Screen Broadly, Diagnose in Context: Hierarchical Time Series Anomaly Detection

Anonymous authors
Paper under double-blind review

Abstract

Time series anomaly detection is essential to the reliability of cloud services, Internet of Things systems, and communication networks. Although many existing methods are sensitive to changes in data, statistical deviation alone does not establish whether a change reflects a genuine operational anomaly or benign variation. This gap between numerical sensitivity and operational meaning remains a key obstacle to industrial deployment. Large language models can incorporate operational semantics, but directly interpreting large volumes of numerical time-series data remains challenging. We introduce VeraTS, a hierarchical framework that addresses this gap through two complementary stages. The first, *Time Series Perception for Signal Detection*, screens broadly for candidate deviations while preserving event coverage on labeled development data. The second, *Expert System for Semantic Judgment*, translates operator policies into explicit rules for harmful direction and persistence, then deterministically verifies candidate deviations against these rules. We evaluate VeraTS with supplied policy programs on operational KPI series from backbone-link monitoring data. The frozen screen overlaps all evaluation events. VeraTS achieves the highest average scores among the evaluated methods and leads ten of twelve category–metric comparisons on real backbone links, with experiments characterizing its robustness and inference cost.

1 Introduction

Cloud services (Shen et al., 2024; Zhou et al., 2024), Internet-of-Things deployments (Liang et al., 2025; Liu & Zhang, 2026), and communication networks (Cui et al., 2026a; Ma et al., 2026) depend on continuous telemetry to reveal changes in their operation. As infrastructure evolves toward more demanding service requirements, including those envisioned for 6G (International Telecommunication Union, 2023), deciding which changes deserve attention becomes increasingly important. Time-series anomaly detection (TSAD) sits at this decision boundary: its outputs influence which signals engineers investigate, which alerts they escalate, and how quickly they respond to degradation. For a detector to be useful in this setting, it must direct limited operator attention toward events that warrant intervention (Ewaschuk, 2016).

Much of the progress in TSAD has come from improving how models represent temporal patterns. Methods such as KAN-AD (Zhou et al., 2025) and CATCH (Wu et al., 2025), alongside forecasting architectures such as iTransformer (Liu et al., 2024a), provide increasingly capable ways to identify departures from expected behavior. Yet identifying a departure leaves an operational question unresolved: does this change violate the conditions under which the service is supposed to operate? A rise in request volume may be expected during an announced campaign but require investigation under a fixed-capacity constraint; similarly, rising success and failure rates can share the same standardized waveform while demanding opposite responses. The missing information in these examples is the operating policy, which cannot be recovered simply by modeling the observed waveform more accurately.

Figure 1 makes this ambiguity explicit by applying two policies to an identical history and trace. When low values are harmful, the drop is an anomaly; when high values are harmful, the rise is an anomaly. Because the numerical input is unchanged, a signal-only detector cannot satisfy both labelings. The same distinction appears in network round-trip time (RTT): an excessive-latency policy treats a sustained increase as harmful, while a decrease may indicate improvement. These examples

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.

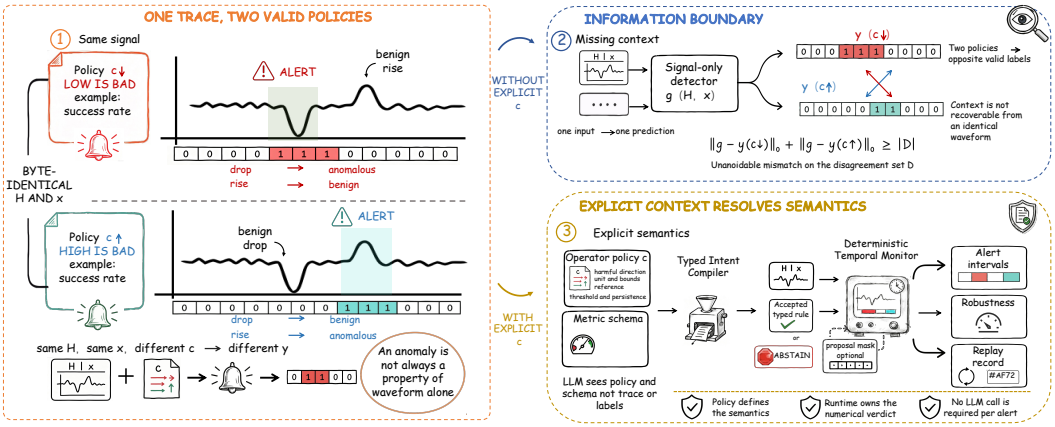


Figure 1: Why a waveform alone does not determine an operational alert. Opposite harmful-direction policies assign different labels to the same history and trace (left), creating an ambiguity that a signal-only detector cannot resolve (middle). Explicit policy and metric context resolve this ambiguity through a checked rule and deterministic execution (right). The policies form a conceptual counterfactual.

suggest that operational detection should combine evidence that a signal has changed with an explicit criterion for deciding whether the change matters.

To bridge the gap between TSAD research and operational deployment, we introduce VERATS, a two-stage framework that connects statistical signal screening with semantic verification. The central challenge is to combine broad sensitivity to numerical changes with the contextual selectivity needed for actionable alerts. Large language models can incorporate metric descriptions and operator requirements (Liu et al., 2025a), but directly interpreting long numerical sequences remains constrained by input length and time-series representation (Alnegheimish et al., 2024; Zhou & Yu, 2025). We address these competing demands through a division of responsibilities: *screen broadly, then diagnose in context*.

The first stage addresses the risk that selective screening discards a true event before diagnosis can assess it. It combines inexpensive, direction-free features to generate broad signal-level candidates, minimizing retained observations subject to zero missed development events, defined as overlap with every labeled event. This coarse screening step identifies where a change may have occurred without prematurely deciding whether that change is operationally harmful.

The second stage addresses the ambiguity between a statistical deviation and an operational anomaly through a rule-based expert system. It combines metric semantics with executable rules for harmful direction, numerical reference, threshold, and persistence; an LLM interface can translate operator requirements into checked, reusable rules for deterministic temporal diagnosis. The aim is to suppress benign fluctuations while preserving recall and avoiding additional missed alerts, so that numerical evidence is qualified by the conditions that matter to the service. Specifically, our contributions are threefold:

- We identify and formalize a deployment gap in TSAD: signal-only methods characterize statistical deviations, whereas site reliability engineers (SREs) need alerts about changes that violate operational requirements. We formulate a detection workflow that combines evidence from the statistical domain with verification rules from the semantic domain.
- We develop VERATS, a two-stage solution that couples inexpensive, coverage-constrained coarse screening with semantic anomaly detection. The first stage targets zero missed development events; the second supports LLM-assisted rule construction and applies metric semantics and checked rules to filter candidates while limiting further misses.
- We evaluate the framework with supplied policy programs on operational KPI series from real backbone-link monitoring data, with explicit metric semantics and policies. Under the reported comparison protocols, VERATS leads Event-F1* in all four morphology categories. The anonymous project repository is at `https://anonymous.4open.science/r/VeraTS/`.

2 Problem and Design Requirements

To make the distinction between a deviation and an alert precise, consider a reference history H, an evaluation sequence $x_{1:T}$, and context c specifying the metric schema and operating policy. The target labels depend on all three:

$$y_{1:T} = f(H, x_{1:T}, c). \tag{1}$$

The role of context is substantive: it determines which numerical changes count as violations, while independent incident evidence would be needed to establish whether those violations correspond to actual operational failures. We therefore formulate detection as a policy-conditioned decision and ask what information each stage must preserve to support it.

Supplying the missing decision criterion. Suppose the same (H, x) is evaluated under two policies with labels $y^{(1)}$ and $y^{(2)}$, and let $D = \{t : y_t^{(1)} \neq y_t^{(2)}\}$. A deterministic signal-only detector produces the same prediction $g = g(H, x)$ in both settings, so

$$\|g - y^{(1)}\|_0 + \|g - y^{(2)}\|_0 \geq |D|. \tag{2}$$

Every disagreement forces an error under at least one policy, regardless of how accurately the detector models the signal. This observation motivates an explicit context interface; whether that interface is implemented by an LLM, a deterministic parser, or a supplied rule is a separate question.

Preserving evidence for the decision. In a hierarchy, context alone is insufficient if screening discards the observations that reveal a violation. Let $q \in \{0,1\}^T$ be the candidate mask, $Q = \{t : q_t = 1\}$ its retained positions, and $\mathcal{E}$ the labeled event set. We distinguish proposal overlap recall from the retained observation fraction:

$$R_{\text{prop}}(q) = \frac{1}{|\mathcal{E}|} \sum_{E \in \mathcal{E}} \mathbf{1}\{E \cap Q \neq \varnothing\}, \qquad \kappa(q) = \frac{|Q|}{T}. \tag{3}$$

Although $R_{\text{prop}} = 1$ ensures that every event is touched, it permits most of an event's extent to be removed. The consequence is visible in a point-level hard gate before postprocessing: with $\epsilon_s = \Pr(q = 0 \mid y = 1)$ and $\epsilon_d = \Pr(\text{diagnosis negative} \mid y = 1, q = 1)$, the total miss probability is

$$\epsilon_{\text{total}} = \epsilon_s + (1 - \epsilon_s)\epsilon_d. \tag{4}$$

Improving diagnosis cannot recover points excluded by this gate, which motivates evaluating retained support alongside overlap recall. Our implementation can subsequently fill short gaps through closing, so final interval localization is measured directly rather than inferred from this decomposition. Throughout the paper, zero development misses denotes an empirical selection constraint, not a guarantee for arbitrary future incidents; the theory appendix gives the precise assumptions and finite-sample limits.

3 A Two-Stage Monitoring Architecture

The formulation above suggests two requirements for the architecture: screening should preserve possible events without committing to their meaning, and diagnosis should evaluate that meaning using sufficient temporal evidence. VeraTS implements them through a numerical screening path and a reusable policy path, which meet in the temporal monitor shown in Figure 2.

3.1 Stage 1: Coarse Signal Screening

The screening stage is designed to avoid discarding an event before its operational meaning can be assessed. For a finite family of masks q^λ indexed by threshold λ, we therefore choose the smallest retained fraction compatible with full development-event coverage:

$$\lambda^* \in \arg\min_{\lambda \in \Lambda} \kappa_{\text{dev}}(q^\lambda) \quad \text{s.t.} \quad R_{\text{prop,dev}}(q^\lambda) = 1. \tag{5}$$

The search grid and tie order are fixed before evaluation, with an all-ones mask as a fallback if no screened configuration satisfies the constraint. This makes coverage the primary screening objective

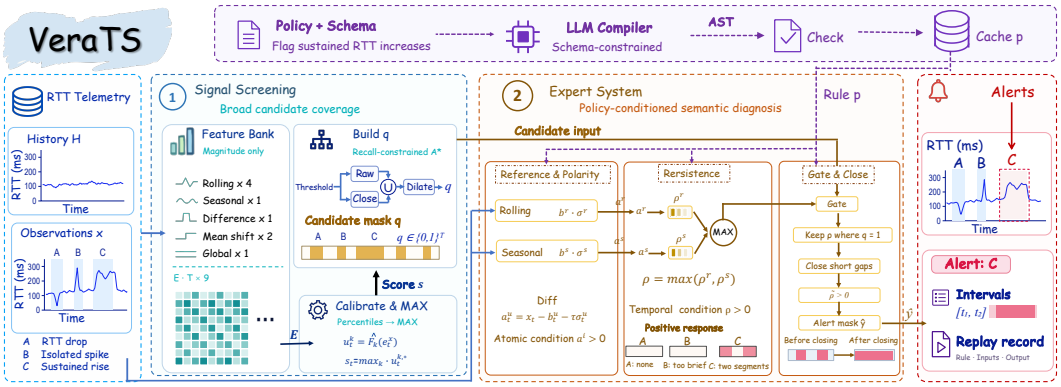


Figure 2: Screening and diagnosis in VERATS. The numerical path combines nine direction-free features into a candidate mask. The policy path compiles, checks, and caches the rule governing reference choice, harmful direction, persistence, and closing. Rolling and seasonal evidence are qualified separately before their temporal margins are combined, gated, and closed into alert intervals. The schematic RTT example illustrates how a drop and an isolated spike can become candidates while only the sustained rise satisfies the policy.

and selectivity a secondary one; it is empirical constrained selection, without the sampling assumptions needed for conformal risk control (Angelopoulos et al., 2024; Farinhas et al., 2024).

To construct these masks, we combine complementary views of numerical change: four rolling residuals, one seasonal residual, a first difference, two mean-shift features, and a global-reference residual. Each feature magnitude $e_t^k \geq 0$ is mapped through a historical empirical CDF, allowing deviations measured on different scales to contribute to a common score:

$$s_t = \max_k \widehat{F}_k(e_t^k), \qquad q^\lambda = \mathrm{Dilate}_r([s \geq \lambda] \vee \mathrm{Close}_g[s \geq \lambda]) \,. \tag{6}$$

The maximum admits evidence from any feature, while remaining agnostic to whether an increase or a decrease is harmful. We retain both the raw threshold crossings and their closed mask so that finite-array closing cannot erase boundary triggers, then dilate the union to include nearby observations. At fixed calibrations and operators, adding a feature can only enlarge the candidate set, matching the stage's emphasis on broad coverage.

From point proposals to temporal support. A mask that touches an event may still retain too little of it for reliable diagnosis. To study this interface, we also partition the trace into non-overlapping packets $I_B(t)$ of width B and admit each packet containing a proposal:

$$q_t^{\lambda,B} = \max_{u \in I_B(t)} q_u^\lambda. \tag{7}$$

Because the expanded mask contains the original proposals, this intervention adds temporal support while leaving the rule and reference history unchanged. Reference state continues across packet boundaries, allowing the experiment to isolate what is gained by retaining more observations rather than by changing the diagnosis itself.

3.2 Stage 2: Semantic Anomaly Detection

Having identified where a change may have occurred, the second stage asks whether the retained evidence satisfies the operating policy. We represent that policy as a closed abstract syntax tree specifying metric units and bounds, harmful direction, numerical references, a robust threshold, persistence, gap closing, and candidate use. Algorithm 1 translates operator requirements and a metric schema into a checked rule and deterministic monitor; observations and candidate masks enter only the monitor. These checks establish an executable contract, while fidelity to the operator's intent remains a separate property to evaluate.

For example, an operator specifies: "For latency, use a 288-observation rolling median and a seasonal median over seven previous 288-observation periods; flag increases above two robust scales that

Algorithm 1 From an operator policy to an executable rule system

Require: Operator policy p, metric schema S, closed rule grammar $\mathcal{G}$
Ensure: Checked rule R, executable formula Φ_R, hash h_R, monitor M_R; or rejection/failure
1: $D \leftarrow \textsc{CompileLLM}(p, S; \mathcal{G})$ ▷ structured compile-or-reject output
2: Stop on rejection or parsing failure; otherwise set $R \leftarrow D$.rule.
3: Check allowed fields, units, bounds, threshold range, and $1 \le m \le w$.
4: Require R.metric $= S$ and R.source_text $= p$; stop if any check fails.
5: If enabled, review R against (p, S) with at most two repair rounds; recheck and review corrected rules, stopping if unresolved.
6: $\Phi_R \leftarrow \textsc{Lower}(R)$: directional atoms $\rightarrow$ per-reference persistence $\rightarrow$ reference disjunction $\rightarrow$ candidate gate $\rightarrow$ closing.
7: $h_R \leftarrow$ SHA256($\textsc{CanonicalJSON}(R)$); save the accepted rule and audit record for reuse.
8: $M_R \leftarrow \textsc{RuleMonitor}(R)$; **return** (R, Φ_R, h_R, M_R).

persist for five consecutive observations within nine, require candidate support, and apply closing width 48." The resulting fields are direction: increase; reference: dual; $\tau = 2$; $(m, w) = (5, 9)$; closing width $g = 48$; and candidate gating: true. The numerical experiments use supplied rules; this interface supports authoring them from operator requirements.

The contract first turns a numerical deviation into directional evidence. For reference u with baseline b_t^u, positive scale σ_t^u, and threshold τ, we define the atomic margin

$$a_t^u = \begin{cases} x_t - b_t^u - \tau\sigma_t^u, & \text{increase is harmful,} \\ b_t^u - x_t - \tau\sigma_t^u, & \text{decrease is harmful,} \\ |x_t - b_t^u| - \tau\sigma_t^u, & \text{either is harmful.} \end{cases} \tag{8}$$

A positive margin indicates a violation of the directional threshold, but an isolated violation need not satisfy the policy's duration requirement. We therefore require m consecutive positive margins within the last w observations:

$$\rho_t^u = \max_{j=0}^{w-m} \min_{i=0}^{m-1} a_{t-j-i}^u, \qquad \rho_t = \max_u \rho_t^u. \tag{9}$$

Persistence is evaluated separately for each reference before the branch maximum is taken, so a qualifying run must be supported by one reference rather than assembled from alternating branches. The resulting robustness is a signed margin, not a probability, and can remain positive while a recently completed run lies within the temporal window.

To turn this evidence into alerts, the monitor applies the candidate mask after persistence, replaces excluded margins with $-\infty$, and closes short gaps before thresholding at zero. Closing can restore positions outside the mask, which is why candidate support and final alert support are accounted for separately. Although the temporal core is causal, symmetric closing makes the complete evaluated pipeline offline; Appendix D specifies the operators and initialization conditions.

Figure 3 follows this sequence from a proposed deviation to a contextual verdict. The latency increase in A supplies persistent evidence in the harmful direction and survives the full rule, whereas the dip in C opposes that direction and is rejected. The constructed request increase in B exposes the interaction with support: five seasonal temporal positives occur outside the candidate mask, leaving their intersection empty. Removing direction instead produces 49 positive observations in the request window and 8 in the latency-dip window, compared with none under the full rule. These examples show how direction, persistence, and support jointly determine an alert; the rolling and seasonal branches are complementary numerical references within this rule, rather than separate language agents.

3.3 Amortized Computation and Replay

Separating the policy from its execution also separates two sources of cost. Policy compilation can be amortized over streams sharing the same rule, while numerical screening and monitoring operate on the observations themselves. For N observations, K policy versions, screening cost c_s, and diagnosis cost c_d per retained observation, an idealized model is

$$C_{\text{hier}} = Nc_s + \kappa Nc_d + KC_{\text{compile}}. \tag{10}$$

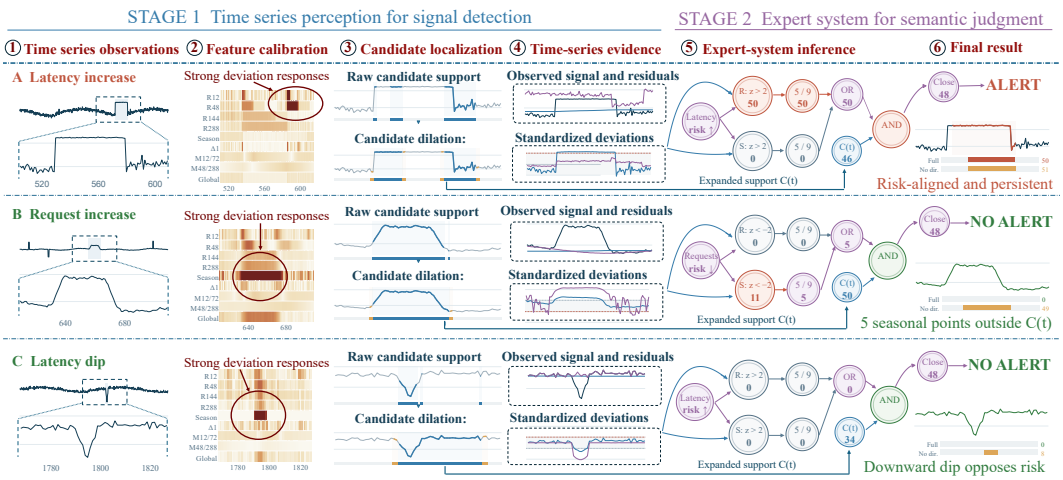


Figure 3: How similar numerical evidence leads to different decisions. The six columns trace observations through feature responses, candidate support, reference-based evidence, rule inference, and final output. A sustained latency increase triggers an alert (A), while a constructed request increase under a low-is-bad policy (B) and a latency dip under a high-is-bad policy (C) do not. Node values count positive observations in the displayed window; $5/9$ requires five consecutive violations within nine observations, followed by closing width 48.

Relative to dense diagnosis with the same compiled policies, routing reduces this cost when $c_s < (1-\kappa)c_d$, before packaging and context overhead. The current implementation still computes reference features over the full trace, so this condition motivates selective execution while actual runtime is established by measurement in Section 4. For reproducibility, each alert is accompanied by the rule and monitor version, input hashes, candidate mask, robustness, and output intervals, allowing its numerical computation to be replayed in the recorded environment.

4 Experiments

Our experiments test the central design of VeraTS: screen broadly to preserve event coverage, then diagnose in context to produce selective, coherent alerts.

Dataset. We evaluate on real backbone-link monitoring data comprising KPI series and evaluation events. Intervals of actual network failures are labeled anomalous. Separately, 29 annotators assign the anomaly morphologies to four categories: Mutation, Step, Trend, and Wave; these annotations do not affect the binary failure labels.

Evaluation protocol. We distinguish score ranking, event discovery, and interval localization to assess different parts of the detection task. Table 1 reports VUS-PR and VUS-ROC for ranking, alongside Event-F1* for discovery. For Event-F1*, we compute the category mean of per-series EasyTSAD event-PA F1 at each integer percentile from 90 to 99, then select the largest mean using evaluation labels.

4.1 Detection across Morphology Categories

Broad screening preserves access to relevant events while concentrating subsequent diagnosis: the frozen candidate mask overlaps all 66 evaluation events and retains 32.8% of observations. Within this support, policy-conditioned diagnosis selects deviations with the prescribed direction and duration, turning broad numerical coverage into targeted alerts. The complete VeraTS pipeline achieves the highest Event-F1* in all four categories and leads both VUS metrics on Mutation, Step, and Trend, ranking first in ten of the twelve category–metric columns. These results support the division of labor between coverage-oriented screening and selective temporal diagnosis across the four evaluated morphologies.

The threshold curves in Figure 4 show that the resulting scores support strong event discovery across a broad range of operating percentiles. Among the plotted baselines, VeraTS leads at every tested

Table 1: Detection quality by morphology category (↑). Red and blue mark the largest and second-largest positive displayed values, respectively; subscripts give the selected percentiles.

Paradigm	Method	Mutation			Step			Trend			Wave		
		VUS PR	VUS ROC	Event F1*	VUS PR	VUS ROC	Event F1*	VUS PR	VUS ROC	Event F1*	VUS PR	VUS ROC	Event F1*
Statistical	MatrixProfile	0.110	0.656	0.105_{99}	0.356	0.897	0.274_{99}	0.199	0.803	0.165_{99}	0.172	0.521	0.035_{96}
	SAND	0.167	0.600	0.138_{99}	0.423	0.802	0.231_{99}	0.075	0.481	0.071_{99}	0.166	0.552	0.195_{98}
	SubLOF	0.052	0.653	0.033_{96}	0.039	0.510	0.031_{98}	0.101	0.595	0.099_{99}	0.140	0.461	0.154_{99}
	KMeansAD	0.083	0.454	0.037_{99}	0.203	0.651	0.218_{99}	0.153	0.576	0.182_{99}	0.156	0.555	0.126_{99}
Prediction	AR	0.092	0.741	0.244_{99}	0.094	0.684	0.280_{99}	0.076	0.588	0.242_{99}	0.108	0.387	0.017_{92}
	LSTMAD	0.222	0.839	0.269_{99}	0.177	0.742	0.306_{99}	0.130	0.669	0.268_{99}	0.108	0.377	0.032_{94}
	POLY	0.171	0.854	0.062_{97}	0.249	0.805	0.035_{95}	0.202	0.753	0.042_{99}	0.128	0.217	0.000_{90}
	KAN-AD	0.072	0.670	0.246_{99}	0.092	0.689	0.252_{99}	0.070	0.618	0.261_{99}	0.109	0.395	0.030_{97}
	TShape	0.066	0.624	0.223_{99}	0.093	0.694	0.255_{99}	0.066	0.568	0.172_{99}	0.156	0.579	0.098_{98}
	DLinear	0.056	0.633	0.249_{99}	0.071	0.681	0.252_{99}	0.055	0.598	0.244_{99}	0.105	0.398	0.040_{90}
	TFC	0.054	0.606	0.248_{99}	0.072	0.574	0.247_{99}	0.053	0.513	0.249_{99}	0.110	0.397	0.029_{90}
	TimeXer	0.127	0.789	0.234_{99}	0.136	0.755	0.277_{99}	0.100	0.700	0.223_{99}	0.106	0.378	0.030_{97}
Reconstruction	AE	0.088	0.702	0.037_{99}	0.170	0.643	0.127_{98}	0.133	0.652	0.079_{99}	0.247	0.601	0.411_{99}
	EncDecAD	0.190	0.872	0.151_{99}	0.204	0.840	0.109_{99}	0.164	0.802	0.093_{99}	0.141	0.566	0.050_{96}
	SRCNN	0.068	0.654	0.030_{90}	0.097	0.748	0.095_{98}	0.053	0.640	0.070_{98}	0.153	0.566	0.140_{99}
	TranAD	0.364	0.960	0.289_{98}	0.353	0.945	0.333_{98}	0.398	0.754	0.365_{99}	0.426	**0.829**	0.342_{99}
	Donut	0.043	0.569	0.016_{96}	0.099	0.669	0.010_{97}	0.045	0.562	0.013_{91}	0.169	0.618	0.092_{98}
	FCVAE	0.139	0.845	0.265_{99}	0.183	0.832	0.304_{99}	0.129	0.772	0.280_{99}	0.098	0.303	0.027_{95}
	ModernTCN	0.228	0.934	0.150_{99}	0.241	0.902	0.175_{99}	0.159	0.861	0.068_{99}	0.103	0.365	0.032_{97}
	SEA-FGT	0.440	0.935	**0.368**$_{99}$	0.511	0.951	0.444_{99}	0.268	0.769	0.254_{99}	0.213	0.618	0.288_{99}
Foundation models	TimesNet	0.162	0.867	0.101_{99}	0.236	0.804	0.118_{99}	0.161	0.737	0.072_{99}	0.090	0.232	0.007_{92}
	FITS	0.276	0.901	0.178_{99}	0.270	0.873	0.222_{99}	0.207	0.846	0.132_{99}	0.085	0.209	0.011_{90}
	Sundial	0.291	0.923	0.188_{99}	0.300	0.858	0.283_{99}	0.173	0.786	0.106_{99}	0.280	0.656	0.283_{99}
	Timer	0.313	0.952	0.201_{99}	0.260	0.916	0.181_{99}	0.163	0.820	0.089_{98}	0.180	0.594	0.169_{98}
	Timer-XL	0.301	0.942	0.226_{99}	0.209	0.907	0.140_{99}	0.162	0.799	0.092_{99}	0.153	0.549	0.161_{98}
	Time-MoE	0.298	0.931	0.202_{99}	0.285	0.853	0.208_{99}	0.161	0.772	0.109_{99}	0.245	0.662	0.230_{98}
	TimeMixer	0.217	0.806	0.260_{99}	0.257	0.834	0.318_{99}	0.175	0.789	0.242_{99}	0.107	0.386	0.034_{90}
	Time-o1	0.118	0.752	0.257_{99}	0.109	0.708	0.268_{99}	0.094	0.673	0.217_{99}	0.110	0.400	0.024_{92}
LLM	kimi-k3	0.440	0.962	0.093_{93}	0.463	**0.959**	0.222_{96}	0.395	0.877	0.194_{96}	0.493	0.645	0.148_{90}
	glm-5.3	**0.480**	**0.969**	0.134_{93}	0.495	0.955	0.333_{96}	0.360	0.891	0.167_{97}	0.481	0.697	0.222_{90}
	Qwen-3.8	0.312	0.679	0.291_{90}	0.437	0.765	0.397_{90}	0.401	0.798	0.340_{90}	0.455	0.572	0.298_{90}
	step-5	0.034	0.502	0.103_{90}	**0.633**	0.831	**0.567**$_{90}$	**0.550**	0.763	**0.528**$_{96}$	**0.595**	**0.709**	0.333_{90}
	DeepSeek-v4	0.293	0.897	0.219_{90}	0.511	0.933	0.392	0.443	0.835	0.206_{90}	0.468	0.635	0.304_{90}
	GPT-6	0.373	0.964	0.049_{90}	0.317	0.942	0.119_{97}	0.349	**0.942**	0.119_{93}	0.467	0.697	**0.431**$_{90}$
Hierarchy	**VeraTS (ours)**	**0.934**	**0.992**	**0.542**$_{97}$	**0.958**	**0.996**	**0.741**$_{98}$	**0.688**	**0.964**	**0.561**$_{99}$	**0.551**	0.700	**0.963**$_{99}$

threshold on Step, Trend, and Wave, and at nine of ten thresholds on Mutation. Deterministic numerical execution also makes this detection quality computationally practical: VeraTS takes 62.2 milliseconds per 1,000 observations with 149.4 MiB peak RSS, combining the highest average VUS-PR and lowest peak RSS among the twelve plotted implementations.

How the rule components shape evidence. Direction and persistence make the operating policy effective at the evidence level. Figure 5 isolates their contributions before candidate gating and closing. Among normal observations with initially positive unsigned evidence, direction reduces the positive-evidence fraction from 100.0% to 29.4% while preserving the 96.4% fraction among anomalous observations; point F1 rises from 0.361 to 0.652. Persistence further reduces this normal positive-evidence fraction to 10.4%, retains positive evidence for 91.2% of anomalous observations, and raises point F1 to 0.740. Direction concentrates evidence on harmful changes, and persistence promotes sustained violations. Their successive gains explain how explicit policy conditions improve selectivity while retaining substantial anomalous evidence.

Matched ablations connect these benefits to the complete rule. Adding direction alignment raises VUS-PR from 0.448 to 0.783, while incorporating persistence raises event-weighted F1-E from 0.779

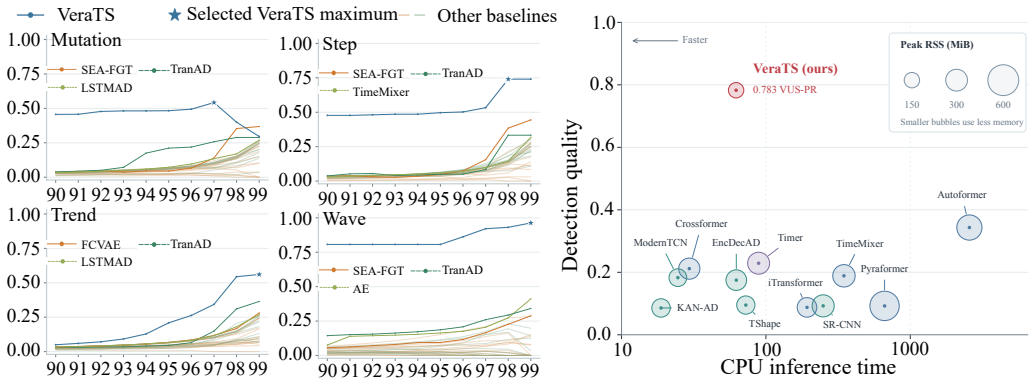


Figure 4: Detection quality across thresholds and inference cost. Left: category-mean EasyTSAD event-PA F1 at percentiles 90–99 for VeraTS and 28 baselines; each category's three strongest baselines are highlighted, with stars at VeraTS's label-selected maxima. Right: twelve implementations on one CPU thread. Axes show milliseconds per 1,000 observations (log scale) and category-balanced VUS-PR; bubble area denotes peak RSS (MiB). Times are medians of three passes.

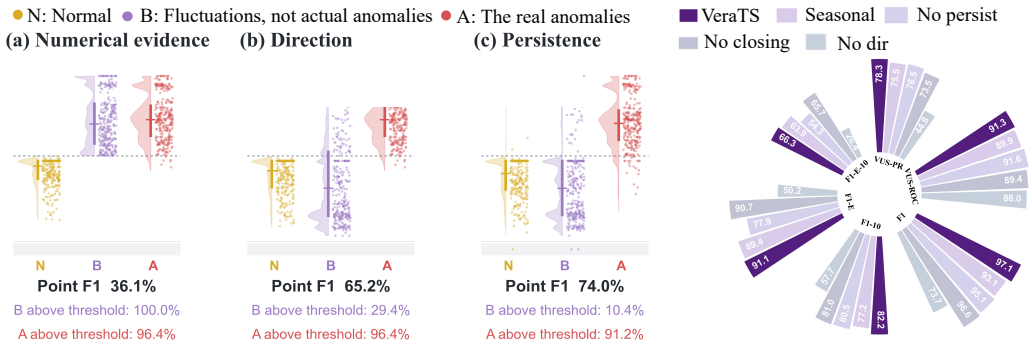


Figure 5: Effects of direction and persistence. Left: unsigned evidence, direction alignment, and persistence before gating and closing. A: anomalous observations; B: normal observations with initially positive unsigned evidence; N: remaining normal observations. F1 is unadjusted point F1. Right: category-balanced metrics ($\times 100$); four F1 metrics use unrestricted per-series maxima. F1 uses point adjustment, E adds logarithmic event weights, and -10 requires a hit in the first ten event observations.

to 0.911. Complementary references and bounded closing further strengthen the score representation: the full rule reaches VUS-PR 0.783, compared with 0.755 for seasonal-only evidence and 0.735 before closing. Rolling evidence extends the available reference context, and closing consolidates qualified temporal support into coherent intervals. Together, these controls show how policy meaning, temporal qualification, reference context, and interval continuity contribute to the measured detection gains.

4.2 Sensitivity of Temporal Diagnosis

The sensitivity study tests whether effective temporal diagnosis is supported by a neighborhood of parameter choices. Figure 6 varies the reference window, minimum persistence, and sieve threshold in three pairwise grids, linking the evidence baseline, duration requirement, and candidate admission. Several neighboring settings yield similarly high strict event F1, providing flexibility in the choice of reference and temporal scales. The high-scoring regions highlight the value of a sustained-evidence requirement: persistence supports both selective alerts and coherent localization under strict interval matching. Here m and $w = m + 4$ vary together; the grids characterize sensitivity without changing the previously selected operating point.

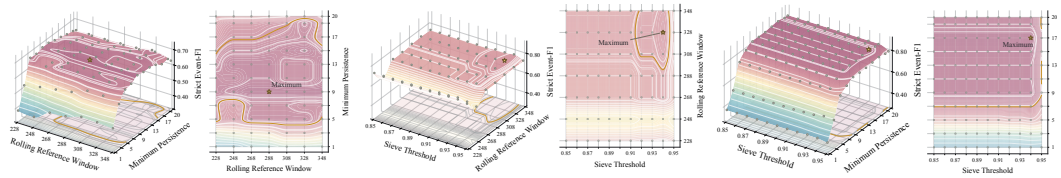


Figure 6: Sensitivity of strict event F1 under one-to-one IoU ≥ 0.5 matching. Left to right: reference window versus minimum persistence, sieve threshold versus reference window, and sieve threshold versus minimum persistence; other settings are fixed, with $w = m + 4$. White points are measurements, surfaces are interpolated, and stars mark representative, possibly tied maxima. Gold contours enclose scores within 10% of each grid's observed score range below its maximum.

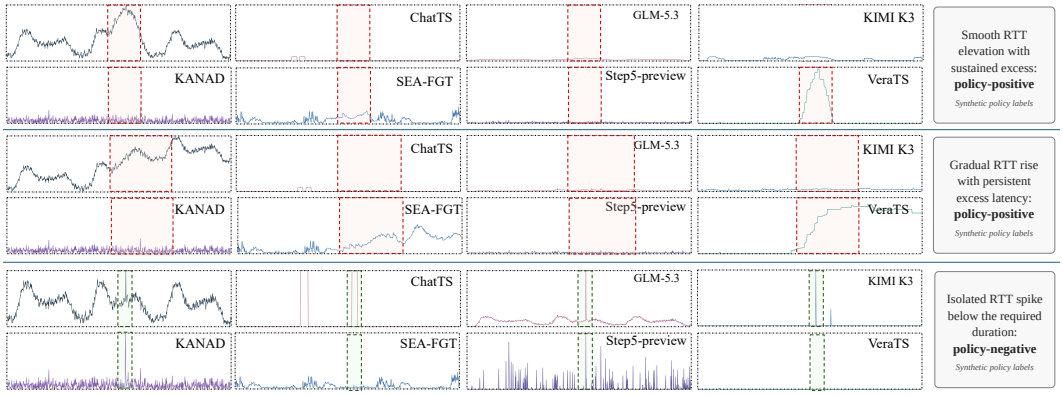


Figure 7: Responses to controlled RTT patterns. Rows show a smooth hump, a gradual rise, and an isolated spike over the first 864 evaluation observations on a shared periodic background. Methods differ in their inputs and output scales: positive robustness for VeraTS, interval masks for ChatTS, scaled native scores for SEA-FGT, and 0–1 ratings for LLM models; the figure compares response shapes.

4.3 Policy Behavior on Controlled RTT Traces

The controlled RTT cases in Figure 7 illustrate how a fixed high-latency rule responds to different temporal patterns. A smooth hump, a gradual rise, and an isolated spike share a synthetic periodic background. The first two are policy-positive, whereas the spike is policy-negative. Without retuning on RTT labels, VeraTS uses its rolling and seasonal references, threshold 2, five consecutive violations within nine observations, and closing width 48. Under the same fixed rule, VeraTS responds to the hump and gradual rise, while the isolated spike adds no native alerts relative to the control trace.

5 Conclusion

Operational TSAD requires both sensitivity to signal changes and an explicit criterion for deciding whether those changes warrant an alert. VeraTS connects these requirements through recall-constrained screening and executable semantic diagnosis: the first stage identifies candidate regions in numerical streams, and the second evaluates them against rules for harmful direction and persistence. With supplied policy programs on real backbone-link data, the complete pipeline ranks first in ten of the twelve category–metric comparisons under the reported protocols, supporting contextual diagnosis across the evaluated morphologies. Ablations and controlled RTT cases show how direction and persistence suppress benign deviations while preserving evidence of sustained violations. Together, these findings support a practical approach to operational detection in which broad numerical screening preserves candidates for contextual judgment, and reusable rules make the resulting alerts explicit and reproducible through deterministic execution.

Reproducibility Statement

The appendix and saved artifacts document splits, frozen settings, rule semantics, per-series results, sweeps, and replay checks.

Ethics Statement

The operational data studied in this paper were collected from backbone links in production systems and released to the research team by the Site Reliability Engineering (SRE) organization of the participating company. The release was approved through the company's internal data governance review, which examines the sensitivity of the requested records, the scope of access granted, and the conditions under which data may be shared with external collaborators. Before leaving the production environment, all data were de-identified: IP addresses, hostnames, account names, credentials, and free-text fields that may carry sensitive information were removed or replaced with irreversible surrogate keys. The released dataset therefore contains no customer information and no personal identifiers of any employee.

Morphology annotations were produced by 29 annotators who were hired and paid as research assistants for this project. Binary anomaly labels were determined by actual network failures, separately from the morphology annotations. Each annotator received an hourly rate of $500, which exceeds both the local minimum wage and the standard compensation for research assistants at the authors' institutions, and was paid for the full amount of time spent on the task, including training and adjudication meetings. Annotators were informed in advance of the nature of the task, the expected workload, and the intended research use of their labels, and participated voluntarily with the option to decline or withdraw at any time. The annotation interface exposed only sanitized operational metadata; no annotator was asked to inspect material that could cause distress, and no annotator was required to install or execute any untrusted code.

This work analyzes system-generated telemetry and does not involve human subjects research as defined by the authors' institutional review boards; accordingly, no IRB approval was required. Annotation outcomes are used exclusively in aggregate to evaluate the proposed method. They are never used to assess the performance, conduct, or employment status of any individual, and they are never fed back to the company as an employee-level signal. We further note that our study is defensive in purpose: we release only aggregate statistics, model implementations, and de-identified data, and we disclose no artifacts—such as exploit code, live attack tooling, or identifying infrastructure details—that could facilitate harm to the systems or operators studied.

AI Use Statement

Generative AI tools were used solely for language polishing of the manuscript text and for stylistic enhancement of the figures.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Équille, and Kalyan Veeramachaneni. Can large language models be anomaly detectors for time series? In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics*, pp. 1–10. IEEE, 2024. doi: 10.1109/DSAA61799.2024.10722786. URL https://doi.org/10.1109/DSAA61799.2024.10722786.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Paul Boniol, John Paparrizos, Themis Palpanas, and Michael J. Franklin. SAND: Streaming subsequence anomaly detection. *Proceedings of the VLDB Endowment*, 14(10):1717–1729, 2021. doi: 10.14778/3467861.3467863. URL https://doi.org/10.14778/3467861.3467863.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000. doi: 10.1145/342009.335388. URL https://doi.org/10.1145/342009.335388.

Mingyue Cheng, Jiahao Wang, Daoyu Wang, Xiaoyu Tao, Qi Liu, and Enhong Chen. Can slow-thinking LLMs reason over time? empirical studies in time series forecasting. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, 2026. doi: 10.1145/3773966.3777931. URL https://doi.org/10.1145/3773966.3777931.

Hang Cui, Jingjing Li, Haotian Si, Quan Zhou, Changhua Pei, Gaogang Xie, and Dan Pei. TShape: Rescuing machine learning models from complex shapelet anomalies. In *Proceedings of the 36th IEEE International Symposium on Software Reliability Engineering Workshops*, pp. 9–14, 2025. doi: 10.1109/ISSREW67781.2025.00038. URL https://arxiv.org/abs/2510.00680.

Hang Cui, Cenjie Hu, Zexin Wang, Jingjing Li, Juncheng Hu, Dan Pei, Changhua Pei, and Gaogang Xie. Predict boldly, recover cautiously: Fast on-router route anomaly prediction and recovery. In *Proceedings of the ACM SIGCOMM 2026 Conference*, pp. 2253–2255, 2026a.

Hang Cui, Zexin Wang, Jingjing Li, Juncheng Hu, Haotian Si, Cenjie Hu, Quan Zhou, Yongchang Hu, Lei Han, Dan Pei, Changhua Pei, and Gaogang Xie. Smart Brain: Semantic Anomaly Detection for Operational Time Series in Large Scale Service Systems—Author Code Artifact. GitHub repository, revision 7789171, 2026b. URL https://github.com/CSTCloudOps/Smart-Brain/tree/77891719a74e769cafee9a669550b9e06b13c179.

Hang Cui, Zexin Wang, Changhua Pei, Juncheng Hu, Haotian Si, Quan Zhou, Cenjie Hu, Jingjing Li, Dan Pei, and Gaogang Xie. Rethinking Time Series Anomaly Detection from a Dynamic Perspective: Temporal-Frequency-Curvature Fusion. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 710–720. Association for Computing Machinery, 2026c. doi: 10.1145/3770855.3817936. URL https://doi.org/10.1145/3770855.3817936.

Alexandre Donzé and Oded Maler. Robust satisfaction of temporal logic over real-valued signals. In *Formal Modeling and Analysis of Timed Systems*, volume 6246 of *Lecture Notes in Computer Science*, pp. 92–106. Springer, 2010. doi: 10.1007/978-3-642-15297-9_9. URL https://doi.org/10.1007/978-3-642-15297-9_9.

EasyTSAD contributors. EasyTSAD: Event-based point-adjustment evaluation. GitHub source code, 2026. URL https://github.com/dawnvince/EasyTSAD. Accessed September 12, 2026; source snapshot a135190.

Rob Ewaschuk. Monitoring distributed systems. In Betsy Beyer, Chris Jones, Jennifer Petoff, and Niall Richard Murphy (eds.), *Site Reliability Engineering: How Google Runs Production Systems*, chapter 6. O'Reilly Media, 2016. URL https://sre.google/sre-book/monitoring-distributed-systems/.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/de04896f011beff76c91e094f72727f4-Abstract-Conference.html.

Wei Huang, Zhihong Wang, Yanyong Huang, Jia Liu, and Xiaocao Ouyang. SEA-FGT: Frequency-Guided Transformer with Semantic Expert Augment for Time Series Anomaly Detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 1910–1921. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817823. URL https://doi.org/10.1145/3770855.3817823.

International Telecommunication Union. Framework and overall objectives of the future development of IMT for 2030 and beyond. Recommendation ITU-R M.2160-0, ITU Radiocommunication Sector, 2023. URL https://www.itu.int/rec/R-REC-M.2160.

Yue Jiang, Chenxi Liu, Yile Chen, Qin Chao, Shuai Liu, Cheng Long, and Gao Cong. STRATA-TS: Selective knowledge transfer for urban time series forecasting with retrieval-guided reasoning. *arXiv preprint arXiv:2508.18635*, 2025. URL https://arxiv.org/abs/2508.18635.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *The Eighth International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgNKkHtvB.

Zhi Li, Hong Ma, and Yongbing Mei. A unifying method for outlier and change detection from data streams based on local polynomial fitting. In *Advances in Knowledge Discovery and Data Mining*, volume 4426 of *Lecture Notes in Computer Science*, pp. 150–161. Springer, 2007. doi: 10.1007/978-3-540-71701-0_17. URL https://link.springer.com/chapter/10.1007/978-3-540-71701-0_17.

Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, Bin Yang, et al. Air quality prediction with physics-guided dual neural odes in open systems. In *International Conference on Learning Representations*, volume 2025, pp. 55132–55154, 2025.

Aoyu Liu and Yaying Zhang. A general spatio-temporal backbone with scalable contextual pattern bank for urban continual forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 72080–72103, 2026.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4623–4634. Association for Computing Machinery, 2025a. doi: 10.1145/3711896.3737239. URL https://doi.org/10.1145/3711896.3737239.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3437. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c3f3c690b7a99fba16d0efd35cb83b2c-Abstract-Datasets_and_Benchmarks_Track.html.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X. Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=0EXmFzUn5I.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=JePfAI8fah.

Yong Liu, Haoran Zhang, Chenyu Li, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer: Generative pre-trained transformers are large time series models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32369–32399, 2024b. URL https://proceedings.mlr.press/v235/liu24cb.html.

Yong Liu, Guo Qin, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Timer-XL: Long-Context Transformers for Unified Time Series Forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=KMCJXjlDDr.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39295–39317, 2025c. URL https://proceedings.mlr.press/v267/liu25be.html.

Donghao Luo and Xue Wang. ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vpJMJerXHU.

Yingjie Ma, Bijal Bharadva, Xin Zhang, and Joann Qiongna Chen. Pategail++: Utility optimized private trajectory generation with imitation learning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Oded Maler and Dejan Nickovic. Monitoring temporal properties of continuous signals. In *Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems*, volume 3253 of *Lecture Notes in Computer Science*, pp. 152–166. Springer, 2004. doi: 10.1007/978-3-540-30206-3_12. URL https://doi.org/10.1007/978-3-540-30206-3_12.

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, and Puneet Agarwal. Long short term memory networks for anomaly detection in time series. In *Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 89–94, 2015. URL https://www.esann.org/sites/default/files/proceedings/legacy/es2015-56.pdf.

Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv:1607.00148*, 2016. URL https://arxiv.org/abs/1607.00148.

Andrew Ng. Sparse autoencoder. CS294A lecture notes, Stanford University, 2011. URL https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3009–3017, 2019. doi: 10.1145/3292500.3330680.

Peter J. Rousseeuw and Annick M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, 2003. ISBN 978-0-471-48855-2. URL https://www.wiley-vch.de/en/areas-interest/mathematics-statistics/robust-regression-and-outlier-detection-978-0-471-48855-2.

Ya Shen, Gang Chen, Hui Ma, and Mengjie Zhang. Cost-aware dynamic cloud workflow scheduling using self-attention and evolutionary reinforcement learning. In *International Conference on Service-Oriented Computing*, pp. 3–18. Springer, 2024.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=e1wDDFmlVu.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022. doi: 10.14778/3514061.3514067. URL https://www.vldb.org/pvldb/vol15/p1201-tuli.pdf.

Hao Wang, Licheng Pan, Zhichao Chen, Xu Chen, Qingyang Dai, Lei Wang, Haoxuan Li, and Zhouchen Lin. Time-o1: Time-series forecasting needs transformed label alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/0cd62dea69635f4c5b569848267fe5a8-Abstract-Conference.html.

Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. MICN: Multi-scale local and global context modeling for long-term series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zt53IDUR1U.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=7oLshfEIC2.

Xin Wang, Yunshi Wen, Yanan He, Haotian Xu, Youlan Zhao, Michel Ferreira Cardia Haddad, and Tengfei Ma. CAAD: Causality-Aware Multivariate Time Series Anomaly Detection via Multi-Scale Alignment and Structural Causal Consistency. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5091–5101. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817978. URL https://doi.org/10.1145/3770855.3817978.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering Transformers for Time Series Forecasting with Exogenous Variables. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://papers.nips.cc/paper/2024/hash/0113ef4642264adc2e6924a3cbbdf532-Abstract-Conference.html.

Zexin Wang, Changhua Pei, Minghua Ma, Xin Wang, Zhihan Li, Dan Pei, Saravan Rajmohan, Dongmei Zhang, Qingwei Lin, Haiming Zhang, Jianhui Li, and Gaogang Xie. Revisiting VAE for unsupervised time series anomaly detection: A frequency perspective. In *Proceedings of the ACM Web Conference 2024*, pp. 3096–3105, 2024c. doi: 10.1145/3589334.3645710. URL https://doi.org/10.1145/3589334.3645710.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ju_Uqw384Oq.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.12261.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. ChatTS: Aligning time series with LLMs via synthetic data for enhanced understanding and reasoning. *Proceedings of the VLDB Endowment*, 18(8):2385–2398, 2025. doi: 10.14778/3742728.3742735. URL https://doi.org/10.14778/3742728.3742735.

Haowen Xu, Wenxiao Chen, Nengwen Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Ying Liu, Youjian Zhao, Dan Pei, Yang Feng, Jie Chen, Zhaogang Wang, and Honglin Qiao. Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In *Proceedings of the 2018 World Wide Web Conference*, pp. 187–196, 2018. doi: 10.1145/3178876.3185996. URL https://doi.org/10.1145/3178876.3185996.

Xiongxiao Xu, Haoran Wang, Yueqing Liang, Philip S. Yu, Yue Zhao, and Kai Shu. Can multimodal LLMs perform time series anomaly detection? In *Proceedings of the ACM Web Conference 2026*, pp. 5392–5403. Association for Computing Machinery, 2026. doi: 10.1145/3774904.3792376. URL https://doi.org/10.1145/3774904.3792376.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with 10k parameters. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bWcnvZ3qMb.

Yiyuan Yang, Zichuan Liu, Lei Song, Kai Ying, Stephen Wang, Joshua Thomas Bamford, Svitlana Vyetrenko, Jiang Bian, and Qingsong Wen. Time-RA: Towards time series reasoning for anomaly diagnosis with LLM feedback. In *Findings of the Association for Computational Linguistics: ACL*

2026, pp. 11591–11616. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.562. URL https://aclanthology.org/2026.findings-acl.562/.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317. URL https://ojs.aaai.org/index.php/AAAI/article/view/26317.

Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. TFAD: A decomposition time series anomaly detection architecture with time-frequency analysis. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management*, pp. 2497–2507, 2022a. doi: 10.1145/3511808.3557470. URL https://doi.org/10.1145/3511808.3557470.

Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented MLP structures. *arXiv preprint arXiv:2207.01186*, 2022b. URL https://arxiv.org/abs/2207.01186.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vSVLM2j9eie.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021. doi: 10.1609/aaai.v35i12.17325. URL https://ojs.aaai.org/index.php/AAAI/article/view/17325.

Quan Zhou, Changhua Pei, Fei Sun, Han Jing, Zhengwei Gao, Haiming Zhang, Gaogang Xie, Dan Pei, and Jianhui Li. KAN-AD: Time series anomaly detection with kolmogorov–arnold networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79136–79149, 2025. URL https://proceedings.mlr.press/v267/zhou25u.html.

Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. Webarena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, volume 2024, pp. 15585–15606, 2024.

Tian Zhou, Ziqing Ma, Xue Wang, Qingsong Wen, Liang Sun, Tao Yao, Wotao Yin, and Rong Jin. FiLM: Frequency improved legendre memory model for long-term time series forecasting. In *Advances in Neural Information Processing Systems*, volume 35, 2022a. URL https://openreview.net/forum?id=zTQdHSQUQWc.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286, 2022b. URL https://proceedings.mlr.press/v162/zhou22g.html.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. One fits all: Power general time series analysis by pretrained LM. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2302.11939.

Zihao Zhou and Rose Yu. Can LLMs understand time series anomalies? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/05774fb74e863308c4b460c9f49f6918-Abstract-Conference.html.

Yan Zhu. *The Matrix Profile: Scalable Algorithms and New Primitives for Time Series Data Mining*. PhD thesis, University of California, Riverside, 2018. URL https://escholarship.org/uc/item/2x4419fp.

A Reader's Guide and Evidence Contract

This appendix separates the new initialized experiment from archival evidence. Sections B and K document the new run. The theoretical results state conditional properties rather than empirical guarantees. The executable-semantics section is retained from the audited implementation specification. Historical compiler and baseline artifacts remain historical even when they have been re-rendered. The dataset atlas in Section E describes all four morphology folders and shows both splits of every supplied series. Section N adds measurement-corruption experiments, parameter sensitivity, held-out percentile transfer, score-resolution diagnostics and a complete comparative score atlas.

Question	Current evidence	Missing evidence
Does context matter?	Opposite-policy construction; fixed-rule direction controls	Independently verified context-conditioned incidents
Does broad screening retain events?	Frozen development constraint; event support; measurement-corruption stress tests	Lineage-disjoint prospective recall; independently observed drift incidents
How do semantic components affect diagnosis?	Selected matched-rule component ablations	A matched-budget advantage over independently tuned alternatives
How much input is retained?	Observation and 96-point-tile accounting	End-to-end LLM token and service-cost measurements
What is the inference cost?	CPU time and peak memory in Figure 4	Hardware scaling and deployment measurements
Does an LLM help?	Historical compiler audit; template parser control	Fresh model runs on independently authored, sealed policies
Can outputs be replayed?	567 numerical equivalence checks in one environment	Independent implementations and environment-bound release manifests

Table 2: Evidence ledger for the hierarchy. The right column is not a completed experiment.

B New Initialization-Corrected Experimental Protocol

B.1 Split boundaries and calibration

Every supplied series has 4,032 training and 2,016 evaluation observations. With zero-based half-open indexing, training values $[0, 2304)$ initialize the period-288, seven-lookback feature path; $[2304, 2880)$ supply at least 576 initialized percentile-reference values; $[2880, 4032)$ provide 1,152 development observations. This development suffix contains 72 labeled events across 63 series. Labels are used for constrained operating-point selection and rule selection, so this is a supervised-development experiment, despite the unsupervised form of the deviation features.

During evaluation, the full 4,032-point training history is available and the initialized CDF reference extends over $[2304, 4032)$, or 1,728 points. This is an unlabeled historical-reference update, not a test-label update. Each feature reference discards values above its 99th percentile only if at least 32 reference values remain. The CDF uses right-side rank insertion and the union takes the maximum percentile over absolute features. A threshold of 1.0 consequently retains values at or above the upper end of a trimmed empirical reference; it does not mean an empty proposal set.

The evaluated candidate morphology differs explicitly from the archive: raw threshold triggers are OR-ed with width-24 binary closing, then dilated by radius two. The OR preserves raw finite-boundary triggers. The downstream verifier retains its existing false-padded numerical closing semantics, which can remove boundary positives. These distinct operator choices are bound in the new script and protocol.

B.2 Selection grid and objective

The proposal grid contains $-1, 0.90, 0.95, 0.975, 0.985, 0.99, 0.995, 0.997, 0.999, 1.0$, and 1.000001. The -1 threshold routes all points and supplies a feasible fallback; 1.000001 exceeds all percentile

Method	Event P	Event R	Event F1	Point F1
Signal sieve only	0.037	0.561	0.070	0.147
Recall-constrained hierarchy	0.573	0.833	0.679	0.593
Without semantic direction	0.182	0.803	0.296	0.343
Reversed direction	0.000	0.000	0.000	0.002
Without persistence	0.323	0.818	0.464	0.552
Without output closing	0.158	0.712	0.259	0.573
Rolling reference only	0.662	0.803	0.726	0.562
Seasonal reference only	0.604	0.439	0.509	0.453

Table 3: Selected semantic-component comparisons under the initialization-corrected reference AST, with a first-stage-only diagnostic. All event columns use strict one-to-one IoU ≥ 0.5 matching. Component controls retain the remaining full-method settings. This table does not enumerate all archived variants. No row is a fresh LLM result.

scores and is an empty-proposal diagnostic. Among thresholds reaching literal overlap recall 1 on the development events, the primary selector minimizes retained point fraction. The selected threshold is 1.0, with 72/72 events overlapped, 63.248% point coverage, 72.278% anomalous-point support, and 545 candidate intervals.

For a fixed selected mask, numerical rule selection searches six thresholds $\tau \in \{1.5, 2, 2.5, 3, 4, 5\}$, two persistence contracts $(m, w) \in \{(3, 6), (5, 9)\}$, and three output-closing widths $\{5, 24, 48\}$. This is 36 rule settings. The selector maximizes development strict event F1, then point F1 and event recall. The selected rule uses $\tau = 2, m = 5, w = 9, g = 48$ and a dual reference. Its development strict F1 is 0.826.

B.3 Selected semantic-component comparisons

The following tables connect each semantic component to its role in selective diagnosis at the fixed operating point. Direction-free replaces polarity with an absolute deviation; reversed direction swaps increase and decrease. Rolling-only and seasonal-only retain one reference with the same threshold and temporal settings. Removing persistence sets $(m, w) = (1, 1)$, and removing output closing disables the closing wrapper. Together with the first-stage-only diagnostic and full method, these matched controls reveal how policy meaning, sustained evidence, reference context and interval continuity contribute to the recorded outputs.

B.4 Label access and artifact binding

The new script reads raw value arrays, computes training features, selects settings, creates all evaluation predictions, and seals scores and selection hashes before opening evaluation label files. The hash-bound output snapshot contains the protocol, selection, all rule variants, training grid, evaluation threshold sweep, scores and masks, per-series tables, input and output manifests, public-API equivalence checks, and timing. Historical detector outputs are not inputs to the new experiment. This order prevents within-run test-label selection; it cannot erase prior human or model exposure to the same retrospective corpus.

