ADAPTIVE GATING IS JUST COUNTING: SCALAR-MEDIATED ENERGY-IDENTITY COLLAPSE IN FEDERATED CONTINUAL LEARNING UNDER CONCEPT DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

When adaptive inference systems (early exits, token pruning, gated adaptation) report energy savings, what should the savings be attributed to: the adaptive policy, or merely the operating point it induces? Under event-driven binary observation with a constant per-event cost (INA219 stage constants on Raspberry Pi 3B), we show that a controller's policy enters the energy identity only through a single scalar, the trigger count. By a Le Cam two-point argument the policy is unidentifiable from this channel, and gated adaptation stays within $2c_{\mathrm{ad}}$ of the best count-matched open-loop schedule—on hardware (-0.27 J, Wilcoxon $p = 7.1 \times 10^{-4}$; paired, n=20). The collapse carries falsifiable content: across all 680 arrival orderings the feasible grid ($V \in \{4, 8, 16\}$, $\Theta = V\, c_{\mathrm{ad}}^{\mathrm{harness}} \geq 1$, $c_{\mathrm{ad}}^{\mathrm{harness}} = 0.2896$ J) returns $N_{\mathrm{fire}} \equiv \lfloor A(T)/\Theta \rfloor$ exactly; at V=2 ($\Theta = 0.5792 < 1$) 134/680 (19.7%) deviate at (A, T)=(3, 17), diluting to 1.67% at the hardware window (A, T)=(5, 300). The policy can matter; it cannot bypass one scalar. The claim is falsified precisely where $\Theta < 1$. Given the event stream, the ledger is auditable without knowing the policy (against under-reporting and replay), and it locates where adaptation can still pay: state-dependent costs (Thm. 2(iii)).

1 INTRODUCTION

Early exits, token pruning and gated adaptation routinely report energy savings. But when the observation channel is compressed to event-driven binary signals, can the energy benefit of adaptive inference be attributed to the policy itself, or only to the number of events it happens to trigger?

We study this question on federated continual learning (FCL) nodes under concept drift, where a detector watches the stream and a gate decides whether to spend energy on adaptation. Two facts motivate the question. First, the cost of getting it wrong is physical: on our Pi 3B testbed a detection–gating controller that fires on drift noise burns 247.4 J over 395 triggers (a loop-level total over all seven phases, not a per-event cost, which is 0.29–0.47 J; Table 7) while its held-out F1 stays at chance, whereas on a silent stream it spends 0 J adapting. Second, reported savings are polarized for the very same controller (46.9%: 247.4 J $\rightarrow$ 131.3 J over 2000 rounds in a single metered batch; 0 J adapting on the silent stream), so the savings evidently track the operating point, not the controller. Concept drift also has a mechanism-level signature: representation norms spike at drift onset (Chen et al., 2024), which is exactly when a trigger-happy gate spends the most.

Meanwhile the energy numbers are unaudited: Green-FL carbon accounting is estimate-based and uncertified against wall power (Tapp et al., 2026), MLPerf Power standardizes whole-system benchmarks (Tschand et al., 2024), but per-phase metered ledgers of the deployment loop with uncertainty budgets and audit predicates have, to our knowledge, not been published. Continual learning is shifting from algorithm design to deployed systems (Hou et al., 2026), where energy is the deployment gate. The double gap: no regret guarantee for gated adaptation, and no way to check a self-reported joule.

This paper. We instrument the adaptation loop end-to-end (INA219, per-phase ledgers, SHA-256 anchors) and prove that, at constant per-trigger cost, the controller's policy enters the energy identity

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

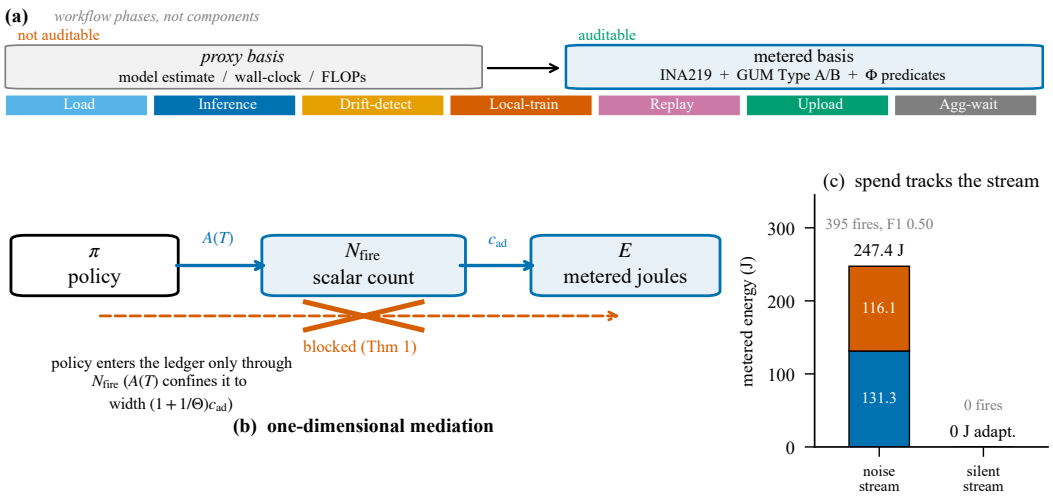


Figure 1: **The auditable substrate.** (a) proxy vs. metered basis (labels only) over the seven workflow phases; (b) scalar mediation: the policy π affects metered energy E only through N_{fire}, which $A(T)$ confines to width $(1 + 1/\Theta)c_{\text{ad}}$; the direct channel is blocked (Thm. 1); (c) the same controller on two streams (247.4 J over 395 triggers on a noise stream, 0 J on a silent one, 46.9% of it avoidable at parity F1), so spend tracks the operating point, not the policy (the chance-level F1 is a null-stream control, unlike RQ2's detection F1 on genuine-drift streams).

only through the scalar N_{fire}: the policy can change how many times the meter ticks, not what each tick costs (Fig. 1b). The claim is falsifiable (Fig. 4): exact on the feasible grid ($0/680$ orderings deviate), not below it ($134/680$). Our contributions:

- **C1 (identity).** The policy enters the energy identity only through the scalar N_{fire}; in the feasible region $N_{\text{fire}} \equiv \lfloor A(T)/\Theta \rfloor$ exactly, timing-included (Lemma 1).
- **C2 (equivalence).** Gated and count-matched open-loop schedules differ by at most $2c_{\text{ad}}$, with the caliber named (Thm. 2(i)); hardware agrees: $|\Delta E| = 0.27$ J, inside the bound (paired n=20, $p = 7.1 \times 10^{-4}$).
- **C3 (unidentifiability), in the energy channel and at matched trigger count.** A Le Cam two-point construction gives total variation 0 on the scalar count channel (Thm. 1); nothing about the policy beyond its trigger count is recoverable from the ledger.
- **C4 (hardware verification).** Pi 3B+INA219 evidence chain: -0.27 J head-to-head, $Q_{\max} \equiv V$ exact on the metered trace, gate-credibility AUROC $= 1.0$.

Auditing here is an instrument: the predicates validate the measurement chain (§5). The findings stand on the identity, the bound, and the metered agreement; savings are attributable to round reduction, verified against oracle fixed-schedule and plateau early-stop controls.

The structural claim. What is at issue is the structure of the identity: the policy's energy footprint factorizes through a single dimension, so attribution questions have one answer on this channel—not the bookkeeping identity $E = \text{triggers} \times \text{constant}$. Timing, magnitude, or co-adaptation entering the ledger separately would falsify it, and Fig. 4 is that test. The shape channel ρ stays partially informative ($38/60 = 0.6333$, Wilson CI $[0.507, 0.744]$): the policy can be informative and still cannot bypass the scalar in the energy channel, where Thm. 2(iii) identifies where adaptation can pay.

2 Related Work

Table 1 positions this paper along six axes; Appendix J gives the full grid. Three sentences on the theory lineage: the price of adaptivity has been quantified at the statistical layer (Chakraborty & Maity, 2026) and the optimization layer (Kreisler et al., 2026); we work at the deployment layer, where the quantity being metered is a physical joule.

Table 1: Six-axis positioning. "J?" = are physical joules metered; "Aud?" = is the measurement path auditable. TIMEGATE is a workshop paper (ICML'26 WS) plus arXiv.

Axis	Representative work	Energy accounting	Caliber	J?	Difference of ours
FCL/CL	Li et al. (2025a), Yu et al. (2026), Yoon et al. (2021), Shoham et al. (2019), Dong et al. (2022)	RAM/FLOPs budgets	abstract	✗	joule-level budget primitive
Drift-gated adaptation	FLAME (Mavromatis et al., 2024), FDAL survey (Mahdi et al., 2025)	none	—	✗	adaptation cost made auditable
Budgeted Lyapunov sched.	RCCDA (Piaseczny et al., 2025), Park et al. (2026)	abstract λ	—	✗	metered constants + regret bound
Price of adaptivity	Chakraborty & Maity (2026), Kreisler et al. (2026), Hor et al. (2024)	gain bounds	—	✗	deployment layer: joules, indistinguishability
Energy measurement	Shi et al. (2022), Tschand et al. (2024), Tapp et al. (2026), Yan et al. (2024)	estimates / whole-system	wall / model	✓ bench.	per-phase metered ledger + GUM + Φ
Provenance	Longpre et al. (2025)	data licensing	—	✗	energy-domain provenance (hash anchors)

Direct competitors. Cost-matched comparison is not new: Hor et al. (2024) bound the achievable gain of adaptive inference under instance-varying per-unit cost. What is new here is that the quantity being matched is the *trigger count* under *constant physical cost*, which yields an *indistinguishability* rather than a gain bound, and thereby delineates the constant-cost/variable-cost boundary; Thm. 2(iii) is the bridge: making c_{ad} queue-state-dependent moves deployment from our degenerate regime into theirs. Concurrent work gates the promote/hold decision of periodic re-evaluation with a calibrated metric-availability signal (Chakraborty et al., 2026); we gate adaptation triggers and meter each trigger's joules with phase-level ledgers, calibrated uncertainty, and self-contained regret bounds—an audit of the adaptation layer rather than the evaluation layer. O-FCL (Serra & Buettner, 2025) is drift-free and unmetered; FOT (Bakman et al., 2024) needs known task boundaries (ours is task-free); FedCCFA (Chen et al., 2024) evaluates in simulation on RTX-4090s with zero joule accounting; FedSSI (Li et al., 2025b) counts wall-clock time only. Sample-level mistake gating (Pache & van Rossum, 2026) skips updates on correctly classified samples to save energy implicitly: it reports update-count savings, never measured joules; we gate and meter physical joules per event. Among FCL systems under tight resource budgets, existing methods degrade uniformly as budgets shrink (Li et al., 2025a), yet budgets remain RAM/FLOPs abstractions, not joules.

Measurement and provenance. Optimized energy numbers in TinyML/FCL pipelines are not audited for how they were measured (Shi et al., 2022; Solowjow & Trimpe, 2020; Anastasi et al., 2009); heavyweight training-integrity proofs (TEE/ZK; Proof-of-Learning (Jia et al., 2021) is spoofable at reduced cost (Zhang et al., 2022)) are impractical at Pi-class budgets, which is why we use physical metering with hash anchors as the lightweight alternative. Data provenance is an established ICLR subject (Longpre et al., 2025); we instantiate provenance in the metered-energy domain, separately from data provenance. ProFL (Wu et al., 2025) is the closest sibling on resource-constrained heterogeneous FL.

