

Online Resource 1: Detailed Results and Evidence Scope

Ali Hamza, Chih-Yi Chiu, Aziz Ahmad, Shehzad, Hassan Iftikhar

Article: Request Segmentation and Metadata Leakage in Low-Rate GPU Power Telemetry: A Single-Platform Audit

Target journal: Journal of Hardware and Systems Security

Correspondence: Ali Hamza, HSE University, Moscow, Russia; akhamza@edu.hse.ru

1 Session inventory and aggregation

Table 1 gives the ten archived test sessions at intersection over union (IoU) 0.3 using the archived detector. Every session has 60 scheduled events; the open-world condition has 43 monitored requests and 17 distractors. Counts refer to monitored-request detection. False alarms are predicted windows minus matches. These are archived observations, not an independent recollection.

Table 1: Test sessions and monitored-request matches

Condition	Requests	Seconds	Windows	Matches	F1
gpu heavy	60	210.5	48	6	0.111
open world	43	81.4	13	5	0.179
back to back	60	34.2	12	11	0.306
idle gaps	60	248.6	27	1	0.023
clean cold	60	94.4	15	6	0.160
clean warm	60	83.5	15	9	0.240
clean hot	60	82.9	16	3	0.079
cpu contention	60	93.7	14	3	0.081
gpu light	60	98.7	15	5	0.133
gpu medium	60	106.9	52	11	0.196

Mean session recall is not exactly the ratio of pooled matches to pooled requests because denominators vary. Similarly, pooled conditional balanced accuracy (0.6875 for power plus time) differs from the average of eligible session balanced accuracies (0.7117). All 60 matched monitored requests are generation requests. There are 466 generation requests, giving coverage of 12.88%; prefill-only coverage is zero out of 117.

2 Conditional inference results

Table 2 reports session means and 95% percentile bootstrap intervals. Detected-window classification uses seven sessions containing both model classes; regression uses nine sessions with at least two generation targets and nonzero target variance. Oracle scores use ten sessions. Each feature set has a separately fitted forest trained on development data. These analyses are post-hoc.

Negative R^2 means larger squared error than a constant equal to the evaluated session’s mean target. The test-session mean is not itself an attacker-accessible predictor. Regression metrics omit requests without matched predictions and do not measure recovery for the whole workload.

Table 2: Conditional session metrics; BA denotes balanced accuracy

Windows	Features	Conditional BA	Token-count R^2
Detected	Time	0.591 [0.425, 0.769]	-0.303 [-0.882, 0.214]
Detected	Amplitude	0.693 [0.538, 0.871]	-0.151 [-0.516, 0.183]
Detected	Power + time	0.712 [0.569, 0.865]	0.300 [-0.006, 0.557]
Detected	All passive	0.676 [0.558, 0.808]	0.310 [-0.018, 0.559]
Oracle	Time	0.843 [0.732, 0.946]	0.698 [0.534, 0.828]
Oracle	Amplitude	0.800 [0.749, 0.842]	0.472 [0.349, 0.587]
Oracle	Power + time	0.890 [0.821, 0.953]	0.922 [0.871, 0.969]
Oracle	All passive	0.884 [0.806, 0.950]	0.918 [0.859, 0.969]

2.1 Training-controlled boundary comparison

The archived detected-window and controlled oracle classifiers use the exact same 144 development event IDs. With power-and-time features, their all-request scores are 0.0757 and 0.7636, respectively. The full-data oracle, trained on 1,043 observable development events, reaches 0.8867. Thus additional oracle training data contribute to the full contrast but do not explain the large controlled gap. On the 60 test event IDs located by the archived detector, detected-window and controlled-oracle balanced accuracy is 0.6875 and 0.6944. Across the seven sessions containing both classes, the mean oracle-minus-detected difference is 0.0255 (95% session-bootstrap interval -0.0459 to 0.1148). This paired conditional result indicates that coverage, rather than classification of already located requests, supplies most of the controlled all-request difference.

3 Feature definitions

The two time features are duration and sample count. The 13 amplitude features are power mean, standard deviation, maximum, minimum, range, 25th/75th/90th percentiles, median, baseline power, and baseline-relative mean, standard deviation, and maximum. The baseline normally uses the preceding 0.5 s; the fallback uses the full trace’s low-power quartile.

The 25-feature power-and-time representation combines these 15 features with integrated power, transition count, transition rate, distinct rounded power states, median and maximum plateau durations, rising and falling transition counts, and mean rising and falling step magnitudes. Power states are rounded to three decimals for plateau identification. Integration uses a trapezoidal rule over timestamps.