The archived numerical replay covers nine monitored variants on 63 series, giving 567 exact comparisons of cached computation against the guarded public `RuleMonitor`. All checks passed; an additional 63 selected-development-rule checks also passed. A separate artifact-verification script rechecked 1,260 stored arrays and recomputed 120 aggregate endpoint values using the same metric implementation. This exact agreement demonstrates a practical benefit of deterministic policy execution: cached analyses, public monitor outputs and reported metrics can be checked against the same numerical record.

B.5 What the measured endpoints say

The primary gate preserves overlap with all 66 evaluation events while retaining 32.838% of all points and 66.094% of anomalous points. This concentrated support realizes the broad-screening objective.

Contextual diagnosis preserves all 66 event hits in the final output and yields 55 one-to-one IoU-0.5 matches, linking the hierarchy to both event discovery and interval localization. The tables report each endpoint separately so that the corresponding design benefit can be assessed directly.

The sieve emits 987 candidate intervals, and the complete monitor consolidates the qualified evidence into 96 alert intervals. These counts describe successive segmentations as persistence and closing act on the evidence. Final point recall is 0.672; output closing adds 433 positive points outside the gate. This explicit support accounting illustrates how the hierarchy combines concentrated candidate admission with reconstruction of coherent event intervals.

B.6 Additional endpoints for the selected semantic components

Variant	Point P	Point R	AUPRC	Alerts	Matches
Signal sieve only	0.083	0.661	0.088	987	37
Recall-constrained hierarchy	0.530	0.672	0.515	96	55
Without semantic direction	0.230	0.678	0.177	292	53
Reversed direction	0.002	0.003	0.048	199	0
Rolling reference only	0.653	0.493	0.416	80	53
Seasonal reference only	0.509	0.409	0.399	48	29
Without persistence	0.444	0.731	0.507	167	54
Without output closing	0.555	0.593	0.438	297	47

AUPRC here follows the existing pooled micro evaluator on concatenated observations. Its raw numerical robustness can have different units and scales across metrics, so this pooled ranking is a descriptive diagnostic rather than a calibrated cross-metric ranking. Per-series ranking values remain in the CSV artifact; the main comparison instead computes VUS per series before averaging within each family.

B.7 Candidate workload and measured inference cost

The observation-level reduction is 67.162%. A packet proxy partitions each evaluation series into 21 non-overlapping 96-point tiles and routes a tile if it contains any candidate. Across 63 series, 1,046 of 1,323 tiles are routed: 79.063% retained, or 20.937% removed. These counts describe the input admitted for diagnosis; context padding, packaging, and the execution strategy determine the resulting computational cost.

Figure 4 supplies the complementary execution evidence: 62.2 milliseconds per 1,000 observations and 149.4 MiB peak RSS. Among the twelve measured implementations, VeraTS combines the highest category-balanced VUS-PR with the lowest peak memory. Candidate retention quantifies the concentration of admitted evidence, while the CPU benchmark demonstrates the practicality of deterministic numerical inference.

C What a Hierarchy Can Guarantee

The purpose of a hierarchy is to allocate different information and computation to different decisions. Its first stage should preserve potentially actionable events; its second stage should distinguish policy violations from harmless deviations. These objectives do not, by themselves, imply zero missed events, an advantage from using a language model, or a reduction in measured runtime. This appendix states the conditions under which each conclusion would follow.

C.1 Implemented hierarchy and proposed extensions

In the released implementation, an LLM translates a policy and metric schema into an accepted typed rule or rejects the request *offline*. The runtime computes a direction-free candidate mask and executes an accepted numerical rule deterministically. It does not call an LLM for each candidate. In particular, statistical evidence is not forwarded to an online language agent for an open-ended diagnosis. Reusing a compiled rule is the implemented way to amortize semantic interpretation across observations.

The revised numerical experiment selects a threshold by minimizing development coverage subject to zero observed development-event overlap misses. Its proposal postprocessor unions raw triggers with closing before dilation, so finite-array closing cannot remove a raw boundary trigger. This changes the archived selection and proposal operator; it is an empirical development constraint, not a population-risk guarantee.

Three extensions discussed below remain research designs: selecting a Sieve threshold under a separately calibrated population event-miss constraint; executing expensive expert analysis only on candidate neighborhoods; and auditing discarded episodes with a predefined fallback. Their mathematical analyses specify testable interfaces, not measurements of released functionality.

We also distinguish static metric context from changing operational context. The current grammar supports explicit direction, units, declared bounds, reference choice, robust threshold, and temporal persistence. It does not interpret deployment logs, maintenance schedules, cross-metric dependencies, or undocumented units. A rule expressing such a condition must be rejected unless the grammar and its evaluator have first been extended. Declared input bounds are runtime validity checks, not a general absolute-threshold anomaly language.

C.2 Separate the statistical and operational targets

Write $Z = (H, X)$ for available numerical observations and C for a policy and its metric schema. An operational target is $Y = f(Z, C)$. Its definition must be fixed independently of the algorithm whose accuracy is measured. A policy-conformance target is the result of an independently adjudicated formal policy; an incident target additionally requires independent incident evidence. Neither an LLM rationale nor agreement with an author-selected rule establishes incident truth.

An unusual shape is useful routing evidence, but is not identical to either target. A rise in a success metric and the same rise in a failure metric can have different meanings. Conversely, a policy violation need not be a large relative deviation: a service may remain persistently outside an absolute operating limit. Thus, the set of admissible incidents must be specified before asserting that a direction-free deviation detector covers it.

C.3 Exact miss decomposition for a hard routing interface

Fix an evaluation unit before observing predictions. A unit can be a point or an independently defined incident episode. Let $Y \in \{0, 1\}$ denote its true status, $R \in \{0, 1\}$ indicate admission to a downstream decision, and $D \in \{0, 1\}$ be the downstream alert decision when $R = 1$. The hard routing contract is $\widehat{Y} = RD$: a discarded unit cannot be reopened. The downstream decision may use a compiled rule or another expert system; no independence between R and D is assumed.

Proposition 1 (Irrecoverable routing loss)**.** *Suppose* $\Pr(Y = 1) > 0$*. Define*

$$\alpha = \Pr(R = 0 \mid Y = 1), \qquad \beta = \Pr(D = 0 \mid R = 1, Y = 1).$$

When $\Pr(R = 1, Y = 1) = 0$*, assign any value to* β*, whose coefficient below is zero. Then*

$$\Pr(\widehat{Y} = 0 \mid Y = 1) = \alpha + (1 - \alpha)\beta \geq \alpha. \tag{11}$$

If additionally $\Pr(Y = 0) > 0$*, define* $\gamma = \Pr(R = 1 \mid Y = 0)$ *and* $\eta = \Pr(D = 1 \mid R = 1, Y = 0)$*. When* $\Pr(R = 1, Y = 0) = 0$*, assign any value to* η*, whose coefficient below is zero. Then*

$$\Pr(\widehat{Y} = 1 \mid Y = 0) = \gamma\eta. \tag{12}$$

Proof. Conditional on $Y = 1$, the two disjoint miss events are $\{R = 0\}$ and $\{R = 1, D = 0\}$. The law of total probability gives Eq. (11). The false-alert event is exactly $\{R = 1, D = 1\}$ conditional on $Y = 0$, giving Eq. (12). □

The second stage can suppress harmless proposals, but under this interface it cannot recover a true incident discarded by the first stage. Its recall objective is therefore to avoid adding missed incidents while improving precision. For a total miss budget ϵ and a routing miss rate $\alpha < 1$, it must satisfy

$$\beta \leq \frac{\epsilon - \alpha}{1 - \alpha}, \qquad \alpha \leq \epsilon. \tag{13}$$

Improving language reasoning cannot compensate for $\alpha > \epsilon$ without changing the routing contract.

Abstention is an operational outcome. If the compiler or expert can abstain, its possible outputs are $\{\text{alert}, \text{normal}, \text{unresolved}\}$. An unresolved case must not be silently interpreted as normal. If the evaluator defines failure to emit an automatic alert as a miss, let

$$u = \Pr(D = \text{unresolved} \mid R = 1, Y = 1), \quad b = \Pr(D = \text{normal} \mid R = 1, Y = 1).$$

Then the automatic-alert miss probability is $\alpha + (1 - \alpha)(u + b)$. A human escalation route changes this outcome only when its response and delay are measured. Syntactic fail-closed rejection prevents execution of an invalid program; it does not, by itself, provide a safe monitoring outcome for the underlying service.

C.4 The actual runtime: gating followed by closing

The hard-routing abstraction requires care when applied to this code. For a fixed accepted AST, let $P_t = \mathbb{1}\{\rho_t > 0\}$ be the positive persistence core before candidate gating. Let Q be the candidate mask and $\mathcal{C}_d$ the Boolean closing used by the runtime, including its fixed anchoring and false boundary padding. Its predictions are

$$\widehat{Y}_Q = \mathcal{C}_d(P \wedge Q), \qquad \widehat{Y}_{\mathbf{1}} = \mathcal{C}_d(P). \tag{14}$$

The first equality follows from sign equivalence of the quantitative and Boolean implementations. Closing may fill a gap at a location with $Q_t = 0$. Consequently, $\widehat{Y}_t = Q_t D_t$ is not a correct formula for the complete postprocessed output. Equation (11) applies directly to the pre-closing point core, or to an episode routing abstraction that explicitly includes every allowed reopening and neighborhood expansion.

Proposition 2 (Gating cannot enlarge the final positive set)**.** *Fix the history, trace, accepted AST, and morphology parameters. If* $Q \leq Q'$ *pointwise, then*

$$\widehat{Y}_Q \leq \widehat{Y}_{Q'} \leq \widehat{Y}_{\mathbf{1}}.$$

If $P \leq Q$*, then* $\widehat{Y}_Q = \widehat{Y}_{\mathbf{1}}$*.*

Proof. $P \wedge Q \leq P \wedge Q' \leq P$. Boolean dilation and erosion with a fixed footprint and fixed padding are increasing functions of their inputs, so their composition preserves both inequalities. If $P \leq Q$, then $P \wedge Q = P$. □

For fixed truth, raw point recall and literal-overlap event recall therefore cannot increase by adding this gate to the same verifier. Strict one-to-one interval IoU recall and F1 need not be monotone: removing points can change interval boundaries or separate merged predictions. The proposition

also says nothing about a refitted rule or a reselected threshold. These distinctions separate candidate admission from final interval localization.

Candidate overlap with a true event is weaker than preservation of useful decision support. The candidate may intersect the event only where $P_t = 0$; the positive persistence endpoint may lie after an event; and finite-array closing may remove a boundary candidate. Evaluation should therefore report both raw candidate-event overlap and whether routing preserves the relevant ungated decision. Morphology must be included in the evaluated routing operator, rather than treating a threshold crossing as guaranteed delivery.

C.5 What context adds, and what it does not prove about LLMs

For fixed z, suppose two admissible contexts c_0, c_1 assign labels $y^{(0)}, y^{(1)} \in \{0,1\}^T$ to the same numerical observations. Let $D_z = \{t : y_t^{(0)} \neq y_t^{(1)}\}$. Every context-free predictor $g(z)$ obeys

$$\|g(z) - y^{(0)}\|_0 + \|g(z) - y^{(1)}\|_0 \geq |D_z|. \tag{15}$$

At each differing coordinate, one output bit cannot equal both targets. If the contexts occur with probabilities π and $1-\pi$, the conditional expected Hamming error is at least $\min(\pi, 1-\pi)|D_z|$, including for a randomized predictor after averaging over its randomness.

More generally, for a binary target define $p(z) = \Pr(Y = 1 \mid Z = z)$. A context-free binary decision has conditional error at least $\min\{p(z), 1-p(z)\}$. This is an information limitation, not an architectural lower bound for a particular signal model. A handcrafted policy engine, structured metadata, a deterministic parser, or a language model can all supply missing context. Demonstrating an LLM-specific advantage requires an independent language task and equal-information non-LLM controls. Opposite-polarity templates alone can demonstrate the need for the polarity input, but cannot establish the need for an LLM.

C.6 Why unconditional zero-miss selective screening is impossible

Proposition 3 (A selective gate needs a coverage assumption)**.** *Let a hard gate $R(z)$ discard some admissible numerical input z_0. If an admissible context can assign $Y = 1$ to z_0, there exists a distribution on admissible inputs for which the gate's conditional miss rate is one.*

Proof. Put probability one on $(z_0, c_0, Y = 1)$, where c_0 is such a context. The gate discards every positive example, so $\Pr(R = 0 \mid Y = 1) = 1$. □

To obtain a deterministic zero-miss statement, one must restrict the anomaly class so that $Y = 1 \Rightarrow R = 1$, or admit every unit. For instance, assume every admissible event has a score witness above a fixed threshold and that the complete routing postprocessor preserves such a witness. Zero literal-overlap routing misses then follows directly from that assumption. It remains a conditional completeness result, not evidence that all operational incidents satisfy the witness condition.

For the implemented postprocessed verifier, a sufficient preservation condition is $P \leq Q$ from Proposition 2. If the ungated verifier is complete for a stated truth class and this condition holds, the gated verifier retains that completeness. Neither premise is established for arbitrary production incidents by the present case study. The all-ones mask removes gate-induced loss but does not eliminate compiler, policy, numerical-rule, or label errors.

C.7 Zero observed misses is a finite-sample statement

Suppose a frozen gate is evaluated on n independent positive units drawn from a fixed target distribution, with common miss probability p. If no misses are observed, the exact one-sided binomial upper confidence limit at confidence $1-\delta$ is

$$p_U = 1 - \delta^{1/n}. \tag{16}$$

Indeed, the probability of observing zero misses is $(1-p)^n$; solving $(1-p_U)^n = \delta$ gives the endpoint. Thus zero observed misses can support $p \leq \epsilon$ at this confidence only when

$$n \geq \left\lceil \frac{\log \delta}{\log(1-\epsilon)} \right\rceil. \tag{17}$$

For illustration, a one-sided 95% statement with zero observed misses needs 299 independent positive units to bound p by 1%, and 2,995 to bound it by 0.1%. These are prospective sample-size calculations, not evidence from the current corpus.

Dependence changes the effective unit of this calculation. Events from a duplicated trace, common incident, shared generator, or repeated template cannot simply be counted as independent trials. A cluster-level experiment can instead define one binary failure per independent source episode, but that bounds episode-level failure, not automatically pooled point or event risk. The current 63-series corpus has unresolved dependence and does not justify substituting its raw event count into Eq. (16). A test set used to select the gate is also not an independent fixed-gate audit. Repeated threshold searches require a separate audit or a valid simultaneous procedure.

C.8 Proposed event calibration with the complete routing operator

This subsection specifies a possible future calibration method. It is not the protocol used to select the archived operating point. Fit every feature, reference distribution, morphology parameter, and threshold grid without the calibration labels. Suppose the resulting family of complete routing masks $\{Q_\lambda\}$ is nested: increasing λ can only remove proposals. For each positive evaluation unit E, define a survival score $M(E)$ so that

$$E \text{ is delivered by } Q_\lambda \quad \Longleftrightarrow \quad M(E) \geq \lambda.$$

On a finite predeclared threshold grid this can be constructed from the highest threshold that still delivers E. Include an explicit admit-all sentinel below the grid, so the lowest operating point delivers every unit. The definition includes closing, padding, neighborhood expansion, and the chosen delivery criterion. A raw maximum anomaly score is a substitute only if its equivalence to that delivery criterion has been proved.

Proposition 4 (Split rank calibration for marginal delivery)**.** *Conditional on all fitted components, let the n calibration survival scores and one fresh positive-unit survival score be exchangeable. For a desired miss level $\epsilon \in [0, 1)$, let $k = \lfloor (n+1)\epsilon \rfloor$. If $k = 0$, use the admit-all sentinel. Otherwise choose $\widehat{\lambda} = M_{(k)}$, the kth smallest calibration survival score. Then*

$$\Pr\{M_{\text{new}} < \widehat{\lambda}\} \leq \frac{k}{n+1} \leq \epsilon.$$

Proof. For $k = 0$ every unit is admitted by construction. Otherwise, break any tied scores using independent continuous random ranks. Exchangeability makes the new score's rank uniform on $\{1, \ldots, n+1\}$. Its being strictly below the kth calibration score implies a rank among the first k, which has probability $k/(n+1)$. Ignoring the auxiliary tie breaking and retaining an inclusive threshold can only reduce the miss probability. □

This is marginal coverage over calibration and a new positive unit. It is not a high-probability bound on the conditional risk of every realized gate, groupwise coverage, a guarantee under arbitrary drift, or guaranteed delivery of every future incident. At $\epsilon = 0$ the stated distribution-free rule uses the admit-all sentinel. Group-disjoint splits and representative positive-unit sampling are required before the exchangeability assumption is plausible. Sampling more events from one episode does not repair a shortage of independent episodes.

The implementation contains a helper based on an empirical quantile of raw training-event score maxima, but the archived main experiment selected Sieve and verifier parameters jointly by training strict event F1. The revised initialized experiment instead minimizes development candidate coverage subject to complete development-event overlap, and then selects verifier parameters. Both are empirical development procedures. That helper is not the calibration protocol above: raw score maxima do not include the full routing morphology, and a generic empirical quantile does not automatically implement the finite-sample rank correction. Neither code existence nor a high retrospective recall licenses a conformal-coverage claim.

C.9 Cost requires work to be avoided, not just masks to be sparse

Consider a proposed system with N fixed windows, a cheap screening cost c_s per window, downstream expert cost c_e per admitted window, and admitted fraction κ. Ignoring explicitly separated

fixed costs, its cost is

$$C_{\text{route}} = Nc_s + N\kappa c_e + C_{\text{overhead}}, \qquad C_{\text{dense}} = Nc_e. \tag{18}$$

It is cheaper precisely when $Nc_s + C_{\text{overhead}} < N(1-\kappa)c_e$ under this cost model. With variable expert inputs, the correct term is $\sum_{i:R_i=1} c_e(i)$, which includes context padding, prompts, retrieval, retries, and serialization. Point coverage alone does not determine number of requests or total tokens.

The current system has a different budget. If K policies are compiled at costs $c_{\text{compile}}(j)$ and then reused, its measured work should be written as

$$C_{\text{current}} = \sum_{j=1}^{K} c_{\text{compile}}(j) + C_{\text{features}}(T) + C_{\text{sieve}}(T) + C_{\text{verifier}}(T) + C_{\text{records}}. \tag{19}$$

The feature and verifier paths currently process the full trace, so candidate coverage measures retained support rather than a multiplicative reduction in those runtime terms. The amortized compilation cost per observation is $\sum_j c_{\text{compile}}(j)/T$ for the stated reuse pattern; this is a consequence of policy reuse, separately from candidate filtering.

A candidate-only verifier must additionally recover the same reference state and temporal support as dense execution. Restarting a rolling or seasonal reference at a candidate boundary changes the algorithm. An efficiency claim requires a sparse implementation, full-support accounting, and equality or explicit approximation tests against the dense reference. An online claim also requires a causal candidate operator and measured alert delay; symmetric closing and dilation do not become causal because their windows are bounded.

C.10 Proposed audits and fallback behavior

Periodic auditing can diagnose screening blind spots, but is not implemented or evaluated in the current pipeline. A prospective protocol should freeze an audit unit, sampling rule, independent adjudicator, review delay, trigger, and fallback before monitoring starts. Samples must come from discarded units; reviewing accepted alerts cannot estimate the loss outside the gate.

For a fixed epoch with N discarded units of which K are truly positive, a uniform sample of n distinct discarded units finds no positives with probability

$$\frac{\binom{N-K}{n}}{\binom{N}{n}},$$

where the numerator is zero when $n > N - K$. This finite-population expression can be inverted for an upper confidence bound on K. Its interpretation requires the specified sampling and truthful adjudication; it does not apply to convenience samples of visually conspicuous discards.

A predefined failure trigger could switch to $Q \equiv 1$ or an independently validated escalation policy. Such fallback removes subsequent gate-induced loss during the fallback interval. It does not retroactively prevent a missed alert, certify the expert, or prove that the trigger detects every distribution shift. A future deployment study must report audit cost and intervention delay alongside any coverage benefit.

C.11 What would establish the necessity of a module?

The context collision proves that some tasks need context. It does not prove that a particular semantic model is required. Gate monotonicity proves an inclusion relation. It does not prove a gate improves a detector. A module is empirically useful only for a declared objective under a matched comparison. For a proposed constrained system, that objective can be

$$\min \ \mathbb{E}[\text{measured total cost}] \quad \text{subject to} \quad R_{\text{miss}} \le \epsilon, \quad R_{\text{false alert}} \le \eta,$$

with the risk definition, confidence procedure, and operating-point selection frozen before testing. Comparisons in such a deployment study should hold inputs and budgets fixed and report both improvements and regressions. The applicable conditions of each component must be established empirically.

Object / field	Domain	Additional validation
Rule	`schema_version`={`1.0`}; nonempty `rule_id`, `source_text`, and `rationale`	Exactly one metric, direction, baseline, threshold, persistence block, and Boolean `candidate_required`; no extra fields.
Metric	nonempty name; unit $\in$ {percent, count, milliseconds, unknown}; explicit real-or-null lower/upper bound keys; integer $1 \leq$ `sample_period_points`$\leq$ 4096	If both bounds exist, lower $<$ upper. Percent requires exact $[0, 100]$. Runtime arrays must be nonempty, finite, one-dimensional, and within supplied bounds (tolerance 10^{-9}).
Direction	{increase, decrease, either}	No direction is inferred from the metric name by the verifier.
Baseline	kind $\in$ {rolling_median, seasonal_median, dual_reference}; $2 \leq P \leq 4096$; $1 \leq L \leq 28$; $4 \leq h \leq 4096$	Rolling rules require $n \geq h$ history points. Seasonal and dual rules require $n \geq \max\{h, P(L+1)\}$; see below.
Threshold	kind={`robust_z`}; $0 < \tau \leq 20$	No absolute-threshold or implicit unit conversion is supported.
Persistence	integers $1 \leq m, w \leq 4096$, $0 \leq d \leq 4096$	$m \leq w$. Width $d \in \{0, 1\}$ means identity closing.
Candidate gate	Boolean `candidate_required`	If true, a Boolean mask with exactly the test-trace shape is mandatory.

Table 5: Exact schema accepted by the TYPED INTENT COMPILER boundary. Braces denote a closed enumeration; brackets in the last column denote validation bounds.

The rank argument is a standard split-calibration construction related to Conformal Risk Control; reject-option risk–coverage analysis is established in Selective Classification for Deep Neural Networks. LLM-generated reusable anomaly rules are already studied by Argos. The proposed contribution must therefore lie in a precisely evaluated composition of context, routing risk, and computational budget, not priority over these constituent ideas.

D TYPED RULE SCHEMA AND EXACT EXECUTABLE SEMANTICS

D.1 CLOSED SCHEMA

Table 5 is the complete accepted rule schema. Every object forbids unknown fields. Thus "typed" means closed enums, field bounds, metric-range validation, and executable-history checks; it does not mean a general unit algebra or automatic time conversion.

The metric sample period is stored as a count of observations, not a physical duration. A request such as "for five minutes" is rejected when no mapping from minutes to observations is supplied. Percent values require explicit bounds; count and latency policies in the local study set a nonnegative lower bound. The archived author-set reference operating point uses $(P, L, h, \tau, m, w, d) = (288, 7, 288, 2, 5, 9, 48)$ and the dual reference.

D.2 REFERENCE AND ATOMIC ROBUSTNESS

Let $v = (H, x)$ be the concatenation of $n = |H|$ history observations and the test trace, indexed from zero, and let $k = n + t$ denote test time t. All test references use values with indices strictly below k. For rolling window h,

$$b_t^{\text{roll}} = \text{median}(v_{k-h:k}), \tag{20}$$

$$s_t^{\text{roll}} = \max\{(Q_{0.75} - Q_{0.25})/1.349, 10^{-8}\}, \tag{21}$$

where both quantiles are taken over $v_{k-h:k}$ using linear interpolation. For period P and L lookbacks,

$$b_t^{\text{season}} = \text{median}\{v_{k-P}, v_{k-2P}, \ldots, v_{k-LP}\}. \tag{22}$$

Define the absolute seasonal residual $e_i = |v_i - b_i^{\text{season}}|$, where early training indices without a seasonal lag use the rolling baseline. At test time,

$$S_H = 1.4826 \text{ median}\,|H - \text{median}(H)| + 10^{-8}, \tag{23}$$

$$s_t^{\text{season}} = \max\{1.4826 \text{ median}(e_{k-PL:k}),\ 0.05S_H + 10^{-8}\}. \tag{24}$$

The scale window is shifted by one observation: it never includes e_k. The guard $n \geq \max\{h, P(L+1)\}$ ensures that the first test seasonal baseline has all L lags and every residual in its PL-point scale window has at least one genuine seasonal lag. This corrects the weaker $n \geq PL$ condition, for which $P = 24, L = 3, n = 72$ is a counterexample. Rolling-only rules require $n \geq h$. The temporal state described below is initialized anew at the first test point rather than inherited from training violations.

For direction δ and threshold τ, the atomic robustness is

$$a_t^r(\delta) = \begin{cases} x_t - b_t^r - \tau s_t^r, & \delta = \text{increase}, \\ b_t^r - x_t - \tau s_t^r, & \delta = \text{decrease}, \\ |x_t - b_t^r| - \tau s_t^r, & \delta = \text{either}. \end{cases} \tag{25}$$

A positive value satisfies the atom. The dual reference is a logical disjunction, not an average. Dividing by $s_t^r > 0$ yields the standardized margin with the same sign, but varying scales can change rankings across time. Robustness is a signed engineering margin, not a probability or calibrated confidence. Input range, finiteness, scale positivity, and arithmetic overflow are checked; invalid inputs fail rather than being coerced.

D.3 Bounded past-time persistence and gating

With unavailable history defined as $-\infty$, the exact temporal core is

$$h_t^r = \min_{j=0}^{m-1} a_{t-j}^r, \tag{26}$$

$$p_t^r = \max_{k=0}^{w-m} h_{t-k}^r, \tag{27}$$

$$p_t = \max_{r \in R} p_t^r, \tag{28}$$

where R is one selected reference or both references. In past-time STL notation (Maler & Nickovic, 2004; Donzé & Maler, 2010), each branch is

$$\mathbf{O}_{[0,w-m]}\mathbf{H}_{[0,m-1]}(a^r > 0),$$

and the dual reference combines the branches by OR. For a candidate mask $q_t \in \{0, 1\}$, quantitative conjunction with the gate is implemented exactly as

$$g_t = \begin{cases} p_t, & q_t = 1, \\ -\infty, & q_t = 0. \end{cases} \tag{29}$$

If the rule does not require candidates, $g_t = p_t$.

Unavailable test history makes $h_t^r = -\infty$. Hence $m = 1$ reduces the historical minimum to the atom and $w = m$ removes the outer lookback. In general, positivity means that the past w test observations contain a run of m consecutive positive atoms; neither m arbitrary positives nor a currently positive atom is required. For example, with $(m, w) = (2, 4)$ and atomic signs $(+, +, -, -)$, the fourth output remains positive because the run ending at the second point still lies in the four-point window.

D.4 Exact bounded non-causal CLOSE wrapper

For $d > 1$, extend a finite robustness trace by $-\infty$ outside its domain. The implementation applies the following flat dilation and erosion:

$$(D_d g)_t = \max_{u=-\lfloor (d-1)/2 \rfloor}^{\lceil (d-1)/2 \rceil} g_{t+u}, \tag{30}$$

$$(E_d z)_t = \min_{u=-\lfloor d/2 \rfloor}^{\lceil d/2 \rceil - 1} z_{t+u}, \tag{31}$$

$$\mathrm{CLOSE}_d(g)_t = (E_d D_d g)_t. \tag{32}$$

Each operator returns values only on the finite trace domain. Before erosion, the finite dilation output is extended again by $-\infty$; there is no retained off-domain dilation state. This indexing records the exact even-width anchoring: dilation is implemented with `maximum_filter1d(origin=-1)` for even d (origin zero for odd d), and erosion uses `minimum_filter1d(origin=0)`. For $d \leq 1$, `CLOSE` is the identity. The final decision is $\hat{y}_t = \mathbb{1}[\mathrm{CLOSE}_d(g)_t > 0]$.

Because Eq. (32) accesses positive offsets, the complete monitor is offline and non-causal whenever $d > 1$, with at most $d - 1$ observations of look-ahead through the composition. Only the core in Eq. (28) is past-time causal. Closing can also fill a short gap outside the original candidate mask because gating precedes morphology; this is intended and is bound into the replay artifact.

D.5 Independent interpreter and hand calculation

The test suite contains a deliberately scalar interpreter that shares no feature, temporal, morphology, or interval-extraction helper with the production monitor. Random differential tests cross all three reference modes, all directions, candidate/no-candidate execution, persistence edges, and odd/even closing widths. Minimal tests cover $L = 1$, the exact history boundary, and one point below it.

For a hand calculation, take a zero history, rolling reference, increasing direction, $(\tau, m, w, d) = (1, 2, 3, 3)$, test values $(0, 10, 10, 0, 0, 10, 10, 0)$, and candidate mask $(1, 1, 1, 1, 1, 1, 1, 0)$. The atomic signs are $(-, +, +, -, -, +, +, -)$; persistence signs are $(-, -, +, +, -, -, +, +)$; gating changes the final sign to negative; and width-three closing fills the two-point internal gap. The resulting Boolean trace is $(0, 0, 1, 1, 1, 1, 1, 0)$, giving the single half-open interval $[2, 7)$. Both implementations reproduce every intermediate array in this example.

Lemma 1 (Sign equivalence of quantitative closing)**.** *For the same finite support, outside value, width, and anchoring,*

$$\mathbb{1}[\mathrm{CLOSE}_d(g) > 0] = \mathrm{binary_closing}_d(\mathbb{1}[g > 0]).$$

Proof. A finite maximum is positive exactly when at least one input is positive, so thresholded dilation equals Boolean dilation. A finite minimum is positive exactly when all inputs are positive, so thresholded erosion equals Boolean erosion. The $-\infty$ extension maps to Boolean false at the boundary. Composition yields the result. □

With this footprint convention, width d fills an interior Boolean gap of at most $d - 1$ samples; `merge_gap_points=`d does not mean that a gap of d samples is filled. False padding also removes possible boundary positives over the erosion reach: $\lfloor d/2 \rfloor$ positions on the left and $\lceil d/2 \rceil - 1$ on the right. For $d = 48$, these reaches are 24 and 23 positions. These are executable boundary semantics, not asymptotic approximations.

E Dataset Description and Visual Atlas

E.1 Collection structure and array-derived inventory

The dataset consists of real backbone-link monitoring data. Twenty-nine annotators annotated four morphology categories; binary anomaly labels are determined by actual network failures, independently of those morphology annotations. The collection comprises 63 univariate KPI series, organized under nine scenario prefixes in four morphology folders: Mutation, Step, Trend, and Wave. Mutation and Step each contain one scenario with nine metrics; Trend contains four scenarios with nine metrics each; Wave contains three scenarios with only Request Count, Success Count, and Success Rate. The other six metric names are Avg Latency, Max Latency, Failure Count, Failure Rate, Retries, and Timeouts. A scenario prefix is a directory identifier, not evidence of synchronized multivariate measurements or an independent source service. Family comparisons consequently reflect differences in both waveform structure and metric composition.

Each series supplies 4,032 training and 2,016 evaluation observations, with matching one-dimensional binary label arrays. The value arrays contain finite numeric values. Both splits contain positive labels for every series, so the training data are not an anomaly-free reference set. We define an event as a maximal contiguous positive run, using zero-based half-open bounds $[s, e)$ and length $e - s$. These intervals are derived from the binary labels for actual network failures, rather than from the morphology annotations. Tables 6 and 7 are recomputed from the arrays rather than copied from metadata.

Family	Split	Series	Points	Positive	Rate (%)	Events
Mutation	Train	9	36,288	1,350	3.72	27
	Evaluation	9	18,144	450	2.48	9
Step	Train	9	36,288	2,700	7.44	54
	Evaluation	9	18,144	450	2.48	9
Trend	Train	36	145,152	6,204	4.27	114
	Evaluation	36	72,576	2,103	2.90	39
Wave	Train	9	36,288	8,928	24.60	36
	Evaluation	9	18,144	2,232	12.30	9
Total	Train	63	254,016	19,182	7.55	231
Total	Evaluation	63	127,008	5,235	4.12	66

Table 6: Array-derived dataset inventory. A positive point has supplied binary label 1; an event is a maximal contiguous run of positive points. The 63 series are not independent incident replications.

The event-length distribution is unusually discrete. Training contains 189 events of length 50, six of length 134, 12 of length 168, and 24 of length 288. Evaluation contains 48 events of length 50, nine of length 67, three of length 168, and six of length 288. Length 134 occurs only in training. The prevalence is also uneven: Wave accounts for only nine series but 2,232 of the 5,235 positive evaluation points. Event counts and point counts therefore weight the corpus differently. Figure 8 visualizes these differences without implying independent samples.

Scenario	Series	Positive points		Events		Eval. lengths
		Train	Eval.	Train	Eval.	(points)
Mutation_dataset	9	1,350	450	27	9	50
Step_dataset	9	2,700	450	54	9	50
trend_dataset1	9	1,350	450	27	9	50
trend_dataset2	9	1,350	450	27	9	50
trend_dataset3	9	1,350	552	27	9	50, 67
trend_dataset4	9	2,154	651	33	12	50, 67
Wave_dataset2	3	2,016	504	12	3	168
Wave_dataset3	3	3,456	864	12	3	288
Wave_dataset4	3	3,456	864	12	3	288

Table 7: Scenario coverage. Mutation, Step, and each Trend scenario contain all nine metric names; each Wave scenario contains Request Count, Success Count, and Success Rate only. A shared prefix does not establish synchronized multivariate measurements.

E.2 Metric meaning and the local policy

The metric names identify supplied measurements; their physical units, source services, and clock alignment are not documented. Latency values are therefore not assigned milliseconds, and rate values are not automatically converted to percentages. The metadata field `intervals=300` has no documented unit. All trace plots use observation indices and supplied numeric values rather than dates or elapsed time.

The evaluated semantic contract supplies a harmful direction for each metric group (Table 8). This context is distinct from signal morphology. It does not establish that the supplied labels were independently adjudicated against that policy, or that a language model inferred the policy from the waveform. In particular, request-volume polarity depends on the operating objective and is not universally downward.

Supplied metric names	Harmful direction in the local contract
Avg Latency; Max Latency	Increase
Failure Count; Failure Rate	Increase
Retries; Timeouts	Increase
Success Count; Success Rate	Decrease
Request Count	Decrease (context-specific)

Table 8: Author-supplied policy inputs. These are operational assumptions of the controlled study, not physical units or independently established incident causes.

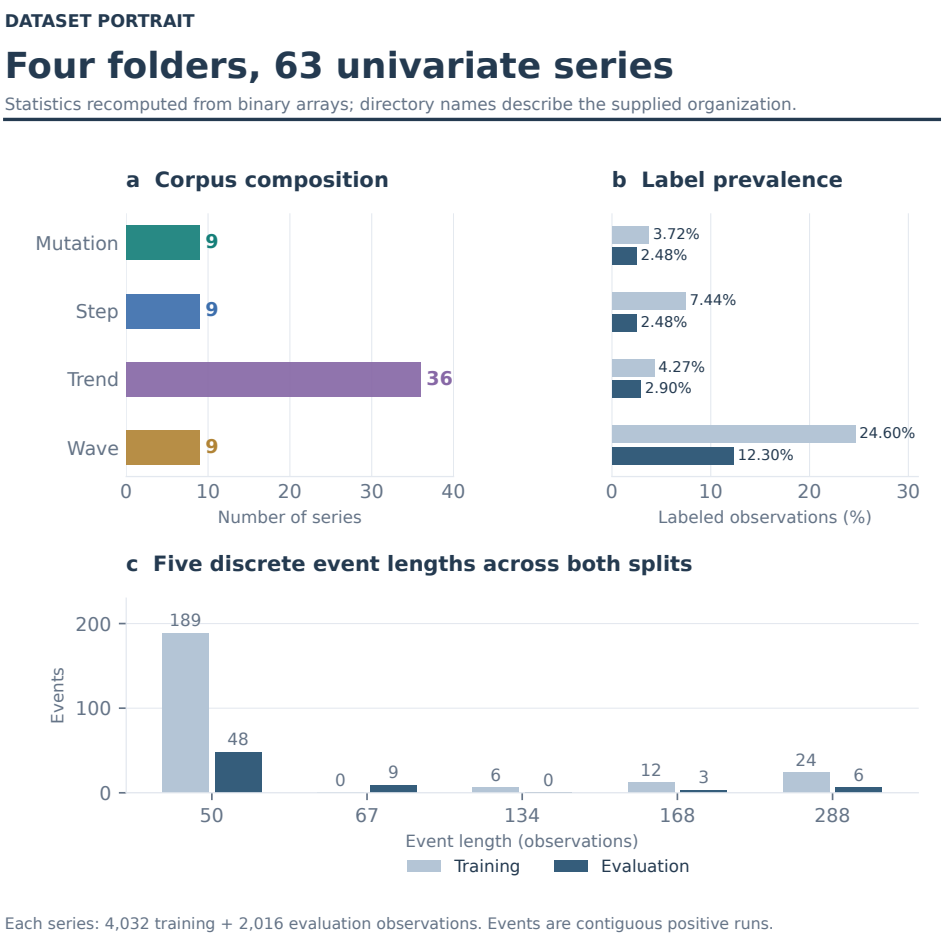


Figure 8: Dataset portrait from the supplied arrays. (a) Series counts are uneven across morphology folders. (b) Positive-point prevalence differs between training and evaluation, with Wave containing much longer labeled intervals. (c) Only five event lengths occur across both splits; zero-height bars are annotated with zero. All counts are descriptive, and all events are contiguous binary-label runs.

E.3 Observed morphology versus supplied annotations

The four directory names organize the morphology annotations separately from the binary failure labels; a morphology category does not determine whether a particular observation is anomalous. A visually conspicuous excursion can be unlabeled, and a labeled interval can end while the changed background continues. The following descriptions characterize visible structure in the arrays; they do not identify an undocumented injection mechanism or generation process.

Mutation: point excursions with rapid recovery. Mutation denotes isolated or short groups of extreme observations followed by a rapid return to the surrounding trajectory. The focus plate selects these visible point excursions, rather than selecting a positive binary-label interval. In Avg Latency, samples 779–781 rise to approximately 178–184 while the adjacent values are about 75; sample 782 has already recovered. Request Count and Success Rate provide four-point downward excursions. These are actual supplied observations, without smoothing or replacement.

Step: sustained changes of level. A step changes the level across a subsequent interval rather than only a few isolated observations. In Step/Avg Latency, the shifted level persists for samples 534–583 and returns at sample 584. The depicted steps last 50 observations; a temporary level shift need not persist to the end of the sequence. Figure 10 shows the onset, sustained interior, and recovery for three metrics.

Shape and detection-target mismatch. The selected Mutation point excursions have binary target 0, whereas its positive target intervals are 50-observation level departures. Thus a positive-label-driven crop would display a step-shaped segment under a Mutation heading. The two annotations must not be conflated: the focus plate marks point-shape examples separately, and the complete atlas retains the supplied binary labels. Reported scores for the Mutation folder cannot, by themselves, establish successful point-anomaly detection. This distinction follows from the separate annotation roles: network failures determine the detection target, while researchers describe morphology; labels are not rewritten to match a visual expectation.

Trend: local departures on changing backgrounds. The four Trend scenarios contain evolving or changing backgrounds together with locally annotated deviations. A Trend folder name does not imply that every labeled event is a gradual drift. In scenario 1, the Avg Latency evaluation label covers a short elevated plateau; in scenario 3, its label covers the initial 67 observations of a higher regime that continues afterward. The three decrease-is-harmful metrics in scenario 4 each contain two evaluation events, and both appear in the complete atlas. Figure 11 shows why event boundaries must be read from the supplied annotations, rather than reconstructed from a morphology name.

Wave: attenuation of recurring structure. The three Wave scenarios show recurring variation with intervals of attenuated or nearly suppressed amplitude. Figure 12 holds the metric name fixed at Request Count to expose differences across scenarios: scenario 2 has a 168-observation annotation, while scenarios 3 and 4 each have a 288-observation annotation. These examples illustrate changes in a pattern even when individual values overlap the surrounding numeric range. Wave includes only the three locally decrease-is-harmful metrics, so it is not a nine-metric counterpart of the other families.

E.4 Visual design, selection, and reading conventions

Each focus plate pairs a complete evaluation trace with a local view. Mutation and Step use Avg Latency, Request Count, and Success Rate. Mutation selects raw point excursions $[779, 782)$, $[465, 469)$, and $[499, 503)$, respectively, with 16 surrounding observations on each side. Step uses its first positive-label interval. Trend uses scenario 1/Avg Latency, scenario 3/Avg Latency, and scenario 4/Success Rate; Wave uses Request Count in each scenario. For Step, Trend, and Wave, the first labeled event $[s, e)$ receives context padding $\max(80, e - s)$. All windows are clipped to split bounds and marked by dashed boxes. Selection is descriptive, not based on detector performance.

The complete atlas in Section E.7 displays all 63 series and both stored splits, exactly once per series. Training and evaluation share the same vertical range within each metric row, as do the full and local views in each focus row. Vertical scales differ between rows. Original values are preserved without smoothing, resampling, clipping of data values, or normalization. Horizontal axes use the index within each split and should not be joined into an inferred calendar timeline. Numerical ranges can legitimately be very different across metric rows.

Blue lines show supplied values; terracotta segments and pale bands mark positive supplied binary targets. Teal dots in the Mutation focus plate instead mark the explicitly selected point-shape examples, whose binary targets are zero. Bands follow sample-cell boundaries $[s - 0.5, e - 0.5]$ so that the highlighted samples match the half-open index interval $[s, e)$. Unshaded positions have binary label zero, indicating no recorded network failure at those observations. Pattern annotations and detector scores are absent. These labeled visualizations are evaluator-only explanations and do not enter detector features, language-model prompts, retrieval, or operating-point selection.

E.5 Provenance, dependence, and reproducible artifacts

All 63 metadata files contain the same anomaly-ratio fields, and their training, evaluation, and overall ratios disagree with their own arrays. The array-derived inventory is authoritative for this appendix. The release organizations and annotation roles are described above. A release URL, license, generation or transformation records, timestamps, metric units, and source-instance identifiers are not included in the local artifact. The four morphology categories belong to this single dataset.

The prior raw-array audit reports median exact train/evaluation value reuse over aligned 2,016-observation blocks of approximately 86.7%, 77.4%, 38.1%, and 85.6% for Mutation, Step, Trend, and Wave, respectively. Every Wave evaluation anomaly interval exactly reuses the values at the same position in the first 2,016 training observations. Among the 63 series, only 32 distinct training-label arrays and 34 distinct evaluation-label arrays occur. These dependencies preclude treating the 63 series as independent incident replications. The auxiliary four-column pattern arrays exactly reveal the binary labels and remain excluded from model-facing computation.

The generator `scripts/make_dataset_atlas.py` saves per-series statistics, all 297 event intervals, focus-window selections, and input/figure SHA-256 hashes. Figures are exported as PDF, SVG, and PNG. Original arrays and frozen experimental results are unchanged.

E.6 Four morphology focus plates

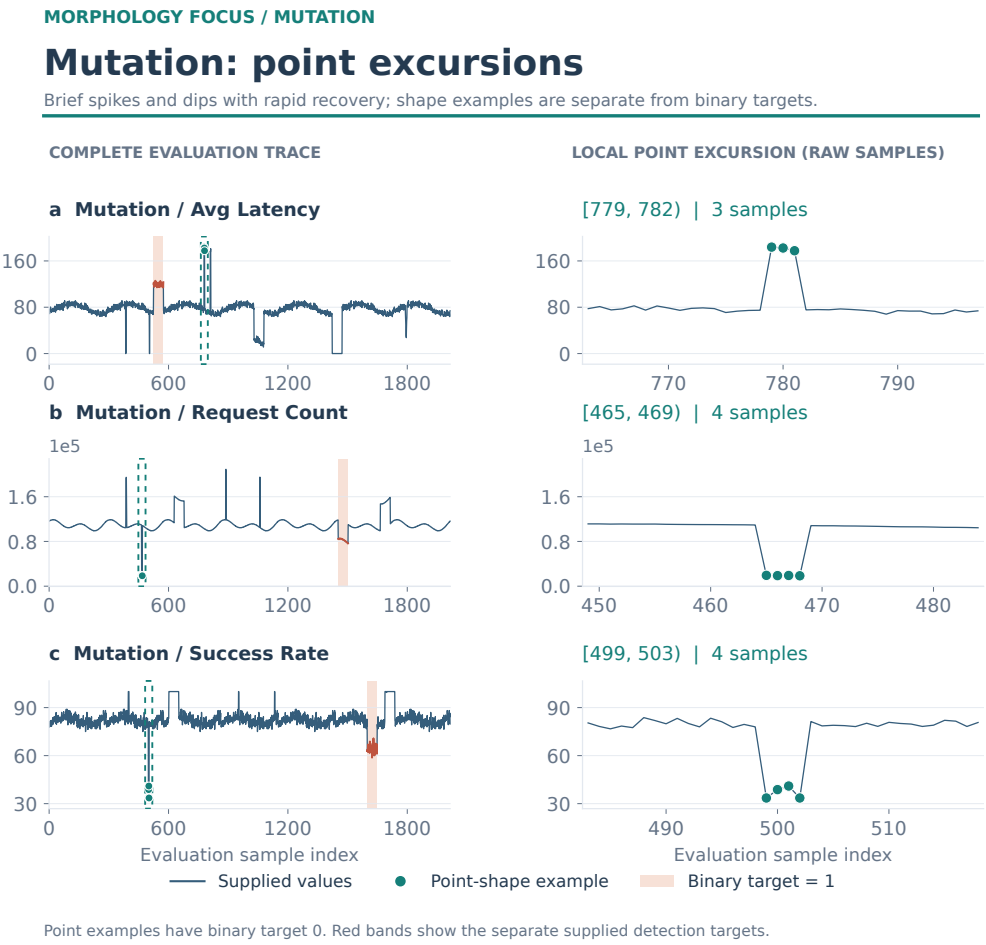


Figure 9: Mutation point morphology. Complete traces (left) and brief excursions with immediate recovery (right): three high observations in Avg Latency, four low observations in Request Count, and four low observations in Success Rate. Teal dots identify these raw-value examples; their supplied binary targets are all zero. Red bands elsewhere retain the positive detection targets, which are distinct 50-observation intervals. Each row shares one vertical scale; no values or labels are changed.

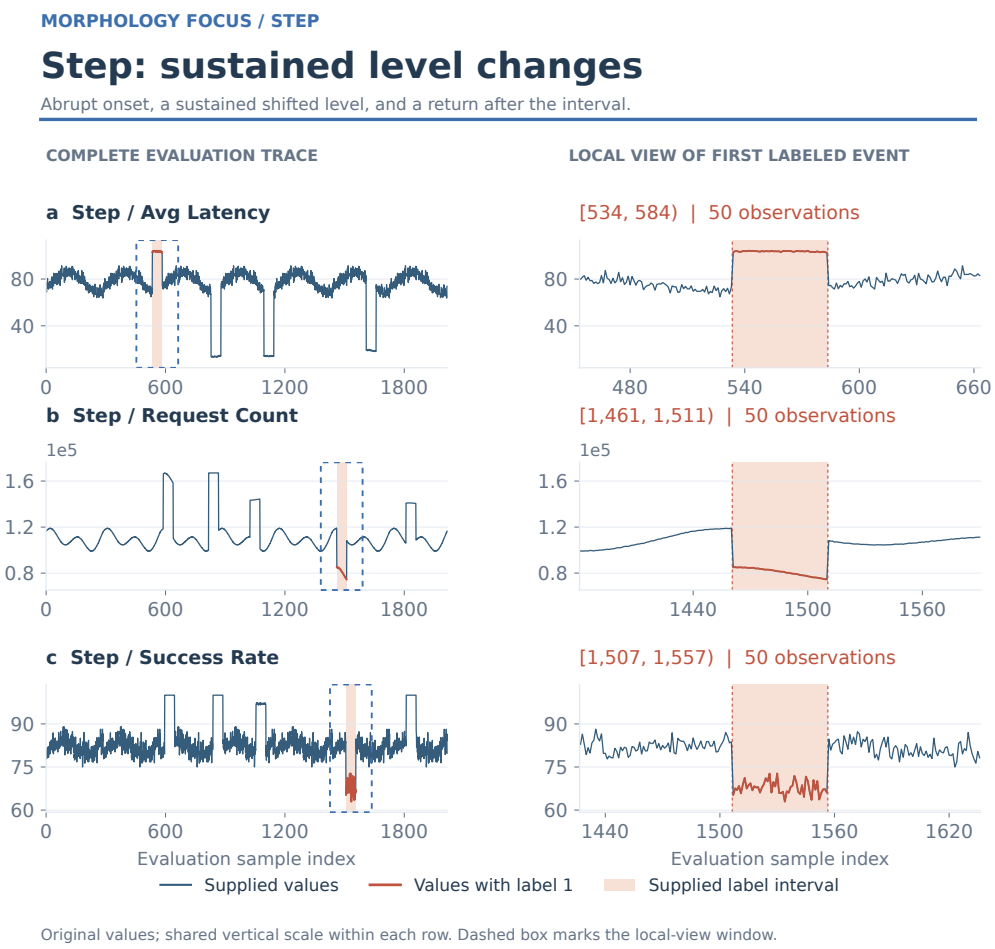


Figure 10: Step morphology. Complete evaluation traces and local views for Avg Latency, Request Count, and Success Rate. Each level change persists throughout a 50-observation interval before recovery, distinguishing sustained step behavior from the brief point excursions in Figure 9. Here the illustrated intervals also have binary target 1, shown in terracotta. Original values are shown on a shared vertical scale within each row.

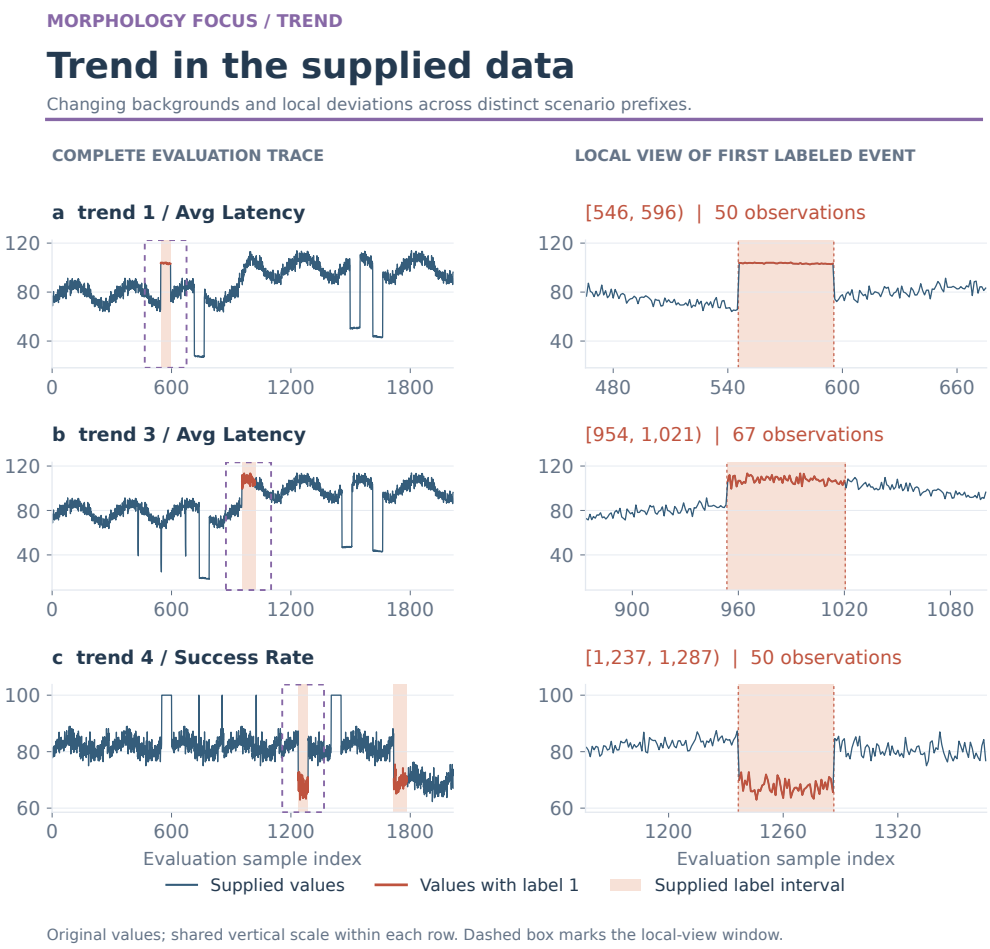


Figure 11: Trend. Named examples from scenarios 1, 3, and 4 distinguish evolving backgrounds from supplied event boundaries. Scenario 3/Avg Latency remains elevated after its positive-label interval ends. Scenario 4/Success Rate has two evaluation annotations, both visible on the left; the right panel enlarges the first. All four Trend scenarios and their nine metrics appear in the complete atlas.

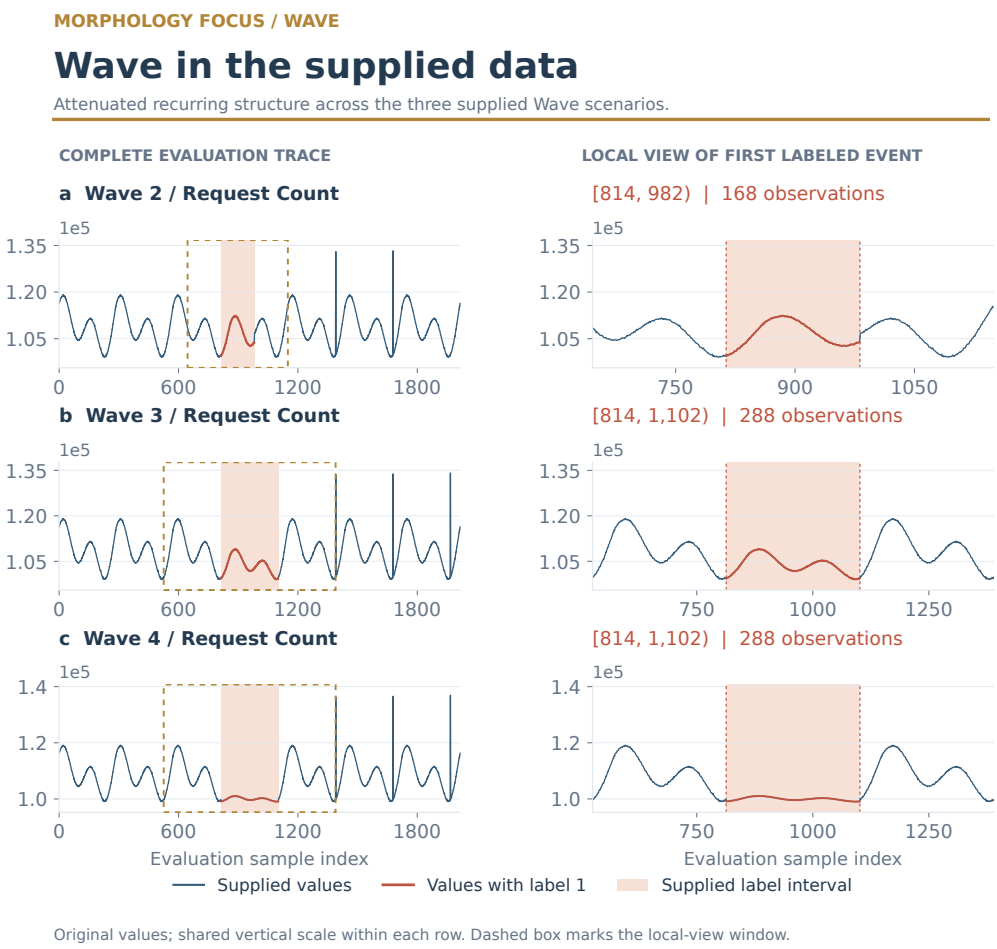


Figure 12: Wave. Request Count in scenarios 2, 3, and 4, with 168-, 288-, and 288-observation labeled intervals, respectively. Holding the metric name fixed helps expose partial and stronger attenuation of recurring variation. The full trace provides the surrounding pattern, while the local view exposes the interval's extent. These intervals reuse same-position training values and are not independent incident evidence.

E.7 Complete training and evaluation trace atlas

Every supplied series appears exactly once below, grouped by its source scenario. Training and evaluation traces share a vertical scale within each metric row. Metrics use their original numeric values; scales differ between rows. Both horizontal axes count observations within their own split, so the panels do not assert a verified calendar-time alignment. Complete event intervals and source checksums accompany the figure artifacts.