3 Problem and Model

Deployment and threat model. A Pi-class edge node runs inference continuously. A drift detector observes the stream and emits binary flags; an energy-gated controller converts flagged events into adaptation triggers; each trigger spends joules measured by an INA219 sensor on the supply rail. An adversary (or simply a noisy world) controls the drift arrival process $A(T)$, the number of binary flag events up to time T. We audit the energy ledger, not the detector: the ledger must remain interpretable even when the detector is fooled. Bounds below are stated for every $A(T)$; the hardware run instantiates one arrival window ($A{=}5$), so no measured ratio is extrapolated to arrival rates not evaluated.

Assumptions.

Assumption 1 (metered cost identity)**.** *Every monitored step costs c_{mon} and every adaptation trigger costs c_{ad} (Table 7), both INA219 stage constants. On hardware the per-event cost has dispersion (n=173 events, mean 0.4712 J, CV 18.0%, range $2.94\times$); we state when this dispersion is load-bearing.*

Assumption 2 (binary evidence, no decay)**.** *Flag events are binary and the debt queue does not decay between events.*

Assumption 3 (deployment grid)**.** *The deployment grid is $V \in \{2, 4, 8, 16\}$ with $\Theta = V\, c_{\mathrm{ad}}^{\mathrm{harness}}$ ($c_{\mathrm{ad}}^{\mathrm{harness}} = 0.2896$ J), giving $\Theta \in \{0.5792, 1.1584, 2.3168, 4.6336\}$. Feasibility requires $\Theta \geq 1$, i.e. $V \geq 1/c_{\mathrm{ad}}^{\mathrm{harness}} = 3.4530$: one step admits at most one arrival but can absorb Θ, so the queue is stable iff $\Theta \geq 1$. V here is the deployment tier; other uses of V in our codebase (boundary-aware switching, tsweep grids) are time-varying and are excluded from the identity's scope.*

Assumption 4 (bounded reward)**.** *Per-step reward is bounded; the cost functional is $J(T) = \mathrm{Regret}(T) + \mu\, E(T)$ with metered $E(T)$.*

Assumption 5 (constant per-trigger cost within a run)**.** *c_{ad} does not depend on the queue state within a run. §6 (RQ6) measures what happens when it does.*

Controller under audit. DACT (Drift-Aware Cost-bounded Trigger) is a three-layer scheduler: a passive monitor (ADWIN over the drift-sensitive stream), an always-on adaptation layer, and an energy-debt gate. The gate maintains a debt queue Q; on each flagged event $a_t \in \{0, 1\}$ it updates $Q \leftarrow \max(0,\, Q + a_t)$, fires iff $Q \geq \Theta$, and on firing absorbs $Q \leftarrow \max(0, Q - \Theta)$ (at most one absorption per step). The closed-form refractory boundary $\delta^* = e_{\max}(\varrho + \alpha B)$ and the two-tier switching rule for $V(t)$ are given in Appendix A (symbols defined there).

4 Theory: the ledger sees one scalar

All proofs are in Appendix A. Energetic claims are run-level (E vs. $A(T)$); level is flagged wherever phase-level observations appear. The identity itself is definitional under Assm. 1; its falsifiable content is the exactness boundary: $N_{\mathrm{fire}} \equiv \lfloor A(T)/\Theta \rfloor$ whenever $a_{\max} \leq \Theta$, so $\Theta < 1$ is the escape regime, violated by a single run with $\Theta \geq 1$ in which $N_{\mathrm{fire}} \neq \lfloor A(T)/\Theta \rfloor$ at some event—verified by 680-order enumeration ($0/680$ vs. $134/680$) and $36/36$ hardware reruns.

Lemma 1 (counting identity)**.** *Under Assumptions 2–5 with $\Theta \geq a_{\max}$ ($\Theta \geq 1$ under Assm. 2; the general form is Prop. 2, App. H): $Q(T) \in [0, \Theta]$ (invariant, verified exhaustively), $N_{\mathrm{fire}}(T) = \lfloor A(T)/\Theta \rfloor$ exactly, and*

$$E(T) \;=\; T\, c_{\mathrm{mon}} \;+\; c_{\mathrm{ad}}\, \frac{A(T)}{\Theta} \;+\; r(T), \quad -c_{\mathrm{ad}} \leq r(T) \leq 0,$$

a path-determined function of the drift-arrival count alone: an identity, not an expectation or a worst case. The policy enters only through N_{fire}, which $A(T)$ confines to an interval of width $(1{+}1/\Theta)$—a one-dimensional mediation, not an absence.

Remark 1. (i) Since $\Theta = V c_{\mathrm{ad}}$, the leading term satisfies $c_{\mathrm{ad}}\, A(T)/\Theta = A(T)/V$: under the harness calibration c_{ad} cancels exactly and $E = T c_{\mathrm{mon}} + A(T)/V + r(T)$; under the metered calibration a residual factor 1.2607 $(= 0.3651/0.2896)$ remains—the caliber must be named whenever the two are combined. (ii) This identity is the literal meaning of a metering-debt gate: threshold price and per-trigger cost are one constant in the implementation. (iii) Counting is independent of the price: writing κ for the gate's service quantum and c_{ad} for the measured per-event energy, $N_{\mathrm{fire}} = \lfloor A(T)/(V\kappa) \rfloor$ remains a deterministic function of $A(T)$ for every $\kappa > 0$. (iv) Conversely, with a non-energy-calibrated threshold neither the counting identity nor the $2c_{\mathrm{ad}}$ equivalence holds automatically; the circularity objection is itself the finding.

Theorem 1 (strategy non-identifiability; energy channel)**.** *Let the observation space be the scalar count channel N_{fire} with per-trigger cost c_{ad} constant. There exists a pair of policies π_0, π_1 with identical trigger counts (count-matched) such that the induced distributions on the count channel coincide: $d_{\mathrm{TV}} = 0$. Consequently no test on this channel achieves power above $1/2$ for distinguishing them, and the policy is unidentifiable from the ledger. The construction is a two-point Le Cam argument; the observation space explicitly excludes the shape channel ρ of the underlying signal.*

Theorem 2 (early-stopping equivalence)**.** *We first characterize when gating necessarily degenerates to open-loop early stopping, then give the condition under which it separates. (i)* Degeneration. *There exists an open-loop periodic schedule P^* (constructed from $A(T)$, oracle-matched; the*

comparison direction favors the gate) with $|E_{\text{gate}} - E_{P^*}| \le 2c_{\text{ad}}$, caliber named. (ii) Hardware agreement. *On the paired n=20 head-to-head the measured gap is Hodges–Lehmann -0.27 J (exact Wilcoxon $p = 7.1 \times 10^{-4}$, 95% CI $[-0.32, -0.21]$); the robust estimator is chosen because one cold-start point pulls the arithmetic mean gap to -0.13 J, which is $46.6\%/37.0\%/34.4\%$ of $2c_{\text{ad}}$ under the harness/metered/M5 calibers and smaller than one service quantum: the observed gap lies inside the discretisation residual predicted by Lemma 1. On a static stream the gate saves the most ($+85.6\%$)—the $\varphi \to 0$ limit of P^*. (iii)* Separation. *Departure from early stopping occurs only when per-trigger cost or benefit is a function of the queue state; with $\Delta E = \sum_{t \le T} \Delta c_{\text{ad}}(Q(t))\,\mathbf{1}[\text{fire}]$, constant cost gives $O(c_{\text{ad}})$. Two regimes divide the literature by trigger source: constant-cost (this paper, and the counting regime) versus state-dependent (adaptive-inference gains (Hor et al., 2024)); the partition applies to our own scheduler as well.*

Theorem 3 (constants-explicit regret)**.** *Under Assumptions 2–4, the sparse-adaptive-regret bound of the gated scheduler, with each term anchored in INA219-metered constants ($c_{\text{mon}} = 0.0517$ J/step, $c_{\text{ad}}^{\text{metered}} = 0.3651$ J/event, CI $[0.3547, 0.3758]$), is $O\big(\sqrt{T \log T} + D\, N_{\text{switch}}\big)$ with the switching term $\le D \cdot A(T)/\Theta$ in the feasible region. Robustness: if $|c_{\text{ad}} - \hat{c_{\text{ad}}}| \le \delta$ then the bound inherits an additive correction $O(\delta\, A(T)/\Theta)$ with $\delta \approx 0.011$ J from the calibrated interval. The closed-form structure is stated as such; the numeric substitution is illustrative.*

Theorem 4 (fixed-threshold gap and time-varying thresholds)**.** *(i) A fixed threshold Θ with time-varying continuation value (drift makes the marginal value μ_{DP} non-stationary) incurs a closed-form approximation gap relative to the reservation-value benchmark; we verify both regimes (pre- and post-onset) explicitly. (ii) Under piecewise-stationary binary arrivals and the first-order trigger rule, a time-varying threshold $\Theta(t)$ achieves $O(\sqrt{T \log T})$ regret, and $O(\sqrt{(C_T + 1)T \log T})$ with C_T detected change-points, by reduction to the proved regret result of Wang (2026). Under time-varying $\Theta(t)$ the identity $A(T) = \Theta\, N_{\text{fire}} + Q$ no longer holds: Lemma 1 and Thm. 2 constrain fixed-Θ deployment systems; Thm. 4 concerns a hypothetical variant.*

Theorem 5 (closed-form refractory boundary)**.** *The gate admits an adaptation round only if the estimated drift magnitude satisfies $\delta \ge \delta^* = e_{\max}(\varrho + \alpha B)$. Empirically the predicted kink is not resolvable at n=3 (below the noise floor); we state agreement with the model, not a confirmed phase transition, and refer the empirical counterpart to the drift-strength sweep of RQ6.*

Predictions. The identity makes point predictions: $E = 103.19 + N \times 0.3651$ gives 157.955 J at N=150 and 194.465 J at N=250 (three-point verification, preregistered; the third point is the intercept anchor, Fig. 2a). Its per-round form $N_{\text{fire}} = \lfloor A(T)/\Theta \rfloor$ is now verified directly on hardware (§H): a per-event rerun (Pi 3B+INA219, T=300, 3 runs $\times$ 3 seeds $\times$ $V \in \{2, 4, 8, 16\}$) matches it in all $36/36$ node-runs at this window (A=5).

Relation to adaptive-inference theory. Function-space results locate catastrophic forgetting in a few NTK feature modes (Hidekel & Raviv, 2026), consistent with the per-trigger value $\mu\Delta R$ being time-varying under drift; we use this as a compatibility remark, not a dependency.

5 Auditable energy substrate

The identity of Lemma 1 is only useful if the meter is trustworthy. We make every joule carry its provenance.

Proposition 1 (single-scalar accounting is forgeable)**.** *Any accounting that reduces a run to a single scalar (total joules, total steps) cannot distinguish honest execution from replay or under-reporting at matched totals: the map from execution traces to scalars is many-to-one. Audit therefore requires phase-level structure.*

Phase-level ledger. Each run is recorded as a seven-phase workflow (Load, Inference, Drift-detect, Local-train, Replay, Upload, Agg-wait) with per-phase duration, energy, debt-gate events, accuracy and communication bytes; the published ledger has 938 rows (Table 5). The ledger is a workflow decomposition, not an energy component model. Audit and attribution are different layers (Prop. 1 vs. Thm. 1; App. H).

