AUDITING SUBJECT-LEVEL CONFORMAL RELIABILITY IN MEDICAL TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical time-series (MedTS) classification often uses subject-independent evaluation, where models generalize to unseen patients who each contribute multiple potentially dependent observations. This creates a mismatch between the subject as the unit of deployment and the window as the unit commonly used for conformal calibration and evaluation. We conduct a systematic audit of subject-level conformal reliability in this setting. Across four EEG and ECG benchmarks, pooled evaluation can conceal substantial undercoverage for individual unseen subjects; on ADFTD, near-nominal pooled and class-stratified coverage conceal the same failure. We show that pooled coverage forms a window-count-weighted average of subject-level coverage, while pooled calibration targets a quantile of a mixture of heterogeneous subject-specific score distributions. Across our diagnostic probes, training-time conformal optimization shows inconsistent effects on classification behavior, while similarity-weighted recalibration and selection-aware inference do not consistently resolve subject-level miscoverage. Established error-driven quantile adaptation can substantially improve reliability in sequential deployment when feedback is sufficient and representative, but its effectiveness depends on within-subject feedback volume and, for a shared-state controller, stream organization, and can reverse under prediction-dependent outcome observation. These findings concern continuous online adaptation rather than a threshold calibrated once and then fixed, and show that nominal pooled reliability need not reflect the reliability experienced by individual unseen subjects.

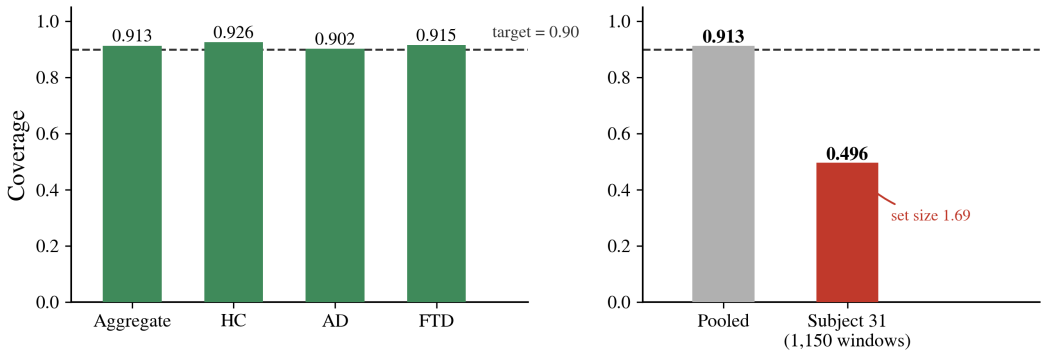
1 INTRODUCTION

Medical time-series (MedTS) classification supports clinical applications ranging from neurological assessment with electroencephalography (EEG) to cardiac diagnosis with electrocardiography (ECG) (Wang et al., 2024; Fan et al., 2025; Mu et al., 2025; Ye et al., 2026b; Li et al., 2025). MedTS data are naturally hierarchical: each patient can contribute multiple segments, trials, or recordings (Wang et al., 2023; Yuan et al., 2025; Jin et al., 2023; Zhang et al., 2023; Chambon et al., 2017). Accordingly, many studies adopt *subject-independent* evaluation, where all observations from one patient remain within a single split and models are evaluated on unseen subjects (Wang et al., 2023; 2024; Fan et al., 2025; Liu et al., 2026; Liang et al., 2024). This makes the subject the natural unit of generalization and deployment, although each subject contributes multiple prediction windows.

Conformal prediction (CP) provides finite-sample marginal coverage when the calibration and test prediction units satisfy the required exchangeability conditions (Angelopoulos & Bates, 2021; Oliveira et al., 2024). Conformal Risk Control extends conformal calibration to broader loss-based guarantees (Angelopoulos et al., 2024; Angelopoulos, 2026), while conformal training methods incorporate conformal objectives into model optimization (Stutz et al., 2022; Yeh et al., 2025). Despite these differences, their validity arguments depend on the statistical units to which exchangeability applies.

Subject-independent MedTS therefore raises two distinct issues. First, even if subject clusters are exchangeable population-level draws, observations within a subject may be dependent; subject-level exchangeability does not by itself imply exchangeability after pooling within-subject windows. Second, even when prediction-level exchangeability is sufficient for marginal conformal validity, marginal coverage controls an average over prediction units and does not guarantee comparable

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.



(a) The standard report: nominal, nothing to see here

(b) Audited patient-by-patient: confidently wrong, not uncertain

Figure 1: **Motivating example: aggregate reliability can conceal a severe patient-level failure.** On ADFTD, aggregate and class-stratified coverage remain near the 0.90 target, yet subject 31 attains only 49.6% coverage. Its average prediction-set size is 1.69, so the failure is not accompanied by inflated uncertainty sets.

reliability for every unseen subject. Our audit focuses on this second issue while explicitly separating it from the hierarchical exchangeability problem.

Motivating example. Figure 1 illustrates the discrepancy on ADFTD, a three-class EEG benchmark for dementia classification. At target coverage 0.90, the pooled conformal predictor attains 0.913 aggregate coverage, while class-stratified coverage is 0.926 for healthy controls, 0.902 for Alzheimer's disease, and 0.915 for frontotemporal dementia. These summaries provide little indication of a reliability problem.

A patient-level audit reveals a different outcome. Subject 31, an Alzheimer's patient contributing 1,150 test windows, attains only 0.496 coverage. The predictor does not compensate with large uncertainty sets: its average prediction-set size for this subject is 1.69. Thus, severe undercoverage can remain hidden by both aggregate and class-stratified reporting.

This motivates our central question:

> *When conformal reliability is calibrated and evaluated over pooled observations, what does nominal aggregate coverage imply about reliability for individual unseen subjects?*

We address this problem theoretically and empirically. Pooled empirical coverage is a window-count-weighted average of subject-level coverages, so subjects contributing more observations receive greater weight in the aggregate. At the score level, pooled calibration targets a quantile of a mixture of heterogeneous subject-specific score distributions, whereas a particular subject generally requires a different quantile to attain subject-level coverage. These results explain how nominal aggregate reliability can conceal substantial subject heterogeneity without claiming a new subject-conditional coverage guarantee (Vovk, 2013; Teneggi et al., 2023).

We next examine what information can identify or correct these failures. Training-time conformal optimization, similarity-weighted recalibration, selection-aware inference, and sequential error feedback probe pooled objectives, representation similarity, selection status, and realized coverage errors, respectively. The ConfTr probe shows no consistent cross-dataset direction in its auxiliary classification diagnostic, while similarity-weighted recalibration and selection-aware inference do not consistently resolve subject-level miscoverage (Stutz et al., 2022; Liu et al., 2024; Bao et al., 2025). By contrast, established error-driven quantile adaptation can substantially improve subject-level reliability when sufficient and representative within-subject feedback is available (Gibbs & Candès, 2021; Angelopoulos et al., 2023; Farinhas et al., 2024).

We further study four boundary conditions for feedback-driven adaptation. *Feedback sufficiency* determines how much subject-specific evidence is available, while *feedback exogeneity* determines

whether observed outcomes represent the underlying error process (Perdomo et al., 2020). *Stream organization* determines how subjects interact through a shared controller state, and *controller-state granularity* determines whether such cross-subject coupling occurs at all.

Our contributions are: we formalize the calibration–deployment unit mismatch in subject-independent MedTS and characterize how pooled calibration and window-level aggregation can conceal subject-level undercoverage. We audit this failure across four EEG and ECG datasets, with additional seed checks and cross-architecture dispersion analyses on ADFTD and second-seed checks on APAVA and PTB (Appendix B.8). We compare four information sources for subject-level correction and identify four boundary conditions for feedback-driven adaptation: feedback sufficiency, feedback exogeneity, stream organization, and controller-state granularity.

2 Problem Setup

2.1 Subject-Independent MedTS

Consider a MedTS classification dataset with subjects $\mathcal{S} = \{1, \dots, M\}$. Subject m contributes a cluster

$$D_m = \{(x_{m,i}, y_{m,i})\}_{i=1}^{n_m}, \tag{1}$$

where $x_{m,i}$ is a physiological time-series window and $y_{m,i} \in \{1, \dots, K\}$ is its class label.

We partition subjects into disjoint training, calibration, and test sets $\mathcal{S}_{\text{tr}}$, $\mathcal{S}_{\text{cal}}$, and $\mathcal{S}_{\text{test}}$, assigning all observations from the same subject to a single split. Test subjects are therefore unseen during both training and calibration.

At the population level, we treat subject clusters $D_1, \dots, D_M$ as exchangeable draws from a population of subjects, but make no exchangeability assumption about observations within a subject. Subject-level exchangeability therefore does not by itself imply exchangeability after pooling within-subject windows.

This hierarchy distinguishes two statistical units: subjects are the unit of generalization and deployment, whereas windows are the unit of individual prediction and pooled evaluation.

2.2 Pooled Conformal Calibration

Let f_θ output class probabilities $p_\theta(y \mid x)$. We use the nonconformity score

$$s(x, y) = 1 - p_\theta(y \mid x). \tag{2}$$

Our baseline pools windows from the calibration subjects,

$$\mathcal{D}_{\text{cal}} = \bigcup_{m \in \mathcal{S}_{\text{cal}}} D_m, \tag{3}$$

and computes a single global threshold. Let $s_1, \dots, s_{n_{\text{cal}}}$ denote the pooled calibration scores. For target miscoverage α, define

$$k = \lceil (n_{\text{cal}} + 1)(1 - \alpha) \rceil, \qquad \hat{q} = (s_1, \dots, s_{n_{\text{cal}}}, +\infty)_{(k)}. \tag{4}$$

The resulting prediction set is

$$C(x) = \{y : s(x, y) \le \hat{q}\}. \tag{5}$$

Conditional on the fitted model, classical split conformal prediction guarantees

$$\mathbb{P}(Y \in C(X)) \ge 1 - \alpha \tag{6}$$

when the calibration examples and a new test example satisfy the required exchangeability condition (Angelopoulos & Bates, 2021; Overman & Bayati, 2025). Subject-independent splitting alone does not establish this condition for pooled within-subject windows. We therefore keep the classical marginal coverage guarantee conceptually separate from the subject-level reliability audit below.

2.3 SUBJECT-LEVEL RELIABILITY

Let

$$M_{\text{test}} = |\mathcal{S}_{\text{test}}|, \qquad N_{\text{test}} = \sum_{m \in \mathcal{S}_{\text{test}}} n_m. \tag{7}$$

We define pooled window-level coverage as

$$\widehat{\text{Cov}}_{\text{win}} = \frac{1}{N_{\text{test}}} \sum_{m \in \mathcal{S}_{\text{test}}} \sum_{i=1}^{n_m} \mathbf{1}[y_{m,i} \in C(x_{m,i})], \tag{8}$$

and subject-level coverage as

$$\widehat{\text{Cov}}_m = \frac{1}{n_m} \sum_{i=1}^{n_m} \mathbf{1}[y_{m,i} \in C(x_{m,i})]. \tag{9}$$

For tolerance $\delta \geq 0$, we define the subject failure rate

$$R_{\text{subj}}(\delta) = \frac{1}{M_{\text{test}}} \sum_{m \in \mathcal{S}_{\text{test}}} \mathbf{1}\left[\widehat{\text{Cov}}_m < 1 - \alpha - \delta\right]. \tag{10}$$

We use $\delta = 0.10$ in the main analysis.