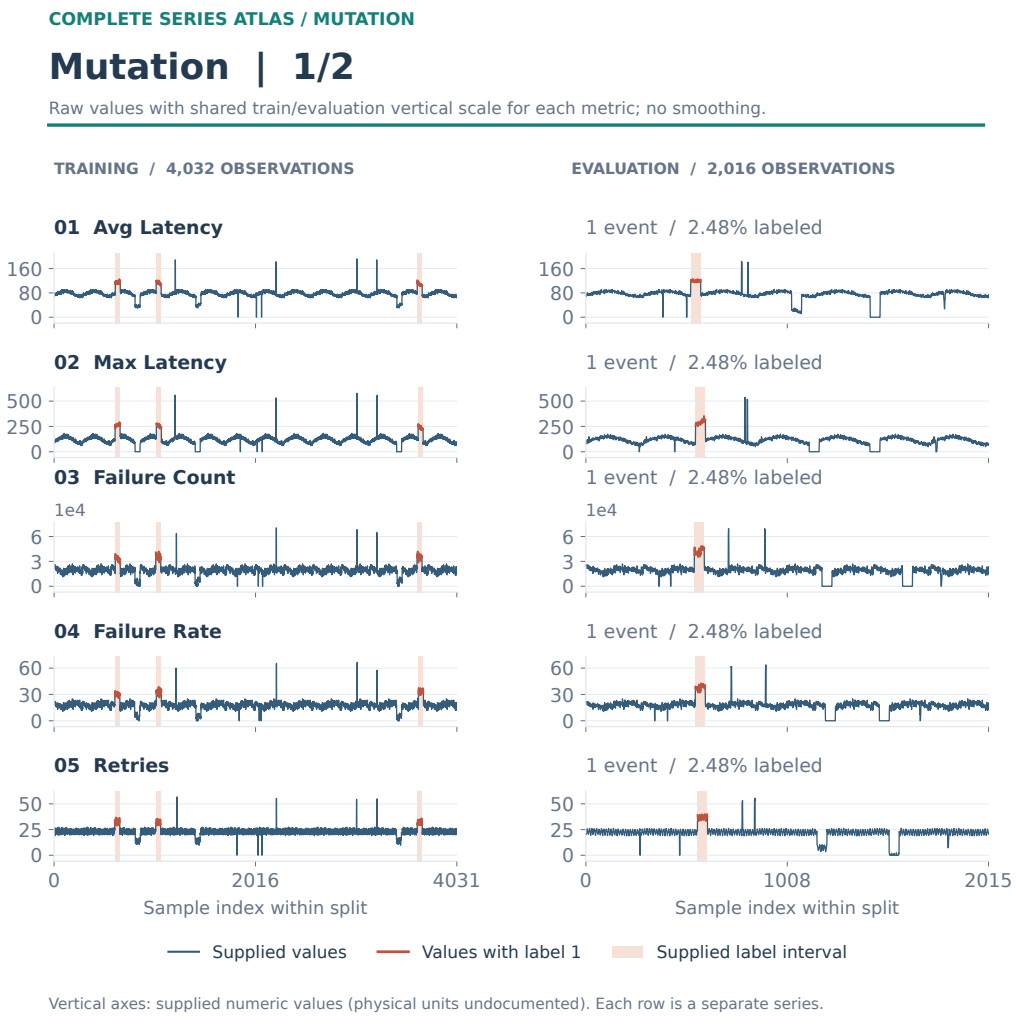


Figure 13: Mutation_dataset, part 1 of 2. Complete traces for Avg Latency, Max Latency, Failure Count, Failure Rate, Retries. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

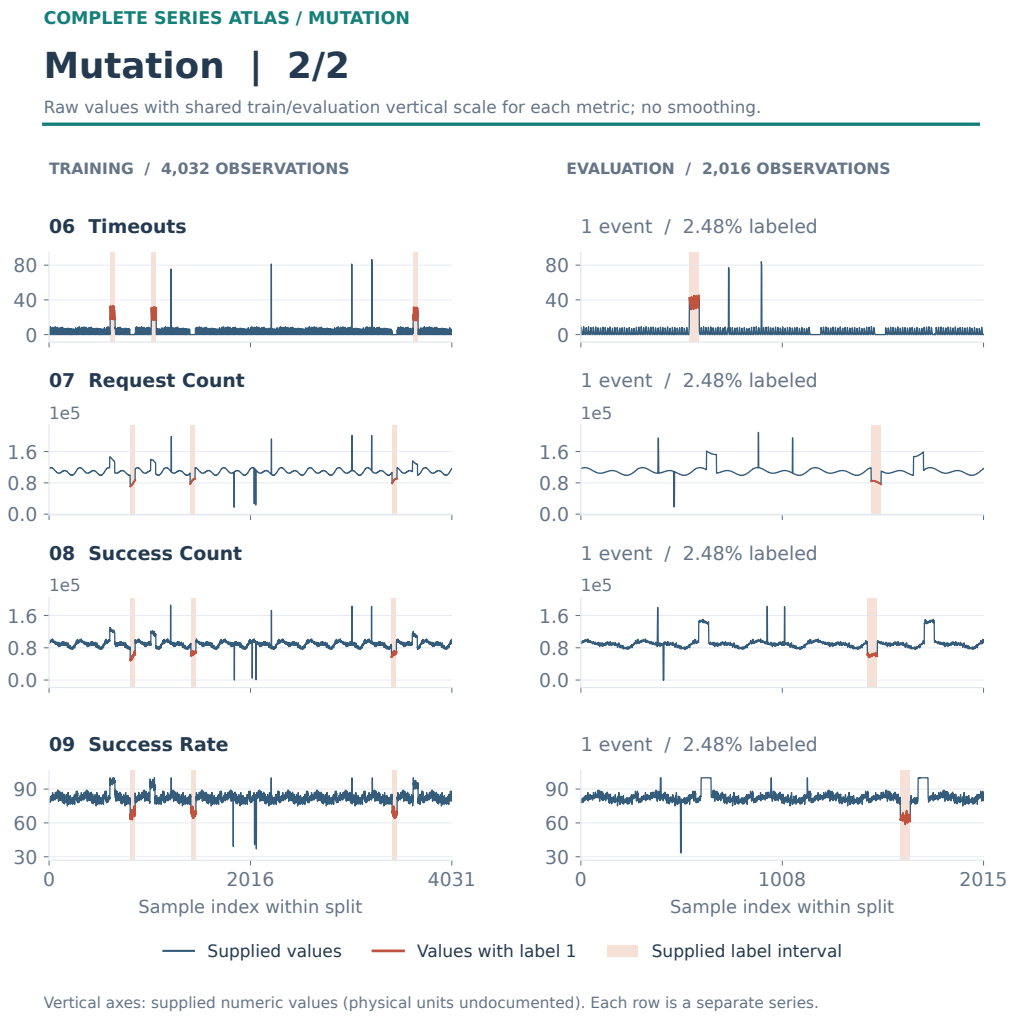


Figure 14: Mutation_dataset, part 2 of 2. Complete traces for Timeouts, Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.



Figure 15: Step_dataset, part 1 of 2. Complete traces for Avg Latency, Max Latency, Failure Count, Failure Rate, Retries. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.



Figure 16: Step_dataset, part 2 of 2. Complete traces for Timeouts, Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

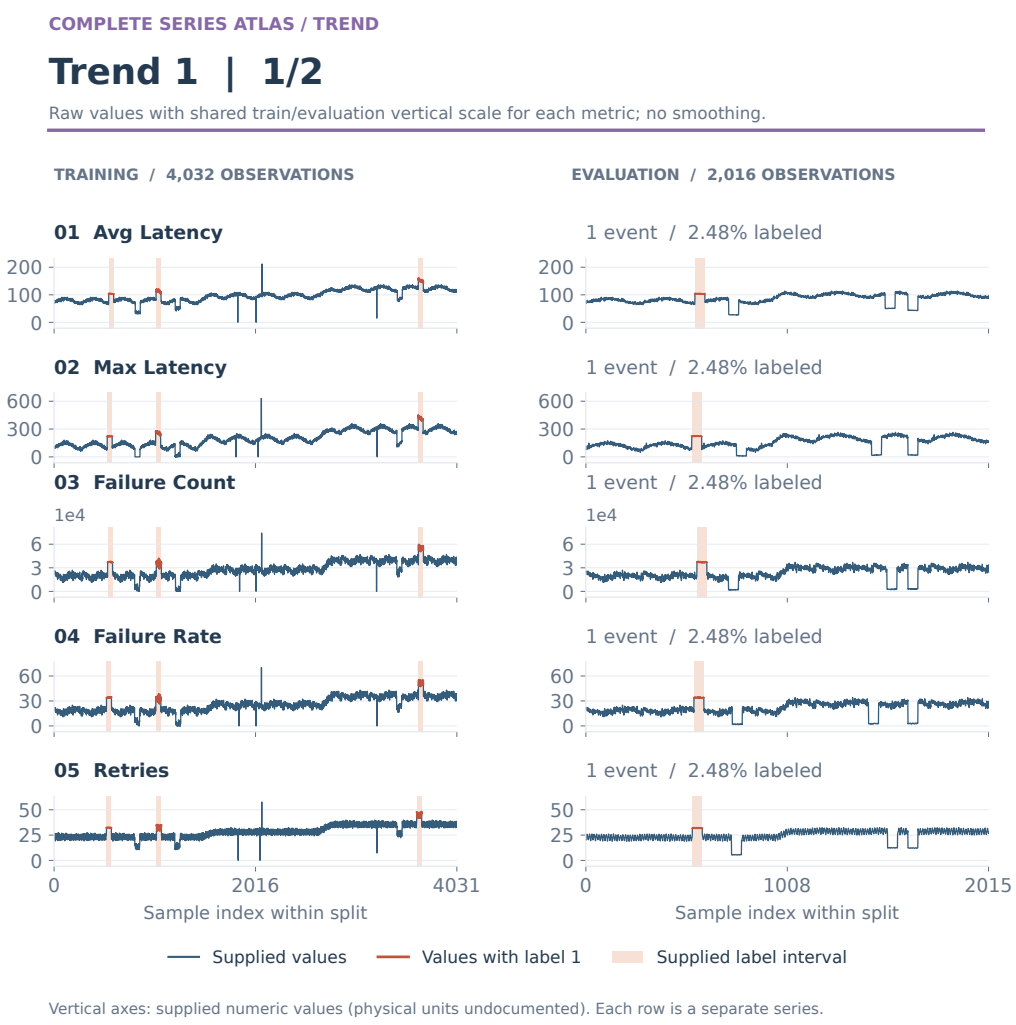


Figure 17: trend_dataset1, part 1 of 2. Complete traces for Avg Latency, Max Latency, Failure Count, Failure Rate, Retries. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

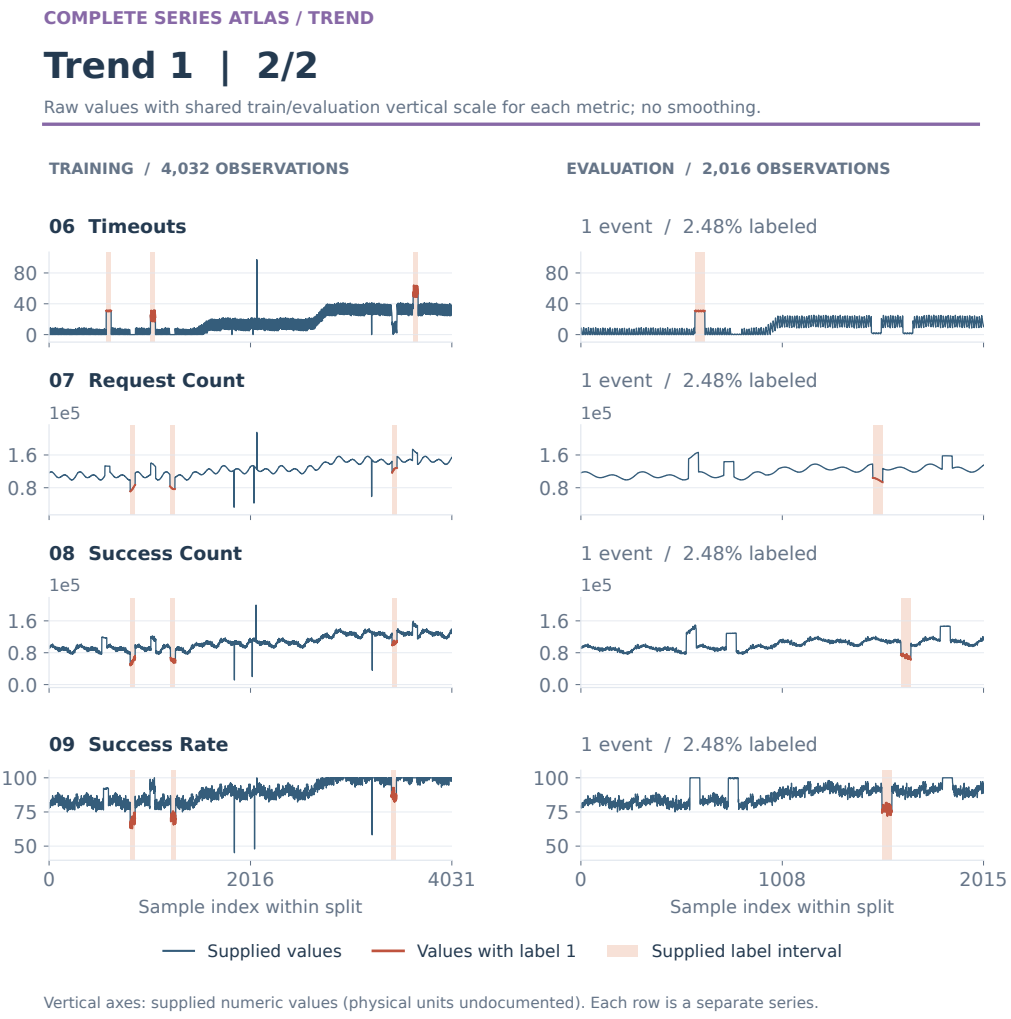


Figure 18: trend_dataset1, part 2 of 2. Complete traces for Timeouts, Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

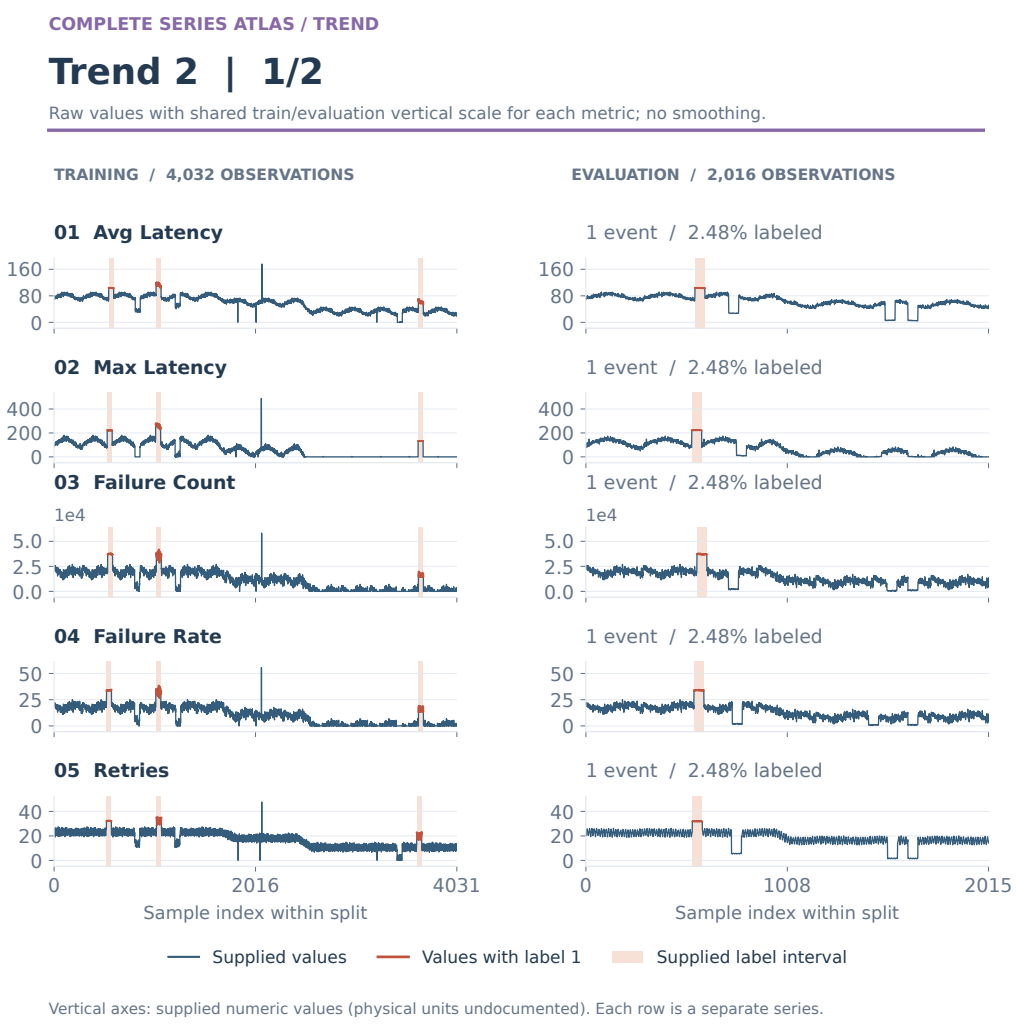


Figure 19: trend_dataset2, part 1 of 2. Complete traces for Avg Latency, Max Latency, Failure Count, Failure Rate, Retries. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.



Figure 20: trend_dataset2, part 2 of 2. Complete traces for Timeouts, Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

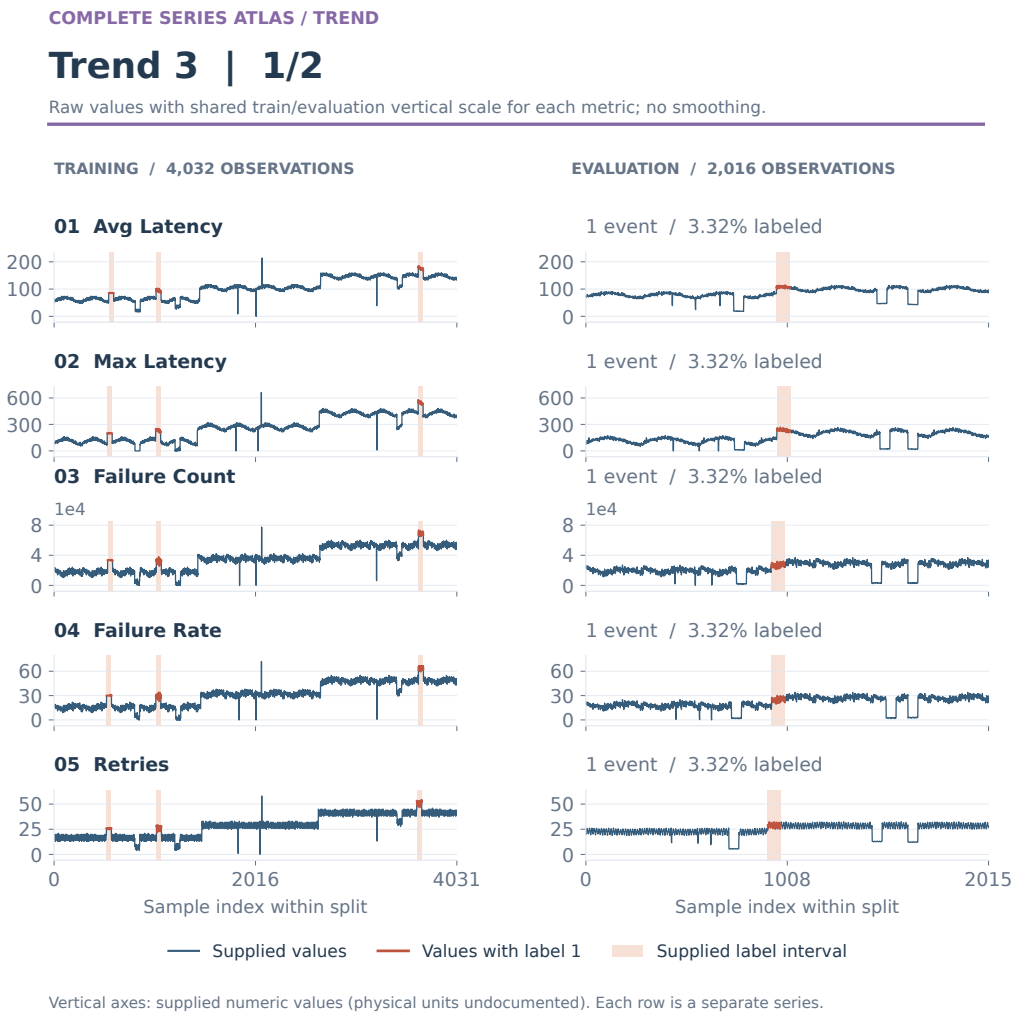


Figure 21: trend_dataset3, part 1 of 2. Complete traces for Avg Latency, Max Latency, Failure Count, Failure Rate, Retries. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

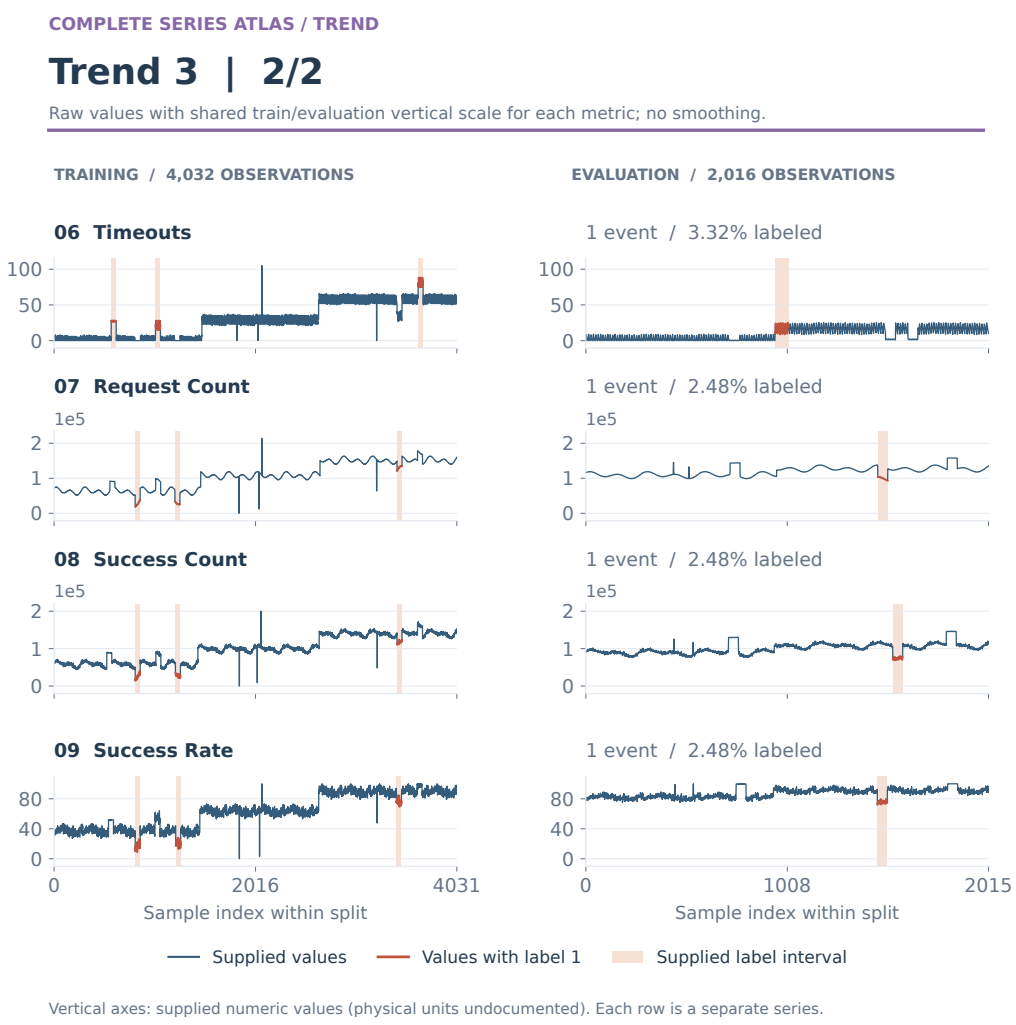


Figure 22: trend_dataset3, part 2 of 2. Complete traces for Timeouts, Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

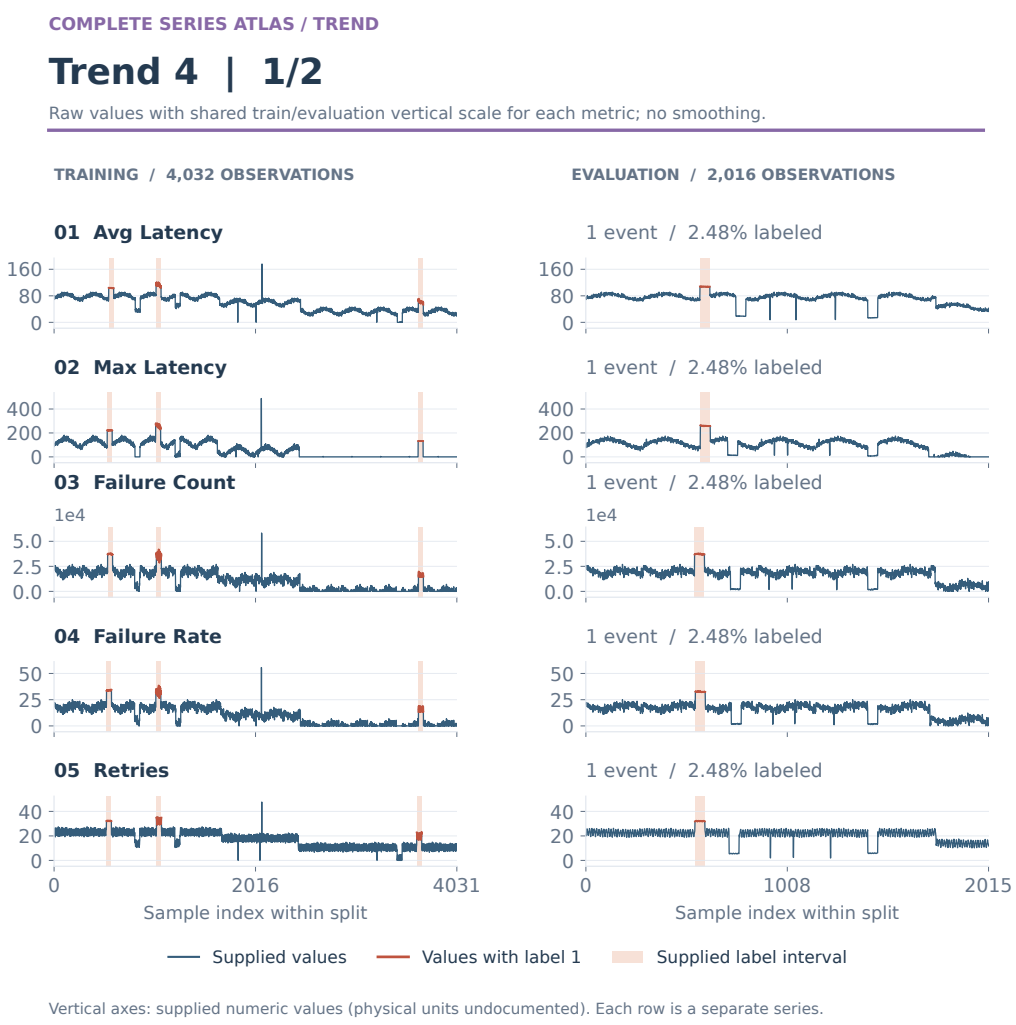


Figure 23: trend_dataset4, part 1 of 2. Complete traces for Avg Latency, Max Latency, Failure Count, Failure Rate, Retries. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

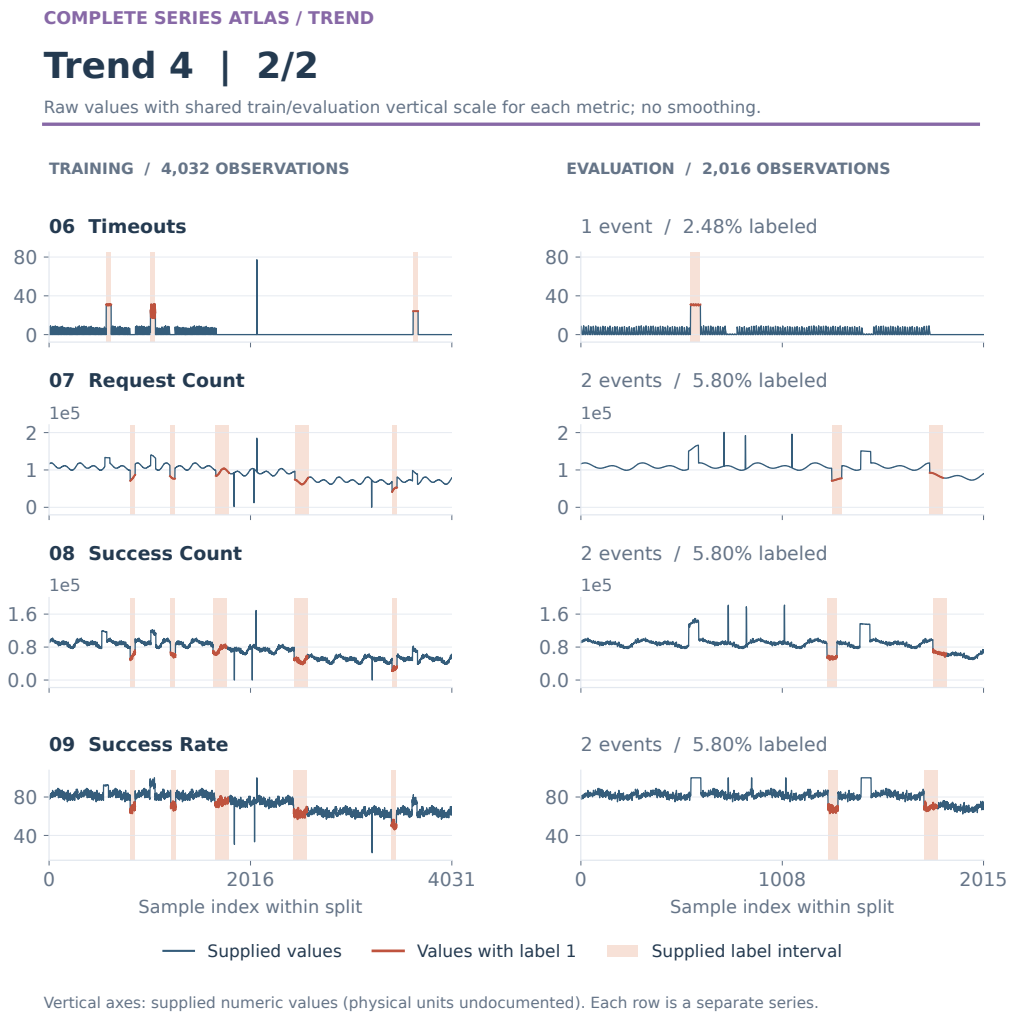


Figure 24: trend_dataset4, part 2 of 2. Complete traces for Timeouts, Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

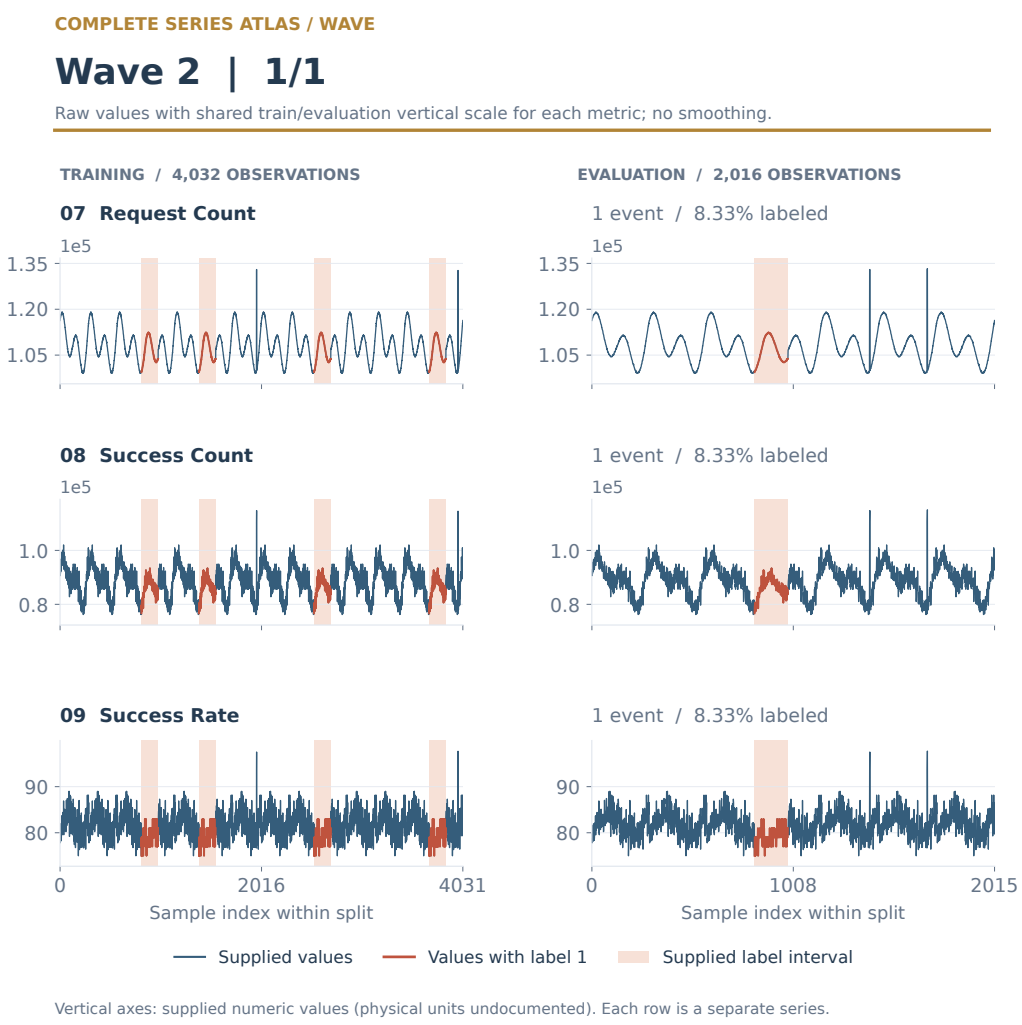


Figure 25: Wave_dataset2, part 1 of 1. Complete traces for Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

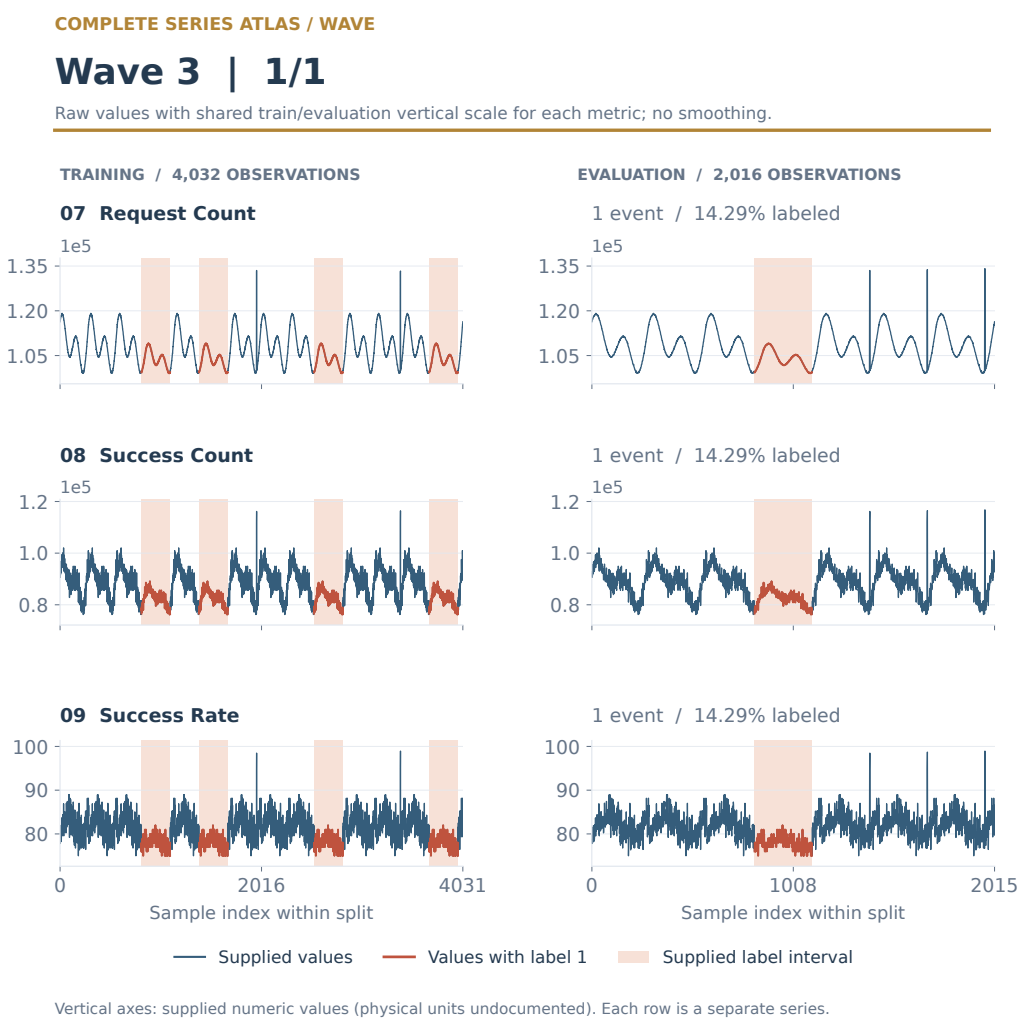


Figure 26: Wave_dataset3, part 1 of 1. Complete traces for Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

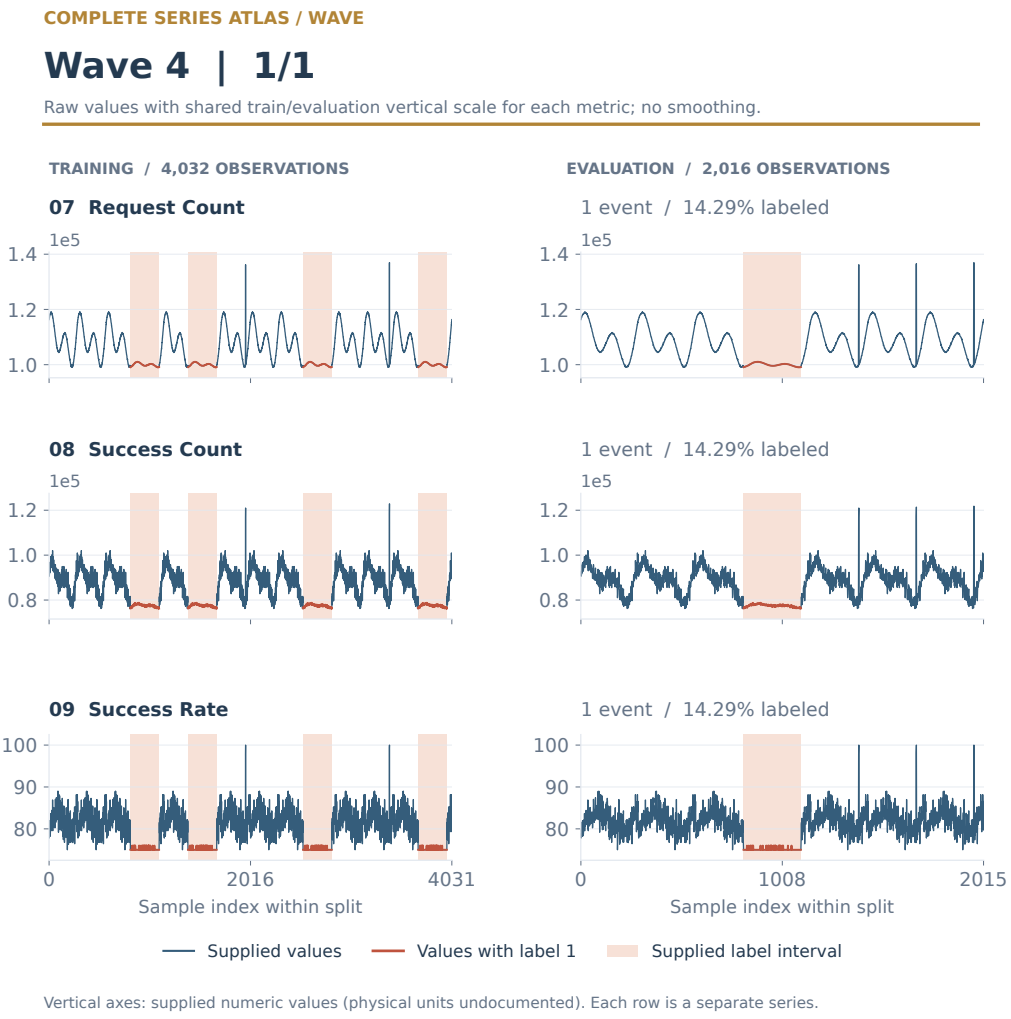


Figure 27: Wave_dataset4, part 1 of 1. Complete traces for Request Count, Success Count, Success Rate. Shading and colored segments reproduce supplied labels for illustration only. No detector outputs or pattern annotations are shown.

F A Successful Detection, Explained Step by Step

We trace one successful detection from the supplied values to its final interval, connecting each design component to its numerical contribution. The primary example is `trend_dataset1` / Max Latency (`series_021`). Its supplied evaluation annotation is $[530, 580)$ and its final prediction is $[534, 584)$: 46 true-positive, four false-positive, and four false-negative observations. The resulting intersection over union is $46/54 = 0.852$, strict event F1 is 1.000, and point F1 is 0.920. This accurately localized alert provides a concrete example of broad proposals becoming a coherent policy-qualified event.

An explicit explanatory selection. Among the 63 evaluated series, 27 have full-method strict event F1 equal to one and point F1 at least 0.90. The archived selection ranks these candidates by the number of strict-F1 decreases under six matched controls, then by their mean decrease, then by full-method point F1, all in descending order. The six selection controls remove direction, persistence, closing, the seasonal reference, the rolling reference, or the sieve. This selects the primary example because several mechanisms are visible in one accurately detected trace; it does not select the largest point F1 regardless of explanatory value. Table 9 reports the selected semantic-component controls used in the following explanation.

Configuration	Output intervals	Strict event F1	Point F1
Full method	1	1.000	0.920
Without direction	5	0.333	0.347
Flipped direction	4	0.000	0.000
Without persistence	2	0.667	0.971
Without output closing	2	0.000	0.615
Rolling reference only	1	1.000	0.920
Seasonal reference only	0	0.000	0.000

Table 9: Selected frozen semantic-component controls on the primary example. Strict event F1 uses one-to-one IoU-0.5 matches; point F1 evaluates individual observations. Other settings remain unchanged. The direction-flipped diagnostic is shown but is not one of the six controls used in the archived example selection.

This retrospective, label-informed illustration executes an author-set reference AST on the supplied corpus. The matched controls make the design benefits inspectable: correct direction concentrates evidence on adverse changes, persistence selects a single event, and closing restores interval continuity. The table retains both event and point endpoints, while the supporting Wave example demonstrates the complementary seasonal path.

F.1 Step 1: Locate the result in its complete signal context

The full trace contains local departures together with a changing background. Figure 28 shows how contextual diagnosis identifies one event within that complete signal. The final prediction begins and ends four observations after the supplied label and overlaps it for 46 observations. Its IoU of 0.852 comfortably exceeds the 0.5 matching threshold, illustrating the interval coherence achieved by the complete rule.

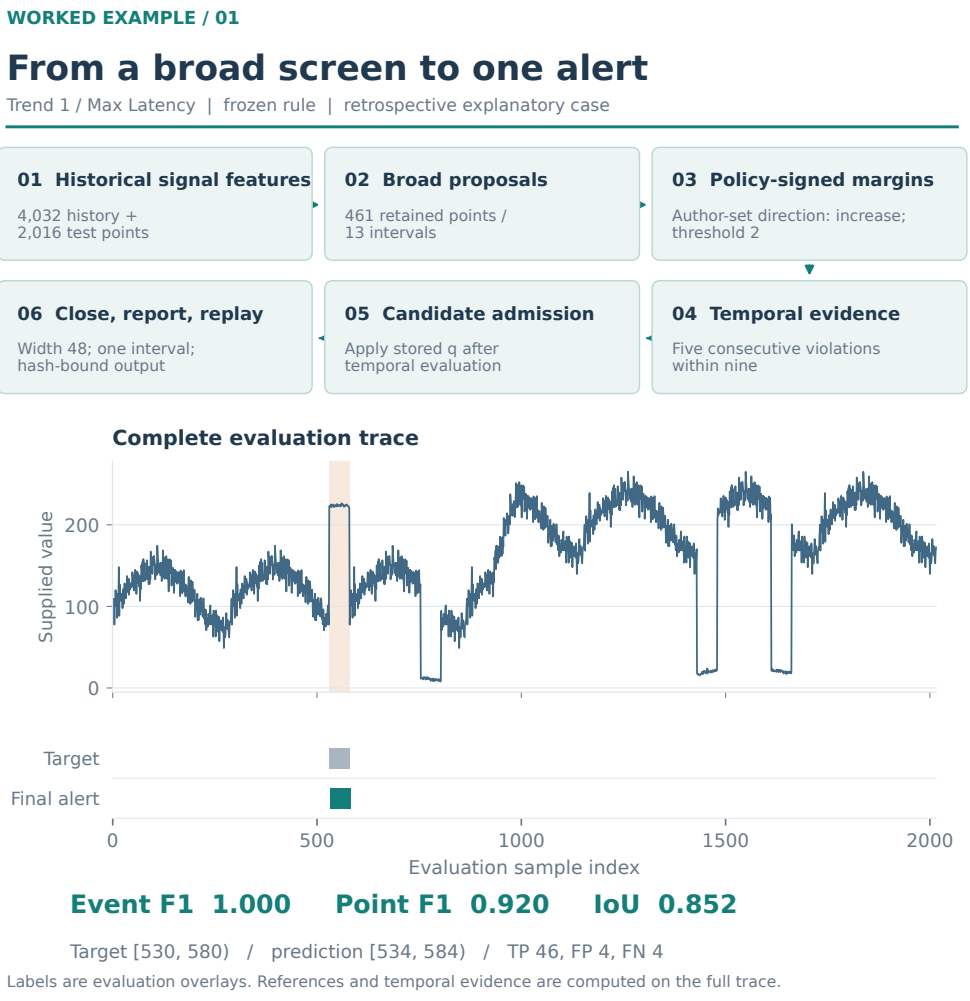


Figure 28: Primary worked example: Trend scenario 1 / Max Latency. The pipeline and complete evaluation context retain the supplied numeric scale. The label $[530, 580)$ and prediction $[534, 584)$ overlap for 46 observations and have a 54-observation union. Numerical counts report the four initial misses and four extra tail observations; Figure 32 enlarges these boundaries.

Throughout this case study, horizontal coordinates are zero-based evaluation observation indices and interval bounds are half open. Raw-signal panels retain the supplied numeric scale; physical units are undocumented. Label overlays assess the frozen output and are absent from the numerical inference inputs. All primary-case views use the same stored series, training history, and selected settings.

F.2 Step 2: Propose broad deviations without choosing their meaning

The proposal score combines nine absolute features: rolling deviations at windows 12, 48, 144, and 288; a seasonal deviation; a first difference; two mean-shift features; and a global-reference deviation. Historical empirical CDFs put these magnitudes on percentile scales after the specified upper-one-percent reference trimming. Their maximum forms the union score. At the selected threshold 1.0, a point can qualify by reaching the upper end of any retained historical reference; 1.0 does not imply an empty proposal.

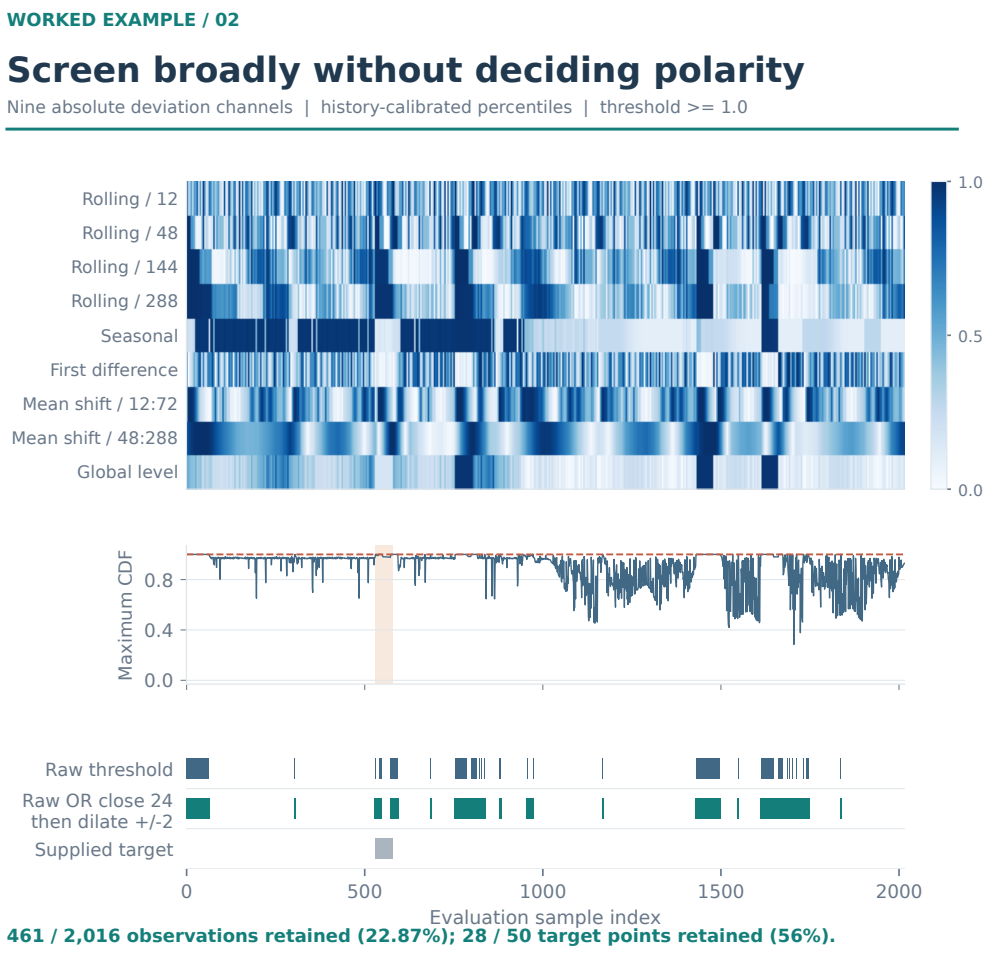


Figure 29: Direction-free proposal construction on the primary trace. The nine calibrated feature magnitudes contribute through a maximum, followed by thresholding, width-24 proposal closing unioned with the original triggers, and radius-two dilation. The resulting mask contains 461 of 2,016 positions, arranged in 13 intervals, including 28 of the annotation's 50 observations.

The proposal stage retains 22.867% of observations and 56% of labeled points, preserving access to the event within a concentrated candidate set. Its direction-free union admits evidence from several temporal scales before policy meaning is applied. Reference and persistence calculations retain the complete numerical history and trace, so selective admission can use full contextual evidence. Retention measures candidate workload; the runtime benchmark separately measures execution cost.

F.3 Step 3: Turn an explicit policy into signed evidence

The author-supplied contract for Max Latency declares increases harmful. With the frozen threshold $\tau = 2$, each reference contributes an atomic margin $a_t^u = x_t - b_t^u - 2\sigma_t^u$. A positive margin means the observation exceeds that reference's adverse-direction boundary. The rolling and seasonal references retain separate margins and scales; the numerical margin is not a probability. This step gives meaning to the direction-free deviations admitted by the proposal stage.



Figure 30: An explicit policy and its numerical consequences. The reference AST supplies direction, references, threshold, persistence, admission, and closing. Lower panels display the maximum atomic margin for a direction comparison; execution applies persistence separately to each reference before taking their maximum. The policy is an author-set experimental input, not a fresh language-model compilation.

Correct polarity converts the five intervals of the absolute-residual control into one matched alert, raising strict F1 from 0.333 to 1.000. The reversed-direction control provides an additional check on this policy dependence. With the waveform and all remaining settings fixed, the comparison demonstrates the benefit of assigning explicit operational meaning to numerical deviations.

F.4 Step 4: Require a sustained run, then inspect its interaction with admission

Persistence requires some run of five consecutive positive atomic margins within the latest nine observations. It is evaluated separately for each reference, and the two persisted margins are then combined by a maximum. Five scattered positive observations are insufficient. Conversely, a recently completed qualifying run can keep the temporal verdict positive after the current atomic margin has ceased to be positive. These semantics explain the distinction between instantaneous evidence and the later temporal verdict.

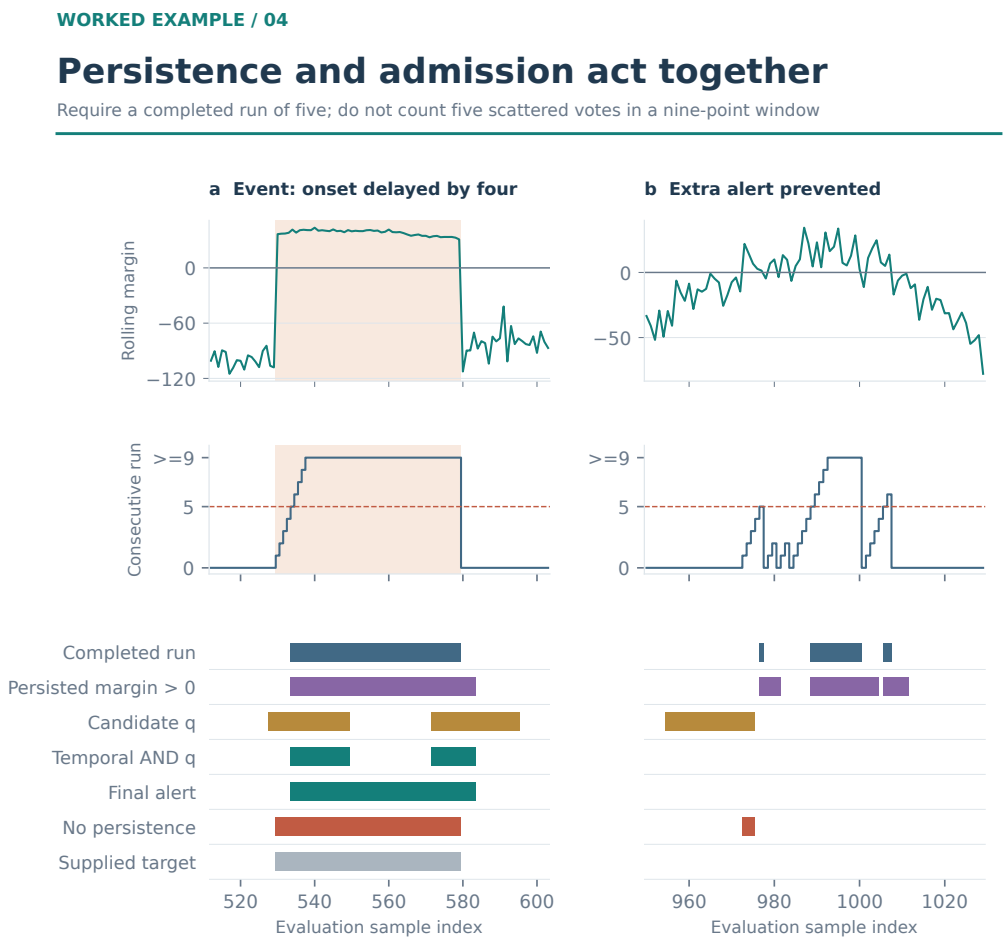


Figure 31: Persistence and admission on the labeled event and the diagnostic region $[950, 1030)$. Without persistence, the latter region yields the extra output $[973, 976)$. With persistence, the temporal verdict first becomes positive at index 977, after the local proposal interval $[955, 976)$ ends. This is an interaction between temporal qualification and admission, not evidence that the entire underlying atomic departure lasts only three observations.

The five-within-nine condition filters the extra three-observation interval $[973, 976)$ and retains one matched event, raising strict event F1 from 0.667 to 1.000. This is the benefit of evaluating temporal qualification together with candidate admission: a numerical departure becomes an alert only when its sustained evidence and admitted support align. Table 9 reports the accompanying pointwise measurements.

F.5 Step 5: Preserve interval extent through bounded closing

The candidate mask acts after persistence: excluded margins become $-\infty$. Before output closing, the positive support consists of $[534, 550)$ and $[572, 584)$, containing 28 observations in total. Their 22-observation gap, $[550, 572)$, lies inside the supplied annotation. Width-48 numerical closing joins these fragments into $[534, 584)$, restoring 22 labeled positions that the admission mask excluded.

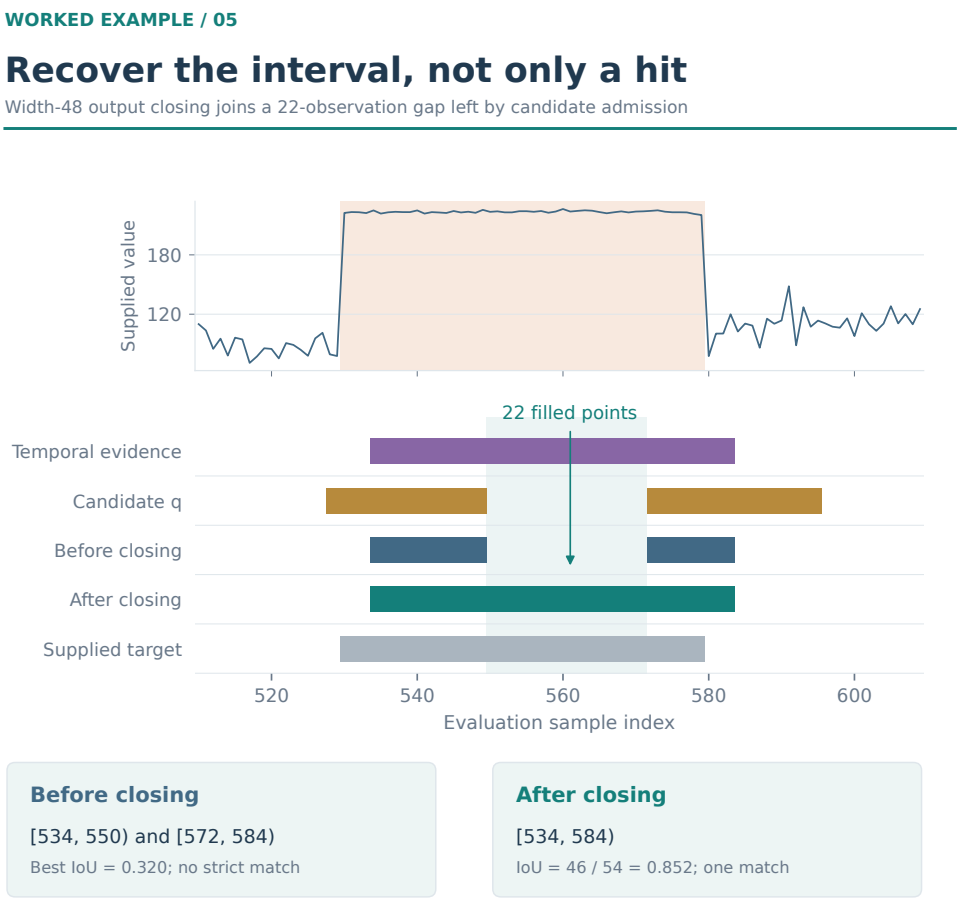


Figure 32: Exact support accounting through admission and output closing. The pre-closing fragments contain 24 labeled and four unlabeled observations. Closing fills the 22-position separation, producing 46 labeled and four unlabeled observations in one interval. Without closing, neither fragment reaches IoU 0.5; with closing, the combined interval reaches IoU 0.852.

