CERTIFIED ADAPTIVE REFRESH: ANYTIME-VALID MONITORING FOR FEDERATED CONFORMAL RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Question-answering services built on retrieval-augmented generation (RAG), in which a language model answers from retrieved documents, are inspected continuously and upgraded repeatedly, so their reliability guarantee must survive both. We study federated conformal RAG: retrieval nodes holding private corpora score a shared list of candidate answers with a shared language model and send only compressed scores to a hub, which returns an answer set; a *miss* is a set that omits the true answer. We formulate monitoring as a sequential test: raise an alarm when misses exceed a certified upper bound on the miss rate (the *allowance*), while keeping the probability of any false alarm over the whole deployment below a preset budget; an alarm is false when no deployed configuration has drifted above its allowance. Existing tools solve only parts of this problem. A conformal certificate bounds the miss rate of one frozen configuration at one look fixed in advance, so it supplies no *anytime-valid* alarm, one that stays valid however often it is checked; a threshold calibrated for one model certifies nothing about its replacement; and re-certifying each upgrade at the full error level compounds failures. We propose Anytime-FC-RAG, a protocol with three auditable rules: pick the next configuration *before* drawing its fresh calibration data, charge every certificate attempt to one trajectory-wide budget δ_{cal}, and fix the deployed threshold, allowance and monitoring bet before each query arrives. One betting wealth process E_t, evidence that grows in expectation only when misses exceed the allowance, then runs across every deployment without reset. Our *certified adaptive-refresh composition theorem* shows that, for an evidence budget δ_e, alarming when E_t first reaches $1/\delta_e$ has false-alarm probability $\mathbb{P}(\sup_{t\geq 1} E_t \geq 1/\delta_e) \leq \delta_e + \delta_{\mathrm{cal}}$ under a random number of history-selected, evidence-driven refreshes of the model, retriever, corpus, score or threshold; δ_{cal} pays for certificates that are wrong by chance. With an exact order-statistic certificate the guarantee is distribution-free and architecture-agnostic; it concerns misses against the allowance, not per-input coverage or distribution shift as such. Replaying Qwen2.5 on MMLU-Pro at a 5% total budget, all 210 runs whose shift an independent audit rated harmful alarm within $5{,}000$ queries, and none of 222 benign or unshifted runs does. In a swap of the deployed model, calibrating the incoming model on fresh data holds a downgrade's miss rate at .089, under its .10 target, whereas calibrating it on the outgoing model's scores lets it rise to .154.

1 INTRODUCTION

Motivating application. A deployed language-model service needs a reliability guarantee that survives the two things its operators actually do: look at it repeatedly, and change it. In retrieval-augmented generation (RAG), a language model answers a question using documents retrieved from a corpus. Consider a domain-specialized question-answering service whose corpus is split across several *retrieval nodes* that cannot pool their documents, each holding a private corpus and a copy of a shared scoring model. For each question every node scores a common finite list of candidate answers against its locally retrieved context and uploads only compressed scores, not documents, to a central *hub*, which returns an answer set; once the correct answer is adjudicated, the operator learns whether that set *missed* it. Persistent misses invite action (widen the communication budget, re-fit the retriever, swap the scoring model, recalibrate the threshold), and each changes the monitored system.

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.

The challenge. The standard guarantee for such a service breaks under both repeated inspection and change. That guarantee comes from *conformal prediction*, which sets a score threshold on held-out labeled questions so that the returned sets omit the true answer at most a target fraction of the time. In its calibration-conditional form it supplies a *certificate*: a high-probability promise that one configuration's miss rate stays under a stated *allowance*. Three difficulties arise. First, a certificate describes one look fixed in advance, not repeated checks that stop when misses seem high, so it supplies no *anytime-valid* alarm, one whose false-alarm guarantee holds however often and whenever the operator looks. Second, a threshold calibrated for one model certifies nothing about the model that replaces it. Third, re-certifying at the full error level after each change compounds failures: with 100 independent certificates at 5% each, the chance that any is wrong is $1 - .95^{100} = .9941$, against $.0488$–$.0490$ when the same 5% is split across them (Figure 2(a)).

Problem formulation. We formulate reliability monitoring as a sequential test of one hypothesis over the whole deployment: every deployed configuration keeps its miss rate within its certified allowance. The operator, who may replace any component between questions using everything observed so far, wants an alarm whose probability of ever firing while this hypothesis holds (a *false alarm*) is at most a preset budget δ; Section 2 makes this precise.

State of the art. To our knowledge, no existing family of methods solves this problem, because each certifies only a part of it (Table 5 in Appendix E). Federated conformal RAG (FC-RAG) (Dubey & Huo, 2026) and other federated conformal methods (Lu et al., 2023; Plassier et al., 2023; Wen et al., 2026; Xu et al., 2025) calibrate across nodes for one fixed configuration and raise no sequential alarm. Conformal test martingales (CTM) (Vovk et al., 2021; Prinster et al., 2025) are anytime-valid but test *exchangeability* of the score stream (roughly, that its distribution has not changed), so they can alarm on score changes that leave every miss intact. Sequential risk monitors (Podkopaev & Ramdas, 2022; Timans et al., 2025) are anytime-valid only for one *fixed* deployed model. Online and adaptive conformal methods (Gibbs & Candès, 2021; Angelopoulos et al., 2024b) re-adjust the threshold and so have no certificate to preserve, and post-selection conformal calibration (Bao et al., 2025) calibrates after an adaptive pick without budgeting a trajectory of refreshes.

What we propose. We propose *Anytime-FC-RAG*, a deployment protocol that lets the operator change the system as often as desired, provided it meets three conditions that can be audited after the fact. *Pick, then calibrate*: the next configuration is selected from past history, and only afterwards is a fresh calibration block (new labeled questions) drawn for it. *Budget every attempt*: each certificate attempt, including one computed and then discarded, spends from a single trajectory-wide budget δ_{cal}. *Act before the outcome*: the deployed configuration, threshold, communication action, allowance and bet are all fixed before the next question arrives. In order,

$$\text{select configuration} \;\longrightarrow\; \text{fresh calibration} \;\longrightarrow\; \text{certificate + threshold} \;\longrightarrow\; \text{deploy and monitor.}$$

Monitoring uses one *betting wealth process* $(E_t)_{t\geq 0}$, where E_t is the wealth, after the t-th question, of a gambler who starts with one unit ($E_0 = 1$) and at each question stakes part of it on the claim that the system misses more often than its allowance permits. The wealth's expected value cannot grow while the claim is false, so large wealth is evidence of excess misses. The same process runs across every configuration and is never reset, and the alarm fires when E_t first reaches $1/\delta_e$, for an *evidence budget* δ_e.

Advantages of the proposed method. The main advantage is a guarantee that does not weaken with the number of inspections or refreshes. Our composition theorem shows that the alarm is anytime-valid, with its false-alarm probability split between the evidence budget and the calibration budget:

$$\mathbb{P}\left(\sup_{t\geq 1} E_t \;\geq\; 1/\delta_e\right) \;\leq\; \delta_e + \delta_{\mathrm{cal}}, \tag{1}$$

however many times, and at whatever data-dependent moments, the operator refreshes the system. The number of refreshes may be random and unbounded, each new configuration may depend on everything observed so far, and no independence across refreshes is assumed; an operator who changes nothing pays δ_{cal} once, however long the deployment runs. Three further advantages follow. The certificate we use is exact in finite samples and distribution-free, and it assumes nothing about how the scoring model was built, so any model or retriever can be swapped in. The monitor targets

misses rather than scores, so changes that leave reliability intact keep its false-alarm rate within budget. And the guarantee sees only a certificate, an allowance and a budget, so it transfers to any sequential conformal protocol that supplies them before each query. On Qwen2.5 scores for the MMLU-Pro benchmark the alarm separated audited-harmful from benign shifts completely at a 5% budget (Section 5).

What is inherited, what we use, and what is new. The composition is new. We inherit FC-RAG's fixed-configuration inference and communication layer (Dubey & Huo, 2026) and use betting e-processes (nonnegative evidence processes whose mean is at most one at any stopping time), Ville's inequality (Section 4.2), conditional union bounds and mixture boundaries (Ville, 1939; Howard et al., 2020; Waudby-Smith & Ramdas, 2024; Ramdas et al., 2023). One frozen certified configuration monitored by one e-process is a special case of our result (Corollary 4.3(a)), and we do not claim it as new. What is new is the *certified composition* of these pieces across history-selected configurations, which is why Anytime-FC-RAG is the only row of Table 5 with every column checked.

Organization. Section 2 formulates the problem, Section 3 presents Anytime-FC-RAG, Section 4 establishes its guarantees, and Section 5 tests them in simulation.

2 Problem formulation

One round of bandwidth-constrained federated RAG. The monitored system answers a question by pooling compressed scores from the nodes rather than their documents. Let $\mathcal{X}$ be the space of questions and $\mathcal{Y}$ a finite, agreed-upon answer list. A synchronous star network has $K < \infty$ retrieval *nodes*, indexed by $i = 1, \dots, K$, and one *hub*. Node i is the tuple $(\mathcal{C}_i, R_i, \widehat{P}, \mathrm{Enc}_{B_i})$: a private corpus $\mathcal{C}_i$, a local retriever R_i, an inference copy $\widehat{P}$ of the shared *student* (following FC-RAG's terminology, the one scoring model deployed at every node), and an encoder Enc_{B_i} that compresses each score to B_i bits. Corpora and retrievers differ across nodes, but the student does not. Queries are indexed by $t = 1, 2, \dots$. At query t a question $X_t \in \mathcal{X}$ arrives, whose correct answer $Y_t \in \mathcal{Y}$ is revealed after the answer set is returned. Node i retrieves the context $Z_{i,t} = R_i(X_t, \mathcal{C}_i)$ and scores every candidate answer $y \in \mathcal{Y}$ by a *nonconformity score* $s_{i,t}(y)$, a number that is larger the less compatible y is with X_t and $Z_{i,t}$; it then quantizes the scores with Enc_{B_i} and uploads them. The hub aggregates the decoded scores into one score $s_t(y)$ per candidate, compares it with the active threshold q, and returns the answer set $C_t(X_t)$; once Y_t is known, it records the miss indicator M_t:

$$C_t(X_t) = \{y \in \mathcal{Y} : \, s_t(y) \le q\}, \qquad M_t = \mathbf{1}\{Y_t \notin C_t(X_t)\} \in \{0, 1\}.$$

Here $\mathbf{1}\{\cdot\}$ is the indicator function. The *coverage event* is $\{M_t = 0\}$, and M_t is the *miss* the operator monitors. Figure 1 shows the round. Raw documents, training examples, gradients and model weights are not part of this uplink; this is a message-content property of the protocol, not a privacy guarantee.

Quantization costs the guarantee nothing by default. By default the compressed, aggregated score is calibrated directly, so quantization is already reflected in the threshold. Alternatively, a certificate can cover a class of bit budgets by adding to its allowance a communication term Δ_{comm} that, under FC-RAG's dithered quantizer (one that adds node-independent random noise before rounding), shrinks with more bits and as $K^{-1/2}$ in the number of nodes, whereas one node gains nothing from averaging (Appendix D.1; a model property, as our system experiments use $K \le 8$).

Configurations, attempts and epochs. A deployment is a sequence of configurations, each certified on its own calibration data and served until it is replaced. A *score state* θ makes "configuration" precise: it collects every component that determines the deployed score, namely student model, retriever and corpus, candidate rule, score semantics, hub aggregation, threshold rule, and any class of communication actions (e.g. per-node bit budgets) the certificate covers uniformly, meaning its allowance holds whichever action in the class is used. *Calibration attempts* are indexed by $r = 1, 2, \dots$; attempt r proposes a state θ_r, acquires a fresh *calibration block* for it (new labeled questions not used to build or choose the state), and from that block computes a *certificate*, a high-probability bound on the miss rate that state will incur. An *epoch* is a maximal run of queries served by one state θ_r, its realized threshold q_r and its matching certificate, whose allowance u_r is defined in Section 3.1; r_t is the attempt active at query t, and a_t is the communication action used at query t, fixed before the query. Table 3 (Appendix A) collects the notation.

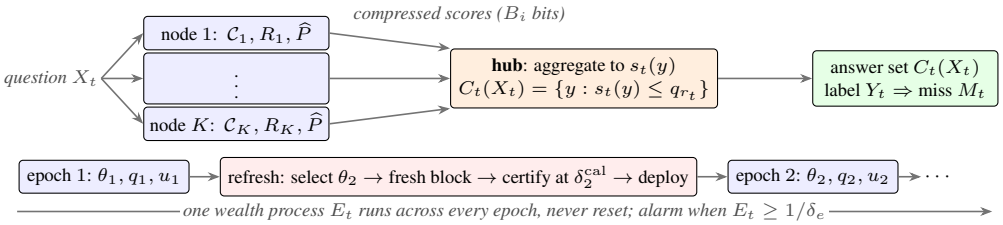


Figure 1: Top: the monitored system. Every node receives the question and scores the answer list $\mathcal{Y}$ against its private corpus with the shared scoring model $\widehat{P}$, then uploads compressed scores; the hub returns $C_t(X_t)$, and the adjudicated Y_t yields the miss M_t. Bottom: a deployment trajectory; epoch r serves state θ_r with threshold q_r and allowance u_r, certified on a fresh block with budget share δ_r^{cal} (Section 3, Algorithm 1).

What is known, and when. Three nested collections of information fix what each decision may depend on. All randomness lives on one probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and each collection of information is a σ-field, a family of events whose occurrence is known at that moment. Let $\mathcal{F}_0$ contain the pre-deployment history. For each query $t \ge 1$,

$$\mathcal{F}_{t-1} \;\subseteq\; \mathcal{F}_t^{\mathrm{pre}} \;\subseteq\; \mathcal{F}_t. \tag{2}$$

Here $\mathcal{F}_{t-1}$ holds everything known after query $t-1$; the *pre-query field* $\mathcal{F}_t^{\mathrm{pre}}$ adds every intervening operator decision, state selection, calibration block, certificate, threshold, action, allowance and bet, but closes *before* (X_t, Y_t); and $\mathcal{F}_t$ adds that pair, the returned set and M_t. A quantity is *fixed before query* t when it is determined by $\mathcal{F}_t^{\mathrm{pre}}$ ($\mathcal{F}_t^{\mathrm{pre}}$-measurable). Evaluating a rule fixed in this sense at X_t is allowed and necessary; choosing any of these quantities after seeing X_t is not.

The deployment null. The hypothesis under test is that answer sets keep honoring their certified allowance; breaching it, not any change in the data, is the operational failure. Write b_t for the *allowance* of query t, the certified upper bound on that query's miss rate, fixed before the query. The allowance bounds a miss rate; δ_e and δ_{cal} bound error probabilities. The *deployment null* is that $\mathbb{E}[M_t \mid \mathcal{F}_t^{\mathrm{pre}}] \le b_t$ for every query t whose certificate is correct (formally, Equation (5) in Section 3.2). The data may change so long as this holds, and certificates that are wrong by chance are paid for separately by δ_{cal}. Two consequences follow: coverage is *marginal over a fresh question*, not conditional on each particular question value $X = x$, and a change that leaves certified reliability intact should leave the monitor silent. That silence is correct, not a missed detection.

The goal. We seek a certificate for each configuration and an alarm rule such that, under the deployment null, the probability that the alarm *ever* fires is at most a preset total budget δ, for every refresh policy that obeys the timing constraints above, while sustained excess misses are detected. The number of refreshes may be random, and each new state may depend on all history.

Scope. The guarantee covers the round just described (finite fixed K, synchronous star, finite answer list, marginal coverage for a fresh question) together with a history-selected state followed by fresh state-matched calibration, and actions fixed before each outcome, with *sticky* (never reset) evidence. Each query the monitor bets on needs its adjudicated answer; a query chosen in advance to go unlabeled gets a zero bet, and late labels are covered for a fixed state (Corollary 4.3(e)). Asynchronous, peer-to-peer, Byzantine or privacy-preserving federation, open-ended generation, per-x coverage, general shift detection, recovery after an alarm and optimal control are outside it (Table 4, Appendix A).

3 Proposed method: Anytime-FC-RAG

Sections 3.1–3.2 implement the three conditions of Section 1 in order; Section 3.3 adds the monitor.

3.1 Pick, then calibrate: a certificate for each configuration

Each state is certified on a block drawn after the state is chosen. Before attempt r, the decision to attempt, the state θ_r it proposes, the calibration size, the threshold rule and the attempt's error

allocation are all fixed ($\mathcal{H}_r$-measurable) by the *pre-calibration history* $\mathcal{H}_r$, everything observed before the attempt. Only then is a *conditionally honest* block drawn, one not used to train or select θ_r. Absorb a fixed communication action into θ_r and write $s_{\theta_r}(X, Y)$ for the aggregate score that state θ_r gives the *true* answer Y of a question X, so $M_t = \mathbf{1}\{s_{\theta_{r_t}}(X_t, Y_t) > q_{r_t}\}$; let F_r be its distribution function for a fresh calibration question, given $\mathcal{H}_r$.