We additionally define average prediction-set size as

$$\widehat{\text{Size}} = \frac{1}{N_{\text{test}}} \sum_{m \in \mathcal{S}_{\text{test}}} \sum_{i=1}^{n_m} |C(x_{m,i})|. \tag{11}$$

For the motivating ADFTD analysis, we also report class-stratified coverage.

Our audit does not claim subject-conditional coverage. Instead, it measures how much subject-level undercoverage can be concealed by pooled window-level summaries.

3 METHODOLOGY

We study which information can identify and correct the subject-level reliability failures defined in Section 2. Rather than treating existing conformal methods as interchangeable baselines, we use them as diagnostic probes that access different signals: pooled training objectives, representation similarity, sequential error feedback, and selection status. Appendix C formalizes the underlying aggregation, mixture-calibration, and feedback mechanisms.

3.1 DIAGNOSTIC PROBES FOR SUBJECT-LEVEL CORRECTION

Pooled training objectives. We first probe how training-time conformal optimization changes failure-relevant classification behavior before post-hoc calibration. We use conformal training (ConfTr), which incorporates conformal set objectives into model optimization (Stutz et al., 2022). Because ConfTr optimizes an objective aggregated over the training population, its effects need not align with the calibration requirements of particular unseen subjects. In our experiments, we summarize this training-time intervention using hardest-class point-prediction recall. We therefore treat ConfTr as an auxiliary diagnostic of how pooled conformal objectives redistribute classification behavior, rather than as a direct test of subject-level conformal coverage.

Representation similarity. We next test whether representation proximity identifies calibration examples with score behavior relevant to the current test input. Let $z(x)$ denote the learned representation of x. For calibration example j, we assign

$$w_j(x) = \frac{\kappa(d(z(x), z_j)/h)}{\sum_{\ell=1}^{n_{\text{cal}}} \kappa(d(z(x), z_\ell)/h)}, \tag{12}$$

where d denotes a representation distance, κ a similarity kernel, and h controls locality. We construct the weighted empirical score distribution

$$\widehat{F}_x(q) = \sum_{j=1}^{n_{\text{cal}}} w_j(x)\mathbf{1}[s_j \le q], \tag{13}$$

and choose the local threshold

$$\hat{q}(x) = \inf\left\{q : \widehat{F}_x(q) \ge 1-\alpha\right\}. \tag{14}$$

This probe tests whether representation similarity provides useful information about subject-specific calibration requirements. We vary the locality parameter to distinguish insufficient localization from a mismatch between representation similarity and subject-specific score behavior.

Sequential error feedback. Unlike pooled objectives or representation similarity, realized errors directly indicate whether the current threshold is too small or too large for the observations entering the test stream. We therefore use an established quantile-tracking/P-control update, closely related to adaptive conformal inference, as a deliberately non-novel positive control (Gibbs & Candès, 2021; Angelopoulos et al., 2023).

We order test observations by subject so that all windows from the same subject remain contiguous. Let $t = 1, \ldots, N_{\text{test}}$ index this global stream, and let subject m occupy

$$\mathcal{I}_m = \{\tau_m, \ldots, \tau_m + n_m - 1\}. \tag{15}$$

We initialize the controller once from the pooled threshold,

$$q_1 = \hat{q}, \tag{16}$$

and construct

$$C_t(x) = \{y : s(x,y) \le q_t\}. \tag{17}$$

After observing

$$e_t = \mathbf{1}[y_t \notin C_t(x_t)], \tag{18}$$

we update

$$q_{t+1} = q_t + \eta(e_t - \alpha), \qquad t = 1, \ldots, N_{\text{test}} - 1, \tag{19}$$

where $\eta > 0$ controls the adaptation rate.

The controller maintains one running state throughout the subject-ordered stream and never resets at subject boundaries. We evaluate subject m over its contiguous segment:

$$\widehat{\text{Cov}}_m^{\text{adapt}} = \frac{1}{n_m}\sum_{t=\tau_m}^{\tau_m+n_m-1} \mathbf{1}[y_t \in C_t(x_t)]. \tag{20}$$

Subject m therefore inherits q_{τ_m} from all preceding observations and contributes n_m consecutive updates before the stream moves to the next subject.

Established quantile-tracking results motivate this update through long-run coverage control for bounded score sequences (Angelopoulos et al., 2023; Overman et al., 2024). Since $s(x,y) = 1 - p_\theta(y \mid x) \in [0,1]$, we use this controller to test whether sequential error feedback can correct subject-level reliability failures. We use this update as an established diagnostic tool rather than proposing it as a new conformal algorithm.

Selection status. Finally, we test whether information about the selection process resolves subject-level miscoverage. We use CAP as a selection-aware conformal probe (Bao et al., 2025). Selection status provides information about which observations enter the relevant selection process, whereas subject identity describes a different axis of heterogeneity. This experiment therefore tests whether accounting for selection also resolves reliability failures concentrated across subjects.

Together, these probes compare four information sources—pooled training objectives, representation similarity, sequentially realized errors, and selection status—and test how directly each source informs subject-level calibration.

3.2 Boundary Conditions for Feedback-Driven Adaptation

Sequential error feedback provides a direct signal of realized miscoverage, but its effectiveness depends on how that signal enters the controller. We distinguish four boundary conditions: *feedback sufficiency*, which controls how much subject-specific evidence is available; *feedback exogeneity*, which controls whether observed errors represent the underlying error process; *stream organization*, which determines how observations from different subjects interact through a shared state; and *controller-state granularity*, which determines whether such cross-subject interaction occurs at all.

Feedback sufficiency. Under the shared sequential controller, subject m enters its segment with the inherited state q_{τ_m} and contributes n_m observations before the stream moves on. Its influence on the threshold therefore grows with the amount of within-subject feedback. When n_m is small, the inherited state can dominate the subject's short trajectory, leaving little opportunity for realized errors to correct the threshold toward that subject's score behavior.

We isolate this effect by varying the number of available windows per subject while holding the trained model, pooled initialization, subject order, and controller parameters fixed. This intervention directly tests how within-subject feedback volume limits sequential correction.

Feedback exogeneity. Clinical workflows may reveal outcomes selectively. Let $o_t \in \{0, 1\}$ indicate whether the outcome at step t becomes observable. We then update

$$q_{t+1} = q_t + \eta o_t(e_t - \alpha). \tag{21}$$

If outcome observation depends on prediction behavior, the observed errors need not represent the errors over all predictions from the current subject. The controller then tracks the selectively observed subset rather than the subject's full error process. Appendix C.5 formalizes this bias.

Stream organization. A shared controller also couples subjects through their ordering in the test stream. In Eq. 19, subject m inherits a threshold determined by all preceding observations. Subject-contiguous ordering gives one subject several consecutive updates, whereas interleaving repeatedly mixes updates from different subjects.

We test this dependence with a round-robin stream that preserves each subject's internal window order but interleaves subjects across rounds. On ADFTD, the reduction in per-subject coverage standard deviation falls from $23.5\times$ under subject-contiguous ordering to $1.0\times$ under interleaving, while subject 31 improves only from 0.4957 to 0.5426. Appendix B.1.2 reports the full comparison. Thus, a single shared controller can depend strongly on how deployment schedules subjects.

Controller-state granularity. We distinguish two uses of error-driven quantile tracking. The *shared-state controller* is our primary positive control: it maintains one global state across the test stream and tests whether realized error feedback can correct subject-level miscoverage under the deployment order used in our main experiments. The *per-subject controller* is a diagnostic refinement used to isolate the effect of cross-subject state coupling: it assigns each subject an independent state updated only from that subject's own realized errors.

Stream-order sensitivity arises because the shared controller uses one state for all subjects. To separate this effect from the error-feedback mechanism itself, we maintain an independent controller state for each subject. For subject m, let $t_{m,1}, \ldots, t_{m,n_m}$ denote the global positions of its windows under a given schedule, in within-subject order. The prediction at $t_{m,i}$ uses

$$C_i^{(m)}(x) = \{y : s(x, y) \le q_i^{(m)}\}, \qquad e_i^{(m)} = \mathbf{1}[y_{t_{m,i}} \notin C_i^{(m)}(x_{t_{m,i}})].$$

Starting from $q_1^{(m)} = \hat{q}$, we update

$$q_{i+1}^{(m)} = q_i^{(m)} + \eta(e_i^{(m)} - \alpha), \qquad i = 1, \ldots, n_m - 1. \tag{22}$$

Because the state uses only subject m's preceding errors, its trajectory is unchanged by reordering other subjects while preserving m's within-subject order.

Accordingly, we use the shared-state controller for the main positive-control results in Table 1, and use the per-subject controller only to test whether those results depend on shared state and stream organization.

Empirically, the per-subject controller remains invariant to stream ordering and matches or improves on the shared controller under contiguous deployment. Its gains still depend on the amount of subject-specific feedback: the reduction ranges from $30.8\times$ on ADFTD, where subjects contribute hundreds of windows, to $2.8\times$ on PTB-XL, where subjects contribute about eleven. This pattern links controller-state granularity back to feedback sufficiency rather than introducing a separate correction mechanism. Appendix C.6 gives the formal construction, while Appendix B.1.2 reports the full scheduling results.

4 EXPERIMENTS

We organize the experiments around three questions. **Q1:** Does subject-level undercoverage persist across datasets? **Q2:** Which conditioning signals provide useful information for subject-level correction? **Q3:** When does feedback-driven correction succeed or fail? We answer these questions across four EEG and ECG benchmarks. Appendix B provides full per-subject results, diagnostic analyses, worst-case coverage, and robustness experiments.

4.1 EXPERIMENTAL SETUP

Datasets and protocol. We evaluate four subject-independent MedTS benchmarks: ADFTD (Miltiadous et al., 2023), APAVA (Escudero et al., 2006), PTB (PhysioBank, 2000), and PTB-XL (Wagner et al., 2020). We assign all windows from one subject to the same split, so training and calibration never contain test subjects. Unless stated otherwise, we use target coverage $1 - \alpha = 0.90$ and the nonconformity score $s(x, y) = 1 - p_\theta(y \mid x)$.

Evaluation metrics. We use pooled window-level coverage as the standard aggregate coverage summary. To expose heterogeneity hidden by marginal coverage (Vovk, 2013), we report subject-level coverage $\widehat{\mathrm{Cov}}_m$ and the subject failure rate $R_{\mathrm{subj}}(0.10)$ from Section 2. For ADFTD, we additionally report class-stratified coverage and average prediction-set size as an efficiency diagnostic. For adaptive procedures, we summarize subject-level heterogeneity using the standard deviation of per-subject coverage and examine minimum coverage separately in Appendix B.3.

Diagnostic probes. Following Section 3.1, we examine four information sources: pooled training objectives, representation similarity, sequentially realized errors, and selection status. We use ConfTr (Stutz et al., 2022), similarity-weighted recalibration, closed-loop quantile tracking, and CAP (Bao et al., 2025), respectively. These methods probe different signals and, in some cases, different evaluation quantities or scopes; we therefore do not treat Table 1 as an algorithmic leaderboard.

4.2 SUBJECT-LEVEL FAILURES PERSIST ACROSS DATASETS (Q1)

Subject-level undercoverage appears across all four datasets. At target coverage 0.90, $1/19$ ADFTD subjects, $1/4$ APAVA subjects, $13/61$ PTB subjects, and $520/3521$ PTB-XL subjects fall more than ten percentage points below the target. PTB and PTB-XL further contain subjects with zero empirical coverage. Thus, pooled evaluation alone does not characterize the lower tail of subject-level reliability. Appendix B.2 reports the full cross-dataset audit and per-subject coverage distributions.