The 32-column representation adds utilization mean and standard deviation, mean clock frequency, mean temperature, baseline-relative mean temperature, and two optional probe-latency summaries. Both probe columns are zero in these sessions. Amplitude statistics are an operational feature restriction, not proof of independence from duration or request-phase composition.

4 Additional detector selection

Power hysteresis tests eight configurations: rises of 5, 10, 20, or 40 W crossed with merge gaps of 0.1 or 0.5 s. Release threshold is half the rise; baseline is the tenth percentile of smoothed power; smoothing spans approximately 50 ms; minimum duration is 0.05 s. Validation selects a 5-W rise and a 0.1-s merge gap (F1 0.1735).

The candidate-change grid has 16 configurations: the parameter named penalty is 4 or 16, active-score threshold multiplier is 0.5 or 1, minimum duration is 0.1 or 0.2 s, and merge gap is 0.1 or 0.5 s. Smoothing

uses three samples. Selection gives 4, 1, 0.1 s, and 0.1 s, respectively (validation F1 0.1756). The penalty controls a jump-score quantile here, not a fitted likelihood penalty. Although the legacy detector function has PELT in its name, this path does not invoke the separate PELT-style dynamic program.

The audit JSON retains every validation candidate score and per-session test result. Choosing the added detector families was post-hoc even though selecting their configurations used validation scores. Each selected detector is also evaluated as a complete pipeline: a separate power-and-time classifier is fitted to its matched development windows and scored on its matched test windows.

Table 3: Complete detector/classifier pipelines at IoU 0.3; false-alarm rates are session means

Detector	Train	Test matches	All-request score	False alarms/h
Archived	144	60	0.076 [0.046, 0.108]	488.8
Power hysteresis	135	59	0.091 [0.043, 0.147]	443.9
Candidate change	140	102	0.127 [0.097, 0.159]	1280.2
Normalized threshold	143	71	0.100 [0.065, 0.138]	565.3

The candidate-change pipeline matches 102 test requests and reaches an all-request score of 0.1271, compared with 60 and 0.0757 for the archived pipeline. Its conditional session BA is 0.7900, while its false-alarm rate is 1,280.2 per hour. The power-hysteresis pipeline matches 59 requests and scores 0.0909. These post-hoc results show that detector choice changes the complete identification outcome; none includes a learned online rejection mechanism for false-positive windows.

5 Normalization and duration diagnostics

The archived score divides baseline-centered channels by the larger of 1.4826MAD and 10^{-6} . Across development/validation/test sessions, zero baseline MAD occurs for clock in 7/4/5 sessions and utilization in 2/2/1; power has no zero baseline MAD. Four of the five affected test clock channels later change. The fifth, back-to-back, remains constant and contributes zero rather than a large score. The idle-gap trace reaches unsmoothed weighted contributions of 2.475×10^8 for clock and 2.07×10^7 for utilization, compared with 17.89 for power. These magnitudes motivate a scale diagnostic, not an attribution of all misses to one channel.

For each channel, the median positive baseline scale across development sessions supplies a fixed reference: 8.2810623 W, 2.2239 utilization percentage points, and 333.585 MHz. Zero-MAD fallback uses that reference only for a zero session MAD. Development scale floor uses the maximum of the session scale and reference. Session centering still uses the full trace’s lowest quartile. No test trace contributes to the reference scales.

The normalized-threshold grid has 36 entries: two rules, thresholds 0.75/1.25/2, minimum durations 0.01/0.05/0.2 s, and merge gaps 0.1/0.5 s. Smoothing remains three samples. Mean validation F1 against monitored requests selects zero-MAD fallback, threshold 2, minimum duration 0.2 s, and gap 0.1 s (F1 0.173235). Grid order resolves ties. Earlier detector families retain their original all-event validation criterion, which includes open-world distractors; the normalized grid explicitly targets monitored requests. All test detection results use the same monitored-request target. Family design remains post-hoc.

Table 4 retains every fixed control, including unfavorable outcomes. The zero-MAD fixed control equals the validation-selected configuration. The scale-floor control has a higher test point estimate, but was not selected by validation. Every row independently refits the power-and-time classifier on its matched development windows; confidence intervals hold fitted models fixed and resample test sessions. No superiority test between these controls is claimed.