Bounded closing restores continuity across qualified fragments, increasing labeled support from 24 to 46 observations and producing a single interval with IoU 0.852. This explains how broad screening and interval reconstruction work together: the gate admits event evidence, and closing restores its connected extent. The symmetric operator uses numerical support with a fixed width, making the complete evaluated detector an offline procedure.

F.6 Complementary evidence from the reference paths

The supporting example, Wave scenario 4 / Success Rate, demonstrates the value of making both reference paths available. Under the locally decrease-is-harmful policy, the seasonal path supports strict event F1 1.000 and point F1 0.966 in the full rule, compared with 0.000 and 0.099 for rolling-only execution. The rolling path supports the accurate primary Trend result. Together, the examples show how complementary reference contexts let the same rule architecture respond to different signal structures.

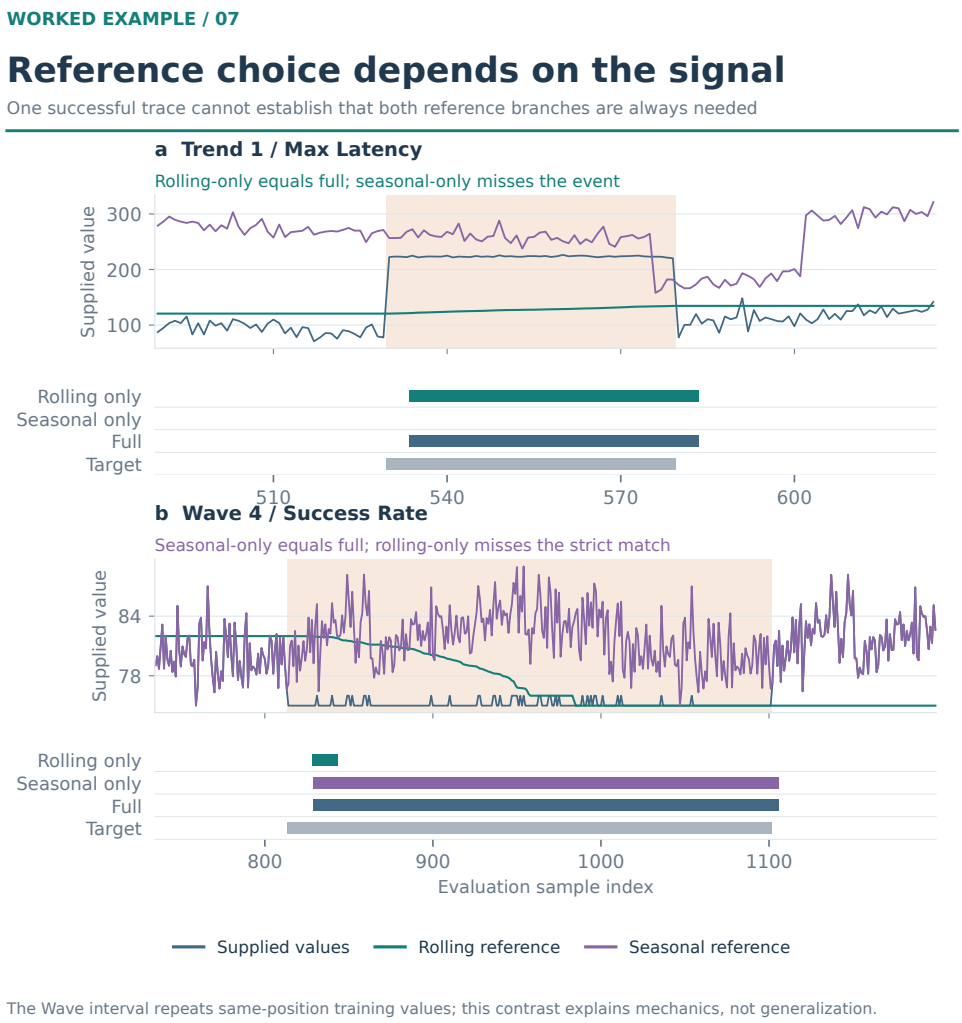


Figure 33: Complementary applicability of the numerical references. Primary Trend 1 / Max Latency succeeds through the rolling path; supporting Wave 4 / Success Rate succeeds through the seasonal path. The latter has label $[814, 1102)$ and full prediction $[829, 1106)$. The two paths provide alternative contextual evidence within the same executable rule.

The union accepts a qualifying path from either reference, preserving the useful evidence supplied by each context. The supporting Wave interval exactly reuses same-position training values, making it a transparent illustration of the recorded seasonal-reference behavior. These paired traces connect the dual-reference design to its intended benefit: contextual coverage across the supplied Trend and Wave patterns.

Evidence and replay. All 18 archived case–configuration reconstructions reproduce the frozen predictions, scores, and direct monitor outputs. This exact replay links the visual explanation to

executable evidence, allowing each highlighted design contribution to be checked against the stored computation.

G Language Evaluation: Historical Audit and a New Test Design

G.1 Existing outcomes and their denominators

The controlled-English development suite contains 48 executable cases and 16 required rejections. Executable cases form 24 pairs of opposite policies sharing fixed waveform probes. Historical semantic case IDs encoded target-relevant attributes and a common rejection prefix. The revised model-facing contract removes those attributes, uses independent opaque identifiers if a join key is needed, shuffles cases, and keeps evaluator fields client-side. No fresh LLM run on that revised interface is represented here.

System	Valid	False accepts	Outcome	Pairs
Archived compiler (rescored)	12/48	0/16	28/64	6/24
Archived reviewer (rescored)	10/48	0/16	26/64	5/24
Deterministic policy/schema parser	48/48	0/16	64/64	24/24

Table 10: Development diagnostics. FA denotes false acceptance. The reviewer attempts 40 supported cases and rejects eight; ten of the 40 attempted ASTs validate. The rows are not a matched fresh-model benchmark, and the parser shares the author-templated evaluation language.

The original compiler output was 48/48 under a legacy schema. Requiring previously omitted nullable bound keys makes 36 of those saved ASTs invalid. That change measures backward compatibility under a stricter interface, not a fresh estimate of the model's ability to follow the current schema. Separately, the original metadata confound prevents attributing the old perfect result to policy comprehension. A fair current comparison must rerun every method under the same schema and information budget.

A nine-policy local integration archive reported 63/63 reproduced outputs under the old schema. Re-evaluation under the current schema accepts only 2/9 policies and fully executes 15/63 series. One additional prediction is all-negative and matches vacuously, giving 16/63 prediction-array identity. These counts should never be collapsed into a successful semantic pipeline score.

G.2 Independent policy suite: specified, not collected

A confirmatory suite should contain independently authored operational policies, with source-service groups sealed before evaluation. Two annotators should independently specify supported semantics, uncertainty, and executable target rules; disagreements should be adjudicated before model outputs are examined. Numeric probes should be generated from adjudicated semantics, not from the model's own predictions. Ambiguous policies should have rejection or clarification targets rather than guessed programs.

The key factorial tests are: same waveform with different contexts; same policy with different waveforms; meaning-preserving paraphrases; polarity conflicts; missing unit or sampling interval; unsupported cross-metric conditions; deployment-state revisions; adversarially irrelevant text; and changes of domain. Context ablations must distinguish metric-name removal, explicit-policy removal, and policy contradiction. They test different information failures.

Equal-information controls should include a deterministic slot parser, a richer grammar parser, direct structured LLM generation, the same LLM with static checks, and the same LLM with execution probes. A separate dense-versus-candidate LLM comparison should hold model snapshot, prompt content, candidate evidence, and token budget fixed as far as possible. Reference-AST execution is an oracle-interface upper control, never a learned result.

Every rejection, timeout, malformed AST, runtime failure, and absent answer remains in the denominator. Report false rejection among supported policies, false acceptance among unsupported policies, semantic error conditional on acceptance, coverage, overall outcome, paired context consistency,

and downstream miss rate. Repeat model runs using frozen snapshots and retain exact payloads and caches. The existing independent-annotation template is empty, so none of these prospective endpoints has a measured value.

H Baseline Implementations and Historical Archive

Main-table layout and metric contract. Table 1 assigns one method to each row and three metric columns to each of Mutation, Step, Trend, and Wave. It contains 34 baselines and VeraTS, giving 35 rows and 420 metric values. The baselines comprise four statistical, eight prediction, eight reconstruction, eight foundation-model, and six LLM rows. TFC and TimeXer appear under prediction, SEA-FGT under reconstruction, and Timer-XL under foundation models. The LLM rows are kimi-k3, glm-5.3, Qwen-3.8, step-5, DeepSeek-v4, and GPT-6; step-5 is the display name of the archived step-5-preview adapter. The four archived API adapters receive each complete unlabelled evaluation trace and metric name. GPT-6 returns a separate rating for every point, while the other three return scored intervals. These offline adapters do not share the information budget of causal forecasting or train-only baselines; exact requested/returned IDs also do not independently verify the serving snapshot. This display grouping does not imply that TimesNet, FITS, and TimeMixer are pretrained, nor that every adapter retains its original detection protocol. Pretraining, future context, adapters, and availability are recorded separately. AnomalyTransformer and OFA are excluded from the current comparison. Table 1 is the authoritative manuscript snapshot; Qwen-3.8 and DeepSeek-v4 are supplied result rows without complete local score and evaluation records in this release. No single-run result receives an invented standard deviation.

GPT-6 pointwise scoring. We request and receive the exact API ID `gpt-6` for all 63 series, using one complete 2,016-point unlabelled trace plus its metric name per request, with no training values, labels, morphology annotations, or other detector outputs. To meet the requested one-minute response budget, the model emits one integer rating in 0–9 for each point, serialized as 63 strings of 32 digits; each rating is divided by 9 to produce the final ten-level score. No padding, interpolation, smoothing, or interval expansion is applied. Requests are strictly serial, with the minimum supported reasoning effort (`low`), a 2,048-token completion limit, and a 60-second response deadline. Two series require format-only repairs after three malformed replies; strict structured output enforces the original length contract during these repairs. All 66 responses take 33.0–44.6 seconds (median 35.7 seconds). The complete 127,008-point score artifact is verified before evaluation labels are opened, and the same VUS and Event-F1* protocols are then applied.

VUS is computed per series with the pinned TSB-AD evaluator, 250 thresholds, and a rank-1 tolerance derived from training values, then averaged within a family. Our VUS uses the corrected full-method scores, including the stored -10^6 masked sentinel, without recombining old and new outputs. The main table colors the largest and second-largest distinct positive values after three-decimal rounding; displayed ties share a color and zeros remain neutral. These descriptive ranks do not control for different information budgets, missing methods, or dependence between series.

Preserved q95 diagnostic. The historical adapters save continuous evaluation scores but generally lack the training scores, binary outputs, or checkpoints needed to recover a common training-calibrated operating point. The first common re-evaluation therefore used an explicitly post hoc, unlabelled full-trace diagnostic for every method, including VeraTS:

$$\theta_{95} = \text{Quantile}_{\text{higher}}(s_{1:T}, 0.95), \qquad \widehat{y}_t^{95} = \mathbf{1}\{s_t > \theta_{95}\}. \tag{33}$$

The quantile is fixed without optimizing against evaluation labels. All ties at the threshold are rejected; constant scores produce no alerts. We apply no closing, dilation, or point adjustment. This rule retains at most 5% of points, using future unlabelled scores to set the threshold, so it is an offline score diagnostic rather than a deployable calibration procedure. The archived-score re-evaluation seals scores and masks before opening labels. New neural adapters seal continuous scores first; their event evaluator can load labels before constructing the mask, but the shared mask function accepts only scores and the fixed quantile. Neither ordering erases prior exposure to the retrospective corpus.

Event-F1^{95} uses one-to-one IoU-0.5 interval matches and pools counts within a family: $2\sum_i M_i/(\sum_i G_i + \sum_i P_i)$, where M_i, G_i, P_i denote matched, true, and predicted interval counts. VUS is a per-series macro mean; event F1 is a count-based micro summary. Its sparse alert budget

can make long intervals impossible to match despite useful ranking. The original q95 arrays and table snapshot remain frozen. The revised main table instead uses the EasyTSAD event-PA definition and the explicitly label-selected q90–99 grid in Appendix L; it does not rewrite the strict q95 diagnostic or VUS. Our original frozen-output event F1 of 0.679 remains in the matched-rule ablation table, with its original operating point.

H.1 Baseline identities and artifact availability

The current main table contains 34 baselines. This inventory also retains TFAD and CAAD as unavailable historical requests; they are not numerical rows of Table 1. Coverage denotes archived score availability, not validation of every supplied table entry.

Method	Coverage	Identity and information contract
MatrixProfile (Zhu, 2018)	63/63	full trace unsupervised.
SAND (Boniol et al., 2021)	63/63	online full trace.
SubLOF (Breunig et al., 2000)	63/63	full trace unsupervised.
KMeansAD (Liu & Paparrizos, 2024)	63/63	Native TSB-AD implementation, 20 clusters and window 20; full unlabelled trace used for fitting; overlapping-window scores use future context.
AR (Rousseeuw & Leroy, 2003)	63/63	Local statsmodels AutoReg(32); causal one-step residuals. The cited book is background, not an original implementation.
LSTMAD (Malhotra et al., 2015)	63/63	full test unlabeled.
POLY (Li et al., 2007)	63/63	Pinned TSB-AD local-polynomial adapter; two-sided neighborhoods and full-trace scaling.
KAN-AD (Zhou et al., 2025)	63/63	train values only.
TShape (Cui et al., 2025)	63/63	train values only.
DLinear (Zeng et al., 2023)	63/63	Train-only architectural reconstruction adapter; not the original forecasting task.
TFC (Cui et al., 2026c)	63/63	Official temporal–frequency–curvature fusion; native first differences, batch-coupled forecasting scores, early stopping and a 1,000-epoch cap.
TimeXer (Wang et al., 2024b)	63/63	Official TimeXer forecasting model, fitted per series; native parallel branch with one variable and no independent exogenous inputs; causal one-step residuals.
AE (Ng, 2011)	63/63	TSB-AD autoencoder with an equivalent per-window scaling repair at inference; the cited lecture notes do not uniquely specify this detector.
EncDecAD (Malhotra et al., 2016)	63/63	Third-party DeepADoTS reverse-LSTM/Gaussian adapter, with 10 epochs per series; all-containing-window averaging uses future context.
SRCNN (Ren et al., 2019)	63/63	unsupervised train only.
TFAD (Zhang et al., 2022a)	Pending	Official Yahoo architecture and 100-epoch budget; checkpoint selection uses training-partition validation labels. Test scores use the common unadjusted metrics. No independently verified complete 63-series result is available in this snapshot.
TranAD (Tuli et al., 2022)	63/63	Transformer anomaly-score adapter; implementation and normalization are recorded separately.
Donut (Xu et al., 2018)	63/63	TSB-AD variational reconstruction adapter; empty-validation construction repaired while retaining default training-loss stopping.
FCVAE (Wang et al., 2024c)	63/63	Official FCVAE core, pooled across all 63 training series; native 30-epoch cap/patience 5 and per-partition normalization.
ModernTCN (Luo & Wang, 2024)	63/63	Official model with a fixed local univariate adapter: one block, width 8, kernels 51/5, patches 8/4 and 10 epochs; aligned right-end reconstruction scores.

Method	Coverage	Identity and information contract
SEA-FGT (Huang et al., 2026)	63/63	Frequency-guided transformer with learned Conv/MLP semantic experts (KDD 2026); no textual policies; overlapping-window scoring uses future context.
TimesNet (Wu et al., 2023)	63/63	full test unlabeled.
FITS (Xu et al., 2024)	63/63	full test unlabeled.
Sundial (Liu et al., 2025c)	63/63	Pretrained Sundial-128M causal forecasting-error adapter.
Timer (Liu et al., 2024b)	63/63	Original ICML 2024 pretrained Timer (2.16M); local zero-shot causal forecasting-error adapter, not the original fine-tuning experiment.
Timer-XL (Liu et al., 2025b)	63/63	Official pretrained Timer-XL; zero-shot causal forecast-error scoring with a fixed context and alignment protocol.
Time-MoE (Shi et al., 2025)	63/63	Pretrained Time-MoE-50M causal forecasting-error adapter.
TimeMixer (Wang et al., 2024a)	63/63	Train-only architectural reconstruction adapter; not a pretrained foundation-model result.
Time-o1 (Wang et al., 2025)	63/63	DLinear with the Time-o1 loss adapter; not an independently pretrained model.
kimi-k3	63/63	Exact requested and returned API ID; one complete unlabelled test trace plus metric name per request; model intervals are expanded by maximum stated severity, with zero elsewhere. Serving snapshot is not independently verified.
glm-5.3	63/63	Exact requested and returned API ID; one complete unlabelled test trace plus metric name per request; model intervals are expanded by maximum stated severity, with zero elsewhere. Serving snapshot is not independently verified.
step-5-preview	63/63	Exact requested and returned API ID; one complete unlabelled test trace plus metric name per request; model intervals are expanded by maximum stated severity, with zero elsewhere. Serving snapshot is not independently verified.
GPT-6	63/63	Exact requested and returned API ID gpt-6; one complete unlabelled test trace plus metric name per request, executed serially with a 60-second deadline and minimum supported reasoning effort. Every one of the 2,016 integer ratings in 0–9 is explicitly generated and divided by 9, without interpolation or interval expansion. Serving snapshot is not independently verified.
CAAD (Wang et al., 2026)	Unavailable	Multivariate causal-consistency detector (KDD 2026); no author code located, and no faithful single-channel score established. No independently verified complete 63-series result is available in this snapshot.
Qwen-3.8	Supplied row	Current main-table values are retained as supplied; complete local score arrays and the evaluation protocol are not available in this release.
DeepSeek-v4	Supplied row	Current main-table values are retained as supplied; complete local score arrays and the evaluation protocol are not available in this release.

For archived implementations, VUS is evaluated from local score artifacts rather than transcribed from the cited papers. Supplied result rows are identified separately above. New adapter settings and modifications are retained with their source hashes. Existing archived implementations retain their historical training and information budgets.

New implementations and future context. AR is a local statsmodels AutoReg(32) adapter with causal one-step residuals, rather than an original implementation from the cited robust-regression book. POLY uses the pinned TSB-AD implementation, including two-sided neighborhoods and full-trace scaling. AE retains the pinned architecture and its per-window min–max transform; its inference adapter repairs a scaler binding to the number of training windows. Its centered-window score uses future observations. Donut repairs empty-validation dataset construction while preserving the na-

tive training-loss stopping branch and epoch cap. Donut and TranAD retain normalization over the complete inference trace. These are offline score adapters, not causal monitoring implementations.

EncDecAD uses the third-party DeepADoTS reverse-LSTM encoder–decoder and Gaussian reconstruction-error score. Its averaging over overlapping windows can include future observations. The archived run records installed NumPy 2.5.3 metadata; packages were subsequently changed by other jobs. A separately recorded restored environment with loaded NumPy 1.26.4 replays all 63 checkpoints and reproduces every score and q95 mask bit for bit. This validates score reconstruction, not a replay of the entire training environment.

FCVAE uses the original author implementation's frequency-conditioned VAE core, loss, augmentation, and Monte Carlo scoring, with device and orchestration adaptations recorded in its protocol. One pooled model is fitted across the 63 series, with each designated training sequence split 70/30 for fitting and validation. This run completed the native 30-epoch budget with patience five and produced all 63 score artifacts. Partition-specific scaling includes unlabelled test means and variances; a training-history prefix supplies complete right-aligned evaluation windows. TFAD remains pending a complete 63-series run; probes and partial fits are excluded from the numerical comparison. Recorded default budgets, device ports, and resumable commands accompany the source package.

A directly related semantic artifact. The public Smart Brain artifact (Cui et al., 2026b), associated with a paper listed in the ASE 2026 research program, already connects four types of temporal features to metric-dependent rules and optional LLM diagnosis. This overlap limits any claim that feature-to-semantic processing itself is new. Our evaluated questions concern recall-constrained proposals, typed policy execution, initialization, and preservation of interval support. The comparison here covers the available author code; the associated paper's full text was not retrieved, so absence from the inspected artifact is not evidence that the paper lacks an idea. The source-package memo records this boundary.

Additional statistical and convolutional baselines. KMeansAD uses the pinned TSB-AD wrapper with 20 clusters, 20-point windows, stride one, native window standardization, and mean-overlap score reconstruction. It fits on the concatenated designated training and unlabelled evaluation trace. This transductive information budget and future-context access are disclosed; it is not a causal train-only baseline. All 63 saved-centroid scores are independently reconstructed (maximum absolute error 3.11×10^{-15}), and VUS and q95 are independently replayed.

ModernTCN uses the official ICLR 2024 detection model with a fixed local univariate adapter: one block, eight channels, kernels 51 and 5, 100-point windows, patch size eight, stride four, batch size 128, and ten epochs. The designated training sequence is split 80/20 for fitting and validation; the scaler uses that entire designated sequence. Checkpoint selection uses validation loss. The original code's unlabelled evaluation-loss telemetry is retained, but evaluation labels are unavailable during training. The adapter returns the last-point reconstruction error of each right-aligned window, using training history at the boundary, instead of flattening repeated overlapping windows. The official learning-rate update and tied-checkpoint behavior are preserved. The local ten-epoch, learning-rate 0.0003 configuration is not asserted to be a published univariate default; the official MSL script instead specifies two epochs and 0.0005. Full score and checkpoint replays accompany admitted results.

Recent KDD 2026 identities and compatibility. TFC is the temporal–frequency–curvature fusion method of Cui et al. (2026c), with official source `https://github.com/CSTCloudOps/TFC`. Its 63-series run retains the native first-difference preprocessing, fit-partition min–max scaling, 16-point prediction window, batch size 128, 1,000-epoch cap, and patience-30 early stopping. The native implementation uses the final epoch rather than restoring a best checkpoint. Its Transformer treats the minibatch axis as a sequence, coupling scores across windows; batch order and size are therefore preserved. This author-code behavior is disclosed rather than silently repaired. All VUS and q95 metrics pass independent replay. Across q90–99 its archived strict IoU event F1 is zero in each family. Its revised EasyTSAD event-PA scores are positive: scattered hits can detect an event without recovering its full interval. The result evaluates this local adapter and interval metric, not the method's published benchmark performance.

SEA-FGT is Huang et al. (2026), with official source `https://github.com/Karwei0/SEA-FGT`. Its semantic experts are learned feature-space augmenters, not natural-language operator-policy judgments. We retain three layers of width 256, ten Conv/MLP experts, the native 10^{-9} gradient-clipping limit, and the 20-epoch cap with early stopping. The first 3,226 training points fit the model and scaler; 806 points provide chronological validation. The one-channel adapter uses sparse selection $k = 1$, where the cross-variable relation degenerates. Native 96-point windows, stride 24, and mean overlap aggregation produce aligned dense scores from $|\mathrm{KL}(p\|q) - \mathrm{KL}(q\|p)|$ between the two feature distributions. Partial batches are retained. A point's score can use up to 95 subsequent observations through its covering windows, so this is an offline adapter.

The execution port keeps FFT and discrete expert routing on CPU and dense FGT computation on MPS. It disables automatic differentiation's per-operator debugging checks while retaining finite-loss and finite-gradient guards. Three paired full training steps reproduce losses, gradients, parameter updates, and Adam states exactly; six complete 20-epoch pairs also reproduce their validation-loss histories and all $6\times2{,}016$ scores exactly. These checks do not assert a second complete training replay for every series. All 63 sequences complete the same diagnostic-off protocol, retaining the original model, optimizer, early stopping, and evaluation calls. Independent recomputation reproduces all 63 VUS pairs and q95 masks exactly. Source hashes and modifications are recorded; probes are excluded.

The supplied DOI `10.1145/3770855.3817978` identifies CAAD (Wang et al., 2026), a separate multivariate method. No author implementation was located in this audit. Its cross-variable causal-consistency score cannot be transferred unchanged to a single channel by treating unrelated traces as synchronized variables. CAAD therefore remains explicitly unavailable; a simplified substitute would not constitute a faithful result.

Long-context and exogenous-input forecasting adapters. Timer-XL (Liu et al., 2025b) uses the native OpenLTM model and its linked PyTorch checkpoint, with 67,398,880 parameters: eight layers, width 1,024, eight heads, feed-forward width 2,048, ReLU activation, and 96-point tokens. We retain the native context normalization and choose 2,880 context points before evaluation. Twenty-one nonoverlapping 96-point forecasts cover each evaluation sequence. Each forecast uses only observations before its origin, including the available training history and previously observed evaluation values. Raw-scale squared residuals provide dense anomaly scores; there is no local fitting or test-label-dependent model choice. The checkpoint is reported by the authors as pretrained on 260 billion points; we cannot audit membership in that pretraining corpus.

The historical archive labeled the later Hugging Face `timer-base-84m` release as Timer. The Timer-XL repository also links that release, and the Timer-XL paper distinguishes a base variant; treating the old label as an independent original-2024 result would conflate versions. The revised Timer row uses the original ICML 2024 model with 2,158,176 parameters and the author-released anomaly-detection checkpoint: four layers, width 256, feed-forward width 512, and 96-point tokens. We evaluate it zero-shot with 672 past points predicting the next 96; this is our causal forecast-residual adapter, not the paper's downstream fine-tuning experiment. The two Timer versions have separate architectures, weights, and source hashes. The old 84M measurements remain only in the historical archive. This correction is based on model identity, not comparative scores.

TimeXer (Wang et al., 2024b) normally incorporates exogenous variables, which this univariate corpus does not supply. We use its official parallel forecasting branch with one variable and no timestamp marks. The observed history forms both endogenous patches and a same-series variate token; no independent external signal is invented. Cross-attention retains its single-key softmax special case, which cannot test the benefit of external covariates. A fixed local adapter uses a 96-point context, one-step forecasts, one layer, width 256, eight heads, feed-forward width 2,048, patch length 16, dropout 0.1, batch size 32, Adam at 10^{-4}, and the paper's ten-epoch cap with patience-three early stopping. The first 3,225 designated training points fit the model and scaler; the remaining 807 targets select the checkpoint. Squared residuals in fitting-standardized units use only preceding observations. All 63 checkpoint score replays are bitwise identical. Its archived strict IoU Event-F1 is zero across q90–99 in all four families, whereas the revised EasyTSAD event-PA results are positive. Event detection and interval localization answer different questions. The one-step output and single-variable setting are local adaptations, not the original multivariate forecasting benchmark.

The following tables preserve the previous requested-adapter archive for transparency. They evaluate 33 requested method identities, with 26 complete numerical adapters and seven incompatible or incomplete entries; the additional VeraTS row is a legacy reference-AST artifact. These measurements use a frozen TSB-AD VUS evaluator with 250 curve thresholds, train-derived tolerance, no point adjustment, and equal weighting of the four family means. They are not recomputations of the new initialized experiment.

Adapters have different information and future-context budgets, as marked below. Native forecasting output is not automatically a valid dense anomaly score; missing compatible outputs are left unavailable. The legacy VeraTS VUS row below remains labeled R. The main table separately recomputes VUS under the corrected mask, initialized reference, and selection protocol; those new values do not validate the historical artifact or replace its provenance label.

Table 12: Requested score adapters and one legacy VeraTS artifact on four morphology families of one controlled KPI corpus (part 1 of 3). Each numeric cell is VUS-PR/VUS-ROC. Values appear only for complete 63/63 score artifacts under VUS-250 without point adjustment; artifact completeness does not imply a currently valid operating point. The requested rows are locally executed score-adapter outputs, and VeraTS is a locally executed legacy artifact; they are not results copied from the cited papers. Macro weights the four family means equally.

Group	Method	State	Mutation (9)	Step (9)	Trend (36)	Wave (9)	Macro
Stat.	SR-CNN (Ren et al., 2019)	OK	0.068/0.654	0.097/0.748	0.053/0.640	0.153/0.566	0.093/0.652
	TFAD (Zhang et al., 2022a)	A	–	–	–	–	–
	Subseq.-LOF adapter† (Breunig et al., 2000; Liu & Paparrizos, 2024)	OK	0.052/0.653	0.039/0.510	0.101/0.595	0.140/0.461	0.083/0.555
	SAND† (Boniol et al., 2021)	OK	0.167/0.600	0.423/0.802	0.075/0.481	0.166/0.552	0.208/0.609
	Matrix Profile† (Zhu, 2018)	OK	0.110/0.656	0.356/0.897	0.199/0.803	0.172/0.521	0.209/0.719
Pred.	DLinear+Time-o1 (Wang et al., 2025)	OK	0.118/0.752	0.109/0.708	0.094/0.673	0.110/0.400	0.108/0.633
	TShape (Cui et al., 2025)	OK	0.066/0.624	0.093/0.694	0.066/0.568	0.156/0.579	0.095/0.616
	LSTM-AD MSE adapter† (Malhotra et al., 2015)	OK	0.222/0.839	0.177/0.742	0.130/0.669	0.108/0.377	0.159/0.657
	KAN-AD (Zhou et al., 2025)	OK	0.072/0.670	0.092/0.689	0.070/0.618	0.109/0.395	0.086/0.593

Table 13: Requested score-adapter results (part 2 of 3; continued from Table 12).

Group	Method	State	Mutation (9)	Step (9)	Trend (36)	Wave (9)	Macro
	FITS† (Xu et al., 2024)	OK	0.276/0.901	0.270/0.873	0.207/0.846	0.085/0.209	0.210/0.707
	FiLM§ (Zhou et al., 2022a)	OK	0.053/0.631	0.087/0.701	0.051/0.612	0.100/0.365	0.073/0.577
	FEDformer§ (Zhou et al., 2022b)	OK	0.402/0.967	0.311/0.938	0.313/0.762	0.214/0.670	0.310/0.834
	TimesNet† (Wu et al., 2023)	OK	0.162/0.867	0.236/0.804	0.161/0.737	0.090/0.232	0.162/0.660
Recon.	LightTS-AD§ (Zhang et al., 2022b)	OK	0.067/0.669	0.099/0.685	0.065/0.609	0.106/0.373	0.084/0.584
	Autoformer§ (Wu et al., 2021)	OK	0.449/0.969	0.358/0.943	0.328/0.791	0.238/0.680	0.343/0.846
	DLinear§ (Zeng et al., 2023)	OK	0.056/0.633	0.071/0.681	0.055/0.598	0.105/0.398	0.072/0.578
	Crossformer§ (Zhang & Yan, 2023)	OK	0.245/0.920	0.256/0.910	0.215/0.750	0.131/0.536	0.212/0.779
	iTransformer§ (Liu et al., 2024a)	OK	0.072/0.695	0.105/0.719	0.074/0.651	0.099/0.344	0.088/0.602
	Informer§ (Zhou et al., 2021)	OK	0.218/0.800	0.144/0.760	0.174/0.585	0.101/0.302	0.159/0.612
	MICN§ (Wang et al., 2023)	OK	0.108/0.748	0.150/0.742	0.106/0.702	0.104/0.370	0.117/0.640
	Pyraformer§ (Liu et al., 2022)	OK	0.110/0.700	0.082/0.630	0.063/0.536	0.115/0.429	0.092/0.574
	Reformer§ (Kitaev et al., 2020)	OK	0.304/0.921	0.232/0.896	0.191/0.708	0.218/0.666	0.236/0.798
	PatchTST† (Nie et al., 2023)	OK	0.088/0.825	0.240/0.855	0.161/0.805	0.085/0.171	0.144/0.664
	TimeMixer§ (Wang et al., 2024a)	OK	0.217/0.806	0.257/0.834	0.175/0.789	0.107/0.386	0.189/0.704

Table 14: Requested score-adapter results (part 3 of 3; continued from Table 12).

Group	Method	State	Mutation (9)	Step (9)	Trend (36)	Wave (9)	Macro
TSFM	FPT/OFA (Zhou et al., 2023)	P	–	–	–	–	–
	Time-MoE-50M (Shi et al., 2025)	OK	0.298/0.931	0.285/0.853	0.161/0.772	0.245/0.662	0.247/0.805
	Timer-84M (Liu et al., 2024b)	OK	0.311/0.943	0.281/0.905	0.174/0.817	0.285/0.702	0.263/0.842
	Sundial-128M (Liu et al., 2025c)	OK	0.291/0.923	0.300/0.858	0.173/0.786	0.280/0.656	0.261/0.806
LLM	Time-RA forecast-MSE (Yang et al., 2026)	A	–	–	–	–	–
	ChatTS forecast-MSE (Xie et al., 2025)	A	–	–	–	–	–
	TSAD-Agents-style forecast-MSE (Xu et al., 2026)	A	–	–	–	–	–
	TimeReasoner forecast-MSE (Cheng et al., 2026)	A	–	–	–	–	–
	STRATA-TS-96 forecast-MSE (Jiang et al., 2025)	A	–	–	–	–	–
	VeraTS (legacy reference AST; gated)‡	**R**	0.933/0.986	0.954/0.988	0.681/0.971	0.716/0.782	0.821/0.932

Information budget: † uses the complete unlabeled test trace, future context, or a full-trajectory transformation during scoring. § denotes the common pooled train-only, right-aligned, no-future TSLib architectural-adapter protocol; unmarked SR-CNN and KAN-AD are native train-only adapters. ‡ marks VeraTS, which receives an author-set reference AST and a training-label-selected operating point under a non-equal information budget; its symmetric CLOSE operator is offline. **State:** OK = complete locally executed adapter run; R = archived 63/63 VeraTS score artifact whose 576-point selection is partially initialized and must be rerun under the corrected 2,304-point guard; A = faithful dense-score path unavailable; O = native output is not a threshold-independent point score; P = gated by protocol/runtime audit. Dashes are N/A, never imputed.

I Public Proposal Pilot: An External Boundary

The archived public integration pilot selects the first series of at most 50,000 points for each source/domain pair from a pinned TSB-AD-U list, using neither labels nor detector outputs for that selection. It contains 26 series, 391,799 points, 17 sources, and nine domains. Its reduced fast signal sieve is not bit-identical to the initialized implementation in the main text.

The pilot achieves macro literal event recall 0.902, micro event recall 0.630, and point coverage 0.364. The global-median comparison reaches macro recall 0.811 at coverage 0.246 and is better on several ranking, micro-F1, and throughput endpoints. No matched-coverage advantage was demonstrated. Because the pilot lacks operator policy, metric units, and polarity, it evaluates only numerical proposals. It cannot establish semantic anomaly detection or an official 350-series benchmark ranking.

A full external test should preserve source-disjoint splits, report per-source behavior, and supplement generic anomaly benchmarks with independently adjudicated context-bearing episodes. Data without a context variable cannot by itself identify the benefit of semantic conditioning. Conversely, a manually authored policy test with reused numeric traces cannot establish numerical transfer.

J Reproduction and Release Contents

The numerical experiment is generated by `scripts/hierarchical_experiments.py`, with table-generation utilities in `scripts/hierarchical_paper_tables.py`. The result directory `results/hierarchical_2027/` retains the complete archived experiment. The checked-in paper tables present the selected semantic-component comparisons described above. To compile the reported manuscript snapshot:

```
cd paper
tectonic -X compile VeraTS_ICLR2027_hierarchical.tex
```

Dependencies and exact hashes are recorded in the result manifests. Running the experiment again changes wall-clock timing and may regenerate manifests; the distributed report defines the measured snapshot. The source bundle contains editable figure scripts and vector outputs, rather than relying on screenshots of numerical plots. Source data redistribution remains subject to the unresolved license.

K Every Series in the New Hierarchy Experiment

Series	Event F1	Point R	Point F1	Alerts
mutation/Mutation_dataset / Avg Latency	1.000	0.920	0.920	1
mutation/Mutation_dataset / Failure Count	1.000	0.920	0.920	1
mutation/Mutation_dataset / Failure Rate	1.000	0.920	0.920	1
mutation/Mutation_dataset / Max Latency	1.000	0.940	0.931	1
mutation/Mutation_dataset / Request Count	1.000	0.920	0.920	1
mutation/Mutation_dataset / Retries	1.000	0.920	0.911	1
mutation/Mutation_dataset / Success Count	1.000	0.920	0.920	1
mutation/Mutation_dataset / Success Rate	1.000	0.920	0.920	1
mutation/Mutation_dataset / Timeouts	0.667	0.960	0.914	2
step/Step_dataset / Avg Latency	1.000	0.920	0.920	1
step/Step_dataset / Failure Count	1.000	0.920	0.920	1
step/Step_dataset / Failure Rate	1.000	0.920	0.920	1
step/Step_dataset / Max Latency	1.000	0.920	0.911	1
step/Step_dataset / Request Count	1.000	0.920	0.920	1
step/Step_dataset / Retries	1.000	0.920	0.902	1
step/Step_dataset / Success Count	1.000	0.920	0.920	1
step/Step_dataset / Success Rate	1.000	0.920	0.920	1
step/Step_dataset / Timeouts	1.000	0.960	0.941	1
trend/trend_dataset1 / Avg Latency	0.667	0.920	0.860	2
trend/trend_dataset1 / Failure Count	0.667	0.920	0.639	2
trend/trend_dataset1 / Failure Rate	0.667	0.920	0.681	2
trend/trend_dataset1 / Max Latency	1.000	0.920	0.920	1

Series	Event F1	Point R	Point F1	Alerts
trend/trend_dataset1 / Request Count	0.667	0.920	0.601	2
trend/trend_dataset1 / Retries	1.000	0.920	0.920	1
trend/trend_dataset1 / Success Count	0.667	0.920	0.601	2
trend/trend_dataset1 / Success Rate	0.667	0.920	0.601	2
trend/trend_dataset1 / Timeouts	0.667	0.920	0.458	2
trend/trend_dataset2 / Avg Latency	0.667	0.920	0.639	2
trend/trend_dataset2 / Failure Count	0.667	0.920	0.676	2
trend/trend_dataset2 / Failure Rate	0.667	0.920	0.676	2
trend/trend_dataset2 / Max Latency	0.667	0.920	0.579	2
trend/trend_dataset2 / Request Count	1.000	0.920	0.920	1
trend/trend_dataset2 / Retries	0.667	0.920	0.605	2
trend/trend_dataset2 / Success Count	1.000	0.920	0.920	1
trend/trend_dataset2 / Success Rate	0.667	0.920	0.681	2
trend/trend_dataset2 / Timeouts	0.000	1.000	0.104	1
trend/trend_dataset3 / Avg Latency	1.000	0.627	0.771	1
trend/trend_dataset3 / Failure Count	0.000	0.075	0.139	1
trend/trend_dataset3 / Failure Rate	0.000	0.463	0.633	1
trend/trend_dataset3 / Max Latency	0.000	0.209	0.346	1
trend/trend_dataset3 / Request Count	0.667	0.920	0.652	2
trend/trend_dataset3 / Retries	0.000	0.299	0.460	1
trend/trend_dataset3 / Success Count	0.667	0.920	0.667	2
trend/trend_dataset3 / Success Rate	1.000	0.920	0.920	1
trend/trend_dataset3 / Timeouts	1.000	0.940	0.926	1
trend/trend_dataset4 / Avg Latency	0.667	0.920	0.639	2
trend/trend_dataset4 / Failure Count	0.667	0.920	0.676	2
trend/trend_dataset4 / Failure Rate	0.667	0.920	0.676	2
trend/trend_dataset4 / Max Latency	0.667	0.920	0.579	2
trend/trend_dataset4 / Request Count	1.000	0.932	0.869	2
trend/trend_dataset4 / Retries	0.667	0.920	0.605	2
trend/trend_dataset4 / Success Count	1.000	0.906	0.934	2
trend/trend_dataset4 / Success Rate	1.000	0.889	0.924	2
trend/trend_dataset4 / Timeouts	0.000	1.000	0.083	2
wave/Wave_dataset2 / Request Count	0.000	0.179	0.297	2
wave/Wave_dataset2 / Success Count	0.667	0.542	0.703	2
wave/Wave_dataset2 / Success Rate	1.000	0.899	0.935	1
wave/Wave_dataset3 / Request Count	0.000	0.094	0.171	3
wave/Wave_dataset3 / Success Count	0.000	0.299	0.460	3
wave/Wave_dataset3 / Success Rate	0.667	0.708	0.826	2
wave/Wave_dataset4 / Request Count	0.000	0.073	0.136	2
wave/Wave_dataset4 / Success Count	0.000	0.201	0.335	1
wave/Wave_dataset4 / Success Rate	1.000	0.948	0.966	1

L EasyTSAD Event-F1: Definition and Threshold Sensitivity

The metric correction. The previous main-table event metric required one-to-one interval matches at IoU ≥ 0.5. That measures localization and is not EasyTSAD's EventF1PA. The archived metric-conversion study recomputes 29 complete methods from the same frozen continuous scores, using the pinned EasyTSAD source (EasyTSAD contributors, 2026). No models, score orientations, VUS values, or true intervals change within that study. The current main table adopts the library's default log event weighting and shared q90–99 search. The archived study retains 18,270 method–series–percentile results ($29 \times 63 \times 10$), independently verified. The curves and unrestricted comparator below describe this 29-method subset, not all 35 rows of the current main table.

Event hits and false alarms. For a maximal true interval $e = [a_e, b_e)$ of length L_e, let

$$w_e = \lfloor \log_3(L_e + 3) \rfloor, \quad h_e(q) = \mathbf{1}\left\{ \max_{t \in e} s_t > \theta_i(q) \right\}, \quad \theta_i(q) = \text{Quantile}_{\text{higher}}(s_{i,1:T}, q/100). \tag{34}$$

An event earns its entire weight upon any hit. Normal timestamps remain individual negatives; adjacent false alarms are not merged. For each series,

$$T_i(q) = \sum_e w_e h_e(q), \quad A_i = \sum_e w_e, \quad B_i(q) = \sum_{t: y_t = 0} \mathbf{1}\{s_t > \theta_i(q)\}. \tag{35}$$

We follow the official numerical convention $P_i = T_i/(T_i + B_i + \epsilon)$, $R_i = T_i/(A_i + \epsilon)$, and $F_i = 2P_iR_i/(P_i + R_i + \epsilon)$, with $\epsilon = 10^{-15}$. We explicitly disable label margins: the original labels are used with margins $(0, 0)$, rather than the controller's default post-event expansion. Thresholds use the original, uncompressed score array; label-dependent event compression never sets the percentile. Equal scores at the threshold are all rejected. The point-adjustment is an evaluation definition, not postprocessing deployed predictions.

Shared percentile and macro aggregation. For $\mathcal{Q} = \{90, \ldots, 99\}$ and family f, the main table reports

$$\overline{F}_{m,f}(q) = \frac{1}{n_f} \sum_{i \in f} F_i^m(q), \qquad q^\star_{m,f} = \min \arg\max_{q \in \mathcal{Q}} \overline{F}_{m,f}(q). \tag{36}$$

The per-series mean follows EasyTSAD's dataset aggregation, replacing the old pooled strict-event counts. One q is shared by a method–family pair; its score threshold differs by series. The maximum uses evaluation labels and is a restricted-grid retrospective oracle, not an independently calibrated deployment result. The raw masks are byte-for-byte identical to the previous ten-point sweep. Saved squeeze-mode evaluations ($w_e = 1$) expose sensitivity to the weighting convention.

Why the zero scores change. Of the 116 method–family maxima in that 29-method subset, 61 are zero under the archived strict metric and one remains zero under the revised grid metric. TFC and TimeXer now have positive values in all four families. POLY–Wave remains zero because none of its true events is hit anywhere in this grid; its saved scores are unchanged. A single correctly placed alarm can now detect a long event, so the strict-IoU q99 impossibility bound does not apply to EventF1PA. Conversely, many normal-point alarms can sharply reduce event-PA precision. This explains why the new metric need not increase every method's score. Metric choice changes what success means, not the predictions.

The official unrestricted comparator. For comparison, Table 16 calls the unmodified official `EventF1PA.calc` on each series, after the official `Performance` minimum-score subtraction with margins $(0, 0)$, and then averages per-series maxima. This is a broader search with a separate threshold for every series; it is not the main table's shared-percentile grid. The official routine visits individual equal-score records in stable order. It can therefore select part of a tied group even though no scalar threshold reproduces that subset. For constant scores and labels $(1, 0)$, the official log F1 is approximately 1; a realizable all-ties threshold yields at most $2/3$, and every strict percentile mask is empty. We report both the exact official output and a complete-tie-group search on the same shifted scores. The official implementation is retained unchanged; this distinction prevents attributing a tie-search artifact to detector quality.

Table 16: Unrestricted per-series EasyTSAD reference, averaged within each family. Official: exact log-mode `calc`, including its itemwise tie handling. Grouped: a realizable threshold search that admits entire equal-score groups. Both use official score translation and zero label margins. These per-series maxima have a broader search scope than the main table; they are not deployment results. Squeeze-mode reference outputs are retained in the source tables.

Method	Mutation		Step		Trend		Wave	
	Official	Grouped	Official	Grouped	Official	Grouped	Official	Grouped
MatrixProfile	0.245	0.245	0.349	0.349	0.344	0.344	0.047	0.044
SAND	0.157	0.157	0.389	0.389	0.122	0.122	0.389	0.389
SubLOF	0.055	0.055	0.086	0.086	0.225	0.225	0.206	0.206
KMeansAD	0.143	0.143	0.388	0.388	0.328	0.324	0.331	0.331
AR	0.392	0.392	0.645	0.645	0.551	0.551	0.047	0.047
LSTMAD	0.418	0.418	0.655	0.655	0.515	0.515	0.058	0.058
POLY	0.105	0.096	0.184	0.065	0.139	0.107	0.011	0.010
KAN-AD	0.404	0.404	0.641	0.641	0.601	0.601	0.081	0.081
TShape	0.339	0.339	0.575	0.575	0.383	0.383	0.239	0.239
DLinear	0.443	0.443	0.679	0.679	0.548	0.548	0.069	0.069
TFC	0.445	0.445	0.657	0.657	0.581	0.581	0.054	0.054
TimeXer	0.373	0.373	0.641	0.641	0.524	0.524	0.072	0.072
AE	0.142	0.142	0.149	0.149	0.132	0.132	0.527	0.527
EncDecAD	0.284	0.284	0.269	0.269	0.225	0.225	0.105	0.105
SRCNN	0.060	0.060	0.231	0.231	0.162	0.162	0.369	0.369
TranAD	0.317	0.317	0.360	0.360	0.415	0.415	0.448	0.412
Donut	0.034	0.034	0.028	0.028	0.027	0.027	0.201	0.201
FCVAE	0.617	0.617	0.772	0.772	0.696	0.696	0.054	0.054
ModernTCN	0.240	0.240	0.284	0.284	0.198	0.198	0.066	0.066
SEA-FGT	0.468	0.468	0.465	0.465	0.333	0.333	0.417	0.417
TimesNet	0.406	0.406	0.308	0.308	0.227	0.227	0.019	0.019
FITS	0.251	0.251	0.259	0.259	0.174	0.174	0.026	0.026
Sundial	0.258	0.258	0.415	0.415	0.174	0.174	0.395	0.395
Timer	0.312	0.312	0.359	0.359	0.147	0.147	0.259	0.259
Timer-XL	0.321	0.321	0.377	0.377	0.185	0.185	0.238	0.238
Time-MoE	0.268	0.268	0.386	0.386	0.211	0.211	0.350	0.350
TimeMixer	0.443	0.443	0.620	0.620	0.484	0.484	0.071	0.071
Time-o1	0.414	0.414	0.669	0.669	0.526	0.526	0.049	0.049
VeraTS (ours)	1.000	0.733	1.000	0.867	0.645	0.634	1.000	1.000

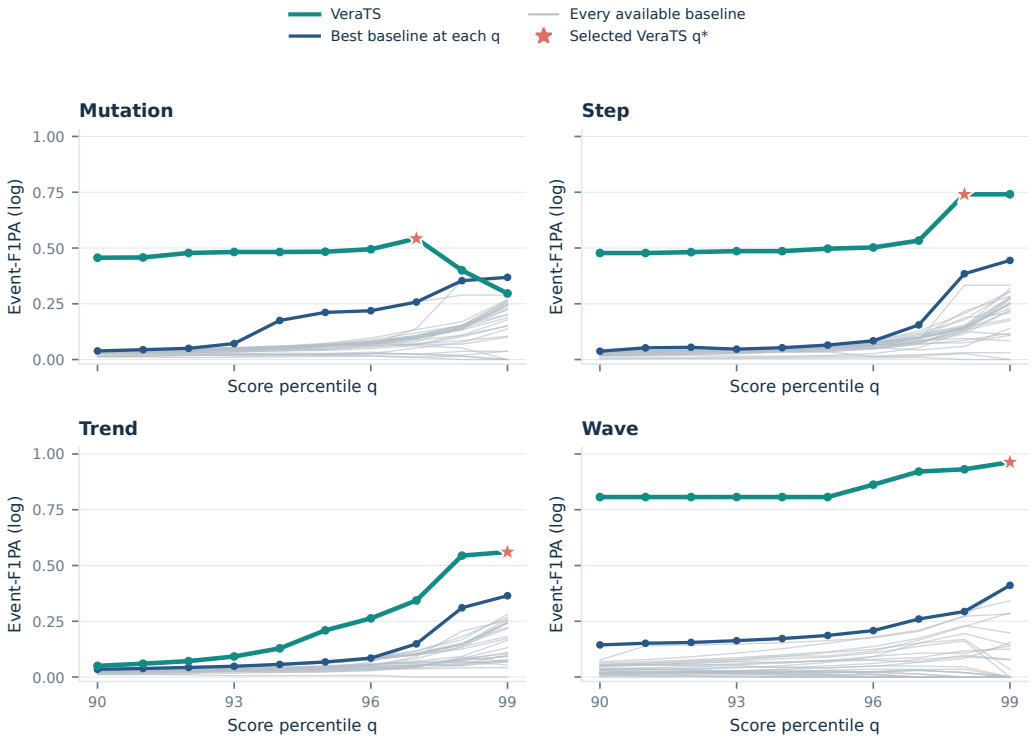


Figure 34: All ten integer percentiles are shown on a common absolute F1 scale. Gray lines are complete baselines; the blue pointwise envelope can switch methods across q. Stars mark the test-label-selected VeraTS maximum (smallest q on ties). Event-F1PA uses the EasyTSAD log-weighted any-hit definition without label margins; curves average per-series F1 at a shared q. Unlike strict IoU matching, sparse q99 alarms can detect long events.

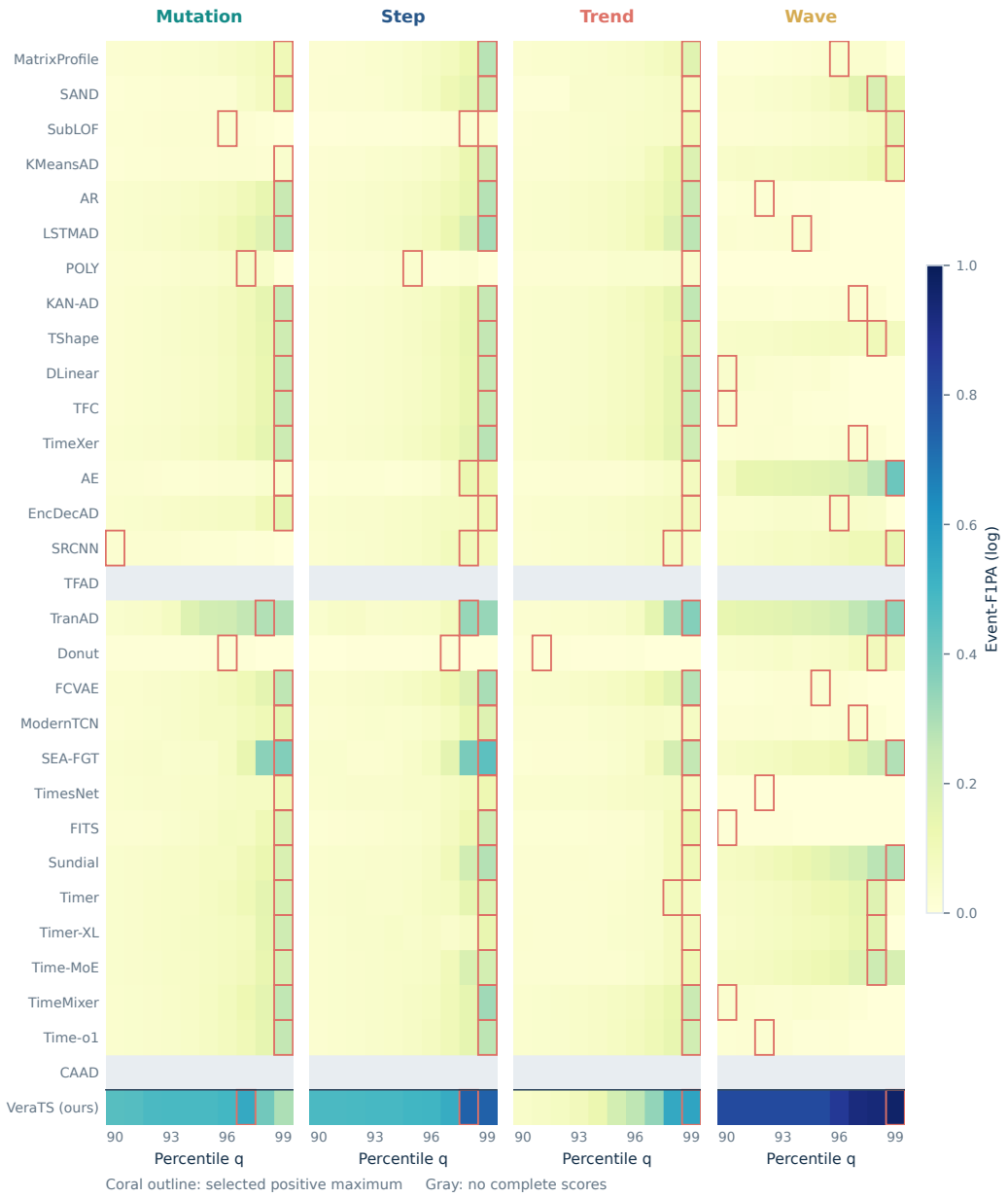


Figure 35: The same zero-to-one color scale is used for all methods and families. Every q90–99 value is retained, including zeros. Gray rows indicate unavailable complete experiments, not zero performance. Coral outlines mark selected positive maxima; tied maxima use the smallest q.

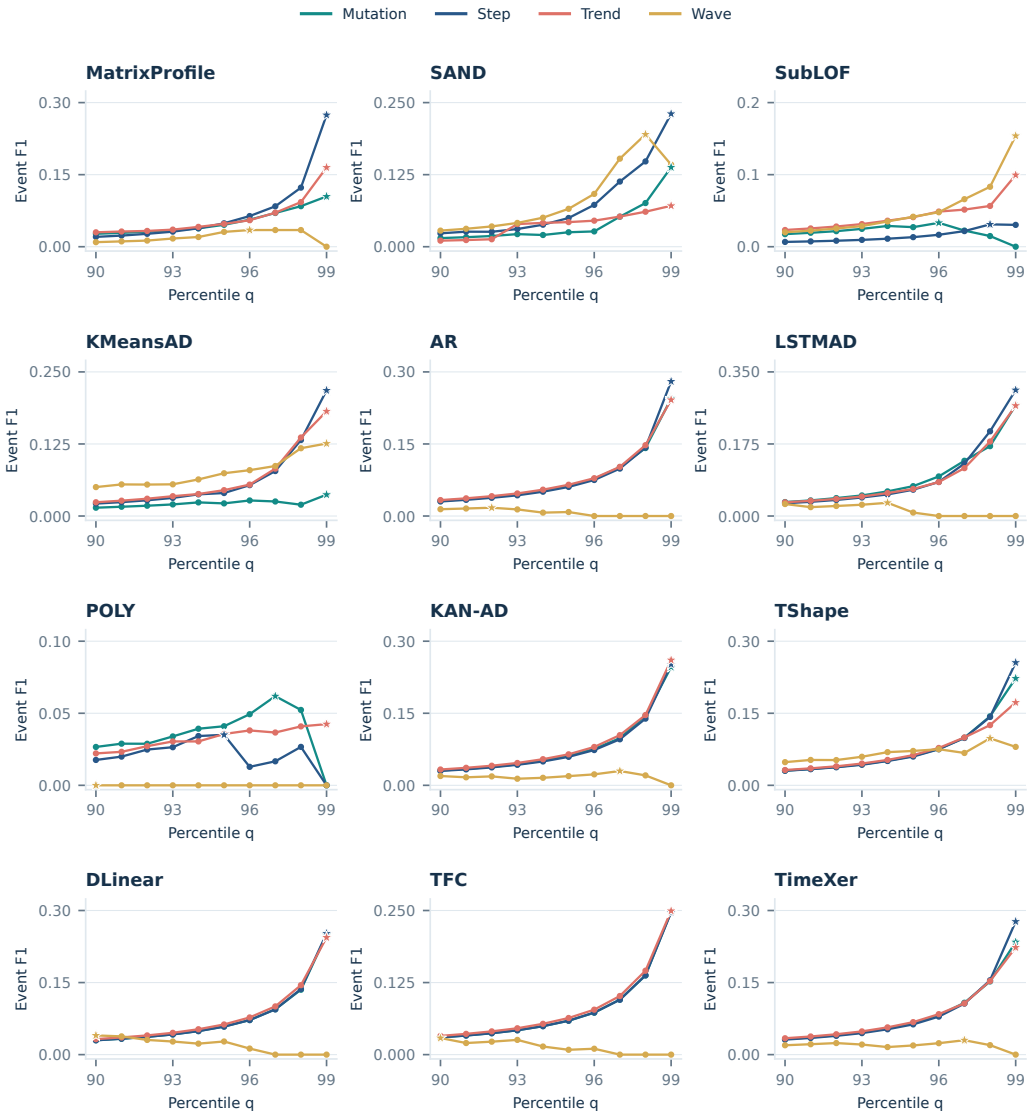


Figure 36: All ten points and four families are shown for each method. Explicit method-specific y-scales keep low curves readable; use the overview and heatmap for absolute comparisons. Stars denote selected maxima. These curves use event hits and normal-point false alarms, without an IoU requirement. No complete scores means unavailable, not zero.

