Cubie-Native Tennessee Eastman Fault Detection

Recovered May Result, Full-Coverage Replay, and
Provider-Bound Opus/Fable Comparison

Nick Venezia
Centillion.AI / TrustFortress
github.com/iamdatanick/cubie-research

Version 8.1 evidence revision — 2026-08-02

Evidence status: VERIFIED_PARTIAL. This revision supersedes the historical reading of "100/100/100 at FAR=0" as a held-out or universal guarantee. The May detector and layout search are reproducible on the four exposed Rieth files. On a complete onset-aware replay the recovered detector raises every fault-active alarm, remains silent on the 960-row d00 file, and also raises alarms in 125 clean-labelled pre-onset rows of the three fault files. A complete sealed 50-cell provider, independent compiler parity, untouched confirmation, and production authorization do not yet exist.

Abstract

This paper records an evidence-preserving revision of the Cubie Tennessee Eastman Process (TEP) study. The historical May 2026 search reproduced its selected 100%/100%/100% result on IDV-3, IDV-9, and IDV-15 and 0/960 alarms on the same d00 trace used for calibration. That is an empirical same-trace result, not an out-of-sample probability statement. Replaying the recovered seed-19 detector continuously across every row of the exact four 960-row files produces 2,400/2,400 fault-active alarms and 125/1,440 clean-labelled alarms, or 3,715/3,840 correct rows (96.745%). All 125 clean alarms occur in fault-file preambles before declared onset sample 161; d00 remains 0/960. On a common schedule of 192 non-overlapping 20-row decisions, Cubie scores 184/192 (95.833%), versus strict-request scores of 96/192 for Claude Opus 4.7, 85/192 for Claude Opus 4.8, and 105/192 for Claude Fable 5. Adaptive-thinking variants score 102/192, 86/192, and 102/192. The comparison uses identical CSV bytes and all 52 process channels, but Cubie retains state within each file whereas each model request is stateless. Exact inputs, outputs, model identifiers, timing, token use, cost estimates, provider-returned thinking summaries, and receipt hashes are preserved in `cubie-research` without API credentials.

Contents

1 Claims this revision supports

This revision deliberately separates four claims that were previously conflated:

1. **Historical recovery.** The preserved May layout-search binary reaches the selected saturation point in all 60 reruns (20 seeds times three search algorithms) on the pinned four-file input set.
2. **Complete exposed-file replay.** The recovered seed-19 source path produces the onset-aware row and window results reported here.
3. **Current provider observation.** The 2026-08-02 Claude API results are bound to returned model IDs and response receipts. They are not immutable claims about future aliases or serving stacks.
4. **Production certification.** This claim is **NOT CERTIFIED**. The corrected research profile is incomplete, fault 9 fails its predeclared grouped-CV sensitivity gate, and no untouched confirmation set has been opened under a sealed provider contract.

The phrase "FAR=0" is therefore used only when its denominator and trace are stated, for example "observed 0/960 alarms on the calibration d00 file."

2 Data and scoring contract

The source identity is the Rieth TEP dataset, DOI 10.7910/DVN/6C3JR1. Four LF-normalized CSV files are used. Metadata columns `faultNumber`, `simulationRun`, and `sample` are excluded from model inputs. The full comparison uses `xmeas_1..41` and `xmv_1..11`.

Table 1: Exact full-replay input identity.

File	Rows	SHA-256
d00	960	`5d9244c34896272766bd8bbb1f42281ba13ca58b9fcff66952a5ea53e59c320b`
d03	960	`f9270212506724230ee9122b1a56c4362b4d09dc0b2c8549ebec1708cae7a254`
d09	960	`045abf14d80376a369969e36a0792d60f66dbacbcac5a4bc61f928b56d5e3978`
d15	960	`811d88dc8f0031816a2641f874c13f7cb7c0e4be00a62858323bb84f3b696067`

The d00 file is clean for all 960 rows. Each fault file has 160 clean-labelled rows and 800 fault-active rows under declared onset sample 161. The aligned comparison uses starts $0, 20, \ldots, 940$: 48 decisions per file and 192 total. Each fault file contributes eight pre-onset and 40 fault-active windows, giving 120 fault and 72 clean decisions.