The threshold is a split-conformal order statistic, and the allowance is its exact Beta bound. Split conformal prediction sets the threshold at the k-th smallest of n true-answer scores on held-out labeled questions; a fresh question is then missed with a probability that depends on the calibration draw, and $k = \lceil (n+1)(1-\alpha) \rceil$ (with $\lceil \cdot \rceil$ rounding up) keeps it at most a target miss rate α *on average over the draw*. An average over the draw does not make each bet's payoff $M_t - b_t$ nonpositive in conditional mean (where b_t is the allowance of query t, Section 2); the monitor needs a bound that holds with high probability over the draw. One classical construction (Vovk, 2012) supplies it exactly in finite samples. Let the block contain n_r labeled questions with true-answer scores $S_{r,1}, \ldots, S_{r,n_r}$, and let $\mathcal{H}_r$ fix the block size n_r, an order $k_r \in \{1, \ldots, n_r\}$ and an error level $\gamma_r \in (0, 1)$, the probability the certificate is allowed to be wrong. Set

$$q_r = S_{r,(k_r)}, \qquad u_r = \mathrm{Beta}^{-1}_{1-\gamma_r}(n_r + 1 - k_r,\ k_r), \tag{3}$$

where $S_{r,(k)}$ is the k-th smallest of the n_r scores and $\mathrm{Beta}^{-1}_p(a, b)$ the p-quantile of the $\mathrm{Beta}(a, b)$ distribution. The state is deployed with threshold q_r and allowance $b_t = u_r$ throughout its epoch. For instance $\alpha = .10$, $n_r = 100$ (so $k_r = 91$) and $\gamma_r = .0125$ give $u_r = .175$, the allowance of the model swap in Section 5 (Appendix F.6). The interface below accepts any certificate that satisfies (4); this one is exact (Proposition 4.1).

3.2 Budget every attempt, and act before the outcome

All attempts share one pathwise budget. Every certificate attempt, including one computed and then discarded, spends a share of one deterministic total calibration budget $\delta_{\mathrm{cal}} \in [0, 1)$. Attempt r receives an allocation $\delta_r^{\mathrm{cal}} \geq 0$, its share of δ_{cal}, fixed by $\mathcal{H}_r$, and yields a *certificate event* G_r determined by its block: the typically unobservable event that its certified bound is correct (for the order-statistic certificate, $G_r = \{p_r \leq u_r\}$ with $p_r = 1 - F_r(q_r)$ the miss probability the threshold leaves, and $\gamma_r \leq \delta_r^{\mathrm{cal}}$). With G_r^c the complement of G_r (the certificate is wrong), the budget condition is

$$\mathbb{P}(G_r^c \mid \mathcal{H}_r) \leq \delta_r^{\mathrm{cal}} \quad \text{and} \quad \sum_{r \geq 1} \delta_r^{\mathrm{cal}} \leq \delta_{\mathrm{cal}} \qquad \text{almost surely.} \tag{4}$$

Allocations may be random and history-dependent, attempts may be dependent, and their number may be random or countably infinite: a known cap R allows $\delta_r^{\mathrm{cal}} = \delta_{\mathrm{cal}}/R$ (the Bonferroni split), and $\delta_r^{\mathrm{cal}} = 6\delta_{\mathrm{cal}}/(\pi^2 r^2)$ (a summable schedule, since $\sum_r r^{-2} = \pi^2/6$) allows unboundedly many. Charging even discarded attempts prevents an unbudgeted search over calibration outcomes.

The deployed state must come with a matching allowance fixed before the query. At query t the active certificate event G_{r_t}, the allowance b_t and the bet λ_t (the stake the monitor places on query t; Section 3.3) are fixed before the query ($\mathcal{F}_t^{\mathrm{pre}}$-measurable), and on that certificate, with $\mathbf{1}_G$ denoting the indicator of an event G,

$$\mathbf{1}_{G_{r_t}}\, \mathbb{E}[M_t \mid \mathcal{F}_t^{\mathrm{pre}}] \ \leq\ \mathbf{1}_{G_{r_t}}\, b_t, \qquad 0 < b_t < 1. \tag{5}$$

This is the deployment null of Section 2, the hypothesis the alarm tests. With (4) it makes a high-probability statement *over the calibration draw* about each deployed query's *conditional* miss rate. The timing rule restricts *when* a choice is made, not what is chosen: without it an operator must either freeze the configuration and forfeit the gains a refresh would buy, or react to M_t and forfeit validity.

3.3 Monitor with one never-reset wealth process

One wealth process runs across every epoch. Choose a bet $0 \leq \lambda_t \leq 1/b_t$ fixed before query t and set

$$E_0 = 1, \qquad E_t = E_{t-1}\big\{1 + \lambda_t(M_t - b_t)\big\}. \tag{6}$$

This is the gambler's wealth of Section 1 with stake λ_t: a miss multiplies it by $1 + \lambda_t(1 - b_t)$ and a covered query by $1 - \lambda_t b_t$. Under the null (5), on a correct certificate each factor has conditional

mean at most one, and the cap $\lambda_t \le 1/b_t$ keeps the wealth nonnegative. Any such bet is valid. Our simulations use the adaptive aGRAPA bettor of Waudby-Smith & Ramdas (2024), which sets each stake from past outcomes to approximately maximize the wealth's growth rate (Appendix D.3). An alarm is declared the first time $E_t \ge 1/\delta_e$, and the wealth is not reset at a refresh.

The whole protocol is one loop. Algorithm 1 (Appendix A) assembles the pieces: its pre-query line closes $\mathcal{F}_t^{\mathrm{pre}}$, its monitor line applies (6), and only its post-outcome line may change the configuration.

4 THEORETICAL GUARANTEES

Three results answer the questions an operator must settle before trusting the alarm. Is each certificate correct with the stated probability? Proposition 4.1 says yes, exactly, in finite samples. Does the alarm stay valid across every inspection and refresh? Theorem 4.2 says yes, at total cost $\delta_e + \delta_{\mathrm{cal}}$. Which moves keep that guarantee, and which break it? Corollary 4.3 and the paragraph after it answer.

4.1 EACH CERTIFICATE IS EXACT IN FINITE SAMPLES

Proposition 4.1 (Finite-sample order-statistic certificate)**.** *Conditional on $\mathcal{H}_r$, let $S_{r,1}, \dots, S_{r,n_r}$ be conditionally i.i.d. true-answer scores from F_r on a fresh labeled block (scores include the deployment randomization), and let the integer $n_r \ge 1$, the order $k_r \in \{1, \dots, n_r\}$ and $\gamma_r \in (0, \delta_r^{\mathrm{cal}}]$ be $\mathcal{H}_r$-measurable. Let q_r and u_r be as in* (3)*, and put $p_r = 1 - F_r(q_r)$. Then $\mathbb{P}(p_r > u_r \mid \mathcal{H}_r) \le \gamma_r$, with equality for continuous F_r; ties can make it conservative.*

Why it matters: the allowance is a finite-sample bound, not an asymptotic one. The miss probability p_r that the threshold leaves has an exact Beta law, so its $1 - \gamma_r$ quantile u_r bounds it with the stated confidence at every block size n_r, whatever the score distribution. Setting $G_r = \{p_r \le u_r\}$ and $b_t = u_r$ through the matching epoch then satisfies (4); (5) needs one further condition.

Assumption 1 (Epoch score stability)**.** Within the epoch served by attempt r, every deployed true-answer score $s_{\theta_r}(X_t, Y_t)$ has conditional law F_r given $\mathcal{F}_t^{\mathrm{pre}}$, the law of the calibration scores; fresh questions drawn i.i.d. from the calibration law, independently of all the operator has observed, suffice.

The assumption is the only place a condition on the data enters. Assumption 1 carries a certificate from its calibration block to the queries it certifies; Theorem 4.2 below needs nothing about the data beyond (4)–(5). It is imposed per epoch, not per trajectory, so a refresh is free to change the score law; drift *inside* an epoch breaks it, and the alarm targets drift that pushes misses above the allowance. The monitor tests misses against the *certified* u_r rather than the target α, as (5) requires (Appendix B.3).

4.2 THE COMPOSITION THEOREM

Theorem 4.2 (Certified adaptive-refresh composition)**.** *Assume the information structure* (2)*, the budget* (4) *and the null* (5) *for the state and action in force, with E_t as in* (6)*; only finitely many calibrations precede any query, and each certificate is completed before its state is deployed. Then, with $\Omega_0 = \Omega$ and $\Omega_t = \bigcap_{s \le t} G_{r_s}$ the event that every certificate in force at queries $1, \dots, t$ is correct, the killed wealth $\widetilde{E}_t = E_t \mathbf{1}_{\Omega_t}$ is a nonnegative supermartingale with respect to $(\mathcal{F}_t)_{t \ge 0}$, and for every $\delta_e \in (0, 1)$,*

$$\mathbb{P}\left(\sup_{t \ge 1} E_t \ge 1/\delta_e \right) \;\le\; \delta_e + \delta_{\mathrm{cal}}.$$

Why it holds: every choice is fixed before the query it affects. Everything the operator may vary (which state, when, how often, on what evidence) enters only through $\mathcal{F}_t^{\mathrm{pre}}$, and $\mathcal{F}_t^{\mathrm{pre}}$ closes before the query it is judged on. What makes that work is the *killed* wealth $\widetilde{E}_t$: the same bets, voided (set to zero) the moment any certificate on the path has failed. When a certificate is wrong, (5) says nothing about the miss rate, and killing discards exactly those paths, so two standard arguments meet at the certificate interface. The killed wealth is a nonnegative supermartingale (its conditional mean never rises), so Ville's inequality (such a process started at 1 reaches $1/\delta_e$ with probability at

most δ_e) contributes the first term; a conditional union bound over attempts, which assumes neither independence nor a fixed attempt count, contributes the second. On the event that no certificate ever fails the two wealths coincide, so δ_{cal} pays for the event on which they differ (Appendix B.2). □

Why it matters: one auditable budget covers the whole deployment. Because (1) holds at every t at once, an operator may inspect the wealth at *any* stopping time: no fixed horizon, no pre-committed stopping rule, no penalty for repeated peeking. For a target total δ, alarm at $1/\delta_e$ with $\delta_e + \delta_{\text{cal}} \le \delta$; $1/\delta$ would overspend whenever $\delta_{\text{cal}} > 0$. Later attempts may get smaller shares of δ_{cal}, hence looser allowances unless their blocks grow, never a weaker guarantee (Appendix B.1).

4.3 Corollaries, and the failure modes they exclude

The same interface delimits a usable design space. Corollary 4.3 lists six operational moves that keep the guarantee, each costing nothing or one allocation from δ_{cal} or δ_e: parts (a)–(c) say what may be done to the deployed state, (d) what is learned beyond the alarm, (e) how late labels are handled, and (f) how a controller may act. Recalibration or a realized threshold change opens a new epoch unless the active certificate covers it uniformly (Appendix A.3); Appendix C proves each part.

Corollary 4.3 (The certified design space)**.** *Under the assumptions of Theorem 4.2, with part (e) relaxing immediate feedback, the following hold.*

(a) Frozen reuse. *One state and certificate G_1 with deterministic $\delta_1^{\text{cal}} \le \delta_{\text{cal}}$ give $\mathbb{P}(\sup_t E_t \ge 1/\delta_e) \le \delta_e + \delta_1^{\text{cal}}$: the calibration term never grows, however many queries or inspections follow.*

(b) Post-selection construction. *θ_r may come from any $\mathcal{H}_r$-measurable randomized, history-dependent training or selection procedure; fresh calibration then needs no multiple-testing correction for the states considered before selection.*

(c) Certified library. *To choose* after *seeing the block, predeclare a finite $\mathcal{H}_r$-measurable library of candidate states indexed by j, where member j has certificate event $G_{r,j}$ and allocation $\delta_{r,j}$, with allocations totalling at most δ_{cal}; then $G_r = \bigcap_j G_{r,j}$ certifies the attempt and any member may be deployed.*

(d) Cumulative envelope. *With $A_t = \sum_{s \le t}(M_s - b_s)$, the cumulative excess of misses over allowances, and a fixed finite mixture of Hoeffding bounds (Appendix C.2), there is a boundary $h_t(\delta_e)$ with $\mathbb{P}(\exists t : A_t \ge h_t(\delta_e)) \le \delta_e + \delta_{\text{cal}}$, and on the complement of that event every almost surely finite stopping time τ (a query index chosen using only information available up to that query) obeys $\tau^{-1}\sum_{t \le \tau} M_t \le \tau^{-1}\sum_{t \le \tau} b_t + h_\tau(\delta_e)/\tau$: where the alarm reports* whether *misses exceed the allowance, the envelope bounds how far they can run above it by chance.*

(e) Exogenous delayed feedback. *For one certified state whose misses are, given its calibration block, independent with mean at most b, and a delay schedule conditionally independent of them, running the wealth in reveal time with the bet fixed before each batch preserves the guarantee.*

(f) Certificate-preserving controller. *A controller may use all of $\mathcal{F}_t$ provided each decision either keeps the state and picks an action the certificate covers uniformly, or completes select–calibrate–certify–deploy before the next affected query. Adaptivity inside this class is unrestricted.*

Five moves break the guarantee. *Same-block selection* chooses among candidate states with the calibration block that then certifies the winner (unless the library was certified simultaneously, part (c)). *Calibrate-then-replace* deploys a new model behind a threshold calibrated for the outgoing one, which would need its own score-transfer bound (Appendix D.2). *Same-query peeking* uses X_t or M_t to choose that query's state, threshold, action, allowance or bet. *Unbudgeted search* tries calibration attempts without charging them to δ_{cal}. *Resetting wealth* without paying for it also breaks validity, though a restart that allocates δ_e across epochs is safe (Appendix C.1).

5 Simulations

The simulations show that dropping any one condition breaks validity and that the monitor detects harmful change within its false-alarm budget. Synthetic studies break one condition of

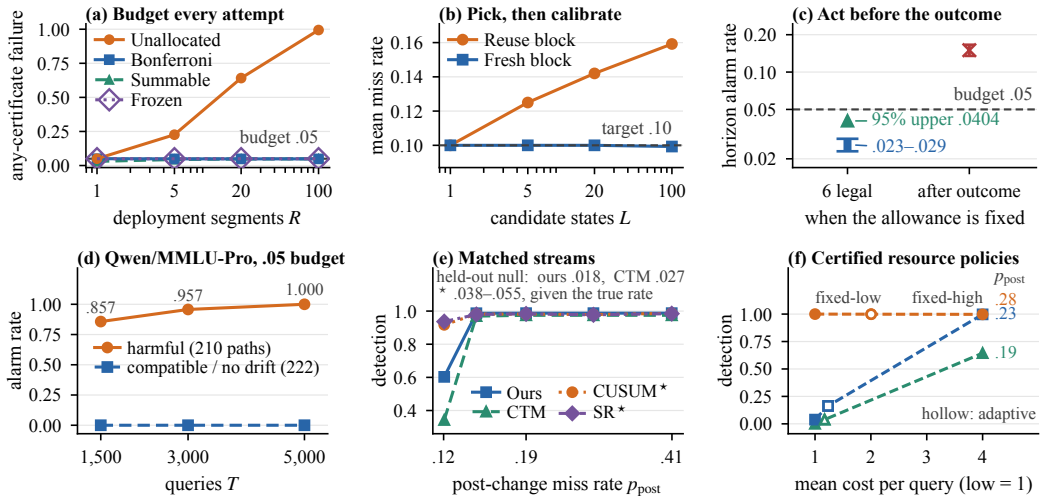


Figure 2: (a)–(c) Synthetic: each panel breaks one condition and prices it. (a) Exact probability that any certificate fails over R segments of 100 queries (failure alone; Table 1's composed rows count alarm *or* failure): .05 per certificate compounds toward certain failure, while the Bonferroni and summable schedules of Section 3.2 hold. (b) Reusing the block that selected the best of L candidate states inflates its miss rate above the .10 target. (c) A monitor whose allowance is chosen after the outcome alarms at three times its budget; six legal schedules fixed before each query hold (triangle: their largest simultaneous 95% upper bound). **(d)–(f)** Monitoring at budget: (d) every audited-harmful Qwen path alarms by horizon $T = 5{,}000$ queries, and no compatible or no-drift path does; (e) without the true post-change rate, ours detects as often as detectors given it, except near the allowance .10 (Table 2); (f) one dashed line per post-change miss rate, from fixed-low through the adaptive policy (middle point) to fixed-high bandwidth.

Table 1: Null-behavior claims beside their budgets; italic rows are deliberately invalid comparators. Composed rows count alarm or certificate failure (Monte Carlo, 5,000 paths; the monitor's own null rate stays $\leq .0262$); the rest use a known allowance. Sentinels are controllers that bet only on high-bandwidth queries scheduled in advance.

Study	Budget	Measured
Alarm, miss rate at / below allowance	.05	.0210/.0025
Envelope breach, 4,000 null paths	.05	.0015
Composed, 100 refreshes (Bonferroni)	.10	.0490
same design, unallocated	—	.9918
Six pre-query allowance schedules	.05	.023–.029
illegal: allowance set after M_t	—	.150
Seven exogenous label-delay laws	.05	.019–.022
Twelve sentinels × two nulls	.05	$\leq .0401$

the protocol at a time and check every guarantee at its stated budget; replays of cached Qwen2.5 scores (computed once by the real model, then fed through the protocol) and matched baselines measure detection on simulated deployment streams (*paths*). Protocols were fixed in advance (Appendices F–G).

