WHITE PAPER · ANOMALY DETECTION & ROOT CAUSE ANALYSIS

The Recurring Anomaly Detection Pattern in All Industries

How a 1981 clinical laboratory framework, biological cybernetics, and a neuro-symbolic AI platform converge on the same answer to anomaly detection and root cause analysis — and why enterprises can eliminate LLM token costs from this workload entirely.

6	3–10×	~$0
industries, one detection-and-RCA pattern	faster time-to-resolution with AI-driven RCA	LLM token cost on the streaming hot path

A deep dive into anomaly detection and root cause analysis

Contents

Executive Summary

Every industry that runs on data has, independently, rediscovered the same problem and arrived at the same shape of solution.

A hospital laboratory watching glucose readings, a bank watching card transactions, a factory watching vibration sensors on a turbine, a security operations centre watching login events, a government agency screening benefit claims, and a military logistics command watching fleet telemetry are — at the level of mathematics — doing the same thing. Each is observing a continuous stream of measurements from a system in operation, comparing those measurements against an expected pattern of behaviour, and trying to distinguish a meaningful deviation from ordinary noise. Each must do this with two competing goals: catch real problems quickly (high sensitivity) and avoid drowning operators in false alarms (high specificity).

And in every one of these industries, detection alone is not enough. An alert that says *"something is wrong"* without saying *"and here is why"* leaves operators hunting through logs, charts, and dashboards while damage compounds. The full pattern is therefore not just anomaly **detection** but **detection followed immediately by root cause analysis** — the diagnostic step that turns a signal into an action.

> **THE RECURRING PATTERN**
> This is the recurring anomaly detection and root cause analysis pattern. It is not industry-specific — it recurs every time a system must be monitored in real time. The clinical chemistry community formalized one version in 1981 with the **Westgard Rules**; control engineering formalized another through **cybernetics**; and the biological sciences observe yet another every time a homeostatic system corrects deviation from a set point. These are three articulations of one pattern.

Today, most enterprises attempt to solve this problem with one of two unsuitable tools:

- **Static machine-learning models** that are trained once on historical data and then deployed, leaving them unable to adapt to a changing reality and unable to explain *why* an anomaly occurred.
- **Large language models (LLMs)** that are expensive per call, latency-bound, opaque in their reasoning, and structurally unsuited to high-volume streaming data.

This white paper argues that the right tool is a **neuro-symbolic AI platform built on biological and cybernetic principles**. The **AsterMind AI EVO Platform** is the concrete realization of this approach. By recognizing that the same multirule-style pattern can be implemented once and reused across finance, healthcare, manufacturing, cybersecurity, government, and military operations, organizations gain four compounding benefits:

- **Higher anomaly-detection accuracy** — a multirule, continuously-learning system catches both subtle systematic drifts and sudden random shocks that single-rule or static models miss.
- **Faster, more accurate root cause analysis** — the symbolic layer reasons over an evolving model of system behaviour, collapsing investigation times from hours to seconds.

- **Explainable, auditable decisions** — the system can articulate exactly which rule fired, on which signals, and which causal hypothesis best explains the deviation.
- **LLM token costs that fall to effectively zero** for the streaming workload, because the work no longer requires an LLM call per event.

> **The thesis is simple: stop building six different anomaly-detection-and-RCA systems for six different industries. Build one pattern, instantiate it once on the right AI platform, and reuse it everywhere.**

1 The Pattern, Stated Plainly

Strip away the vocabulary of any one industry and the problem looks like this:

> *A system generates a continuous stream of measurements over time. Most measurements are normal. Some are not. The job of the monitoring layer is to detect, as early as possible, when the system has departed from its expected behaviour — without raising so many false alarms that operators stop trusting the system. And when an alert fires, the system must immediately help the operator answer the next question: what caused it?*

The solution, also stated plainly, looks like this:

> *Establish a baseline of expected behaviour. Continuously compare new measurements against that baseline. Apply a set of rules — not one rule — that together catch different kinds of deviation. Update the baseline as the world legitimately changes. Explain every alert in terms of which rule fired. And immediately trace the alert back through the correlated signals to the most likely root cause.*

This is the full pattern. It is the same in a clinical lab, a fraud-detection engine, a predictive-maintenance system, an intrusion-detection platform, a government anti-fraud unit, and a military logistics operations centre. The data sources differ. The mathematics do not.