State semantics. Cubie resets once at each 960-row `simulationRun` boundary and retains predictor, Meta, and CUSUM state within the file. Each Anthropic window is an independent request. Matching bytes and decision boundaries therefore make the comparison aligned, but not algorithmically identical.

3 Detector and compiler contract

For active logical cell j, the detector computes a predictor residual and standardized score

$$e_{t,j} = x_{t,j} - \widehat{x}_{t,j}, \qquad z_{t,j} = \frac{e_{t,j}}{\sigma_j}.$$

The executable model must pin means, scales, marginal AR(3) terms, seam and quadratic bases, lag initialization, coefficient order, history rules, active mask, and all CUSUM parameters. A prose equation is not a model bundle.

3.1 Mixed-scale coefficient lowering

Q16.16 cannot represent every recovered predictor term. In particular, $-4.350869883868759 \times 10^{-8}$ rounds to zero at Q16.16 resolution. The research compiler instead represents each coefficient independently:

$$m_i = \mathrm{round}_{\text{ties-even}}(c_i 2^{F_i}), \qquad \widetilde{c}_i = m_i 2^{-F_i},$$

choosing the greatest safe $F_i \in \{0, \ldots, 61\}$ for which m_i fits in `i64`, while rejecting $c_i \neq 0 \Rightarrow m_i = 0$. The per-coefficient bound is

$$|c_i - \widetilde{c}_i| \leq 2^{-(F_i+1)}.$$

Runtime multiplication uses checked `i128` intermediates and checked folding. Numeric representation is necessary but not sufficient: acceptance still requires equality of protected comparisons and final event digests against an independent reference.

3.2 CUSUM semantics

The freshness-aware contract for a future phase-aware profile is

$$S_{t,j} = \begin{cases} \max(0, S_{t-1,j} + |z_{t,j}| - k_j), & m_{t,j} = 1, \\ S_{t-1,j}, & m_{t,j} = 0, \end{cases}$$

where $m_{t,j}$ marks a genuinely fresh analyzer measurement. A confirmation requires strict $S_{t,j} > H_j$ on two consecutive fresh updates by the same cell; different cells cannot combine. Only the confirmed cell resets after emission. Missing or malformed inputs fail closed. Meta and CUSUM histories are separate and all state resets at a run boundary.

The recovered May replay uses its frozen legacy all-samples cadence and sticky latch. Phase-aware behavior must be trained, frozen, and evaluated as a new profile; it is not silently inserted into the historical receipt.

4 Historical result and the onset-aware correction

The May white paper reported a selected 100%/100%/100% result for IDV-3/9/15 and 0/960 d00 alarms. A preserved release binary with SHA-256 `976b9caff511d8c4a9bd68643122859b087c636392e5544cceb6f2802bf7af76` reruns the search; the recovered seed-19 layout has SHA-256 `8fa0fff4b8ac5cb1382a76d45309cd18938adf8cb7458b4f9267e7dbd150a4fc`. All 60 searches again saturate on the pinned files. This proves recovery of the search behavior, not generalization.

Table 2: Recovered Cubie result under complete and historical protocols.

Protocol	Decisions	Fault alarms	Clean alarms	Correct
Continuous rows	3,840	2,400/2,400	125/1,440	3,715/3,840
Aligned 20-row decisions	192	120/120	8/72	184/192
All continuous 20-row starts	3,764	2,400/2,400	125/1,364	3,639/3,764
Historical five-window slice	20	12/12	0/8	20/20

CUSUM is supposed to rise at large residuals; those rises are not suppressed. Under the declared labels the first CUSUM fires in d03, d09, and d15 occur at samples 123, 136, and 99, before onset 161. The three preambles contain 38, 25, and 62 alarmed rows respectively. The historical sparse slice simply did not sample those ranges. Resetting before every window is not a valid remedy: it destroys required history and produces 921/941 d00 window alarms in the retained diagnostic.

5 Full-coverage Cubie versus Anthropic comparison

5.1 Strict final-answer requests

The strict request omits an explicit thinking parameter and requires an exact final token of `FAULT` or `NORMAL`. Fable 5 still uses its provider-mandated hidden thinking.