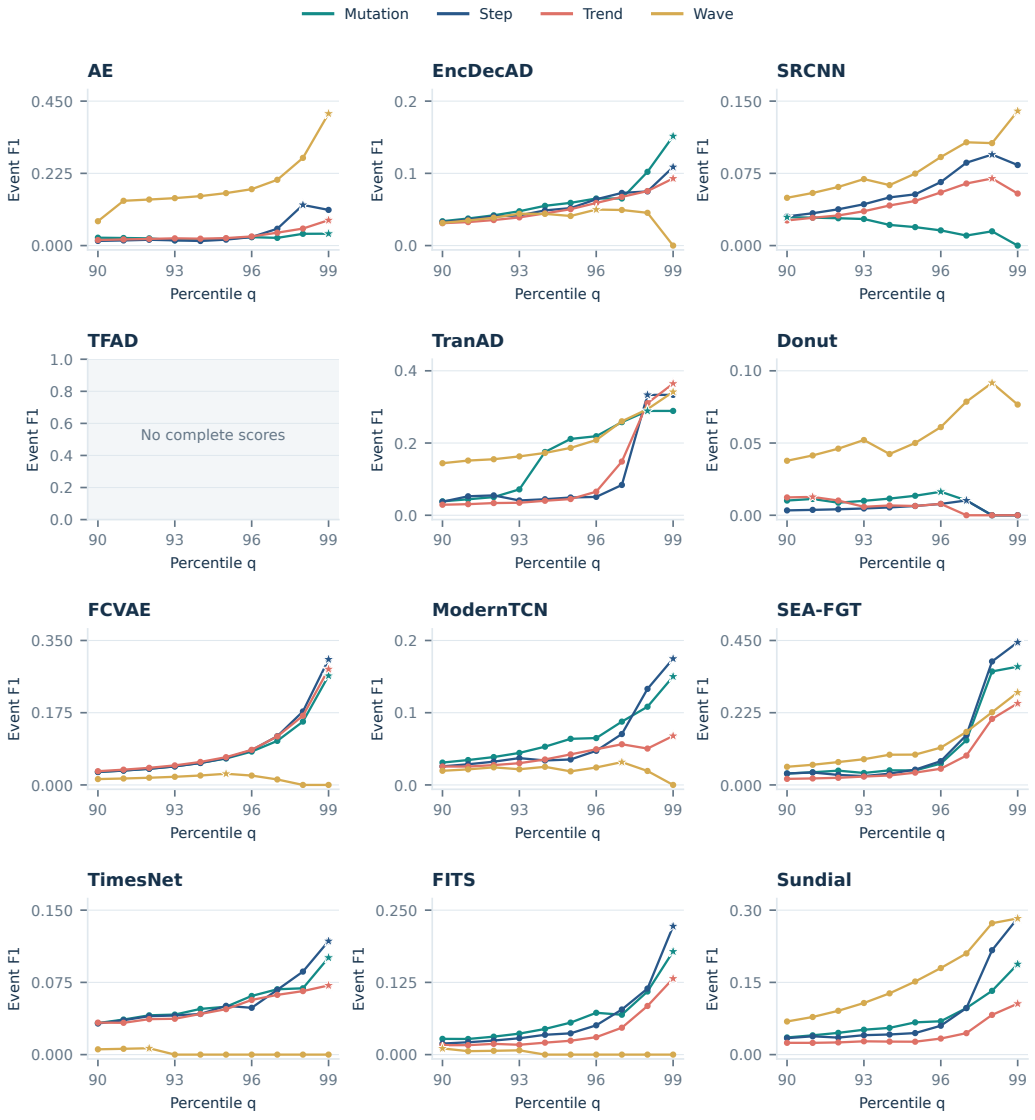


Figure 37: All ten points and four families are shown for each method. Explicit method-specific y-scales keep low curves readable; use the overview and heatmap for absolute comparisons. Stars denote selected maxima. These curves use event hits and normal-point false alarms, without an IoU requirement. No complete scores means unavailable, not zero.

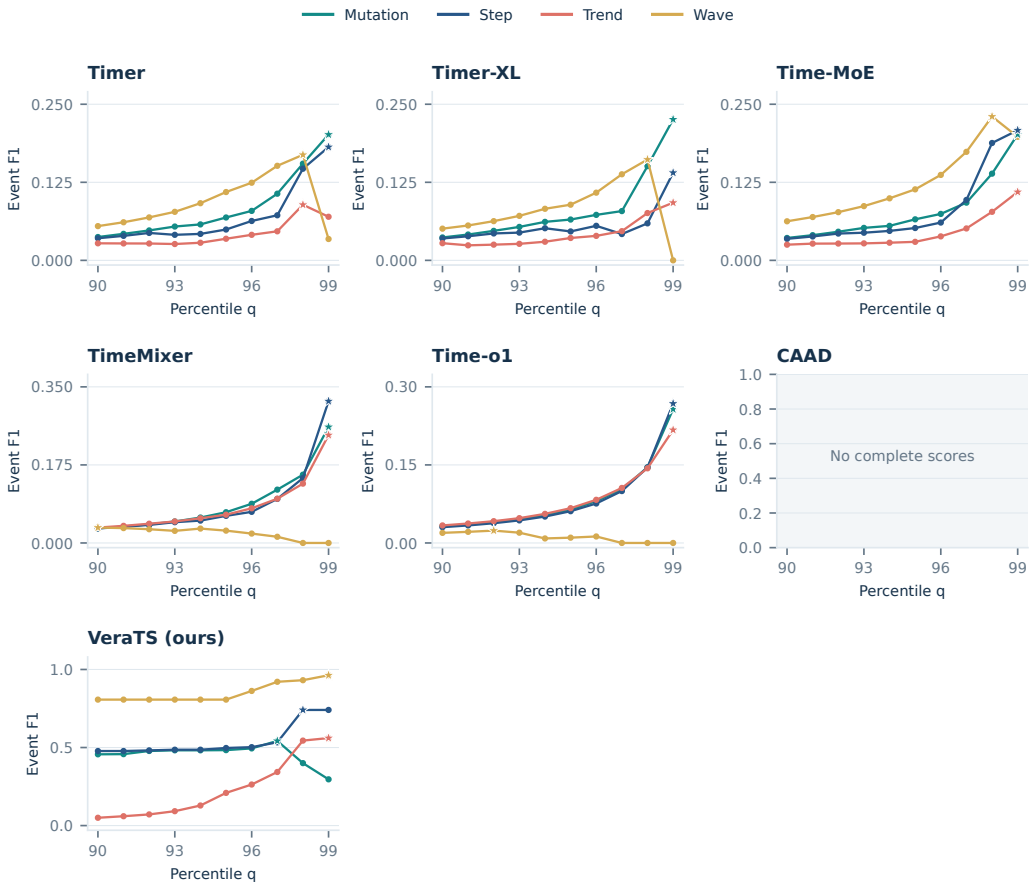


Figure 38: All ten points and four families are shown for each method. Explicit method-specific y-scales keep low curves readable; use the overview and heatmap for absolute comparisons. Stars denote selected maxima. These curves use event hits and normal-point false alarms, without an IoU requirement. No complete scores means unavailable, not zero.

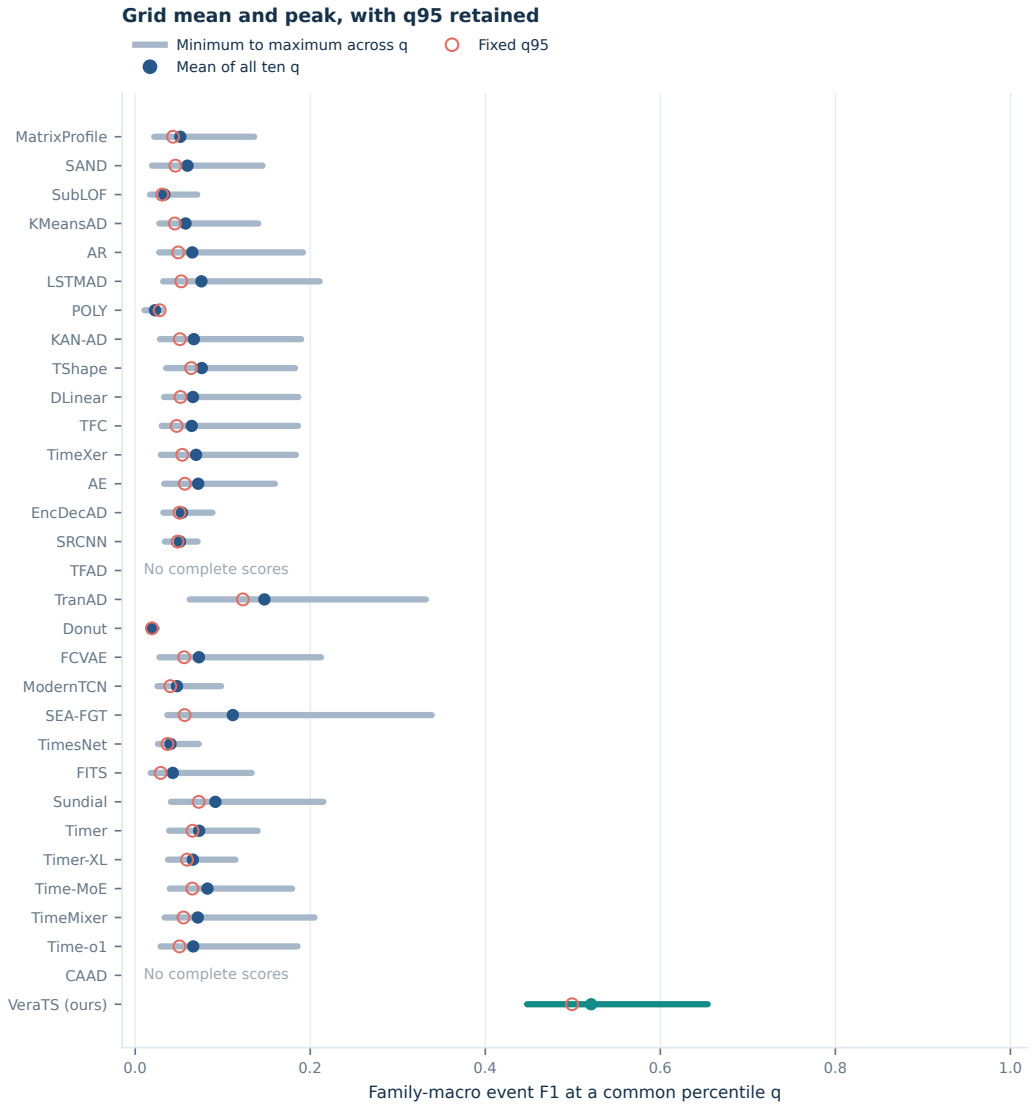


Figure 39: Bars span the minimum and maximum over ten percentiles, not confidence intervals. Dots average all ten values; open circles retain q95. Each point applies one common q to all families before averaging, unlike the family-specific maxima in the table. Ranges quantify sensitivity only over this specified grid. A low range is useful only together with adequate accuracy; it is not a confidence interval or evidence of deployment calibration.

M Archived Strict-IoU Threshold Diagnostic

A distinct localization metric. This section retains the previous strict-IoU results to explain the metric discrepancy; these values are no longer the main-table Event-F1*. Event-PA in the preceding section measures event detection, while this diagnostic requires interval overlap.

Selection and interpretation. For every complete method m, family f, and integer $q \in \mathcal{Q} = \{90, \ldots, 99\}$, we form $\widehat{y}_t^{m,q} = \mathbf{1}\{s_t^m > \text{Quantile}_{\text{higher}}(s_{1:T}^m, q/100)\}$. All ten raw masks are preserved without temporal closing, dilation, or point adjustment. We compute strict one-to-one IoU-0.5 interval matches and pool counts within a family:

$$F_{m,f}(q) = \frac{2\sum_{i\in f} M_i^{m,q}}{\sum_{i\in f} G_i + \sum_{i\in f} P_i^{m,q}}, \qquad q_{m,f}^\star = \min \arg\max_{q\in\mathcal{Q}} F_{m,f}(q). \tag{37}$$

The archived table snapshot reports $F_{m,f}(q_{m,f}^\star)$ with subscript $q_{m,f}^\star$. One percentile is selected for the entire method–family pair, never separately for each series; its absolute score threshold is still computed from each series' own score distribution. Selection uses evaluation labels and is a retrospective oracle over a specified grid. It is not a threshold fixed before evaluation, independent generalization performance, or calibration available without labels. The score/VUS protocol and original q95 artifacts remain unchanged. A separate implementation replays every percentile from sorted scores and maximum bipartite interval matching, including the fixed q95 anchors.

Why some zeros are inevitable. Under the strict higher-quantile rule, at most

$$B_T(q) = T - 1 - \lceil (T-1)q/100 \rceil \tag{38}$$

points can be positive. A true event of length L requires at least $\lceil L/2 \rceil$ positive points to obtain IoU ≥ 0.5. Here $T = 2{,}016$, giving budgets of 201, 100, and 20 points at q90, q95, and q99. Mutation and Step events have length 50, Trend events range from 50 to 67, and Wave events range from 168 to 288. Thus q99 cannot match any event for any method. Wave cannot match any event at q96–99; at q95, six of its nine events are individually impossible to match. The figures retain and mark these structural zeros. Low q95 event F1 does not alone establish poor ranking, and collapse in an impossible region cannot establish model fragility. The length bound is necessary, not sufficient: fragmentation and false events can reduce F1 even within a feasible budget.

Peak performance and sensitivity. VeraTS selects q97 for Mutation and Step, q95 for Trend, and q90 for Wave, yielding event F1 of 0.900, 0.947, 0.455, and 0.286. These are label-selected descriptive maxima. The overview uses one common absolute scale and a pointwise best-baseline envelope, which may switch methods across thresholds. The heatmap and per-method panels retain every measured value, including zero. The range plot applies each common q to all four families and then averages their F1 values; its mean and range describe the ten-point grid, not uncertainty intervals. Every q99 endpoint is necessarily zero, so each range equals its peak and supplies no separate stability evidence. Future deployment comparisons require thresholds frozen on a separate calibration set and event-length-aware operating constraints.

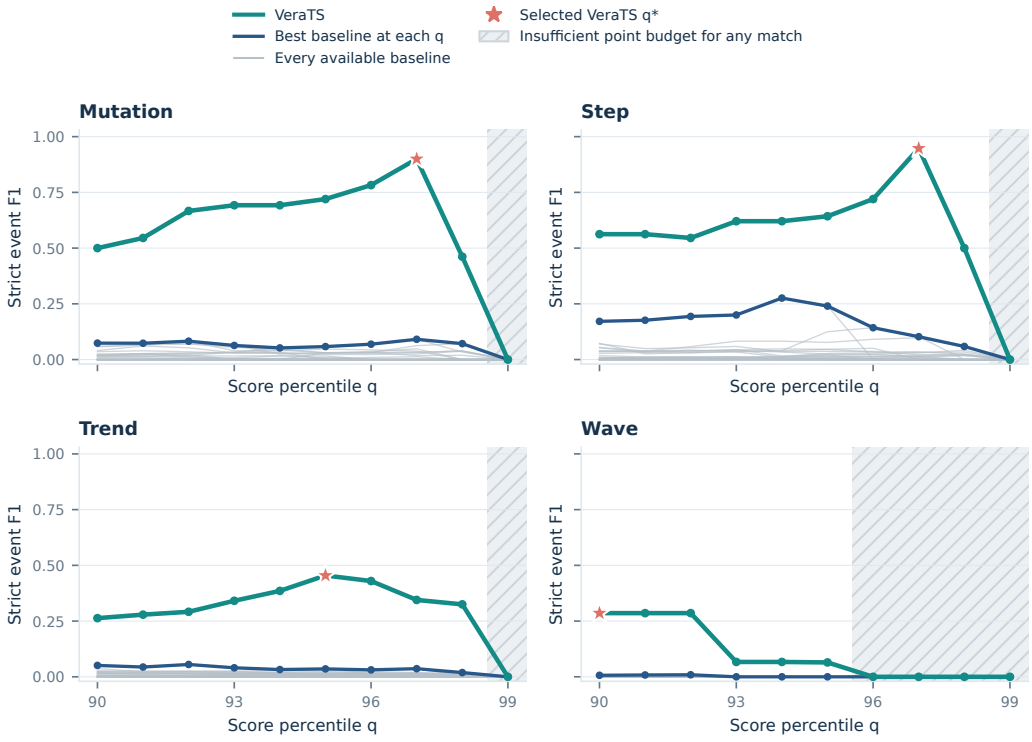


Figure 40: All ten integer percentiles are shown on a common absolute F1 scale. Gray lines are complete baselines; the blue pointwise envelope can switch methods across q. Stars mark the test-label-selected VeraTS maximum (smallest q on ties). Hatched regions have insufficient positive-point budget to match any true event at IoU at least 0.5: q99 for every family and q96–99 for Wave. Those unavoidable zeros are retained, but cannot diagnose model fragility.



Figure 41: The same zero-to-one color scale is used for all methods and families. Every q90–99 value is retained, including zeros. Gray rows indicate unavailable complete experiments, not zero performance. Coral outlines mark selected positive maxima; tied maxima use the smallest q.

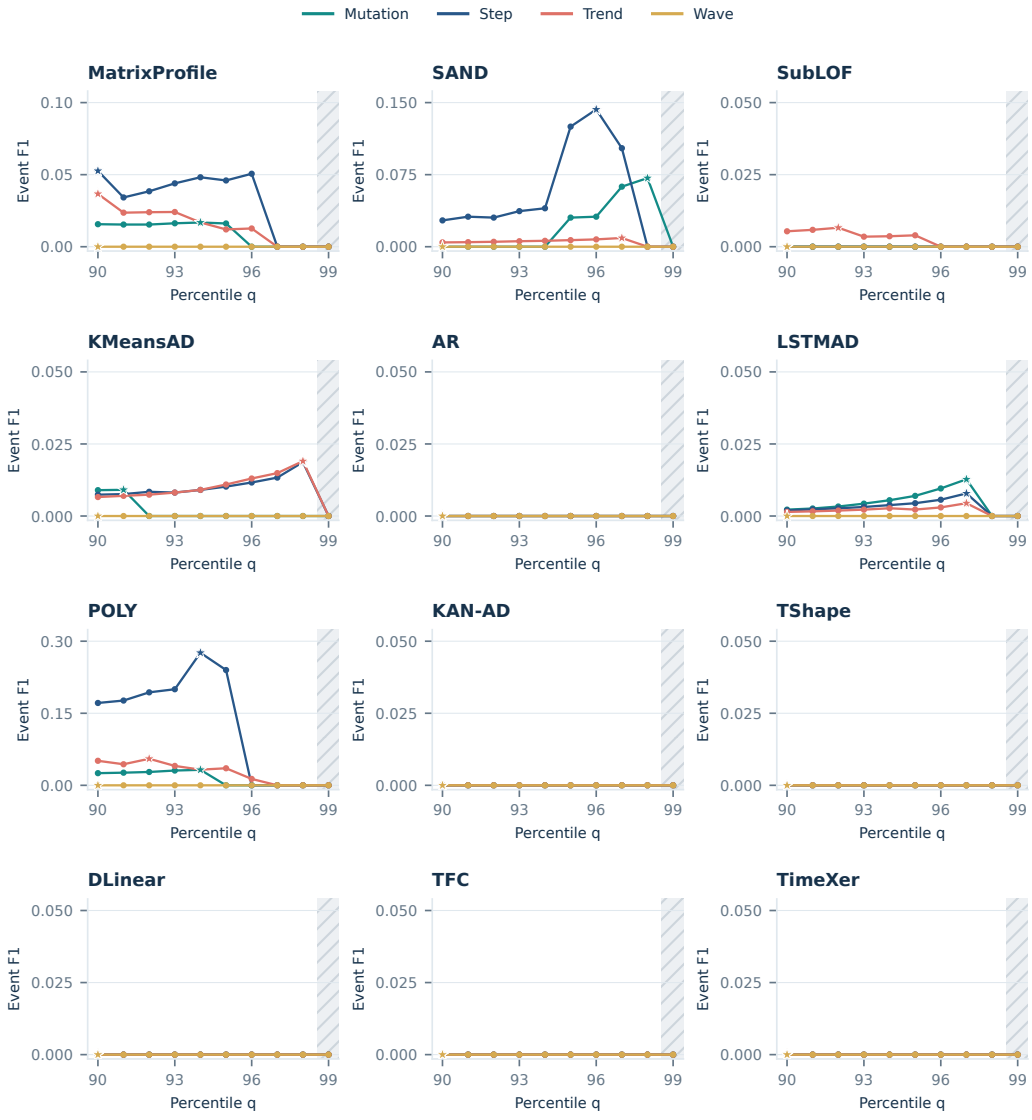


Figure 42: All ten points and four families are shown for each method. Explicit method-specific y-scales keep low curves readable; use the overview and heatmap for absolute comparisons. Stars denote selected maxima. The shaded q99 endpoint is structurally zero for every family; Wave also cannot match at q96–99. No complete scores means unavailable, not zero.

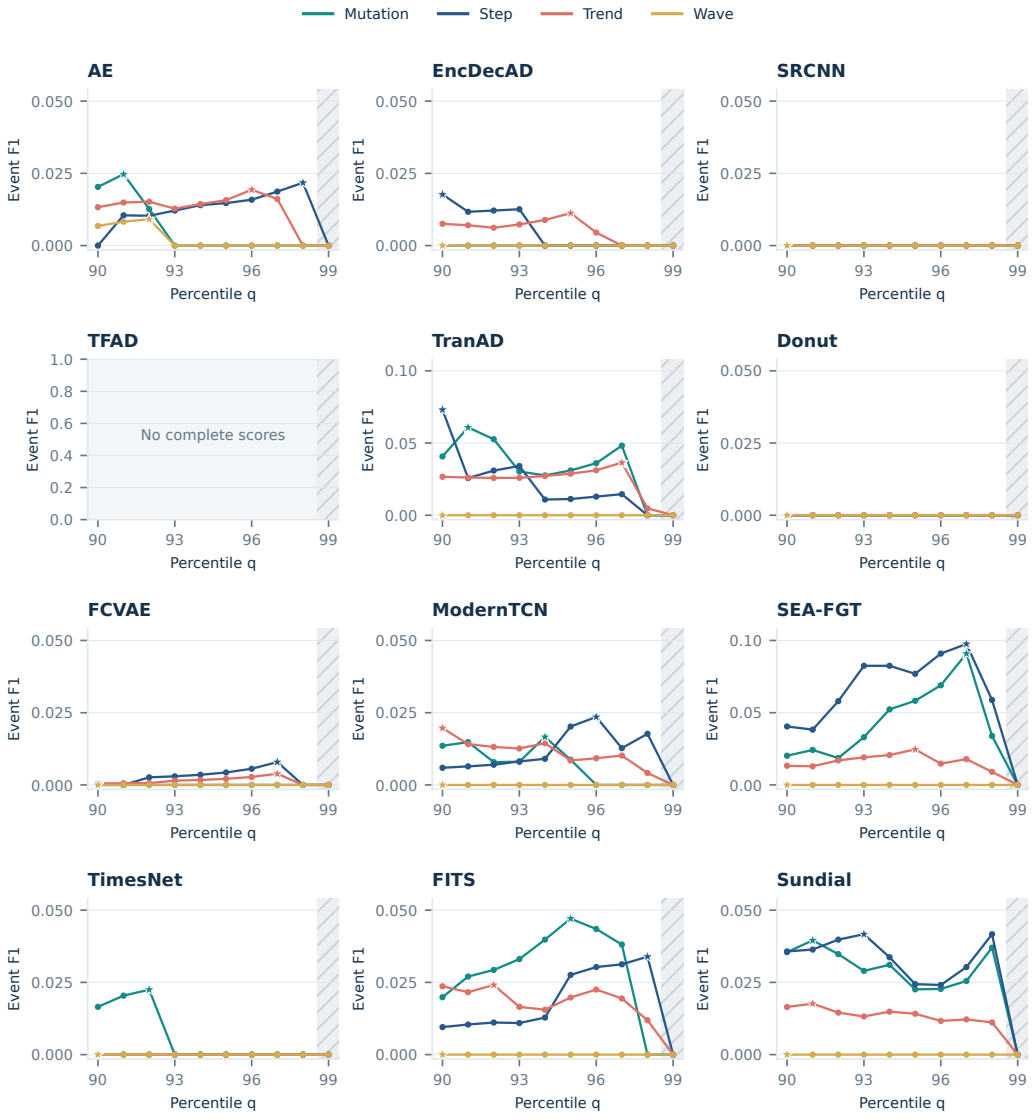


Figure 43: All ten points and four families are shown for each method. Explicit method-specific y-scales keep low curves readable; use the overview and heatmap for absolute comparisons. Stars denote selected maxima. The shaded q99 endpoint is structurally zero for every family; Wave also cannot match at q96–99. No complete scores means unavailable, not zero.



Figure 44: All ten points and four families are shown for each method. Explicit method-specific y-scales keep low curves readable; use the overview and heatmap for absolute comparisons. Stars denote selected maxima. The shaded q99 endpoint is structurally zero for every family; Wave also cannot match at q96–99. No complete scores means unavailable, not zero.

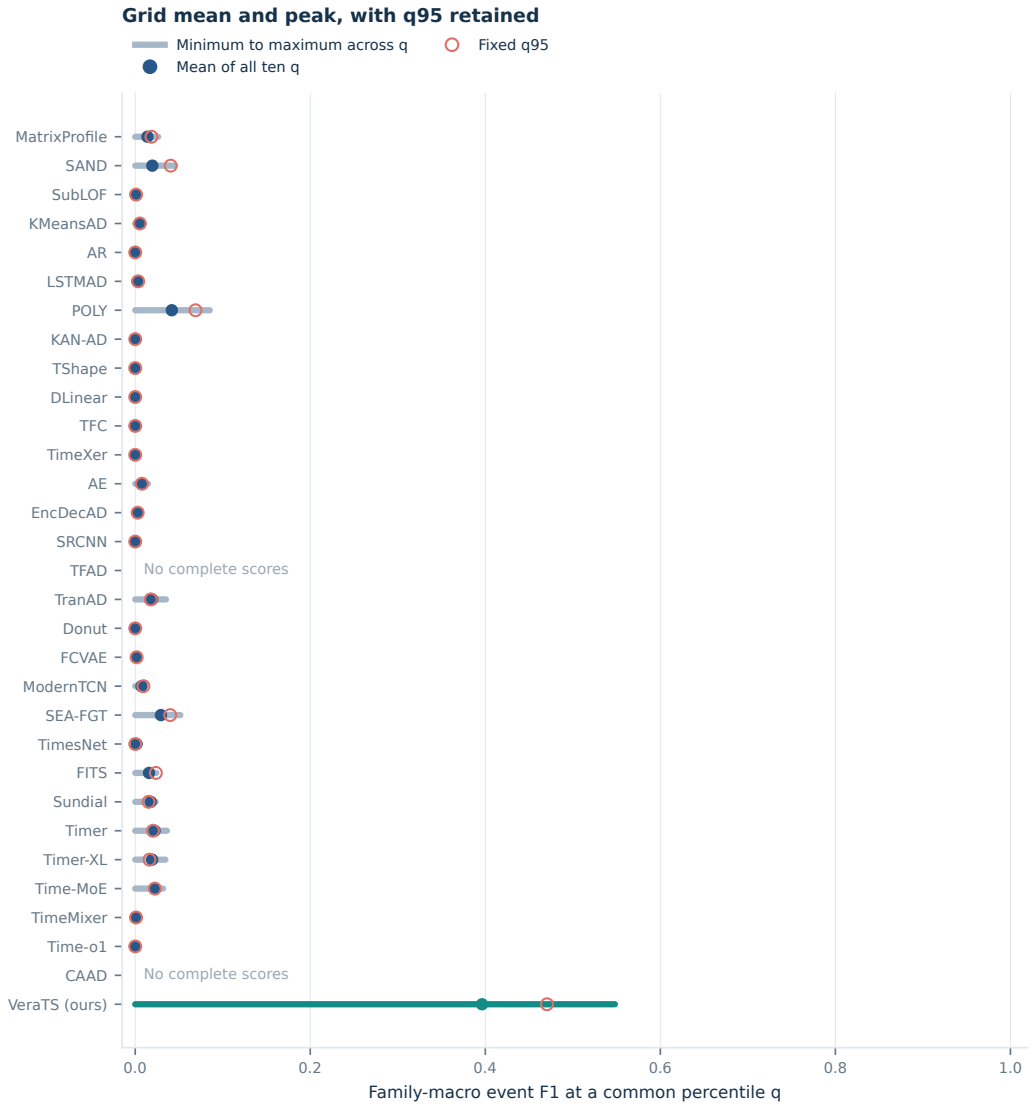


Figure 45: Bars span the minimum and maximum over ten percentiles, not confidence intervals. Dots average all ten values; open circles retain q95. Each point applies one common q to all families before averaging, unlike the family-specific maxima in the table. Every minimum is zero because q99 cannot match any true event under this alert budget. A low range alone does not establish useful robustness.

N Additional Experiments: Robustness, Transfer and Complete Score Evidence

This extension connects the design of VeraTS to five practical benefits: broad event access, selective temporal decisions, useful quantitative scores, stable operation under small measurement changes, and exact numerical replay. Serial perturbation experiments and frozen-score analyses cover all 63 series and all prespecified conditions. The original main-table numbers and training-selected policy remain fixed.

N.1 Experiment map and scope

The new experiments comprise 1,953 condition–series replays (31 conditions on 63 series) and 2,205 parameter–series evaluations (35 grid entries on 63 series, representing 34 unique settings because the selected rule appears in both sweeps). The clean condition is included in those counts and all 63 clean score and prediction arrays reproduce the frozen initialized experiment bit for bit. Reused feature arrays accelerate parameter sweeps without changing the computation. These counts describe evaluations, not independent samples.

The saved-score study covers seven methods: VeraTS, GPT-6, step-5-preview, kimi-k3, glm-5.3, MatrixProfile, and Timer-XL. The three-method trace figures emphasize VeraTS, GPT-6 and step-5-preview; the complete seven-method numerical results accompany them. The three direct API baselines use the already disclosed interval adapters, whereas GPT-6 supplies ten-level pointwise ratings. No new API calls, model retraining, or LLM-generated semantic policies are introduced.

Question	Intervention or analysis	Evidence
Measurement corruption	Four noise levels; three missingness rates; three fixed seeds	1,323 perturbed traces
Drift and units	Six sensor-offset ramps; three equivalent-unit transforms	567 transformed traces
Parameter response	30 threshold/persistence pairs; five closing widths	2,205 evaluations
Threshold transfer	Nine held-out scenario prefixes; seven score methods	63 folds, 441 held-out scores
Score resolution	Realizable tied thresholds; monotone ten-bin control	882 score variants
Complete case evidence	Original values, scores, masks and stress response	63 full-trace plates

Table 17: Prespecified computation inventory. The 63 clean replays are additional to the 1,890 transformed traces. Repeated seeds and methods do not create new independent data.

How to read the additional figures. Seventeen summary plates reuse the visual grammar of the existing style references 28–44, including small multiples, annuli, event flows, rainclouds, paired slopes, score carpets, feature projections, networks and sensitivity matrices. Each plate states a question, experiment, measured result, design benefit and data source. They are alternative encodings of the named experiments, not 17 independent benchmarks. Every generated plot is available as a vector PDF, editable SVG and preview PNG.

N.2 Frozen protocol and interpretation

Label and parameter access. The protocol, all seeds, input hashes and selected policy are recorded before generating the new perturbation scores. The serial scoring process uses only train/test values and saves a hash seal before a separate evaluation stage opens test labels. Earlier human exposure to the corpus remains; this is a controlled retrospective extension. The selected parameters remain candidate cutoff 1.0, semantic threshold 2.0, five consecutive points within nine, and closing width 48.

Perturbations. Let $s_i = \max(1.4826\,\text{median}|x_i^{\text{train}} - \text{median}(x_i^{\text{train}})|, 10^{-8})$. Noise adds $\alpha s_i \epsilon_t$, with $\epsilon_t \sim \mathcal{N}(0,1)$ and $\alpha \in \{.05, .10, .25, .50\}$. Missingness independently masks each evaluation value with probability $p \in \{.01, .05, .10\}$; masked values use the last available observation, initialized by the final training value. The three seeds are 2027, 2028 and 2029, deterministically combined with the canonical series index. The same draws are reused across strengths. Neither operation changes the original fault labels.

Sensor-offset drift adds a linear evaluation ramp from zero to ds_i, for $d \in \{-2, -1, -.5, .5, 1, 2\}$. This models corrupted observations of the same latent labeled trace; it does not assert that a newly introduced operational trend is benign. Unit checks transform both splits by $x \mapsto 0.1x$, $x \mapsto 10x$, or $x \mapsto x + 10s_i$. Every condition rebuilds the initialized numerical references. Units are not converted into undocumented physical quantities.

Metrics. Point average precision (AP) uses original dense scores and original binary labels, without tolerance windows. Point F1 uses the indicated fixed mask. Strict F1 requires one-to-one IoU-0.5 interval matching. EventF1PA uses the preceding appendix definition, with base-three log event weights, no label margins, and normal-point false positives. At q95 the mask is $s_t > \text{Quantile}_{\text{higher}}(s, .95)$. All ties at the threshold are rejected; no arbitrary subset of equal scores is selected. The frozen-policy mask instead uses VeraTS robustness > 0. These operating points are explicitly separated.

Aggregation and uncertainty. Unless labeled otherwise, summaries average per-series metrics within each morphology family and then average the four family means equally. Raw event and point counts retain their stated denominators. Three-seed bands describe perturbation variability on the fixed corpus only; observed-series ranges are not confidence intervals. Scenario prefixes are directory identifiers rather than independently verified source services, and value reuse persists.

What this extension establishes. The results link the fixed hierarchy to complete clean-event coverage, strong common-percentile and transferred-percentile discovery, stable low-noise and missingness responses, and reproducible numerical execution. Native-score and reference controls explain how quantitative resolution and contextual evidence support these outcomes. All claims refer to the supplied corpus, frozen author-set policies and disclosed offline evaluation protocols.

N.3 Compact numerical results

Condition	Point AP	Point F1	Strict F1	Candidate fraction
Clean 0	0.741	0.758	0.745	0.272
Noise 0.05	0.762	0.761	0.728	0.277
Noise 0.1	0.734	0.730	0.668	0.292
Noise 0.25	0.723	0.664	0.553	0.336
Noise 0.5	0.686	0.570	0.404	0.402
Missing 0.01	0.742	0.758	0.748	0.279
Missing 0.05	0.756	0.769	0.751	0.328
Missing 0.1	0.763	0.764	0.737	0.389
Drift -2	0.797	0.615	0.555	0.342
Drift -1	0.782	0.620	0.558	0.269
Drift -0.5	0.757	0.606	0.542	0.251
Drift 0.5	0.696	0.571	0.549	0.275
Drift 1	0.688	0.542	0.510	0.304
Drift 2	0.666	0.467	0.402	0.366
Scale 0.1	0.747	0.758	0.745	0.272
Scale 10	0.741	0.758	0.745	0.272
Offset 10	0.744	0.758	0.745	0.272

Table 18: All perturbation levels under the fixed policy. Noise and missingness rows average their three seeds. All values are equal-family means; candidate fraction is observation retention, not inference cost.

Method	q95 event F1	Transfer F1	Main-table mean	Point AP
VeraTS	0.499	0.628	0.702	0.741
GPT-6	0.098	0.145	0.180	0.501
step-5	0.268	0.267	0.383	0.546
kimi-k3	0.090	0.127	0.164	0.544
glm-5.3	0.117	0.077	0.214	0.577
MatrixProfile	0.043	0.136	0.145	0.170
Timer-XL	0.059	0.081	0.155	0.187

Table 19: Threshold-transfer diagnostics and the current main-table summary. Only the Main-table mean column averages the four displayed Event-F1* values in Table 1, rounded to three decimals. The q95, transfer and point-AP columns are independent evaluations of frozen score artifacts, not reconstructions of the current main table. Transfer excludes held-out scenario labels during q selection but uses the full unlabeled held-out score distribution.

N.4 Style coverage and artifact map

Style	Visual grammar	New scientific role
28	Small-multiple curves	Stable diagnosis under sparse measurement corruption
29	Radial bars and matrix	Contextual robustness supports accurate point ranking
30	Annular fractions	Point support motivates explicit temporal decisions
31	Alluvial event paths	Broad event coverage across shared event identities
32	Correlation triangle and ridges	Trace structure informs contextual reference design
33	Raincloud and family matrix	Paired evidence for the low-noise operating regime
34	Paired slope panels	Strong event discovery with transferred percentiles
35	Lollipop and score carpet	Leading event discovery at a common percentile
36	Faceted scatter comparison	Quantitative evidence connects ranking to event decisions
37	Layered score trace	Traceable evidence from raw values to a localized alert
38	Event matrix and cumulative curves	Complete event coverage with inspectable temporal support
39	Feature projection and composition	Signal heterogeneity motivates complementary references
40	Similarity network and distributions	Auditable data context for reproducible diagnosis
41	Pastel multi-panel dashboard	Consistent decisions across equivalent numerical units
42	Blue/gold budget dashboard	Broad candidate coverage with selective numerical diagnosis
43	Purple/green paired distributions	Native score resolution preserves informative evidence
44	Warm sensitivity surface	The frozen policy lies near the observed performance plateau

The next 17 plates connect each experiment to its design motivation, observed evidence and resulting benefit. The complete comparative atlas provides one page per series in canonical order, with original fault bounds, full-resolution raw values and scores, q95 and policy masks, numerical error counts and three-seed stress summaries. All 63 cases remain available for direct inspection.

Reproduction. The extension is generated by `scripts/run_appendix_extension.py` and `scripts/make_appendix_extension.py`. Detailed CSVs and hashed arrays are under `results/appendix_extension_20260926/`; vector and editable plots are under `output/figures/appendix_extension_20260926/`. The extension manifest binds inputs, intermediate scores and final figure files. Raw credentials are absent; the analysis reads only the existing saved model responses.

N.5 Stable diagnosis under sparse measurement corruption

Question. How does contextual temporal diagnosis preserve useful decisions under small measurement changes?

Experiment. Add zero-mean Gaussian noise to evaluation values in training-MAD units, or remove independent observations and forward-fill from the past. Reuse seeds 2027–2029 across strengths; keep training and labels fixed.

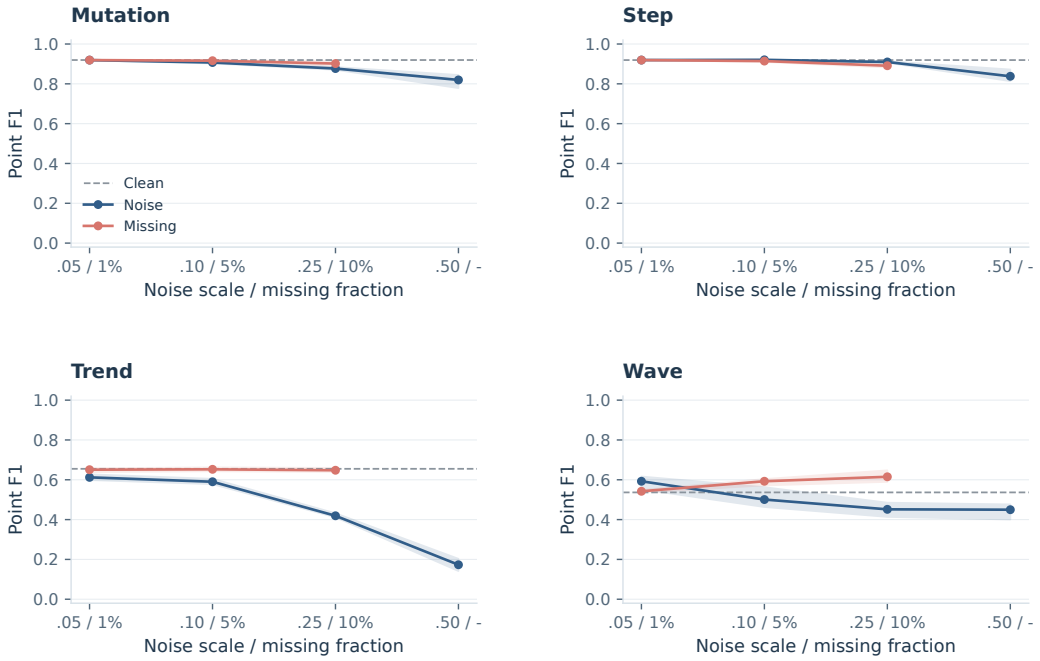


Figure 46: Family means of per-series point F1. Lines average three seeds; bands span their minimum and maximum, not confidence intervals. The two x-label entries give noise and missingness settings; the final missingness setting is absent.

Observed result. The fixed policy achieves family-balanced point F1 0.758 on clean traces, 0.761 at noise 0.05, and 0.764 at 10% missingness. The figure and compact table report every tested strength.

Interpretation. The similar clean, low-noise and missingness scores support the practical value of contextual temporal qualification on these perturbations. Reference-based margins and sustained-evidence rules retain useful event structure with one frozen policy; the full curves document its response across the prescribed measurement conditions.

Audit trail. Style 28; `stress_per_series.csv`.

N.6 Contextual robustness supports accurate point ranking

Question. Does policy-conditioned numerical evidence assign useful priorities to anomalous observations?

Experiment. Compute point average precision on unmodified dense scores and average first within family. Select the four displayed methods before the new diagnostics; all seven methods remain in the numerical tables.

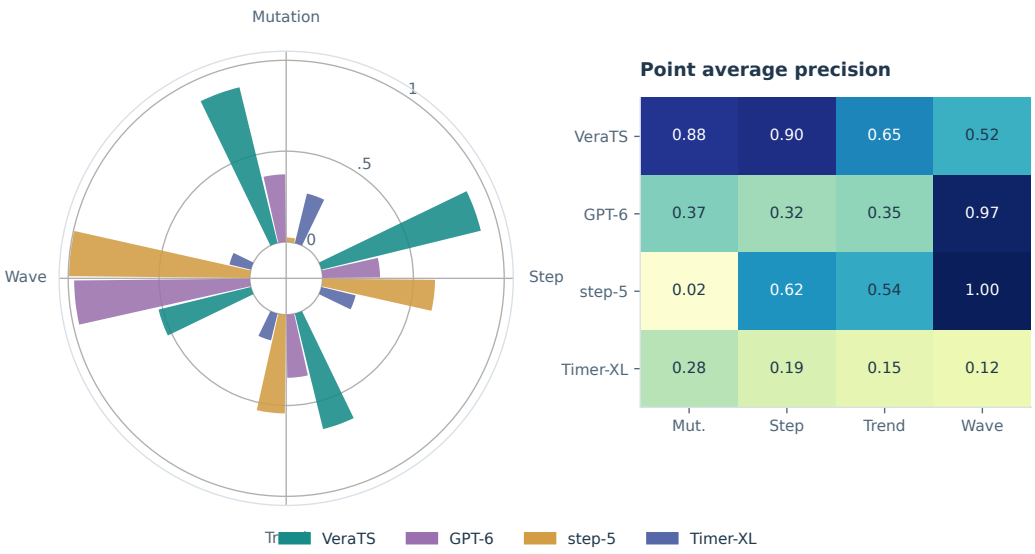


Figure 47: Radial bar length encodes 0–1 AP from a common inner baseline; the companion matrix provides exact values. Sector area is not a second metric.

Observed result. Four-family point AP is VeraTS 0.741, GPT-6 0.501, step-5 0.546, Timer-XL 0.187.

Interpretation. VeraTS has the highest four-family point AP among the displayed methods. This supports the usefulness of its quantitative robustness score for prioritizing individual observations as well as producing temporal decisions. AP evaluates original point labels, complementing the tolerance-aware VUS results; the disclosed adapters retain their respective input contracts.

Audit trail. Style 29; `comparison_per_series.csv`.

N.7 Point support motivates explicit temporal decisions

Question. What does pointwise API scoring reveal about the value of an executable event policy?

Experiment. Use score > the higher empirical 95th percentile of each complete evaluation sequence. Sum true-positive, false-negative and normal-point false-positive counts within each family.

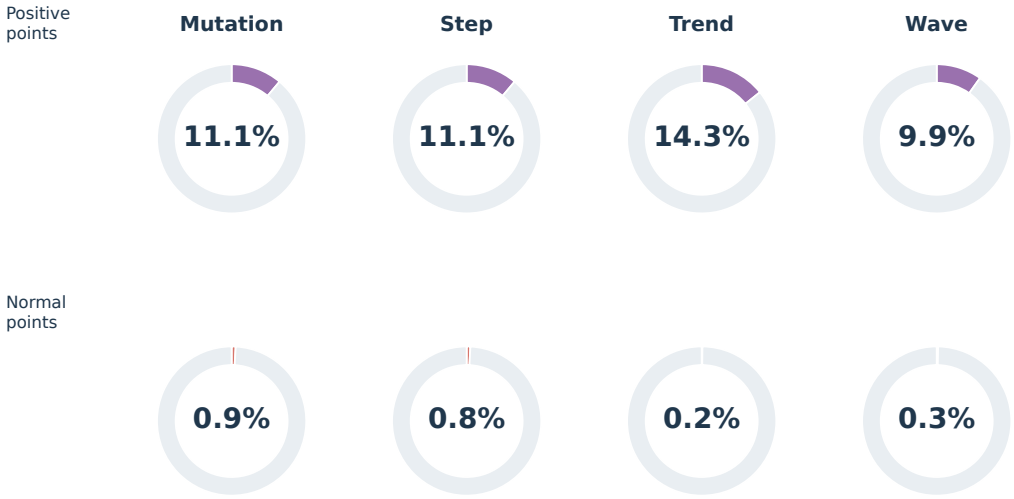


Figure 48: Top annuli: TP/(TP+FN), pooled over original positive points. Bottom: FP/(all normal points). The denominators differ; neither row represents calibrated anomaly probability.

Observed result. At q95 GPT-6 assigns positive decisions to 1,075 points, including 620 labeled and 455 normal observations. The matching all-event audit records 66/66 event hits for VeraTS and 11/66 for GPT-6 under the same percentile convention.

Interpretation. The point-support view motivates the separation between numerical ratings and explicit event decisions. VeraTS makes direction and duration part of the scoring rule, and its complete event coverage in the accompanying audit provides the corresponding detection evidence. The annuli preserve the point-level denominators needed to interpret that comparison.

Audit trail. Style 30; `comparison_per_series.csv`.

N.8 Broad event coverage across shared event identities

Question. Does broad screening followed by contextual diagnosis preserve event discovery across the complete corpus?

Experiment. Enumerate all 66 maximal original label intervals. A hit requires any strict-q95 positive point inside the interval. Link identical event identities across the three predeclared methods.

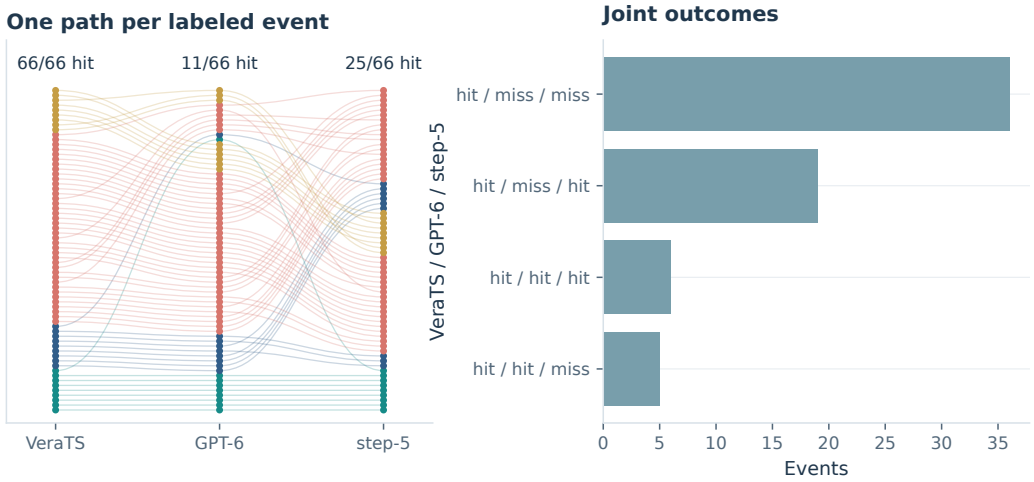


Figure 49: Events are sorted separately into missed then hit groups within each method. Color identifies family. The adjacent bar chart gives exact joint counts, avoiding ambiguity from crossing paths.

Observed result. VeraTS touches all 66 labeled events at q95, compared with 11 for GPT-6 and 25 for step-5. Six events are touched by all three methods. Every path tracks one original event across the three outputs.

Interpretation. The complete set of VeraTS event hits supports the coverage objective of the hierarchy: broad proposals and policy-conditioned scores preserve access to events across the four morphologies. These any-hit outcomes measure discovery; the separate strict-IoU experiments evaluate interval localization.

Audit trail. Style 31; `events.csv`.

N.9 Trace structure informs contextual reference design

Question. Which measured trace properties help interpret the value of adapting evidence to its numerical context?

Experiment. Join value-only train-normalized features with the three-seed mean point-F1 change at noise scale 0.50. Calculate Spearman correlations across the 63 supplied series. Prevalence and F1 enter only this evaluator-side analysis.

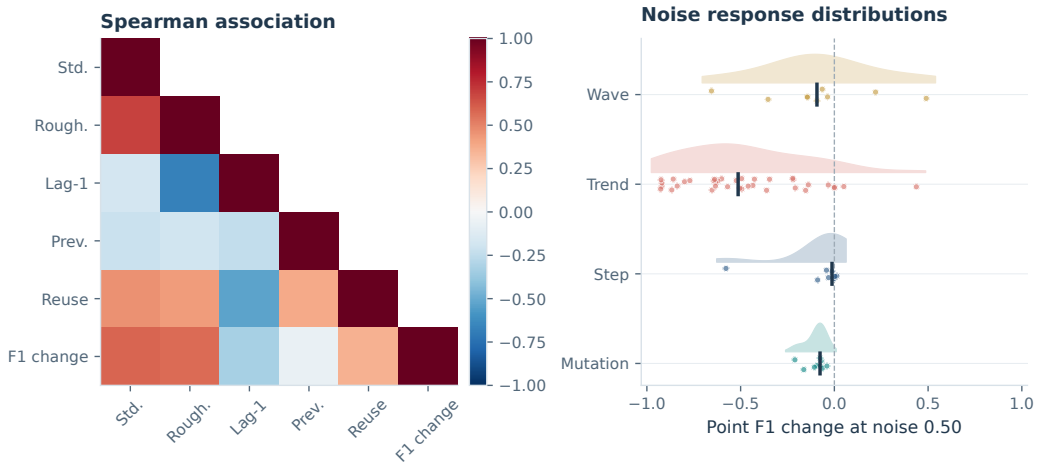


Figure 50: The lower correlation triangle uses a symmetric -1 to 1 scale. Distribution curves smooth the observed changes; raw points and median marks remain visible. There are no significance stars.

Observed result. The Spearman association between train-value reuse and point-F1 change is 0.355; between roughness and the change it is 0.564.

Interpretation. The associations connect the measured stress response to observable trace structure and motivate contextual reference construction. Rolling and seasonal paths provide explicit places to represent that structure within the rule. This descriptive analysis informs the design rationale; the matched reference ablations provide the component-level performance evidence.

Audit trail. Style 32; `series_features.csv`; `stress_per_series.csv`.

N.10 Paired evidence for the low-noise operating regime

Question. How well does the frozen temporal rule preserve detection quality at small noise levels?

Experiment. Subtract each clean point F1 from that same series after perturbation, averaging its three fixed seeds before plotting. Retain increases as well as decreases.

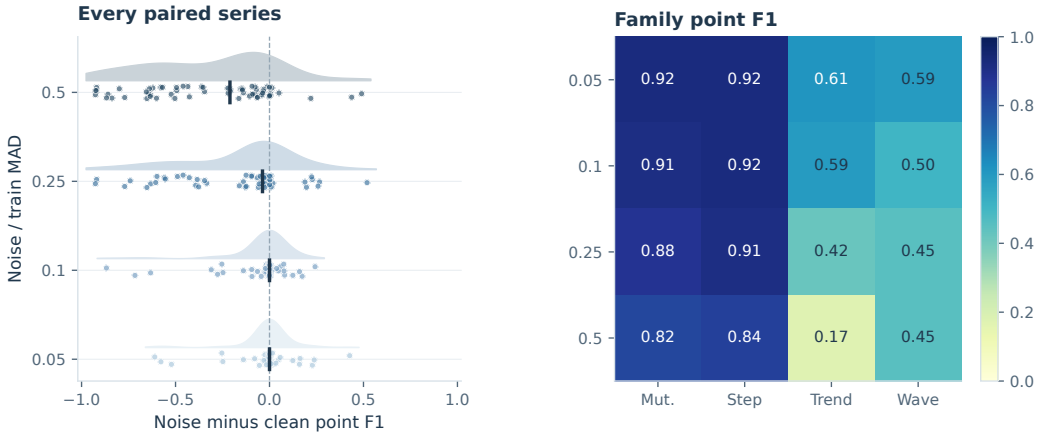


Figure 51: Rainclouds display all 63 paired effects; the matrix averages per-series F1 within each family. Density is a reading aid, not a fitted population model.

Observed result. Family-balanced point F1 is 0.758 on clean data, 0.761 at noise 0.05 and 0.730 at noise 0.10. The paired display retains all 63 series at each of the four noise strengths.

Interpretation. The low-noise results support a useful operating regime for the fixed contextual rule, with point F1 close to its clean value. Pairing every perturbed trace with its own clean counterpart makes this stability directly inspectable while preserving the full range of measured responses.

Audit trail. Style 33; `stress_per_series.csv`.

N.11 Strong event discovery with transferred percentiles

Question. Can the same numerical scores support effective threshold selection using other scenario prefixes?

Experiment. Hold out each directory scenario prefix. Select one shared q90–99 using only other prefixes, with equal-family donor means and smallest-q tie breaks. Compute that q from unlabeled held-out scores; never use its labels for selection.

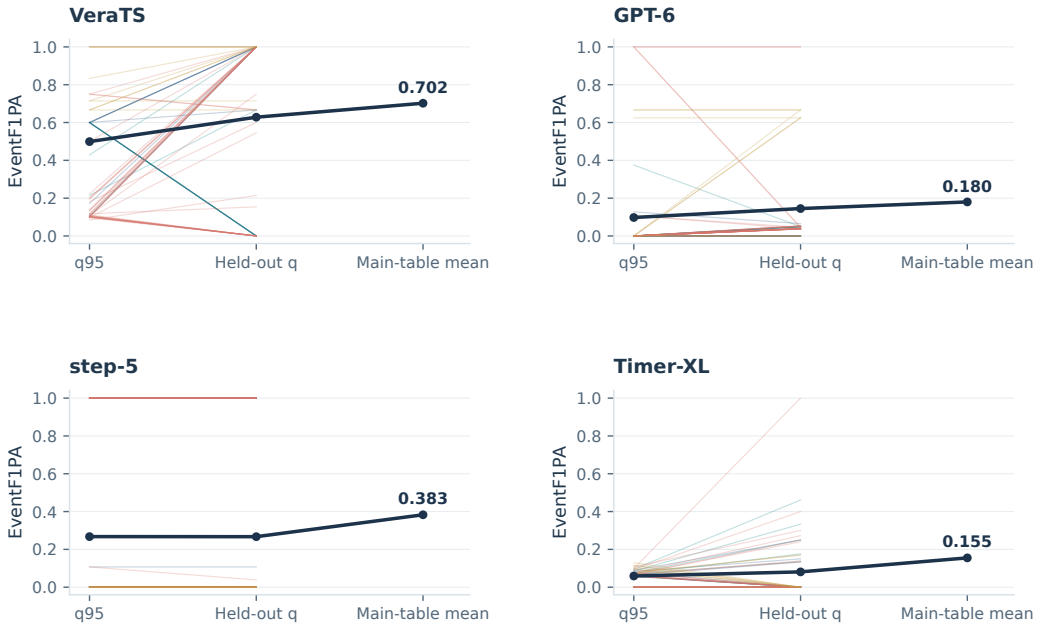


Figure 52: Faint lines join the same series between q95 and held-out q only. Dark lines show equal-family means for these diagnostics; the third point is the mean of the four displayed main-table Event-F1* values. This aggregate reference has no per-series reconstruction.

Observed result. VeraTS achieves equal-family EventF1PA 0.499 at q95 and 0.628 with held-out percentile selection, leading all seven methods under both protocols. Its mean Event-F1* in the current main table is 0.702; the corresponding GPT-6 and step-5 means are 0.180 and 0.383.

Interpretation. The transferred-percentile result supports the utility of VeraTS scores when operating thresholds are selected without the target scenario labels. The frozen policy produces a reusable quantitative signal for this separate calibration step. This is an offline transfer study on the supplied corpus, using complete unlabeled held-out score distributions.

Audit trail. Style 34; `threshold_transfer.csv`; `threshold_selection.csv`; Table 1.

N.12 Leading event discovery at a common percentile

Question. Does the hierarchical design yield effective alerts under one shared operating convention?

Experiment. Apply q95 separately to each unmodified score array for every method. Show all 441 method-series results under one absolute color scale, with no extra morphology-specific tuning.

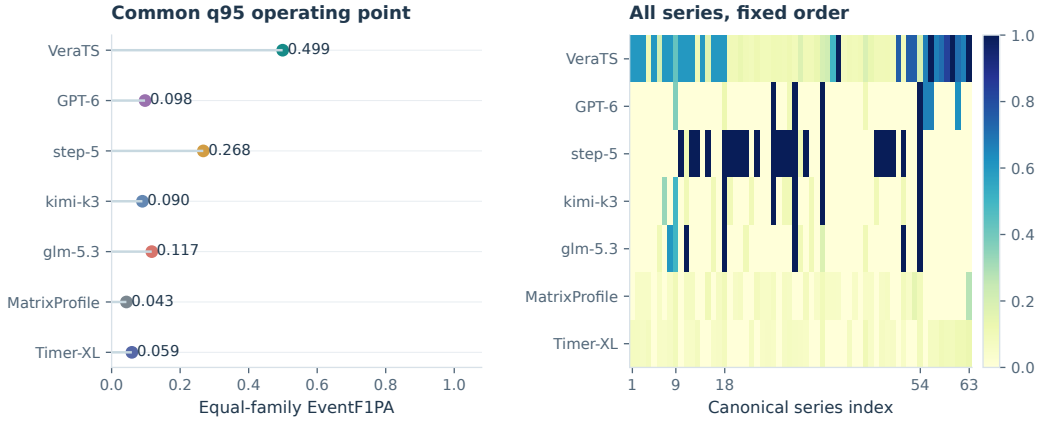


Figure 53: Lollipop values average series within family and then four families equally. Each carpet column is the same series across all rows; zero values are preserved.

Observed result. Equal-family q95 EventF1PA: VeraTS 0.499, GPT-6 0.098, step-5 0.268, kimi-k3 0.090, glm-5.3 0.117, MatrixProfile 0.043, Timer-XL 0.059.