4.3 CONDITIONING SIGNALS PROVIDE UNEQUAL INFORMATION (Q2)

Table 1 summarizes these diagnostic probes. Because ConfTr uses a different metric and CAP uses a different evaluation scope, the table does not define a common algorithmic ranking; we separate directly comparable full-population coverage-dispersion results from auxiliary diagnostics.

Pooled training objectives. ConfTr does not show a consistent direction across datasets. Under its strongest configuration, hardest-class recall decreases from 12.6% to 3.2% on ADFTD and from 47.0% to 38.1% on APAVA, but increases from 32.6% to 43.0% on PTB and from 15.1% to 24.2% on PTB-XL. Thus, changing the pooled training objective can substantially alter classification behavior, but the direction is not consistent across datasets. Because this ConfTr probe is summarized by hardest-class point-prediction recall rather than subject-level conformal coverage, we treat it as an auxiliary diagnostic rather than direct evidence of subject-level correction. Appendix B.4 reports the full ConfTr configurations.

Table 1: **Diagnostic probes for subject-level correction.** The rows probe different conditioning signals and are not directly comparable as algorithms. Panel A reports the standard deviation of per-subject coverage over the full test population at target coverage 0.90; lower is better. Panel B contains auxiliary diagnostics with different metrics or evaluation scopes: ConfTr reports hardest-class point-prediction recall, while CAP evaluates only its deferred subset. "–" denotes not run.

Method	Scope	ADFTD	APAVA	PTB	PTB-XL
A. Full-population subject-level coverage dispersion					
Naive (no adaptation)	full population	0.1200	0.2166	0.2927	0.2710
Reweighting, default τ	full population	0.0781	0.1893	0.2583	0.2670
Reweighting, $\tau = 1.5$ (ADFTD only)	full population	0.1450	–	–	–
Reweighting, $\tau = 0.75$ (ADFTD only)	full population	0.1880	–	–	–
Closed-loop, endogenous ($\eta = 0.05$)	full population	0.1978	0.2806	0.2810	0.3126
Closed-loop, exogenous ($\eta = 0.05$)	full population	**0.0051**	**0.0604**	**0.0617**	0.2044
B. Auxiliary diagnostics—different metric or scope					
ConfTr (training-time)	full population	12.6% →3.2%	47.0% →38.1%	32.6% →43.0%	15.1% →24.2%
CAP	deferred subset only	0.0804	0.0860	0.0678	0.2539

Representation similarity. Similarity-weighted recalibration reduces subject-level dispersion under the default bandwidth, but it does not consistently remove the lower tail. The standard deviation decreases from 0.1200 to 0.0781 on ADFTD, from 0.2166 to 0.1893 on APAVA, and from 0.2927 to 0.2583 on PTB, while PTB-XL changes only from 0.2710 to 0.2670.

Increasing locality does not improve this result. On ADFTD, decreasing τ from the default setting to 1.5 and 0.75 increases dispersion to 0.1450 and 0.1880, respectively. The corresponding minimum coverage also decreases from 0.7215 to 0.3788 and 0.1626. Appendix B.4.2 gives the complete analysis.

Selection status. CAP produces dataset-dependent results within its deferred subset. Because this subset differs from the full test population, we avoid comparing its absolute dispersion directly with the other rows. Its detailed analysis shows that CAP can improve some selected cases while degrading others; on PTB, the binary-class deferred-set geometry also creates a mechanical apparent repair. Appendix B.4 provides the corresponding figure and numerical results.

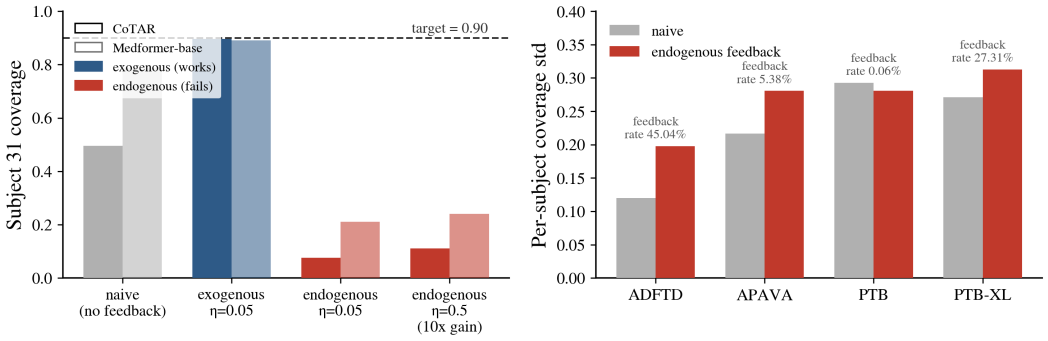
Realized errors. Sequentially observed errors provide the clearest corrective signal. Exogenous feedback reduces subject-level coverage dispersion from 0.1200 to 0.0051 on ADFTD, from 0.2166 to 0.0604 on APAVA, and from 0.2927 to 0.0617 on PTB, corresponding to $23.5\times$, $3.6\times$, and $4.7\times$ reductions. PTB-XL improves by only $1.3\times$, from 0.2710 to 0.2044.

The lower tail follows the same pattern. The companion worst-case analysis in Appendix B.3 shows that exogenous feedback raises minimum coverage from 0.4957 to 0.8910 on ADFTD, from 0.4800 to 0.8089 on APAVA, and from 0 to 0.7557 on PTB. PTB-XL remains the exception: its minimum coverage stays at 0. The full per-subject distributions appear in Appendix B.5.

The contrast with endogenous feedback already suggests that observing errors alone does not suffice. Endogenous feedback increases dispersion above the naive baseline on ADFTD, APAVA, and PTB-XL. We next study why the same sequential update succeeds in some regimes and fails in others.

4.4 When Does Feedback-Driven Correction Succeed or Fail? (Q3)

The results above show that realized coverage errors can provide a strong corrective signal. We next examine the four boundary conditions from Section 3.2: feedback sufficiency, feedback exogeneity, stream organization, and controller-state granularity.



(a) ADFTD: bigger feedback gain cannot rescue the endogenous loop

(b) Endogenous feedback weakens or reverses correction

Figure 2: **Prediction-dependent feedback can reverse error-driven correction.** (a) On ADFTD subject 31, exogenous feedback restores coverage, whereas endogenous feedback fails even with a larger adaptation gain. (b) Across datasets, endogenous feedback weakens or reverses the reduction in subject-level coverage dispersion.

Feedback sufficiency. PTB-XL provides substantially fewer observations per subject than the other datasets and shows the weakest benefit from exogenous adaptation. Varying the adaptation rate does not remove this gap. Instead, when we cap the number of ADFTD windows per subject while holding the trained model, initialization, subject order, and controller parameters fixed, the dispersion reduction falls from $23.5\times$ without capping to $9.6\times$ at 300 windows, $4.6\times$ at 150, $1.9\times$ at 30, and $1.2\times$ at approximately 11 windows. The last regime closely matches PTB-XL's natural $1.3\times$ improvement, indicating that limited within-subject feedback helps explain its weaker correction. Full adaptation-rate and window-cap results are reported in Appendix B.1.1.

Feedback exogeneity. Feedback quantity alone is insufficient when outcome observation depends on the model's predictions. We simulate this setting by withholding feedback for confident singleton predictions. Figure 2(a) shows that ADFTD subject 31 attains 0.4957 coverage without adaptation and 0.8957 with exogenous feedback, but only 0.0757 under endogenous feedback at $\eta = 0.05$. Increasing the adaptation gain tenfold raises coverage only to 0.1104, so stronger updates cannot compensate for systematically unrepresentative feedback.

Across datasets, endogenous feedback also fails to reproduce the reduction in subject-level dispersion obtained with exogenous feedback (Figure 2(b)). Dispersion increases from 0.1200 to 0.1978 on ADFTD, from 0.2166 to 0.2806 on APAVA, and from 0.2710 to 0.3126 on PTB-XL. PTB changes only from 0.2927 to 0.2810, despite feedback being observed for just 0.06% of predictions.

Additional results on stream organization, controller-state handling, and the held-out frozen-state setting are reported in Appendices B.1.2 and B.1.3.

5 Conclusion

We presented an audit of conformal reliability in subject-independent medical time-series classification. Across EEG and ECG benchmarks, pooled evaluation can conceal substantial undercoverage for individual unseen subjects, reflecting a mismatch between window-level calibration and subject-level deployment. Representation similarity and selection status do not consistently resolve this failure, while the ConfTr probe shows no consistent cross-dataset direction. Established error-driven quantile adaptation can substantially improve subject-level reliability in continuous online deployment, but its effectiveness depends on feedback volume, representativeness, and controller state. Per-subject state removes the stream-order sensitivity of a shared controller, whereas limited or prediction-dependent feedback weakens correction. These gains also do not transfer unchanged to a threshold learned from limited feedback and then held fixed. Overall, pooled conformal reliability need not reflect the reliability experienced by individual unseen subjects.

AI USE STATEMENT

Generative AI tools were used for language polishing, grammar checking, and improving clarity and readability.

REFERENCES

Henrique Aguiar, Mauro D. Santos, Peter J. Watkinson, and Tingting Zhu. Learning of cluster-based feature importance for electronic health record time-series. In *ICML*, pp. 161–179, 2022.

Jiahao Ai and Zhimei Ren. Not all distributional shifts are equal: Fine-grained robust conformal inference. In *ICML*, pp. 641–665, 2024.

Léo Andéol, Thomas Fel, Florence de Grancey, and Luca Mossina. Confident object detection via conformal prediction and conformal risk control: an application to railway signaling. In *CPPA*, pp. 36–55, 2023.

Anastasios Angelopoulos, Emmanuel J. Candès, and Ryan J. Tibshirani. Conformal PID control for time series prediction. In *NeurIPS*, pp. 23047–23074, 2023.

Anastasios N. Angelopoulos. Conformal risk control for non-monotonic losses. *CoRR*, abs/2602.20151, 2026.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *ICLR*, 2024.

Yajie Bao, Yuyang Huo, Haojie Ren, and Changliang Zou. CAP: A general algorithm for online selective conformal prediction with FCR control. *J. Mach. Learn. Res.*, 26:287:1–287:74, 2025.

Stanislas Chambon, Mathieu Galtier, Pierrick Arnal, Gilles Wainrib, and Alexandre Gramfort. A deep learning architecture for temporal sleep stage classification using multivariate and multimodal time series. *CoRR*, abs/1707.03321, 2017.

Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544):2491–2502, 2023.

J. Escudero, D. Abásolo, R. Hornero, P. Espino, and M. López. Analysis of electroencephalograms in alzheimer's disease patients with multiscale entropy. *Physiological measurement*, 27(11):1091, 2006.

Wei Fan, Jingru Fei, Dingyu Guo, Kun Yi, Xiaozhuang Song, Haolong Xiang, Hangting Ye, and Min Li. Towards multi-resolution spatiotemporal graph learning for medical time series classification. In *WWW*, pp. 5054–5064, 2025.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *ICLR*, 2024.

Feichen Gan, Youcun Lu, Yingying Zhang, and Yukun Liu. Conformal prediction beyond the horizon: Distribution-free inference for policy evaluation. In *NeurIPS*, 2025.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *NeurIPS*, pp. 1660–1672, 2021.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *CoRR*, abs/1703.07771, 2017.