Audit predicates. Φ is a conjunction of three predicates: channel adequacy (each phase's energy is covered by a calibrated channel), minimality (no phase double-counted), and consistency (per-round sums match the reported totals). Injecting faults into the export pipeline, Φ catches all of

Table 2: Gate-credibility evidence chain (all Pi 3B+INA219, `simulated: false`).

Exp	Question	Result
E0	sensor noise floor	$\sigma_{\mathrm{dyn}} = 0.009320$ J (n=100; n-dependent)
E1	cost model	$\alpha = 1.943 \times 10^{-10}$ J/FLOP single-parameter, CI positive, $R^2 = 0.990$
E2	under-report	AUROC $= 1.0$; honest 7.5116 vs. 2.7543 J ($\approx 503\sigma$)
E3	over-report	undetectable from energy (0.49); replay probe restores AUROC $= 1.0$
E4	evasion matrix	$2/6$ arms undetectable (preregistered); time-dilution detectable ($+57.6\%$)
E5	detection gate	TPR $80/80$ ($G1 \wedge G2 \wedge G3$; exact one-sided 95% LB 0.9632); honest-test FPR $0/80$ (CP upper 3.68%); cross-session $0/200$ (1.49%)
E7	masking	two independent energy contrasts: $+22.03\%$ (perm. p=5×10^{-5}) and $+27.73\%$ ($= 550\,\sigma_{\mathrm{static}}$)

them: $938/938$ at ledger level and $381/381$ at event level, zero leakage (Appendix C). Injected-fault detection is a property of the instrument; it does not enter the measured loop and adds zero joules to reported numbers.

Gate credibility on hardware. Across E0–E7 the audit catches the failures that matter: silent under-reporting is detected at AUROC $= 1.0$ (honest 7.5116 J vs. under-reported 2.7543 J, $\Delta \approx 503\sigma$); over-reporting is not detectable from energy alone (0.49) but a known-cost replay probe restores AUROC $= 1.0$; of six evasion arms, two remain undetectable, and fusing all arms degrades AUROC from 1.000 to 0.753 (dilution), so we do not fuse. Table 2 summarizes the evidence chain.

Calibers. Every single-value energy constant in this paper names its caliber (Table 7); cross-caliber intervals are never formed. Wang (2026)'s proved regret and threshold-optimality results anchor Thm. 4 alone.

6 Experiments

All numbers below are Pi 3B+INA219 metered unless marked. Each headline is tagged with its evidence level: **M**(etered) / **R**(eplayed) / **S**(imulated) / **P**(redicted); simulated numbers never enter an INA219 confidence interval.

Preregistered interpretation rule. Before analysing the three-arm comparison we fixed: if the 95% interval for the energy difference between gated control and vanilla plateau early stopping lies within the interval predicted by Thm. 2(i), we report the contribution as an auditable quantification of adaptation cost, not as superiority of gated control. Statistics: cluster bootstrap with *seed* as the cluster unit (stream, seed) across streams; the random-skip arm matches the gated arm's realized skip rate per cell (B, V, seed) (per-cell binary calibration, 5% tolerance; both matched and unmatched versions reported); three-point extrapolations are labelled *three-point verification*.

The five per-trigger calibers, their objects and their roles are tabulated in Appendix D (Table 7); they are never mixed. Energy is metered on a single Pi 3B node; accuracy comes from a 5-client software-simulated FL driver (2/5 clients drift asynchronously).

RQ1: where do the joules go? Training dominates the loop; drift detection costs 0.0145 J per check (Table 5 in Appendix B). The drift window costs $7.81\times$/$7.57\times$ the steady window (protocol-annotated; both factors reported). With the replay gate set to its floor ($\lambda \leq \lambda_c$), 100-round energy floors at 15.9–16.6 J with marginal adaptation 0.143 ± 0.024 J; λ=0.10 adds 11.1%, and an ungated counter-forecast from the 30.5 J baseline extrapolates outside the measurement.

RQ2: does gating save energy? Yes, and the saving tracks the trigger count, not the controller. On three M5-Pi blocks (100 rounds each, same arms as §1): ADWIN-always spends $28.55/28.74/27.30$ J with 60 triggers; DACT-gated spends $18.40/17.63/16.84$ J with $37/39/37$ triggers (35.6–38.7% less, F1 within 0.001–0.042)—all **M**. On the E6 protocol the metered run drops from 20.27 to 2.06–6.46 J. Ablation moves the whole saving onto the trigger count: ablating

Table 3: Harder-drift three-stream suite (RQ6), metered, bootstrap 95% CIs. Level: M.

Stream	Energy saving	CI	Δacc	Rounds
SEA	$+75.52\%$	$[72.85, 78.19]$	0 (20/20)	—
CovShift	$+65.63\%$	$[60.67, 70.59]$	-0.0012	—
Hyperplane	-14.77%	$[-49.27, 19.73]$	-3.61pp*	$4.0 \to 11.5$

*CI $[-0.0577, -0.0144]$, excludes 0.

Table 4: Main results (metered; calibers named, never mixed). Levels: M/R, acc: S.

Method	Energy [J]	vs. P^*	N_{fire}/trig.	F1 / acc (level)
ADWIN-always (M5-Pi, 3×100 r.)	27.30–28.74	—	60/block	0.932–0.940 (S)
DACT-gated (3×100 r.)	16.84–18.40	-35.6–-38.7%	37–39/block	0.897–0.931 (S)
EBDA (energy-budget debt-aware, paired, n=20)	0.712 (mean)	-0.27 J (HL)	gated	acc_{kept} 0.949 (S)
Oracle P^* = FUS-4 (n=20)	0.587 (mean)	0 (reference)	fixed 4	acc_{kept} 0.948 (S)
FedAvg (no adaptation, 30 r.)	186.23	—	—	not tracked (M)
DP-SGD naïve (R=20)	1.95 ± 0.18	—	—	— (R)
Opacus DP (R=20)	0.64 ± 0.07	—	—	— (R)

the ADWIN (Bifet & Gavaldà, 2007) calibration returns the loop to FedAvg-level (McMahan et al., 2017) energy (185.19 vs. 186.23 J) while the remaining components change little (94.2% attribution, leave-one-out). Gating is the means of changing the count, not a saving mechanism of its own. The gate's robust core is violation-rate control: default-rate $3.18\% \to 1.03\%$ (paired $p \leq 4 \times 10^{-5}$, sign $17^+/0^-$, $6/6$ tiers) with $Q_{\max}$ compressed $\sim$70% and $Q_{\max} \equiv V$ exact on hardware; ***efficiency gain remains a conditional claim*** (merge setting $+11.5\%$ to $+61.5\%$, $5/6$ significant; nomerge default -5.0% to $+9.8\%$, $1/6$).

RQ6: the middle regime (core). Net benefit is modulated by whether accuracy saturates (Table 3): SEA saves $+75.52\%$ $[72.85, 78.19]$ at zero accuracy change (20/20); CovShift saves $+65.63\%$ $[60.67, 70.59]$ at Δacc -0.0012; Hyperplane *costs* -14.77% $[-49.27, 19.73]$ (CI crosses zero) with Δacc -3.61pp (CI $[-0.0577, -0.0144]$, significantly negative) and adaptation rounds $4.0 \to 11.5 \pm 6.68$. Because SEA's accuracy axis has zero variance, we plot a degeneracy-diagnostic scatter (energy vs. accuracy with an inset magnifying the zero-variance axis), not a Pareto front. The drift-strength sweep brackets the boundary: $\delta \leq 2.0$ stays near-oracle at $\sim$18% always-adapt energy; $\delta \geq 3.0$ enters the adversarial regime. Under merge modelling the replay-n=120 paired comparison is fully balanced (energy -1.22 J, p=0.65; count $+0.058$, p=0.78; accuracy $+0.0033$, p=0.58; utility-per-joule $+0.00676$, t=5.94)—strongest evidence for the conditional (merge) half of the claim.

RQ3–RQ5: baselines, audit, ablation. Table 4 gives the head-to-head. The audit results are in Table 2. Fig. 3 is the punchline: count-matched pairs are indistinguishable on the energy channel ($d_{\text{TV}} = 0$) while the shape channel ρ still separates them ($38/60 = 0.6333$, Wilson CI $[0.507, 0.744]$; per-session escape $0.350/0.300/0.450$)—a different channel, not a counterexample. The V-sweep reproduces the tier structure of Assumption 3; gating costs 7.2pp accuracy on the replay harness (negative result).

RQ7: cross-platform (same wall-clock caliber). Android/MatePad 77.1% vs. Pi 78.2% saving, both figures training wall-clock and neither metered energy; the MatePad telemetry link blocks full per-phase capture, so this comparison establishes form transfer only, and no Pi 3B joule claim rests on it.

RQ8: privacy cost. Against the published non-private constant $C_{\text{fix}} = 0.2673$ J: naïve DP-SGD costs $7.3\times$ (1.95 ± 0.18 J, CI $[1.87, 2.04]$), Opacus $2.4\times$ (0.64 ± 0.07 J, CI $[0.61, 0.67]$), R=20, Student-t df=19. The three candidate baselines ($0.0972/0.2673/0.3651$ J) span $3.8\times$, so the ratio alone is never reported without them; the metering-debt budget makes an explicit (ε, σ)-feasibility region (Appendix E). Local-DP results for privatized change-point detection (Zhang et al., 2026) act on the observation stream, a different mechanism from DP-SGD on gradients; we cite the qualitative analogy only and do not compose the two premiums. Calibration validity is time-bounded: the paired detection budget stays valid for ≥ 2.86 h (test-segment pairs; the 2.46 h calib-segment value is the sensitivity bound, Appendix F).

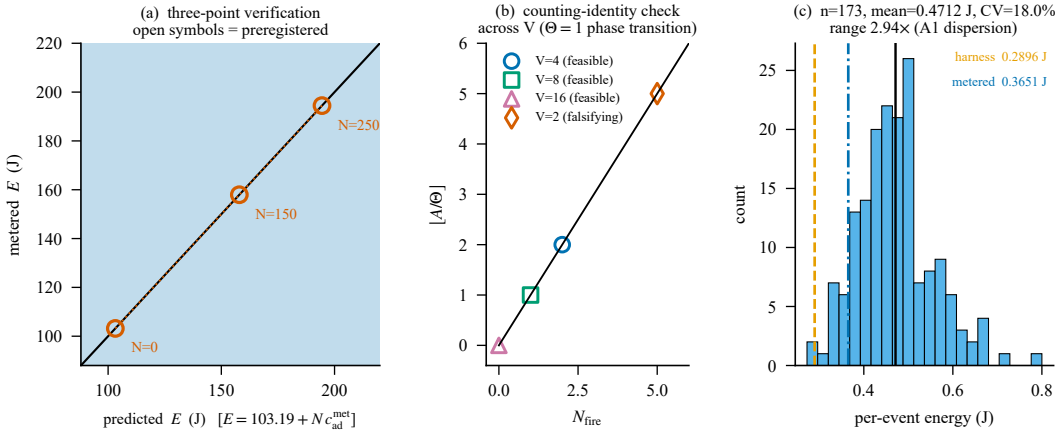


Figure 2: **Identity closure and A1 dispersion.** (a) three-point check of $E = 103.19 + Nc_{\text{ad}}^{\text{met}}$ (open = preregistered; band $\pm c_{\text{ad}}^{\text{met}}$). (b) counting-identity check $N_{\text{fire}} = \lfloor A/\Theta \rfloor$ at the feasibility line Θ=1 (V=3.4530); V=2 is the falsifying cell. (c) per-event energy (n=173, CV 18.0%): rules mark the two calibers used in (a)/(b).