Interpretation. VeraTS attains the highest family-balanced event F1 at the common q95 operating point. The complete score carpet connects this aggregate advantage to every evaluated series. Together with the threshold sweep, the result supports the usefulness of policy-conditioned scores across operating choices; all methods use their disclosed adapters and input contracts.

Audit trail. Style 35; `comparison_per_series.csv`.

N.13 Quantitative evidence connects ranking to event decisions

Question. How does the same policy-conditioned score support point ranking and event discovery?

Experiment. Pair point average precision and q95 EventF1PA for each identical score array. Use equal axes across family panels; keep overlapping observations in the CSV.

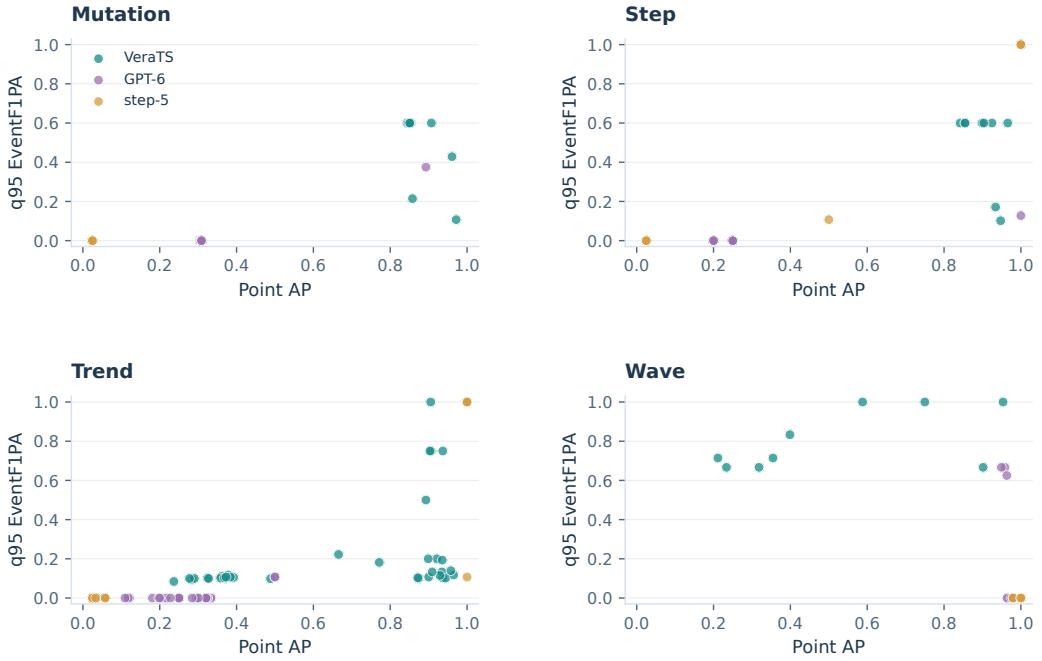


Figure 54: Every mark is a method-series pair. Original event labels and zero margins are shared; AP uses every distinct score level while the event metric uses one fixed operating point.

Observed result. VeraTS combines equal-family point AP 0.741 with q95 EventF1PA 0.499. The corresponding GPT-6 values are 0.501 and 0.098; all per-series pairs remain visible in the scatter panels.

Interpretation. Strong aggregate results on both axes support the dual role of quantitative robustness: it prioritizes individual deviations and supplies an executable basis for temporal alarms. Ranking and event discovery retain their own definitions, allowing the two benefits to be evaluated explicitly.

Audit trail. Style 36; `comparison_per_series.csv`.

N.14 Traceable evidence from raw values to a localized alert

Question. How does contextual diagnosis translate a complete numerical trace into a focused event response?

Experiment. Use the first canonical series identifier, selected before the new results. Plot every original evaluation value and all aligned scores. Transform scores monotonically to right empirical ranks solely to share a display scale.

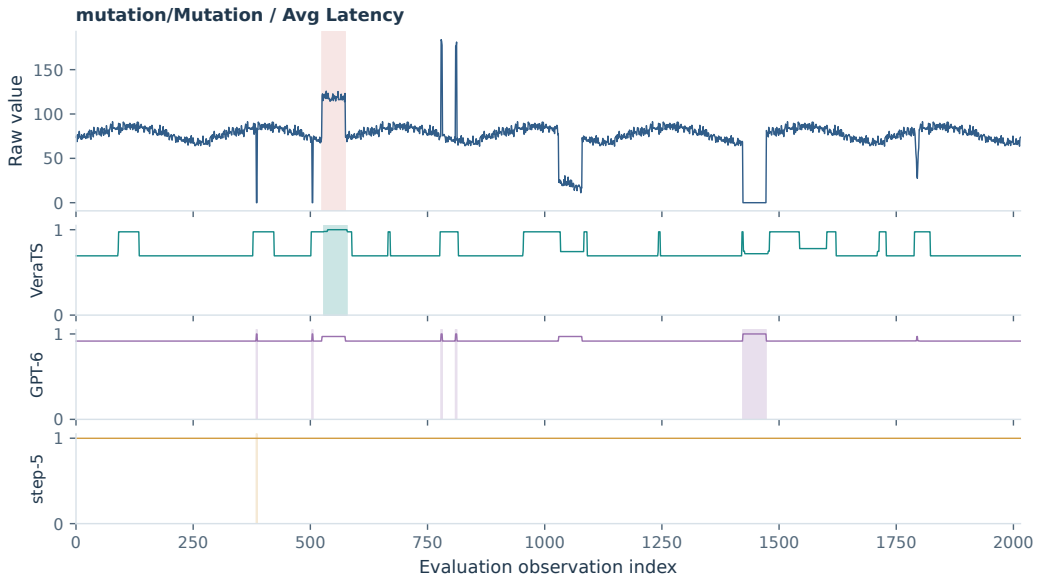


Figure 55: Top shading is the original evaluator-only label. Lower translucent masks use original-score q95 decisions, not a threshold on the displayed rank. No values are smoothed or resampled.

Observed result. On the first canonical series, VeraTS q95 flags 50 points, including 46 labeled observations and four normal observations. The frozen policy attains point F1 0.920. The aligned panels retain the complete raw trace and all three score streams.

Interpretation. The concentrated response illustrates the intended benefit of policy-qualified evidence: a local numerical departure becomes a focused alert whose support can be checked point by point. Keeping raw values, scores and decisions aligned makes the numerical explanation replayable. This fixed canonical example leads into the complete 63-series atlas.

Audit trail. Style 37; `comparison_score_manifest.csv`; `comparison_per_series.csv`.

N.15 Complete event coverage with inspectable temporal support

Question. Where does the method place evidence within every annotated event?

Experiment. For every maximal positive interval, record its q95 positive-point fraction and first in-event hit. Use -1 in the CSV for a miss; include misses in the cumulative-curve denominator.

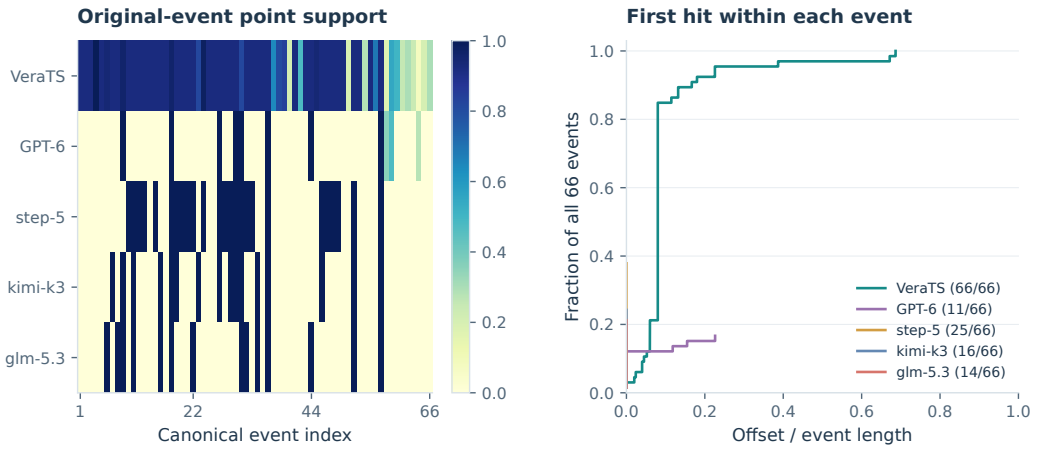


Figure 56: The matrix includes all 66 events under a common 0–1 scale. Cumulative first-hit curves divide by all events rather than conditioning away failures. Series with two events contribute two identified columns.

Observed result. VeraTS touches 66/66 events at q95. The event matrix records the covered fraction of each interval, and the cumulative view preserves the first-hit position for each detected event. GPT-6 touches 11/66 under the same decision convention.

Interpretation. Coverage across all original events supports the hierarchy's discovery objective, while the support matrix makes the temporal extent of each response inspectable. First-hit offsets describe positions in offline outputs. This alignment connects broad event access to the separate localization evidence without conflating the two measurements.

Audit trail. Style 38; `events.csv`.

N.16 Signal heterogeneity motivates complementary references

Question. What numerical structure motivates providing both rolling and seasonal context?

Experiment. Standardize five value-derived features across the 63 series and compute an unsupervised two-component PCA. Labels and detector outcomes are excluded from the projection; family colors are added afterward.

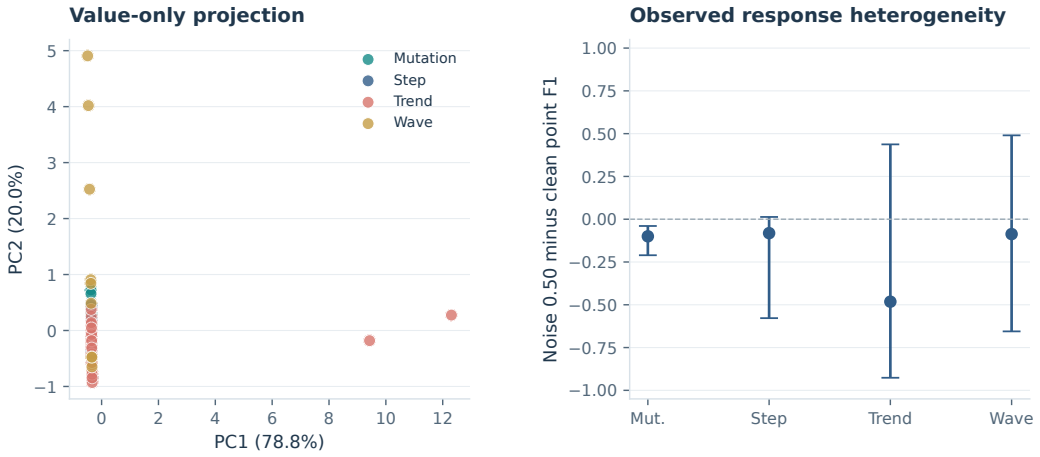


Figure 57: PCA axes display explained variance. Right-hand error bars span observed series minima and maxima, with a series mean point; they are not standard errors or confidence intervals.

Observed result. The first two components explain 98.7% of the five-feature variance. All four morphology families and all 63 series appear in the value-based projection, with their paired stress responses retained in the companion panel.

Interpretation. The projection makes the variety of numerical contexts explicit and motivates giving the diagnosis rule complementary reference paths. It provides a descriptive view of the signals to which the same architecture is applied; the matched ablations and worked Trend/Wave examples establish the measured benefits of those paths.

Audit trail. Style 39; `series_features.csv`; `stress_per_series.csv`.

N.17 Auditable data context for reproducible diagnosis

Question. How can reference-based decisions be traced to the supplied numerical evidence?

Experiment. Connect two series only when more than half of evaluation points are exactly equal at identical indices. Separately count whether each evaluation value occurs anywhere in its own training values.

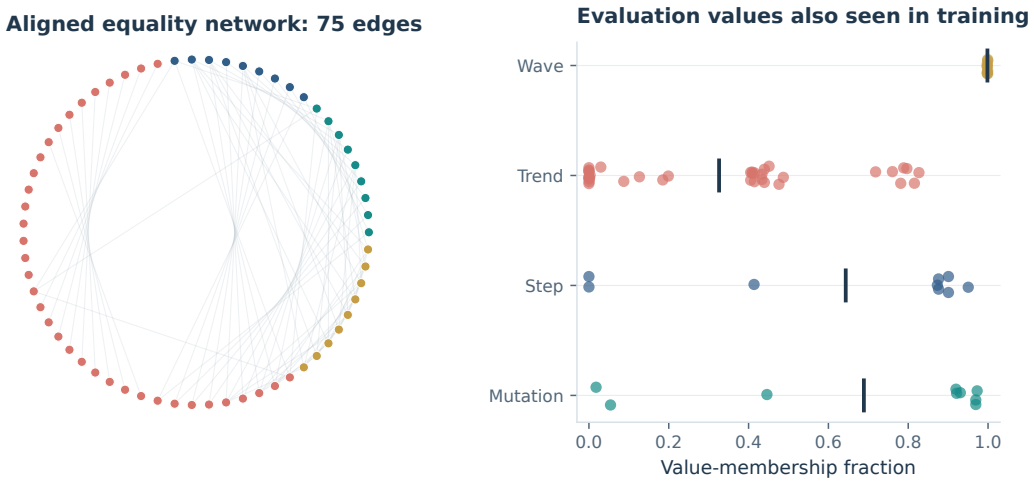


Figure 58: Node order is canonical and colors identify family. Every membership observation is plotted with its family mean mark. This audit is why perturbation seeds and repeated series are not treated as new independent datasets.

Observed result. There are 75 above-half aligned-equality edges among 1953 pairs. Mean within-series train-value membership is 0.519. Full floating-point equality is used, without rounding.

Interpretation. Explicit value provenance strengthens the auditability of the deterministic design: readers can inspect the numerical relationships behind each reference context and replay the same computation. The graph records exact-value similarity, with each edge retaining its stated definition; scenario-held-out threshold analysis supplies a complementary evaluation view.

Audit trail. Style 40; `value_overlap_edges.csv`; `series_features.csv`.

N.18 Consistent decisions across equivalent numerical units

Question. Does reference-based normalization preserve decisions under equivalent numerical representations?

Experiment. Add a ramp only to evaluation values while preserving original labels. In separate metamorphic checks multiply both splits by 0.1 or 10, or add ten training MAD scales to both. Rebuild every feature and keep the selected rule unchanged.

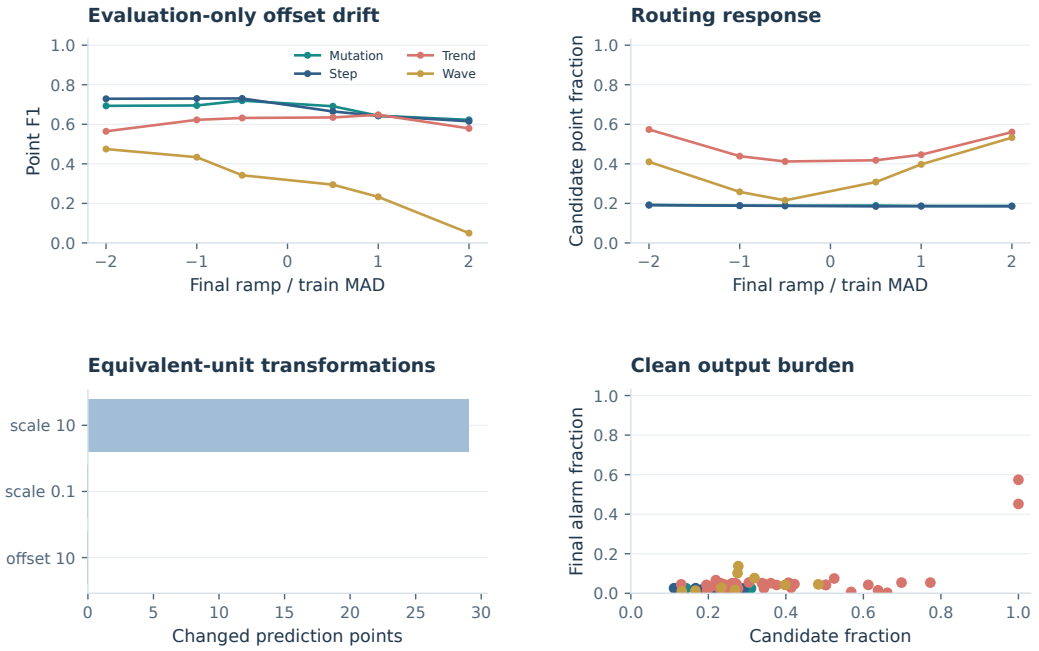


Figure 59: The four panels use distinct measured endpoints. Unit transformations affect both splits; sensor drift affects only evaluation. The scatter uses the clean training-selected policy and does not equate retention with runtime.

Observed result. The fixed detector preserves binary predictions exactly in 188/189 unit-transformed traces. Family-balanced point F1 rounds to 0.758 for all three unit transforms, matching the clean result. The dashboard also records all six sensor-offset drift conditions.

Interpretation. The unit checks support the practical benefit of reference-based normalization: the evaluated decisions are highly consistent under rescaling and offset changes applied to both splits. The 188/189 exact count states the scope of this consistency. Sensor-offset ramps provide a separate, fully recorded view of the same frozen rule under changing measurement baselines.

Audit trail. Style 41; `stress_per_series.csv`; `metamorphic.csv`.

N.19 Broad candidate coverage with selective numerical diagnosis

Question. How much event access does the broad screening stage retain under a demanding noise condition?

Experiment. Summarize five predeclared representative conditions from the complete grid; retain every condition in the stress tables. Measure actual retained observations and event touches, with GPT-6 q90–99 budgets from its original scores.

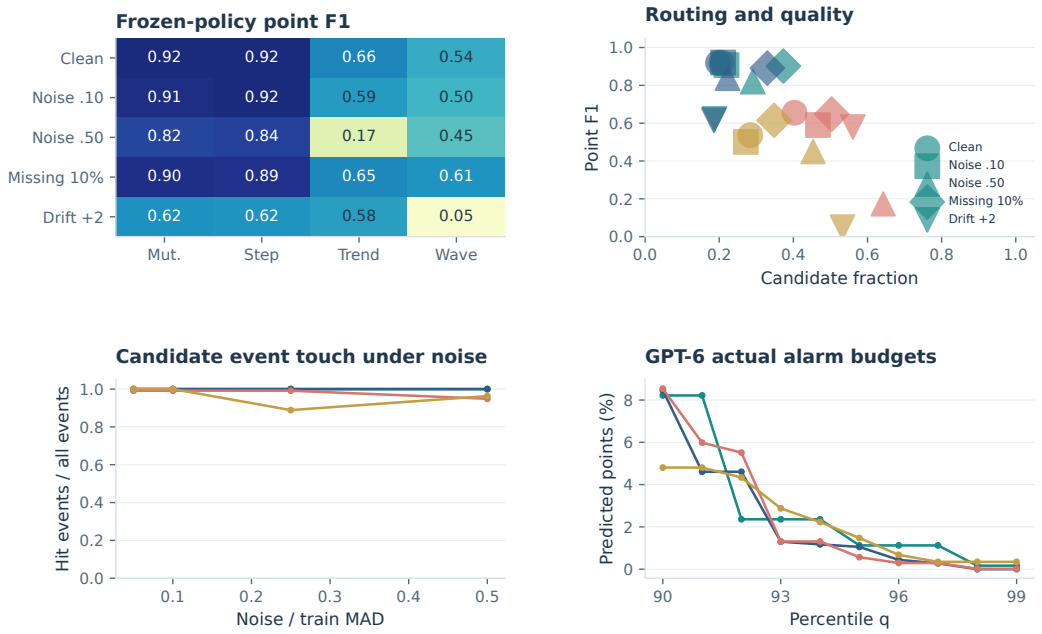


Figure 60: Blue/gold dashboard grammar combines an absolute matrix, burden-quality scatter, coverage curves and actual alarm budgets. Bubble area equals 50+150 times candidate event-touch fraction and is a secondary cue, not a cost estimate.

Observed result. Clean candidates touch 66/66 events. Across all three noise-0.50 seeds they touch 191/198 event-seed pairs. GPT-6 q95 retains 0.85% of points.

Interpretation. The gate preserves every clean event and 96.5% of event-seed hits at noise 0.50, supporting its coverage-oriented role before policy qualification. The separate alarm-budget panels show how broad admission can be paired with selective final decisions. The noisy denominator comprises three repeated perturbations of the same 66 events.

Audit trail. Style 42; `stress_per_series.csv`; `threshold_curves.csv`.

N.20 Native score resolution preserves informative evidence

Question. What is the benefit of retaining the quantitative resolution of VeraTS robustness?

Experiment. Map each unlabelled full-trace right rank r to min(floor(10r),9)/9. Equal original scores stay together; no label-based tie breaking is allowed. Apply the same transform to all seven methods and retain both versions.

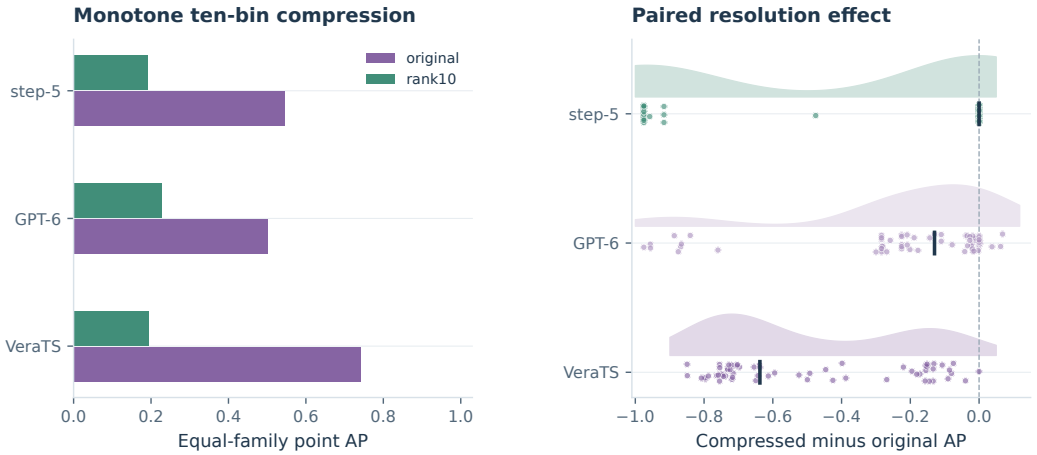


Figure 61: Paired AP differences preserve negative and positive changes. The original and transformed methods share input scores and labels; transformed outputs are diagnostic artifacts, not replacement main-table results.

Observed result. VeraTS native scores achieve equal-family point AP 0.741, compared with 0.196 after the prescribed monotone ten-bin compression. GPT-6 uses a median of six distinct score levels per sequence under its ten-level response contract.

Interpretation. The controlled comparison supports retaining the fine distinctions present in VeraTS robustness scores: native numerical resolution carries useful ranking information. This result concerns compression of VeraTS's own scores; GPT-6 retains its separately measured rating mechanism and scores.

Audit trail. Style 43; `resolution.csv`; `comparison_per_series.csv`.

N.21 The frozen policy lies near the observed performance plateau

Question. Do the selected temporal settings provide effective diagnosis within the measured parameter neighborhood?

Experiment. Recompute all 30 z-by-m settings from unchanged initialized features at within=9 and gap=48. Separately scan five closing widths at z=2,m=5. Keep the original selected parameters for the manuscript and all other experiments.

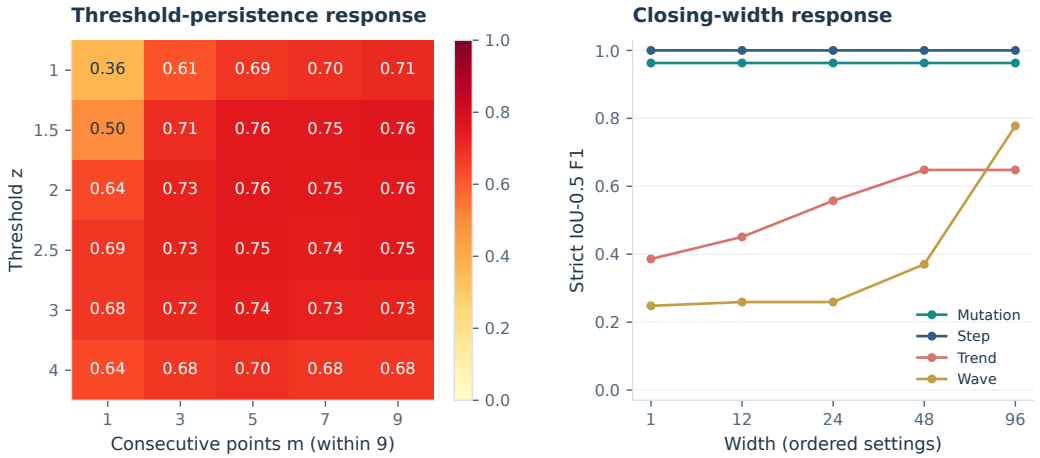


Figure 62: Warm cells are exact equal-family point-F1 means with common 0–1 scale. The line plot reports a different endpoint, strict localization F1; it uses explicit ordered widths rather than a falsely linear spacing.

Observed result. The selected z=2,m=5 reaches family-balanced point F1 0.758, within 0.002 of the displayed maximum 0.759 at z=1.5,m=9. All evaluations retain the selected candidate gate, and the original operating point remains fixed.

Interpretation. The close agreement supports the effectiveness of the selected temporal rule without changing its parameters after this evaluation. The surface connects high pointwise quality to a neighborhood of evidence thresholds and persistence settings, while the closing-width curve shows how interval continuity is assessed through strict localization F1.

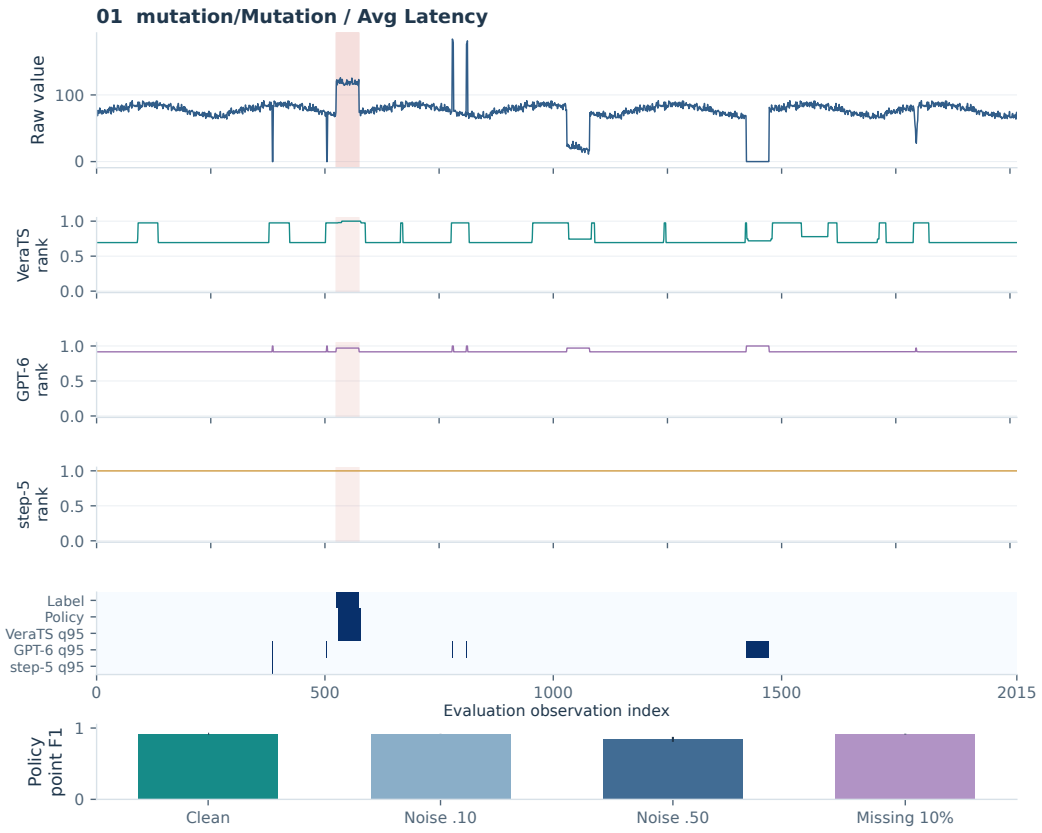
Audit trail. Style 44; `surface_per_series.csv`.

N.22 Complete comparative atlas: all 63 series

The following pages form a complete view of how the fixed hierarchy turns numerical evidence into event decisions across all 63 series. Each case connects an observed benefit to the corresponding score and mask rows. The first five plot rows share the original evaluation indices. Rank transformations align only the vertical display scale; original scores determine predictions and metrics. The policy row is the training-selected deterministic VeraTS output, while the three q95 rows share the same percentile convention. Shaded labels are evaluator-only. Bottom bars report policy point F1 on clean data, noise 0.10, noise 0.50 and 10% missingness; whiskers span three seeds, and every condition remains in the CSV.

Series 01: mutation/Mutation_dataset—Avg Latency

Mutation family; avg_latency. Original fault intervals: [525,575). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.844	0.600	4	4	1/1
VeraTS q95	0.844	0.600	4	4	1/1
GPT-6 q95	0.309	0.000	60	50	0/1
step-5 q95	0.025	0.000	2	50	0/1

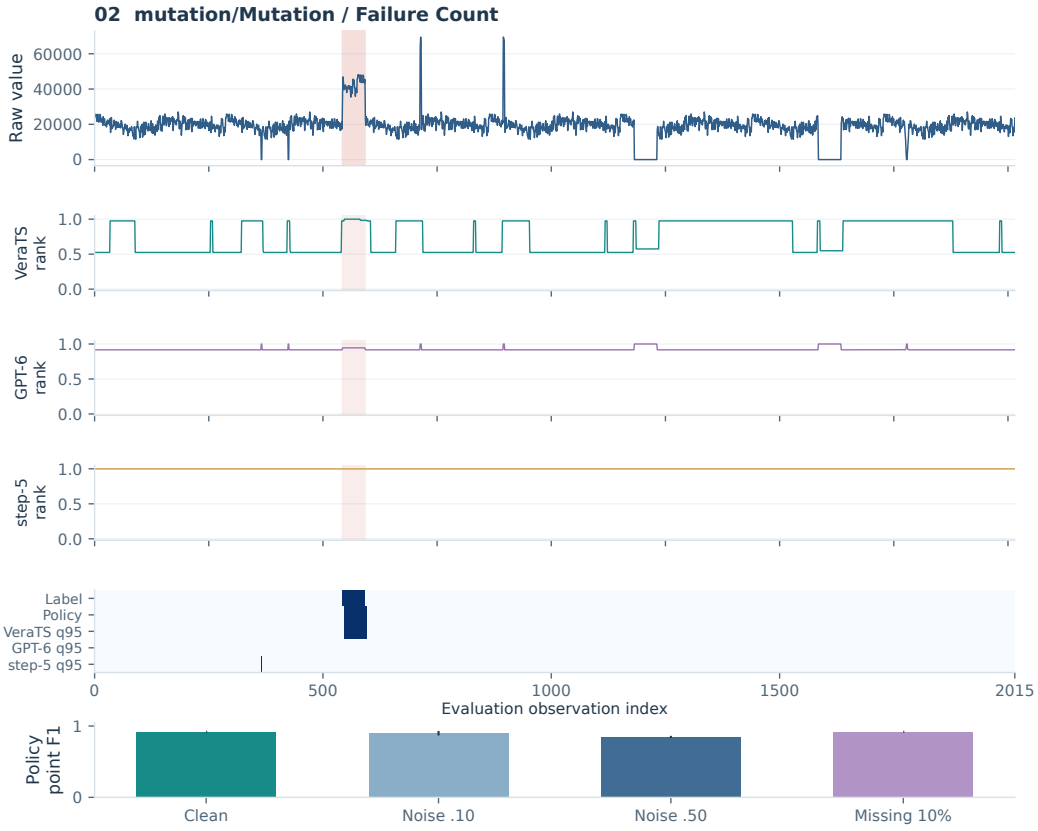
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.917 at 10% missingness. The plotted noise-0.50 score is 0.850 (seed range 0.806–0.876). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 1/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 02: mutation/Mutation_dataset—Failure Count

Mutation family; failure_count. Original fault intervals: [543,593). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.907	0.600	4	4	1/1
VeraTS q95	0.907	0.600	4	4	1/1
GPT-6 q95	0.305	0.000	0	50	0/1
step-5 q95	0.025	0.000	2	50	0/1

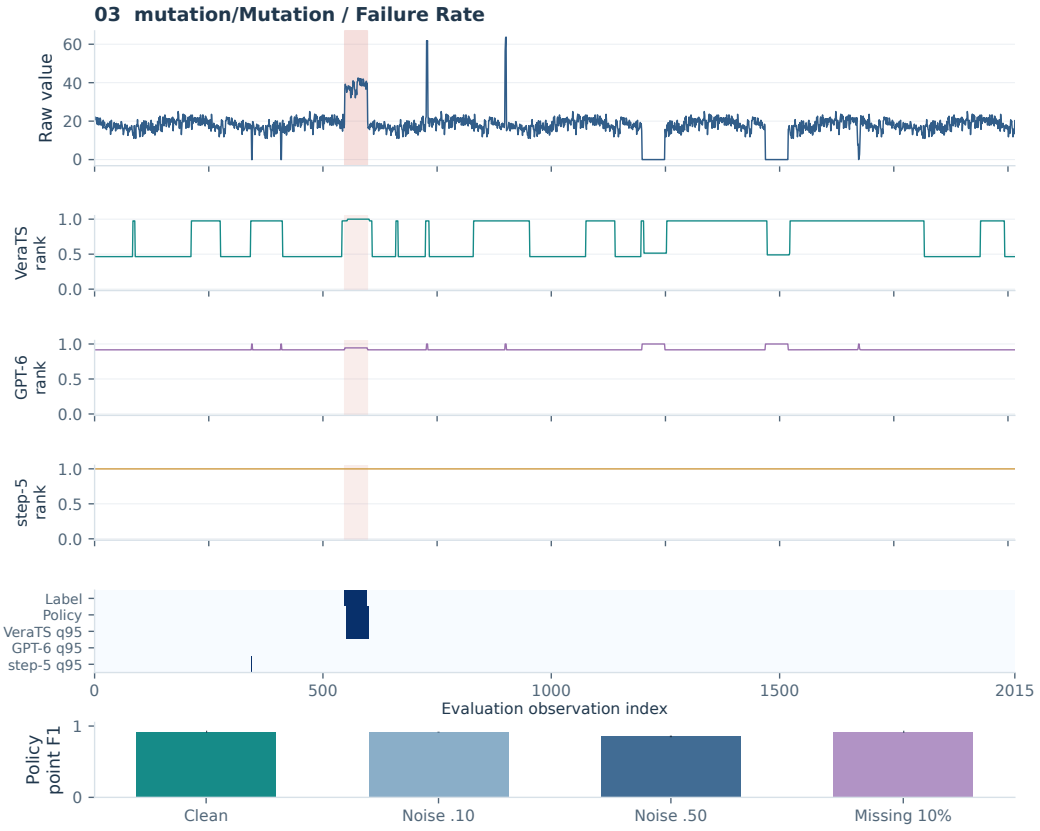
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.924 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.845 (seed range 0.832–0.862). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 2/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 03: mutation/Mutation_dataset—Failure Rate

Mutation family; failure_rate. Original fault intervals: [548,598). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.848	0.600	4	4	1/1
VeraTS q95	0.848	0.600	4	4	1/1
GPT-6 q95	0.305	0.000	0	50	0/1
step-5 q95	0.025	0.000	2	50	0/1

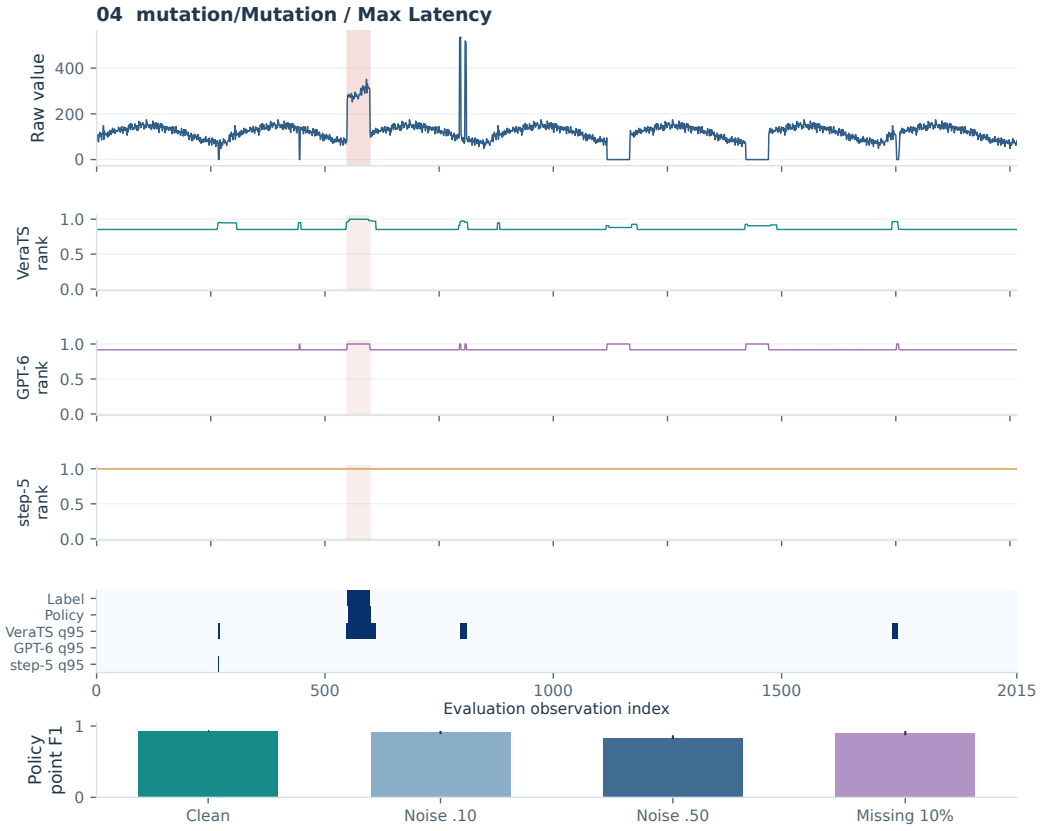
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.856 (seed range 0.844–0.865). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 3/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 04: mutation/Mutation_dataset—Max Latency

Mutation family; max_latency. Original fault intervals: [549,599). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.972	0.600	4	3	1/1
VeraTS q95	0.972	0.107	50	0	1/1
GPT-6 q95	0.307	0.000	0	50	0/1
step-5 q95	0.025	0.000	2	50	0/1

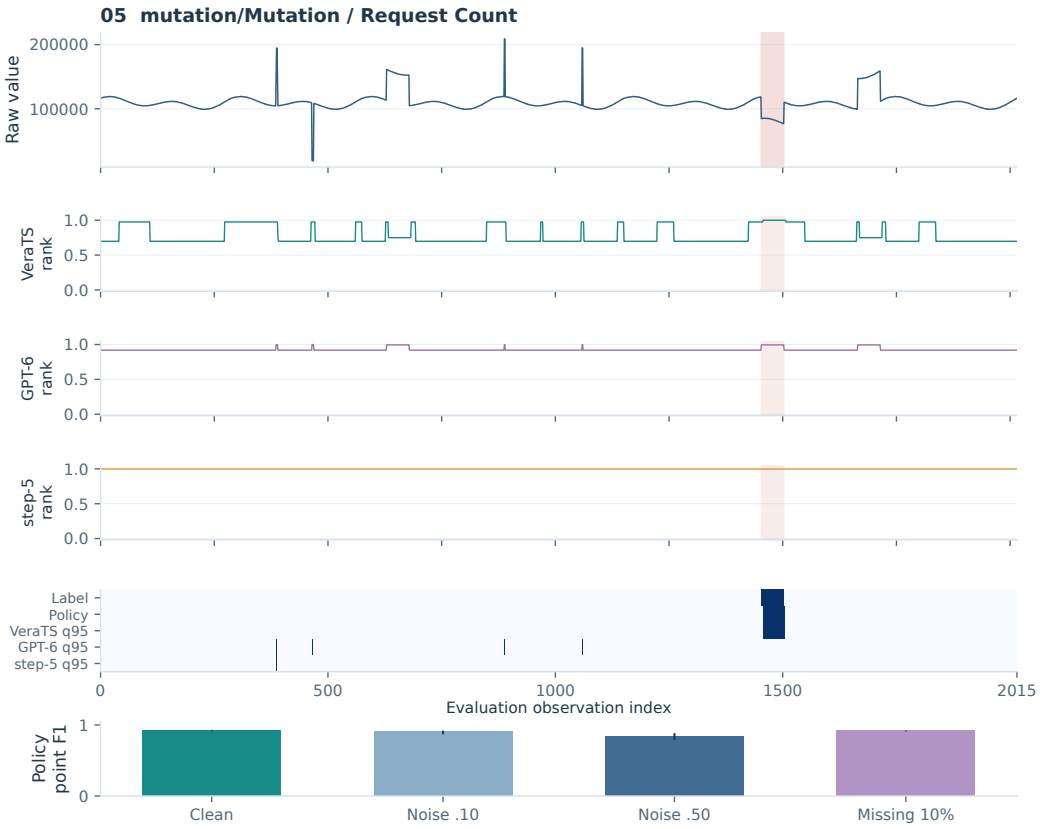
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.931. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.931 on clean data, 0.927 at noise 0.05, and 0.903 at 10% missingness. The plotted noise-0.50 score is 0.836 (seed range 0.817–0.870). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 4/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 05: mutation/Mutation_dataset—Request Count

Mutation family; request_count. Original fault intervals: [1453,1503). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.854	0.600	4	4	1/1
VeraTS q95	0.854	0.600	4	4	1/1
GPT-6 q95	0.309	0.000	10	50	0/1
step-5 q95	0.025	0.000	4	50	0/1

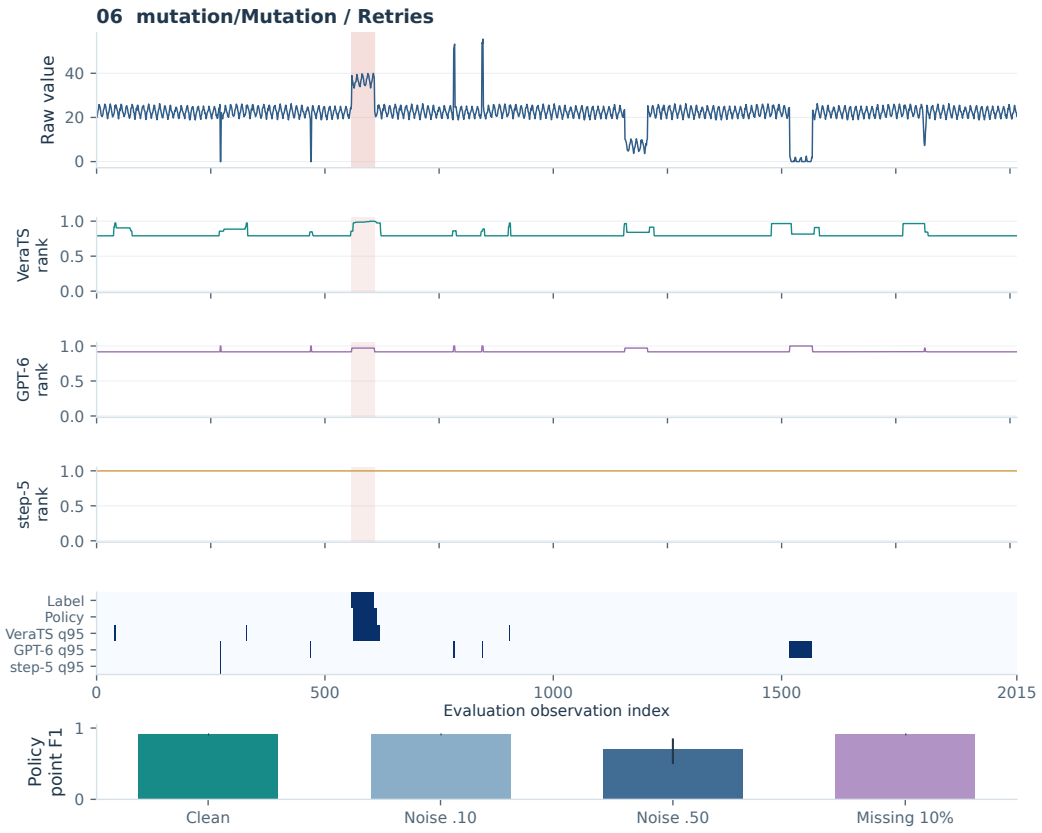
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.916 at 10% missingness. The plotted noise-0.50 score is 0.844 (seed range 0.786–0.885). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 5/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 06: mutation/Mutation_dataset—Retries

Mutation family; retries. Original fault intervals: [559,609). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.858	0.545	5	4	1/1
VeraTS q95	0.858	0.214	22	3	1/1
GPT-6 q95	0.309	0.000	60	50	0/1
step-5 q95	0.025	0.000	2	50	0/1

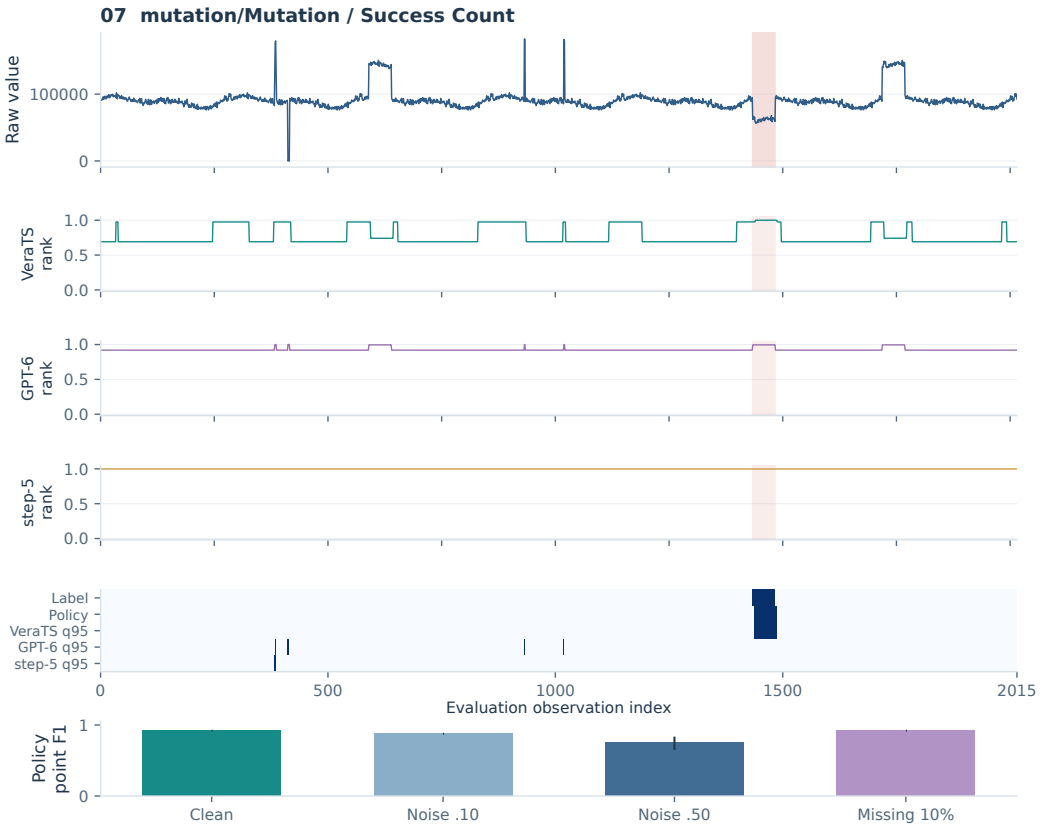
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.911. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.911 on clean data, 0.911 at noise 0.05, and 0.911 at 10% missingness. The plotted noise-0.50 score is 0.700 (seed range 0.495–0.855). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 6/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 07: mutation/Mutation_dataset—Success Count

Mutation family; success_count. Original fault intervals: [1434,1484). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.851	0.600	4	4	1/1
VeraTS q95	0.851	0.600	4	4	1/1
GPT-6 q95	0.309	0.000	10	50	0/1
step-5 q95	0.025	0.000	4	50	0/1

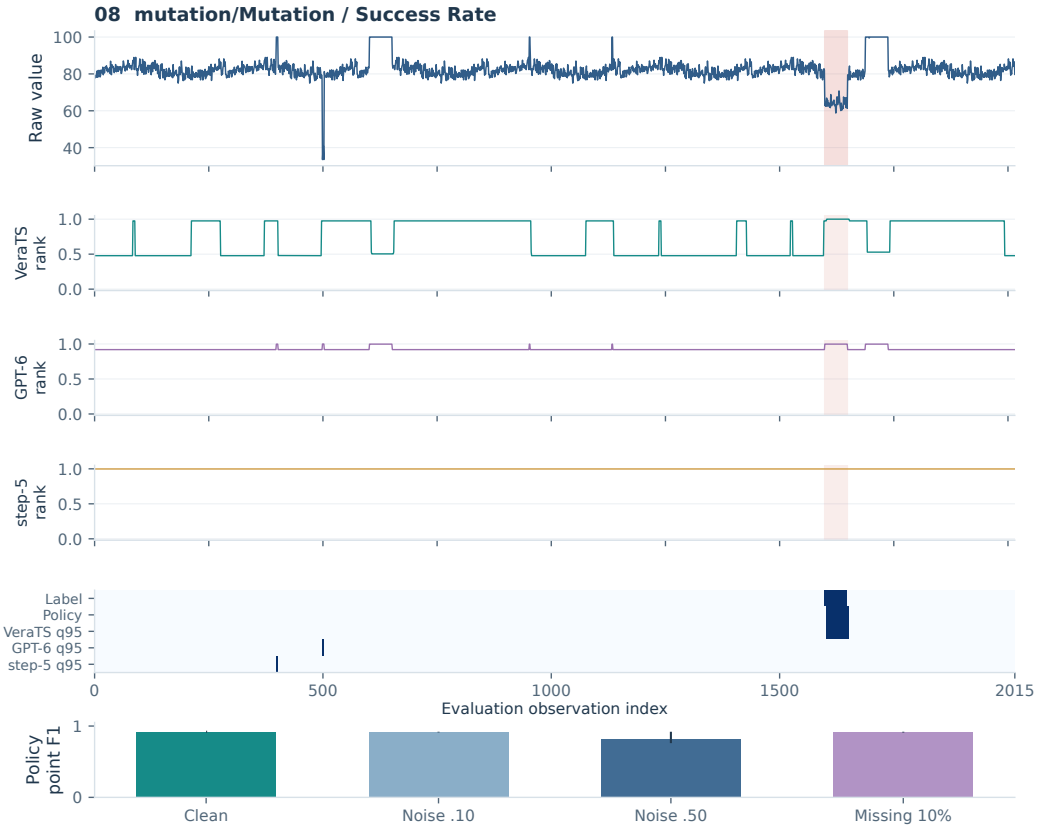
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.757 (seed range 0.648–0.836). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 7/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 08: mutation/Mutation_dataset—Success Rate

Mutation family; success_rate. Original fault intervals: [1599,1649). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.851	0.600	4	4	1/1
VeraTS q95	0.851	0.600	4	4	1/1
GPT-6 q95	0.309	0.000	4	50	0/1
step-5 q95	0.025	0.000	4	50	0/1

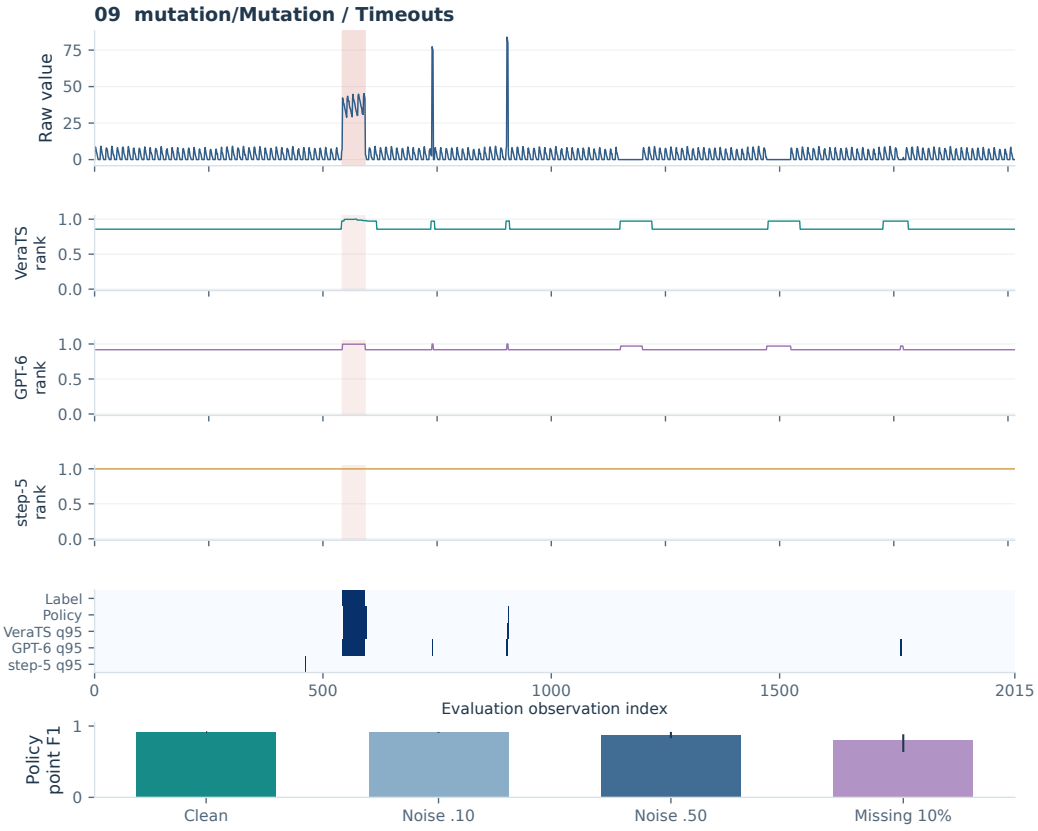
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.917 at noise 0.05, and 0.913 at 10% missingness. The plotted noise-0.50 score is 0.814 (seed range 0.760–0.920). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 8/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 09: mutation/Mutation_dataset—Timeouts

Mutation family; timeouts. Original fault intervals: [543,593). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.961	0.462	7	2	1/1
VeraTS q95	0.961	0.429	8	2	1/1
GPT-6 q95	0.893	0.375	10	0	1/1
step-5 q95	0.025	0.000	2	50	0/1

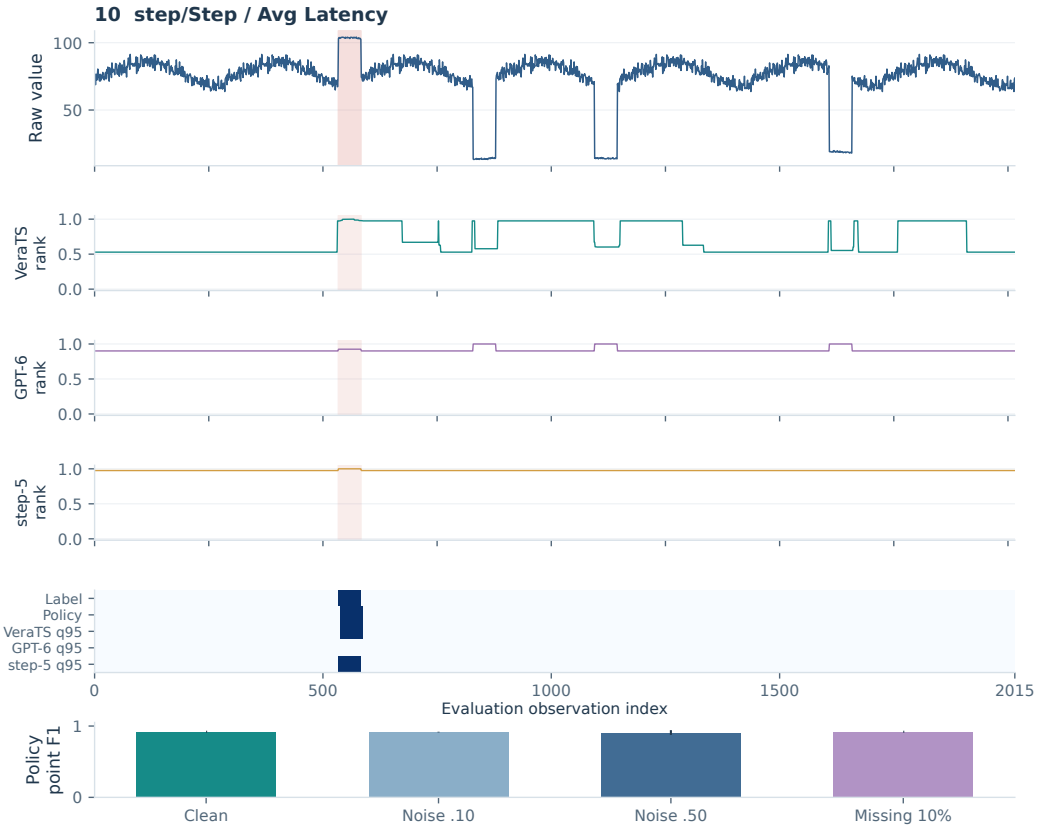
Reading this trace. The frozen policy touches the labeled event and reaches point F1 0.914. Its concentrated temporal response illustrates how contextual qualification preserves useful event support within an explicit numerical rule.