Each condition is load-bearing, and every budget is respected. Breaking any one condition inflates the error far above its budget or target (Figure 2(a)–(c)): for example, re-certifying each of 100 refreshes at the full .05 leads to an alarm or a failed certificate with probability .9918, and a monitor whose allowance is set after the outcome alarms at .150 against a .05 budget. Every valid design in Table 1 stays below its budget (at most .0401 against .05, and .0490 against .10 for one never-reset wealth process across 100 refreshed epochs); allocation, not independence, restores validity (Lemma B.1).

On Qwen2.5/MMLU-Pro, harmful and benign shifts separate completely at a 5% budget (Figure 2(d)). The open language model Qwen2.5-1.5B-Instruct (Qwen Team, 2024) scored the

Table 2: Matched baselines on paired synthetic streams whose pre-change miss rate equals the allowance .10; * marks CUSUM and Shiryaev–Roberts (SR), *given* the true post-change rate p_{post} and a null-calibrated threshold; T is the horizon in queries. Detection (↑) is by $T = 5{,}000$; ↓ columns are alarm rates against a .05 budget: a held-out null, $T = 20{,}000$ under the frozen $T = 5{,}000$ threshold, and three miss-preserving score transforms, which CTM is designed to detect (Appendices F.7, F.9).

	Told the alternative	Held-out null ↓	Detection ↑ p_{post}=.123	Detection ↑ $p_{\mathrm{post}}\geq$.146	Null, T=20,000 ↓ frozen threshold	Miss-preserving shifts ↓
Ours	no	.0175	.603	.987–.988	.036	.003/.003/.003
CTM	no	.0270	.343	.970–.975	.015	.142/.964/.969
CUSUM/SR*	yes	.038–.055	.917–.936	.978–.987	.193–.274	—

ten options of questions from the multiple-choice benchmark MMLU-Pro (Wang et al., 2024) over eight nodes at $\delta_{\mathrm{cal}} = \delta_e = .025$, each allowance being an independent audit's exact Clopper–Pearson bound. Each path switched after 500 queries to one of 14 subjects; a separate audit rated each split–subject cell harmful, compatible with the allowance or (three cells) ambiguous. By $T = 5{,}000$, all 210 paths with an audited-harmful shift alarmed, and none of the 192 compatible, 30 no-drift and 18 ambiguous-cell paths did. A discovery-disjoint confirmation, on questions these paths never used, upheld the strongest harmful shift. In a separate study at a .10 budget, removing any one retrieval node raised no alarm on 240 paths yet left the harmful Computer Science and History shifts detected (Appendices F.5 and F.9).

Calibrating a new model on fresh data keeps its miss rate in check, and the alarm stays within budget across the swap. When the deployed 1.5B model is replaced by the 0.5B model on a two-node system, calibrating the incoming model on a fresh block holds the miss rate at .089, below the .10 target, whereas calibrating it on the outgoing model's scores lets it rise to .154. With one wealth process running across the swap, the alarm rate is only .025 and .015 on paths whose served miss rate is aimed at the allowance, within the .05 budget; under drift, the monitor still fires on .925–.955 of paths after the swap, so the refresh did not hide the shift (Appendix F.6).

Without being told the post-change rate, the monitor detects as often as detectors that are told it, except near the allowance (Figure 2(e)). Table 2 compares our monitor with CUSUM and Shiryaev–Roberts (Page, 1954; Roberts, 1966), classical change-point detectors that are here handed the true p_{post}, and with the conformal test martingale (CTM) of Vovk et al. (2021), which tests exchangeability. Our monitor has the lowest held-out null rate (each against its own null) and detects the stronger alternatives as often as the oracle detectors; knowing the alternative raises detection near the allowance and buys speed elsewhere (Appendix F.7).

The alarm tracks misses rather than scores, and its threshold $1/\delta_e$ needs no recalibration. Under score shifts that leave every miss unchanged, our alarm rate stays at .003 while CTM's reaches .969. Our Ville threshold stays within budget at $T = 20{,}000$, whereas CUSUM and Shiryaev–Roberts thresholds calibrated for $T = 5{,}000$ exceed the budget there. Online conformal prediction, by contrast, re-adjusts its score threshold through a harmful shift, so its own output does not reveal the shift (Appendix F.8).

The default bettor needs no tuning, and monitoring is cheap. With no tuning, detection rises from .310 to .996 as misses exceed the allowance by .02 to .04. Raising bandwidth once evidence builds detects more than fixed-low bandwidth at lower cost than fixed-high (Figure 2(f)), and one update costs 24.0 μs (Appendices F.9 and F.10).

6 Conclusion

A deployment may change its model, retriever, corpus, score or threshold as often as it likes and still report one trajectory-level false-alarm guarantee, provided it follows select–calibrate–certify–deploy, budgets every certificate attempt and acts before each outcome. On Qwen2.5/MMLU-Pro the monitor alarmed on every audited-harmful shift and no benign one. The guarantee is deliberately narrow (Section 2); Appendix G lists what remains open.

Reproducibility statement

The operating model and all assumptions are stated in Section 2, Assumption 1 and Appendix A, with the covered and excluded regimes in Table 4. Complete proofs of Proposition 4.1, Theorem 4.2 and Corollary 4.3 are in Appendices B–C. Every experimental design (models, dataset revisions, eligibility rules, split and seed construction, audit sizes, budget allocations and horizons) is specified in Appendix F, and Appendix G records the compute environment, the cache-integrity checks, and how every number reported in Section 5 traces to a completed run. Figures 2 and 3 are rendered by scripts that read the retained numerical summaries directly, so each plotted value is traceable to a completed run. The anonymized supplementary material contains the code, one configuration per experiment (fixing its seeds and any pinned model and dataset revisions), the retained result summaries, the figure scripts, a verifier and a map from every figure and table to its result files.

Use of generative AI and AI-assisted technologies

We used generative AI tools (ChatGPT, Google Gemini, and Claude) for language editing and code formatting support. All data, results, and mathematical derivations are the authors' own work.

References

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=33XGfHLtZg.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 1616–1630. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/angelopoulos24a.html.

Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. CAP: A general algorithm for online selective conformal prediction with FCR control. *Journal of Machine Learning Research*, 26(287): 1–74, 2025. URL https://www.jmlr.org/papers/v26/24-0452.html.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276. URL https://doi.org/10.1214/23-AOS2276.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Debashish Chakraborty, Eugene Yang, Daniel Khashabi, Dawn Lawrie, and Kevin Duh. Principled context engineering for RAG: Statistical guarantees via conformal prediction. In *Advances in Information Retrieval: 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29–April 2, 2026, Proceedings, Part II*, pp. 537–546, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-21299-3. doi: 10.1007/978-3-032-21300-6_45. URL https://doi.org/10.1007/978-3-032-21300-6_45.

Prasanjit Dubey and Xiaoming Huo. Federated language models under bandwidth budgets: Distillation rates and conformal coverage, 2026. URL https://arxiv.org/abs/2605.09986.

John C. Duchi. Sample-conditional coverage in split-conformal prediction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 86761–86791, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7d83f9e9462c417e208d55e83a5058a8-Abstract-Conference.html.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. E-values expand the scope of conformal prediction, 2025. URL https://arxiv.org/abs/2503.13050.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

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. doi: 10.1214/20-AOS1991. URL https://doi.org/10.1214/20-AOS1991.

Chenfeng Huang, Zixuan Ma, and George Michailidis. Model-agnostic online certificate-driven calibration for time series forecasting under distribution shift. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 2244–2273, 2026. URL https://proceedings.mlr.press/v337/huang26b.html.

Bror Hultberg, Dave Zachariah, and Antônio H. Ribeiro. Anytime-valid conformal risk control, 2026. URL https://arxiv.org/abs/2602.04364.

Hamed Khosravi and Xiaoming Huo. Conformal selective acting: Anytime-valid risk control for RLVR-trained LLMs, 2026. URL https://arxiv.org/abs/2605.20270.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6 (2):167–195, 2015. doi: 10.3233/SW-140134.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3799–3821, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.210. URL https://aclanthology.org/2024.naacl-long.210/.

Charles Lu, Yaodong Yu, Sai Praneeth Karimireddy, Michael Jordan, and Ramesh Raskar. Federated conformal predictors for distributed uncertainty quantification. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22942–22964. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lu23i.html.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100. URL https://doi.org/10.1093/biomet/41.1-2.100.

Vincent Plassier, Mehdi Makni, Aleksandr Rubashevskii, Eric Moulines, and Maxim Panov. Conformal prediction for federated uncertainty quantification under label shift. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27907–27947. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/plassier23a.html.

Aleksandr Podkopaev and Aaditya Ramdas. Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Ro_zAjZppv.

Drew Prinster, Xing Han, Anqi Liu, and Suchi Saria. WATCH: Adaptive monitoring for AI deployments via weighted-conformal martingales. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49830–49859, 2025. URL https://proceedings.mlr.press/v267/prinster25a.html.

Qwen Team. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894. URL https://doi.org/10.1214/23-STS894.

S. W. Roberts. A comparison of some control chart procedures. *Technometrics*, 8(3):411–430, 1966. doi: 10.1080/00401706.1966.10490374. URL https://doi.org/10.1080/00401706.1966.10490374.

Shalev Shaer, Yarin Bar, Drew Prinster, and Yaniv Romano. Testing for distribution shifts with conditional conformal test martingales. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=B8sTBdGvqG.

Alexander Timans, Rajeev Verma, Eric Nalisnick, and Christian A. Naesseth. On continuous monitoring of risk violations under unknown shift. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 4204–4226, 2025. URL https://proceedings.mlr.press/v286/timans25a.html.

Jean Ville. *Étude critique de la notion de collectif*. Number 3 in Monographies des probabilités. Gauthier-Villars, Paris, 1939. URL http://eudml.org/doc/192893. Fascicule III.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk, Ivan Petej, Ilia Nouretdinov, Ernst Ahlberg, Lars Carlsson, and Alex Gammerman. Retrain or not retrain: conformal test martingales for change-point detection. In Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin, and Khuong An Nguyen (eds.), *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, volume 152 of *Proceedings of Machine Learning Research*, pp. 191–210. PMLR, 08–10 Sep 2021. URL https://proceedings.mlr.press/v152/vovk21b.html.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018.

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. doi: 10.1093/jrsssb/qkad009. URL https://doi.org/10.1093/jrsssb/qkad009.

Haifeng Wen, Osvaldo Simeone, and Hong Xing. Efficient federated conformal prediction with group-conditional guarantee, 2026. URL https://arxiv.org/abs/2603.14198.

Rui Xu, Xingyuan Chen, Wenxing Huang, Minxuan Huang, Yun Xie, Weiyan Chen, and Sihong Xie. Federated conditional conformal prediction via generative models, 2025. URL https://arxiv.org/abs/2510.13297.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, pp. 60110–60132, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6eb05d8bc6bd7bb6868c64b5802125bd-Abstract-Conference.html.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28, pp. 649–657. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

A OPERATING MODEL AND PROTOCOL

This appendix states the statistical and operational contract in full, so that a reader can check applicability without reconstructing it from the proofs. Section 2 gives the same model in compressed form.

A.1 ONE ROUND OF BANDWIDTH-CONSTRAINED FEDERATED RAG

There are $K < \infty$ retrieval nodes and one hub in a synchronous star. The nodes share one deployed student model but retain separate private corpora and retrievers. For a question X_t and a finite, agreed-upon answer list $\mathcal{Y}$, every active node scores every $y \in \mathcal{Y}$ using its locally retrieved context; in the Qwen study of Appendix F.5, that local score is the full-option mean token negative log-likelihood. Each node quantizes its scalar scores and uploads them; the hub aggregates them, compares them with the active threshold, and returns a set $C_t(X_t)$. In symbols, one round is

$$\{\text{private corpus } i\}_{i=1}^{K} \xrightarrow[\text{shared student}]{\text{local retrieval}} \{\text{quantized scores}\}_{i=1}^{K} \xrightarrow{\text{hub aggregation}} C_t(X_t),$$

and once the label is adjudicated the monitored miss is $M_t = \mathbf{1}\{Y_t \notin C_t(X_t)\}$. Raw documents, training examples, gradients and model weights are not part of this uplink. The analysis is of the score-message protocol; it does not supply a privacy guarantee merely because documents remain local.

The analysis uses a finite-list, full-scoring convention. If a deployment instead generates a query-dependent candidate list, that generation rule must be placed inside the score state and the failure to place Y_t on the list must be included in the miss event. Open-ended generation is not covered by the present result.

Table 3 collects the notation and terms of the main text.

Table 3: Notation and terms used in the main text; Sections 1–4 define each where it first appears.

Symbol or term	Meaning
X_t, Y_t; $C_t(X_t)$; M_t	Question and true answer at query t; the returned answer set; the miss $\mathbf{1}\{Y_t \notin C_t(X_t)\}$
Score state θ; epoch	Everything determining the deployed score (model, retriever and corpus, candidate rule, score semantics, hub aggregation, threshold rule, and any certified within-epoch communication-action class), and the maximal run of queries sharing one such state, its realized threshold q_r and its matching certificate
Action a_t; covered uniformly	The pre-query communication choice, e.g. per-node bit budgets B_i; a certificate covers a class of actions uniformly when its allowance holds whichever member is used
Attempt r; history $\mathcal{H}_r$; certificate event G_r; allocation δ_r^{cal}	A state and allocation chosen from the pre-calibration history $\mathcal{H}_r$, then a fresh calibration block yielding the (typically unobservable) event G_r on which the epoch's risk bound holds; an attempt is charged even if discarded
δ_{cal}; δ_e	Total calibration budget, bounding the chance that any certificate is wrong; evidence budget, bounding false alarms when every certificate is correct
Allowance b_t; u_r; α	A pre-query upper bound on $\mathbb{E}[M_t \mid \mathcal{F}_t^{\mathrm{pre}}]$ holding on the active certificate; the Beta-quantile allowance of Proposition 4.1; the target miss rate that sets the order k_r
$\mathcal{F}_{t-1} \subseteq \mathcal{F}_t^{\mathrm{pre}} \subseteq \mathcal{F}_t$	Information after query $t-1$; that plus everything fixed before query t; that plus query t's pair, set and miss
Wealth E_t; bet λ_t; sticky	The evidence $E_t = E_{t-1}\{1 + \lambda_t(M_t - b_t)\}$ against the null, with the predictable stake $\lambda_t \in [0, 1/b_t]$; sticky wealth is never reset
Calibrate-then-replace; same-block selection	Deploying a new model behind a threshold calibrated for the outgoing one; selecting a state with the calibration block that then certifies it. Both break the guarantee

A.2 Information structure and the certified transition

All calibration, deployment, protocol, controller and training randomness is defined on the common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ of Section 2. Every $\mathcal{H}_r$, $\mathcal{K}_r$, $\mathcal{F}_t^{\mathrm{pre}}$ and $\mathcal{F}_t$ below is a sub-σ-field of $\mathcal{A}$.

Let $\mathcal{F}_0$ contain the pre-deployment history. For deployment query $t \geq 1$,

$$\mathcal{F}_{t-1} \subseteq \mathcal{F}_t^{\mathrm{pre}} \subseteq \mathcal{F}_t.$$

The first field contains all information after query $t-1$. The pre-query field additionally contains every intervening state selection, completed calibration block, certificate, threshold, communication action, allowance and bet, but is formed *before* the fresh pair (X_t, Y_t). The last field additionally contains that pair, the returned set and M_t. The certified risk is therefore conditional on the pre-query history but marginal over a fresh question. Evaluating a pre-committed retrieval, scoring and set rule at X_t is allowed; using X_t to select the state, threshold, bandwidth, allowance or bet would require a stronger certificate conditional on X_t.

Fix a deterministic total budget $\delta_{\mathrm{cal}} \in [0, 1)$, and before calibration attempt r let $\mathcal{H}_r$ be the available history. The decision to make attempt r, the proposed state θ_r, the calibration size, the threshold rule and the allocation $\delta_r^{\mathrm{cal}} \geq 0$ are $\mathcal{H}_r$-measurable. A fresh block D_r is then drawn and $\mathcal{K}_r = \mathcal{H}_r \vee \sigma(D_r, \text{calibration randomization})$. Because each certificate is completed before its state is deployed, $\mathcal{K}_r \subseteq \mathcal{F}_t^{\mathrm{pre}}$ for every query t the attempt serves. The canonical state-changing operation is:

1. use $\mathcal{H}_r$ to select θ_r and δ_r^{cal};
2. acquire a fresh, conditionally honest calibration block that was not used to train or select θ_r;
3. construct the state-matched threshold, allowance and event G_r; and
4. deploy only after these objects belong to the pre-query information.

Calibrate-then-replace, deploying a new model behind a threshold calibrated for the outgoing one, does not certify the replacement: selection must precede the fresh, state-matched calibration block.