2 The Westgard Rules: A Formalization of the Pattern

In 1981, James O. Westgard and colleagues published *"A multi-rule Shewhart chart for quality control in clinical chemistry"* in *Clinical Chemistry*. The paper introduced what is now universally known as the **Westgard Rules** — a multirule statistical quality-control framework that became the global standard for medical laboratory testing and is embedded in regulations including the U.S. Clinical Laboratory Improvement Amendments (CLIA).

The framework solves a problem anyone working on anomaly detection will recognize. A single rule — "alert whenever a measurement falls more than 2 standard deviations from the mean" — produces too many false alarms (roughly 9% with two control measurements per run, rising to nearly 18% with four). A wider single rule — "alert only at 3 standard deviations" — eliminates false alarms but misses real problems. **No single rule can be both sensitive and specific.**

The Westgard answer is to use several rules together, each designed to catch a different kind of departure from normal:

Rule	What it detects	Plain-English meaning
13s	A large random shock	A single measurement exceeds the mean ±3 standard deviations.
12s	An early warning	A single measurement exceeds ±2σ — used as a warning trigger to inspect.

Rule	What it detects	Plain-English meaning
22s	A sustained bias starting	Two consecutive measurements exceed the same ±2σ limit on the same side.
R4s	A sudden increase in noise	One measurement exceeds +2σ and another exceeds −2σ in the same run.
41s	A small, persistent shift	Four consecutive measurements exceed the same ±1σ limit.
10x	A slow drift	Ten consecutive measurements fall on the same side of the mean.

These rules combine with logic that is both *serial* (a warning rule triggers inspection by stricter rules) and *parallel* (any of several rejection rules can fire independently). The result, in Westgard's own words: false rejections stay low while error detection stays high. You catch the real problems and you stop crying wolf.

Two further properties matter for what follows:

- **Explainability is built in.** When an alert fires, the system says exactly which rule was violated — no black-box probability score. This explicit rule attribution is the foundation on which root cause analysis is built.
- **The pattern is QC-as-diagnosis.** Westgard drew the analogy explicitly: a QC test "is like a diagnostic test." Both detect a deviation against a noisy background; both balance sensitivity against specificity; both improve by combining multiple tests in series and parallel.

That second point is the bridge. If the Westgard framework is structurally identical to a diagnostic test, and a diagnostic test is structurally identical to fraud detection, intrusion detection, predictive maintenance, benefits-fraud screening, and military readiness monitoring — then the framework generalizes far beyond the clinical laboratory.

3 Biological and Cybernetic Principles: The Pattern at the Root

The Westgard Rules formalized one industry's version of the pattern. But the pattern is older than 1981. It is older than statistics. It is, in fact, how living systems have always worked.

Cybernetics, the science founded by Norbert Wiener in 1948, defines a self-regulating system as one that does four things continuously:

- Holds a **goal** (a set point, an expected state).
- **Senses** what is actually happening.
- **Compares** the two and computes the gap.
- **Acts** to close the gap — then loops back to sensing.

A thermostat does this with temperature. A sailor does it with a ship's heading. A body does it with blood glucose, temperature, and blood pressure, each maintained within a narrow band by overlapping

feedback loops. When a deviation exceeds tolerance, a homeostatic response activates — exactly the same shape of decision a Westgard rule makes.

Biological intelligence extends cybernetics with two additional capabilities that turn out to be decisive for anomaly detection and root cause analysis:

- **Continuous learning.** A biological system does not freeze its expectations after one training period. It updates its baseline constantly, distinguishing legitimate change (acclimatization, growth, ageing) from genuine anomaly. Static ML models cannot do this; their world view is fixed at training time.
- **Multi-loop integration.** A body does not monitor blood glucose in isolation. It correlates it with insulin, recent food intake, activity, and circadian rhythm. Any single signal can be ambiguous; multiple signals together resolve the ambiguity and identify the cause.

