix F; GSR covers the full deployment.

Arrival order	TE_{500}	TE_{1250}	$t_{0.035}$	GSR
Random (default)	4.0	3.2	838	92.1%
Easy-first	3.9	3.2	831	92.3%
Hard-first	4.5	3.2	963	91.4%

Sensitivity to the below-reference tolerance ξ. Let $N_{r,T}^{\mathrm{below}} = \sum_{t=1}^{T} \mathbf{1}\{\lambda_{r,t}^{*} < \lambda_{r}^{\mathrm{ref}} - \xi\}$. Table 13 reports $R^{-1} \sum_{r} N_{r,T}^{\mathrm{below}} / T$ for $R = 1{,}000$ splits and $\xi \in \{0.005, 0.01, 0.02\}$. The strict inequality excludes deviations exactly equal to ξ; on the threshold grid it is equivalent to comparing integer grid indices. Smaller tolerances count more deviations; GSR is unchanged because ξ affects only the diagnostic. Percentile deviations alone do not determine realized coverage.

Table 13: **Below-reference round fractions** (TL;DR, Gemma-3-27B, n=200, 1,000 splits).

ξ	Mean fraction (%)	GSR (%)
0.005	42.3	92.1
0.01 (default)	31.0	92.1
0.02	18.7	92.1