Table 4: Fixed normalization and duration controls at IoU 0.3; normalization controls retain a 200-ms minimum, and duration controls retain archived normalization

Configuration	Train	Matches	F1	False alarms/h	All-request score
Archived	144	60	0.151	488.8	0.076 [0.046, 0.108]
Zero-MAD fallback	143	71	0.169	565.3	0.100 [0.065, 0.138]
Development scale floor	144	81	0.196	431.9	0.113 [0.076, 0.151]
Archived, 10 ms	158	65	0.135	1320.7	0.086 [0.051, 0.125]
Archived, 50 ms	156	65	0.137	1191.4	0.084 [0.053, 0.120]
Archived, 100 ms	147	61	0.148	606.1	0.072 [0.045, 0.101]

For truth duration d and predicted duration at least m , intersection is at most d and union is at least m , giving $\text{IoU} \leq d/m$. Consequently, $d < 60$ ms precludes matching at IoU 0.3 for the archived 200-ms minimum. Table 5 separates that mechanical exclusion from observed failures. The 110 short prefill requests are excluded by construction; the remaining seven also miss. Even a 10-ms minimum leaves prefill coverage at zero and increases false alarms. This result does not establish whether sensor lag, merged activity, smoothing, or other detector choices explain the remaining misses.

Table 5: Duration-stratified test requests: excluded counts are mathematically unmatchable under the archived rule; the last two columns are matched-request counts

Phase	Duration (ms)	Requests	Excluded	Archived	10-ms minimum
Prefill	< 60	110	110	0	0
Prefill	60–< 200	5	0	0	0
Prefill	≥ 200	2	0	0	0
Generation	< 60	0	0	0	0
Generation	60–< 200	38	0	1	2
Generation	≥ 200	428	0	59	63

6 Collection environment provenance

The packaged collector tokenizes one prompt per request (batch size one), uses model identifiers gpt2 and gpt2-large, and invokes `AutoModelForCausalLM.from_pretrained` without a dtype override or autocast. It does not record actual model dtype or repository revision. The continuous-session manifests identify the GPU and driver, but omit operating-system build and PyTorch, Transformers, and CUDA versions. Earlier Windows 11 project records concern other collections and cannot supply a complete environment for these sessions. Those historical gaps limit physical recollection. Pinned Python/NumPy/scikit-learn and plotting dependencies support CPU reanalysis of the included traces.

7 Scope audit of auxiliary evidence

The earlier manuscript combined several research questions. The revision focuses on the continuous request-identification corpus. Original auxiliary data remain in the working project; selected summary reports are included in Online Resource 2 under their original paths.

7.1 AES controls

The directories `normal_aes_negative_control` and its `_long` variant contain exploratory 200-trace runs with a rank-zero result for the tested key-byte correlation. The later `_final2` run uses 500 traces and does not recover a key byte under the reported analyses. Neither reporting only the later negative outcome nor dismissing the earlier outcomes as chance establishes channel security.

There is also a sensitivity limitation: the fixed-work implementation encrypts many independently varying blocks, whereas the correlation predictor represents a selected block. Aggregate averaging can attenuate that predictor. Fixed-versus-random tests using all-zero and random-template fixed inputs produce sharply different outcomes. These controls depend on predictor alignment and workload construction and are excluded from the revised main security conclusion.

7.2 Tail smearing and telemetry replay

Physical tail smearing uses ten paired sessions in one clean condition and a small adaptive-training collection. Scores are conditional on sparse matches, and overhead varies. It does not establish negligible overhead or general defense efficacy.

Sanitization transforms stored traces rather than changing a driver. The 36-policy grid uses already inspected sessions. Each policy reselects a detector on transformed validation data and retrains inference on transformed development data; it does not simply reuse a frozen raw-data detector. The utility summary pools splits. Coarser publication or quantization does not monotonically lower attack scores. Selecting a favorable setting after examining the grid would not be a locked policy evaluation. These results do not support a deployed privacy recommendation in the revised paper.

7.3 Hardware implementation and workload diversity

Cycle-accurate firmware and routed FPGA reports concern implementation feasibility, not physical board-level power or defense efficacy. Exploratory backend/workload comparisons have different execution structures and duration distributions. They do not replicate the continuous GPT-2 result across platforms. These materials are omitted from the main article and compact primary-analysis archive.

8 Reproducibility and provenance

Online Resource 2 contains 36 telemetry traces, separate ground truth and manifests, the archived split and detector report, revision code and outputs, and raw records for the two new cadence runs. Filenames containing `locked2` are retained for provenance, not to label these post-hoc analyses as fresh locked tests. Attack-feature comparisons can run on a CPU without the GPU or model weights. GPU recollection is separate and requires hardware and model dependencies.