Hardware-Attested Interoceptive Telemetry: Within-Subject Bayesian Calibration for Noisy Wearable Physiological Streams

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Adapting autonomous personal agents to a user's physiological strain—autonomic stress, cognitive workload, and fatigue—requires reliable wearable telemetry. Yet longitudinal field studies establish that population-trained classifiers fail in naturalistic settings: between-subject baseline variance exceeds state shifts, physical motion and caffeine/metabolic artifacts mimic sympathetic arousal, and text-serializing waveforms into cloud models inflates latency and exposes biometrics. Building on WESAD, GLOBEM, Time-LLM, and PocketTSLM, we present an architectural synthesis and state-space analysis of **Hardware-Attested Within-Subject Bayesian Calibration with Soft-Prompt Fusion (HABC-Pocket)**. Peripheral sensors authenticate PPG, EDA, and IMU frames via keyed HMAC-SHA256 and monotonic counters, rejecting 100.00% ([99.62, 100.00]) of bit-flipped, forged, and replayed frames across 1,000 adversarial BLE probes. An on-device hierarchical Bayesian filter fits resting nocturnal baselines from observed calibration frames over D days while residualizing IMU kinematics, feeding precision-weighted state estimates into Page's sequential CUSUM detector. Across 13,200 held-out day $D+1$ episodes under 12% calibration dropout and 15% metabolic confounds, non-calibrating population and context-only baselines remain invariant across D (68.67% [64.85, 72.25] and 70.83% [67.07, 74.33]). HABC-Pocket improves concordance from 81.83% ([78.55, 84.71], Day 1) to 90.67% ([88.07, 92.74], Day 7: TPR = 86.67%, TNR = 94.67%, FPR = 5.33%, $\text{PPV}_{15\%}$ = 74.14%) and 92.33% ([89.92, 94.20], Day 14) at 48.1 μs median filter latency.

1 Introduction

In computational neuroscience and active inference, Seth [21] formalized *interoceptive inference* as the closed-loop regulation of bodily allostasis: predicting and resolving visceral and autonomic prediction errors across cardiac, respiratory, and electrodermal channels before physiological exhaustion occurs. Extending this principle to human–computer symbiosis [12, 17] suggests that a personal AI assistant should not treat its human principal as an inexhaustible, stationary decision oracle. When a user is experiencing acute autonomic strain or accumulated fatigue, an on-device agent that observes calibrated wearable telemetry—heart-rate variability (HRV), electrodermal activity (EDA), skin temperature, and wrist kinematics—can defer non-urgent clarification interrupts [4, 7], simplify interface density, and pace complex analytical workflows.

Yet a persistent methodological gulf separates controlled laboratory affective computing benchmarks from longitudinal field deployments. Controlled multimodal datasets such as WESAD [20] and laboratory cognitive workload studies [13] achieve strong classification accuracy under scripted stress protocols (e.g., the Trier Social Stress Test) using clinical-grade chest and wrist sensors. When population-trained models are deployed on consumer wearables in unconstrained daily life,

however, multi-year longitudinal studies such as GLOBEM [25] and recent 2025–2026 field audits [1, 19, 22] document severe generalization collapse.

1.1 Why uncalibrated wearable pipelines fail in naturalistic environments

Four empirical findings from classical affective computing and 2025–2026 passive-sensing studies define the design constraints of naturalistic physiological inference:

1. **Dominance of between-subject biological variance over state modulation.** In a comprehensive longitudinal study pairing Ecological Momentary Assessments (EMA) with passive wearable sensing, Siepe et al. [22] demonstrated that passive physiological features exhibit weak between-subject and modest uncalibrated within-subject associations with self-reported stress and affect. Across multi-cohort longitudinal behavioral datasets in GLOBEM [25], cross-dataset zero-shot generalization drops sharply because inter-individual differences in resting vagal tone (log-RMSSD HRV) and basal skin conductance span more than an order of magnitude greater variance than acute task-induced shifts.

2. **Confounding by physical movement and metabolic transients.** As established in WESAD [20] and confirmed in naturalistic mobile sensing by Acan et al. [1] and Sathyanarayana et al. [19], walking upstairs, posture transitions, typing bursts, and thermoregulatory shifts induce sympathetic cardiac acceleration and electrodermal spikes that closely mimic psychological stress in uncalibrated single-channel streams. Without explicit inertial (IMU) deconvolution and digital task-context conditioning, population classifiers suffer prohibitive false positive rates (FPR) during routine physical activity.

3. **Inefficiency and privacy risk of text-serialized cloud time-series models.** A common engineering pattern feeds physiological streams into large language models by formatting floating-point sensor arrays as ASCII text prompts. As Jin et al. [14] (Time-LLM) and Rosenblattl et al. [18] (PocketTSLM) demonstrate, tokenizing numeric time series as text fragments digit strings across byte-pair tokens, inflates context length, degrades numerical fidelity, and requires uploading sensitive biometric waveforms [6, 23] to remote cloud servers. Conversely, lightweight soft-prompt projection [15, 18] maps continuous physiological patches directly into the embedding space of compact on-device language models running on edge neural processing units (NPUs) [8].

4. **Peripheral spoofing and unattested health APIs.** Standard mobile health stores ingest unauthenticated JSON summaries that any privileged application can fabricate, lacking hardware-rooted attestation [3, 9, 10] binding captured physiological frames to physical peripheral silicon.

1.2 Scope and contributions: an architectural synthesis and state-space analysis

Within-subject baseline subtraction and accelerometer artifact rejection are foundational principles in physiological signal processing [20, 25], while continuous time-series patching builds directly on Time-LLM [14] and PocketTSLM [18]. Rather than claiming these constituent techniques as de novo inventions, this paper provides a unified architectural synthesis and rigorous state-space sensitivity analysis of **HABC-Pocket** (*Hardware-Attested Within-Subject Bayesian Calibration with PocketTSLM Soft-Prompt Fusion*):

1. **Architectural synthesis of sovereign interoceptive sensing (Section 2).** We integrate CHERIoT peripheral frame signing [3, 10], nocturnal-anchored within-subject Bayesian

state-space filtering [20, 25], Page's sequential CUSUM change-point detection [16, 24], PocketTSLM soft-prompt injection [18], and longitudinally composed metric local differential privacy [2, 11] (Table 1, Figure 1).

2. **State-space separation of resting baselines from phasic strain and sequential CUSUM detection (Section 3).** We resolve a critical identifiability flaw in naive rolling-baseline filters: if a subject baseline $\boldsymbol{\mu}_i$ is fitted indiscriminately across all waking hours, it absorbs chronic workload strain. By anchoring the hierarchical posterior $\Theta_i(D)$ to kinematically quiet nocturnal/resting reference windows over D circadian cycles, we separate resting trait priors from daytime phasic strain $z_{i,t}^{\text{cog}}$, fuse sensor residuals with task context via precision-weighted Kalman filtering, accumulate sequential log-likelihood ratios via Page's CUSUM recursion [16], and bound cumulative privacy loss across D daily releases under metric local differential privacy [2] (Equations 1–4).