The assumptions used by Theorem 4.2 can now be stated compactly: the budget condition and the deployment null of Section 3.2, repeated here under their original numbers, hold almost surely:

$$\mathbb{P}(G_r^{\mathrm{c}} \mid \mathcal{H}_r) \leq \delta_r^{\mathrm{cal}}, \qquad \sum_{r \geq 1} \delta_r^{\mathrm{cal}} \leq \delta_{\mathrm{cal}}, \tag{4}$$

$$\mathbf{1}_{G_{r_t}} \mathbb{E}[M_t \mid \mathcal{F}_t^{\mathrm{pre}}] \leq \mathbf{1}_{G_{r_t}} b_t, \qquad 0 < b_t < 1. \tag{5}$$

With the information structure and timing above, these two are what Theorem 4.2 needs; a third condition, Assumption 1 of Section 4.1, is needed only to carry a *particular* certificate across the epoch it was drawn for, and is where any distributional condition on the data enters. Allocations may be random and history-dependent, but each is chosen before its block. Attempts need not be independent, and their realized count may be random or countably infinite, provided only finitely many calibration operations occur before any one query. To pad a random collection of attempts to a countable sequence, set $G_r = \Omega$ and $\delta_r^{\mathrm{cal}} = 0$ for every unattempted slot; the same convention ($G_r = \Omega$, $\delta_r^{\mathrm{cal}} = 0$) covers a known, deterministic allowance.

Table 4 sets the covered regime beside the adjacent problems that this paper does not settle; Section 2 states the same split in prose.

Algorithm 1 states the loop the theorem assumes, and the correspondence is line by line: its pre-query step is the closure of $\mathcal{F}_t^{\mathrm{pre}}$ above, its monitor step is the wealth recursion (6), and its two post-outcome branches are exactly the two cases of Appendix A.3. The four-step state-changing operation listed above is what its second branch performs.

A.3 What a controller may change

After query t, a controller may use $\mathcal{F}_t$, including the wealth, the misses, realized costs and its own policy randomization, to choose later actions. The controller remains covered if it either (i) retains the state and chooses an action from a communication-action class that the current certificate covers uniformly, or (ii) completes a new select–calibrate–certify–deploy transition before the next affected query. Outside case (i), a change of model, retriever, corpus, candidate rule, score, aggregation rule

Table 4: Scope of the formal guarantee. The right column lists what the theorem does not establish: adjacent problems it leaves open and, in the middle row, the moves of Section 4.3 that break it.

Covered	Not established
Finite fixed K; synchronous star; shared student copies; private corpora; uplink scalar-score messages; finite answer list; marginal coverage for a fresh question	Asynchronous, peer-to-peer, Byzantine or privacy-preserving federation; open-ended generation; coverage conditional on every $X = x$
History-selected state then fresh state-matched calibration; pre-outcome covered actions; sticky evidence	Same-block selection; calibrate-then-replace; same-query peeking; unbudgeted search; unpaid monitor restarts
Evidence for excess conditional miss rate; fixed-state exogenous-delay corollary	General score or covariate shift detection; guaranteed recovery after an alarm; optimal control; outcome-dependent missingness

Algorithm 1 Anytime-FC-RAG: one certified deployment trajectory

Require: budgets $\delta_e + \delta_{\mathrm{cal}} \le \delta$, an initial state θ_1 and its allocation δ_1^{cal}

draw a fresh, conditionally honest block for θ_1, compute q_1 and its allowance, and set $E_0 \leftarrow 1$, $r \leftarrow 1$ ▷ Eq. (3)

for $t = 1, 2, \ldots$ **do**

  pre-query: fix the active attempt r_t, threshold q_{r_t}, action a_t, allowance b_t and bet $\lambda_t \in [0, 1/b_t]$ ▷ closes $\mathcal{F}_t^{\mathrm{pre}}$; Eq. (2)

  every node retrieves locally, scores $\mathcal{Y}$ with the deployed student and uploads quantized scores; the hub returns $C_t(X_t) = \{y : s_t(y) \le q_{r_t}\}$ ▷ Fig. 1

  observe Y_t; $M_t \leftarrow \mathbf{1}\{Y_t \notin C_t(X_t)\}$; $E_t \leftarrow E_{t-1}\{1 + \lambda_t(M_t - b_t)\}$ ▷ Eq. (6)

  if $E_t \ge 1/\delta_e$ **then**

    raise the alarm; **do not** reset the wealth unless a share of δ_e pays for the restart (App. C.1) ▷ Thm. 4.2

  end if

  post-outcome: using all of $\mathcal{F}_t$, either keep θ_{r_t} and pick an action the active certificate covers uniformly, or select θ_{r+1} and allocation $\delta_{r+1}^{\mathrm{cal}}$ from $\mathcal{H}_{r+1}$, calibrate it freshly, certify it, deploy it unless discarded, and set $r \leftarrow r + 1$ (the allocation is spent either way) ▷ Cor. 4.3(f)

end for

or realized threshold starts a new epoch, and so do recalibration and certificate replacement unless the original event covered the realized choice uniformly. A certified bound $b_t \ge 1$ is uninformative and inadmissible for the betting theorem, and cannot be clipped below one while preserving validity.

Concretely, the controller has three main levers, each covered when it is exercised through case (i) or (ii) before the next query: *recalibration*, which requests fresh compressed calibration summaries from each node and reconstructs the threshold at the hub; *bandwidth escalation*, which raises selected per-node budgets B_i or the calibration-refresh budget; and *student refresh*, which replaces the deployed model. Student refresh always opens a new certified epoch, and recalibration does too unless the active certificate covers the realized threshold uniformly. Bandwidth escalation may stay inside the current epoch when the active certificate covers the new action uniformly, and otherwise opens one too.

The guarantee therefore survives every certificate-preserving controller, and the wealth remains valid evidence for its decisions. It does not prove that an alarm-triggered intervention restores coverage, minimizes communication, or solves an optimal stochastic-control problem.

B Proofs

B.1 Trajectory-level certificate accounting

Lemma B.1 (Conditional familywise control)**.** *If* (4) *holds, then* $\mathbb{P}\big(\bigcup_{r \ge 1} G_r^c\big) \le \delta_{\mathrm{cal}}$.

Proof. For a finite prefix, the union bound and the tower property give

$$\mathbb{P}\left(\bigcup_{r=1}^{R} G_r^c\right) \le \mathbb{E}\left[\sum_{r=1}^{R} \mathbf{1}_{G_r^c}\right] = \mathbb{E}\left[\sum_{r=1}^{R} \mathbb{P}(G_r^c \mid \mathcal{H}_r)\right] \le \mathbb{E}\left[\sum_{r=1}^{R} \delta_r^{\mathrm{cal}}\right] \le \delta_{\mathrm{cal}}.$$

Letting $R \to \infty$ and using continuity from below (equivalently, Tonelli's theorem on the nonnegative indicators) proves the claim. The argument charges discarded attempts, and it assumes neither a fixed number of attempts nor independence among them. □

In words: the chance that *any* certificate the trajectory ever issues is wrong is at most the total δ_{cal} fixed in advance, no matter how many attempts are made, in what order, or how strongly they depend on one another. It is what allows Theorem 4.2 to treat an unbounded, adaptively chosen sequence of refreshes as if it were a single certificate.

When $0 < \delta_{\mathrm{cal}} < 1$, the deterministic schedule $\delta_r^{\mathrm{cal}} = 6\delta_{\mathrm{cal}}/(\pi^2 r^2)$ allows arbitrarily many attempts. For a standard root-n_r conditional certificate, the allocation-dependent term of the allowance scales under this schedule as $O\big(\sqrt{(\log(1/\delta_{\mathrm{cal}}) + 2\log r)/n_r}\big)$, so $n_r \asymp \log r$ with a sufficiently large constant keeps that term bounded at a fixed level. A known cap R instead permits $\delta_r^{\mathrm{cal}} = \delta_{\mathrm{cal}}/R$.

B.2 Proof of Theorem 4.2

Define $\Omega_0 = \Omega$, $\Omega_t = \bigcap_{s \le t} G_{r_s}$ and $\widetilde{E}_t = E_t \mathbf{1}_{\Omega_t}$. Because each certificate is completed before its state is deployed, $\Omega_t \in \mathcal{F}_t^{\mathrm{pre}}$ and $\Omega_{t-1} \in \mathcal{F}_{t-1}$. The constraint $0 \le \lambda_t \le 1/b_t$ gives

$$1 + \lambda_t(M_t - b_t) \;\ge\; 1 - \lambda_t b_t \;\ge\; 0,$$

so both E_t and $\widetilde{E}_t$ are nonnegative. Moreover $\Omega_t \subseteq \Omega_{t-1} \cap G_{r_t}$, and since $E_{t-1}, \mathbf{1}_{\Omega_t}, b_t, \lambda_t$ are all $\mathcal{F}_t^{\mathrm{pre}}$-measurable, Equation (5) gives

$$\mathbb{E}[\widetilde{E}_t \mid \mathcal{F}_t^{\mathrm{pre}}] = E_{t-1}\mathbf{1}_{\Omega_t}\big\{1 + \lambda_t(\mathbb{E}[M_t \mid \mathcal{F}_t^{\mathrm{pre}}] - b_t)\big\} \le E_{t-1}\mathbf{1}_{\Omega_t}.$$

Taking the conditional expectation given $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t^{\mathrm{pre}}$ and using $\Omega_t \subseteq \Omega_{t-1}$ with $\Omega_{t-1} \in \mathcal{F}_{t-1}$ yields

$$\mathbb{E}[\widetilde{E}_t \mid \mathcal{F}_{t-1}] \le E_{t-1}\,\mathbb{E}[\mathbf{1}_{\Omega_t} \mid \mathcal{F}_{t-1}] \le E_{t-1}\mathbf{1}_{\Omega_{t-1}} = \widetilde{E}_{t-1},$$

so the killed process is a nonnegative supermartingale with respect to $(\mathcal{F}_t)_{t \ge 0}$.

Let $\Omega_\infty = \bigcap_{t \ge 1} \Omega_t$. On this event $\widetilde{E}_t = E_t$ for every t, so

$$\Big\{\sup_t E_t \ge 1/\delta_e\Big\} \;\subseteq\; \Big\{\sup_t \widetilde{E}_t \ge 1/\delta_e\Big\} \;\cup\; \Omega_\infty^c.$$

Ville's inequality applied to $\widetilde{E}_t$, which starts at $\widetilde{E}_0 = 1$, bounds the first probability by δ_e. Moreover $\Omega_\infty^c = \bigcup_{t \ge 1} G_{r_t}^c \subseteq \bigcup_{r \ge 1} G_r^c$, so Lemma B.1 bounds the second by δ_{cal}. Adding the two bounds proves the theorem. ■

Remark B.2 (Total target level). For a desired total false-alarm level δ, choose $\delta_e + \delta_{\mathrm{cal}} \le \delta$ and use the threshold $1/\delta_e$. Using $1/\delta$ when $\delta_{\mathrm{cal}} > 0$ would overspend the target. The same decomposition shows that the union $\{\sup_t E_t \ge 1/\delta_e\} \cup \Omega_\infty^c$ itself has probability at most $\delta_e + \delta_{\mathrm{cal}}$, which is the quantity measured directly in Appendix F.3.

Remark B.3 (Anytime validity, precisely). "Anytime-valid" here refers to this time-uniform false-alarm guarantee for evidence under the certified deployment null. It claims neither input-conditional prediction-set coverage nor time-uniform coverage at every stopping time.

B.3 Proof of Proposition 4.1

Let the positive integer n_r and the order $k_r \in \{1, \dots, n_r + 1\}$ be $\mathcal{H}_r$-measurable. For $1 \le k_r \le n_r$, let γ_r also be $\mathcal{H}_r$-measurable with $0 < \gamma_r \le \delta_r^{\mathrm{cal}}$. The main text states the nondegenerate case; the complete definition also allows $k_r = n_r + 1$:

$$q_r = \begin{cases} S_{r,(k_r)}, & 1 \le k_r \le n_r, \\ +\infty, & k_r = n_r + 1. \end{cases}$$

When $k_r = n_r + 1$ the answer set is all of $\mathcal{Y}$ and the conditional miss probability is zero; to satisfy the strict requirement $0 < b_t < 1$ in this degenerate case, use any predeclared $b_t = \varepsilon \in (0, 1)$, the sure event $G_r = \Omega$ and the allocation $\delta_r^{\mathrm{cal}} = 0$.

Otherwise, condition on $\mathcal{H}_r$ and let F_r be the distribution function of the true-answer score $s_{\theta_r}(X, Y)$ of a fresh calibration question, for the already-selected state. Let $V_j \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(0, 1)$ be conditionally independent of the scores and define the randomized probability-integral transforms

$$U_j = F_r(S_{r,j}^-) + V_j\big\{F_r(S_{r,j}) - F_r(S_{r,j}^-)\big\}, \qquad F_r(s^-) := \lim_{z \uparrow s} F_r(z).$$

Then the U_j are i.i.d. uniform on $(0, 1)$ and $U_j \le F_r(S_{r,j})$ for every j. At least k_r of the calibration scores satisfy $S_{r,j} \le S_{r,(k_r)}$, and F_r is nondecreasing, so at least k_r of the U_j are at most $F_r(S_{r,(k_r)})$; hence $U_{(k_r)} \le F_r(S_{r,(k_r)})$. Therefore

$$p_r = 1 - F_r(q_r) \;\le\; 1 - U_{(k_r)}, \qquad 1 - U_{(k_r)} \sim \mathrm{Beta}(n_r + 1 - k_r,\, k_r),$$

using $U_{(k)} \sim \mathrm{Beta}(k, n + 1 - k)$ and $1 - \mathrm{Beta}(a, b) \sim \mathrm{Beta}(b, a)$. Writing u_r for the $(1 - \gamma_r)$ quantile of that Beta law gives the exact certificate

$$\mathbb{P}(p_r > u_r \mid \mathcal{H}_r) \;\le\; \gamma_r \;\le\; \delta_r^{\mathrm{cal}}. \tag{7}$$

For continuous F_r, $p_r = 1 - U_{(k_r)}$ and the conditional failure probability at the stated quantile equals γ_r; atoms and ties give the displayed inequality and can make the bound conservative. The Beta quantile has an equivalent binomial form, usually the more convenient one to compute with:

$$\mathbb{P}\big\{\mathrm{Beta}(n + 1 - k, k) > u\big\} = \sum_{j=k}^{n} \binom{n}{j} (1 - u)^j u^{\,n-j},$$

so u_r is the smallest u making the right-hand side at most γ_r. Setting $G_r = \{p_r \le u_r\}$ and $b_t = u_r$ throughout the matching epoch then gives $\mathbb{P}(G_r^c \mid \mathcal{H}_r) \le \delta_r^{\mathrm{cal}}$, which is the first condition in Equation (4). ■

Two points deserve emphasis. Equation (7) is a high-probability bound, over the fresh calibration draw, on the calibration-conditional risk; a joint marginal coverage statement over both calibration and one test observation is not enough to make each bet's payoff nonpositive in conditional mean, which a sequential test needs. And reuse throughout an epoch also requires Assumption 1: each future deployed score must obey the selected score law conditionally on the pre-query history, for which conditionally i.i.d. sampling of calibration and future scores is a simple sufficient condition. Exchangeability without the corresponding conditional statement is not sufficient after arbitrary history-dependent interventions. This view, a bound that holds with high probability over the calibration draw, is closely related to sample-conditional split-conformal coverage (Duchi, 2025), and for a fixed state it is the estimand controlled by risk-controlling prediction sets (Bates et al., 2021); sequential reuse further needs the pre-query conditional score law, and adaptive refresh needs the trajectory-level allocation of Lemma B.1.

C Corollaries and operational variants

C.1 Frozen reuse, post-selection construction and certified libraries

Corollary 4.3(a): frozen reuse. Suppose one state and one certificate G_1 are reused for every query. Under the remaining conditions of Theorem 4.2 (including (5), admissible predictable bets and sticky wealth), choose a deterministic allocation $\delta_1^{\mathrm{cal}} \le \delta_{\mathrm{cal}}$ with $\mathbb{P}(G_1^c \mid \mathcal{H}_1) \le \delta_1^{\mathrm{cal}}$. The proof of Appendix B.2 applies verbatim with a single attempt and gives $\mathbb{P}(\sup_t E_t \ge 1/\delta_e) \le \delta_e + \delta_1^{\mathrm{cal}} \le \delta_e + \delta_{\mathrm{cal}}$. The calibration term does not grow with the number of queries or inspections that retain that state.

Resetting the monitor, legally. Sticky continuation is the default because it costs nothing extra and keeps the evidence gathered before a refresh, but a deployment that wants a fresh test after each intervention can have one. Split the evidence budget into per-restart allocations $\delta_e^{(m)}$, each fixed before its process starts, with $\sum_m \delta_e^{(m)} \le \delta_e$ almost surely, and start a new wealth process $E^{(m)}$

at each reset. Kill every $E^{(m)}$ on the certificate events as in Appendix B.2: from its start, each killed process is a nonnegative supermartingale, so Ville's inequality applied conditionally on the information at that start bounds its crossing probability by $\delta_e^{(m)}$. Summing over m, and charging certificate failures once through Lemma B.1, gives $\mathbb{P}\big(\exists m : \sup_t E_t^{(m)} \geq 1/\delta_e^{(m)}\big) \leq \delta_e + \delta_{\mathrm{cal}}$. Both variants are safe; what is not safe is restarting *without* allocating, which is the failure mode Section 4.3 lists.