Table 3: All 52 process channels, every row, 192 aligned decisions.

System	IDV-3	IDV-9	IDV-15	d00 FA	pre-onset FA	Correct
Cubie continuous state	**40/40**	**40/40**	**40/40**	**0/48**	**8/24**	**184/192**
Opus 4.7	16/40	20/40	18/40	19/48	11/24	96/192
Opus 4.8	9/40	14/40	5/40	10/48	5/24	85/192
Fable 5	20/40	24/40	20/40	20/48	11/24	105/192

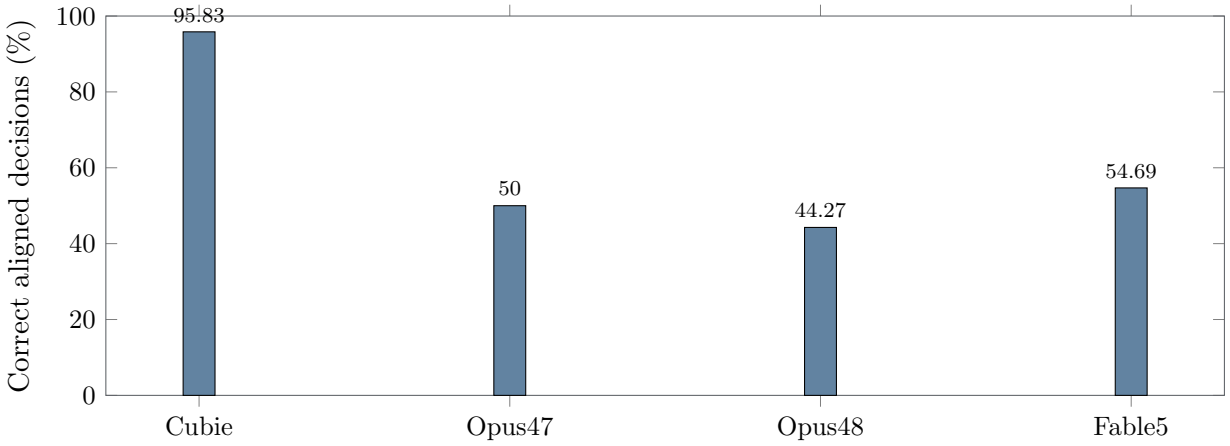

Figure 1: Strict full-coverage comparison. This visual replaces the earlier sparse-window headline. The aggregate coordinates are transcribed into this TeX source from the hash-bound comparison receipt and are independently recomputed from its source receipts during validation.

5.2 Adaptive thinking with provider-returned summaries

The adaptive request sets `thinking.type=adaptive`, requests summarized display, and uses high effort. Only summaries actually returned by the provider are stored.

Table 4: Adaptive-thinking full-coverage comparison.

System	IDV-3	IDV-9	IDV-15	d00 FA	pre-onset FA	Correct
Cubie continuous state	**40/40**	**40/40**	**40/40**	**0/48**	**8/24**	**184/192**
Opus 4.7	19/40	23/40	17/40	18/48	11/24	102/192
Opus 4.8	9/40	15/40	7/40	12/48	5/24	86/192
Fable 5	17/40	24/40	22/40	23/48	10/24	102/192

5.3 Time, tokens, cost, and thinking disclosure

The two API sweeps completed 1,152/1,152 responses without logged request errors. The receipts contain response IDs, resolved model IDs, block types, per-request latency, attempts, usage, signature hashes,

Table 5: Measured full-coverage resource use. USD is a base-price estimate; provider billing remains authoritative.

Mode/model	Seconds	Input	Output	Thinking	Returned summaries	Est. USD
Cubie local	2.746	0	0	0	0	0.000000
Strict Opus 4.7	238.698	945,897	1,644	0	0	4.770585
Strict Opus 4.8	248.760	944,937	1,301	0	0	4.757210
Strict Fable 5	1,582.205	944,937	69,734	68,485	0	12.936070
Adaptive Opus 4.7	225.757	945,897	1,719	77	1	4.772460
Adaptive Opus 4.8	438.422	944,937	9,607	8,311	49	4.964860
Adaptive Fable 5	1,781.090	944,937	68,014	66,766	192	12.850070

and returned summary text. They do not contain an API key. A model's hidden full chain of thought is not available through this API and is not claimed to be recovered. Fable's strict run returned thinking blocks and billed output tokens but no visible summary text.