The AsterMind AI EVO Platform is an AI system based on biological and cybernetic principles. It holds an evolving model of expected behaviour, senses live measurements, computes the gap, applies a hierarchy of rules to classify it, reasons over correlated channels to identify the cause, learns from the outcome, and acts — continuously. The Westgard Rules are one instance of this loop, encoded for clinical chemistry. The **EVO AI Engine** is a general-purpose engine for the same loop, encoded for any data stream.

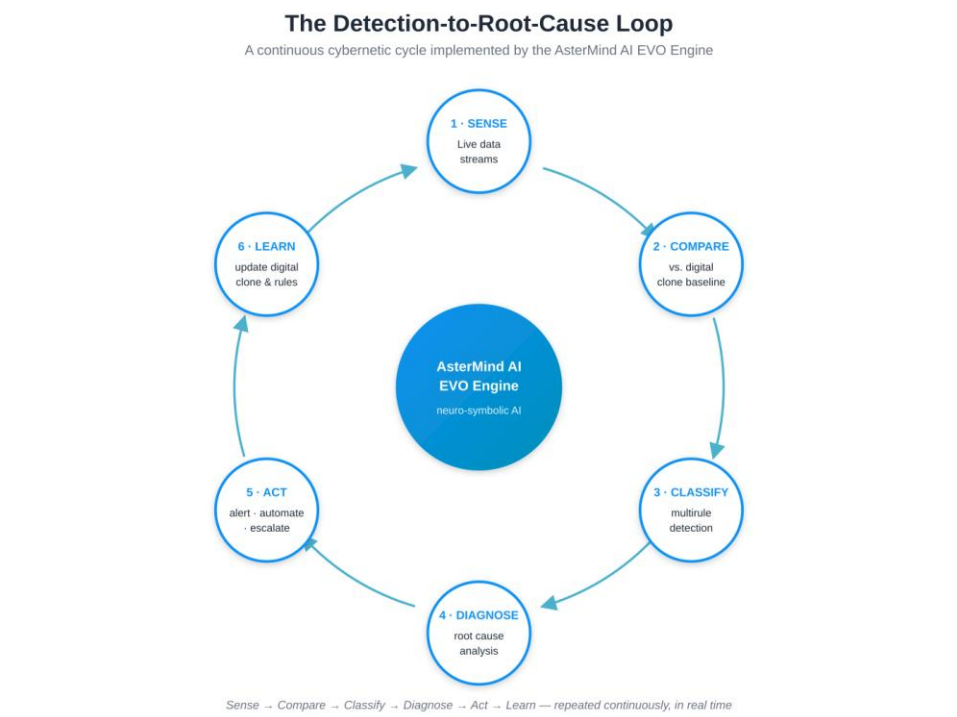

Figure 1. *The detection-to-root-cause loop. Each stage feeds the next, and the LEARN stage closes the loop back to the digital clone — the defining characteristic of a cybernetic, biologically-inspired AI system.*