Nan Huang, Haishuai Wang, Zihuai He, Marinka Zitnik, and Xiang Zhang. Repurposing foundation model for generalizable medical time series classification. *CoRR*, abs/2410.03794, 2024.

Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, Shirui Pan, Vincent S. Tseng, Yu Zheng, Lei Chen, and Hui Xiong. Large models for time series and spatio-temporal data: A survey and outlook. *CoRR*, abs/2310.10196, 2023.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: medical time series foundation model for real-world health data. In *NeurIPS*, 2025.

Hongming Li, Shujian Yu, and José C. Príncipe. Causal recurrent variational autoencoder for medical time series generation. In *AAAI*, pp. 8562–8570, 2023.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *KDD*, pp. 6555–6565, 2024.

Yi Liu, Alexander Levis, Sharon-Lise T. Normand, and Larry Han. Multi-source conformal inference under distribution shift. In *ICML*, pp. 31344–31382, 2024.

Yuanbo Liu, Xiucheng Li, Xinyang Chen, Hongwei Liu, and Zhijun Li. Toward learning shift-invariant representations for healthcare series classification. *IEEE Trans. Knowl. Data Eng.*, 38 (5):3222–3233, 2026.

Andreas Miltiadous, Katerina D. Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G. Tsalikakis, Pantelis Angelidis, Markos G. Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, and Alexandros T. Tzallas. A dataset of scalp EEG recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine EEG. *Data*, 8(6): 95, 2023.

Yang Mu, Muhammad Shahzad, and Xiao Xiang Zhu. Mptsnet: Integrating multiscale periodic local patterns and global dependencies for multivariate time series classification. In *AAAI*, pp. 19572–19580, 2025.

Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and João Vitor Romano. Split conformal prediction and non-exchangeable data. *J. Mach. Learn. Res.*, 25:225:1–225:38, 2024.

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. In *NeurIPS*, 2025.

William Overman, Jacqueline Jil Vallon, and Mohsen Bayati. Aligning model properties via conformal risk control. In *NeurIPS*, 2024.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *ICML*, pp. 7599–7609, 2020.

PhysioToolkit PhysioBank. Physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215 – e215, 2000.

Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *AAAI*, pp. 4091–4098, 2018.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *ICLR*, 2022.

Sophia Sun and Rose Yu. Conformal prediction for time-series forecasting with change points. In *NeurIPS*, 2025.

Jacopo Teneggi, Matthew Tivnan, J. Webster Stayman, and Jeremias Sulam. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *ICML*, pp. 33940–33960, 2023.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92(2-3): 349–376, 2013.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):1 – 15, 2020.

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. In *NeurIPS*, pp. 55694–55717, 2023.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *NeurIPS*, 2024.

Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, Meng Zhao, and Fugee Tsung. MedualTime: a dual-adapter language model for medical time series-text multimodal learning. In *IJCAI*, pp. 7913–7921, 2025.

Jiexia Ye, Jia Li, Chang Xu, and Fugee Tsung. MedTVL: Harnessing vision and language for medical time series classification. In *KDD*, pp. 12691–12702, 2026a.

Jiexia Ye, Weiqi Zhang, Ziyue Li, Jia Li, and Fugee Tsung. Medspaformer: A transferable transformer with multi-granularity token sparsification for medical time series classification. In *AAAI*, pp. 27791–27799, 2026b.

Christopher Yeh, Nicolas Christianson, Adam Wierman, and Yisong Yue. Conformal risk training: End-to-end optimization of conformal risk control. In *NeurIPS*, 2025.

Xiaoyan Yuan, Wei Wang, Junxin Chen, and Xiping Hu. Reading between the channels: Knowledge-augmented medical time series classification. In *MM*, pp. 8978–8987, 2025.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *KDD*, pp. 3273–3285, 2023.

Xiaofan Zhou, Baiting Chen, Yu Gui, and Lu Cheng. Conformal prediction: A data perspective. *ACM Comput. Surv.*, 58(2):49:1–49:37, 2026.

A Related Work

Subject-independent medical time-series modeling. Medical time-series research increasingly treats the patient as a meaningful unit of representation and generalization (Yuan et al., 2025; Li et al., 2025; Ye et al., 2026a; 2025; Li et al., 2023; Harutyunyan et al., 2017). `COMET` explicitly models hierarchical structure across observation, sample, trial, and patient levels (Wang et al., 2023), while `Medformer` and `MedGNN` evaluate models under subject-independent protocols that separate patients across data splits (Wang et al., 2024; Fan et al., 2025; Aguiar et al., 2022). Recent work further studies generalization across heterogeneous datasets, signal configurations, and patient populations (Huang et al., 2024; Li et al., 2025; Song et al., 2018). These settings naturally organize observations into subject-level clusters: subjects define the unit of generalization, while each subject contributes multiple potentially dependent windows. Our audit examines whether conformal reliability follows the same hierarchy when calibration and evaluation pool windows across subjects.

Conformal prediction, risk control, and adaptation. Split conformal prediction provides marginal coverage when calibration and test prediction units satisfy exchangeability (Angelopoulos & Bates, 2021; Zhou et al., 2026; Gan et al., 2025), while prior work distinguishes marginal validity from stronger forms of conditional validity (Vovk, 2013; Andéol et al., 2023). Training-conditional guarantees bound how much a *fixed* conformal procedure's miscoverage varies across independent draws of the calibration set under exchangeability (Vovk, 2013). The failures we study instead arise within a single calibration set, from applying one pooled threshold across heterogeneous subject-specific score distributions. Training-conditional bounds therefore constrain calibration-set-to-calibration-set variability, but do not by themselves address the reliability experienced by a particular unseen subject at deployment. Conformal Risk Control extends conformal calibration from miscoverage to broader bounded monotone losses (Angelopoulos et al., 2024), while conformal training and conformal risk training incorporate conformal objectives into model optimization (Stutz et al., 2022; Yeh et al., 2025). These frameworks derive their validity for the statistical units that satisfy the corresponding exchangeability conditions. In hierarchical MedTS, exchangeability

of subject-level clusters does not by itself imply exchangeability of pooled within-subject windows. This line of work directly addresses this exchangeability question, constructing valid prediction sets for two-layer hierarchical data via CDF pooling or subsampling over subjects (Dunn et al., 2023). Our work instead audits subject-level reliability even when marginal validity holds, treating exchangeability as a separate issue rather than the object of correction. Moreover, even when prediction-level marginal validity applies, it constrains an aggregate over prediction units rather than reliability for every unseen subject.

Other conformal approaches address different sources of heterogeneity. Reweighting methods adapt calibration to heterogeneous populations (Liu et al., 2024; Ai & Ren, 2024); adaptive conformal methods update calibration from sequentially observed errors (Gibbs & Candès, 2021; Angelopoulos et al., 2023; Sun & Yu, 2025); and selection-aware methods account for mechanisms that determine which observations enter calibration or provide feedback (Bao et al., 2025). These methods condition on different sources of information, but those signals need not identify subject-specific calibration requirements. Our audit instead asks a complementary question: how does hierarchical subject-independent deployment interact with the statistical units used for conformal calibration, evaluation, and adaptation?

B Additional Experimental Results

This appendix provides detailed experimental evidence for Section 4. We include the full subject-level audit, worst-case coverage, detailed analyses of the diagnostic conditioning signals, complete closed-loop results, and additional robustness checks. Unless stated otherwise, all experiments use target coverage $1 - \alpha = 0.90$.

B.1 Feedback-Driven Adaptation: Additional Analyses

The main text highlights feedback representativeness as a critical condition for successful error-driven adaptation. We provide three additional controlled analyses here: feedback sufficiency, stream organization and controller-state granularity, and evaluation after freezing the adapted state. Together, these experiments clarify when the gains from continuous online adaptation persist and when they weaken.

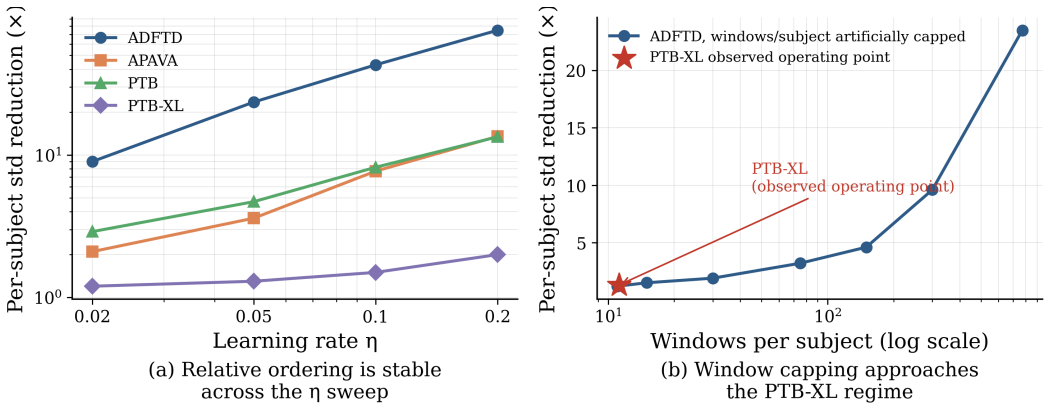
B.1.1 Feedback Sufficiency

We first examine whether the weaker exogenous-adaptation result on PTB-XL is explained by the adaptation rate or by the amount of within-subject feedback.

Adaptation-rate sensitivity. Figure 3(a) varies $\eta \in \{0.02, 0.05, 0.1, 0.2\}$ across all four datasets. The relative cross-dataset ordering remains stable throughout the sweep: PTB-XL continues to show the weakest improvement, indicating that simply increasing the update step does not remove the gap.

Within-subject feedback volume. We next vary the number of observations available from each subject while holding the dataset and trained model fixed. On ADFTD, we cap the number of windows per subject while preserving the pooled initialization, subject order, and controller parameters. Figure 3(b) shows that the benefit weakens progressively as the available feedback decreases. The reduction in subject-level coverage dispersion falls from $23.5\times$ without capping to $9.6\times$ at 300 windows, $4.6\times$ at 150, $3.2\times$ at 75, $1.9\times$ at 30, $1.5\times$ at 15, and $1.2\times$ at approximately 11 windows per subject. The final regime closely matches PTB-XL's natural $1.3\times$ reduction.

Because this intervention holds the dataset, trained model, pooled initialization, and controller fixed, it isolates feedback volume from cross-dataset differences in modality, architecture, or class structure. Under a shared-state controller, short subject segments provide fewer subject-specific updates with which to correct the inherited state. These results support feedback sufficiency as a boundary condition for effective online correction.



(a) Relative ordering is stable across the η sweep

(b) Window capping approaches the PTB-XL regime

Figure 3: **Sensitivity to adaptation rate and within-subject feedback volume.** (a) The relative cross-dataset behavior remains stable across the adaptation-rate sweep. (b) Restricting the number of ADFTD windows per subject progressively weakens correction and approaches the weak-feedback regime observed on PTB-XL.

Table 2: Subject-level coverage-dispersion reduction under blocked and interleaved scheduling. Shared-state adaptation is sensitive to stream order, whereas per-subject state removes cross-subject coupling and is invariant to scheduling.