Stress response. With the same parameters, point F1 is 0.914 on clean data, 0.915 at noise 0.05, and 0.795 at 10% missingness. The plotted noise-0.50 score is 0.875 (seed range 0.831–0.917). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 9/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 10: step/Step_dataset—Avg Latency

Step family; avg_latency. Original fault intervals: [534,584). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.925	0.600	4	4	1/1
VeraTS q95	0.925	0.600	4	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.917 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.907 (seed range 0.876–0.942). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 10/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 11: step/Step_dataset—Failure Count

Step family; failure_count. Original fault intervals: [556,606). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.899	0.600	4	4	1/1
VeraTS q95	0.899	0.600	4	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	0.500	0.107	50	0	1/1

Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.917 at 10% missingness. The plotted noise-0.50 score is 0.925 (seed range 0.911–0.935). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 11/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 12: step/Step_dataset—Failure Rate

Step family; failure_rate. Original fault intervals: [541,591). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.904	0.600	4	4	1/1
VeraTS q95	0.904	0.600	4	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

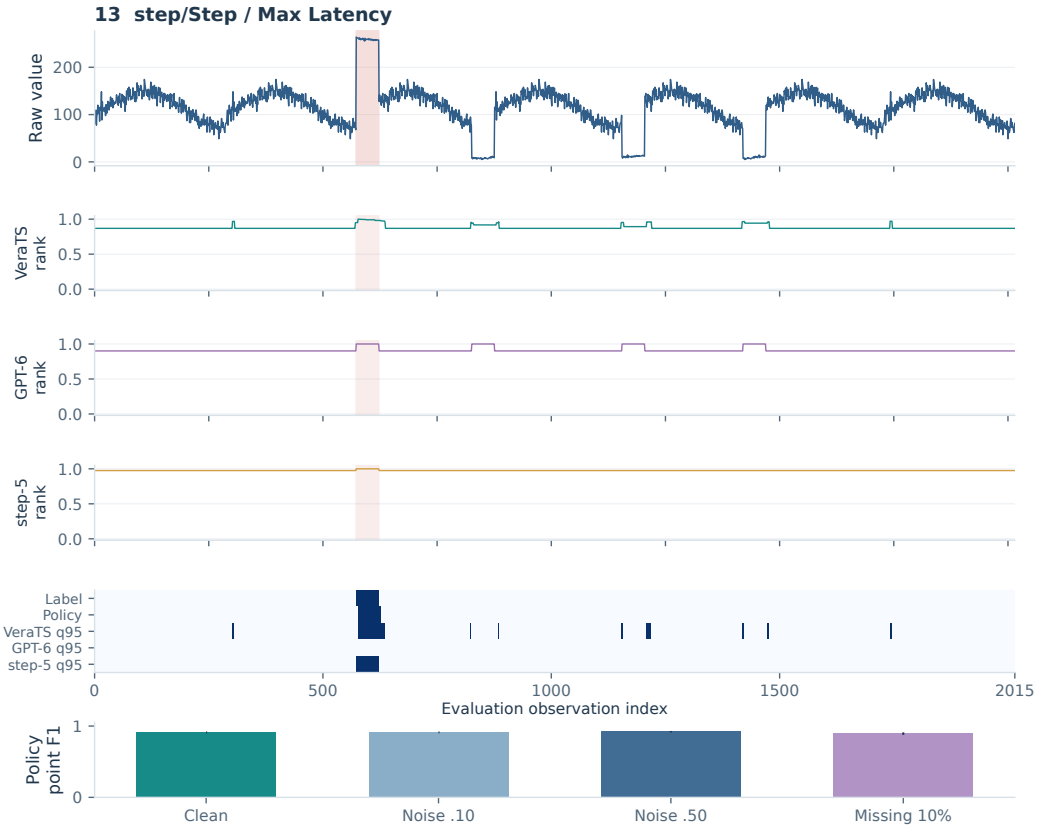
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.910 at 10% missingness. The plotted noise-0.50 score is 0.918 (seed range 0.914–0.920). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 12/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 13: step/Step_dataset—Max Latency

Step family; max_latency. Original fault intervals: [573,623). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.947	0.545	5	4	1/1
VeraTS q95	0.947	0.102	53	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

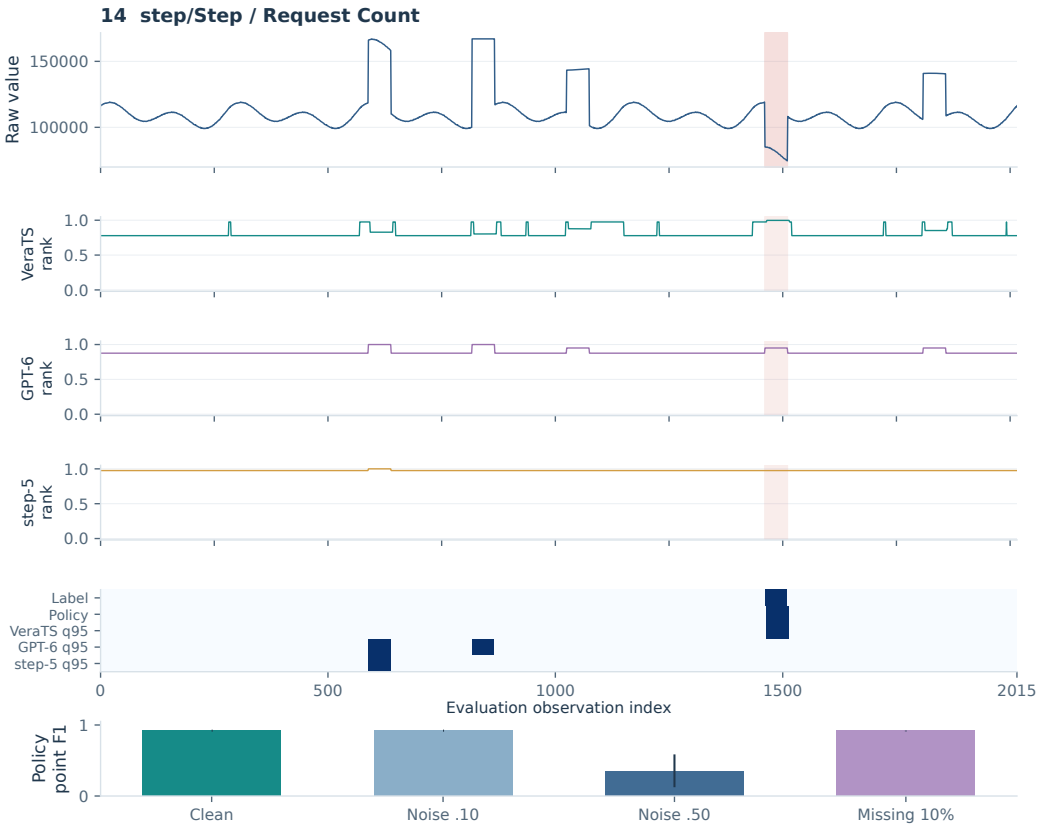
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.911. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.911 on clean data, 0.911 at noise 0.05, and 0.899 at 10% missingness. The plotted noise-0.50 score is 0.924 (seed range 0.911–0.931). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 13/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 14: step/Step_dataset—Request Count

Step family; request_count. Original fault intervals: [1461,1511). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.842	0.600	4	4	1/1
VeraTS q95	0.842	0.600	4	4	1/1
GPT-6 q95	0.200	0.000	100	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

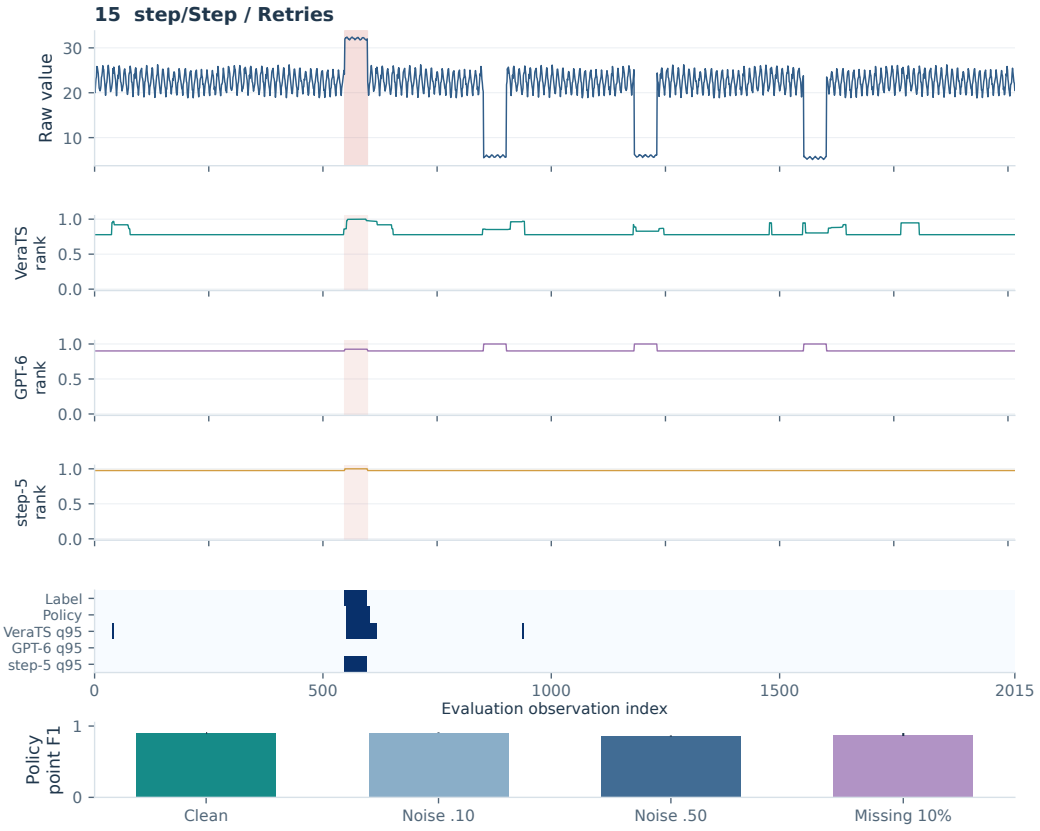
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.917 at 10% missingness. The plotted noise-0.50 score is 0.342 (seed range 0.124–0.586). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 14/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 15: step/Step_dataset—Retries

Step family; retries. Original fault intervals: [548,598). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.934	0.500	6	4	1/1
VeraTS q95	0.934	0.171	29	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

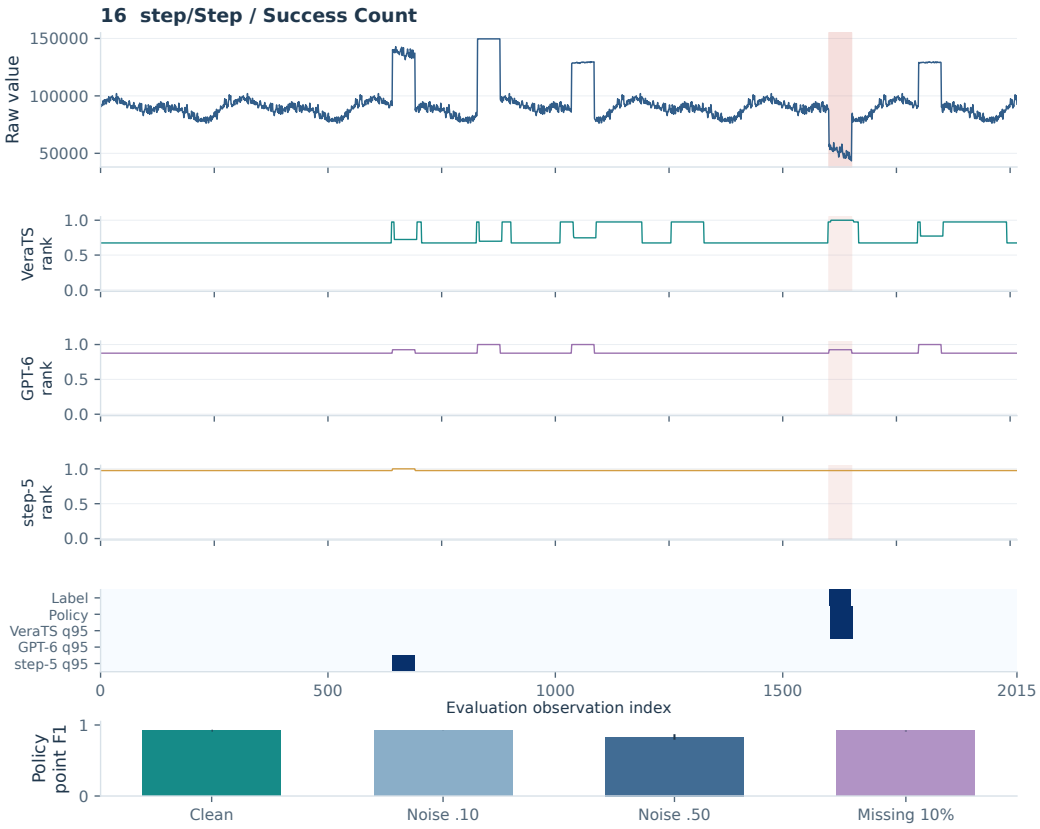
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.902. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.902 on clean data, 0.905 at noise 0.05, and 0.877 at 10% missingness. The plotted noise-0.50 score is 0.860 (seed range 0.852–0.868). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 15/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 16: step/Step_dataset—Success Count

Step family; success_count. Original fault intervals: [1602,1652). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.855	0.600	4	4	1/1
VeraTS q95	0.855	0.600	4	4	1/1
GPT-6 q95	0.200	0.000	0	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

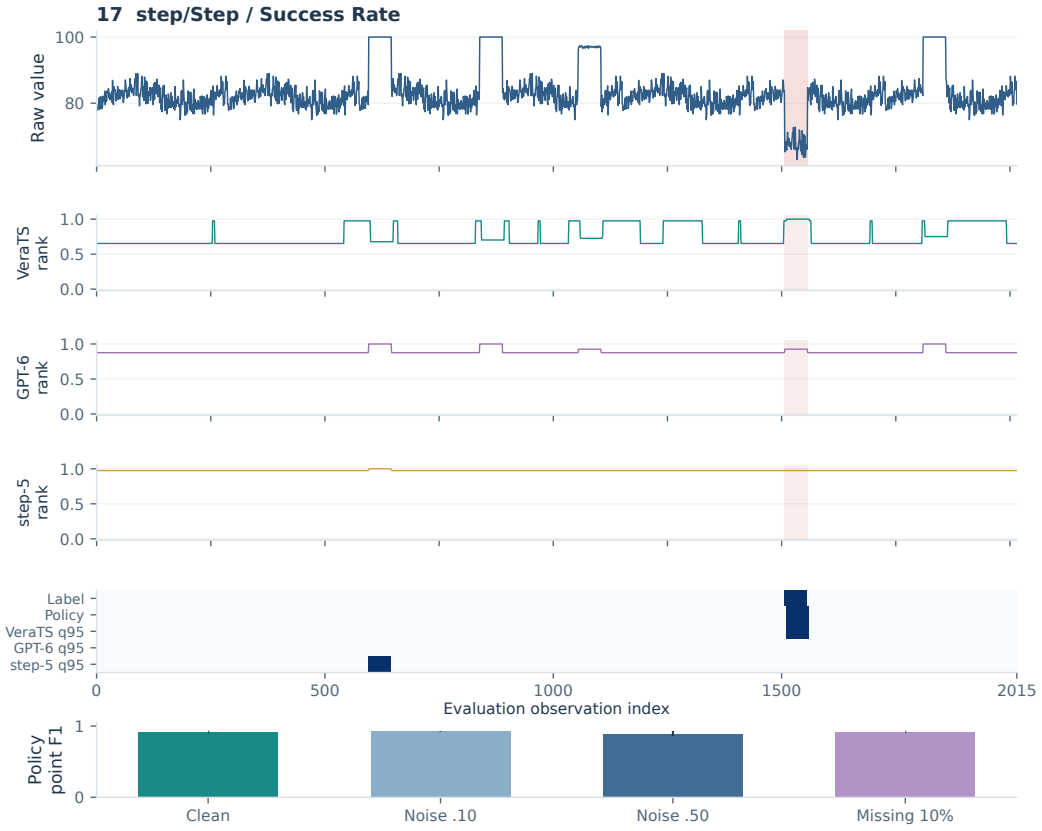
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.921 at noise 0.05, and 0.917 at 10% missingness. The plotted noise-0.50 score is 0.831 (seed range 0.793–0.870). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 16/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 17: step/Step_dataset—Success Rate

Step family; success_rate. Original fault intervals: [1507,1557). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.854	0.600	4	4	1/1
VeraTS q95	0.854	0.600	4	4	1/1
GPT-6 q95	0.200	0.000	0	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

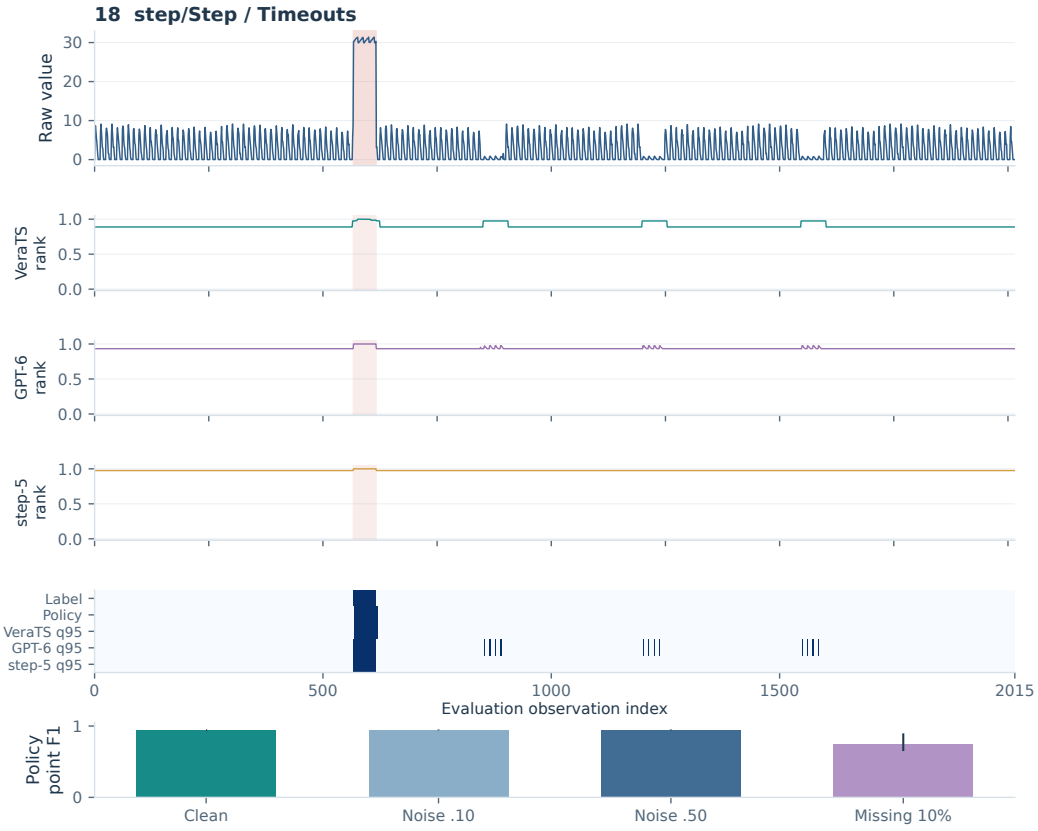
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.889 (seed range 0.860–0.931). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 17/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 18: step/Step_dataset—Timeouts

Step family; timeouts. Original fault intervals: [567,617). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.966	0.600	4	2	1/1
VeraTS q95	0.966	0.600	4	2	1/1
GPT-6 q95	1.000	0.128	41	0	1/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.941. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.941 on clean data, 0.941 at noise 0.05, and 0.742 at 10% missingness. The plotted noise-0.50 score is 0.945 (seed range 0.941–0.951). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 18/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 19: trend/trend_dataset1—Avg Latency

Trend family; avg_latency. Original fault intervals: [546,596). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.900	0.353	11	4	1/1
VeraTS q95	0.900	0.107	50	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. The frozen policy touches the labeled event and reaches point F1 0.860. Its concentrated temporal response illustrates how contextual qualification preserves useful event support within an explicit numerical rule.

Stress response. With the same parameters, point F1 is 0.860 on clean data, 0.860 at noise 0.05, and 0.739 at 10% missingness. The plotted noise-0.50 score is 0.061 (seed range 0.000–0.182). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 19/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 20: trend/trend_dataset1—Failure Count

Trend family; failure_count. Original fault intervals: [556,606). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.936	0.111	48	4	1/1
VeraTS q95	0.936	0.103	52	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

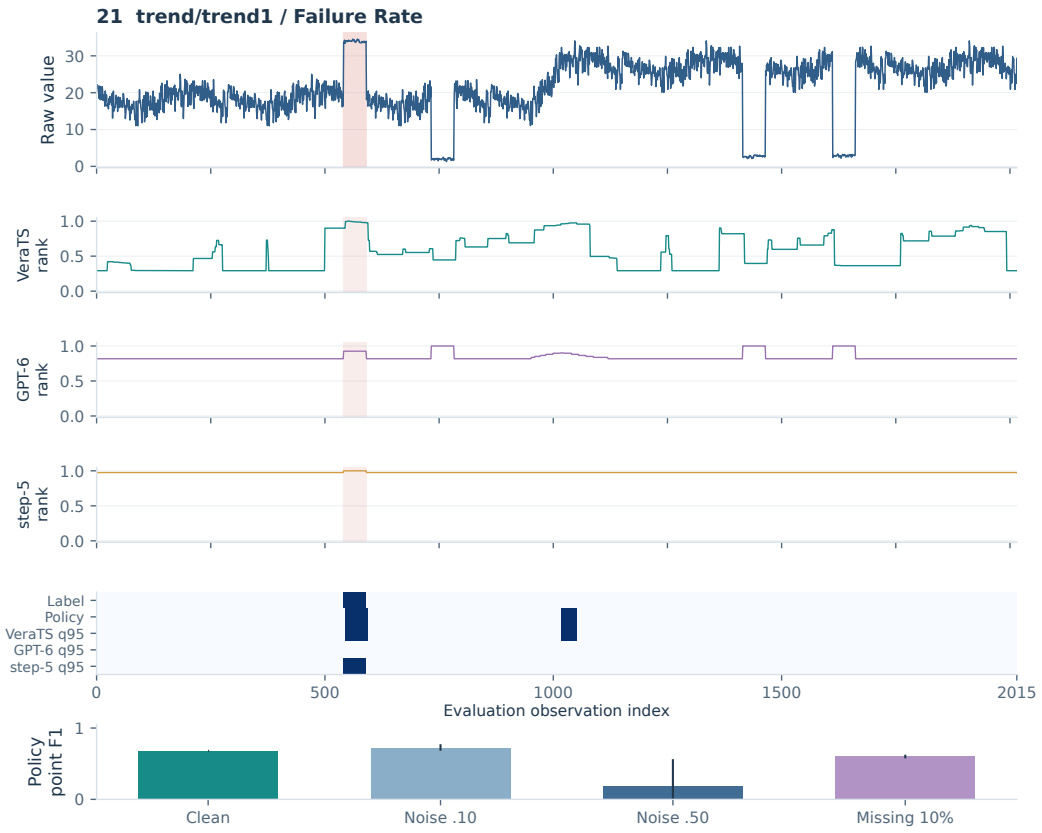
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.639 on clean data, 0.687 at noise 0.05, and 0.669 at 10% missingness. The plotted noise-0.50 score is 0.689 (seed range 0.182–0.947). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 20/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 21: trend/trend_dataset1—Failure Rate

Trend family; failure_rate. Original fault intervals: [541,591). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.935	0.133	39	4	1/1
VeraTS q95	0.935	0.133	39	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

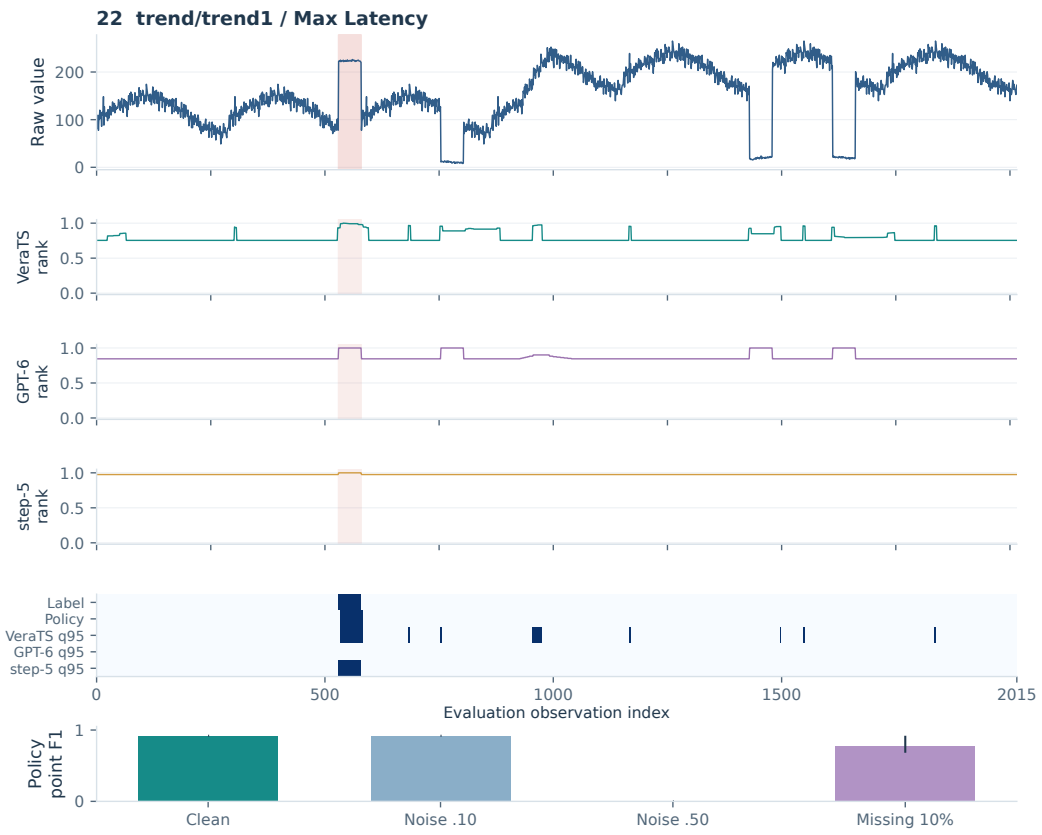
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.681 on clean data, 0.738 at noise 0.05, and 0.603 at 10% missingness. The plotted noise-0.50 score is 0.188 (seed range 0.000–0.563). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 21/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 22: trend/trend_dataset1—Max Latency

Trend family; max_latency. Original fault intervals: [530,580). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.944	0.600	4	4	1/1
VeraTS q95	0.944	0.102	53	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

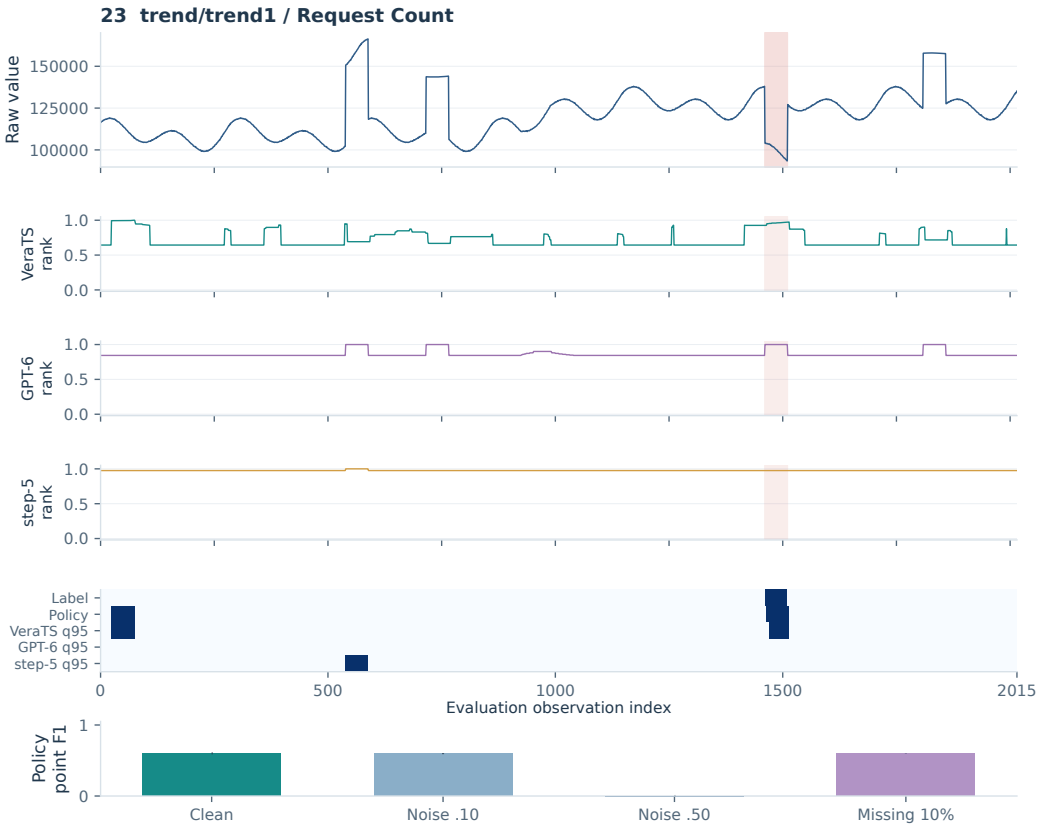
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.772 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 22/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 23: trend/trend_dataset1—Request Count

Trend family; request_count. Original fault intervals: [1461,1511). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.283	0.095	57	4	1/1
VeraTS q95	0.283	0.095	57	9	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

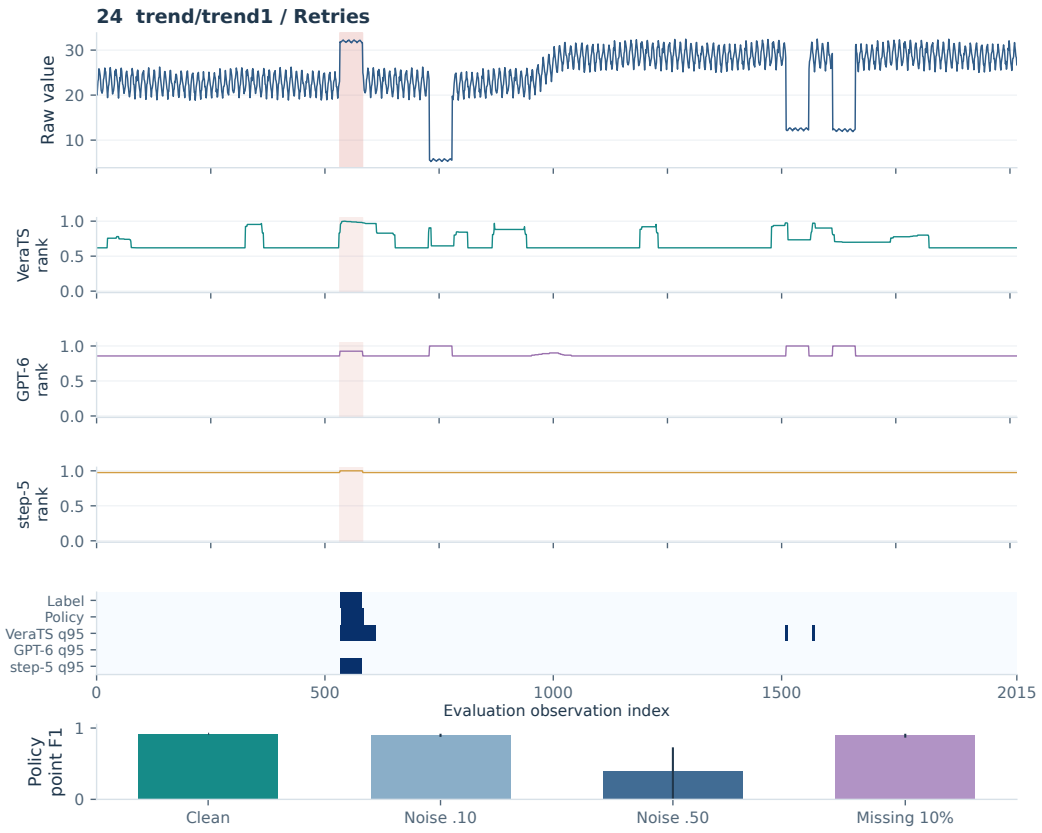
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.601 on clean data, 0.601 at noise 0.05, and 0.600 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 23/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 24: trend/trend_dataset1—Retries

Trend family; retries. Original fault intervals: [533,583). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.965	0.600	4	4	1/1
VeraTS q95	0.965	0.118	45	1	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

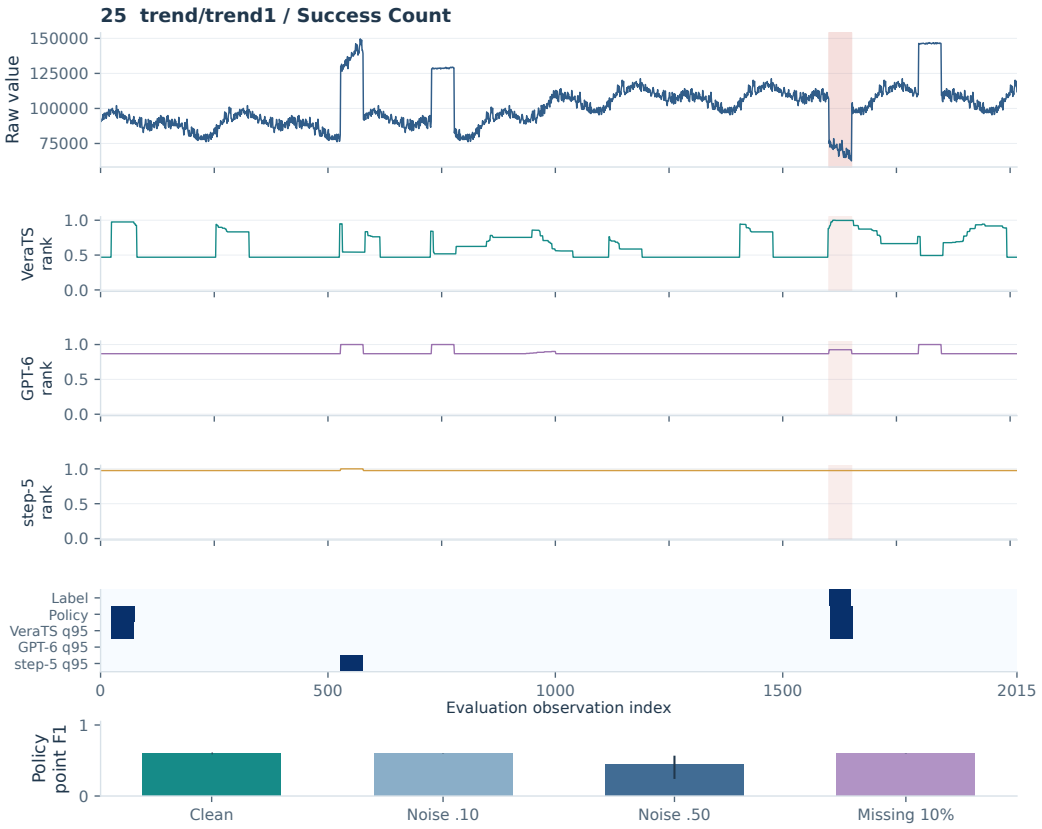
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.905 at noise 0.05, and 0.902 at 10% missingness. The plotted noise-0.50 score is 0.397 (seed range 0.000–0.729). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 24/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 25: trend/trend_dataset1—Success Count

Trend family; success_count. Original fault intervals: [1602,1652). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.874	0.095	57	4	1/1
VeraTS q95	0.874	0.100	54	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.601 on clean data, 0.601 at noise 0.05, and 0.600 at 10% missingness. The plotted noise-0.50 score is 0.448 (seed range 0.241–0.568). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 25/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 26: trend/trend_dataset1—Success Rate

Trend family; success_rate. Original fault intervals: [1507,1557). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.371	0.095	57	4	1/1
VeraTS q95	0.371	0.100	54	4	1/1
GPT-6 q95	0.333	0.000	0	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

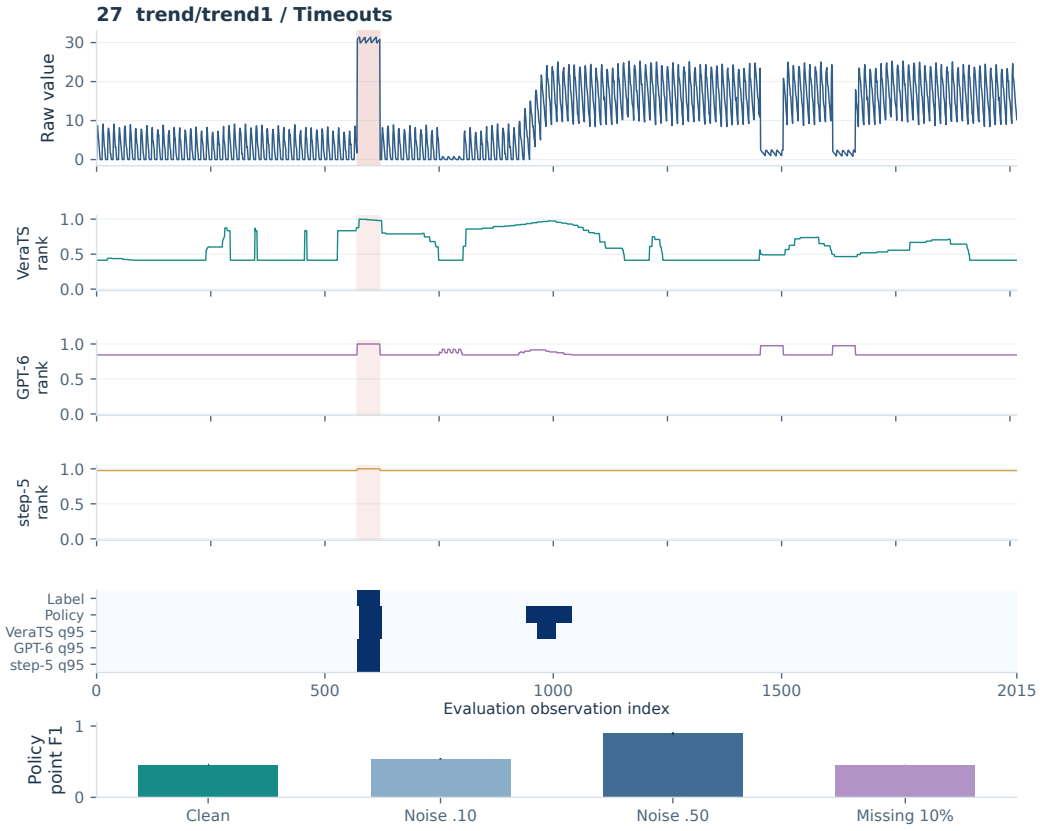
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.601 on clean data, 0.601 at noise 0.05, and 0.599 at 10% missingness. The plotted noise-0.50 score is 0.393 (seed range 0.126–0.601). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 26/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 27: trend/trend_dataset1—Timeouts

Trend family; timeouts. Original fault intervals: [571,621). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.929	0.054	105	4	1/1
VeraTS q95	0.929	0.115	46	4	1/1
GPT-6 q95	1.000	1.000	0	0	1/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.458 on clean data, 0.477 at noise 0.05, and 0.458 at 10% missingness. The plotted noise-0.50 score is 0.895 (seed range 0.876–0.913). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 27/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 28: trend/trend_dataset2—Avg Latency

Trend family; avg_latency. Original fault intervals: [546,596). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.361	0.111	48	4	1/1
VeraTS q95	0.361	0.111	48	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.639 on clean data, 0.639 at noise 0.05, and 0.637 at 10% missingness. The plotted noise-0.50 score is 0.144 (seed range 0.000–0.333). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 28/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 29: trend/trend_dataset2—Failure Count

Trend family; failure_count. Original fault intervals: [556,606). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.379	0.130	40	4	1/1
VeraTS q95	0.379	0.103	52	4	1/1
GPT-6 q95	0.333	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.676 on clean data, 0.676 at noise 0.05, and 0.676 at 10% missingness. The plotted noise-0.50 score is 0.242 (seed range 0.093–0.457). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 29/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 30: trend/trend_dataset2—Failure Rate

Trend family; failure_rate. Original fault intervals: [541,591). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.379	0.130	40	4	1/1
VeraTS q95	0.379	0.118	45	4	1/1
GPT-6 q95	0.500	0.107	50	0	1/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.676 on clean data, 0.678 at noise 0.05, and 0.680 at 10% missingness. The plotted noise-0.50 score is 0.173 (seed range 0.110–0.207). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 30/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 31: trend/trend_dataset2—Max Latency

Trend family; max_latency. Original fault intervals: [530,580). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.290	0.087	63	4	1/1
VeraTS q95	0.290	0.100	54	9	1/1
GPT-6 q95	1.000	1.000	0	0	1/1
step-5 q95	1.000	1.000	0	0	1/1

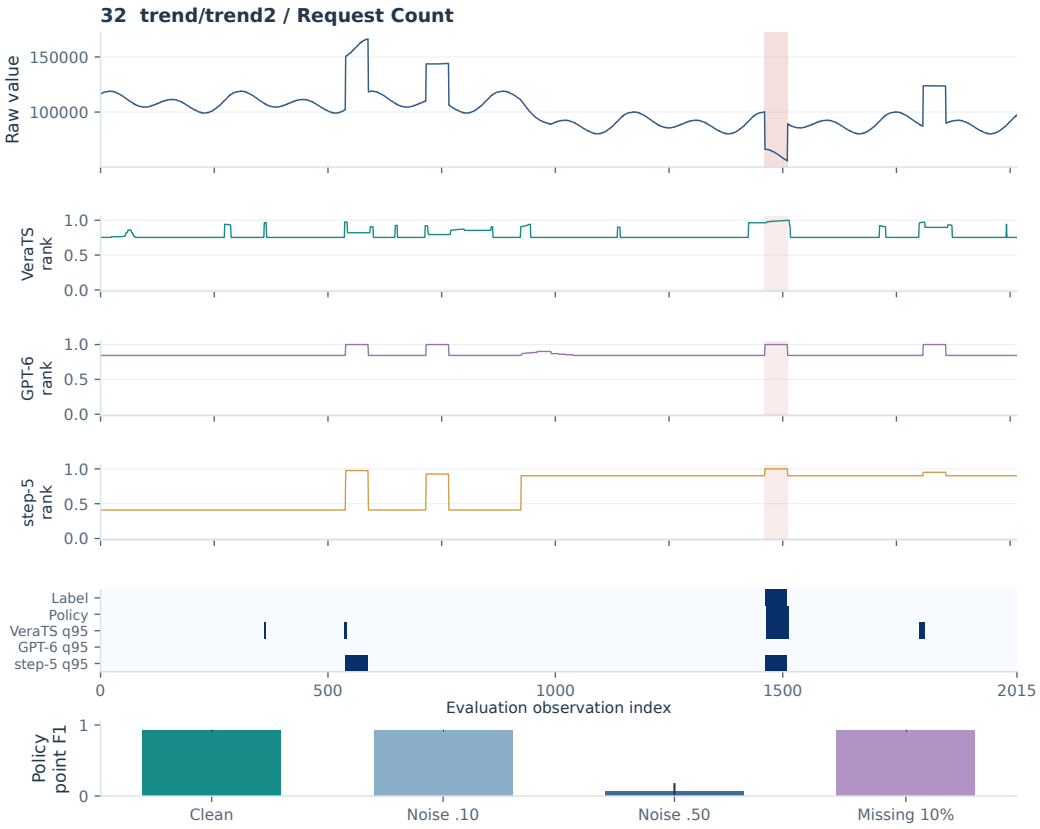
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.579 on clean data, 0.581 at noise 0.05, and 0.579 at 10% missingness. The plotted noise-0.50 score is 0.358 (seed range 0.260–0.512). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 31/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 32: trend/trend_dataset2—Request Count

Trend family; request_count. Original fault intervals: [1461,1511). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.771	0.600	4	4	1/1
VeraTS q95	0.771	0.182	27	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	0.107	50	0	1/1

Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.061 (seed range 0.000–0.182). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 32/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 33: trend/trend_dataset2—Retries

Trend family; retries. Original fault intervals: [533,583). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.324	0.097	56	4	1/1
VeraTS q95	0.324	0.102	53	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

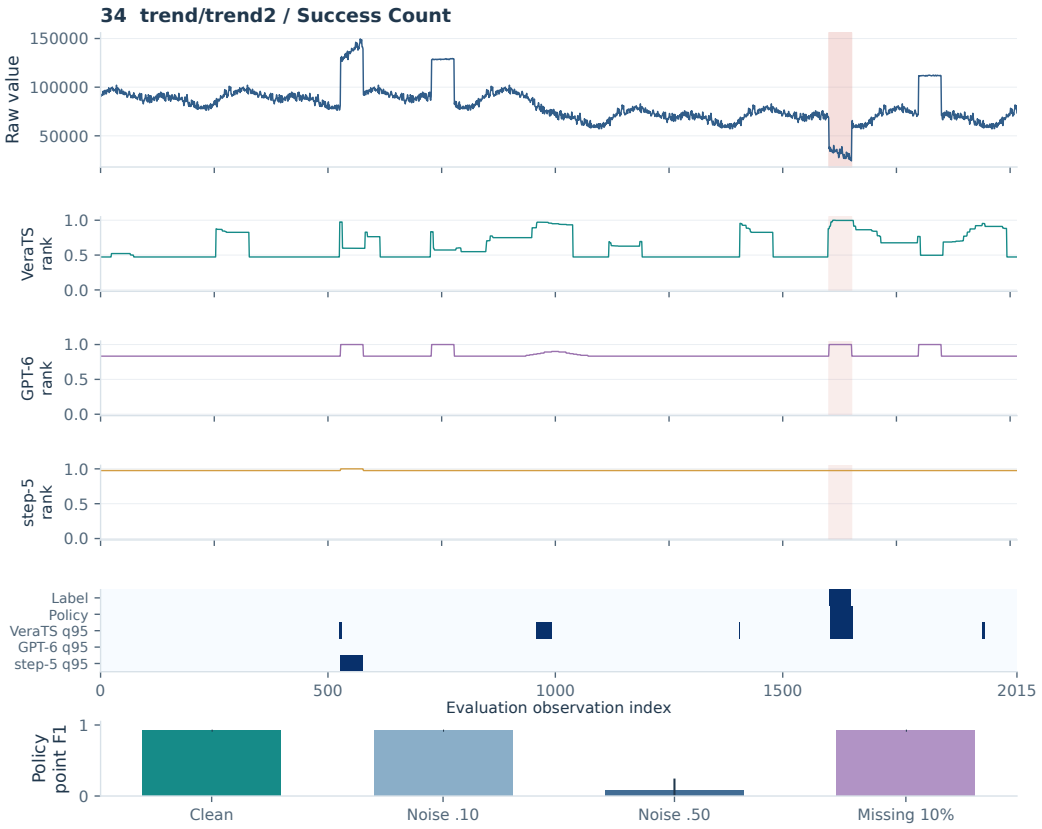
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.605 on clean data, 0.612 at noise 0.05, and 0.605 at 10% missingness. The plotted noise-0.50 score is 0.036 (seed range 0.000–0.107). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 33/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 34: trend/trend_dataset2—Success Count

Trend family; success_count. Original fault intervals: [1602,1652). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.872	0.600	4	4	1/1
VeraTS q95	0.872	0.103	52	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

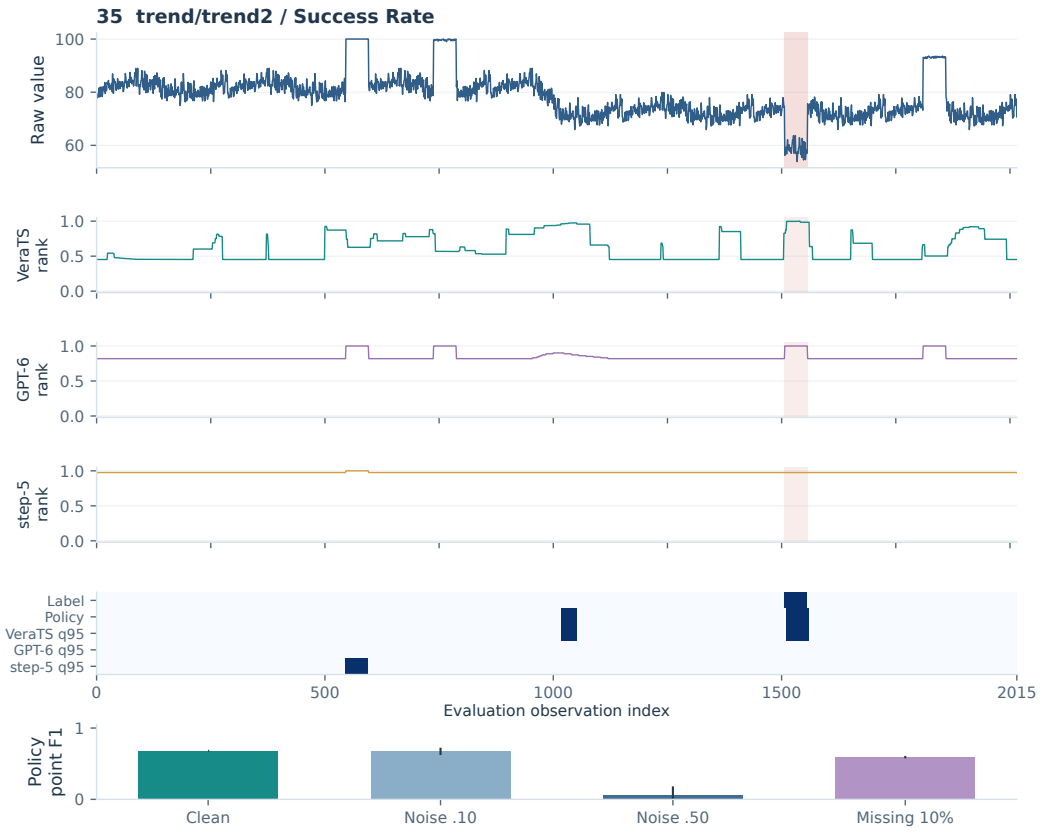
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.082 (seed range 0.000–0.246). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 34/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 35: trend/trend_dataset2—Success Rate

Trend family; success_rate. Original fault intervals: [1507,1557). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.909	0.133	39	4	1/1
VeraTS q95	0.909	0.133	39	4	1/1
GPT-6 q95	0.250	0.000	0	50	0/1
step-5 q95	0.025	0.000	50	50	0/1

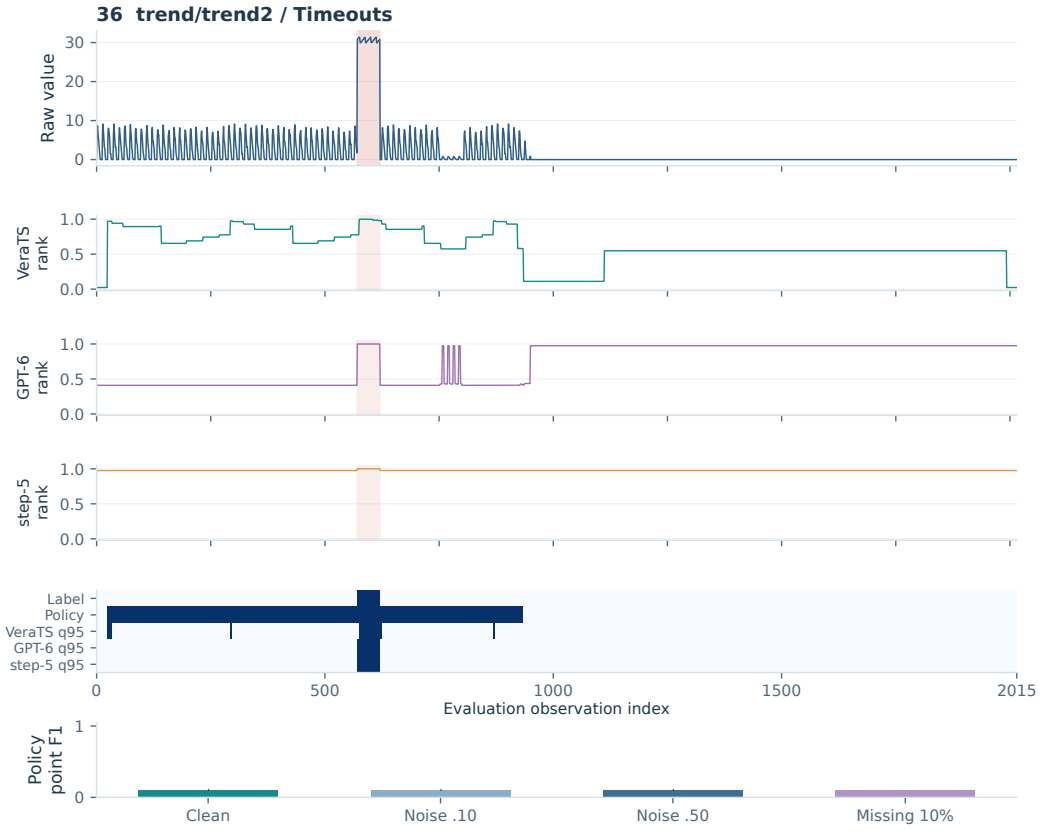
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.681 on clean data, 0.657 at noise 0.05, and 0.588 at 10% missingness. The plotted noise-0.50 score is 0.061 (seed range 0.000–0.182). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 35/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 36: trend/trend_dataset2—Timeouts

Trend family; timeouts. Original fault intervals: [571,621). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.921	0.007	861	0	1/1
VeraTS q95	0.921	0.200	24	4	1/1
GPT-6 q95	1.000	1.000	0	0	1/1
step-5 q95	1.000	1.000	0	0	1/1

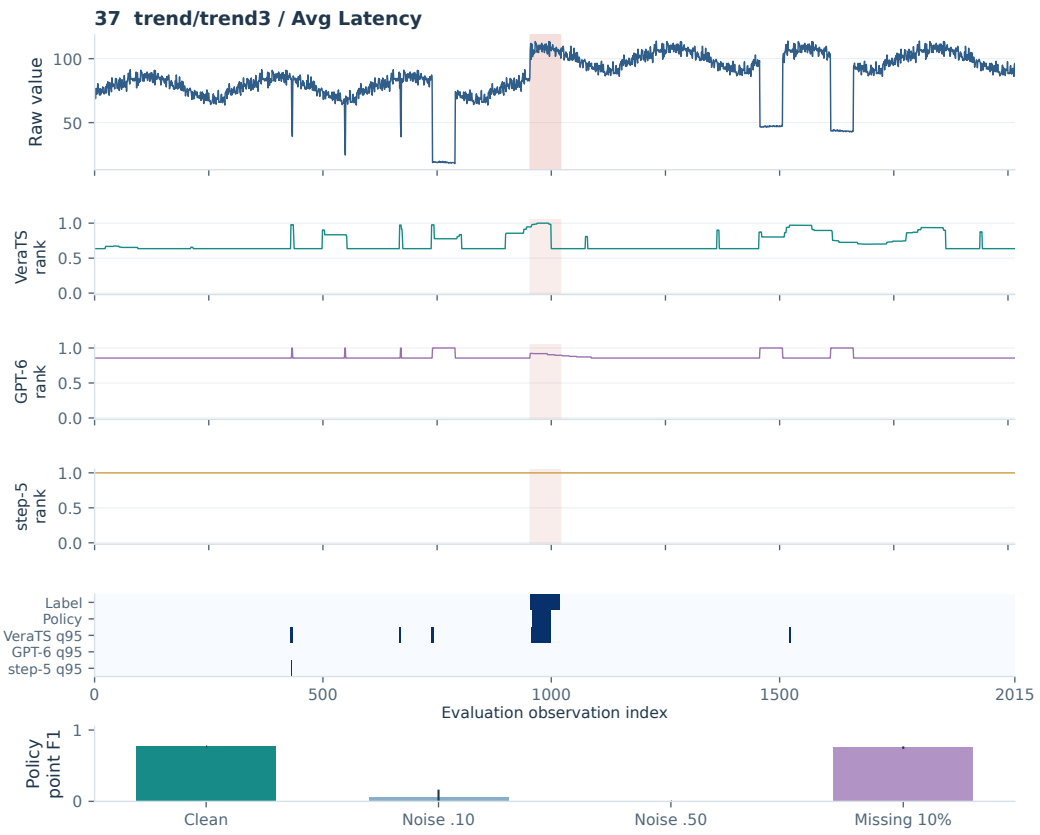
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.104 on clean data, 0.104 at noise 0.05, and 0.103 at 10% missingness. The plotted noise-0.50 score is 0.104 (seed range 0.104–0.104). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 36/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 37: trend/trend_dataset3—Avg Latency

Trend family; avg_latency. Original fault intervals: [954,1021). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.665	1.000	0	25	1/1
VeraTS q95	0.665	0.222	21	24	1/1
GPT-6 q95	0.216	0.000	0	67	0/1
step-5 q95	0.033	0.000	2	67	0/1

Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.771. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.771 on clean data, 0.249 at noise 0.05, and 0.759 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 37/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 38: trend/trend_dataset3—Failure Count

Trend family; failure_count. Original fault intervals: [927,994). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.893	1.000	0	62	1/1
VeraTS q95	0.893	0.500	6	13	1/1
GPT-6 q95	0.181	0.000	0	67	0/1
step-5 q95	0.033	0.000	2	67	0/1