4 The Pattern Recurring Across Six Industries

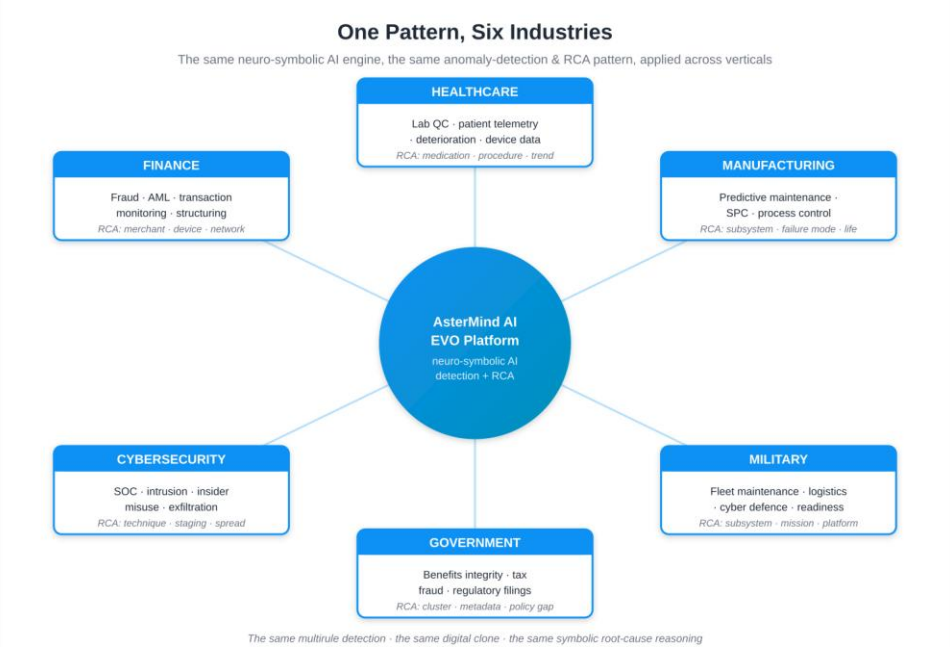

Figure 2. *The same neuro-symbolic AI engine, the same detection-and-RCA pattern, applied across six very different verticals. The data sources differ; the mathematics do not.*

The strongest evidence that this is a recurring pattern is to walk through how it appears in six very different industries — and to notice how similar the structure is in each. In every case, detection is only the first half of the work; root cause analysis is what turns an alert into a decision.

4.1 Finance: Real-Time Transaction Monitoring and AML

A bank's transaction monitoring system observes millions of card transactions per minute. Most are legitimate. A small fraction are fraudulent or part of money-laundering activity. The detection challenge is the same as in a clinical lab:

- A single transaction slightly above a customer's normal spend ≈ a single measurement at +2σ. Possibly noise; do not block the card.
- Two consecutive transactions above the customer's normal pattern ≈ rule 2_{2s}. Investigate.
- A single transaction at 50× the normal spend ≈ rule 1_{3s}. Block immediately.
- Ten consecutive small transactions all just below the alerting threshold ≈ rule 10_{x} (the "structuring" pattern in AML). Escalate to a human investigator.
- A burst of activity on a normally dormant account ≈ rule R_{4s}. Flag.

ROOT CAUSE ANALYSIS IN FINANCE

Detection produces the alert; RCA produces the case file. A fraud investigator needs to know why an account is anomalous — which merchant cluster, IP range, device fingerprint, or prior transactions in the related-account network form the pattern. A neuro-symbolic system maintaining a live model of customer-merchant-device-geography relationships answers "why" in milliseconds, where a human investigator would need hours. Industry studies routinely report investigators spend 60–80% of their

time on data gathering and correlation rather than judgement; collapsing that share is the productivity step-change.

4.2 Healthcare: Patient Monitoring and Device Telemetry

Beyond the laboratory itself, the same pattern governs continuous patient monitoring in intensive care, wearable telemetry for cardiac patients, and predictive deterioration models. A patient's vital signs are a multi-channel stream — heart rate, oxygen saturation, blood pressure, respiratory rate, temperature — and the clinical question at every moment is: is this patient stable, or has something started to go wrong?

- A single elevated heart rate reading ≈ rule 1_{2s}. Warning only.
- Heart rate elevated and oxygen saturation falling for two consecutive readings ≈ rule 2_{2s} across correlated channels. Alert nursing staff.
- A sudden drop in blood pressure of more than 30 mmHg ≈ rule 1_{3s}. Crash team.
- A slow drift downward in oxygen saturation over an hour ≈ rule 10_{x}. Early sign of deterioration; escalate.

ROOT CAUSE ANALYSIS IN HEALTHCARE

A nursing team facing a deterioration alert needs more than "patient is deteriorating." The next question is always why now? A neuro-symbolic AI maintaining a digital clone of the patient's recent state can correlate the deterioration with medication administered in the past hour, fluid balance, recent procedures, comorbidity context, and small earlier signals that did not individually trigger an alert — handing the clinician a triaged hypothesis list ("BP drop most consistent with medication X started 32 minutes ago") at the same moment as the alert. Cutting the time from alert to correct intervention is where lives are saved.