Controller / schedule	ADFTD	APAVA	PTB	PTB-XL
Shared-state, blocked	23.5×	3.6×	4.7×	1.3×
Shared-state, interleaved	1.0×	1.5×	1.3×	1.1×
Per-subject, blocked	30.8×	8.4×	4.8×	2.8×
Per-subject, interleaved	30.8×	8.4×	4.8×	2.8×

B.1.2 Stream Organization and Controller-State Granularity

We next test whether the shared-state controller depends on the organization of subjects in the deployment stream. We compare subject-contiguous ("blocked") scheduling with interleaved scheduling and contrast the shared-state controller in Eq. 19 with the per-subject controller in Eq. 22.

Table 2 shows that the shared-state controller is strongly order-sensitive. On ADFTD, interleaving reduces the subject-level dispersion improvement from $23.5\times$ to $1.0\times$, and the improvement also weakens on APAVA, PTB, and PTB-XL. This behavior follows from cross-subject state coupling: observations from one subject modify the state subsequently used for another.

In contrast, the per-subject controller maintains a separate state for each subject and is therefore invariant to the ordering of observations across subjects. It yields reductions of $30.8\times$, $8.4\times$, $4.8\times$, and $2.8\times$ on ADFTD, APAVA, PTB, and PTB-XL, respectively, under both blocked and interleaved scheduling.

These results distinguish two related boundary conditions. *Stream organization* matters when subjects share controller state, whereas *controller-state granularity* determines whether such cross-subject dependence exists in the first place.

B.1.3 Held-Out Evaluation and Frozen Adaptation

The continuous-adaptation results evaluate subject m over windows that also generate subsequent feedback within the same subject segment. This does not create label leakage for an individual prediction: the threshold q_t in Eq. 19 is fixed before e_t is observed, so the prediction set at time t never depends on its own outcome. It nevertheless raises a separate question: does the benefit persist when adaptation is stopped before the evaluation windows are observed?

Simply evaluating the second half of a subject while continuing to update the controller does not provide a genuinely held-out evaluation, because the state used for later evaluation windows would depend on earlier outcomes from that same half. We therefore split each subject into a feedback

Table 3: **Held-out evaluation after limited subject-specific feedback.** Per-subject coverage standard deviation, with reduction relative to the matching naive baseline in parentheses. "Full segment" uses continuous online adaptation; "Held-out, frozen" snapshots the controller at the feedback/held-out boundary and performs no further updates; "Split-CP" uses the same feedback-half labels to estimate a fixed subject-specific threshold. The split-CP baseline does not depend on the controller and is reported once per dataset. †Falls below the nominal binomial sampling scale for its held-out sample size (mean $n = 5.6$) and is therefore treated as an unreliable effect estimate.

Dataset	Controller	Full segment	Held-out, frozen	Split-CP
ADFTD	Shared-state	0.0051 (23.5×)	0.0666 (1.7×)	0.0169 (6.8×)
ADFTD	Per-subject	0.0039 (30.8×)	0.0743 (1.5×)	
APAVA	Shared-state	0.0604 (3.6×)	0.0523 (4.3×)	0.0301 (7.4×)
APAVA	Per-subject	0.0257 (8.4×)	0.0386 (5.8×)	
PTB	Shared-state	0.0617 (4.7×)	0.2787 (1.0×)	0.0519 (5.6×)
PTB	Per-subject	0.0614 (4.8×)	0.2171 (1.3×)	
PTB-XL	Shared-state	0.2044 (1.3×)	0.2272 (1.2×)	0.2056 (1.4×)
PTB-XL	Per-subject	0.0956 (2.8×)	0.1169† (2.4×)	

half and a held-out half. At the transition, we snapshot the controller state and hold it fixed, with no further updates, throughout the held-out half. We additionally compare against a feedback-matched baseline: an ordinary non-sequential per-subject split-CP threshold calibrated from the same subject-specific feedback-half labels and applied fixed to the held-out half. We use this construction only as an empirical fixed-threshold baseline; because we make no within-subject exchangeability assumption, we do not invoke a split-conformal coverage guarantee for this subject-level half-split. Appendix C.6 formalizes these constructions.

The frozen state transfers only partially. On ADFTD, the dispersion reduction falls from 23.5–30.8× under continuous adaptation to only 1.5–1.7× after freezing. Subject 31 shows the same pattern: its naive held-out coverage of 0.5217 rises to 0.7774 under either frozen controller, compared with 0.9026 under the feedback-matched split-CP baseline and 0.8991 under continuous prequential adaptation over the same second-half windows. On PTB, freezing similarly reduces the benefit to 1.0–1.3×. APAVA is the main exception, where the frozen result remains comparatively strong.

The feedback-matched split-CP baseline recovers substantially more of the held-out correction on ADFTD, APAVA, and PTB, achieving $6.8\times$, $7.4\times$, and $5.6\times$ reductions, respectively. This shows that the feedback half contains useful subject-specific calibration information, but a partially adapted sequential state is not necessarily an effective static threshold once further updates are removed. PTB-XL remains difficult to interpret because its held-out segments are extremely short; the per-subject frozen estimate is below the nominal binomial sampling scale implied by the mean held-out sample size.

These results clarify the scope of the main positive-control result: the large gains from error-driven quantile adaptation characterize *continuous online deployment*. They should not be interpreted as evidence that a threshold learned from limited subject-specific feedback and then held fixed will transfer with the same strength.

B.2 Full Subject-Level Reliability Audit

Figure 4 expands the motivating ADFTD example from Figure 1 to all test subjects. Aggregate coverage reaches 0.913, yet subject-level coverage varies substantially. In particular, subject 31 attains only 0.4957 coverage.

Table 4 summarizes the lower tail across all four datasets. Every benchmark contains subjects whose empirical coverage falls more than ten percentage points below the target. PTB and PTB-XL further contain subjects with zero empirical coverage.

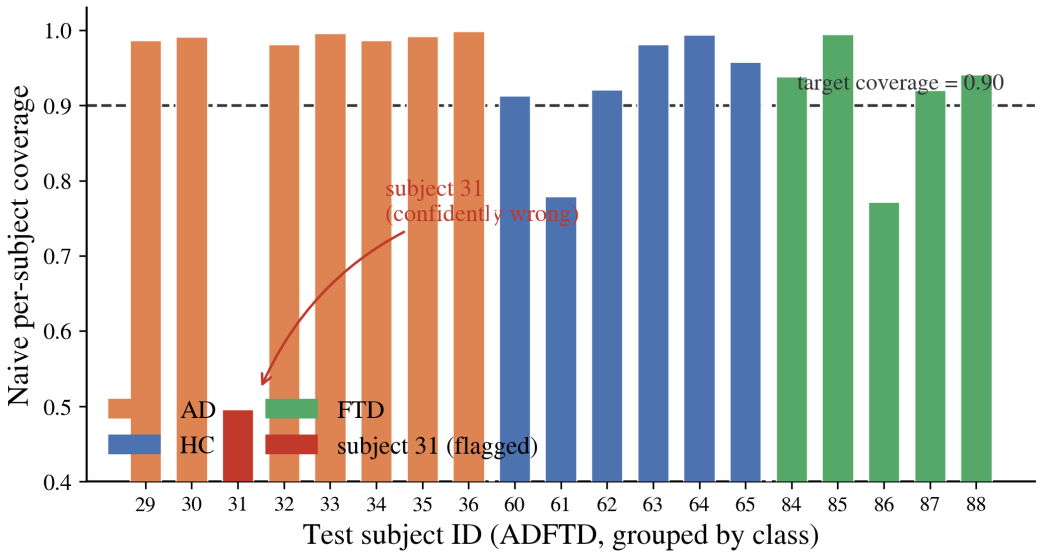


Figure 4: **Full subject-level coverage audit on ADFTD.** Per-subject coverage grouped by class at target coverage 0.90. Aggregate coverage reaches 0.913, while subject 31 attains only 0.4957 coverage.

Table 4: **Cross-dataset subject-level reliability audit.** We count subjects whose empirical coverage falls more than 0.10 below the target 0.90.

Dataset	$R_{\text{subj}}(0.10)$: subjects $>$ 10 pts below target	Worst coverage
ADFTD	1/19 (5.3%)	0.4957
APAVA[a]	1/4 (25.0%)	0.4800
PTB	13/61 (21.3%)	0.0000
PTB-XL	520/3521 (14.8%)	0.0000

[a] APAVA's test split contains only 4 subjects; we report this row as a qualitative case study rather than a population-level estimate.

B.3 Worst-Case Subject Coverage

Table 1 summarizes subject-level dispersion, but a standard deviation can still conceal an extreme lower tail. We therefore complement the main table with minimum per-subject coverage.

ConfTr does not appear in Table 5 because its diagnostic in Table 1 measures hardest-class point-prediction recall rather than subject-level coverage. CAP remains in the table for completeness, but its values apply only to the deferred subset and should not be compared directly with full-population rows.

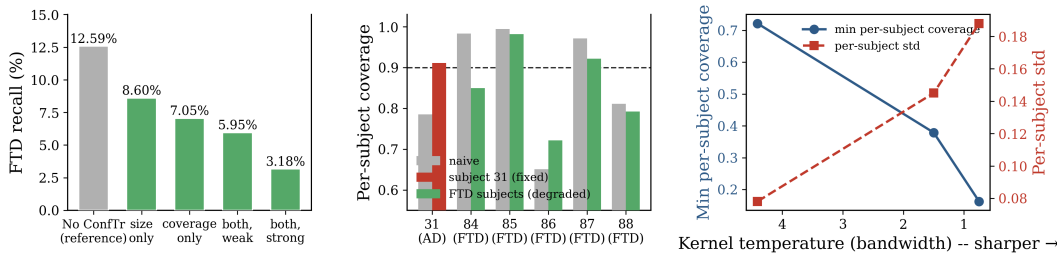
The worst-case results strengthen the conclusion from the dispersion metric. Exogenous feedback raises minimum coverage from 0.4957 to 0.8910 on ADFTD, from 0.4800 to 0.8089 on APAVA, and from 0 to 0.7557 on PTB. PTB-XL remains qualitatively different: its minimum coverage stays at zero even after adaptation.

B.4 Detailed Diagnostic Conditioning Signals

We next provide the detailed experiments behind the ConfTr, similarity-weighting, and CAP rows in Table 1.

Table 5: **Worst-case subject coverage corresponding to Table 1.** We report minimum per-subject coverage at target 0.90; higher is better. CAP applies only to its deferred subset; all other rows use the full test population. "–" denotes not run.

Method	Scope	ADFTD	APAVA	PTB	PTB-XL
Naive (no adaptation) (Angelopoulos & Bates, 2021)	full population	0.4957	0.4800	0.0000	0.0000
Reweighting, default τ (Liu et al., 2024)	full population	0.7215	0.5467	0.0000	0.0000
Reweighting, $\tau = 1.5$ (sharper, ADFTD only)	full population	0.3788	–	–	–
Reweighting, $\tau = 0.75$ (sharpest, ADFTD only)	full population	0.1626	–	–	–
CAP (Bao et al., 2025)	deferred subset only	0.7535	0.7684	0.8182	0.0000
Closed-loop, endogenous ($\eta = 0.05$)	full population	0.0757	0.3022	0.0000	0.0000
Closed-loop, exogenous ($\eta = 0.05$)	full population	**0.8910**	**0.8089**	**0.7557**	0.0000



(a) ConfTr: FTD recall collapses under every loss variant

(b) Fixes subject 31, but borrows calibration mass from the wrong class

(c) Sharper localization is monotonically worse, not better

Figure 5: **Detailed analyses of pooled and representation-based signals.** (a) Stronger ConfTr configurations substantially change hardest-class recall on ADFTD. (b) Similarity-weighted recalibration improves the flagged subject but reduces coverage for several other subjects. (c) Sharper localization decreases minimum subject coverage and increases subject-level dispersion.