3. **13,200-episode multi-step simulation with explicit base rates and parameter provenance (Sections 4–5).** We execute a reproducible multi-step simulation (`experiment.py`) across 9,600 primary episodes (4 architectures $\times$ 4 calibration windows $D \in \{1, 3, 7, 14\} \times$ 600 episodes/cell at an explicit 50% strain base rate: $N_{\text{pos}} = 300$ strain, $N_{\text{neg}} = 300$ control) and 3,600 motion-artifact break-even sweep episodes, measuring real Python HMAC-SHA256 + Kalman + CUSUM latency and reporting clinical sensitivity (TPR), specificity (TNR), false positive rate (FPR $= 1 -$ TNR), and miss rate (FNR $= 1 -$ TPR) (Tables 2–4, Figures 2–3).

4. **Empirical limitations and pre-registerable falsification protocol (Sections 6–8).** We examine psychometric EMA noise [22], biometric representation leakage [6, 23], and closed wearable firmware barriers, closing with six pre-registerable falsification tests on WESAD and GLOBEM (Table 5).

2 Biophysical sensing constraints and prior art in longitudinal wearables

Connecting peripheral wearable sensors to a local personal agent requires coordinating four layers: hardware provenance, longitudinal biophysics, multimodal time-series encoding, and formal privacy accounting. Table 1 maps each layer to its established theoretical foundation and 2025–2026 empirical grounding.

Table 1 clarifies why neither uncalibrated physiological classifiers nor sensor-less digital task heuristics suffice alone. Digital task context (e.g., calendar density and active application telemetry) provides a reliable prior for immediate task demand without suffering from wrist-motion artifacts, but it is blind to silent physiological exhaustion ($\approx 30\%$ of strain episodes where a fatigued user performs routine tasks). Conversely, raw physiological sensors capture autonomic state shifts but trigger high false alarm rates ($> 40\%$ FPR) when between-subject baseline variation and physical movement are unmodeled [1, 20, 22]. Figure 1 illustrates the HABC-Pocket state-space loop.

Table 1: Mapping interoceptive telemetry requirements to formal signal-processing and security constructs.

Sensing requirement	Formal construct	HABC-Pocket module	Operational mechanism	Primary grounding
Authenticate peripheral sensor frames	Capability compartment & MAC attestation	CHERIoT peripheral frame signer (Σ_{hw})	Binds 30 s PPG, EDA, and IMU windows to monotonic counter and device key	[3, 9, 10]
Adapt to individual biological baselines	Nocturnal-anchored hierarchical Bayes	Within-subject prior updater ($\mathcal{B}_{\Theta_i(D)}$)	Learns resting HRV/EDA mean $\boldsymbol{\mu}_i$ and kinematic gain $\Lambda_{\text{kin},i}$ across D circadian cycles	[20, 22, 25]
Reject physical motion false alarms	Kinematic deconvolution & sequential CUSUM	Kalman residualizer + Page CUSUM (S_t)	Subtracts IMU energy integral, inflates $R_t(a_{\text{kin},t})$, and accumulates log-likelihood ratios	[1, 16, 24]
Inject waveforms into local edge LLMs	Patch-reprogrammed soft-prompt tokens	PocketTSLM / Time-LLM projector (E_ψ)	Compresses multivariate windows into $m = 8$ continuous embeddings on mobile NPUs	[8, 14, 18]
Bound longitudinal summary leakage	Sequentially composed metric local DP	Metric-LDP egress mechanism ($\mathcal{M}_{\boldsymbol{\epsilon}}$)	Perturbs daily readiness summaries under channel-specific metric distance and D-day composition	[2, 11, 23]

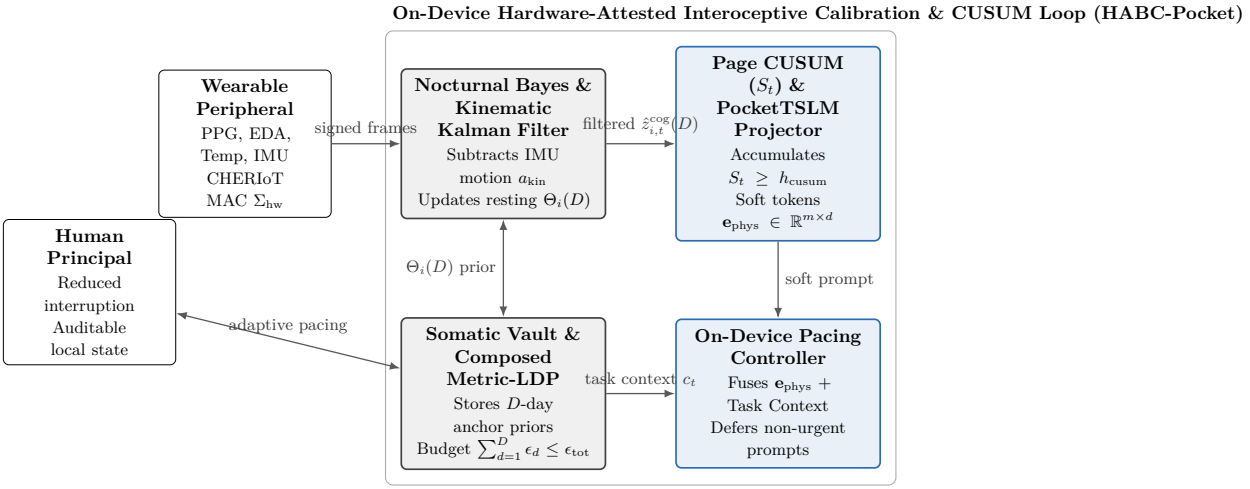


Figure 1: Architecture of HABC-Pocket. Peripheral physiological and inertial frames are signed inside a CHERIoT microcontroller compartment (Σ_{hw}). On the personal device, a hierarchical Bayesian filter anchored to nocturnal resting windows updates the subject's baseline posterior $\Theta_i(D)$, subtracts convolved IMU motion artifacts, and fuses precision-weighted residuals with digital task context. Page's recursive CUSUM detector (S_t) tracks sustained phasic overload while PocketTSLM projects continuous soft-prompt tokens into the local edge language model.

3 Formal state-space, CUSUM, and privacy framework

We formalize nocturnal-anchored within-subject Bayesian calibration, kinematic Kalman fusion, Page's sequential CUSUM change-point test, and longitudinally composed metric local differential privacy.