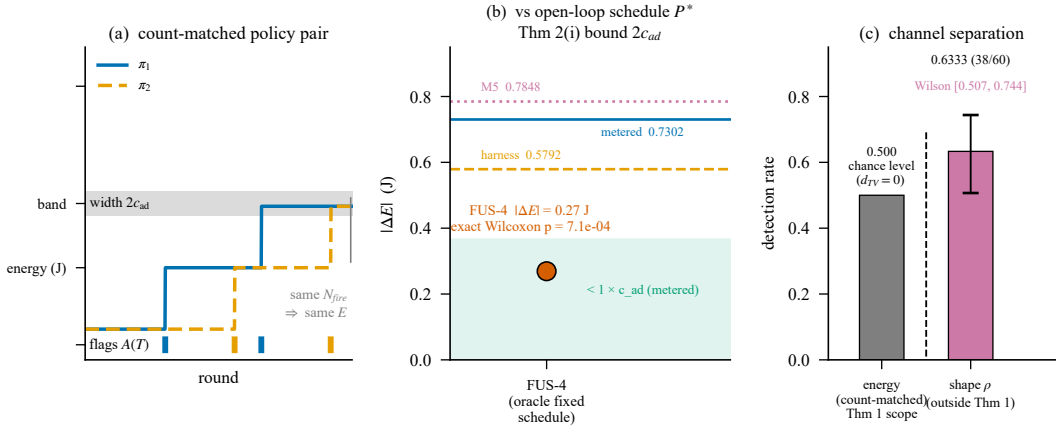


Figure 3: **Policy pairs are count-matched.** (a) two policies with different trigger timing and the same N_{fire} stay inside the $2c_{\text{ad}}$ band. (b) FUS-4 pair (n=20): HL $|\Delta E| = 0.27$ J (95% CI $[0.21, 0.32]$), $p = 7.1 \times 10^{-4}$ (34–47% of $2c_{\text{ad}}$ across calibers). (c) detection rate 0.5 *by construction* ($d_{\text{TV}} = 0$, Thm. 1); the shape channel ρ lies outside the observation space.

6.1 Negative Results

We retain and flag: (i) Hyperplane is the negative result of RQ6—no significant saving is detected; (ii) the L2-slope prediction is rejected (OLS energy_rate $\sim \rho$: $b_1 = -2.2003$, SE 0.4297, CI $[-3.043, -1.358]$, excluding c_{ad}), with a ρ/period collinearity caveat that makes the coefficient a joint design slope rather than a causal effect of ρ, and the cost-model exponent α alone explains only 8.4% of the variance; (iii) two earlier headlines were withdrawn on trace-back: the 87.7% saving fell to 66.7% (cold-start thermal spike; absolute-joule CV ≈32% is throttle-dominated, so we report repeated-saving ratios and mJ/step instead) and the E6-v2 analysis (protocol defect); (iv) the δ^* kink (+8.3%) is not resolvable at n=3; (v) of six evasion arms two are undetectable and fusion dilutes AUROC 1.000 → 0.753, while on the replay harness gating costs 7.2pp accuracy.

7 Limitations

Single board (Pi 3B): the cross-device hardware comparison is not closed (MatePad telemetry blockage) and is future work, and loopback TX bounds the upload-phase floor. The M5 gate is calibrated

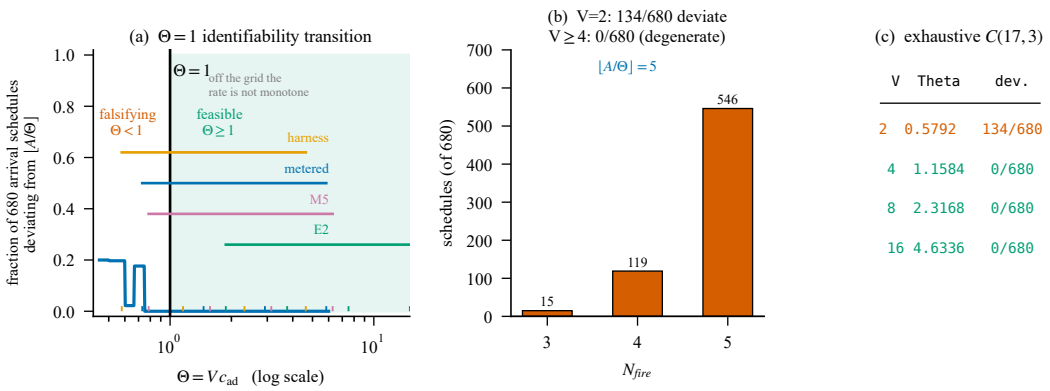


Figure 4: **The claim is falsifiable, and where.** Exhaustive $C(17, 3) = 680$ arrival orderings, $A(T) = 3$, $\Theta = V c_{\text{ad}}^{\text{harness}}$. (a) deviation from $\lfloor A(T)/\Theta \rfloor$ vs. Θ (log axis): the $\Theta = 1$ line separates feasible (shaded) from falsifying; all four calibers fall inside $V \in [1.07, 3.45]$. (b) V=2 ($\Theta = 0.5792$): $134/680$ (19.7%) deviate at this (A, T) window (Appendix G). (c) $V \geq 4$: $0/680$ at every ordering.

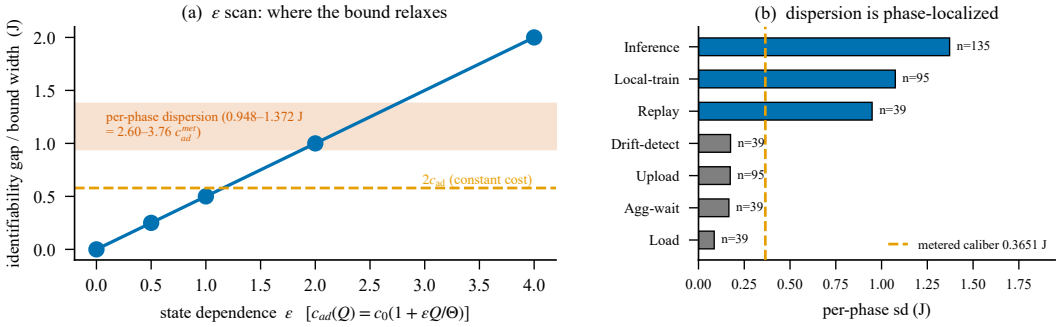


Figure 5: **Where the bound relaxes (Thm. 2(iii)).** (a) ε scan (synthetic, level S) under $c_{\text{ad}}(Q) = c_0(1 + \varepsilon Q/\Theta)$: the gap grows with state dependence and crosses the constant-cost $2c_{\text{ad}}$ bound (red band = (b)). (b) that dependence is phase-localized: per-phase sd (n on bars, dashed = $c_{\text{ad}}^{\text{met}}$) is 0.948–1.372 J in train/replay/inference, below $0.5\, c_{\text{ad}}^{\text{met}}$ in the four fixed-cost phases (938-row ledger).

in-sample. Assm. 5 fails on violation streams, where the claim degrades to a metering statement, and A1 is run-level: under time-varying $\Theta(t)$ the counting identity does not hold, so Lemma 1/Thm. 2 constrain fixed-Θ deployment only. The δ^* kink is below the noise floor at n=3, Φ cannot detect forged-marker attacks (information-theoretically necessary), evasion arm 5 (multi-thread proxy) awaits a faithful re-test, and the audit ROC's κ intervals are not cited pending source verification.

8 Conclusion

We asked whether the energy benefit of adaptive inference can be attributed to the policy when observations are compressed to event-driven binary signals. On metered hardware the answer is channel-specific: in the energy channel the policy enters the ledger only through N_{fire} (it changes how often the meter ticks, not what each tick costs), so gated adaptation is indistinguishable, up to $2c_{\text{ad}}$, from a count-matched open-loop schedule and, given the event stream, the ledger is auditable without knowing the policy (against under-reporting and replay; over-reporting is not detectable from energy alone, Table 2). The claim has an explicit failure threshold ($\Theta < 1$) and escape (state-dependent trigger cost). The M5 saving (46.9%) comes with a negative result (Hyperplane -14.77%). Green-FL standardization (Tapp et al., 2026) made estimates comparable; metering makes them checkable, and extends to on-device LLM fine-tuning, where LoRA detectors report accuracy but no energy (Rondanini et al., 2026). The joule is still missing from the edge-CL ledger.

REFERENCES

Martin Abadi et al. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016. doi: 10.1145/2976749.2978318.

Giuseppe Anastasi, Marco Conti, Mario Di Francesco, and Andrea Passarella. Energy conservation in wireless sensor networks: A survey. *Ad Hoc Networks*, 7(3):537–568, 2009.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Yahya H. Ezzeldin, and Salman Avestimehr. Federated orthogonal training: Mitigating global catastrophic forgetting in continual federated learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Albert Bifet and Ricard Gavaldà. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM International Conference on Data Mining (SDM)*, pp. 443–448. SIAM, 2007. doi: 10.1137/1.9781611972771.42.

Abhijit Chakraborty, Suddhasvatta Das, Yash Shah, Vivek Gupta, and Kevin A. Gary. TIMEGATE: Sustainable time-boxed promotion gates for continual ML adaptation under resource constraints. *arXiv preprint arXiv:2605.29183*, 2026. ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI"; not a main-conference publication.

Abhinav Chakraborty and Subha Maity. The statistical cost of adaptation in multi-source transfer learning. *arXiv preprint arXiv:2605.09471*, 2026.

Junbao Chen, Jingfeng Xue, Yong Wang, Zhenyan Liu, and Lu Huang. Classifier clustering and feature alignment for federated learning under distributed concept drift. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. doi: 10.52202/079017-2586.

Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10164–10173, 2022. doi: 10.1109/cvpr52688.2022.00992.

Aurélien Garivier and Éric Moulines. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory (ALT)*, volume 6925 of *Lecture Notes in Computer Science*, pp. 174–188, 2011. doi: 10.1007/978-3-642-24412-4_16.

Ido Nitzan Hidekel and Dan Raviv. Catastrophic forgetting is low-rank: A function-space theory for continual adaptation. In *ICML 2026 Workshop "Continual Adaptation at Scale"*, 2026. arXiv:2606.18024.

Soheil Hor, Ying Qian, Mert Pilanci, and Amin Arbabian. Adaptive inference: Theoretical limits and unexplored opportunities. *arXiv preprint arXiv:2402.04359*, 2024. Preprint; no peer-reviewed venue to date.