B.4.1 Pooled Conformal Training

Figure 5(a) examines how increasing ConfTr strength changes classification behavior on ADFTD. FTD recall decreases from 12.59% without ConfTr to 8.60% with the size-loss configuration, 7.05% with the coverage-loss configuration, 5.95% with the weak joint objective, and 3.18% with the strongest joint objective.

The direction does not remain consistent across datasets. Under the strongest ConfTr configuration, hardest-class recall changes from 12.6% to 3.2% on ADFTD and from 47.0% to 38.1% on APAVA, but increases from 32.6% to 43.0% on PTB and from 15.1% to 24.2% on PTB-XL. Thus, the pooled training objective can substantially change classification behavior without consistently identifying subject-specific reliability failures.

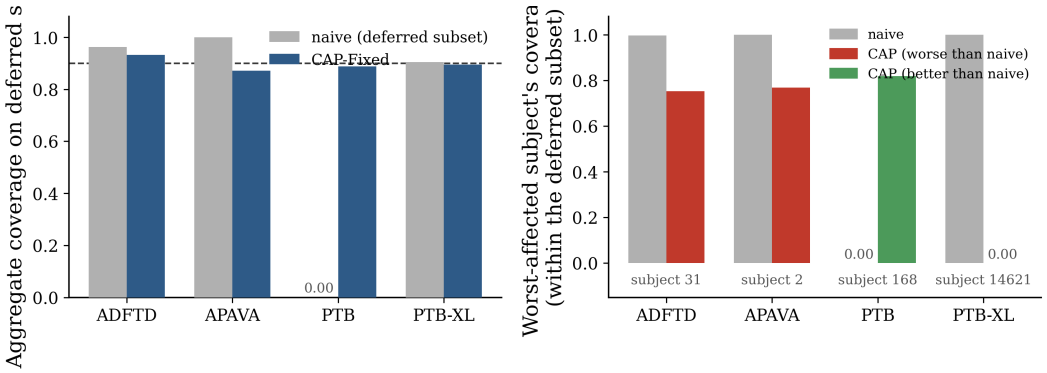
B.4.2 Representation Similarity

Figure 5(b) shows that similarity-weighted recalibration improves the flagged ADFTD subject but simultaneously reduces coverage for several other FTD subjects. Figure 5(c) tests whether stronger localization resolves this failure.

Table 6 gives the numerical ADFTD locality sweep. Sharper kernels consistently worsen both metrics over the tested settings.

Table 6: **Similarity-weighting locality sweep on ADFTD.** Smaller τ produces sharper localization. Lower standard deviation and higher minimum coverage indicate better subject-level reliability.

Kernel temperature τ	Coverage std.	Minimum coverage
4.41 (default)	0.0781	0.7215
1.50	0.1450	0.3788
0.75	0.1880	0.1626



(a) CAP's own scope: efficiency effect is inconsistent across datasets

(b) CAP helps or hurts the specific flagged subject unpredictably

Figure 6: **Selection-aware calibration has dataset-dependent effects.** (a) Aggregate coverage within CAP's deferred subset. (b) Coverage of the worst-affected subject within the same subset. CAP improves some selected cases and degrades others, indicating that selection status and subject-specific calibration requirements provide different information.

We choose the default bandwidth with the same median-pairwise-distance rule on each dataset. This rule yields numerical bandwidths 4.41, 11.63, 10.93, and 12.04 on ADFTD, APAVA, PTB, and PTB-XL, respectively. The values differ because the representation scales differ across datasets and backbones.

B.4.3 Selection Status

CAP conditions on selection status rather than subject-specific error feedback. We evaluate it only within its intended deferred subset.

Figure 6 shows dataset-dependent behavior. CAP decreases coverage for the flagged ADFTD and APAVA subjects, gives an apparent improvement on PTB, and does not resolve the selected PTB-XL failure.

Table 7 reports the corresponding values.

The PTB result requires particular caution. PTB has only two classes, so the deferred-set construction can produce empty or full prediction sets. The apparent improvement therefore largely reflects this binary-set geometry rather than a general correction of subject-level miscoverage.

B.5 Full Closed-Loop Subject-Level Results

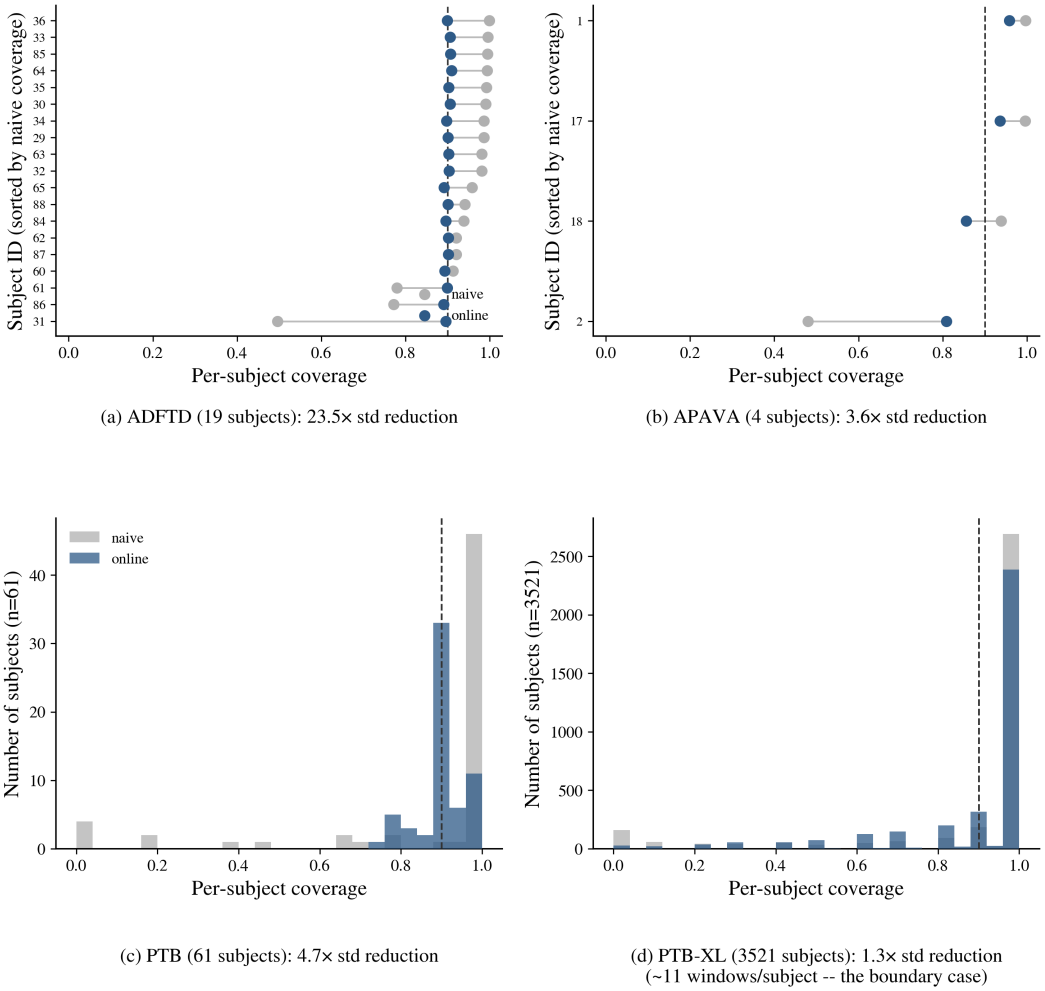
Figure 7 gives the full per-subject comparison between pooled calibration and exogenous closed-loop quantile tracking across all four datasets.

The standard deviation decreases from 0.1200 to 0.0051 on ADFTD, from 0.2166 to 0.0604 on APAVA, and from 0.2927 to 0.0617 on PTB, corresponding to 23.5×, 3.6×, and 4.7× reductions. PTB-XL improves from 0.2710 to 0.2044, a substantially weaker 1.3× reduction.

On ADFTD, the improvement in subject-level reliability also coincides with smaller prediction sets. Across the tested configurations, average prediction-set size decreases from approximately 2.48–

Table 7: **Detailed CAP results within the deferred subset.** The last column gives the naive and CAP coverage of the selected worst-affected subject within CAP's evaluation scope.

Dataset	Naive deferred	CAP	Worst subject: naive → CAP
ADFTD	0.9620	0.9316	0.9965 → 0.7535
APAVA	1.0000	0.8711	1.0000 → 0.7684
PTB	0.0000	0.8878	0.0000 → 0.8182
PTB-XL	0.9046	0.8950	1.0000 → 0.0000



(a) ADFTD (19 subjects): 23.5× std reduction

(b) APAVA (4 subjects): 3.6× std reduction

(c) PTB (61 subjects): 4.7× std reduction

(d) PTB-XL (3521 subjects): 1.3× std reduction (~11 windows/subject -- the boundary case)

Figure 7: **Full per-subject results for exogenous error feedback.** Naive pooled calibration and closed-loop quantile tracking across ADFTD, APAVA, PTB, and PTB-XL. Error feedback substantially tightens subject-level coverage on the first three datasets, while PTB-XL retains considerable residual heterogeneity.

2.51 under pooled calibration to 1.87–1.89 under sequential adaptation. We treat this efficiency change as an empirical observation rather than a theoretical consequence.

B.6 Additional Robustness Analyses

We test whether the closed-loop result depends on the nominal coverage target or the choice of nonconformity score. Adaptation-rate and within-subject feedback-volume analyses are reported in Appendix B.1.1.

Table 8: **Robustness across target coverage levels.** We report the standard deviation of per-subject coverage before and after exogenous closed-loop adaptation.

Dataset	Target coverage	Naive std.	Closed-loop std.
ADFTD	0.80	0.1769	0.0064
ADFTD	0.90	0.1200	0.0051
ADFTD	0.95	0.0683	0.0040
PTB	0.80	0.3589	0.0692
PTB	0.90	0.2927	0.0617
PTB	0.95	0.2160	0.0475

Table 9: **Robustness with APS nonconformity scores.** We report the naive pooled APS threshold together with the naive and closed-loop per-subject coverage standard deviations under exogenous adaptation; the reduction column is their ratio. The naive threshold rounds to 1.0000 on three datasets, but the underlying naive std. is nonzero in every case, so the reduction is well defined.

Dataset	Naive APS threshold	Naive std.	Closed-loop std.	Std. reduction
ADFTD	1.0000	0.0138	0.0009	$15.1\times$
APAVA	1.0000	0.0150	0.0014	$11.1\times$
PTB	1.0000	0.2809	0.0048	$58.5\times$
PTB-XL	0.9888	0.2318	0.0829	$2.8\times$

B.6.1 Target Coverage

Table 8 varies target coverage over $\{0.80, 0.90, 0.95\}$ on ADFTD and PTB. Exogenous feedback substantially reduces subject-level dispersion at every target.

B.6.2 Alternative Nonconformity Score

We additionally replace $1 - p_\theta(y \mid x)$ with APS. The same qualitative pattern persists across all four datasets.

The naive APS threshold is numerically close to its upper score limit on ADFTD, APAVA, and PTB, rounding to 1.0000 in Table 9. Despite this near-saturation, sequential error feedback still substantially reduces subject-level coverage dispersion.