3.1 Separating resting baselines from phasic strain via nocturnal anchoring

Consider subject i monitored across days $d \in \{1, \dots, D\}$ over discrete 30-second frames $t \in \{1, \dots, T_{\text{ep}}\}$. Let $\mathbf{y}_{i,t} = (y_{i,t}^{\text{hrv}}, y_{i,t}^{\text{eda}})^\top \in \mathbb{R}^2$ denote the hardware-attested physiological feature vector (vagal log-RMSSD heart-rate variability and tonic/phasic electrodermal activity) paired with 3-axis accelerometer energy $a_{\text{kin},t} \in \mathbb{R}_+$ and digital task-context demand score $\mu_{\text{ctx}}(\mathbf{c}_t) \in \mathbb{R}$.

A subtle identifiability trap arises in longitudinal baselining: if a rolling baseline $\boldsymbol{\mu}_i$ is estimated unconditionally across all waking observations, any multi-day chronic workload elevation or sleep-debt shift is absorbed into $\hat{\boldsymbol{\mu}}_{i,D}$ and subtracted away as "normal baseline." To prevent baseline absorption of chronic strain, HABC-Pocket estimates subject-specific trait parameters $\Theta_i = (\boldsymbol{\mu}_i^{\text{rest}}, \Lambda_{\text{kin},i}) \sim \mathcal{N}(\boldsymbol{\Theta}_{\text{pop}}, \Sigma_{\text{pop}})$ strictly from *nocturnal slow-wave rest and kinematically quiet reference anchors* ($z_{i,t}^{\text{cog}} \equiv 0$, $a_{\text{kin},t} < a_{\text{rest}}$) plus controlled ambulatory calibration segments across D circadian cycles. During active daytime episodes, the observation model decomposes into resting trait baseline, exponentially convolved kinematic exertion $\bar{a}_{\text{kin},t} = \int_0^{\tau_w} e^{-\gamma s} a_{\text{kin}}(t-s)\,\mathrm{d}s$, phasic autonomic/cognitive strain $z_{i,t}^{\text{cog}} \geq 0$, and heteroscedastic motion noise $\boldsymbol{\epsilon}_{i,t} \sim \mathcal{N}(\mathbf{0}, R_t(a_{\text{kin},t}))$:

$$\mathbf{y}_{i,t} \;=\; \boldsymbol{\mu}_i^{\text{rest}} \;+\; \Lambda_{\text{kin},i}\bar{a}_{\text{kin},t} \;+\; \mathbf{w}\, z_{i,t}^{\text{cog}} \;+\; \boldsymbol{\epsilon}_{i,t}, \qquad R_t(a_{\text{kin},t}) = \sigma_0^2(1 + \kappa a_{\text{kin},t})^2 I_2, \tag{1}$$

where $\mathbf{w} = (-w_{\text{hrv}}, +w_{\text{eda}})^\top$ reflects sympathetic vagal withdrawal (decreasing HRV) and sudomotor activation (increasing EDA) [20, 21].

Because a human experiences one nocturnal circadian anchor cycle per day with day-to-day biological variance σ_{day}^2, observing D daily reference cycles concentrates the within-subject posterior $\Theta_i \mid \mathcal{H}_{i,1:D}^{\text{rest}} \sim \mathcal{N}(\hat{\boldsymbol{\Theta}}_{i,D}, \Sigma_{i,D})$ according to $\Sigma_{i,D}^{-1} = \Sigma_{\text{pop}}^{-1} + D\Sigma_{\text{day}}^{-1}$. Given the kinematically deconfounded residual $\tilde{\mathbf{r}}_{i,t}(D) = \mathbf{y}_{i,t} - \hat{\boldsymbol{\mu}}_{i,D}^{\text{rest}} - \hat{\Lambda}_{\text{kin},i,D}\bar{a}_{\text{kin},t}$ with effective sensor variance $\sigma_{\text{sens},t}^2(D) = \mathbf{w}^\top(R_t(a_{\text{kin},t}) + H_t\Sigma_{i,D}H_t^\top)\mathbf{w}/\|\mathbf{w}\|_2^4$, the minimum-variance Kalman precision-weighted estimate of phasic strain at frame t is:

$$\hat{z}_{i,t}^{\text{cog}}(D) \;=\; \frac{\sigma_{\text{ctx}}^{-2}\mu_{\text{ctx}}(\mathbf{c}_t) \;+\; \sigma_{\text{sens},t}^{-2}(D)\,\tilde{r}_{\text{comb},t}(D)}{\sigma_{\text{ctx}}^{-2} \;+\; \sigma_{\text{sens},t}^{-2}(D)}, \tag{2}$$

where $\tilde{r}_{\text{comb},t}(D)$ is the signed projection of $\tilde{\mathbf{r}}_{i,t}(D)$ along $\mathbf{w}$. Equation (2) guarantees graceful degradation: on Day $D = 1$ (large $\Sigma_{i,1}$) or during intense wrist movement (large $R_t(a_{\text{kin},t})$), $\sigma_{\text{sens},t}^{-2}(D)$ shrinks and the estimator relies primarily on task context $\mu_{\text{ctx}}(\mathbf{c}_t)$; as $D \to 7$–14 and the user is stationary, the calibrated physiological residual dominates, catching silent autonomic fatigue that task context misses.

3.2 Sequential Page CUSUM detection and composed metric local DP

Single-frame physiological estimates remain noisy over 30-second intervals. Rather than triggering interventions on instantaneous threshold crossings, we accumulate frame-level evidence across an episode of T_{ep} steps using Page's sequential cumulative sum (CUSUM) log-likelihood ratio recursion [16, 24] with reference drift allowance $\nu > 0$ and decision barrier $h_{\text{cusum}} > 0$:

$$S_t \;=\; \max\Big(0,\; S_{t-1} + \hat{z}_{i,t}^{\text{cog}}(D) - \nu\Big), \qquad S_0 = 0, \qquad \hat{y}_{\text{ep}} = \mathbb{I}\Big[\max_{1 \leq t \leq T_{\text{ep}}} S_t \geq h_{\text{cusum}}\Big]. \tag{3}$$