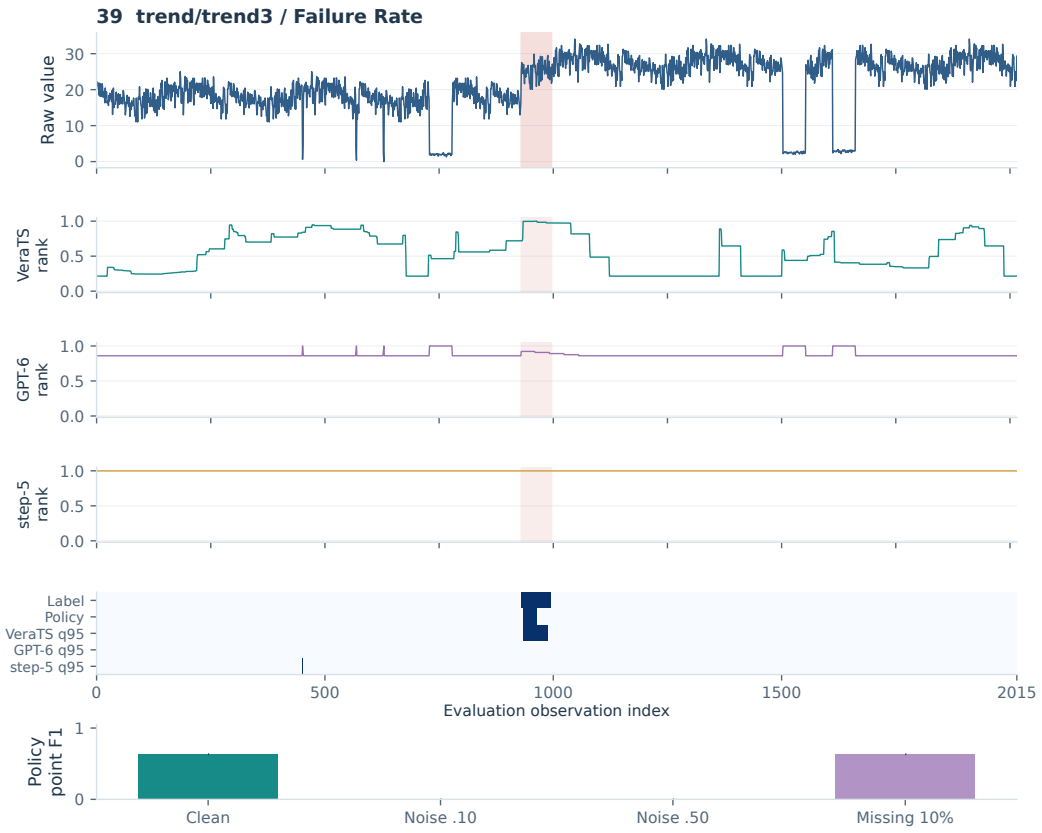
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.139 on clean data, 0.000 at noise 0.05, and 0.308 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 38/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 39: trend/trend_dataset3—Failure Rate

Trend family; failure_rate. Original fault intervals: [930,997). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.905	1.000	0	36	1/1
VeraTS q95	0.905	1.000	0	11	1/1
GPT-6 q95	0.228	0.000	0	67	0/1
step-5 q95	0.033	0.000	2	67	0/1

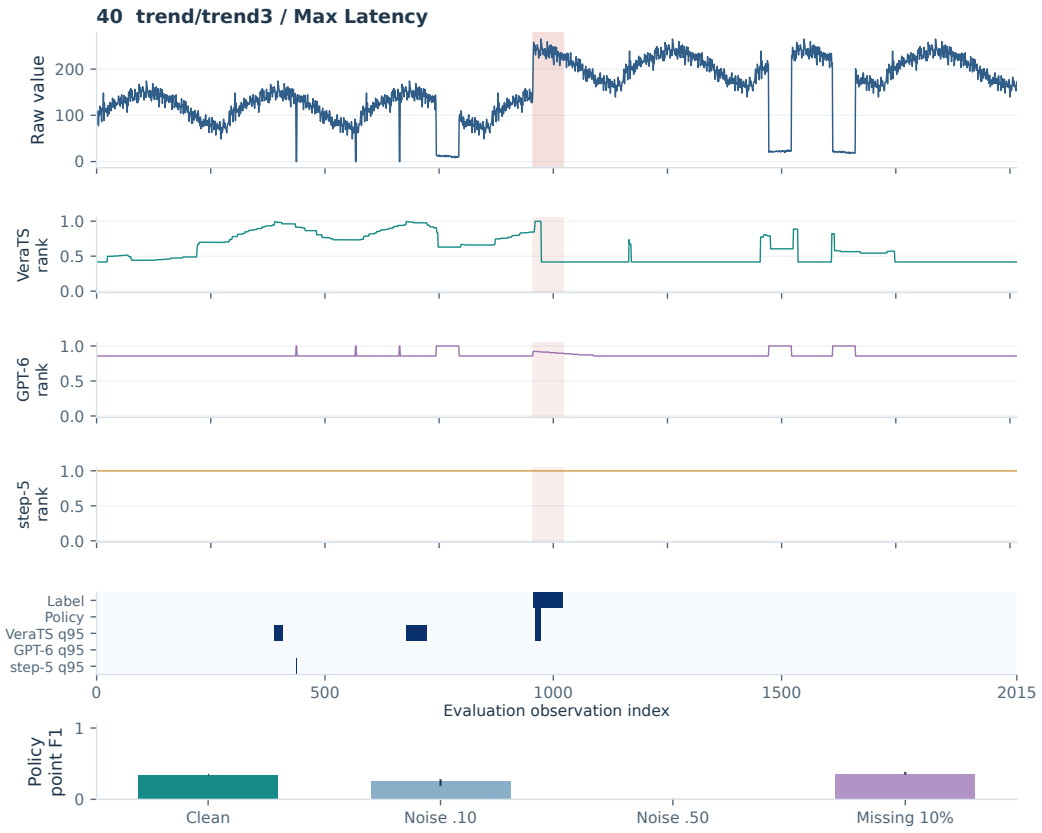
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.633 on clean data, 0.055 at noise 0.05, and 0.633 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 39/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 40: trend/trend_dataset3—Max Latency

Trend family; max_latency. Original fault intervals: [956,1023). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.237	1.000	0	53	1/1
VeraTS q95	0.237	0.085	65	53	1/1
GPT-6 q95	0.198	0.000	0	67	0/1
step-5 q95	0.033	0.000	2	67	0/1

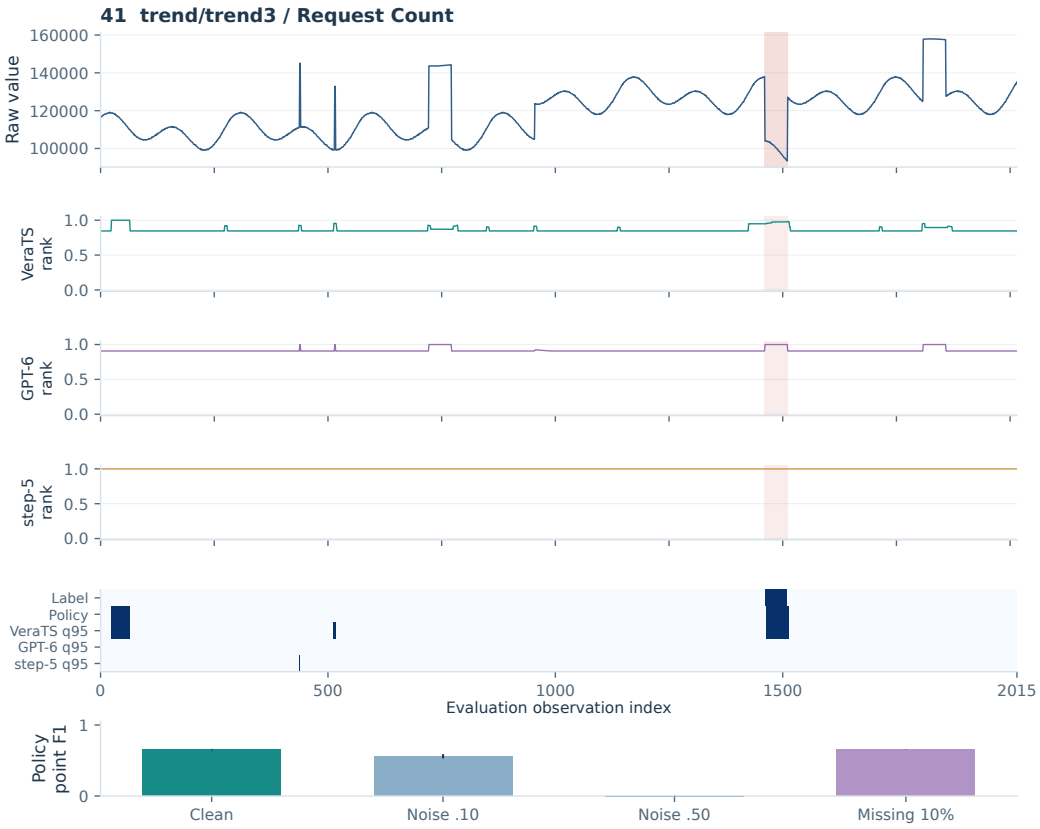
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.346 on clean data, 0.508 at noise 0.05, and 0.359 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 40/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 41: trend/trend_dataset3—Request Count

Trend family; request_count. Original fault intervals: [1461,1511). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.392	0.118	45	4	1/1
VeraTS q95	0.392	0.105	51	4	1/1
GPT-6 q95	0.325	0.000	0	50	0/1
step-5 q95	0.025	0.000	2	50	0/1

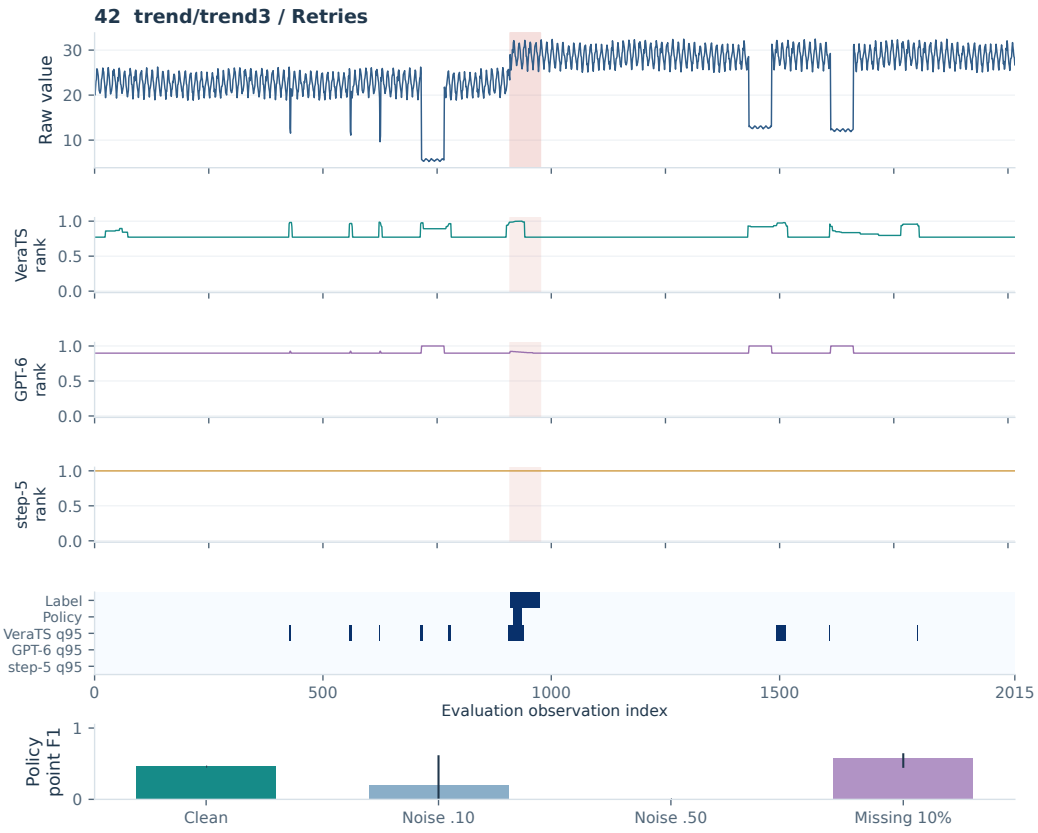
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.652 on clean data, 0.599 at noise 0.05, and 0.651 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 41/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 42: trend/trend_dataset3—Retries

Trend family; retries. Original fault intervals: [910,977). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.488	1.000	0	47	1/1
VeraTS q95	0.488	0.098	55	35	1/1
GPT-6 q95	0.121	0.000	0	67	0/1
step-5 q95	0.033	0.000	2	67	0/1

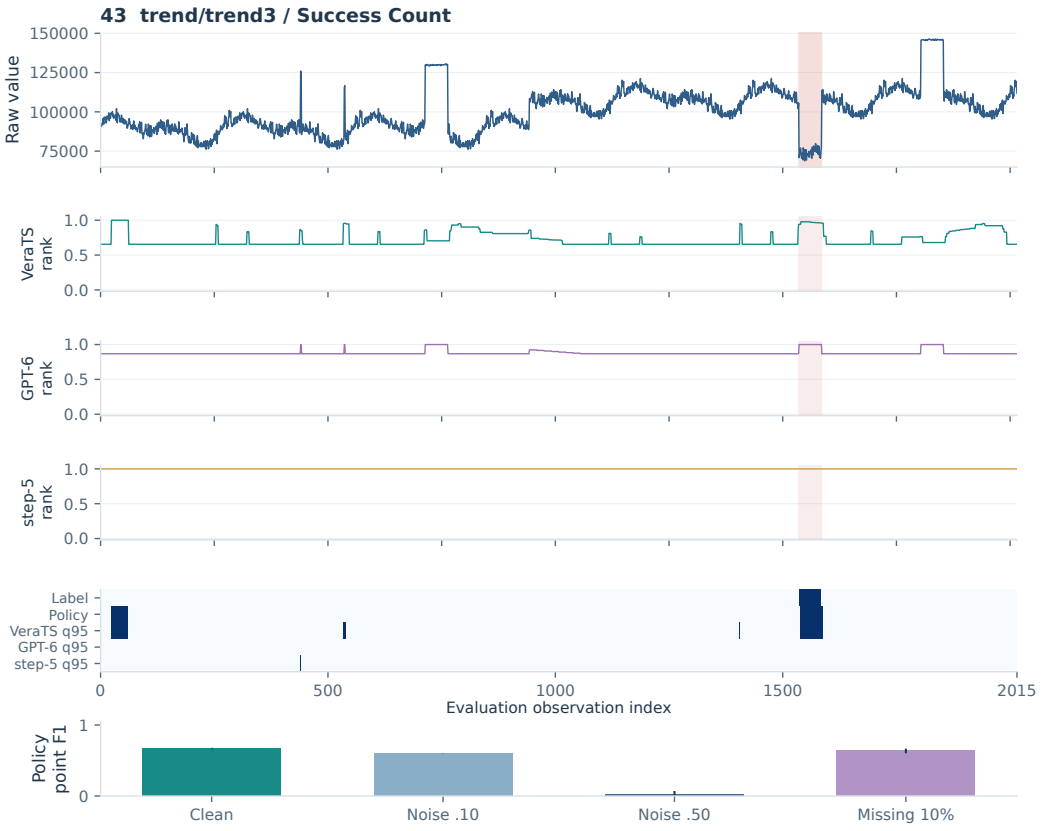
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.460 on clean data, 0.206 at noise 0.05, and 0.574 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 42/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 43: trend/trend_dataset3—Success Count

Trend family; success_count. Original fault intervals: [1536,1586). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.384	0.125	42	4	1/1
VeraTS q95	0.384	0.107	50	4	1/1
GPT-6 q95	0.325	0.000	0	50	0/1
step-5 q95	0.025	0.000	2	50	0/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.667 on clean data, 0.637 at noise 0.05, and 0.642 at 10% missingness. The plotted noise-0.50 score is 0.023 (seed range 0.000–0.070). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 43/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 44: trend/trend_dataset3—Success Rate

Trend family; success_rate. Original fault intervals: [1459,1509). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.935	0.600	4	4	1/1
VeraTS q95	0.935	0.194	25	4	1/1
GPT-6 q95	0.500	0.107	50	0	1/1
step-5 q95	0.025	0.000	2	50	0/1

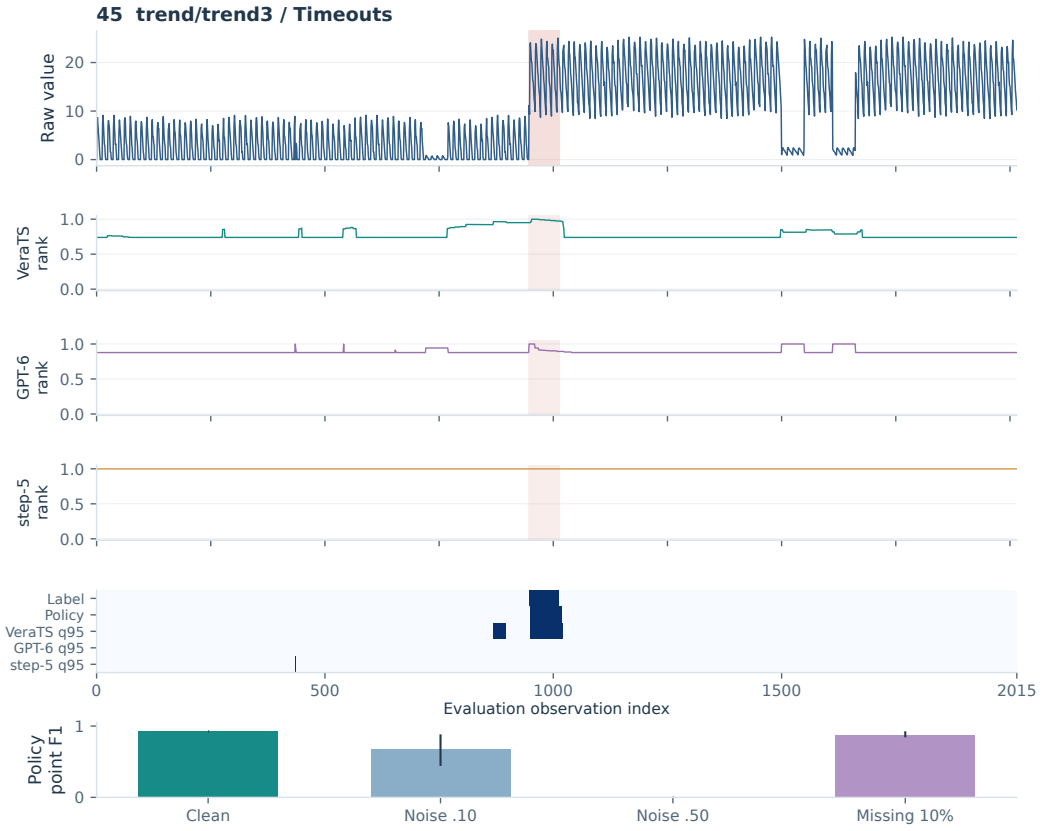
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.920. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.920 on clean data, 0.920 at noise 0.05, and 0.920 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 44/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 45: trend/trend_dataset3—Timeouts

Trend family; timeouts. Original fault intervals: [947,1014). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.958	0.500	6	4	1/1
VeraTS q95	0.958	0.140	37	4	1/1
GPT-6 q95	0.200	0.000	0	67	0/1
step-5 q95	0.033	0.000	2	67	0/1

Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.926. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.926 on clean data, 0.678 at noise 0.05, and 0.875 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 45/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 46: trend/trend_dataset4—Avg Latency

Trend family; avg_latency. Original fault intervals: [571,621). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.359	0.111	48	4	1/1
VeraTS q95	0.359	0.102	53	4	1/1
GPT-6 q95	0.300	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

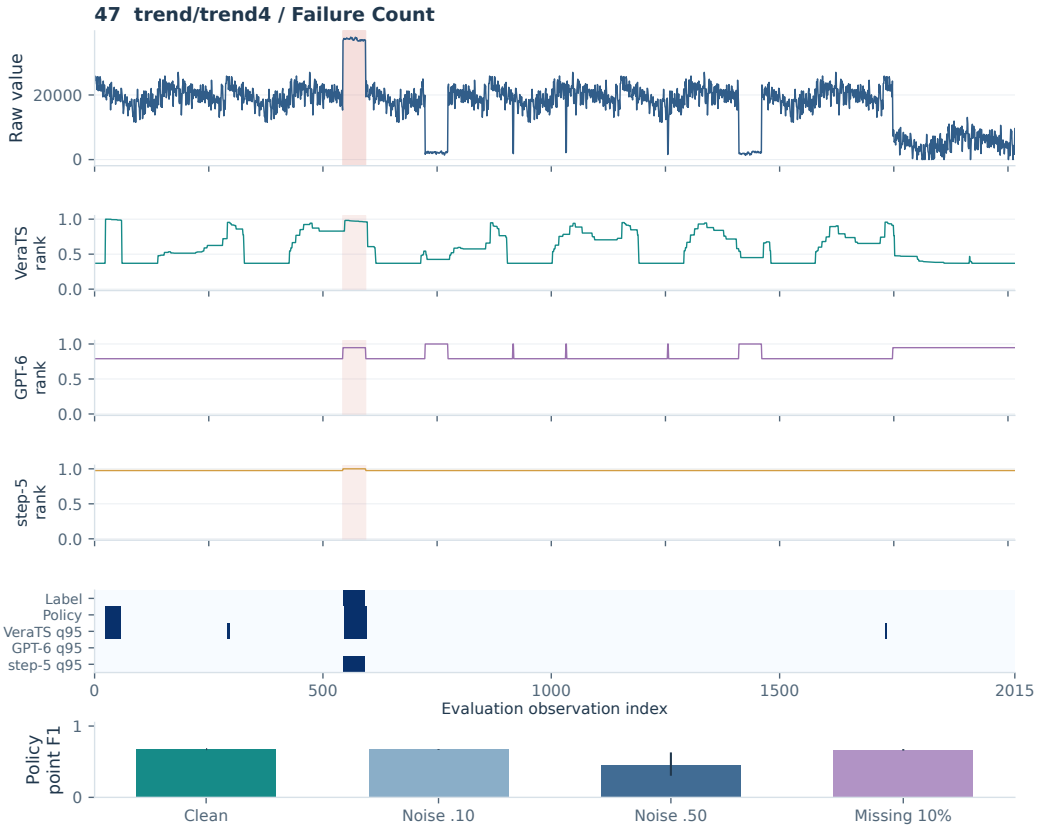
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.639 on clean data, 0.639 at noise 0.05, and 0.636 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 46/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 47: trend/trend_dataset4—Failure Count

Trend family; failure_count. Original fault intervals: [544,594). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.368	0.130	40	4	1/1
VeraTS q95	0.368	0.103	52	4	1/1
GPT-6 q95	0.118	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.676 on clean data, 0.670 at noise 0.05, and 0.668 at 10% missingness. The plotted noise-0.50 score is 0.455 (seed range 0.303–0.630). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 47/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 48: trend/trend_dataset4—Failure Rate

Trend family; failure_rate. Original fault intervals: [548,598). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.373	0.130	40	4	1/1
VeraTS q95	0.373	0.107	50	4	1/1
GPT-6 q95	0.321	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

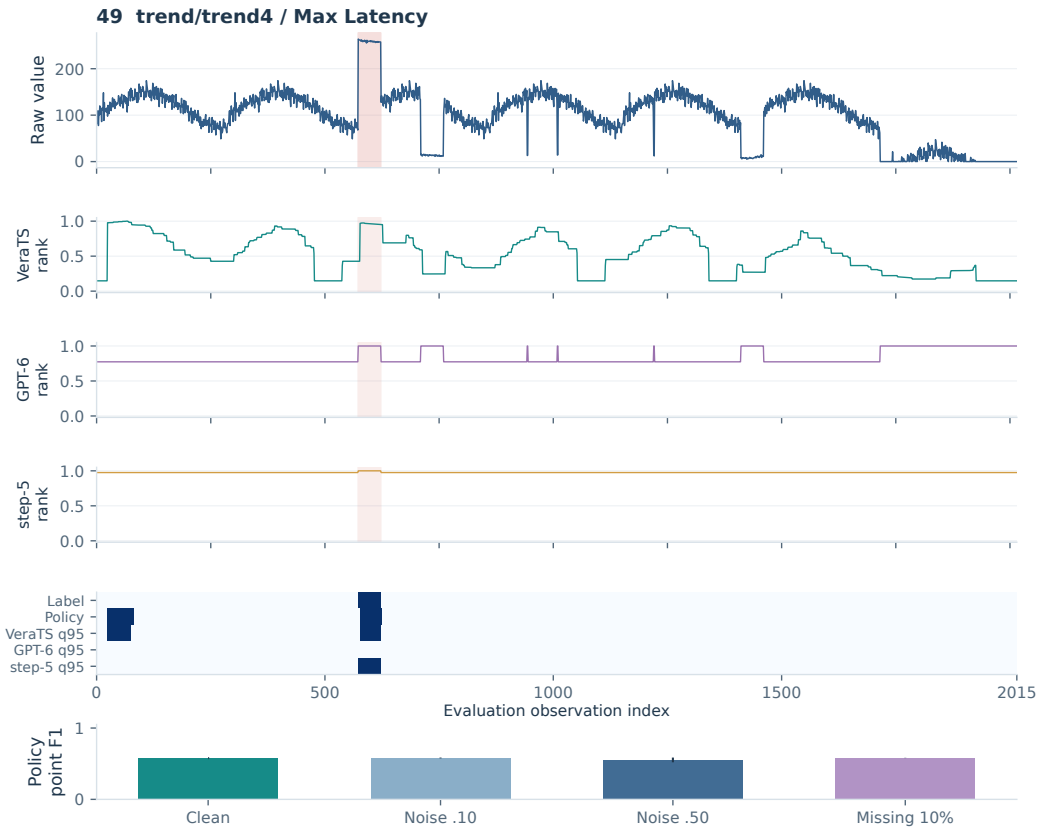
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.676 on clean data, 0.678 at noise 0.05, and 0.668 at 10% missingness. The plotted noise-0.50 score is 0.315 (seed range 0.000–0.493). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 48/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 49: trend/trend_dataset4—Max Latency

Trend family; max_latency. Original fault intervals: [573,623). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.277	0.087	63	4	1/1
VeraTS q95	0.277	0.100	54	4	1/1
GPT-6 q95	0.110	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

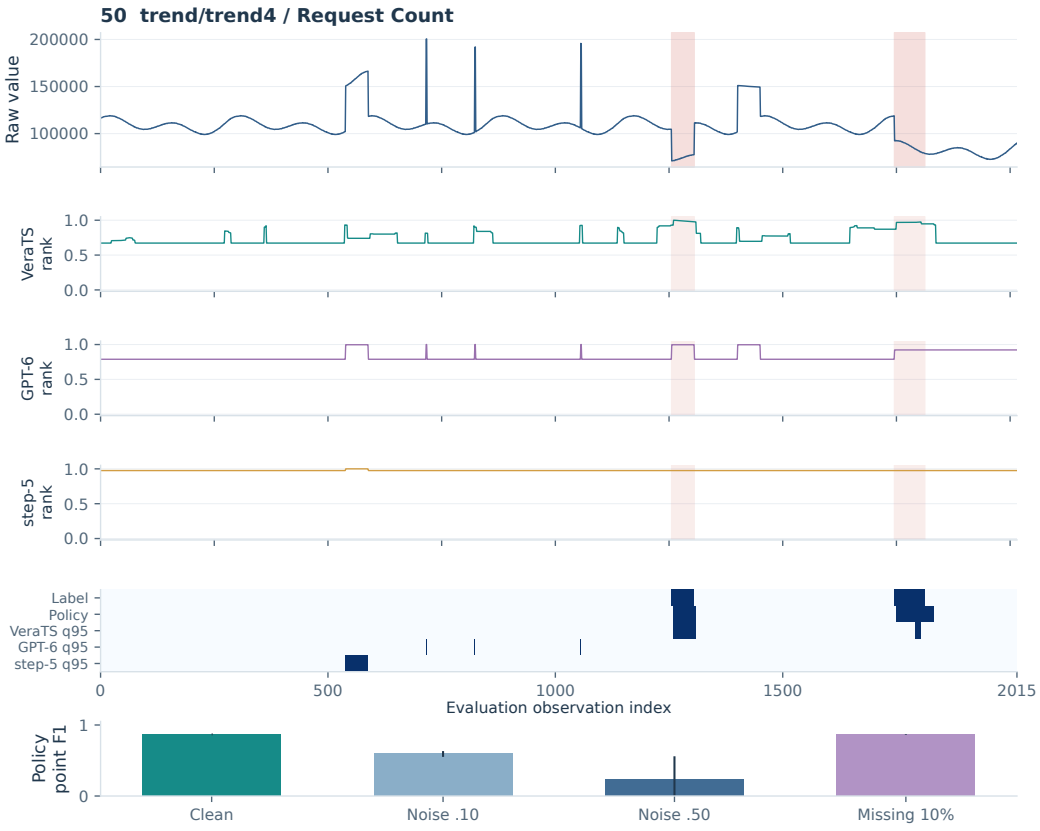
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.579 on clean data, 0.581 at noise 0.05, and 0.580 at 10% missingness. The plotted noise-0.50 score is 0.546 (seed range 0.520–0.586). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 49/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 50: trend/trend_dataset4—Request Count

Trend family; request_count. Original fault intervals: [1256,1306), [1746,1813). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.937	0.324	25	8	2/2
VeraTS q95	0.937	0.750	4	58	2/2
GPT-6 q95	0.294	0.000	6	117	0/2
step-5 q95	0.058	0.000	50	117	0/2

Reading this trace. VeraTS converts contextual evidence into 2 matched alert intervals, reaching strict event F1 1.000 and point F1 0.869. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.869 on clean data, 0.902 at noise 0.05, and 0.865 at 10% missingness. The plotted noise-0.50 score is 0.234 (seed range 0.000–0.560). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 50/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 51: trend/trend_dataset4—Retries

Trend family; retries. Original fault intervals: [546,596). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.327	0.097	56	4	1/1
VeraTS q95	0.327	0.100	54	4	1/1
GPT-6 q95	0.321	0.000	0	50	0/1
step-5 q95	1.000	1.000	0	0	1/1

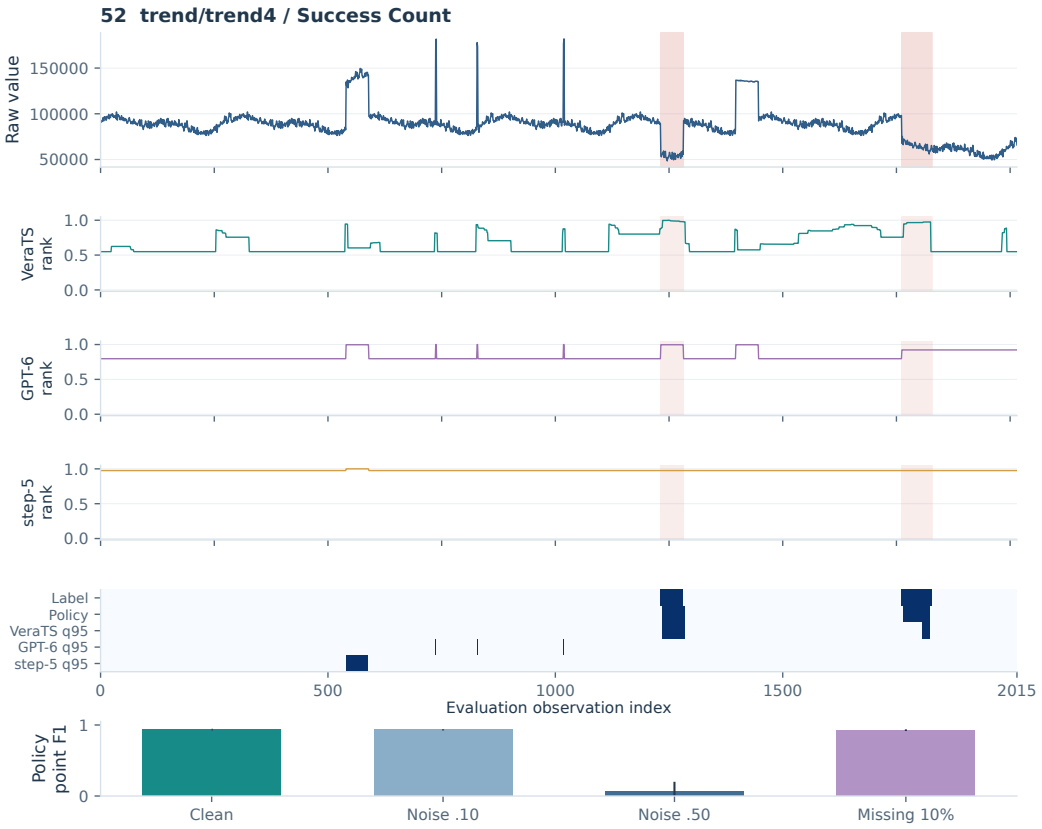
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.605 on clean data, 0.614 at noise 0.05, and 0.604 at 10% missingness. The plotted noise-0.50 score is 0.179 (seed range 0.000–0.455). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 51/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 52: trend/trend_dataset4—Success Count

Trend family; success_count. Original fault intervals: [1232,1282), [1762,1829). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.907	0.750	4	11	2/2
VeraTS q95	0.907	0.750	4	52	2/2
GPT-6 q95	0.300	0.000	6	117	0/2
step-5 q95	0.058	0.000	50	117	0/2

Reading this trace. VeraTS converts contextual evidence into 2 matched alert intervals, reaching strict event F1 1.000 and point F1 0.934. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.934 on clean data, 0.917 at noise 0.05, and 0.923 at 10% missingness. The plotted noise-0.50 score is 0.067 (seed range 0.000–0.200). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 52/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 53: trend/trend_dataset4—Success Rate

Trend family; success_rate. Original fault intervals: [1237,1287), [1717,1784). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.903	0.750	4	13	2/2
VeraTS q95	0.903	0.750	4	25	2/2
GPT-6 q95	0.284	0.000	0	117	0/2
step-5 q95	0.058	0.000	50	117	0/2

Reading this trace. VeraTS converts contextual evidence into 2 matched alert intervals, reaching strict event F1 1.000 and point F1 0.924. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.924 on clean data, 0.921 at noise 0.05, and 0.924 at 10% missingness. The plotted noise-0.50 score is 0.000 (seed range 0.000–0.000). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 53/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 54: trend/trend_dataset4—Timeouts

Trend family; timeouts. Original fault intervals: [549,599). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.899	0.005	1107	0	1/1
VeraTS q95	0.899	0.200	24	4	1/1
GPT-6 q95	1.000	1.000	0	0	1/1
step-5 q95	1.000	1.000	0	0	1/1

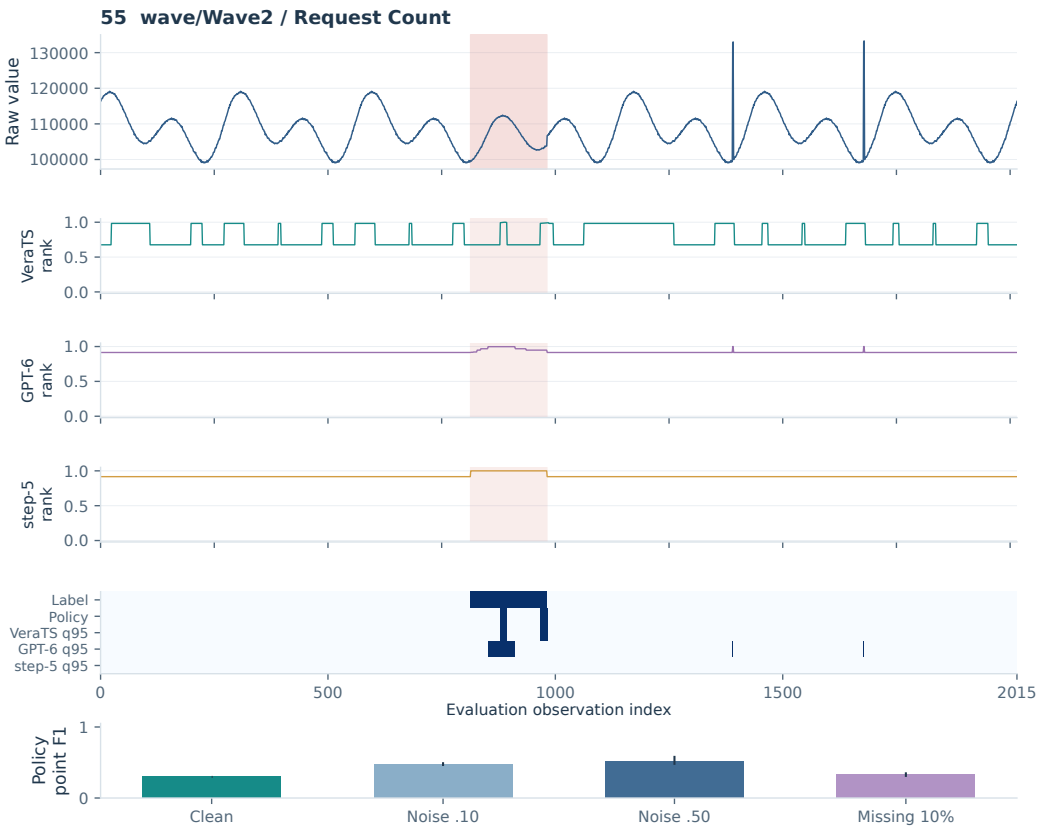
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.083 on clean data, 0.083 at noise 0.05, and 0.083 at 10% missingness. The plotted noise-0.50 score is 0.083 (seed range 0.083–0.083). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 54/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 55: wave/Wave_dataset2—Request Count

Wave family; request_count. Original fault intervals: [814,982). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.234	0.667	4	138	1/1
VeraTS q95	0.234	0.667	4	138	1/1
GPT-6 q95	0.958	0.667	4	108	1/1
step-5 q95	1.000	0.000	0	168	0/1

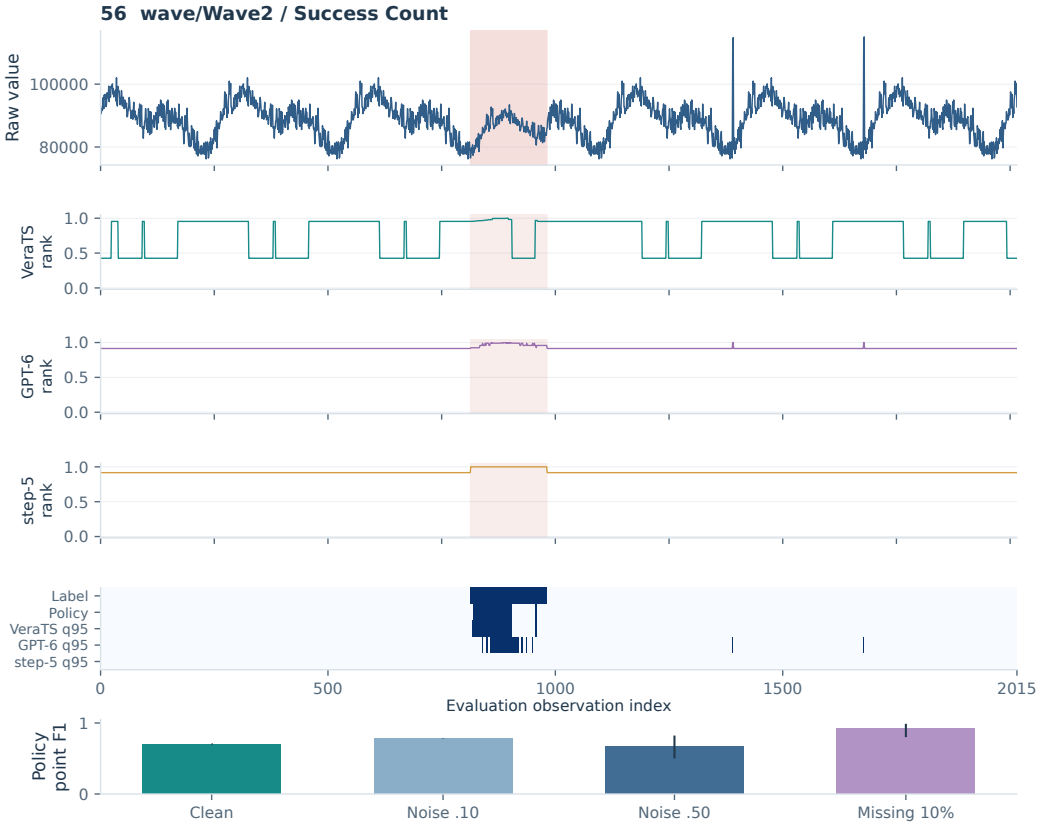
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.297 on clean data, 0.525 at noise 0.05, and 0.324 at 10% missingness. The plotted noise-0.50 score is 0.517 (seed range 0.467–0.594). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 55/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 56: wave/Wave_dataset2—Success Count

Wave family; success_count. Original fault intervals: [814,982). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.588	1.000	0	77	1/1
VeraTS q95	0.588	1.000	0	76	1/1
GPT-6 q95	0.949	0.667	4	88	1/1
step-5 q95	1.000	0.000	0	168	0/1

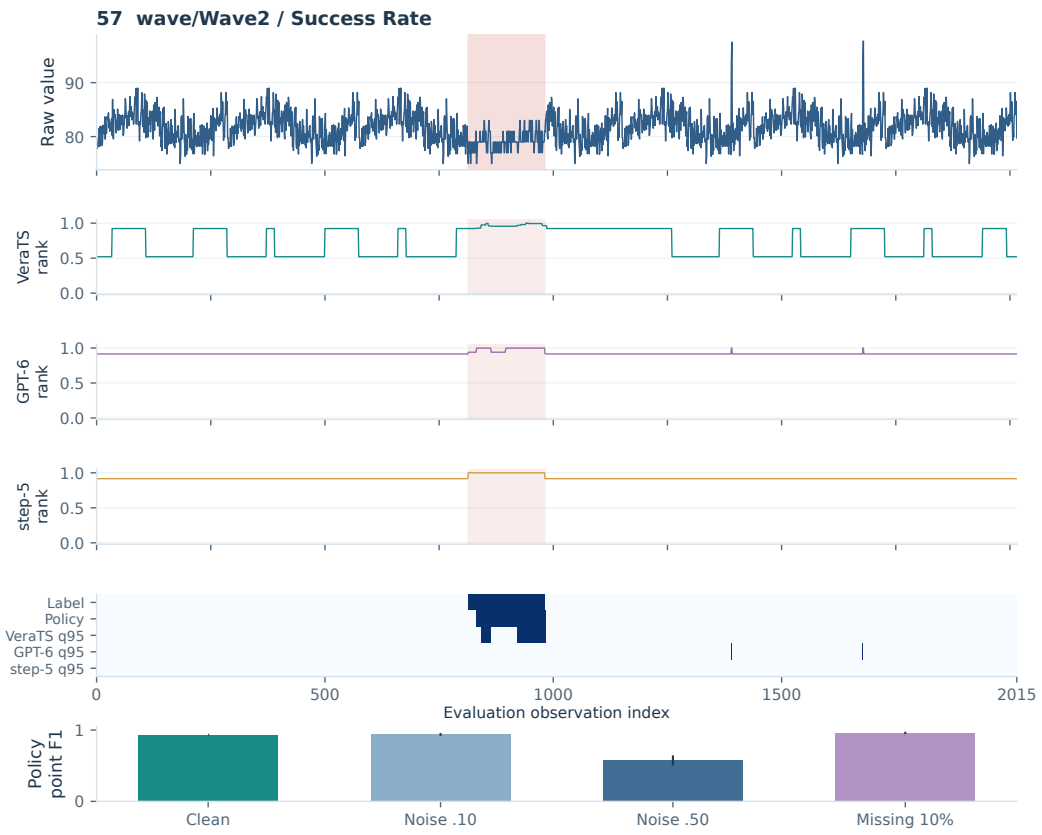
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.703 on clean data, 0.941 at noise 0.05, and 0.921 at 10% missingness. The plotted noise-0.50 score is 0.666 (seed range 0.502–0.823). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 56/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 57: wave/Wave_dataset2—Success Rate

Wave family; success_rate. Original fault intervals: [814,982). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.902	0.667	4	17	1/1
VeraTS q95	0.902	0.667	4	84	1/1
GPT-6 q95	0.970	0.000	4	168	0/1
step-5 q95	1.000	0.000	0	168	0/1

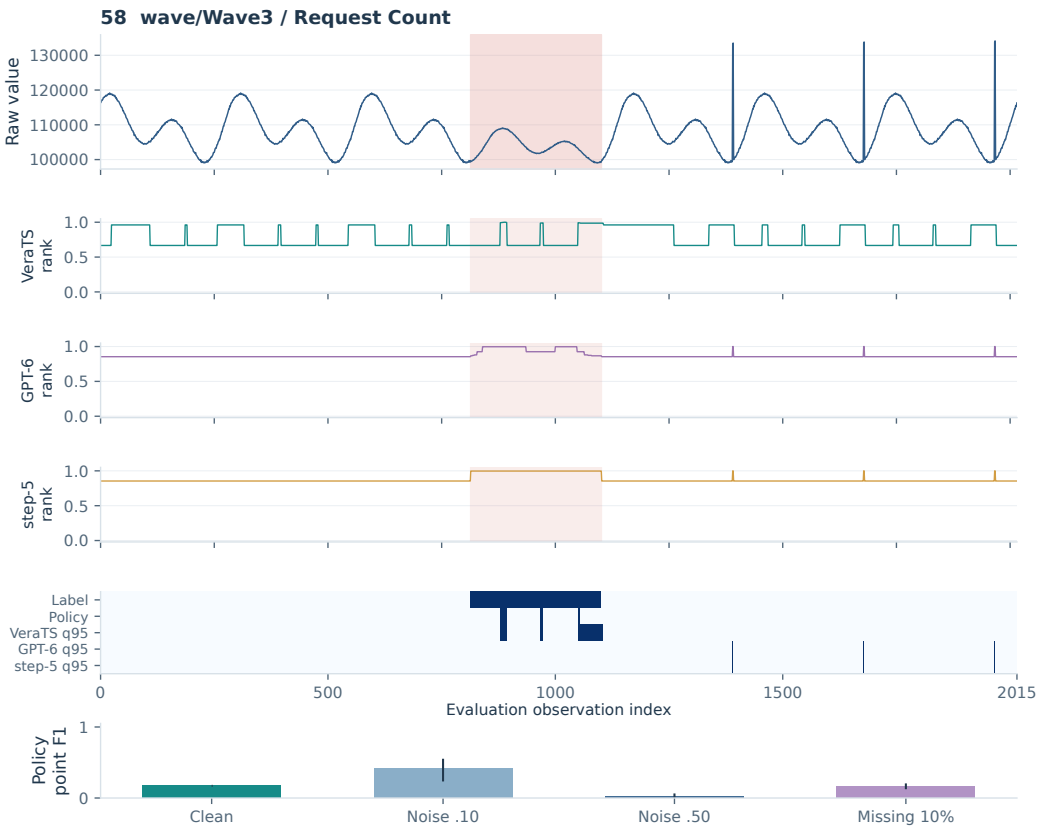
Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.935. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.935 on clean data, 0.936 at noise 0.05, and 0.961 at 10% missingness. The plotted noise-0.50 score is 0.581 (seed range 0.503–0.645). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 57/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 58: wave/Wave_dataset3—Request Count

Wave family; request_count. Original fault intervals: [814,1102). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.355	1.000	0	261	1/1
VeraTS q95	0.355	0.714	4	214	1/1
GPT-6 q95	0.968	0.000	6	288	0/1
step-5 q95	0.980	0.000	6	288	0/1

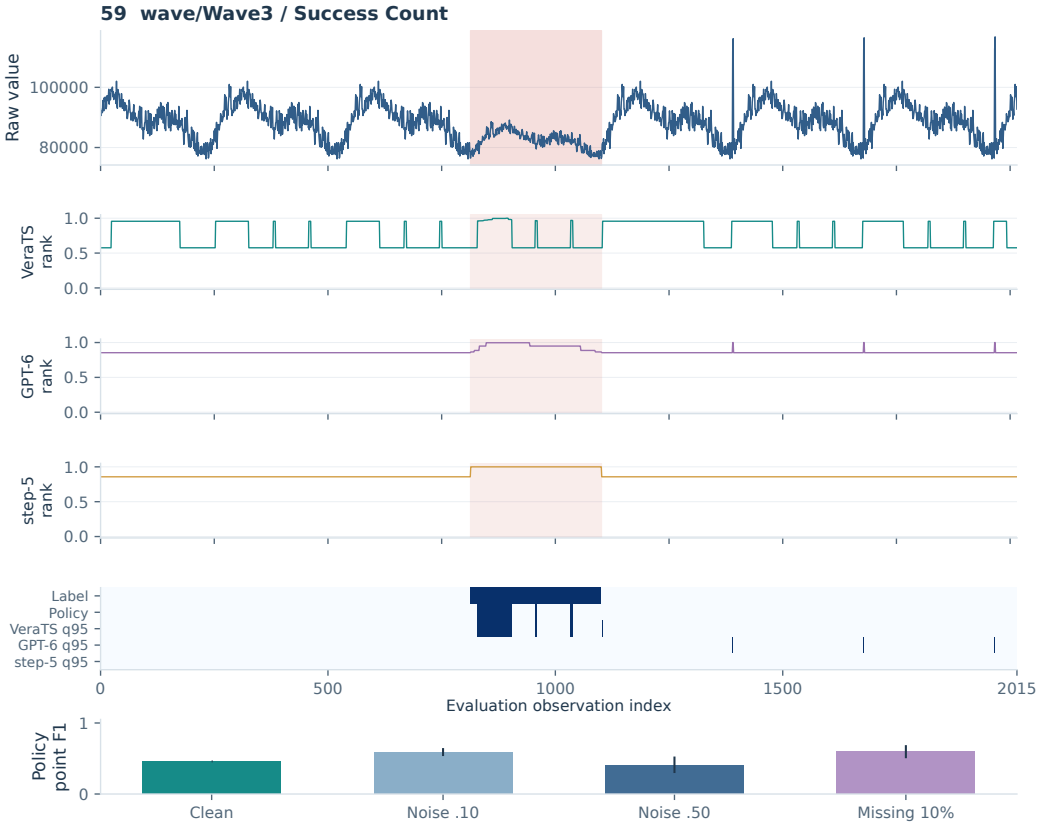
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.171 on clean data, 0.598 at noise 0.05, and 0.159 at 10% missingness. The plotted noise-0.50 score is 0.026 (seed range 0.000–0.064). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 58/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 59: wave/Wave_dataset3—Success Count

Wave family; success_count. Original fault intervals: [814,1102). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.399	1.000	0	202	1/1
VeraTS q95	0.399	0.833	2	202	1/1
GPT-6 q95	0.964	0.000	6	288	0/1
step-5 q95	1.000	0.000	0	288	0/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.460 on clean data, 0.619 at noise 0.05, and 0.599 at 10% missingness. The plotted noise-0.50 score is 0.395 (seed range 0.296–0.528). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 59/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 60: wave/Wave_dataset3—Success Rate

Wave family; success_rate. Original fault intervals: [814,1102). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.750	0.833	2	84	1/1
VeraTS q95	0.750	1.000	0	218	1/1
GPT-6 q95	0.980	0.000	6	288	0/1
step-5 q95	1.000	0.000	0	288	0/1

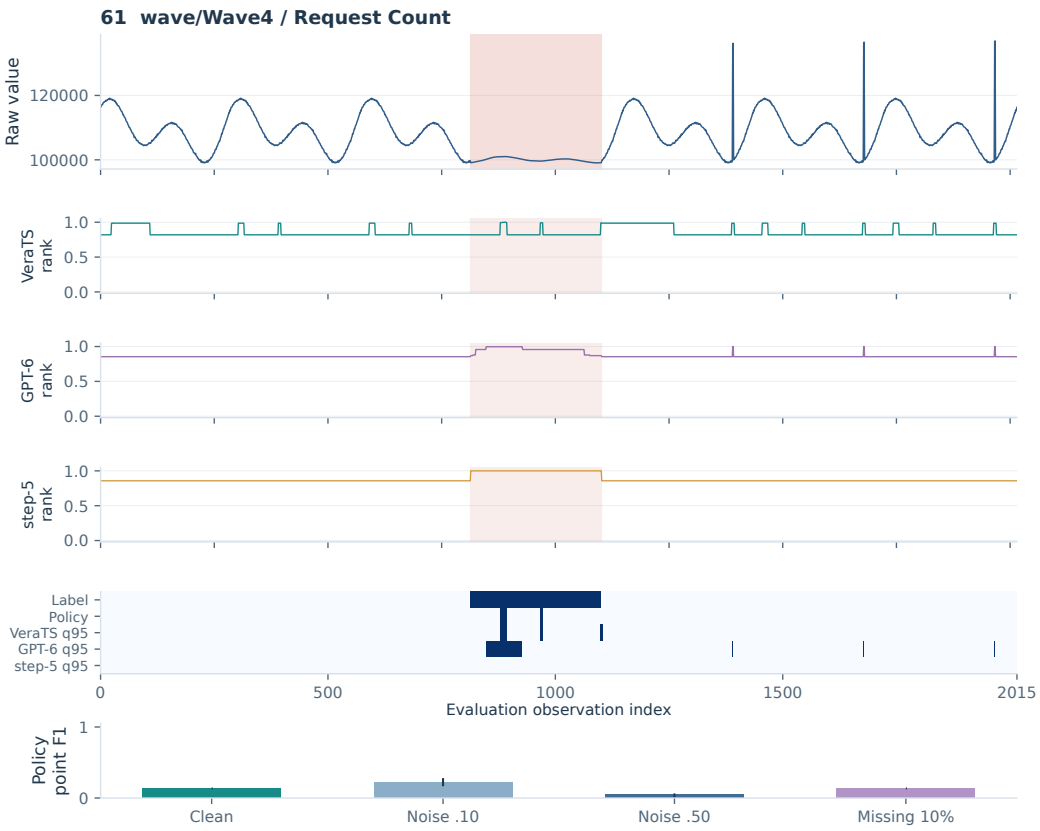
Reading this trace. The frozen policy touches the labeled event and reaches point F1 0.826. Its concentrated temporal response illustrates how contextual qualification preserves useful event support within an explicit numerical rule.

Stress response. With the same parameters, point F1 is 0.826 on clean data, 0.747 at noise 0.05, and 0.897 at 10% missingness. The plotted noise-0.50 score is 0.170 (seed range 0.034–0.282). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 60/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 61: wave/Wave_dataset4—Request Count

Wave family; request_count. Original fault intervals: [814,1102). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.211	1.000	0	267	1/1
VeraTS q95	0.211	0.714	4	265	1/1
GPT-6 q95	0.963	0.625	6	208	1/1
step-5 q95	1.000	0.000	0	288	0/1

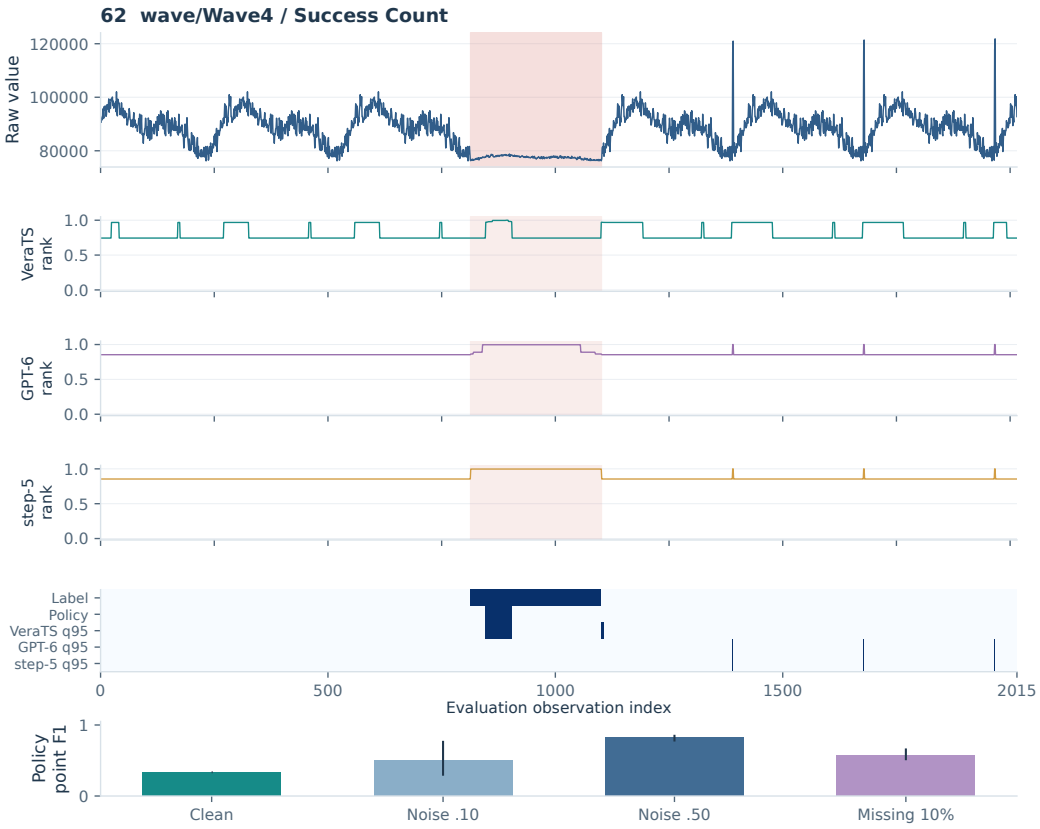
Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.136 on clean data, 0.152 at noise 0.05, and 0.136 at 10% missingness. The plotted noise-0.50 score is 0.043 (seed range 0.000–0.065). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 61/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 62: wave/Wave_dataset4—Success Count

Wave family; success_count. Original fault intervals: [814,1102). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.318	1.000	0	230	1/1
VeraTS q95	0.318	0.667	5	229	1/1
GPT-6 q95	0.974	0.000	6	288	0/1
step-5 q95	0.980	0.000	6	288	0/1

Reading this trace. The frozen policy preserves access to the labeled event. This directly illustrates the hierarchy's event-discovery objective; aligned score and mask rows make the resulting temporal decisions inspectable under the same rule.

Stress response. With the same parameters, point F1 is 0.335 on clean data, 0.459 at noise 0.05, and 0.572 at 10% missingness. The plotted noise-0.50 score is 0.825 (seed range 0.767–0.862). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 62/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.

Series 63: wave/Wave_dataset4—Success Rate

Wave family; success_rate. Original fault intervals: [814,1102). The detector target is the original binary fault annotation; the folder name describes morphology separately.



Method / rule	Point AP	Event F1	FP points	FN points	Hits
VeraTS policy	0.954	0.714	4	15	1/1
VeraTS q95	0.954	1.000	0	202	1/1
GPT-6 q95	0.980	0.000	6	288	0/1
step-5 q95	1.000	0.000	0	288	0/1

Reading this trace. VeraTS converts contextual evidence into 1 matched alert interval, reaching strict event F1 1.000 and point F1 0.966. This illustrates coherent event localization under the fixed direction, persistence and closing rule.

Stress response. With the same parameters, point F1 is 0.966 on clean data, 0.356 at noise 0.05, and 0.966 at 10% missingness. The plotted noise-0.50 score is 0.823 (seed range 0.810–0.837). The saved rule makes each measurement condition reproducible; all raw observations and original-score decisions remain available.

Canonical series 63/63. Dense scores, masks, and all perturbation conditions are retained in the companion artifacts. Event F1 here is log-weighted EventF1PA, without label margins.