Corollary 4.3(b): arbitrary post-selection construction. The state θ_r may be produced by any randomized, history-dependent training, model-selection, retriever-selection or corpus-selection procedure that is $\mathcal{H}_r$-measurable. Once that single state is selected, fresh calibration incurs no multiplicity cost for the states considered before selection, because Proposition 4.1 conditions on $\mathcal{H}_r$ and the selection is $\mathcal{H}_r$-measurable.

Corollary 4.3(c): simultaneously certified library. If selection is to occur *after* viewing the fresh block, predeclare a finite $\mathcal{H}_r$-measurable library $\Theta_r = \{\theta_{r,1}, \ldots, \theta_{r,L_r}\}$. For each member construct a matching $\mathcal{K}_r$-measurable event $G_{r,j}$ on which its threshold and allowance satisfy (5) whenever that member is deployed, and choose $\mathcal{H}_r$-measurable allocations $\delta_{r,j}$ before viewing the block, with

$$\mathbb{P}(G_{r,j}^c \mid \mathcal{H}_r) \leq \delta_{r,j}, \qquad \sum_r \sum_{j=1}^{L_r} \delta_{r,j} \leq \delta_{\mathrm{cal}} \quad \text{a.s.}$$

The conditional union bound makes $G_r = \bigcap_j G_{r,j}$ a valid certificate event with allocation $\delta_r^{\mathrm{cal}} = \sum_j \delta_{r,j}$, so Lemma B.1 and the killed-process argument apply unchanged. The controller may then inspect the block and deploy any library member with its matching certificate. Same-block selection across uncorrected pointwise certificates thus has two valid alternatives: either select one state before calibration and pay no candidate multiplicity cost, or certify a predeclared finite library and select afterward while paying an explicit within-library allocation.

C.2 Cumulative envelope

Corollary 4.3(d): cumulative envelope. Under the assumptions of Theorem 4.2, let $Z_t = M_t - b_t$ and $A_t = \sum_{s=1}^t Z_s$. For a fixed $\eta > 0$ define $L_t(\eta) = \exp\{\eta A_t - \eta^2 t/8\}$. On the active certificates, Z_t has conditional mean at most zero and lies in an interval of length one, so the conditional Hoeffding lemma gives

$$\mathbf{1}_{\Omega_t}\, \mathbb{E}\big[e^{\eta(M_t - b_t)} \mid \mathcal{F}_t^{\mathrm{pre}}\big] \leq \mathbf{1}_{\Omega_t}\, e^{\eta^2/8},$$

and the tower argument of Appendix B.2 makes $L_t(\eta)\mathbf{1}_{\Omega_t}$ a nonnegative supermartingale. For a deterministic finite grid $\eta_1, \ldots, \eta_J > 0$ and weights $w_j > 0$ summing to one, the mixture $L_t^{\mathrm{mix}} = \sum_j w_j L_t(\eta_j)$ is again a nonnegative supermartingale when killed. Define the boundary $h_t(\delta_e)$ as the unique solution of

$$\sum_{j=1}^J w_j \exp\Big\{\eta_j h - \frac{\eta_j^2 t}{8}\Big\} = 1/\delta_e.$$

On Ω_∞ a crossing of A_t through $h_t(\delta_e)$ is a crossing of the killed mixture through $1/\delta_e$, so Ville's inequality followed by the split on Ω_∞ and Lemma B.1 gives $\mathbb{P}\{\exists t \geq 1 : A_t \geq h_t(\delta_e)\} \leq \delta_e + \delta_{\mathrm{cal}}$. Consequently, on the complementary event, every almost surely finite $(\mathcal{F}_t)$-stopping time $\tau \geq 1$ satisfies

$$\frac{1}{\tau}\sum_{t=1}^{\tau} M_t \;\leq\; \frac{1}{\tau}\sum_{t=1}^{\tau} b_t + \frac{h_\tau(\delta_e)}{\tau}.$$

Computing this finite-mixture boundary exactly avoids relying on a stitched closed form; mixture boundaries and time-uniform supermartingale bounds are standard (Howard et al., 2020; 2021). The Hoeffding step is, however, loose: replacing it with an empirical-Bernstein boundary (Howard et al., 2021), which adapts to the realized variance of $M_t - b_t$ rather than to its range, can tighten the envelope in low-variance regimes while retaining time-uniform validity. We keep the Hoeffding form because it is self-contained and its boundary is computed exactly, and flag the substitution as a natural

refinement. Where the alarm reports *whether* the allowance has been exceeded, the envelope bounds, at every stopping time, how far realized misses can run above it by chance. For a joint guarantee at the same total budget, allocate $\delta_e^{\mathrm{alarm}} + \delta_e^{\mathrm{env}} \le \delta_e$ to their two crossing events; the shared calibration event still costs δ_{cal} only once.

C.3 A LIMITED DELAYED-FEEDBACK RESULT

Corollary 4.3(e): exogenous delayed feedback. Theorem 4.2 assumes M_t is available before a later action uses it. A narrower feedback-arrival result covers exogenous delays. Fix one state, a certificate $G_1 \in \mathcal{K}_1$ and a $\mathcal{K}_1$-measurable allowance $b \in (0,1)$; let $\delta_1^{\mathrm{cal}} \le \delta_{\mathrm{cal}}$ be deterministic with $\mathbb{P}(G_1^c \mid \mathcal{H}_1) \le \delta_1^{\mathrm{cal}}$. Conditional on the completed calibration field $\mathcal{K}_1$, assume the misses are mutually independent with $\mathbf{1}_{G_1}\mathbb{E}[M_t \mid \mathcal{K}_1] \le \mathbf{1}_{G_1} b$ for every t, and assume the complete nonnegative integer-valued delay schedule $(\ell_t)_{t\ge 1}$ is conditionally independent of the entire miss sequence given $\mathcal{K}_1$. Let

$$\mathcal{B}_u = \{t : t + \ell_t = u\}, \qquad \mathcal{R}_u = \mathcal{K}_1 \vee \sigma((\ell_t)_{t\ge 1}) \vee \sigma(M_t : t + \ell_t \le u).$$

Before the batch $\mathcal{B}_u$ is observed, select an $\mathcal{R}_{u-1}$-measurable $\lambda_u \in [0, 1/b]$ and hold it fixed within that batch. Define $E_0^{\mathrm{rev}} = 1$, $E_u^{\mathrm{rev}} = \prod_{v\le u}\prod_{t\in\mathcal{B}_v}\{1 + \lambda_v(M_t - b)\}$ and $\widetilde{E}_u^{\mathrm{rev}} = E_u^{\mathrm{rev}}\mathbf{1}_{G_1}$. Batch membership, G_1, b and λ_u are all measurable before the batch, so conditional independence factorizes the batch expectation and every factor has conditional expectation at most one:

$$\mathbb{E}[\widetilde{E}_u^{\mathrm{rev}} \mid \mathcal{R}_{u-1}] = E_{u-1}^{\mathrm{rev}}\mathbf{1}_{G_1}\prod_{t\in\mathcal{B}_u}\big\{1 + \lambda_u(\mathbb{E}[M_t \mid \mathcal{K}_1] - b)\big\} \le \widetilde{E}_{u-1}^{\mathrm{rev}}.$$

Every factor is nonnegative because $\lambda_u \le 1/b$. Hence $(\widetilde{E}_u^{\mathrm{rev}})_{u\ge 0}$ is a nonnegative supermartingale in reveal time that is constant between nonempty reveal batches, and Ville's inequality with the certificate split gives $\mathbb{P}\{\sup_u E_u^{\mathrm{rev}} \ge 1/\delta_e\} \le \delta_e + \delta_1^{\mathrm{cal}} \le \delta_e + \delta_{\mathrm{cal}}$.

In words: labels that arrive late do not break the guarantee, provided the lateness has nothing to do with the outcomes. Run the wealth in the order the labels arrive rather than the order the queries were served, fix the bet before each arriving batch, and the same budget covers the trajectory. Delay slows detection but leaves validity intact, as the delay-law study of Appendix F.9 shows. The case to avoid is outcome-dependent missingness, where lateness depends on whether the query was a miss: there the factorization above fails, and the corollary no longer applies. It also does not cover adaptive state changes or a bet changed inside a feedback-arrival batch.

C.4 CERTIFICATE-PRESERVING CONTROLLER

Corollary 4.3(f): certificate-preserving controller. The statement is immediate from the theorem once the action class is named. Case (i) of Appendix A.3 leaves r_t, G_{r_t} and the epoch unchanged and selects an action the current certificate covers uniformly, so (5) holds for the realized a_t and no new allocation is spent. Case (ii) opens a new attempt, which Lemma B.1 already budgets, and completes it before the next affected query so that the new G_r, q_r, b_t and λ_t are $\mathcal{F}_t^{\mathrm{pre}}$-measurable. In both cases the hypotheses of Theorem 4.2 hold at every query, so its conclusion is unchanged. Inside this class, adaptivity is restricted only by pre-query measurability, $0 < b_t < 1$, the pathwise calibration budget, and state–certificate matching. The practical reading is that the constraint is on timing rather than on ambition. A controller may use everything it knows (the wealth, the miss history, realized costs, its own randomization) to decide what to do next, and may change the model, the retriever, the corpus, the threshold or the bandwidth on that basis; what it may not do is use the query it is about to answer, or deploy a state before the certificate that covers it exists.

D FC-RAG INSTANTIATION, TRAINING AND BETTING

D.1 VALID COMMUNICATION DESIGNS

At attempt r, place the implemented student, retrievers, corpora, candidate rule, local score, hub aggregation, quantizer and threshold rule inside θ_r. There are two direct ways to certify communication.

1. *Direct calibration.* Calibrate the end-to-end compressed score at a fixed communication action. Quantization and aggregation are then already part of the score law, so Proposition 4.1 needs no separate compression-failure event. This is the self-contained default of the Qwen studies and the model swap, and it certifies one communication action per epoch; the resource-policy studies of Appendices F.9 and F.10 instead let the allowance vary with bandwidth.
2. *Certified action class.* A separately verified static FC-RAG certificate supplies a base allowance c_r, and a uniform predictable communication term Δ_{comm} supplies $b_t = c_r + \Delta_{\mathrm{comm}}(a_t) < 1$ for every allowed action a_t. If γ_r denotes the calibration-certificate allocation and ρ_r the allocation charged to any stochastic reconstruction-error bound, require $\gamma_r + \rho_r \le \delta_r^{\mathrm{cal}}$.

A third option is a finite simultaneously certified action library (Corollary 4.3(c)). Changing bandwidth without a uniform communication term or such a library requires a fresh calibrated epoch.

Why the communication term shrinks as $K^{-1/2}$. The communication term Δ_{comm} of Section 2 comes from the dither calculation inherited from FC-RAG. Under the subtractively dithered quantizer of FC-RAG, the hub's aggregate score decomposes as $s_t(y) = s_t^\star(y) + \bar{\xi}_t(y)$, where $s_t^\star$ aggregates the unquantized node scores and $\bar{\xi}_t = K^{-1}\sum_{i\le K}\xi_{i,t}$; the per-node reconstruction noise satisfies $\mathbb{E}[\xi_{i,t} \mid \mathcal{F}_t^{\mathrm{pre}}, X_t] = 0$, $\mathbb{E}[\xi_{i,t}^2 \mid \mathcal{F}_t^{\mathrm{pre}}, X_t] \le v(B_i) = O(2^{-2B_i/\kappa})$ for a protocol-specific constant κ, and is conditionally independent across nodes. Independence removes the cross terms, so

$$\mathbb{E}[\bar{\xi}_t^2 \mid \mathcal{F}_t^{\mathrm{pre}}, X_t] \;\le\; V_K := K^{-2}\textstyle\sum_{i\le K} v(B_i).$$

For the true-answer score and any $\mathcal{F}_t^{\mathrm{pre}}$-measurable threshold u, assume the conditional law of $s_t^\star$ given $(\mathcal{F}_t^{\mathrm{pre}}, \bar{\xi}_t)$ has density at most $f_{\max}$ between u and $u - \bar{\xi}_t$. Comparing its CDF at these two endpoints and averaging over the fresh query and the noise gives

$$\left|\mathbb{P}(s_t \le u \mid \mathcal{F}_t^{\mathrm{pre}}) - \mathbb{P}(s_t^\star \le u \mid \mathcal{F}_t^{\mathrm{pre}})\right| \le f_{\max}\,\mathbb{E}[|\bar{\xi}_t| \mid \mathcal{F}_t^{\mathrm{pre}}] \le f_{\max}\sqrt{V_K},$$

by Cauchy–Schwarz and the conditional variance bound. Thus $\Delta_{\mathrm{comm}} = f_{\max}\sqrt{V_K}$. With equal budgets $B_i \equiv B$ this is $f_{\max}\sqrt{v(B)/K}$: the $K^{-1/2}$ rate, whereas a single node gains nothing from averaging. The density clause is what converts a second-moment bound on the noise into a bound on the *miss probability*; without anti-concentration, a vanishing bound on the miss probability need not follow. Appendix F.4 checks the arithmetic numerically over $K \in \{1, 2, 4, \ldots, 128\}$.

Communication is reported either as an abstract per-query cost or in fully specified logical scalar-message units. The GPT-2 frozen-system studies of Appendix F.10 use $\sum_i B_i$, the sum of configured per-node quantizer widths, as their source-defined abstract cost. With K_t active nodes, m answer scores and B bits per quantized scalar, the logical score payload is $K_t m B$ bits per query; it excludes answer identifiers, framing, retransmission, transport and other protocol overhead unless explicitly stated. The Qwen configuration, with $K_t = 8$, $m = 10$ and $B = 12$, therefore has a logical score payload of 960 bits per query, not measured network traffic. This is an inference-side charge: a calibration refresh adds its own uplink B^{cal} on the rounds where one occurs, which the per-query figure above does not include.

D.2 Training, model replacement, and why validity is training-agnostic

Federated Probe-Logit Distillation (FPLD) is a modular training layer: nodes evaluate public probe contexts, compress and upload probe logits, and the hub aggregates those messages to distill a shared student, without receiving private examples, gradients or node model weights (Dubey & Huo, 2026). If that student is selected first and freshly calibrated, the certificate controls its deployed miss rate under the score-law conditions of Appendix B.3. Coverage is therefore agnostic to the training algorithm: FPLD quality affects accuracy, answer-set size and resource efficiency, but adds no training-error term to the validity budget.

Reusing an old threshold after replacing a model is a different problem. It requires an explicit old-to-new score-transfer theorem under additional density and training assumptions; the original FC-RAG analysis gives one possible route, whose strengthened conditional-density clause buys a two-term $f_{\max}(\varepsilon_{\mathrm{KL}} + \sqrt{2\varepsilon_{\mathrm{KL}}})$ bound on the transfer term, where $\varepsilon_{\mathrm{KL}}$ is FC-RAG's high-probability bound on the expected Kullback–Leibler divergence between the true answer distribution and the trained model. Without that clause the chain breaks, and Markov truncation leaves only a weaker

$O(f_{\max}\varepsilon_{\mathrm{KL}}^{1/4})$ rate. That would change the scale of an old-threshold transfer bound and nothing else, since neither our theorem nor any experiment in this paper relies on such a transfer.

D.3 Bettors, power, and the evidence scale

The theorem permits any predictable bettor satisfying $0 \le \lambda_t \le 1/b_t$; betting choices affect power but not validity. We use the cumulative form of approximate growth-rate-adaptive-to-the-particular-alternative (aGRAPA) betting (Waudby-Smith & Ramdas, 2024) without supplying the monitor with a post-change rate, and Appendix F.9 reports a fixed finite mixture that spreads power over effect sizes at a horizon-free threshold.

For power intuition, consider the idealized stationary alternative $M_t \overset{\text{iid}}{\sim} \mathrm{Bern}(p)$ with constant $b < p < 1$. The admissible oracle bet $\lambda^\star = (p-b)/\{b(1-b)\}$, which satisfies $\lambda^\star \le 1/b$ for every $p \le 1$, makes the one-step wealth factor equal to the Bernoulli likelihood ratio:

$$1 + \lambda^\star(M_t - b) = \Big(\frac{p}{b}\Big)^{M_t}\Big(\frac{1-p}{1-b}\Big)^{1-M_t},$$

so that, with $\mathbb{E}_p$ denoting expectation under $\mathrm{Bern}(p)$ and KL the Kullback–Leibler divergence,

$$\mathbb{E}_p \log\big\{1 + \lambda^\star(M_t - b)\big\} = p\log\frac{p}{b} + (1-p)\log\frac{1-p}{1-b} = \mathrm{KL}\big\{\mathrm{Bern}(p) \,\|\, \mathrm{Bern}(b)\big\}.$$

Thus $\log(1/\delta_e)/\,\mathrm{KL}$, the number of queries after which an oracle bettor's expected log-wealth reaches $\log(1/\delta_e)$, is the natural evidence scale. This is an oracle calculation within the admissible cap; it does not claim that the empirical aGRAPA monitor attains oracle growth, and it is not a universal finite-sample or trajectory-wide delay theorem.