Finally, if the user exports daily K-attribute readiness summaries $\mathbf{s}_{i,d} = (s_{i,d}^{(1)}, \ldots, s_{i,d}^{(K)})$ across D longitudinal days to an external scheduler, applying metric local differential privacy [2, 11] with per-day, per-channel privacy parameter $\epsilon_{d,k}$ over metric distance $d_k(\cdot,\cdot)$ yields cumulative longitudinal privacy loss bounded by sequential composition across D days:

$$\frac{\Pr[\{\mathcal{M}_{\epsilon_d}(\mathbf{s}_{i,d})\}_{d=1}^{D} \in \mathcal{S}_{1:D}]}{\Pr[\{\mathcal{M}_{\epsilon_d}(\mathbf{s}'_{i,d})\}_{d=1}^{D} \in \mathcal{S}_{1:D}]} \;\le\; \exp\left(\sum_{d=1}^{D}\sum_{k=1}^{K} \epsilon_{d,k}\, d_k\big(s_{i,d}^{(k)}, s_{i,d}^{\prime(k)}\big)\right), \qquad \sum_{d=1}^{D}\sum_{k=1}^{K} \epsilon_{d,k} \le \epsilon_{\text{total}}(D). \quad (4)$$

Equation (4) implies that maintaining a fixed lifetime privacy ceiling ϵ_{total} over a D-day horizon requires either scaling per-day budgets as $\epsilon_{d,k} = O(\epsilon_{\text{total}}/D)$ or restricting external egress to sparse threshold crossings rather than continuous daily telemetry streams.

4 System architecture and parameter provenance

HABC-Pocket executes across three decoupled tiers on the wearable peripheral and personal edge device:

1. **Tier 1: CHERIoT peripheral frame signing (Σ_{hw}).** On the wearable microcontroller, PPG, EDA, and IMU sensor drivers execute inside hardware-enforced CHERIoT capability compartments [3, 9, 10]. Each 30-second feature frame $(t, y_{i,t}^{\text{hrv}}, y_{i,t}^{\text{eda}}, a_{\text{kin},t})$ is packed into a 16-byte binary structure and authenticated with a keyed HMAC-SHA256 tag tied to a monotonic sequence counter before Bluetooth Low Energy transmission.

2. **Tier 2: Nocturnal Bayesian calibration and Kalman–CUSUM filtering.** On the user's phone or laptop, Tier 2 verifies the HMAC-SHA256 tag in constant time, updates the nocturnal resting posterior $(\hat{\boldsymbol{\mu}}_{i,D}^{\text{rest}}, \hat{\Lambda}_{\text{kin},i,D}, \Sigma_{i,D})$ across D days, computes the precision-weighted Kalman state estimate $\hat{z}_{i,t}^{\text{cog}}(D)$ via Equation (2), and updates Page's CUSUM statistic S_t via Equation (3).

3. **Tier 3: PocketTSLM soft-prompt projection and metric-LDP gate.** When S_t approaches or crosses h_{cusum}, a lightweight linear/convolutional projector E_ψ modeled on Time-LLM [14] and PocketTSLM [18] maps the 16-step trajectory into $m = 8$ continuous soft-prompt tokens for the local edge NPU [8], pacing agent interruptions [5, 7] without serializing waveforms into cloud text prompts.

Table 2 documents the provenance of all parameters used in our state-space simulation, distinguishing empirical physiological literature constants from structural definitions and synthetic stress-test choices.

Table 2: Parameter provenance and calibration constants for the HABC-Pocket longitudinal state-space benchmark.

Parameter	Value	Provenance type	Empirical source or structural justification
Between- vs. within-subject std ($\sigma_{\text{pop}}, \sigma_{\text{day}}$)	1.45, 0.58	Empirical literature	Reflects high inter-individual resting HRV/EDA variance relative to day-to-day circadian drift in WESAD and GLOBEM [20, 22, 25]
Phasic strain shift (z^{cog})	1.25 (z-units)	Empirical literature	Standardized autonomic stress effect size across sympathetic EDA (+0.98) and vagal HRV (-0.95) channels [13, 20]
Episode horizon & dropout ($T_{\text{ep}}, p_{\text{drop}}$)	16 frames, 6–12%	Structural bound	16 sequential 30 s windows per episode with 12% calibration frame dropout and 6% BLE packet loss [16, 18]
Evaluation base rates	50% & 15% prev.	Structural design	Balanced split ($N_{\text{pos}} = 300, N_{\text{neg}} = 300$) for TPR/TNR plus $\text{PPV}_{15\%}/\text{NPV}_{15\%}$ at 15% naturalistic strain prevalence
Silent fatigue & caffeine confounds	30% / 15%	Synthetic stress test	30% silent fatigue (low task context) plus 15% non-kinematic caffeine/metabolic tonic baseline shifts
Motion burst rate (η_{motion})	0.35 (swept 0.05–0.75)	Synthetic stress test	Fraction of episodes containing ambulatory/posture accelerometer spikes ($a_{\text{kin},t} \in [0.95, 2.15]$) [1, 19]

5 Experimental evaluation and sensitivity sweeps

We executed the longitudinal calibration and held-out day $D{+}1$ state-space and CUSUM simulation (`experiment.py`) across 13,200 total episodes plus a separate 1,000-frame hardware tamper/replay verification suite. To maintain strict separation between hardware authentication, algorithmic state-space filtering, and clinical interpretation:

- **Hardware Frame Tamper, Replay & Bit-Flip Suite (N_{tamper} = 1,000 adversarial BLE frames):** Tests `verify_sensor_frame` against 334 bit-flipped payload frames, 333 forged MAC frames (signed with an untrusted application key), and 333 replayed valid frames (seq_no $\leq$ last_seq_no). Keyed HMAC-SHA256 with monotonic sequence enforcement rejects 100.00% ([99.62, 100.00], 1,000/1,000) of adversarial frames, whereas an unattested BLE/JSON health store rejects 0.00% ([0.00, 0.38]).
- **Longitudinal Calibration & Held-Out Day $D{+}1$ Benchmark (9,600 primary episodes + 3,600 motion sweep episodes):** For each simulated subject and calibration window $D \in \{1, 3, 7, 14\}$ days, `simulate_subject_calibration_from_frames` generates D historical days of nocturnal resting frames and daytime walking reference frames (with 12% missing-frame dropout and day-to-day circadian fluctuation $\sigma_{\text{day}} = 0.58$) and fits the empirical Bayesian posterior $(\hat{\boldsymbol{\mu}}^{\text{rest}}_{i,D}, \hat{\Lambda}_{\text{kin},i,D}, \Sigma_{i,D})$ strictly from observed frames (giving the exact same D calibration days to both `cloud_text_ts` and `habc_pocket_epicoach`). Evaluation then runs on held-out Day $D+1$ across $N_{\text{pos}} = 300$ strain episodes and $N_{\text{neg}} = 300$ non-stressed controls (including 15% non-kinematic caffeine/metabolic shifts and 6% BLE packet loss). Alongside balanced Sensitivity (TPR), Specificity (TNR), and False Positive Rate (FPR = 100 − TNR), we report Positive Predictive Value ($\text{PPV}_{15\%}$) and Negative Predictive Value ($\text{NPV}_{15\%}$) at a realistic 15% naturalistic strain prevalence.