6 Current corrected profile estimate

6.1 Separate fixed-profile 20-fault evidence stream

The four-file May replay above is distinct from a later fixed-profile, 20-fault run-level receipt. At that source cut, the enforced base/Meta path detected 4,618/10,000 faulty runs; the advisory CUSUM path detected 8,500/10,000 (85.00%) and observed 0/500 alarmed fault-free runs. It detected 17 of 20 fault classes but missed every IDV-3, IDV-9, and IDV-15 run. Advisory specialists detected 8,094/10,000 but alarmed on 94/500 fault-free runs, so they do not support a zero-observed-alarm operating point.

That receipt is planning evidence, not current certification. Its execution manifest marks required replays and provider preflight pending, execution identity `VERIFIED_PARTIAL`, and source rebuildability `NOT_RECOVERABLE`. The receipt SHA-256 is `e4749f4b61ee1e51c327a55c0224a7d6a2b8c2bc17720c6fe4e904cdeeb0294f`. The 0/500 observation is reported with the finite-sample bound in Section 7; it is not a population FAR claim.

6.2 Corrected grouped-training channels

The historical recovered detector is not the corrected schema-v3 provider. Whole-run grouped cross-validation on the corrected training data currently supports only the following advisory estimates:

Table 6: Five-fold grouped-training evidence for corrected specialist channels.

Fault	Detected validation runs	Clean alarms	Median/p95 delay	Gate
IDV-3	463/500 (92.6%)	1/500	329 / 455	Passes 80% floor
IDV-9 best candidate	4/500 (0.8%)	1/500	154 / 193.45	Fails 1% floor
IDV-15	148/500 (29.6%)	0/500	438.5 / 474.65	Passes 25% floor

The partial emitted profile includes only IDV-3 and IDV-15 advisory channels and is not confirmation-eligible. No single corrected whole-detector percentage is defensible until fault 9 is resolved, the 50-cell profile is frozen, the product consumes every compiled coefficient, and a sealed replay plus untouched confirmation exists.

7 Statistical acceptance contract

Rows from one `simulationRun` must never cross folds. Tuning occurs on grouped training folds only. The complete profile is frozen and hashed before an untouched confirmation set is opened once. Sensitivity

and latency floors remain predeclared so the trivial $H \to \infty$ solution cannot pass.

For zero alarmed runs among n representative independent runs, the exact one-sided 95% upper confidence bound is

$$p_{\text{upper}} = 1 - 0.05^{1/n}.$$

For $n = 500$ this is 0.005973551516, approximately 0.5974% per run. Accordingly, the supported statement is "observed 0/500; upper bound about 0.5974% per run," not "the FAR is zero." A bound below 0.1% requires at least 2,995 independent zero-alarm runs; below 0.01% requires 29,956. Fault 21 remains right-censored at the 480-row horizon unless a separately identified 960-row receipt is recovered and bound.

8 Formal evidence and authority boundary

The research repository contains CUB-3240 and CUB-3251–3256 proof triples and a dependency-free reference crate covering coordinate identity, exact layout validation, mixed-scale lowering, checked predictor folds, and freshness-aware same-cell CUSUM semantics. These artifacts establish their stated mathematical and reference-code properties. They do not establish dataset provenance, source-to-product compiler parity, empirical sensitivity, runtime invocation, or production authorization.

The unified registry therefore publishes `cubie-research` rows fail-closed with `deployable=false`. Product deployability may be established only by an audited `cubie-tf` consumer projection. At the time of this revision the research migration is **VERIFIED_PARTIAL** and the corrected provider is not production eligible.