An alarm is evidence that the certified conditional-mean allowance has become untenable; it is not a general distribution-shift alarm. A score distribution can move substantially while every miss stays unchanged, in which case the evidence should remain unchanged, and empirically does (Appendix F.9).

E Related work

This is the paper's related-work section; Section 1 summarizes the closest families in one sentence each, and Table 5 compares capabilities. Here Table 6 instead groups the adjacent literature by inferential target, including conformal test martingales, sequential risk monitoring, online and adaptive conformal prediction, conformal risk control, post-selection calibration, conformal e-values, non-exchangeable conformal, federated and conformal RAG, and selective acting. The last column describes each method's role in this paper rather than asserting a capability ordering.

Table 5: What each family of methods certifies; the last row is what this paper supplies, and Table 6 compares the wider literature. Columns: valid at every time at once; prices the chosen bandwidth into the certified bound; calibrates across nodes; certifies states selected from past data; adapts without voiding the certificate.

Approach	Time-uniform	Comm.-aware allowance	Federated calibration	History-sel. refresh	Cert.-pres. control
FC-RAG[a]	×	✓	✓	×	×
CTM, WATCH[b]	✓	×	×	×	×
Sequential risk monitoring[c]	✓	×	×	×	×
Online / adaptive conformal[d]	partial*	×	×	×	—§
Post-selection conformal, CAP[e]	×	×	×	partial†	×
Federated conformal[f]	×	partial‡	✓	×	×
Anytime-FC-RAG (this work)	✓	✓	✓	✓	✓

* Long-run average coverage via an adaptive step, not per-step time-uniform. † Calibrates after an adaptive pick, without budgeting a trajectory of refreshes. ‡ Compresses calibration summaries, not deployed score messages. § No certificate to preserve.

[a] Dubey & Huo (2026). [b] Vovk et al. (2021); Prinster et al. (2025). [c] Podkopaev & Ramdas (2022); Timans et al. (2025). [d] Gibbs & Candès (2021); Angelopoulos et al. (2024b). [e] Bao et al. (2025). [f] Lu et al. (2023); Plassier et al. (2023); Wen et al. (2026); Xu et al. (2025).

Table 6: Nearby methods have different inferential targets.

Method	Primary target	Relation to this work
Conformal test martingales (CTM) / weighted-conformal monitoring (WATCH) (Vovk et al., 2021; Prinster et al., 2025; Shaer et al., 2026)	Score-stream exchangeability, or adaptive and fixed-reference distribution-shift monitoring	Useful deployment monitors with different nulls; our certificate interface instead targets excess conditional miss means across history-selected score states
Sequential risk monitoring (Podkopaev & Ramdas, 2022; Timans et al., 2025)	Time-uniform evidence for deployed risk increases or bounded-risk violations	The closest fixed-state risk-monitoring target; our result additionally composes fresh certificates across history-selected federated score states
Active anytime-valid risk control (Xu et al., 2024)	Anytime-valid risk-controlled prediction sets with active label acquisition	Handles adaptive label collection under a label budget; our interface instead composes fresh state certificates across refreshes
Non-exchangeable conformal (Barber et al., 2023)	Coverage under non-exchangeability and drift	A complementary coverage target; ours is trajectory-level certificate accounting across refreshes
Sample-conditional split conformal (Duchi, 2025)	Sample-conditional split-conformal coverage	A fixed-state form of the statement, holding with high probability over the calibration draw, that Equations (4)–(5) require
Conformal e-values (Gauthier et al., 2025)	Conformal e-value constructions	A complementary conformal-inference target; our monitor tests certified conditional miss-mean nulls across state changes
Adaptive/online conformal and calibration (Gibbs & Candès, 2021; Angelopoulos et al., 2024b; Huang et al., 2026)	Long-run coverage or online calibration under shift	Adjusts prediction sets or forecasting models rather than budgeting fresh state certificates or testing a fixed conditional-mean null
Conformal risk control (Angelopoulos et al., 2024a) and risk-controlling prediction sets (Bates et al., 2021)	A general monotone risk, controlled in expectation and with high probability over the calibration draw respectively	Complementary risk-control targets, distinct from budgeting every certificate attempt and running one sticky monitor across refreshes; for a fixed state, Equations (4)–(5) require the high-probability statement that Bates et al. (2021) control
CAP (Bao et al., 2025)	Post-selection false coverage rate	Directly motivates calibration honesty after adaptive selection
Anytime-valid conformal risk control (Hultberg et al., 2026)	Risk control from a cumulatively growing held-out calibration set	A complementary calibration regime, distinct from separately budgeted fresh-state certificates
Conformal selective acting (Khosravi & Huo, 2026)	Threshold-specific selective-action risk	A complementary selective-decision problem rather than federated score-state refresh
Federated conformal / conformal RAG (Lu et al., 2023; Plassier et al., 2023; Wen et al., 2026; Xu et al., 2025; Li et al., 2024; Chakraborty et al., 2026)	Distributed calibration; RAG answer reliability	Supply federation and RAG context. The cited conformal-RAG works assume one model, corpus and calibration set; the cited federated works take the score as given. FC-RAG (Dubey & Huo, 2026) supplies the fixed-state communication layer composed here

F Experimental designs and full results

F.1 Evidence structure and reporting conventions

Three kinds of study answer the two questions Section 5 separates: whether each condition of the protocol is load-bearing, and how the resulting monitor performs. *Modular synthetic studies* answer the first: they isolate certificate multiplicity, post-selection calibration, monitoring power, predictable

dynamic allowances, delayed feedback and a resource policy, because they can violate one condition of the protocol at a time. A *cached replay* of Qwen2.5-1.5B-Instruct scores answers the second, evaluating fixed-state monitoring at a 5% total budget, and a GPT-2 *frozen-system study*, run before the Qwen studies, checks the full system end to end. One experiment swaps the deployed student, with fresh state-matched calibration and continued monitoring (Appendix F.6); no experiment refreshes the retriever or corpora, and none deploys over a network, so the remaining evidence for adaptive refresh comes from the stress tests of the theorem's interfaces. Monte Carlo evidence illustrates the theorem and its failure modes and does not substitute for its proof.

Unless a row states otherwise, "detection" means unconditional post-change detection by the stated horizon, so a pre-change false alarm counts against it. Where power is reported among pre-alarm survivors, the text says so. The risk-audit confidence statements and the monitor's false-alarm guarantee belong to separate inferential families and use separate error budgets.

F.2 What breaks when a refresh condition is dropped

The first stress test uses an exact independent-certificate model, with $50{,}000$ Monte Carlo trials as a numerical check. Table 7 reports the analytic probability of at least one certificate failure across R deployment segments. Each attempt used calibration size $n_r = 200$, target miss rate .10, and a bet cap equal to one half of the admissible $1/b_t$ cap; each epoch contained 100 queries. With .05 spent at each refresh the failure probability approaches one, whereas both trajectory allocations stay at or below .05, and a frozen certificate stays at .05 regardless of how many deployment epochs reuse it. The $50{,}000$-trial Monte Carlo check agreed with the exact values within .002 everywhere.

Table 7: Exact any-certificate-failure probability across R deployment segments at $\delta_{\mathrm{cal}} = .05$. The refresh schemes use R certificates: "unallocated" spends .05 on each, Bonferroni spends $.05/R$, and summable uses $6(.05)/(\pi^2 r^2)$. Frozen reuses one certificate. Bonferroni and summable additionally admit dependence-robust union bounds of at most .05.

R	Frozen	Unallocated	Bonferroni	Summable
1	.0500	.0500	.0500	.0304
5	.0500	.2262	.0490	.0440
20	.0500	.6415	.0488	.0478
100	.0500	.9941	.0488	.0490

The second stress test generates independent uniform score blocks for each of L candidate states, selects the smallest empirical threshold, and compares reuse of that selection block with fresh post-selection calibration. Each cell uses $50{,}000$ trials; Table 8 reports the sweep. The exact no-selection marginal miss rate for this design is .0995. At $L = 100$ the unrounded means were .1592 under reuse and .0993 after fresh calibration. Each reported value averages over calibration trials and therefore estimates marginal post-selection miss rate, not a guarantee for every realized calibration sample. These are mechanism stress tests, not estimates of any deployed model's risk; they isolate the theorem's two load-bearing certificate requirements: budget every attempt, and pick, then calibrate.

Table 8: Mean marginal post-selection miss rate over $50{,}000$ trials at target level .10.

Candidate states L	Reuse selection block	Fresh after selection
1	.100	.100
5	.125	.100
20	.142	.100
100	.159	.099

F.3 The composed guarantee under repeated refresh

Table 7 isolates the certificate side of the guarantee; its design also exercises the composed statement of Theorem 4.2 directly. On the first $5{,}000$ of the same trials, fresh Bernoulli outcomes drove one sticky wealth process across all R epochs of 100 queries each, using the epoch-specific allowance supplied by that epoch's own certificate. The scores are uniform, so the true conditional miss probability in an epoch is exactly one minus its realized threshold. Here $\delta_e = \delta_{\mathrm{cal}} = .05$, so the

composed bound is .10 and the alarm threshold is $1/\delta_e = 20$. Table 9 reports how often the monitor alarms or some certificate on the trajectory fails; Remark B.2 bounds that union by $\delta_e + \delta_{\text{cal}}$ as well.

Table 9: Rate of alarm or certificate failure over 5,000 null paths of $100R$ queries, at $\delta_e = \delta_{\text{cal}} = .05$ and composed bound .10. One sticky wealth process runs across all R epochs with epoch-specific allowances.

R	Frozen	Unallocated	Bonferroni	Summable
1	.0580	.0580	.0580	.0360
5	.0526	.2282	.0492	.0438
20	.0504	.6554	.0498	.0486
100	.0524	.9918	.0490	.0512

Every allocated scheme stayed between .0360 and .0580, well inside the composed .10 bound, at every R; at $R = 100$ the Bonferroni rate was .0490 with a pointwise 95% Wilson interval of $[.0434, .0553]$. Under the unallocated scheme the rate instead reached .9918, and the excess was driven by certificate failures rather than by the monitor, whose null alarm rate was at most .0262 anywhere in this sweep. These are Monte Carlo rates over 5,000 paths, with a standard error of about .001 for rates near one, which is why the unallocated $R = 100$ entry, .9918, sits slightly below the exact certificate-failure probability .9941 of Table 7 although its event is larger. This is a null-simulation diagnostic of the composed guarantee under an exact independent-certificate model; it illustrates Theorem 4.2 rather than substituting for its proof.

F.4 Testing each guarantee separately

A deterministic, analytically known allowance is the $\delta_{\text{cal}} = 0$ case of Section 3.2 ($G_r = \Omega$, no calibration-failure event), so the whole budget is δ_e and each guarantee can be tested on its own. All six studies below run on CPU over Bernoulli or Gaussian variance-surrogate streams with the default allowance $b = .155$, the target $\alpha = .10$ plus FC-RAG's additive slack terms for a calibration size of 200; none involves a language model. Table 10 collects them. The envelope check uses the fixed 15-point mixture $\eta_j = 2^{j/2}$, $j = -8, \ldots, 6$, with equal weights; its boundary is the numerical root in Appendix C.2.

F.5 Qwen/MMLU-Pro fixed-state monitoring

Table 11 specifies the protocol. Of the 12,032 rows in the pinned MMLU-Pro test split (Wang et al., 2024), 9,981 had exactly ten answer options and were eligible. The complete grid crossed 14 subjects with five fixed calibration/pool splits, producing 70 risk cells and six paired monitoring paths per cell. The replay assigned $\delta_{\text{cal}} = .025$ to the state certificate and $\delta_e = .025$ to evidence, so the alarm threshold was $1/\delta_e = 40$ and the per-monitor total was .05. Under that allocation the independent risk audit classified 35 cells as harmful, 32 as compatible and three as ambiguous; the five frozen allowances were .1514, .0754, .1428, .1989 and .1578.

For subjects with a unanimous audit class the separation is clean: six subjects (Computer Science, Health, History, Other, Philosophy and Psychology) are audited harmful in *all five* splits, and all six reach detection 1.000 by $T = 5{,}000$, three of them already by $T = 1{,}500$. Table 12 gives the complete fixed grid. The detection columns average all 30 paths for a subject and therefore mix its harmful and compatible cells; the relevant aggregation is by audit class, and it gives $210/210$ harmful cell–path pairs detected by $T = 5{,}000$ (.857 at $T = 1{,}500$, .957 at $T = 3{,}000$), zero alarms among the 192 compatible cell–path pairs, and zero among the 30 distinct no-drift paths, at every horizon. Ambiguous cells count as neither power nor specificity evidence; none of their 18 paths alarmed.

Discovery-disjoint confirmation. Before any confirmation question was scored, a selection rule that uses audited risks only was frozen: among the subjects harmful in all five discovery splits (the grid of Table 12), select the one whose smallest simultaneous lower-bound margin above its split-specific allowance is largest. This selected Philosophy uniquely, at a minimum discovery margin of .279 against .079 for the next-best subject, without using detection outcomes. Of 350 eligible Philosophy questions, the discovery pools had used 127 distinct rows; a fixed label-free allocation seed placed 150 of the 223 untouched rows into five mutually disjoint 30-question pools with zero discovery overlap. Separately seeded 10,000-draw query/dither audits gave conditional risks from

Table 10: Each guarantee tested separately at the default $\delta_e = .05$ with a deterministic known allowance ($\delta_{\text{cal}} = 0$).

Study	Target	Result
Alarm validity	Theorem 4.2 at the exact boundary and in the interior	2,000 paths of length 5,000: alarm rate .0210 at $\mathbb{E}[M_t \mid \mathcal{F}_t^{\text{pre}}] = b_t$ and .0025 strictly inside, against the $\delta_e = .05$ budget; median and 95th-percentile $\sup_t E_t$ are 1.32 and 10.28 at the boundary; a separate interior check at $p = .15 < b$ gives .0105.
Detection and delay	Power of the same alarm under a sustained excess	500 paths of length 8,000, change at 2,000: detection .310/.996/.998/.998/.998/.998 at excess $+.02$ through $+.15$, with median post-change delay 5,076 $\rightarrow$ 687 (3,047 at $+.04$). No universal delay rate is claimed; the KL scale of Appendix D.3 is the informal guide.
Envelope coverage	Corollary 4.3(d)	Breach rates .0025/.0005 on 2,000 boundary/interior paths each (6/4,000 overall); the median maximum excursion of $A_t - h_t(\delta_e)$ is -1.68.
Communication averaging	The $K^{-1/2}$ shrinkage of Section 2	Over $K \in \{1, 2, 4, \ldots, 128\}$ at $B_i = 8$ bits, the log-log slope of the predicted Δ_{comm} against K is -0.5000 and the communication term falls from 3.91×10^{-3} to 3.45×10^{-4}, a factor $\sqrt{128} \approx 11.3$. The Gaussian variance-surrogate draws give empirical mean-absolute-noise slope -0.5002, checking the variance model.
Bettor ablation	Does power need a tuned stake?	1,000 paths: the adaptive bettor gives null .009/.002 and detection .340/.992 on small/large excess, against .017/.001 and .054/.980 for λ=.323, .036/.010 and .000/.000 for λ=1.613, and .029/.015 and .000/.000 for λ=3.226. Every fixed stake loses power and, on small excess, has a higher null rate than the adaptive bettor.
Hyperparameter sensitivity	Are α, δ_e and the bet cap delicate?	Sweeping $\alpha \in \{.05, .10, .15, .20\}$, $\delta_e \in \{.01, .025, .05, .10\}$ and the bet cap over $\{.25, .50, .75, 1.00\}\times$ the admissible $1/b_t$: null rates track δ_e (.0020/.0060/.0180/.0520) and detection stays in .814–.984, its minimum at $\alpha = .20$ with half-cap; the full cap at the default $\alpha = .10$ gives .832.

.231 to .469 (mean .400), and with two-sided exact Clopper–Pearson tails Bonferroni-allocated over all five cells at familywise level .05, every simultaneous lower bound exceeded its frozen allowance, the smallest margin being .144. As a secondary operating characteristic, all 30 newly seeded shifted paths alarmed by $T = 1{,}500$ with median and 90th-percentile delays of 149 and 200 queries, while 0/30 paired no-drift paths alarmed through $T = 5{,}000$. The five score states, thresholds, calibration blocks and pre-change pools remained frozen from discovery, and the allowances are the five listed above, at $\delta_{\text{cal}} = .025$. The 150-question evidence and the seeded query/dither draws are fresh.

Estimand and scope. At $T = 5{,}000$ the no-drift answer sets had mean size 8.63 of ten options across the five splits. The sets are therefore inefficient, and these results are monitoring evidence rather than a competitive question-answering result. They are conditional on the fixed pools and thresholds, and the 10,000 draws, taken uniformly with replacement from each fixed 30-question pool, estimate that pool's query/dither risk, not population risk over MMLU-Pro or risk over 10,000 distinct questions. The run is a static cache-replay monitoring stress test rather than measured network federation, and the grid is a deterministic replay of immutable cached scores, pools, seeds and paths.