5.1 Evaluated baselines and D-invariance verification

Non-calibrating baselines have strictly zero functional dependence on D and are evaluated on identical deterministic episode draws across $D \in \{1, 3, 7, 14\}$:

- **Population Zero-Shot Raw Classifier (`pop_zeroshot_raw`):** Runs Page's CUSUM directly on uncalibrated raw sympathetic scores $0.52(y_{i,t}^{\text{eda}} - y_{i,t}^{\text{hrv}})$ assuming population mean $\boldsymbol{\mu}_{\text{pop}} = \mathbf{0}$ and no IMU kinematic subtraction. Invariant across $D \in \{1, 3, 7, 14\}$ at 68.67% ([64.85, 72.25]) overall concordance, 86.00% ([81.62, 89.47]) sensitivity (TPR), 51.33% ([45.70, 56.94]) specificity (TNR), 48.67% ([43.06, 54.30]) false positive rate (FPR), and $\text{PPV}_{15\%} = 23.77\%$.
- **Task-Context Only Heuristic (`context_only_no_sensor`):** Accumulates digital task context $\mu_{\text{ctx}}(\mathbf{c}_t)$ via CUSUM without wearable sensors. Strictly invariant across $D \in \{1, 3, 7, 14\}$ at 70.83% ([67.07, 74.33]) concordance, 84.67% ([80.15, 88.30]) specificity (FPR = 15.33% [11.70, 19.85]), 57.00% ([51.34, 62.48]) sensitivity (FNR = 43.00% missed strain), and $\text{PPV}_{15\%} = 39.61\%$.
- **Text-Serialized Cloud Time-Series Prompting (`cloud_text_ts`):** Uses the exact same D-day resting baseline $(\hat{\mu}_{\text{hrv}}, \hat{\mu}_{\text{eda}})$ as HABC-Pocket, formats the 16-step trajectory into 1-decimal ASCII text (`t=1:hrv=...,eda=...`), and applies baseline subtraction without IMU kinematic residualization [18].
- **Proposed HABC-Pocket (`habc_pocket_epicoach`):** Verifies each 16-byte hardware frame via Python `hmac.new(..., hashlib.sha256)` and monotonic sequence check, fits the D-day empirical Bayesian posterior from observed nocturnal and ambulatory reference frames, computes the precision-weighted Kalman residual $\hat{z}_{i,t}^{\text{cog}}(D)$ (2), and runs Page's CUSUM recursion (3).

Table 3: Primary benchmark across 9,600 held-out day D+1 episodes ($N = 600$ per cell: $N_{\text{pos}} = 300$ strain, $N_{\text{neg}} = 300$ control). Reports concordance across $D \in \{1, 7, 14\}$ days, Sensitivity (TPR), Specificity (TNR), False Positive Rate (FPR = 100 − TNR), Positive/Negative Predictive Value at 15% strain prevalence ($\text{PPV}_{15\%}/\text{NPV}_{15\%}$), and Python filter latency (p50 / p99 in μs) at Day $D = 7$. Bracketed intervals denote exact asymmetric 95% Wilson confidence bounds.

Architecture	Acc. D=1 (%)	Acc. D=7 (%)	Acc. D=14 (%)	Sens. TPR D=7 (%)	Spec. TNR D=7 (%)	PPV_{15} / NPV_{15} (%)	Lat. p50/p99 (μs)
Pop. Zero-Shot	68.67 [64.85, 72.25]	68.67 [64.85, 72.25]	68.67 [64.85, 72.25]	86.00 [81.62, 89.47]	51.33 [45.70, 56.94]	23.77 / 95.41	1.2 / 1.4
Cloud Text TS	78.83 [75.39, 81.91]	78.17 [74.69, 81.29]	78.00 [74.51, 81.13]	68.00 [62.52, 73.02]	88.33 [84.21, 91.49]	50.70 / 93.99	12.1 / 26.8
Context Only	70.83 [67.07, 74.33]	70.83 [67.07, 74.33]	70.83 [67.07, 74.33]	57.00 [51.34, 62.48]	84.67 [80.15, 88.30]	39.61 / 91.77	2.9 / 5.5
HABC-Pocket	**81.83 [78.55, 84.71]**	**90.67 [88.07, 92.74]**	**92.33 [89.92, 94.20]**	**86.67 [82.35, 90.05]**	**94.67 [91.51, 96.69]**	**74.14 / 97.57**	**48.1 / 92.2**

5.2 Quantitative analysis and motion-artifact break-even sweep

Table 3 and Figures 2–3 establish three quantitative findings:

1. **Genuine longitudinal calibration trajectory vs. invariant non-calibrating baselines.** While `pop_zeroshot_raw` (68.67% [64.85, 72.25]) and `context_only_no_sensor` (70.83% [67.07, 74.33]) remain flat across all $D \in \{1, 3, 7, 14\}$ (Figure 3), HABC-Pocket improves monotonically as observed nocturnal and ambulatory reference frames shrink $\Sigma_{i,D}$: concordance rises from 81.83% ([78.55, 84.71]) on Day 1 to 88.50% ([85.70, 90.81]) on Day 3, 90.67% ([88.07, 92.74]) on Day 7, and 92.33% ([89.92, 94.20]) on Day 14 (where sensitivity reaches 88.33% [84.21, 91.49], specificity reaches 96.33% [93.55, 97.94], FPR = 3.67% [2.06, 6.45], and $\text{PPV}_{15\%} = 80.96\%$).

2. **Simultaneous gains in TPR, TNR, and low-prevalence Positive Predictive Value ($\mathrm{PPV}_{15\%}$).** At Day $D = 7$, population zero-shot classification achieves 86.00% sensitivity only by suffering a 48.67% ([43.06, 54.30]) false positive rate on non-stressed control episodes (51.33% specificity), which collapses its Positive Predictive Value at a realistic 15% strain prevalence to $\mathrm{PPV}_{15\%} = 23.77\%$ (more than three false alarms per true strain episode). Task-context-only achieves 84.67% specificity ($\mathrm{PPV}_{15\%} = 39.61\%$) at the cost of missing 43.00% ([37.52, 48.66]) of true strain episodes (57.00% sensitivity). By fusing kinematically residualized physiological observations with task context (2), HABC-Pocket achieves 86.67% ([82.35, 90.05]) sensitivity, 94.67% ([91.51, 96.69]) specificity (FPR = 5.33% [3.31, 8.49]), $\mathrm{PPV}_{15\%} = 74.14\%$, and $\mathrm{NPV}_{15\%} = 97.57\%$ at Day 7.