Zhiyan Hou, Dan Zhang, Tao Feng, Liyuan Wang, Wei Li, Xiangzhao Hao, Hongyan An, Junfeng Fang, Haokai Ma, Zhaohui Xu, Xinyu Tang, Haiyun Guo, Jinqiao Wang, and Tat-Seng Chua. Continual learning in transition. *arXiv preprint arXiv:2608.06216*, 2026.

Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 1039–1056, 2021. doi: 10.1109/SP40001.2021.00106.

James Kirkpatrick, Razvan Pascanu, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Daniel Kreisler, Yair Carmon, and Oliver Hinder. Clipping the price of adaptivity at the tail. In *Conference on Learning Theory (COLT)*, volume 336 of *Proceedings of Machine Learning Research*, pp. 4267–4307, 2026.

Yichen Li, Yuying Wang, Jiahua Dong, Haozhao Wang, Yining Qi, Rui Zhang, and Ruixuan Li. Resource-constrained federated continual learning: What does matter? In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025a.

Yichen Li, Yuying Wang, Haozhao Wang, Yining Qi, Tianzhe Xiao, and Ruixuan Li. FedSSI: Rehearsal-free continual federated learning with synergistic synaptic intelligence. In *International Conference on Machine Learning (ICML)*, 2025b.

Shayne Longpre et al. Bridging the data provenance gap across text, speech, and video. In *International Conference on Learning Representations (ICLR)*, 2025.

Osamah A. Mahdi, Eric Pardede, Savitri Bevinakoppa, and Nawfal Ali. Federated learning under concept drift: A systematic survey of foundations, innovations, and future research directions. *Electronics*, 14(22):4480, 2025. doi: 10.3390/electronics14224480. URL `https://www.mdpi.com/2079-9292/14/22/4480`.

Ioannis Mavromatis, Stefano De Feo, and Aftab Khan. FLAME: Adaptive and reactive concept drift mitigation for federated learning deployments. In *Proceedings of the 2024 International Conference on Embedded Wireless Systems and Networks (EWSN) — Workshop on Enabling Machine Learning Operations for next-Gen Embedded Wireless Networked Devices (EMERGE)*, pp. 336–341, 2024. doi: 10.3217/7c2g-rq14.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Aaron Pache and Mark C. W. van Rossum. Mistake gating leads to energy and memory efficient continual learning. *arXiv preprint arXiv:2604.14336*, 2026.

Junsoo Park, Ikjune Yoon, and Dong Kun Noh. Stability-controlled continual federated learning for energy-harvesting AIoT systems. *Sensors*, 26(11):3325, 2026. doi: 10.3390/s26113325.

Adam Piaseczny, Md Kamran Chowdhury Shisher, Shiqiang Wang, and Christopher G. Brinton. RCCDA: Adaptive model updates in the presence of concept drift under a constrained resource budget. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Christian Rondanini, Barbara Carminati, Elena Ferrari, Niccolò Lardo, and Ashish Kundu. LoRA-based parameter-efficient LLMs for continuous learning in edge-based malware detection. *arXiv preprint arXiv:2602.11655*, 2026.

Giuseppe Serra and Florian Buettner. Federated continual learning goes online: Uncertainty-aware memory management for vision tasks and beyond. In *International Conference on Learning Representations (ICLR)*, 2025.

Yimin Shi, Haihan Duan, Lei Yang, and Wei Cai. An energy-efficient and privacy-aware decomposition framework for edge-assisted federated learning. *ACM Transactions on Sensor Networks*, 18(4):53:1–53:24, 2022. doi: 10.1145/3522741.

Neta Shoham, Tomer Avidor, Aviv Keren, Nadav Israel, Daniel Benditkis, Liron Mor-Yosef, and Itai Zeitak. Overcoming forgetting in federated learning on non-IID data. *arXiv preprint arXiv:1910.07796*, 2019. NeurIPS 2019 Workshop on Federated Learning for Data Privacy and Confidentiality.

Friedrich Solowjow and Sebastian Trimpe. Event-triggered learning. *Automatica*, 117:109009, 2020. doi: 10.1016/j.automatica.2020.109009.

Austin Tapp, Holger R. Roth, Ziyue Xu, Abhijeet Parida, Hareem Nisar, and Marius George Linguraru. Standardized methods and recommendations for green federated learning. In *ACM e-Energy*, 2026.

Arya Tschand, Arun Tejusve Raghunath Rajan, Sachin Idgunji, Anirban Ghosh, Jeremy Holleman, Csaba Kiraly, Pawan Ambalkar, Ritika Borkar, Ramesh Chukka, Trevor Cockrell, Oliver Curtis, Grigori Fursin, Miro Hodak, Hiwot Kassa, Anton Lokhmotov, Dejan Miskovic, Yuechao Pan, Manu Prasad Manmathan, Liz Raymond, Tom St. John, Arjun Suresh, Rowan Taubitz, Sean Zhan, Scott Wasson, David Kanter, and Vijay Janapa Reddi. MLPerf Power: Benchmarking the energy efficiency of machine learning systems from microwatts to megawatts for sustainable AI. *arXiv preprint arXiv:2410.12032*, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794.

Zhaohui Wang. Uncertainty-aware sequential decision rules for event-triggered LLM invocation in streaming systems. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, Lecture Notes in Computer Science. Springer, 2026. doi: 10.5281/zenodo.20783298. Research Track. Six results proved: inter-event chattering bound; threshold-policy optimality via smooth pasting; approximate SPRT; $O(\sqrt{T \log T})$ regret extending to $O(\sqrt{(C_T+1)T \log T})$ under C_T changepoints; OGD convergence; calibration-to-miss-rate transfer. Code artifact: Zenodo 10.5281/zenodo.20783298.

Yebo Wu, Li Li, and Cheng-zhong Xu. Breaking the memory wall for heterogeneous federated learning via progressive training. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025. doi: 10.1145/3690624.3709284.

Kang Yan, Nina Shu, Tao Wu, Chunsheng Liu, and Panlong Yang. A survey of energy-efficient strategies for federated learning in mobile edge computing. *Frontiers of Information Technology & Electronic Engineering (FITEE)*, 25(5):645–663, 2024. doi: 10.1631/FITEE.2300181.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12073–12086, 2021. URL `https://proceedings.mlr.press/v139/yoon21b.html`.

Qingyang Yu, Yang Hua, Qizhen Zhang, and Hao Wang. GFedCL: Graph-based federated continual learning with spatial and temporal awareness. In *International Conference on Machine Learning (ICML)*, 2026. Poster.

Lixing Zhang, Xuran Liu, Ruizhi Zhang, and Liyan Xie. Sequential change detection with local differential privacy. *Entropy*, 28(4):402, 2026. doi: 10.3390/e28040402.

Rui Zhang, Jian Liu, Yuan Ding, Zhibo Wu, Qingbiao Wang, and Kui Ren. "adversarial examples" for proof-of-learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2022. doi: 10.48550/arXiv.2108.09454.

A PROOFS

Lemma 1. The gate adds each arrival before testing and admits at most one absorption per step, so the debt queue follows the cumulative–discharge recursion $Q_{t+1} = Q_t + a_t - \Theta\,\mathbf{1}[Q_t + a_t \geq \Theta]$ with $Q_0 = 0$, equivalently the classical form $Q_{t+1} = (Q_t + a_t - \Theta)^+$ when at most one absorption is allowed per step. Work conservation is exact and holds bitwise on the metered trace: $\Theta\, N_{\mathrm{fire}}(T) + Q(T) = A(T)$. Two consequences. (a) *Granularity.* If $a_t \leq a_{\max} \leq \Theta$ then $Q_t < \Theta$ for all t (induction: $Q_t + a_t < 2\Theta$, so at most one absorption fires, leaving $Q_{t+1} = Q_t + a_t - \Theta < \Theta$), hence $Q_T < \Theta$ and $N_{\mathrm{fire}}(T) = \lfloor A(T)/\Theta \rfloor$ exactly; the paper's "$\Theta \geq 1$" is the binary-evidence ($a_{\max}$=1) instance of the bound $a_{\max} \leq \Theta$ (Prop. 2). (b) *Deficit.* In general $\lfloor A(T)/\Theta \rfloor - N_{\mathrm{fire}}(T) = (Q(T) - (A(T) \bmod \Theta))/\Theta \in \mathbb{Z}_{\geq 0}$: any shortfall is one-sided downward, integral, and vanishes exactly when the terminal backlog attains its minimum $A(T) \bmod \Theta$—a truncated busy period, not an identity failure. Energy decomposes as $E = T c_{\mathrm{mon}} + c_{\mathrm{ad}}\, N_{\mathrm{fire}} + r(T)$; work conservation gives $r(T) = c_{\mathrm{ad}}\,(N_{\mathrm{fire}} - A(T)/\Theta) = -c_{\mathrm{ad}}\, Q_T/\Theta \in (-c_{\mathrm{ad}}, 0]$, so the residual is *non-positive* rather than non-negative, with magnitude $Q_T/\Theta < 1$ of one service quantum, i.e. $|E - (T c_{\mathrm{mon}} + c_{\mathrm{ad}}\, A(T)/\Theta)| \leq c_{\mathrm{ad}}$; the bound on $r(T)$ is tight. □

Theorem 1 (Le Cam two-point, (Tsybakov, 2009)). Take any two policies π_0, π_1 and any arrival stream with the same flag count $A(T)$. On the count channel the observed statistic is N_{fire} alone; by Lemma 1 N_{fire} is a deterministic function of $A(T)$ and Θ. Hence the induced distributions on the observation space coincide, $d_{\text{TV}} = 0$, and by the Le Cam two-point lemma no test achieves power above $\alpha + (1 - \alpha) = 1/2$ at equal mass. The construction is finite (two policies, matched counts); the shape channel ρ is not in the observation space. □

Theorem 2(i). Construct P^* from $A(T)$ (oracle-matched; direction favors the gate): fire at the same count N_{fire}, spreading triggers evenly. Total adaptation energy is $c_{\text{ad}}\, N_{\text{fire}}$ for both, so $|E_{\text{gate}} - E_{P^*}| \le c_{\text{ad}}\, |Q_T/\Theta| + c_{\text{ad}} \le 2c_{\text{ad}}$. (ii) is the measured statement (paired, n=20). (iii) With $c_{\text{ad}}(Q)$, $\Delta E = \sum_t \Delta c_{\text{ad}}(Q(t))\mathbf{1}[\text{fire}]$; under constant cost every term is zero, $O(c_{\text{ad}})$. □

Theorem 3. Standard sparse-adaptive-regret analysis in the switching-bandit lineage (Garivier & Moulines, 2011) with metered constants substituted; the uncertainty correction follows from $|c_{\text{ad}} - \hat{c_{\text{ad}}}| \le \delta$ propagating linearly through the switching term $\le D\, A(T)/\Theta$. □

Theorem 4. (i) Fixed-threshold gap: with continuation value $V(t)$ piecewise-constant across the onset, the reservation-value benchmark threshold is $\Theta^{\text{RV}}(t)$; the gap is the closed-form one-period option value of waiting under the two regimes. (ii) By Wang (2026) (their proved regret result): their assumptions map structurally: threshold policy $\leftrightarrow$ $\Theta(t)$ gate; bounded parameter set $\leftrightarrow$ $\Theta \in$ bounded grid; stationarity within regime $\leftrightarrow$ between flag events; bounded risk $\leftrightarrow$ $Q(t) \in [0, \Theta + 1)$; binary increments $\leftrightarrow$ $a(t) \in \{0, 1\}$; restart-at-detection $\leftrightarrow$ flag events. Regret $O(\sqrt{T \log T})$, with change-points $O(\sqrt{(C_T + 1)T \log T})$. We rely only on their proved results. □