F.6 One state-matched model swap

Every study outside this section keeps the deployed model fixed. This one swaps it. On the same Qwen2.5/MMLU-Pro stack at $K = 2$, the deployed student is replaced mid-trajectory with state-matched calibration, and one wealth process runs across the swap without being reset. Table 13 summarizes the result.

Table 11: Protocol details for the Qwen replay and the matched synthetic baselines.

Component	Fixed specification
Qwen assets	Pinned Qwen2.5-1.5B-Instruct model, retriever, MMLU-Pro test split and prompt revisions. Rows without exactly ten options were removed before splitting.
Score messages	Under each of eight subject contexts, pinned MiniLM retrieved the top three of five passages, and the shared student scored all ten choices by full-option mean token negative log-likelihood. Node scores were clipped to $[0, 50]$, independently subtractive-dithered at 12 bits and averaged; the logical payload was 960 bits/query, not measured traffic.
Threshold and allowance	In each of five fixed splits, 80 calibration examples (ten per context) set the threshold at target miss rate .10. An independent 1,000-draw pre-deployment audit under the same pre-change fixed-pool query/dither law set the allowance to a one-sided exact Clopper–Pearson upper bound at certificate-error level .025.
Shift and paths	The exhaustive design crossed 14 prespecified post-change subject concentrations with five splits and six paired paths per cell. Each subject used a fixed 30-question pool; change occurred after query 500 and the horizon was $T = 5{,}000$.
Risk classification	A separate 10,000-draw query/dither audit per split–subject cell used two-sided exact Clopper–Pearson intervals with a Bonferroni correction over both tails of all 70 cells at familywise level .05. A cell was harmful if its lower endpoint exceeded b, compatible if its upper endpoint was at most b, and otherwise ambiguous.
Monitor	The protocol-frozen cumulative aGRAPA monitor, not supplied with any post-change rate. The evidence allocation $\delta_e = .025$ gives alarm threshold 40.
Synthetic baselines	Each comparison used 2,000 paired streams, a change at query 2,000 and horizon 5,000. A full-history randomized-rank CTM and independently null-calibrated CUSUM and Shiryaev–Roberts procedures received the same streams; CUSUM/SR were also given the true post-change Bernoulli rate.

Table 12: Complete Qwen/MMLU-Pro grid at $\delta_{\mathrm{cal}} = \delta_e = .025$. "Risk" is the mean fixed-pool audited risk over the five splits; "H/C/A" counts the splits in which the independent audit called the cell harmful, compatible or ambiguous; detection averages all 30 paths for that subject.

Post-change subject	Risk	H/C/A	$T = 1{,}500$	$T = 3{,}000$	$T = 5{,}000$
Biology	.119	2/3/0	.067	.267	.400
Business	.067	0/5/0	.000	.000	.000
Chemistry	.086	0/4/1	.000	.000	.000
Computer Science	.236	5/0/0	.967	1.000	1.000
Economics	.119	1/3/1	.100	.200	.200
Engineering	.052	1/4/0	.200	.200	.200
Health	.280	5/0/0	1.000	1.000	1.000
History	.233	5/0/0	.667	.900	1.000
Law	.067	0/5/0	.000	.000	.000
Math	.061	0/4/1	.000	.000	.000
Other	.323	5/0/0	.967	1.000	1.000
Philosophy	.484	5/0/0	1.000	1.000	1.000
Physics	.102	1/4/0	.033	.133	.200
Psychology	.319	5/0/0	1.000	1.000	1.000

Protocol. Two retrieval nodes carry disjoint corpora (Biology, History), and Philosophy is held out as the drift target; the audit of Appendix F.5 classified Philosophy as harmful in all five splits of that design, so the drift target was fixed by that prior evidence rather than by anything measured here. Each of the two calibration subjects is cut into three mutually disjoint parts of $50/50/100$ items, so the calibration blocks D_1 and D_2 hold 100 items each and the served pool holds 100 per subject: D_1 calibrates the first epoch, D_2 calibrates the second, and neither is ever served. Item identity, used for these disjointness checks, is the pair (question stem, option tuple), because MMLU-Pro reuses stems across rows with different option sets. Epoch one deploys Qwen2.5-0.5B-Instruct behind a threshold and the Proposition 4.1 Beta allowance computed from D_1. The incoming Qwen2.5-1.5B-Instruct is preselected and its threshold and allowance are computed from D_2 alone before replay, then activated at the swap. The *downgrade* arms run the same trajectory with the two students exchanged. Budgets are $\delta_e = \delta_{\mathrm{cal}} = .025$, so each of the two attempts is allocated $\gamma_r = .0125$ and the specified total

budget is .05; $\alpha = .10$ gives an allowance of $u_r = .175$ at $n_r = 100$. The horizon is $T = 4{,}000$ queries, with 200 paths per arm.

Evidence selects the refresh time under drift. In the drift arms, the controller swaps on the query after wealth reaches 10% of the alarm threshold, or at query $2{,}001$, whichever comes first. The incoming state and its disjoint calibration block are preselected; the swap time uses only past evidence. Wealth triggered the swap on 186 of 200 paths in each legal drift arm, with median swap queries of $1{,}145$ and $1{,}222.5$ in the two arms. No-drift and boundary-targeted arms swap at query $2{,}001$.

Why a boundary-targeted arm is needed. The no-drift arm sits empirically far inside the allowance: a realized miss rate near .09 against an allowance of .175 leaves the predictable bettor staking on about one step in a thousand, so its wealth barely moves and a zero alarm count would reflect near-zero stakes, not validity. The *boundary* rows therefore target the allowance. Before deployment, $4{,}000$ query/dither draws estimate the miss rate in each of the calibration-subject and drift pools. The served stream then mixes the two pools with a plug-in weight chosen so that its estimated miss rate equals the epoch's allowance, clamped to $[0, 1]$. Monte Carlo fitting does not establish exact equality or certify the null inequality, so these rows are boundary-targeted stress tests. In the legal upgrade, the mean pathwise maximum of $\log E_t$ is 0.59 in the boundary-targeted arm, against 0.09 in the no-drift arm and an alarm threshold of $\log 40 = 3.689$.

Table 13: One state-matched model swap at $K = 2$ on Qwen2.5/MMLU-Pro (0.5B and 1.5B Instruct), 200 paths per row, allowance $u_r = .175$ ($\alpha = .10$, $n_r = 100$), budget .05; Miss is the mean before $\rightarrow$ after the swap. *Legal* is select–calibrate–certify–deploy; *illegal* is calibrate-then-replace, scoring D_2 with the outgoing student and deploying the incoming one behind it. *Boundary* targets the allowance by Monte Carlo mixing; *drift* lets the held-out subject take over after query 500 and the evidence trigger the refresh.

Swap	Deployment	Legal ordering			Illegal ordering		
		Alarm	95% CI	Miss	Alarm	95% CI	Miss
0.5B → 1.5B	no drift	.000	[.000, .019]	.119 → .075	.000	[.000, .019]	.119 → .054
1.5B → 0.5B	no drift	.000	[.000, .019]	.089 → .089	.000	[.000, .019]	.089 → .154
0.5B → 1.5B	boundary	.025	[.011, .057]	.176 → .175	.000	[.000, .019]	.176 → .139‡
1.5B → 0.5B	boundary	.015	[.005, .043]	.176 → .175	.010	[.003, .036]	.176 → .175
0.5B → 1.5B	drift	.925	[.880, .954]	.207 → .210	.000	[.000, .019]	.207 → .140‡
1.5B → 0.5B	drift	.955	[.917, .976]	.203 → .220	1.000	[.981, 1.000]	.203 → .303

Intervals are 95% Wilson intervals for the alarm rate over the 200 paths of that arm. The miss columns are arm means over those same paths. The retained summaries keep the arm mean rather than the per-path miss rate, so no dispersion is quoted for them; the text of this appendix instead reports the two epoch-two arm-mean thresholds. ‡ A stale threshold from the *weaker* student is too loose for the better one, so these cells stay below their allowance and alarm at .000: this conservatism is uncontrolled and is no evidence of safety.

What the rows say. Table 13 reports the six rows. *Monitoring continues across a real refresh*: boundary-targeted alarm rates are .025 and .015 with the bettor active. *State-matched calibration held the downgrade's miss rate steady across the swap*: on the same 200 paired paths, differing in which student scores D_2, the legal ordering holds a downgrade's miss rate at .089 before and after the swap, while calibrate-then-replace raises it to .154.

No interval accompanies that contrast: the retained summaries keep the arm mean and not the per-path miss rate. The retained arm-mean thresholds show the two students' different score scales. Across the four no-drift arms their epoch-two means take two values, according to which student scored D_2: 5.582 when the 0.5B student did, 4.921 when the 1.5B student did. The legal downgrade therefore deploys the 0.5B student behind a state-matched threshold averaging 5.582 and holds the miss rate at .089; calibrate-then-replace deploys the same student behind an outgoing-model threshold averaging 4.921, tighter by 0.661, and the miss rate rises to .154. The same two values, in the opposite order, explain the ‡ cells: a threshold of 5.582 inherited from the weaker student is too loose for the better one, so the estimated pool mixtures do not reach the allowance. The allowance itself is $u_r = .1753$ in every arm, so none of this is a change in the allowance, only in which student supplied the calibration scores.

The refresh did not hide the distribution shift: after the refresh under drift (triggered by the evidence on 186 of 200 paths), the refreshed state's empirical miss rate still exceeded its allowance on $200/200$ paths in both arms, and the monitor still fired on .925 and .955 of paths; in the downgrade arm the illegal ordering shortened the median post-change delay from 1,086 to 763.5 queries while driving the mean pathwise maximum $\log E_t$ from 16.8 to 97.5.

A second model pair, and what it does not reach. The same runner on an MMLU stack with DistilGPT-2 and GPT-2 ($K = 2$, blocks of 80, allowance .185) yields a legal-downgrade post-swap miss rate of .176 in the boundary-targeted arm; $198/200$ mixture estimates are clamped below the allowance. Its sole alarm occurs at query 6, before the swap. The illegal arm reaches .184; neither ordering establishes an exact boundary null. The ordering effect is present but small (.131 $\rightarrow$ .131 legal against .131 $\rightarrow$ .137 illegal, and 27 against 130 allowance breaches under drift), plausibly because the two students' score scales are much closer than the two Qwen students'. Its held-out drift subject is not independently audited, so its drift arms say nothing about detection and are not reported as a null.

Scope. This executes a refresh of the deployed *student*; the retriever, corpora and communication design are unchanged across the swap, and nothing runs on a network. Model forward passes are cached per (model, revision, node, corpus, item), so the 200-path sweep replays them rather than recomputing them, exactly as the Qwen studies of Appendix F.5 do. One trajectory per configuration is additionally executed with the cache bypassed, scoring every query as it arrives; for the Qwen swap, that run scored 8,000 option sequences in 800 language-model forward calls, including calibration, in 297 seconds, with no alarm over its 200-query horizon. Deployment draws with replacement from fixed 100-item pools, so these are conditional statements about those pools, as in Appendix F.5. Disjoint, stratified calibration blocks and separate fixed deployment pools do not by themselves establish the premise of Assumption 1 that deployed scores are conditionally i.i.d. with the calibration law; the Beta allowances here implement the certificate rule, while the model experiment tests state matching and continuous monitoring across a swap. Detection here is below that at the stronger alternatives of Appendix F.7, plausibly because the excess over the allowance is smaller; this study is evidence about *refresh*, not about power.

F.7 Matched sequential baselines

The matched synthetic comparison used 2,000 paired continuous-score streams, a change at query 2,000 and horizon 5,000. Scores were $\mathrm{Unif}(0, 1)$ before the change, the threshold was .9, and post-change scores followed $\mathrm{Beta}(a, 1)$ for $a \in \{1.25, 1.5, 2, 3, 5\}$; threshold crossings formed the binary misses. Here the null allowance was known exactly, so $\delta_{\mathrm{cal}} = 0$ and the full monitor budget was $\delta_e = .05$. CTM saw the continuous scores and mixed power martingales at exponents .1, .3, .5, .7, .9; it therefore tests exchangeability. CUSUM (Page, 1954) and Shiryaev–Roberts (Roberts, 1966) received the true post-change Bernoulli rate and used thresholds calibrated on 5,000 independent null streams at the same horizon and target level .05. Table 14 reports the resulting detection rates.

Table 14: Unconditional post-change detection on matched streams. * denotes a procedure given the true post-change Bernoulli rate and an independently null-calibrated threshold at target .05.

p_{post}	Ours	CTM	CUSUM*	SR*
.123	.603	.343	.917	.936
.146	.987	.970	.982	.980
.190	.988	.975	.982	.983
.271	.988	.975	.979	.978
.410	.988	.975	.987	.985

On a separate held-out null our monitor alarmed at .0175 (pointwise 95% Wilson interval $[.0126, .0242]$) and CTM at .0270; the oracle CUSUM/SR null rates ranged from .038 to .055 across their five separately calibrated thresholds. Oracle knowledge materially improved the near-boundary case and bought speed elsewhere: median post-change delays were 24–442 for the oracle procedures against 139–1,101 for ours over the four stronger alternatives. This is an operational comparison at each method's own stated target, not an identical-null ranking; CTM's null rate pertains to its exchangeability target.

F.8 ONLINE CONFORMAL: A DIFFERENT QUESTION ON THE SAME SCORES

Online conformal prediction is a further comparison class, and it is not a monitor: it adapts the threshold to hold long-run coverage rather than testing whether coverage has broken. Running it on the *identical* Qwen per-query scores makes the division of labor concrete. Under the no-drift sampler its post-change miss rate is $.101$ against a $.10$ target, with mean answer sets of 8.84 of ten options, against $.126$ and 8.63 for our fixed threshold. Under the audited-harmful Philosophy concentration it reports $.107$, within $.006$ of its no-drift value, by re-adjusting its threshold to sets of 8.89, while the fixed threshold that our certificate covers admits only 5.65 options and misses $.486$ of the time. On Computer Science the pattern repeats at $.102$ with sets of 8.54, against our $.236$ and 7.71; on the compatible Business shift both stay below $.10$, at $.097$ and $.067$, and our alarm stays silent.

Table 15: Online conformal on identical Qwen scores re-adjusts its threshold to hold its $.10$ target through a harmful shift, keeping sets of 8.54–8.89 of ten options where the fixed threshold's sets shrink to 5.65–7.71; its miss rate barely moves, while the certified fixed state's rises and our alarm fires. † Audited harmful.

Shift	Online conformal	Certified state	Our alarm
No drift	.101	.126	silent
Business	.097	.067	silent
Comp. Sci.†	.102	.236	fires
Philosophy†	.107	**.486**	fires

This is online conformal working as designed: it holds coverage. But its coverage output is close to invariant across a shift that roughly quadruples the fixed-state miss rate, so an operator watching only that output sees little sign that anything has changed. The alarm answers the complementary question and detects every path on both harmful subjects. The two consume one set of scores and can run together at no extra retrieval cost.

The GPT-2-small/MiniLM system of Appendix F.10 showed the same division of labor at a coarser resolution in a separately seeded head-to-head run: alarm rate $.27/1.00/.00$ on MMLU, DBpedia and AG News against online-conformal post-change miss rates $.118/.121/.108$, with our monitor firing on every DBpedia path of that run.

F.9 ADDITIONAL MODULAR STRESS TESTS

Table 16 condenses the power and operating-cost results behind the takeaways of Section 5; Tables 17 and 18 give the statistical and operational modular studies in full. All conclusions are scoped to the listed designs.

The delayed-feedback and controller studies support their deliberately limited formal counterparts; they do not justify outcome-dependent delay handling or same-query decisions. The resource-policy and node-removal studies describe observed cost–evidence tradeoffs, not causal effects of bandwidth on model quality, and no communication optimality is claimed.

F.10 GPT-2 FROZEN-SYSTEM STUDY

This study evaluated a shared GPT-2-small student with MiniLM retrieval on MMLU (Hendrycks et al., 2021), DBpedia (Lehmann et al., 2015) and AG News (Zhang et al., 2015) before the Qwen studies were run. It kept the model and retriever fixed and used three calibration splits crossed with five deployment paths per split. Its allocations were $\delta_e = \delta_{\text{cal}} = .05$, a total level of $.10$, so it is reported separately from the Qwen replay's $.025 + .025 = .05$. Figure 3 and Table 19 are an end-to-end check of the full system, not evidence for a model swap with fresh state-matched calibration.

Across the three benchmarks the evidence-triggered controller, in its own runs at $T = 3{,}000$, costs 34.4, 41.5 and 24.0 mean per-query $\sum_i B_i$ on MMLU, DBpedia and AG News at alarm rates $.33$, 1.00 and $.00$: it escalates most on DBpedia, where the shift is harmful, and never on AG News. The sudden-drift arm of the separately seeded drift-type sweep below alarms at the same $.33$, 1.00 and $.00$; both differ from the monitoring run of Table 19, which alarms on 5 of its 15 DBpedia paths.

Table 16: Power and operating cost at each study's stated budget, condensed from Tables 10, 17 and 18; each value from one completed study. *Excess* is the post-change miss rate minus the allowance; *survivor power* is power among the paths that had not already alarmed; timing schedules fix the allowance before each query; the five-scale mixture is a precommitted mixture of bettors with different half-lives; sentinels bet only on scheduled high-bandwidth queries.