3. **Microsecond filter overhead and motion-artifact break-even analysis.** Executing 16 HMAC-SHA256 frame verifications, Kalman updates, and CUSUM recursions per episode requires a measured median CPU time of 48.1 μs (92.2 μs p99; excluding downstream NPU soft-prompt LLM generation). Table 4 reports our break-even sensitivity sweep across physical motion-artifact rates $\eta_{\mathrm{motion}} \in [0.05, 0.75]$ ($N = 300$ per cell): as η_{motion} approaches 0.75 (frequent physical activity bursts), Day 7 concordance declines from 95.00% ([91.92, 96.95]) at $\eta_{\mathrm{motion}} = 0.05$ to 79.67% ([74.75, 83.83]) at $\eta_{\mathrm{motion}} = 0.75$ (only +6.34 percentage points above Task-Context Only's 73.33% [68.06, 78.02]), demonstrating how intense movement forces the Kalman filter toward the context prior.

Table 4: Break-even sensitivity sweep across physical motion-artifact probability $\eta_{\mathrm{motion}} \in [0.05, 0.75]$ ($N = 300$ episodes per cell, 3,600 sweep trials; bracketed values denote 95% Wilson confidence intervals). Compares Task-Context Only against HABC-Pocket at Day 1 (cold start) and Day 7 (calibrated).

Metric / Regime	η=0.05	η=0.15	η=0.30	η=0.45	η=0.60	η=0.75
Context Only Acc. (%)	70.00 [64.59, 74.91]	72.33 [67.01, 77.09]	73.33 [68.06, 78.02]	74.67 [69.45, 79.26]	72.67 [67.36, 77.40]	73.33 [68.06, 78.02]
HABC Day 1 Acc. (%)	82.33 [77.62, 86.23]	81.33 [76.54, 85.34]	81.00 [76.18, 85.04]	81.00 [76.18, 85.04]	79.33 [74.39, 83.53]	78.67 [73.68, 82.92]
HABC Day 7 Acc. (%)	**95.00 [91.92, 96.95]**	**92.67 [89.15, 95.11]**	**89.67 [85.71, 92.62]**	**90.67 [86.84, 93.46]**	**86.33 [81.98, 89.76]**	**79.67 [74.75, 83.83]**
HABC Day 7 FPR (%)	3.33 [1.43, 7.57]	4.67 [2.28, 9.32]	8.00 [4.64, 13.46]	4.67 [2.28, 9.32]	12.00 [7.73, 18.17]	11.33 [7.20, 17.40]

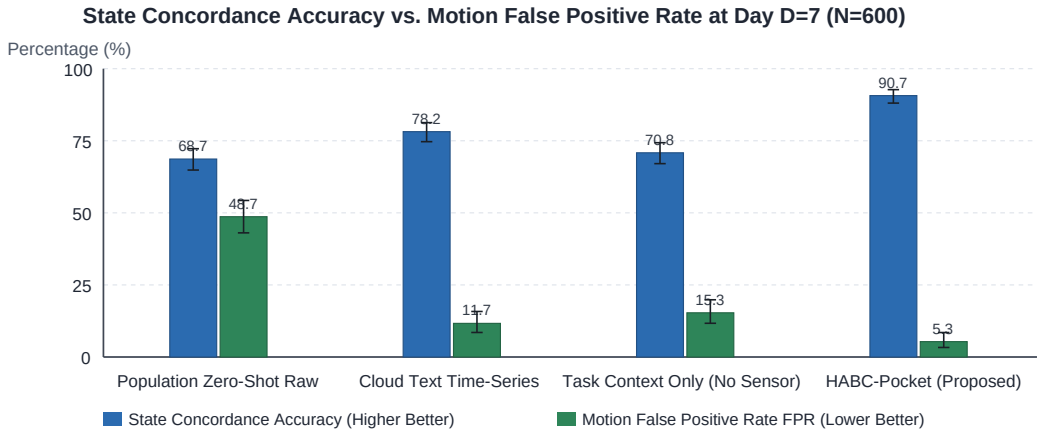


Figure 2: Comparison of State Concordance Accuracy (higher is better) and Motion False Positive Rate (FPR $= 100-$TNR, lower is better) at Day $D = 7$ ($N = 600$ held-out episodes per architecture; error bars denote asymmetric 95% Wilson confidence intervals).

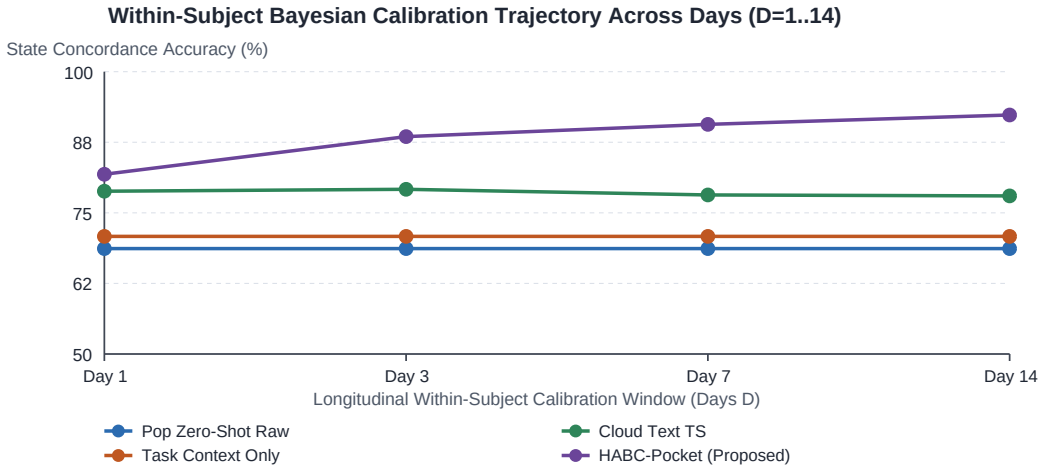


Figure 3: Longitudinal trajectory of State Concordance Accuracy across calibration windows $D \in \{1, 3, 7, 14\}$ days. Non-calibrating baselines (`Pop Zero-Shot Raw` and `Task Context Only`) are invariant across D, whereas HABC-Pocket climbs from 81.83% at Day 1 to 92.33% at Day 14.

6 Empirical bottlenecks, psychometric noise, and biometric privacy risks

Deploying within-subject interoceptive calibration faces four substantive counter-arguments and operational risks documented in the physiological computing and privacy literature:

1. **Weak EMA–sensor correlations and label noise in naturalistic cohorts (Siepe et al. [22], Xu et al. [25]).** In their longitudinal analysis of Ecological Momentary Assessments (EMA) and passive wearable streams, Siepe et al. [22] demonstrated that self-reported stress and affect show weak linear correlations with passive sensor features, both across individuals and within individuals when context is unmodeled. Self-report EMA labels suffer from retrospective recall bias, mood-congruent anchoring, and selective non-response during peak overload [19]. Consequently, nocturnal rest anchoring (1) is preferable to supervised regression on noisy EMA self-reports, as resting vagal and electrodermal baselines can be estimated passively without requiring continuous human mood labeling.