B.7 Additional Endogenous-Feedback Statistics

Section 4.4 reports the main endogenous-feedback results. Table 10 gives the fraction of predictions whose outcomes enter the controller under the prediction-dependent observation mechanism.

The feedback rates vary substantially across datasets, but feedback quantity alone does not explain the failure. For ADFTD subject 31, exogenous feedback raises coverage from 0.4957 to 0.8957, whereas endogenous feedback gives only 0.0757 at $\eta = 0.05$. Increasing the gain to $\eta = 0.5$ raises coverage only to 0.1104. This result supports the distinction between feedback quantity and feedback representativeness.

B.8 Robustness Across Architectures and Random Seeds

We report additional robustness checks beyond the single-seed results used in the main tables. On ADFTD, we evaluate two architectures and five independently trained models. On APAVA and PTB, we additionally evaluate a second independently trained seed. We do not extend the architecture- or seed-robustness claim to PTB-XL.

ADFTD. As described in Section 4.1, the main conformal analyses use the CoTAR-based backbone, while the ConfTr diagnostic uses a Medformer-based backbone. Here we evaluate robustness across both architectures and five independently trained models—three CoTAR seeds and two Medformer-

Table 10: **Endogenous-feedback statistics.** We report subject-level coverage dispersion before adaptation and under endogenous feedback, together with the observed feedback rate.

Dataset	Naive std.	Endogenous std.	Feedback rate
ADFTD	0.1200	0.1978	45.04%
APAVA	0.2166	0.2806	5.38%
PTB	0.2927	0.2810	0.06%
PTB-XL	0.2710	0.3126	27.31%

base seeds. Table 11 reports per-subject coverage dispersion for all five models and additionally reports subject 31's naive and closed-loop coverage when available.

All three CoTAR seeds show severe naive undercoverage for subject 31 and restore its coverage to approximately the nominal target under online adaptation. Across all five trained models, closed-loop adaptation yields an order-of-magnitude reduction in overall subject-level coverage dispersion.

Table 11: ADFTD closed-loop correction across two architectures and five independently trained models. Subject 31 coverage is available for the three CoTAR runs but was not isolated individually in the logged output of the two Medformer-base runs, hence the dashes. Overall subject-level coverage dispersion is available for all five models.

Architecture	Seed	Subject 31 (naive $\rightarrow$ online)	Overall (online)	Std (naive $\rightarrow$ online)
CoTAR	42	$0.4957 \rightarrow 0.8957$	0.9001	$0.1200 \rightarrow 0.0051$ $(23.5\times)$
CoTAR	43	$0.5139 \rightarrow 0.8974$	0.9001	$0.1167 \rightarrow 0.0050$ $(23.3\times)$
CoTAR	44	$0.5374 \rightarrow 0.8965$	0.9001	$0.1118 \rightarrow 0.0056$ $(20.0\times)$
Medformer-base	42	—	0.9000	$0.0934 \rightarrow 0.0062$ $(15.1\times)$
Medformer-base	43	—	0.9000	$0.0974 \rightarrow 0.0067$ $(14.5\times)$

APAVA and PTB. A second independently trained seed is available for the single architecture used on these datasets. Table 12 matches the direction and order of magnitude of the seed-42 numbers in Table 1.

Table 12: APAVA and PTB closed-loop correction under a second random seed.

Dataset	Seed	Naive std	Online std (reduction)
APAVA	42 (Table 1)	0.2166	0.0604 $(3.6\times)$
APAVA	43	0.1786	0.0613 $(2.9\times)$
PTB	42 (Table 1)	0.2927	0.0617 $(4.7\times)$
PTB	43	0.2898	0.0508 $(5.7\times)$

PTB-XL. Only one trained model is available for this dataset; the alternate-seed checkpoint was not retained, and retraining was outside this submission's timeline. We report PTB-XL as single-seed, single-architecture and do not extend the seed/architecture-robustness claim to it.

C Theoretical Analysis

This appendix formalizes four mechanisms underlying the subject-level reliability failures identified in our audit: window-weighted aggregation, pooled calibration across heterogeneous subject distributions, inherited state in sequential adaptation, and selective feedback. The aggregation results operate directly on empirical coverage and therefore do not require window-level exchangeability.

C.1 Window-Weighted Aggregation

Recall the empirical subject-level coverage $\widehat{\mathrm{Cov}}_m$, the number of test subjects M_{test}, and the total number of test windows N_{test} from Section 2. Define the fraction of test windows contributed by

subject m as

$$\pi_m = \frac{n_m}{N_{\text{test}}}. \tag{23}$$

Proposition 1 (Window-weighted coverage)**.** *Pooled empirical coverage satisfies*

$$\widehat{\text{Cov}}_{\text{win}} = \sum_{m \in \mathcal{S}_{\text{test}}} \pi_m \widehat{\text{Cov}}_m. \tag{24}$$

Proof. By definition,

$$\widehat{\text{Cov}}_{\text{win}} = \frac{1}{N_{\text{test}}} \sum_{m \in \mathcal{S}_{\text{test}}} \sum_{i=1}^{n_m} \mathbf{1}[y_{m,i} \in C(x_{m,i})] \tag{25}$$

$$= \sum_{m \in \mathcal{S}_{\text{test}}} \frac{n_m}{N_{\text{test}}} \left(\frac{1}{n_m} \sum_{i=1}^{n_m} \mathbf{1}[y_{m,i} \in C(x_{m,i})] \right) \tag{26}$$

$$= \sum_{m \in \mathcal{S}_{\text{test}}} \pi_m \widehat{\text{Cov}}_m. \tag{27}$$

□

Proposition 1 shows that pooled evaluation weights subjects according to the number of windows they contribute. To compare this quantity with an equal-subject summary, define

$$\widehat{\text{Cov}}_{\text{subj}} = \frac{1}{M_{\text{test}}} \sum_{m \in \mathcal{S}_{\text{test}}} \widehat{\text{Cov}}_m. \tag{28}$$

Proposition 2 (Effect of unequal subject sizes)**.** *Let $\mathbb{E}_m$ and Cov_m denote expectation and covariance under the uniform distribution over test subjects. Then*

$$\widehat{\text{Cov}}_{\text{win}} - \widehat{\text{Cov}}_{\text{subj}} = \frac{\text{Cov}_m(n_m, \widehat{\text{Cov}}_m)}{\mathbb{E}_m[n_m]}. \tag{29}$$

Proof. Since

$$N_{\text{test}} = M_{\text{test}} \mathbb{E}_m[n_m], \tag{30}$$

we obtain

$$\widehat{\text{Cov}}_{\text{win}} = \frac{\mathbb{E}_m[n_m \widehat{\text{Cov}}_m]}{\mathbb{E}_m[n_m]}. \tag{31}$$

Using

$$\text{Cov}_m(n_m, \widehat{\text{Cov}}_m) = \mathbb{E}_m[n_m \widehat{\text{Cov}}_m] - \mathbb{E}_m[n_m] \mathbb{E}_m[\widehat{\text{Cov}}_m] \tag{32}$$

and

$$\mathbb{E}_m[\widehat{\text{Cov}}_m] = \widehat{\text{Cov}}_{\text{subj}}, \tag{33}$$

rearrangement gives Eq. 29. □

Positive association between subject size and coverage raises pooled coverage relative to the equal-subject average. Equal subject sizes remove this weighting effect, but they do not force individual subject coverages to coincide.

C.2 How Much Failure Can Pooled Coverage Conceal?

We next quantify how much subject-level undercoverage remains compatible with nominal pooled coverage.

Proposition 3 (Individual subject bound)**.** *Suppose*

$$\widehat{\mathrm{Cov}}_{\mathrm{win}} = \sum_{m \in \mathcal{S}_{\mathrm{test}}} \pi_m \widehat{\mathrm{Cov}}_m \geq 1 - \alpha, \tag{34}$$

where $\widehat{\mathrm{Cov}}_m \in [0, 1]$*. Then every subject* m *satisfies*

$$\widehat{\mathrm{Cov}}_m \geq \max\left\{0,\, 1 - \frac{\alpha}{\pi_m}\right\}. \tag{35}$$

Hence, if $\pi_m \leq \alpha$*, nominal pooled coverage remains compatible with zero coverage for subject* m*.*

Proof. All subjects other than m contribute at most $1 - \pi_m$ to pooled coverage. Therefore,

$$1 - \alpha \leq \pi_m \widehat{\mathrm{Cov}}_m + (1 - \pi_m). \tag{36}$$

Rearranging gives

$$\widehat{\mathrm{Cov}}_m \geq 1 - \frac{\alpha}{\pi_m}. \tag{37}$$

Combining this inequality with $\widehat{\mathrm{Cov}}_m \geq 0$ proves the result. □

We extend the same argument to the subjects counted as failures by Eq. 10.

Proposition 4 (Hidden undercoverage mass)**.** *Let*

$$\mathcal{A}_\delta = \left\{ m \in \mathcal{S}_{\mathrm{test}} : \widehat{\mathrm{Cov}}_m < 1 - \alpha - \delta \right\}, \qquad \delta > 0, \tag{38}$$

and define its test-window mass as

$$\Pi_{\mathcal{A}_\delta} = \sum_{m \in \mathcal{A}_\delta} \pi_m. \tag{39}$$

If pooled coverage reaches the nominal target, then

$$\Pi_{\mathcal{A}_\delta} < \frac{\alpha}{\alpha + \delta}. \tag{40}$$

The quantity $\alpha/(\alpha + \delta)$ *is the supremum of the admissible failure mass in the continuous relaxation of the coverage vector.*

Proof. For every $m \in \mathcal{A}_\delta$, $\widehat{\mathrm{Cov}}_m < 1 - \alpha - \delta$, while every remaining subject has coverage at most 1. Hence,

$$\begin{aligned} \widehat{\mathrm{Cov}}_{\mathrm{win}} &= \sum_{m \in \mathcal{S}_{\mathrm{test}}} \pi_m \widehat{\mathrm{Cov}}_m && (41)\\ &< \Pi_{\mathcal{A}_\delta}(1 - \alpha - \delta) + (1 - \Pi_{\mathcal{A}_\delta}) && (42)\\ &= 1 - \Pi_{\mathcal{A}_\delta}(\alpha + \delta). && (43) \end{aligned}$$

Since $\widehat{\mathrm{Cov}}_{\mathrm{win}} \geq 1 - \alpha$, we obtain

$$\Pi_{\mathcal{A}_\delta} < \frac{\alpha}{\alpha + \delta}. \tag{44}$$

The upper bound can be approached arbitrarily closely as the coverages of subjects in $\mathcal{A}_\delta$ approach $1 - \alpha - \delta$ from below while the remaining subjects approach coverage 1. □

For $\alpha = \delta = 0.10$,

$$\frac{\alpha}{\alpha + \delta} = \frac{0.10}{0.20} = 0.5. \tag{45}$$

Thus, subjects with empirical coverage below 0.80 can account for arbitrarily close to half of the test-window mass while pooled coverage still reaches 0.90.

Propositions 1–4 do not invoke window-level exchangeability. They characterize what an observed pooled coverage value implies, and does not imply, about individual subjects.

C.3 Pooled Calibration as a Mixture Distribution

We next analyze the calibration side of the same statistical-unit mismatch. For subject m, define the population score CDF

$$F_m(q) = \mathbb{P}\left(s(X,Y) \le q \mid S = m\right). \tag{46}$$

Pooling windows induces a subject distribution P_{win} associated with sampling a window from the pooled population. The resulting score distribution is

$$F_{\mathrm{pool}}(q) = \mathbb{E}_{S\sim P_{\mathrm{win}}}[F_S(q)]. \tag{47}$$

When subjects contribute different numbers of windows, P_{win} weights high-volume subjects more heavily than an equal-subject distribution.