Detection, delay, late labels. At excess $+.02/+.04/+.15$: detection $.310/.996/.998$, median delay $5{,}076/3{,}047/687$; survivor power .945–.955 under seven delay laws at $+.04$ and .855–.922 under six legal timing schedules at $+.02$.

Betting and sensitivity. The adaptive bettor reaches $.340/.992$ on small/large excess at null $.009/.002$, against $.054/.980$ for the best fixed stake; sweeping the target miss rate α, δ_e and the bet cap keeps the null rate tracking δ_e and detection in .814–.984; a five-scale mixture lifts mean survivor power $.925 \to .958$ and worst-cell survivor power $.627 \to .789$ at null .035 (paired unconditional gain $+.154$ at the near-boundary alternative), reaching 1.000 in all five cells at $T = 20{,}000$ on the same horizon-free threshold.

Evidence-triggered control. Escalating at 10% of the threshold adds $+.125$ detection over fixed-low at $p_{\text{post}} = .23$ (95% interval $[.064, .185]$) at 69% less cost than fixed-high, though at $p_{\text{post}} = .28$ fixed-low already detects every path and is cheapest; three sentinel controllers (periods $5, 10, 20$) add $+.112/.127/.123$ over budget-matched randomized policies at $p_{\text{post}} = .17$.

Node removal and operating cost. At a .10 total budget, with any one retrieval node removed: $0/40$ audited cells harmful, $0/240$ paths alarming, harmful Computer Science and History shifts still caught; a $250 \to 20{,}000$-draw audit cuts allowance overhead $.0397 \to .0039$ ($10.2\times$); the monitor costs 24.0 μs per update.

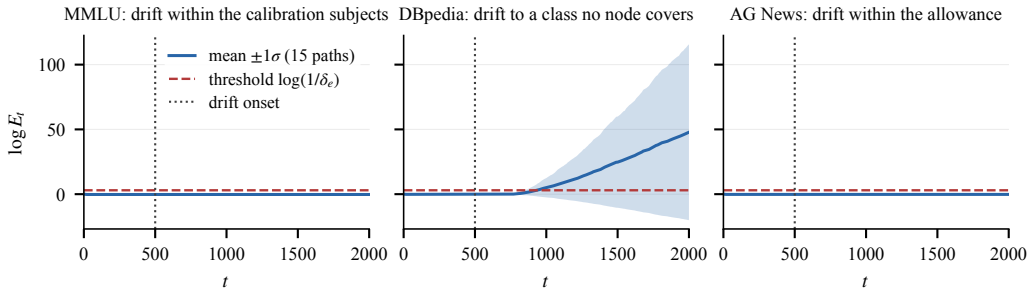


Figure 3: The alarm tracks misses against the certified allowance, not distribution shift as such. The fixed-state GPT-2-small/MiniLM system: $\log E_t$ under a sudden change at $t = 500$, alarm threshold dashed, $\pm 1\sigma$ over 15 paths. Two shifts leave the miss rate under its allowance and keep the monitor silent: redistribution *within* the calibration subjects (MMLU, .149 against a .201 allowance) and drift to the AG News class no node covers, which stays within its allowance (.127 against .346). Drift to a class no node covers (DBpedia, .370 against .252) drives the mean evidence through the threshold; 5 of its 15 paths alarm in this run (Table 19).

Three ablations on the same frozen system are worth recording, because each tracks a dependence the allowance predicts. Under *drift type* on DBpedia, alarm rates rank by severity (no drift .00, sudden 1.00, gradual .67, periodic .33): gradual onset slows evidence accumulation, and a periodic schedule returns to the calibration subjects and dilutes it. MMLU alarms at .33 under sudden drift and stays near zero otherwise, while AG News stays near zero throughout. MMLU runs nearest its allowance, so this independently seeded sweep alarms on one of its three splits where the run behind Table 19 alarms on none. Under *heterogeneous bandwidth* on MMLU, in another separately seeded sweep, alarms occur only at the highest bandwidth (.27 at $B_i \equiv 10$, and .00 for $(3, 10, 10, 10)$, $(3, 3, 10, 10)$ and $B_i \equiv 4$), which is the direction Section 2 predicts, since more bandwidth shrinks the communication term and tightens b_t. Under *multi-refresh* on MMLU, where two simulated student refreshes lower the standard deviation of a noise perturbation on the student's logits along $0.5 \to 0.25 \to 0$ at $t \in \{500, 1500\}$, the allowance falls at each refresh and the evidence growth rate adjusts to it; the final miss rate is .120 (standard deviation .027) over 15 paths. These $K \leq 4$ studies keep the model and retriever fixed, apart from that simulated perturbation, and are reported as mechanism illustrations, not as evidence for a certified model swap.

Under the scalar-message accounting of Appendix D.1, the MMLU/DBpedia/AG News monitoring rows correspond to $160/160/96$ bits per query, and the DBpedia fixed-low/adaptive/fixed-high controller rows to $160/207.26/240$ bits per query. The monitor stayed silent on MMLU and AG

Table 17: Statistical modular studies.

Question	Result
Multiscale betting	A precommitted five-component analytic mixture (half-lives 32, 128, 512, 2,048, plus cumulative; equal fixed wealth, no component, rate or horizon selection) had null alarm rate .035, mean power .958 and worst-cell power .789 among pre-alarm survivors on the fixed grid, against .027/.925/.627 for the cumulative bettor and .020/.873/.368 for CTM. At the near-boundary cell the gain is statistically significant: paired against the cumulative bettor at $p_{\mathrm{post}} = .1234$, the unconditional improvement is $+.154$ (95% Hoeffding interval $[.068, .240]$; 168 multiscale-only against 14 cumulative-only detections). The analytic threshold is horizon-free, and at $T = 20{,}000$ the mixture reached power 1.000 in all five cells.
Horizon transfer	In a nested-path study the unchanged threshold of 20 gave alarm rates .030/.036 for our monitor and .015/.015 for CTM at $T = 5{,}000$ and $T = 20{,}000$. In this nested-path sample CUSUM/SR were already at .050–.077 at their own design horizon, and once those $T = 5{,}000$ thresholds were frozen they alarmed at .193–.274 by $T = 20{,}000$; horizon-specific recalibration brought them to .039–.067 but required 2,500 null paths for each of five design rates (12,500 maximum-horizon paths, 250 million generated null observations). At the near-boundary alternative $p_{\mathrm{post}} = .1234$, extending the horizon raised our survivor power from .616 to 1.000 without being given the post-change rate.
Target specificity	Three monotone post-change score transforms had Kolmogorov–Smirnov distances .095/.225/.425 and Wasserstein-1 distances .058/.135/.243 over six million post-change scores each, while preserving every miss exactly. Our alarm times were bit-for-bit identical to the unshifted reference; the post-change alarm rate was .003 in each condition and the whole-horizon null rate remained .026, whereas CTM detected the score changes at .142/.964/.969. Harmful positive controls reverse the ordering: among pre-alarm survivors our monitor detected .634 at $p_{\mathrm{post}} = .123$ against CTM's .374, and 1.000 at every stronger alternative.
Predictable allowances	Across 2,000 paths of length 10,000, six legal pre-outcome schedules (fixed-low, fixed-high, 50-query alternating blocks, independent pre-query randomization, a lagged previous-miss rule and next-query evidence-triggered escalation) had exact-null whole-horizon alarm rates .023–.029, with Bonferroni–Clopper–Pearson simultaneous upper endpoints at most .0404. An intentionally illegal same-query rule, which selects the allowance after observing M_t, alarmed at .150 (Wilson $[.135, .166]$). At conditional excess .02, power among pre-alarm survivors was .855–.922; it was 1.000 at excess .04 and .08.
Exogenous delays	Seven delay laws (immediate; fixed 1, 5, 20, 100; $\mathrm{Unif}\{0, \ldots, 100\}$; capped geometric with mean 25) kept null alarm rates at .019–.022. At the query horizon, power among paths surviving to the first post-change feedback arrival was .945–.955 at excess .04 and 1.000 at excess .12. A fixed delay left the detection delay counted in source queries unchanged and added exactly that lag in calendar time.

News, whose measured post-change miss rates were below their source-reported mean allowances, and alarmed on 5 of 15 DBpedia paths, where the miss rate was above the allowance. On the severe DBpedia shift all three controller policies alarmed; predictable escalation reduced mean $\sum_i B_i$ by 13.6% relative to fixed-high but remained more expensive than fixed-low, so it is an intermediate operating point and not a Pareto improvement. This is an end-to-end mechanism illustration with 15 correlated paths per benchmark.

G Reproducibility, compute and licenses

Theory. Every assumption is stated in Section 2, Assumption 1 and Appendix A.2, with the covered and excluded regimes in Table 4. Lemma B.1, Theorem 4.2 and Proposition 4.1 are proved in Appendix B; Corollary 4.3 in Appendix C. No proof step is deferred and every constant is named.

Experimental provenance. Every number in Section 5 traces to a completed run, and the rendered panels are produced by scripts that read the retained numerical summaries directly, so no plotted value is recomputed at figure time. Every retained summary comes from a completed run whose protocol

Table 18: Operational and resource modular studies.

Question	Result
Synthetic resource policy	Across 2,000 paired streams with pre-change risk .13 and allowances $b_{\text{low}} = .185$, $b_{\text{high}} = .150$, policies shared the exogenous Bernoulli outcomes within each seed. At $p_{\text{post}} = .23$, fixed-low, evidence-triggered ($w = .10$, meaning escalation once the wealth reaches w/δ_e) and fixed-high had detection $.039/.163/.998$ at mean costs $1.00/1.23/4.00$; pre-drift alarms were at most .0025 throughout. The adaptive policy raised detection over fixed-low by .125 (paired 95% Hoeffding interval $[.064, .185]$; 249 adaptive-only and no low-only detections) for .234 more cost per query ($[.052, .417]$), while spending 69% less than fixed-high. At $p_{\text{post}} = .28$ fixed-low already saturates at 1.000 and is cheapest. The complete study covered four warning policies and eight drift magnitudes; bandwidth changed the allowance and the cost, not the generated errors.
Evidence-triggered sentinels	Twelve prespecified sentinel controllers updated wealth only on high-bandwidth queries at a fixed period, with the bet set to zero on low-bandwidth queries, and a warning changed bandwidth only from the next query. Across both an interior and a least-favorable boundary null, the largest of all 24 simultaneous null-alarm upper endpoints was .0401. At $p_{\text{post}} = .17$, the $w = .10$ controllers with sentinel periods $5, 10, 20$ improved unconditional detection by $.112/.127/.123$ over randomized policies whose high-bandwidth rate was matched to each controller's expected budget in an independent pilot run, with 95% intervals simultaneous over a family of 360 comparisons: $[.014, .209]$, $[.029, .224]$ and $[.025, .220]$.
Node removal	At $\delta_{\text{cal}} = \delta_e = .05$ (total .10), eight node-removal scenarios (one per node) crossed with five splits yielded 40 audited cells, none classified harmful; the paired independent risk-audit changes relative to the healthy system were only $-.00456$ to $+.00210$. Neither our monitor nor CTM alarmed on any of the 240 node-removal paths through $T = 5{,}000$, while active nodes fell from eight to seven and the logical payload from 960 to 840 bits/query; paired realized $T = 5{,}000$ path miss-rate changes lay in $[-.0042, +.0024]$. Audited-harmful Computer Science and History concentrations were still detected in every path by $T = 5{,}000$, even with the node holding that subject's corpus removed. This is evidence for the synchronous protocol under node removal only, not for asynchronous or Byzantine operation.
Audit size	At $\delta_{\text{cal}} = \delta_e = .05$ (total .10), independent audits of $250/1{,}000/5{,}000/20{,}000$ draws gave mean allowances $.1685/.1418/.1354/.1311$ and mean Clopper–Pearson overheads $.03970/.01816/.00789/.00389$, a $10.2\times$ reduction from the smallest to the largest audit. Each audit size is a standalone design, not a data-dependent stopping rule. Every cell the independent audits called compatible, and every pre-change path, remained alarm-free.
Monitor overhead	In a benchmark of six methods (the five below and online-conformal threshold adaptation) on one Intel Xeon Gold 6226 core, the cumulative aGRAPA loop processed one million updates in 24.0 seconds (24.0 μs/update); the five-component mixture took 75.0 seconds and the full-history CTM 116.1 seconds, against 0.6 and 1.7 seconds for CUSUM and Shiryaev–Roberts. Process-wide peak resident set size across the whole benchmark was about 100 MB (100,171,776 bytes). The memory figure is not attributable to any one method, and these are absolute single-implementation measurements rather than a cross-method speed ranking.

was fixed in advance; the six single-guarantee studies of Appendix F.4 are CPU-only and were run from pinned configurations, and their runner and summaries are included in the supplementary material. The modular studies (Appendices F.2–F.3, F.7–F.9) run on CPU; each fixes its seeds in its configuration file, and the tables above give its stream counts, horizons and budgets. The Qwen fixed-state studies of Appendix F.5 are cache-only replays: the base cache holds 17,744 entries and the confirmation cache 1,472 entries (1,200 fresh target vectors plus 272 prespecified controls). Both were verified by SHA-256 before and after use, opened read-only and never modified, and incurred zero cache misses and zero model forward calls; the model swap of Appendix F.6 additionally executes one trajectory per configuration with the cache bypassed. In the upstream cache audit, all 272 prespecified controls (2,720 scalar option scores) agreed exactly between Tesla V100-PCIE-

Table 19: GPT-2 frozen-system results. The monitoring study used $T = 2{,}000$ and the controller study $T = 3{,}000$, both with a configured drift index of 500; the source generator switched at $t = 500$. The monitoring post-change mean is computed over queries $t = 501, \ldots, 2{,}000$. The final column is the source implementation's mean per-query $\sum_i B_i$, an abstract bandwidth-cost unit.

Study and benchmark	Alarm rate	Mean post-change miss / allowance	Mean $\sum_i B_i$
Monitoring: MMLU	.00	.149/.201	40.00
Monitoring: DBpedia	.33	.370/.252	32.00
Monitoring: AG News	.00	.127/.346	24.00
Controller, DBpedia: fixed-low	1.00	n/a	32.00
Controller, DBpedia: adaptive	1.00	n/a	41.45
Controller, DBpedia: fixed-high	1.00	n/a	48.00

16GB and V100-PCIE-32GB scoring; the hardware claim is exact agreement on those controls, and the checks verify cache integrity rather than constituting a new model evaluation.

Compute. The synthetic sweeps run on commodity CPU nodes; the largest studies, the predictable-allowance and sentinel sweeps, dominate the cost and remain within a few hundred CPU-hours. Scoring the Qwen caches used single-GPU V100-family nodes on a shared academic cluster; the cache studies replay those scores on CPU, except the cache-bypassed model-swap trajectories of Appendix F.6, which recompute them. The GPT-2 frozen-system study used single-GPU nodes (mostly RTX 6000, with one single-seed bandwidth sweep on an A100) for at most 60 GPU-hours. No proprietary cloud or multi-node job was used. The monitoring layer itself costs 24.0 μs per update on one CPU core (Appendix F.9).

Released artifacts. The supplementary material contains the `anytime_fcrag` package, one YAML configuration per experiment, the experiment runners and retained result JSONs, the three figure-generation scripts, and a non-interactive verifier that checks the retained summaries against their recorded SHA-256 digests. Because the bundle is anonymized, its copies differ from the production artifacts in filesystem paths, package metadata and the recording host; the digests are recorded over those copies and attest to the bundle. No reported quantity differs: across the retained summaries every numeric value is identical to the production file, and the only changes are redacted strings. The two large model-score caches are not redistributed, so the Qwen numbers can be checked against their retained summaries and provenance records, but the replay cannot be rerun from the bundle alone. A de-anonymized repository link will be added on acceptance.

Licenses for existing assets. Benchmarks are used as-is, without redistribution, under their original licenses: MMLU-Pro (MIT; Wang et al., 2024), MMLU (MIT; Hendrycks et al., 2021), DBpedia ontology classification (CC BY-SA 3.0 / GFDL; Lehmann et al., 2015) and AG News topic classification (a research-use academic corpus released by the original maintainer; Zhang et al., 2015). Pretrained checkpoints are used as-is for inference: `Qwen/Qwen2.5-1.5B-Instruct` (Apache 2.0), `Qwen/Qwen2.5-0.5B-Instruct` (Apache 2.0), `openai-community/gpt2` (MIT), `distilbert/distilgpt2` (Apache 2.0) and `sentence-transformers/all-MiniLM-L6-v2` (Apache 2.0). All datasets and models are used in compliance with their terms.

Open questions. One concrete open problem is worth naming, because it is the natural strengthening of Corollary 4.3(f): a communication-optimality oracle inequality bounding a certified controller's expected cumulative uplink against the best validity-preserving controller in the same class, up to an $o(T)$ overhead. We do not attempt it. Appendix F.6 executes a swap of the deployed student; what remains untested is a retriever and corpus refresh, and a deployment over a real network rather than a cached replay. Beyond those lie observation models with outcome-dependent missing labels, formal privacy accounting for the score messages, and a decision-theoretic analysis of intervention and communication policies. None of these is needed for the guarantee proved here.