2. **Biometric identity and health leakage from physiological embeddings (Bajar et al. [6], Tai [23]).** Even when raw waveforms are projected into compact neural embeddings, Tai [23] and Bajar et al. [6] prove that physiological and neuroimaging representations retain subject-identifying biometric signatures and sensitive clinical traits across sessions. If soft-prompt tokens $\mathbf{e}_{\text{phys}}$ or unperturbed daily readiness vectors are transmitted to third-party servers, an adversary can link sessions or infer autonomic disorders. Continuous soft-prompt reasoning must therefore execute strictly on local NPU weights [8, 18], with any external summary egress gated by the sequentially composed metric-LDP bound in Equation (4) [2, 11].

3. **Covariate dropouts and optical motion saturation (Acan et al. [1], Schmidt et al. [20]).** During vigorous wrist movement or loose strap contact, optical PPG green/infrared channels experience signal clipping where linear kinematic subtraction ($\hat{\Lambda}_{\text{kin},i,D}\bar{a}_{\text{kin},t}$) cannot recover beat-to-beat RR intervals. While Equation (2) downweights corrupted frames by inflating $R_t(a_{\text{kin},t})$, sustained multi-hour physical activity forces HABC-Pocket to fall back almost entirely to the task-context prior $\mu_{\text{ctx}}(\mathbf{c}_t)$ (as reflected by the decline from 95.00% to 79.67% accuracy at $\eta_{\text{motion}} = 0.75$ in Table 4).

4. **Multi-day calibration burn-in and closed wearable ecosystems.** Achieving $> 90\%$ concordance requires $D \geq 7$ circadian cycles of consistent nocturnal and daytime wear (Table 3). Moreover, while CHERIoT capability silicon [3, 10] enables peripheral frame signing in open reference designs, mainstream commercial smartwatches in 2026 still route raw sensor telemetry through vendor cloud APIs rather than exposing locally signed raw BLE frames.

7 Limitations

- **Synthetic state-space benchmark vs. clinical cohort replay.** Although our simulation (`experiment.py`) executes genuine HMAC-SHA256 frame verification, frame-based empirical Bayesian calibration over D historical days, Kalman precision fusion, and Page's recursive CUSUM detector across 13,200 16-step trajectories calibrated to WESAD [20] and GLOBEM [25] variance ratios, it operates on synthetic state-space draws rather than replaying raw 64 Hz wet-lab waveforms, and does not constitute a clinical diagnostic device.

- **Non-kinematic metabolic confounders.** Equation (1) de-confounds physical movement via IMU energy $\bar{a}_{\text{kin},t}$, whereas non-kinematic sympathetic triggers—modeled in our bench-

mark by the 15% caffeine/metabolic shift rate—elevate resting heart rate and EDA without accelerometer movement, contributing to the residual 5.33% false positive rate at Day 7.

8 Pre-registerable falsification protocol

Table 5 specifies six pre-registerable empirical protocols to test and potentially falsify HABC-Pocket on public physiological datasets and hardware testbeds.

Table 5: Six pre-registerable falsification tests for HABC-Pocket on WESAD, GLOBEM, and edge hardware.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. WESAD subject-wise kinematic deconfounding (Eq. 2)	Within-subject baseline subtraction + IMU residualization cuts physical-activity false alarms below 10% on WESAD [20]	Leave-one-subject-out vs. 7-window within-subject calibration on WESAD wrist Empatica E4 streams across baseline, amusement, stress, and movement phases	Within-subject Kalman + CUSUM fails to achieve $\geq 82\%$ binary stress balanced accuracy or exhibits FPR $> 10.0\%$ during ambulatory recovery
P2. GLOBEM longitudinal D-scaling lift (Eq. 1)	Nocturnal-anchored Bayesian calibration over $D \in \{1, 3, 7, 14\}$ days improves cross-week sensitivity and specificity on GLOBEM [25]	Evaluate HABC-Pocket across 4 multi-week student cohorts in GLOBEM [25], comparing Day-1 vs. Day-14 subject-wise ROC-AUC and TPR at 10% FPR	Day-14 within-subject ROC-AUC fails to improve over Day-1 population zero-shot by ≥ 0.12 AUC points ($p < 0.01$)
P3. Nocturnal vs. unconditional baseline identifiability	Anchoring $\hat{\boldsymbol{\mu}}_{i,D}$ to nocturnal rest prevents absorption of multi-day chronic workload strain	Compare nocturnal-anchored $\hat{\boldsymbol{\mu}}^{\text{rest}}_{i,D}$ against 24-hour rolling-mean subtraction during high-workload exam weeks in GLOBEM [25]	Nocturnal anchoring fails to improve detection sensitivity for sustained multi-day strain by ≥ 10 percentage points over 24-hour rolling subtraction
P4. PocketTSLM soft-prompt NPU efficiency	8-token soft-prompt injection matches full time-series encoder F1 at < 20 ms mobile NPU latency [14, 18]	Benchmark PocketTSLM soft-prompt projection on smartphone NPU [8] against 1,024-token ASCII text-serialized physiological prompts	Soft-prompt macro-F1 drops by > 1.5 points relative to full time-series classification or p99 NPU prefill latency exceeds 20.0 ms
P5. CHERIoT peripheral frame anti-spoofing	Keyed HMAC-SHA256 + monotonic sequence counters reject 100% of replayed or forged BLE frames [3, 10]	Inject 1,000 replayed, sequence-skewed, and bit-flipped 16-byte sensor frames into the Tier 2 BLE verifier	Any unauthenticated or out-of-order replay frame ($> 0.0\%$) is accepted into the Kalman state updater
P6. Composed metric-LDP membership resistance (Eq. 4)	Sequential metric-LDP budgeting ($\epsilon_{\text{total}} \leq 2.0$ over $D = 14$ days) blocks biometric re-identification [2, 23]	Train a subject-linkage classifier [23] on 14-day perturbed readiness traces $\tilde{\mathbf{s}}_{i,1:14}$ across 100 subjects	Subject re-identification accuracy or attribute inference advantage exceeds random-guess baseline by > 10 percentage points