Theorem 5. Symbols. μ is the energy weight of the cost functional $J(T) = \text{Regret}(T) + \mu E(T)$ (Assm. 4); ΔR is the per-round regret improvement obtained by adapting; ϱ is the state-dependent part of the per-trigger cost, as a fraction of c_{ad}; α is the FLOP-normalised energy coefficient of E1 (J/FLOP) and B the extra FLOPs a trigger adds, so αB is an extra per-trigger energy in joules; $e_{\max}$ is the worst-case overshoot $q_{\text{after}}/\Theta - 1$ beyond the threshold at a fire. Both terms inside the bracket are taken in units of c_{ad}. The boundary follows from requiring the expected per-round gain to clear the trigger cost, $\mu \Delta R \ge c_{\text{ad}}\varrho + \alpha B$, with the overshoot inflating the effective threshold by $e_{\max}$; solving gives $\delta^* = e_{\max}(\varrho + \alpha B)$. The empirical kink test at n=3 is underpowered (below the noise floor). □

Proposition 1. Two traces (honest and replayed) with identical totals map to the same scalar; separating them requires a statistic that varies across the trace, i.e. phase-level structure. □

B METROLOGY DETAILS

Uncertainty budget for the INA219 chain follows the GUM, but with a stated composition: every interval reported in this paper is a *Type A* component (replication-based, computed from the per-phase repeats at the stated n). The *Type B* components—ADC quantisation of the INA219, its gain and offset error, and the shunt tolerance—are named here but deliberately *not* folded into the reported intervals, because converting datasheet limits would require a shunt and gain assumption that this deployment does not pin down; the quoted ranges are therefore Type A only and do not include the Type B contribution. Sampling and idle-baseline calibration follow; the INA219 is sampled at a nominal 50 Hz (55.6 Hz achieved in the E0 `sample_rate_scan`), well above the loop's per-phase energy timescale (0.05–0.47 J over multi-second phases), so transient capture is not limiting; the throttling-artifact discovery and correction history (absolute-joule CV $\approx$32% under throttle, motivating repeated-saving ratios and mJ/step as honest indicators). Sensor-noise decomposition: σ_{dyn} is replication-count dependent (0.009320 J at n=100 vs. 0.026505 at n=3, 2.84×); the block-internal $\sigma_{\text{static}} = 0.040651$ J must not be read as cross-block drift (cross-block std 0.006508). Temperature correction is not applied: the sign of the inferred correction is not stable across the three measurement sessions we formed, so no single correction is defensible. The energy rate α of E1 is a regression of measured step energy against an analytic FLOP count of the deployed graph (each multiply–accumulate counted as two FLOPs; harness and data-movement overhead excluded), so α is model-relative rather than a hardware constant. Per-phase dispersion (938-row ledger, sd by phase) is phase-localized: 0.948–1.372 J = 2.60–3.76 $c_{\text{ad}}^{\text{met}}$ in train_local/replay/inference ($n = 95/39/135$), i.e. in the phases whose cost depends on state, while the four fixed-cost phases (load/upload/drift-

detect/agg-wait, $n = 39/95/39/39$) stay below 0.175 J $= 0.48\, c_{\mathrm{ad}}^{\mathrm{met}}$. A1 is a run-level statement, so this is state dependence, not a violation.

Re-calibration recipe. Because c_{ad} is a session- and board-specific INA219 constant, porting the ledger to a different board requires reusing the procedure rather than the numbers: lock the CPU clock, sample the rail at $\geq$50 Hz, subtract a per-session idle baseline, then estimate c_{ad} over $n \geq 100$ identical single-trigger events (n=3 inflates σ_{dyn} by $2.84\times$), take that estimate as $c_{\mathrm{ad}}^{\mathrm{harness}}$, and rebuild the Θ grid as $\{2, 4, 8, 16\} \times c_{\mathrm{ad}}^{\mathrm{harness}}$ (here 0.2896 J, giving 0.5792–4.6336 J, cf. Assm. 3). The numerical values in Table 7 must not be carried over to another board.

Table 5: Seven-phase workflow structure. Shares from the v3 fixed-window ledger (52 phase rows, 704.61 J total; drift-window observation rows excluded from shares). *Workflow phases, not energy components.* Calibers are never mixed: the real-operation ledger (2.89 J/loop, adaptive share 79.1%; train_local+replay) is reported separately from the v3 fixed-window ledger (66.30 J/rep, adaptive share 19.46%).

Phase	Share (%)	Gated?
Local-train (incl. DP variant)	44.06	yes (trigger-gated)
Inference	14.47	no
Drift-window observation (steady/adapt)	14.67	observation only
Upload	8.48	partly (replay bytes)
Replay	5.58	yes (trigger-gated)
Load	4.83	no
Drift-detect	4.76	no (always-on monitor)
Agg-wait	3.15	no
ADAPT_PHASES two-way split (real operation)	adaptive 79.1 / base 20.9	—

C Fail-closed ledger and Φ

Ledger schema (per phase: duration, energy, debt-gate events, accuracy, comm bytes); Φ = channel adequacy $\wedge$ minimality $\wedge$ consistency. Fault injection follows the fail-closed discipline: eleven defect classes were injected into the export pipeline and every one was caught (`placeholder_zero`, `fired_trigger_mismatch`, `ts_constant`, `queue_recursion_break`, `missing_arrival_mode`, `theta_caliber_dropped`, `invariant_violation`, `pre_step_over_bound`, `segment_sum_break`, `phase_name_collision`, `stream_state_missing`). The two reported rates differ in their unit of account, not in stringency: $938/938$ counts *ledger rows* and $381/381$ counts *per-round event rows*, with zero leakage at either level.

Table 6: Evasion matrix (E4): six arms, n=30 trajectories per arm, K=8, 50 Hz sampling, hardware (`simulated:false`). "Total" is AUROC over the whole ledger; "Probe" is AUROC once the known-cost replay probe is added. Arms 3 and 5 are software proxies (no hardware DVFS or coprocessor was available).

Arm	Mechanism	Total	Probe	Verdict
1 `dummy`	unaccounted idle energy	0.623	0.887	needs joint κ
2 `dilute`	$\int P\,dt$ conserving	1.000	0.753	detectable (energy)
3 `dvfs_proxy`	J/op shift, Φ departure	0.942	1.000	detectable (energy)
4 `reorder`	ρ change (G2)	0.974	0.981	detectable (energy)
5 `thread`	ρ+κ change	0.549	0.546	**undetectable**
6 `replay`	$\int P\,dt$=L replay	0.657	0.602	**undetectable**

D Caliber cross-table and retired constants

Full caliber $\times$ indicator matrix. Caliber-to-grid map used in Fig. 4: $V = 3.4530/2.7390/2.5484/1.0651$ for harness/metered/M5/E2. Retired constants with reasons:

87.7% (throttle artifact), 9.78× (superseded by 7.81 × /7.57× protocol pair), kink +8.3% (n=3), E6-v2 (protocol defect), C1 155.08 J (dirty row).

Table 7: Per-trigger energy constants. The theorem-facing default is the metered caliber $c_{\mathrm{ad}}^{\mathrm{met}} = 0.3651$ J; the others are retained with suffixes as robustness. Runner naming: `C_ADAPT` is the module constant (= harness 0.2896); `c_adapt` is the config parameter name.

Caliber	Value	Object	Source	Role
harness	0.2896 J/fire	closed-loop fire	$2 \times c_{\mathrm{replay}}$=0.1448	Θ grid (Assm. 3)
metered ($c_{\mathrm{ad}}^{\mathrm{met}}$)	0.3651 J/event	adaptation trigger	INA219; CI $[0.3547, 0.3758]$	theorem default, Thm. 3
M5	0.3924 J/event	M5 pipeline	M5 calibration	sensitivity
M5-Pi per-event	0.4712 J/event	single events (n=173, CV 18.0%)	INA219 per-round	A1 dispersion (only per-event caliber)
E2 per-step	0.9389 J/step	real adaptation step	E2 calibration	step-level audits

E BUDGET-FEASIBLE PRIVACY

Given a per-step energy budget $E_{\max}$, the feasible (ε, σ) region is $\{(\varepsilon, \sigma) : \bar{J}(\sigma) \leq E_{\max}\}$ with $\bar{J}$ the DP-SGD energy curve measured at R=20; points: naïve 1.95 ± 0.18 J, Opacus 0.64 ± 0.07 J against $C_{\mathrm{fix}} = 0.2673$ J, sensitivity span $0.0972/0.2673/0.3651$ (3.8×). The privacy configuration is fixed and identical across both replays: noise multiplier $\sigma = 8.0$, clipping norm $C = 1.0$, batch size 64, $R = 20$ repetitions. We report σ as the recorded primitive and state no ε, because the accountant configuration is not part of the metered record and an unconverted noise multiplier is the honest quantity; the axis above is therefore parameterised by σ, with ε left to the reader's accountant. Both replayed sources were measured on a second Pi 3B+INA219 deployment (2026-07-24) and transferred with SHA-256 verification against the packaging manifest.

F CALIBRATION VALIDITY TIMELINE

$\tau_\alpha \geq 2.86$ h (test-segment paired, gap $\Delta t = 1801$ s, $\Delta_{\mathrm{adv}} = 0.034676$ J); calib-segment sensitivity 2.46 h; cross-session pairs $2.86/34.62/34.23$ h ($\tau \propto \Delta t$: the 12× spread is arithmetic, not three independent measurements); Welch $p = 0.867/0.761/0.686$; metered-debt margin 20× (10,279 s / 500 s).

G DEVIATION-RATE WINDOW SENSITIVITY

The deviation statistic reported in Fig. 4b is (A, T)-dependent, and the qualitative conclusion is invariant across windows: $\Theta \geq 1$ is exact at every window and every arrival ordering, while $\Theta < 1$ admits a one-sided downward budget shortfall that dilutes as the window grows (e.g. A=5: 31.2% → 10.1% → 1.67% at T=17/50/300). The condition behind "$\Theta \geq 1$" is the granularity bound of Prop. 2: exactness is guaranteed whenever $a_{\max} \leq \Theta$, which binary evidence ($a_{\max}$=1) reduces to $\Theta \geq 1$; the mechanism is developed in §H.

H QUEUE-THEORETIC STRUCTURE OF THE COUNTING IDENTITY

Per-round hardware form of the identity. The deployment ledger records per-phase gate decisions; a dedicated rerun of the same node (Raspberry Pi 3B+INA219) writes a per-round trace (a_t, q_{before}, q_{after}, `fired`, round energy) for 3 runs × 3 seeds × $V \in \{2, 4, 8, 16\}$ at T=300 with A=5 arrivals per seed. On all 36/36 node-runs the recorded fire count equals $\lfloor A/\Theta \rfloor$ exactly (V=2/4/8/16 → 8/4/2/1 fires per seed), the recurrence reproduces the trace bitwise, and the terminal backlog is $A \bmod \Theta = 0.3664$ J $< \Theta$. Fig. 2b is therefore a hardware-checked statement rather than a prediction, and the energy residual is one-sided, matching $N_{\text{fire}} \leq \lfloor A/\Theta \rfloor$.