Define the pooled population quantile

$$q^\star_{\mathrm{pool}} = \inf\left\{q : F_{\mathrm{pool}}(q) \ge 1-\alpha\right\}, \tag{48}$$

and the subject-specific target quantile

$$q^\star_m = \inf\left\{q : F_m(q) \ge 1-\alpha\right\}. \tag{49}$$

Proposition 5 (Pooled and subject-specific calibration)**.** *If*

$$q^\star_m > q^\star_{\mathrm{pool}}, \tag{50}$$

then

$$F_m(q^\star_{\mathrm{pool}}) < 1-\alpha. \tag{51}$$

Hence, the pooled population quantile undercovers subject m.

Proof. By definition, $q^\star_m$ is the smallest threshold at which F_m reaches $1-\alpha$. Therefore, every $q < q^\star_m$ satisfies $F_m(q) < 1-\alpha$. Since $q^\star_{\mathrm{pool}} < q^\star_m$, the result follows. □

The discrepancy

$$\Delta_m = q^\star_m - q^\star_{\mathrm{pool}} \tag{52}$$

summarizes the subject-specific calibration gap relative to the pooled population target. A positive Δ_m indicates that subject m requires a larger threshold than the pooled target, whereas a negative value indicates the opposite. This motivates the conditioning-signal experiments in Section 3.1: useful signals should reveal whether the current subject's calibration requirement differs from the pooled one.

This population analysis does not claim that the finite-sample threshold $\hat{q}$ enjoys classical split-conformal validity when pooled windows fail the required exchangeability conditions. Instead, it isolates the population target induced by window pooling and distinguishes that target from the quantile required by an individual subject.

C.4 Sequential Feedback and Inherited Controller State

We now analyze a consequence of the sequential controller already defined in Section 3.1. Recall that test observations are arranged in a subject-contiguous stream, subject m occupies $\mathcal{I}_m = \{\tau_m, \ldots, \tau_m + n_m - 1\}$, and the shared controller starts from the pooled threshold and updates after each realized error according to Eq. 19. The controller maintains one running state throughout the stream and does not reset at subject boundaries.

For any $r \in \{1, \ldots, n_m\}$, unrolling Eq. 19 over the first r observations of subject m gives

$$q_{\tau_m+r} = q_{\tau_m} + \eta \sum_{u=\tau_m}^{\tau_m+r-1} (e_u - \alpha). \tag{53}$$

Eq. 53 separates two contributions to the threshold during subject m: the inherited state q_{τ_m} produced by all preceding observations and the correction accumulated from the current subject. Because subject m contributes only n_m consecutive updates, its segment length determines how much opportunity its realized errors have to move the controller away from the inherited state.

This identity explains the feedback-sufficiency intervention in Section 3.2. When n_m is small, the controller receives fewer consecutive updates from subject m, so the threshold used over that segment can remain more strongly influenced by preceding subjects. Increasing n_m provides more within-subject feedback with which to modify the inherited state.

The same decomposition also shows why adaptive subject-level coverage can depend on stream organization. Changing the observations preceding subject m can change q_{τ_m} even when subject m and its internal window order remain unchanged. Thus, subject-contiguous and interleaved schedules can produce different trajectories under a shared controller. Our windows-per-subject intervention therefore holds subject order and controller parameters fixed when varying within-subject feedback volume.

C.5 Feedback Bias under Selective Observation

We finally analyze selective outcome observation under the same running controller. Let S_t denote the subject occupying stream position t, and let $o_t \in \{0, 1\}$ indicate whether the corresponding outcome becomes observable. Under the selective-feedback update in Eq. 21, only observations with $o_t = 1$ contribute an error signal to the controller.

Proposition 6 (Selective-feedback bias)**.** *For any subject m with $\mathbb{P}(o_t = 1 \mid S_t = m) > 0$,*

$$\mathbb{E}[e_t \mid o_t = 1, S_t = m] - \mathbb{E}[e_t \mid S_t = m] = \frac{\mathrm{Cov}(e_t, o_t \mid S_t = m)}{\mathbb{P}(o_t = 1 \mid S_t = m)}. \tag{54}$$

Proof. Conditioning on $S_t = m$ gives

$$\mathbb{E}[e_t \mid o_t = 1, S_t = m] = \frac{\mathbb{E}[e_t o_t \mid S_t = m]}{\mathbb{P}(o_t = 1 \mid S_t = m)}. \tag{55}$$

Using

$$\mathrm{Cov}(e_t, o_t \mid S_t = m) = \mathbb{E}[e_t o_t \mid S_t = m] - \mathbb{E}[e_t \mid S_t = m]\mathbb{P}(o_t = 1 \mid S_t = m) \tag{56}$$

and rearranging proves Eq. 54. □

Eq. 54 shows that observed feedback reproduces the subject's overall mean miscoverage rate whenever

$$\mathrm{Cov}(e_t, o_t \mid S_t = m) = 0. \tag{57}$$

This zero-covariance condition suffices for unbiased mean-error feedback; it does not require full conditional independence between e_t and o_t. When prediction behavior affects outcome observation, this covariance can become nonzero, and the update signal can reflect the selectively observed subset rather than the subject's overall error process.

Together, these results characterize four aspects of the statistical-unit mismatch identified in our audit. Pooled evaluation weights subjects by their numbers of windows; pooled calibration combines heterogeneous subject-specific score distributions; sequential adaptation inherits state across subject boundaries; and selective outcome observation can bias the feedback used for correction.

C.6 Per-Subject Controllers and a Held-Out Evaluation Split

The shared controller analyzed in Appendix C.4 maintains one state across the global test stream. Under subject-contiguous scheduling, the inherited state q_{τ_m} in Eq. 53 depends on observations from preceding subjects. We now formalize a per-subject alternative that removes this cross-subject state dependence.

Per-subject controller. For subject m, let $t_{m,1}, \ldots, t_{m,n_m}$ denote the global positions of its windows under a given deployment schedule, listed in within-subject order. We initialize an independent state for subject m as

$$q_1^{(m)} = \hat{q}. \tag{58}$$

At its ith window, the prediction set is

$$C_i^{(m)}(x) = \{y : s(x, y) \le q_i^{(m)}\}, \tag{59}$$

with corresponding error

$$e_i^{(m)} = \mathbf{1}\left[y_{t_{m,i}} \notin C_i^{(m)}(x_{t_{m,i}})\right]. \tag{60}$$

The state then evolves as

$$q_{i+1}^{(m)} = q_i^{(m)} + \eta\big(e_i^{(m)} - \alpha\big), \qquad i = 1, \dots, n_m - 1. \tag{61}$$

Thus, $q_i^{(m)}$ depends only on $\hat{q}$ and subject m's preceding realized errors $e_1^{(m)}, \dots, e_{i-1}^{(m)}$, and not on observations from any other subject. Consequently, any two schedules that preserve subject m's within-subject order produce the same state sequence $q_1^{(m)}, \dots, q_{n_m}^{(m)}$. The per-subject controller is therefore invariant to whether subject m's windows are contiguous or interleaved with those of other subjects.

In contrast, the shared-state controller depends on all observations preceding each global stream position through its inherited state. It can therefore change when subjects are reordered or interleaved.

Relation to adaptive conformal inference. For a fixed subject m, Eq. 61 is the same quantile-tracking recursion applied only to that subject's own observation sequence. Existing long-run coverage-control results for adaptive conformal inference and P-control therefore motivate the behavior of each per-subject trajectory (Gibbs & Candès, 2021; Angelopoulos et al., 2023). Because our nonconformity scores satisfy $s(x, y) \in [0, 1]$, the bounded-score setting considered by these results applies here.

This connection does not imply finite-sample subject-conditional coverage or a particular reduction in cross-subject coverage dispersion. Rather, it clarifies why the amount of subject-specific feedback remains important: longer subject sequences provide more updates through which cumulative miscoverage can approach the target. Consistent with this feedback-sufficiency effect, the empirical dispersion reduction ranges from $30.8\times$ on ADFTD to $2.8\times$ on PTB-XL; Appendix B.1.2 reports the full scheduling comparison.

Held-out evaluation. For this analysis, we return to the subject-contiguous schedule used in the main experiments. Eq. 20 evaluates subject m over the same windows $t \in \mathcal{I}_m$ whose realized errors are used to update the controller during that subject's segment. To distinguish continual online adaptation from evaluation under a fixed adapted threshold, we partition $\mathcal{I}_m$ into a feedback half

$$\mathcal{I}_m^{\mathrm{fb}} = \{\tau_m, \dots, \tau_m + \lfloor n_m/2 \rfloor - 1\} \tag{62}$$

and a held-out half

$$\mathcal{I}_m^{\mathrm{ho}} = \mathcal{I}_m \setminus \mathcal{I}_m^{\mathrm{fb}}. \tag{63}$$

Simply reporting coverage on $t \in \mathcal{I}_m^{\mathrm{ho}}$ while continuing to update the controller throughout that half is *prequential*, rather than genuinely held out. For any held-out position after the first, q_t already depends on realized errors from earlier windows in $\mathcal{I}_m^{\mathrm{ho}}$.

We therefore freeze the shared controller at the first held-out position. Let

$$t_m^* = \min \mathcal{I}_m^{\mathrm{ho}}, \qquad q_m^{\mathrm{frozen}} = q_{t_m^*}. \tag{64}$$

We then apply this threshold without further updates to all windows in $\mathcal{I}_m^{\mathrm{ho}}$:

$$C_m^{\mathrm{frozen}}(x) = \{y : s(x, y) \le q_m^{\mathrm{frozen}}\}. \tag{65}$$

The corresponding held-out subject-level coverage is

$$\widehat{\mathrm{Cov}}_m^{\mathrm{frozen}} = \frac{1}{|\mathcal{I}_m^{\mathrm{ho}}|} \sum_{t \in \mathcal{I}_m^{\mathrm{ho}}} \mathbf{1}\left[y_t \in C_m^{\mathrm{frozen}}(x_t)\right]. \tag{66}$$

By construction, q_m^{frozen} depends only on the pooled initialization and realized errors observed before the held-out half; it does not depend on any outcome in $\mathcal{I}_m^{\mathrm{ho}}$.

For the per-subject controller, let

$$h_m = \left\lfloor \frac{n_m}{2} \right\rfloor \tag{67}$$

denote the number of feedback windows. We analogously freeze its subject-specific state after those windows:

$$q_{m,\mathrm{ps}}^{\mathrm{frozen}} = q_{h_m+1}^{(m)}. \tag{68}$$

This state is then applied without further updates throughout $\mathcal{I}_m^{\mathrm{ho}}$.

As a feedback-matched comparison that uses the same feedback-half labels without a sequential controller, we additionally calibrate an ordinary split-CP threshold from the feedback half,

$$q_m^{\mathrm{split}} = \mathrm{SplitCP}\left(\{s(x_t, y_t)\}_{t \in \mathcal{I}_m^{\mathrm{fb}}}, \alpha\right), \tag{69}$$

and apply it as a fixed threshold to $\mathcal{I}_m^{\mathrm{ho}}$.

Table 3 in Appendix B.1.3 reports the resulting held-out coverage quantities, aggregated across subjects, alongside the full-segment online quantity from Eq. 20.