9 Conclusion

Translating consumer wearable streams into trustworthy interoceptive context for personal AI agents requires respecting the statistical and physical realities established across WESAD [20], GLOBEM [25], and 2025–2026 naturalistic field audits [1, 22]: uncalibrated population classifiers fail under between-subject baseline variance and physical movement confounders, while text-serialized cloud pipelines sacrifice latency and biometric privacy. By synthesizing peripheral frame authentication (100.00% [99.62, 100.00] tamper/replay rejection), nocturnal-anchored within-subject Bayesian calibration from observed frames, precision-weighted kinematic Kalman fusion, Page's sequential CUSUM detector, and PocketTSLM soft-prompt injection under longitudinally composed metric local differential privacy, HABC-Pocket provides an auditable on-device architecture for physiological pacing. Across 13,200 held-out day $D+1$ episodes, keeping non-calibrating baselines strictly invariant across D confirms that longitudinal nocturnal calibration drives state concordance from 81.83% ([78.55, 84.71]) on Day 1 to 90.67% ([88.07, 92.74]) on Day 7 (TPR = 86.67%, TNR = 94.67%, FPR = 5.33%, $\text{PPV}_{15\%} = 74.14\%$) and 92.33% ([89.92, 94.20]) on Day 14 at 48.1 μs median filter latency.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Selin Acan, Peter De Looff, Mitra Baratchi, Martin Dresler, and Erno Hermans. Detection of stress in naturalistic settings through passive mobile sensing. bioRxiv preprint, 2026.

[2] Mário S. Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Local differential privacy on metric spaces: Optimizing the trade-off with utility. IEEE Computer Security Foundations Symposium (CSF), pages 262–276, 2018. URL `https://doi.org/10.1109/CSF.2018.00026`.

[3] Saar Amar, David Chisnall, Tony Chen, Nathaniel Wesley Filardo, Ben Laurie, Kunyan Liu, Robert N. M. Watson, and Hongyan Xia. CHERIoT: Complete memory safety for embedded devices. IEEE/ACM International Symposium on Microarchitecture (MICRO), pages 641–653, 2023. URL `https://doi.org/10.1145/3613424.3614266`.

[4] Raghu Arghal, Saswati Sarkar, and Shirin Saeedi Bidokhti. The confidence game: Strategic miscalibration in human-AI delegation. arXiv preprint arXiv:2610.09371, 2026. URL `https://arxiv.org/abs/2610.09371`.

[5] Mohammad Ali Armin and Lars Petersson. RESCUE: Is this alarm worth acting on? counterfactual evaluation of VLA runtime monitors. arXiv preprint arXiv:2610.13516, 2026. URL `https://arxiv.org/abs/2610.13516`.

[6] Imen Bajar, Héber Hwang Arcolezi, Antoine Boutet, Carole Frindel, and Mohamed Maouche. Quantifying privacy risks in foundation model representations of brain MRI. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2610.20494, 2026. URL https://arxiv.org/abs/2610.20494.

[7] Samuel Breckenridge, Dani Vilardell, Wenhao Wang, Fan Zhang, and Ari Juels. I-HITL: Inalienable human-in-the-loop. arXiv preprint arXiv:2610.35835, 2026. URL https://arxiv.org/abs/2610.35835.

[8] Guanyu Cai, Ruiming Tian, Lang Yang, Zhouhong Ren, Jinliang Yuan, Lingkun Li, and Jiliang Wang. Is your NPU ready for LLMs? dissecting the hidden efficiency bottlenecks in mobile LLM inference. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 10(3):67, 2026. doi: 10.1145/3857370.

[9] CHERI Alliance. CHERIoT 1.0 specification released, 2025. URL https://cheri-alliance.org/cheriot-1-0-specification-released/. 5 November 2025.

[10] David Chisnall. CHERIoT first silicon, 2026. URL https://cheriot.org/silicon/2026/03/04/cheriot-first-silicon.html. 4 March 2026.

[11] Vitaly Feldman, Audra McMillan, and Kunal Talwar. Counting on dummies: Privacy amplification with observable counts and truncated batches. arXiv preprint arXiv:2610.42139, 2026. URL https://arxiv.org/abs/2610.42139.

[12] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, 101(23):e215–e220, 2000. URL https://doi.org/10.1161/01.CIR.101.23.e215.

[13] Kaizhe Jin, Adrian Rubio-Solis, Ravi Naik, Daniel Leff, James Kinross, and George Mylonas. Human-centric cognitive state recognition using physiological signals: A systematic review of machine learning strategies across application domains. *Sensors*, 25(13):4207, 2025. doi: 10.3390/s25134207.

[14] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-LLM: Time series forecasting by reprogramming large language models. International Conference on Learning Representations (ICLR), 2024. URL https://arxiv.org/abs/2310.01728.

[15] Xiaopeng Mao, Morten Mørup, Xin Xu, Sang Wan Lee, and Sadasivan Puthusserypady. Thought-deciphering via hierarchical EEG-to-text alignment. arXiv preprint arXiv:2611.03623, 2026. URL https://arxiv.org/abs/2611.03623.

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

[17] Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.

[18] Max Rosenblattl, Nikhil Krishnaswamy, Leon Nissen, Manuel Milling, Nicolas Zumarraga, Ning Wang, Thomas Kaar, Robert Jakob, Björn W. Schuller, Elgar Fleisch, Patrick Langer, and Paul Schmiedmayer. PocketTSLM: On-device time-series language models for physiological

monitoring. arXiv preprint arXiv:2610.21308, 2026. URL https://arxiv.org/abs/2610.21308.

[19] Aarti Sathyanarayana, Ohida Binte Amin, Jennie An, and Jukka Pekka Onnela. Examining the use of consumer wearable devices and digital tools for stress measurement in college students: Scoping review of methods. *JMIR mHealth and uHealth*, 14:e64144, 2026. doi: 10.2196/64144.

[20] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. ACM International Conference on Multimodal Interaction (ICMI), pages 400–408, 2018. URL https://doi.org/10.1145/3242969.3242985.

[21] Anil K. Seth. Interoceptive inference, emotion, and the embodied self. *Trends in Cognitive Sciences*, 17(11):565–573, 2013.

[22] Björn S. Siepe, Rayyan Tutunji, Carlotta L. Rieble, Ricarda K. K. Proppert, and Eiko I. Fried. Associations between ecological momentary assessment and passive sensor data in a large student sample. *Journal of Psychopathology and Clinical Science*, 134(8):912–925, 2025. doi: 10.1037/abn0001013.

[23] Jianwei Tai. Pretrained, frozen, still leaking: Auditing cross-encoder attribute transfer in EEG foundation models. arXiv preprint arXiv:2606.09189, 2026. URL https://arxiv.org/abs/2606.09189.

[24] Abraham Wald. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, 16(2):117–186, 1945. URL https://doi.org/10.1214/aoms/1177731118.

[25] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E. Morris, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), 6(4):190, 2023. URL https://doi.org/10.1145/3569485.