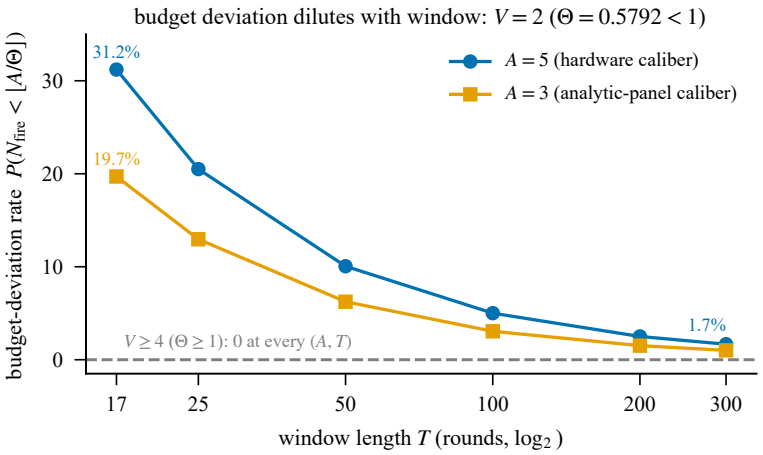


Figure 6: **Budget deviation dilutes with the observation window.** At $V{=}2$ ($\Theta{=}0.5792 < 1$, the falsifying cell) the budget-deviation rate $P(N_{\text{fire}} < \lfloor A/\Theta \rfloor)$ is a function of the window (A, T), not a universal constant: the analytic-panel window $(A{=}3, T{=}17)$ gives $134/680 = 19.7\%$ (Fig. 4b) whereas the hardware window $(A{=}5, T{=}300)$ gives 1.67% — consistent with the three hardware runs' zero observed deviations ($0.983^3 \approx 95.1\%$). Deviation is one-sided downward ($N_{\text{fire}} \leq \lfloor A/\Theta \rfloor$): it reflects worst-case budget conservatism, *not* identity violation — the counting identity holds exactly per event (hardware bitwise match). $V{\geq}4$ ($\Theta{\geq}1$): 0 at every (A, T). All plotted values are exact (dynamic program over the queue-state lattice, $\Theta = 362/625$); exhaustive enumeration ($C(T, A) \leq 2{\times}10^5$) and seeded Monte Carlo ($n{=}2{\times}10^4$) agree at every grid point.

Proposition 2 (granularity condition)**.** *Let $a_t \in [0, a_{\max}]$ and $Q_{t+1} = Q_t + a_t - \Theta\,\mathbf{1}[Q_t + a_t \geq \Theta]$ with $Q_0 = 0$. If $a_{\max} \leq \Theta$, then $Q_t < \Theta$ for every t; hence $Q_T < \Theta$ and $N_{\text{fire}}(T) = \lfloor A(T)/\Theta \rfloor$ exactly. Under binary evidence (Assm. 2, $a_{\max}{=}1$) the condition reads $\Theta \geq 1$. Under graded evidence the same inequality $a_{\max} \leq \Theta$ remains necessary and sufficient, so $\Theta \geq 1$ alone is no longer enough once $a_{\max} > 1$: only the ratio $a_{\max}/\Theta$ enters.*

Proof. By induction $Q_t < \Theta$. If $Q_t + a_t < \Theta$, no absorption fires and $Q_{t+1} = Q_t + a_t < \Theta$; if $Q_t + a_t \geq \Theta$, exactly one fires (since $Q_t + a_t < \Theta + a_{\max} \leq 2\Theta$) and $Q_{t+1} = Q_t + a_t - \Theta < \Theta$. With $Q_T < \Theta$, the conservation identity $\Theta N_{\text{fire}} + Q_T = A(T)$ gives $N_{\text{fire}} = \lfloor A(T)/\Theta \rfloor$; the escape cells of Table 8 are exactly those with $a_{\max}/\Theta > 1$. □

Table 8: The grid's granularity ratio $g = a_{\max}/\Theta$ from the per-round rerun trace (a_t is exactly binary, $a_{\max}{=}1.0000$). The single cell with $g > 1$ is the single cell that escapes.

V	2	4	8	16
Θ (J)	0.5792	1.1584	2.3168	4.6336
$g = a_{\max}/\Theta$	**1.727**	0.863	0.432	0.216
$\max q_{\text{before}}/\Theta$	**1.633**	0.863	0.863	0.863

Assumption 5 as a tested proposition. The assumption asserts that $\mathbb{E}[c_{\text{ad}} \mid Q{=}q]$ is free of q. We test it on the per-round rerun by stratifying fire events by q_{before}/Θ within each run; since the three runs carry different calibrations, the test is run three times independently and the calibers are never pooled.

Regressor convention and its audit trail. In the logged traces q_{before} is Q_t, the backlog *before* the arrival is added, whereas the cost model multiplies the *overshoot* $\tilde{q}_t = Q_t + a_t$, so two regressor conventions are available and Table 9 reports the second. Running the *identical* stratified test under the overshoot regressor $\tilde{q}_t/\Theta$ moves `195417` from $\rho = +0.059$, KW $p = 0.869$, gap 0.5% to $\rho = +0.215$, KW $p = 0.277$, gap 8.4%; `001535` from $+0.041/0.769/6.9\%$ to $+0.127/0.379/6.3\%$; and `175805` from $-0.158/0.178/9.3\%$ to $-0.132/0.582/6.1\%$. Both rows are reproduced by

Table 9: Stratified constancy test (n=45 fires per run, $V \in \{2, 4, 8, 16\}$). ρ: Spearman correlation of c_{ad} against the stratifier; gap: largest stratum-to-stratum mean difference as a fraction of the run mean; SD: relative dispersion of c_{ad} within the run. **Regressor:** the tabulated ρ uses the *overshoot* $\tilde{q}_t/\Theta = (Q_t + a_t)/\Theta$, the quantity that the cost model of Thm. 2(iii) actually multiplies, not the logged q_{before}, which is the pre-arrival backlog Q_t alone. Because the strata are tertiles of the regressor, the two conventions re-sort the pooled n=45 sample and yield different numbers; the superseded q_{before} row ($\rho = -0.158, +0.059, +0.041$; KW $p = 0.178, 0.869, 0.769$; gap $9.3\%, 0.5\%, 6.9\%$) is retained in the erratum below for audit only. **Pooling:** c_{ad} is z-scored within each V before pooling, so the statistics are scale-free across calibrations; the three runs are never pooled with one another.

Run	n	ρ (p)	KW p	gap	SD
`175805`	45	-0.132 (0.387)	0.582	6.1%	14.9%
`195417`	45	$+0.215$ (0.157)	0.277	8.4%	14.9%
`001535`	45	$+0.127$ (0.404)	0.379	6.3%	14.2%

the same script (`_a5_table8_align.py`), with the backward-looking row matching the earlier tabulation to the printed precision in every cell, so the difference is attributable to the regressor alone and not to a change of stratification.

A third variant, and why it is rejected. A re-implementation that omitted the within-V z-scoring, used quartiles instead of tertiles, and grouped the Kruskal–Wallis statistic by the raw regressor value rather than by stratum yields $\rho = +0.365$, gap 17.6% for `195417`; that variant is retained here for audit and is rejected, because it does not reproduce the stratification whose numbers it claims to revise. Against the reproduction above, the figures are $\rho = +0.215$ and gap 8.4%.

Three caveats bound what this test supports. (i) Only V=2 has variation in a_t ($a_t \in \{0, 1\}$ on fire rounds); at $V \in \{4, 8, 16\}$ every fire round has a_t=1, so $\tilde{q}_t = Q_t + 1$ is a *constant* offset within those cells and cannot re-order them—the entire change in the pooled statistic comes from re-sorting *across* V, not from any within-V evidence. (ii) At V=16 the regressor has *no variation at all* ($q_{\text{before}}/\Theta \equiv 0.863$, $\tilde{q}/\Theta \equiv 1.08$ for all 3 fires), so 3 of the 45 events carry no information about Q-dependence. (iii) The sign of ρ is still not consistent across runs ($-, +, +$ under either convention) and no run rejects ($p \geq 0.157$).

We accordingly restate the finding. Under the corrected regressor the authoritative run moves closer to the noise floor (gap $0.5\% \rightarrow 8.4\%$ against a dispersion of 14.9%), but still does *not* exceed it, so the qualitative summary survives the correction: no run rejects, no consistent monotone Q-dependence appears, and the largest stratum gap (6.1–8.4%) remains smaller than the per-event dispersion (14.2–14.9%). What the correction changes is the *margin*, not the verdict: the test is nearer to its resolution limit than the superseded table suggested, and with n=45 events (of which 3 carry no within-V variation at all, 18 (V=4, 8) vary within V but keep the same ordering under either regressor, and only the 24 in the V=2 cell can actually be re-ordered), it is a failure to reject with little power rather than evidence of constancy. It cannot distinguish a small genuine state-dependence from session/thermal drift correlated with the round on which a policy fires, because the fire operation is a frozen, fixed-workload batch (Sec. D) in which no cost-channel dependence can arise by construction. Resolving exactly that distinction (a real state-dependent cost versus the same nuisance) requires an experiment that *manipulates* the state dependence at a known amplitude, which is the ε-manipulation ablation. The V=4 subsample (n=12) still reverses sign across runs and is not quotable alone. The violation regime of Thm. 2(iii) (a state-dependent workload) remains the escape.

Internal structure of the escape region. The shortfall is not uniform in $A(T)$. Let $r = (A(T) - 1) \bmod \Theta$ and $\kappa = \lfloor (r + a_{\max})/\Theta \rfloor - 1$; to first order in the dilute ratio $A(T)/T$,

$$P(N_{\text{fire}} \neq \lfloor A(T)/\Theta \rfloor) \;\approx\; \kappa\,(A(T)/T).$$

After the last arrival the queue needs $\kappa + 1$ steps to drain, so a horizon shorter than that leaves a deficit, and whether $\kappa = 0$ is decided by the arithmetic of $A(T) \bmod \Theta$ rather than by anything the policy does.

All 17/17 predicted cells match, the tightest being $A(T)$=11, whose residual sits $0.008\,\Theta$ from the boundary. A model-only check outside the hardware grid (V=1, Θ=0.2896, predictions $\kappa = 2, 3$)

Table 10: Comb law at $V{=}2$ ($\Theta{=}0.5792$), $T{=}400$, exact enumeration over all arrival orderings. *pred*: $\kappa \geq 1$ (ON) or $\kappa = 0$ (off); *obs*: measured rate divided by $A(T)/T$.