9 Credential-safe reproduction

The replay harness reads `ANTHROPIC_API_KEY` only from the current process environment, never reads a repository `.env`, and never logs or serializes the key. Validate inputs and the schedule before spending tokens:

```
pwsh ./replay_opus_windows.ps1 ‘
  -DatasetDir C:\path\to\tep-csv ‘
  -Models claude-opus-4-7,claude-opus-4-8,claude-fable-5 ‘
  -MeasurementCount 52 -FullCoverage -DryRun
```

Then use a process-scoped key for strict and adaptive variants:

```
$env:ANTHROPIC_API_KEY = ’<your own key>’
pwsh ./replay_opus_windows.ps1 ‘
  -DatasetDir C:\path\to\tep-csv ‘
  -Models claude-opus-4-7,claude-opus-4-8,claude-fable-5 ‘
  -MeasurementCount 52 -FullCoverage ‘
  -OutFile ./full52-strict.json

pwsh ./replay_opus_windows.ps1 ‘
  -DatasetDir C:\path\to\tep-csv ‘
  -Models claude-opus-4-7,claude-opus-4-8,claude-fable-5 ‘
  -MeasurementCount 52 -FullCoverage ‘
  -AdaptiveThinking -ThinkingEffort high ‘
  -OutFile ./full52-adaptive.json
Remove-Item Env:ANTHROPIC_API_KEY
```

Cubie search and source replay commands, exact prerequisites, and expected hashes are documented in `evidence/tep/may-2026-opus47-opus48/README.md`. Model aliases and prices can change. The harness embeds the price schedule effective 2026-08-01 and records that effective date separately from receipt generation time; a later schedule must be explicitly dated rather than silently relabeling these constants. Provider billing remains authoritative.

10 Receipt inventory

Table 7: Canonical 2026-08-02 receipts. Each has a sibling `.sha256`.

Receipt	SHA-256
`retest_20260801_strict_opus47_opus48_fable5.json`	`a7f8c4f2d195b9ededa837c46a764cef562957c0e0d9f55e91ad8acb6a47e87d`
`retest_20260801_adaptive_thinking_opus47_opus48_fable5.json`	`0bc7d7e43b8447ed878d2d71cdff36c7adf9f274cff831774172648b49084359`
`retest_20260801_cubie_same_data.json`	`22c041fe2c80e9918a67fd28e67203c400fad6c0e3e0a4f121eed756e66d2e9c`
`retest_20260801_cubie_seed19_windows.json`	`c7dcaff176cdd66ad2caf5d49a8cbadde69cab9371dc38ccc4b8ed5b81540f3e`
`retest_20260801_full52_strict_opus47_opus48_fable5.json`	`fc4873983a047fac59489a76819bdbf7469966b7043a4c76bc88a10dabb460d1`
`retest_20260801_full52_adaptive_thinking_opus47_opus48_fable5.json`	`321518514e240018578f9c74d66c606aaefb01f903aa1d40e6e591505ffcf0ae`
`retest_20260801_full52_comparison.json`	`d9bf7233e8f416e682644be8d75bb56a5b020f20ccda8533e9ceec7ecbfbe4cb`

11 Remaining certification gates

1. Recover or independently rebuild and cross-check the missing frozen selector, evaluator, policy, and receipt artifacts.
2. Resolve IDV-9 without weakening the predeclared sensitivity floor; freeze and hash a complete corrected 50-active-cell profile.
3. Establish source-float, independent compiler, reference runtime, and product runtime parity for the full mixed-scale predictor set.
4. Seal `source_manifest.json`, `model_manifest.json`, and `execution_receipt.json`; replay all exposed runs with exact reset, cadence, censoring, and malformed-input behavior.
5. Open a genuinely untouched confirmation set exactly once under the frozen acceptance contract.
6. Produce CPU and target-hardware latency/throughput, inventory, soak, and operational authorization receipts.

12 Conclusion

The recovered Cubie detector is materially stronger than the tested stateless LLM protocols on the four exposed files, and its selected May search behavior is reproducible on the pinned files. The scientifically important correction is that the historical "perfect" slice hid pre-onset alarms and reused the selection/calibration traces. The complete replay is 96.745% correct by row and 95.833% on the aligned window schedule, not universal perfection. The corrected profile is currently incomplete and remains non-production. By preserving source bytes, model and dataset identities, full response receipts, token

and timing accounting, formal boundaries, and the outstanding gates, this revision turns the original demonstration into a reproducible research artifact without overstating what has been certified.