$A(T)$	2	3	4	5	6	7	8	9	10
pred	ON	ON	off	ON	ON	ON	off	ON	ON
obs	1.00	1.01	0.01	1.00	1.01	1.02	0.02	1.00	1.02
$A(T)$	11	12	13	14	15	16	19	20	
pred	off	ON	ON	ON	off	ON	off	ON	
obs	0.03	1.00	1.03	1.03	0.04	1.00	0.05	1.00	

matches $6/6$. The law is a dilute-regime statement: it requires $A(T)/T \lesssim 0.3$, and in the $\kappa{=}0$ branch it fails at $A(T)/T \geq 0.4$ (checked $47/50$; the three failures are $\kappa{=}0$ cells at $A(T)/T = 0.40$–0.53, where arrival clustering, a second-order term, dominates). Two consequences. The deviation rate of Fig. 6 is $O(A(T)/T)$ by construction rather than an empirical accident; and the law is falsifiable—at $V{=}2$, $T{=}300$, raising $A(T)$ from 5 to 8 should cut the deviation rate $\approx 26\times$ ($1.67\% \to 0.065\%$), whereas at $T{=}17$ the same change does not help ($31.2\% \to 41.0\%$), because the dilute precondition fails there.

Where the collapse would break. Let $c_{\mathrm{ad}}(Q) = c_{\mathrm{ad}}^0(1 + \varepsilon Q/\Theta)$. The ledger then observes only $\sum_{\text{fires}} c_{\mathrm{ad}}(Q_t) = N_{\mathrm{fire}} c_{\mathrm{ad}}^0 \langle Q/\Theta \rangle$, still a scalar: the collapse is re-expressed, not broken. Separating two count-matched policies requires the induced signal to clear the noise, $|\Delta E| \leq N_{\mathrm{fire}} c_{\mathrm{ad}}^0 \varepsilon g$ against $\sigma_E \approx \sigma_c \sqrt{N_{\mathrm{fire}}}$, giving the order-of-magnitude boundary

$$\varepsilon^* \;\approx\; 2\,\mathrm{CV} \big/ \big(g\sqrt{N_{\mathrm{fire}}}\big), \qquad g = a_{\max}/\Theta.$$

With the stratified CVs of Table 9 (pooled within two strata over the three seeds), $\varepsilon^* \approx 0.03$ at $V{=}2$ (0.028 before rounding) and ≈ 0.13 at $V{=}4$ (0.128); the scaling $\varepsilon^* \propto \mathrm{CV}\, V^{3/2}$ reproduces the $4.6\times$ ratio between them ($0.128/0.028$).

Convention for N_{fire} (disambiguation). In the display above, N_{fire} counts *independent* observations of the per-fire cost, i.e. the stratified sample pooled over the three seeds of the per-round rerun: $N_{\mathrm{fire}} = 24$ at $V{=}2$ and $N_{\mathrm{fire}} = 12$ at $V{=}4$ (8 and 4 fires per deployment, respectively). It does *not* count the fires of a single deployment. The two readings differ by $\sqrt{3}$, and the single-deployment reading gives the larger boundary $\varepsilon^* \approx 0.05$ at $V{=}2$ and 0.22 at $V{=}4$. We quote the pooled value because the question it answers (whether the signal is resolvable *given the sample we collected*) is the one posed above; the single-deployment value, which we report only for completeness, is the one an online controller would face when it must decide from one trajectory. Both use the same measured stratified CV (0.118 at $V{=}2$, 0.192 at $V{=}4$; overshoot convention) and the same $g = a_{\max}/\Theta$, so the $\sqrt{3}$ gap is purely a convention and not a discrepancy in the data. (The superseded q_{before} convention gives 0.125 at $V{=}2$ and leaves $V{=}4$ unchanged at 0.192: only $V{=}2$ has a_t variation, so only its CV moves. Either convention rounds to the same quoted ε^*.) At $V{=}8/16$ the stratified samples ($n{=}6/3$ fires per run) are too small to quote. Equivalently, the energy-channel capacity of the observation map, $C = \sup_{\pi_0,\pi_1 \text{ count-matched}} d_{\mathrm{TV}}(\mathbb{P}^E_{\pi_0}, \mathbb{P}^E_{\pi_1})$, is 0 at constant cost (Thm. 1 is the two-point witness) and becomes positive once $\varepsilon > \varepsilon^*$. The estimate is heuristic (it uses a two-sigma separability convention and ignores correlated drift) and is stated to make a boundary explicit; a hardware realisation at fixed $V{=}4$ with a state-dependent workload is the direct test and remains open.

Audit versus attribution. Prop. 1 and Thm. 1 act at different layers and do not compete: aggregate scalars cannot certify execution, which is why the phase-level ledger exists, whereas at matched trigger count even a complete phase-level ledger cannot identify the policy in the energy channel. Auditable execution does not imply attributable policy.

Scale boundary of the constant-cost scope. The identity is calibrated on a Pi-class node running a 21,154-parameter student, and nothing in the argument requires that particular size; what it does require is that the per-trigger cost stay free of the queue state. As model size and accelerator width grow, the adaptation step acquires a state-dependent component (Assm. 5 fails), which places the deployment in the violation regime of Thm. 2(iii) rather than in the collapse regime. The same caveat covers the instrument: the 50 Hz sampling and the 0.05–0.47 J per-phase scale are the regime

in which the metered ledger of App. B is the binding instrument, and a data-centre GPU would need a different accounting layer.

I FULL RESULTS, SEEDS, HYPERPARAMETERS

Seeds, hyperparameters and the device/workload specification for E0–E7 follow. The per-run tables themselves will ship with the artifact bundle upon acceptance rather than inside this PDF; the SHA-256 anchors of the three replayed privacy ledgers are listed in Table 11.

Table 11: SHA-256 anchors for the replayed privacy ledgers behind RQ8 (Table 4). The three files were produced on a second Pi 3B+INA219 deployment (2026-07-24), packaged 2026-09-13, and re-verified digit-for-digit against the packaging manifest on 2026-09-17.

File	SHA-256
`e5_fus4_R20.json`	`96414283210c4711b293848f13ffbd8153af8bd88b167410e0cd80b2e287067a`
`naive_dp_bench_result.json`	`bff6161adf7590aab835f3d25d5623bc86f01bc4e63d4c4692e3ffe9ddb08bce`
`opacus_dp_bench_result.json`	`993d2734c49c4f1851205f42178f783806d16feb1dd1b5b77ca0a68fc7995fc2`

Device and workload. Every metered number is produced on a single Raspberry Pi 3B (1 GB) node at locked 1.2 GHz with an INA219 shunt meter, on a 115-dimensional N-BaIoT slice. The deployed workload is a 1D student network: Conv1d channels $[32, 64, 64]$ (kernel 3, BatchNorm, ReLU, MaxPool, dropout 0.2), global average pooling, then $\mathrm{Linear}(64{\to}32)$ $\to$ReLU$\to$ $\mathrm{Linear}(32{\to}2)$—21,154 parameters, batch 64, replay batch 32, `torch.manual_seed`(42), two CPU threads.

Objective and optimiser. Adaptation minimises cross-entropy (`nn.CrossEntropyLoss`). Two lineages are used and are *not* interchangeable. (a) The energy harness that yields every per-trigger constant drives the adaptation step with plain SGD ($\mathrm{lr} = 0.01$) and, by construction, *random labels*: it measures joules, not accuracy. This is the operational form of Thm. 1—the ledger cannot see label content, only the trigger count. (b) Every accuracy-bearing number comes from the federated driver, which trains with Adam ($\mathrm{lr} = 10^{-3}$, batch 64, 10 epochs, EWC $\lambda = 10$ (Kirkpatrick et al., 2017)); the DP-SGD runs (Abadi et al., 2016) wrap that same optimiser with Opacus. Where scheduler *decision quality* rather than deployment accuracy is measured, the in-stream head is an MLP ($115{\to}64{\to}2$, ReLU hidden layer) and not the deployed student, so those numbers carry no accuracy claim for the deployed model.

Seeds and anchors. Seeds $[0, 1, 2]$ for the 100-round M5 sweeps ($T{=}100$, round size 500, drift rounds $[30, 50, 70, 90]$); 20 paired repetitions for the head-to-head comparison; three sessions (X/Y/Z, $80{+}80{+}40$ honest trials and 3×20 escape trials) for the E5 protocol; all three seed sweeps for the Fig. 5 λ-grid. Each ledger file carries a SHA-256 anchor registered in the artifact manifest, and every constant in Table 7 is re-derived from raw traces by the scripts in the artifact bundle.

J COMPETITION BLANK MAP

The full grid promised in §2 (Table 1 shows the axes only): rows are the competing lines of work, columns the six audit axes. "—" marks a quantity the paper does not report.

REPRODUCIBILITY STATEMENT

All metered experiments run on a single Raspberry Pi 3B (1 GB) node with an INA219 power meter at 1.2 GHz locked frequency; no simulated energy numbers enter any headline claim. We will release, upon acceptance, (i) the turnkey hardware metering protocol (wiring, sampling, idle-baseline calibration), (ii) one-command scripts that re-derive every constant in Table 7 from raw traces, and (iii) all seeds, hyperparameters and SHA-256 anchors for each ledger, including the preregistration ledger that fixes the three-point energy verification; the training setup (workload, objective, optimiser and the two training lineages) is specified in Appendix I. The exhaustive $C(17, 3)$ enumeration of Fig. 4 is pure Python and re-runs in seconds.

Table 12: Six-axis competition grid. "J?" = are physical joules metered; "Aud?" = is the measurement path auditable; "Regret?" = is a regret or budget guarantee proved.

Work	Drift protocol	Energy caliber	J?	Aud?	Regret?	Difference from ours
O-FCL (Serra & Buettner, 2025)	drift-free	none	✗	✗	✗	we run task-free under real drift
FOT (Bakman et al., 2024)	known boundaries	none	✗	✗	✗	we need no task boundary
FedCCFA (Chen et al., 2024)	simulated, RTX-4090	none	✗	✗	✗	zero joule accounting there
FedSSI (Li et al., 2025b)	—	wall-clock	✗	✗	✗	time is not energy
TIMEGATE (Chakraborty et al., 2026)	periodic re-eval	H200 wall-clock	✗	✗	✗	evaluation layer, not trigger layer
GFedCL (Yu et al., 2026)	—	RAM/FLOPs	✗	✗	✗	joule-level budget primitive
RCCDA (Piaseczny et al., 2025)	—	abstract λ	✗	✗	✓	metered constants in the bound
Li et al. (Li et al., 2025a)	shrinking budgets	RAM/FLOPs	✗	✗	✗	physical joules, not abstractions
this paper	real-device drift	INA219 per-phase + GUM	✓	✓	✓	Thm. 1–4

ETHICS STATEMENT

This work measures and audits energy accounting on user-operated edge devices; all data used are network-intrusion benchmark traffic (BaIoT/N-BaIoT) with no personal information. Every reported saving is the one the metered ledgers support, reported alongside the negative results. We see no dual-use risk specific to honest energy accounting.

AI-USE STATEMENT

AI tools were used for copy-editing to improve clarity and readability, and for creating or modifying the figures. AI tools were NOT used for any activity in the required-disclosure category: not to devise or describe the theoretical model or conceptual framework; not to select or generate hypotheses, models, variables, or study designs; not to design the experiment; not to formulate the mathematical propositions, or to devise or write the key ingredients of proofs; not to write proofs; not to implement the data-collection or analysis code; not to clean, generate, process, or restructure the data; and not to analyze the results or produce their interpretation. Those contributions are entirely the authors' own. All experiments were performed by the authors on the hardware pipeline described above, and all numerical claims were verified against raw traces by the authors.