4.3 Manufacturing: Predictive Maintenance and Process Control

A modern factory floor instruments its machines with sensors for vibration, temperature, current draw, acoustic emissions, lubricant condition, and dozens of other signals. The maintenance question is again the same: is this machine operating normally, or heading toward failure?

- A single vibration spike ≈ rule 1_{3s}. Inspect.
- Vibration steadily climbing over hours, all on the high side ≈ rule 10_{x}. Bearing wear; schedule maintenance.
- Temperature and current draw both elevated together ≈ multirule across correlated channels. Overload condition.
- A persistent small bias in process output ≈ rule 4_{1s}. Tool drift; recalibrate.

ROOT CAUSE ANALYSIS IN MANUFACTURING

Detection says "vibration is abnormal on pump 3." RCA says "vibration + temperature + lubricant viscosity together are consistent with bearing degradation, expected residual life 72 hours, recommended action: schedule replacement at next planned downtime." An engine maintaining a digital clone of each machine — its normal signature, correlated channels, and historical failure

> modes — produces a ranked causal hypothesis at the moment of detection. The Shewhart control charts on which Westgard built his framework were invented in this very industry, at Bell Labs; we can now apply the matured framework back to manufacturing with continuous learning and causal reasoning added.

4.4 Cybersecurity: SOC Monitoring and Intrusion Detection

A security operations centre watches authentication events, network flows, DNS queries, process executions, file accesses, and privilege changes across thousands of endpoints. The question is once again the same: is this normal user behaviour, or is an attacker present?

- A single failed login ≈ noise.
- Five consecutive failed logins from the same source ≈ rule 4_{1s}. Possible brute-force attempt.
- One successful login from a country the user has never been in ≈ rule 1_{3s} on a categorical signal. Possible account takeover.
- Data-egress volume drifting upward over weeks, never tripping a hard threshold ≈ rule 10_{x}. Possible slow-drip exfiltration — the most dangerous pattern.
- Two normally-uncorrelated processes communicating ≈ correlated-channel violation. Possible lateral movement.

ROOT CAUSE ANALYSIS IN CYBERSECURITY

SOC analysts famously drown in alerts. The single most valuable thing an AI system can do is not raise another alert but contextualize the alert it does raise: “This egress anomaly on endpoint X correlates with a privilege change three days ago and a connection pattern resembling MITRE ATT&CK technique T1041, with two other endpoints showing the early-stage signature.” That is RCA — turning a 30-minute investigation into a 30-second triage decision. The most pernicious threats (slow exfiltration, low-and-slow brute-forcing, insider misuse) are precisely those no single threshold can catch and no human can correlate at machine speed.

4.5 Government: Benefits Integrity, Tax Fraud, and Public-Service Operations

Modern government agencies run enormous data operations: tax authorities processing returns, benefits agencies processing claims, customs processing declarations, regulators processing filings. Each is a near-real-time stream where most submissions are legitimate and a small fraction warrant investigation. The pattern maps cleanly:

- A single claim with an unusual amount ≈ rule 1_{2s}. Warning only.
- Multiple claims from the same household with subtly different identifying information ≈ correlated-channel violation. Possible identity-fraud ring.
- A burst of claims with identical metadata fingerprints ≈ rule R_{4s}. Probable organized fraud.
- A geographic cluster of claims drifting upward over months ≈ rule 10_{x}. Possible policy gap or emerging scheme.
- A small persistent over-claim bias on a category code ≈ rule 4_{1s}. Systemic misuse or coding error.

ROOT CAUSE ANALYSIS IN GOVERNMENT

Public-sector investigators operate under acute constraints: limited staffing, strict due-process requirements, audit obligations, and political sensitivity around false accusations. A bare “suspicious” score cannot be defended in an appeal hearing. A system that produces “flagged under rule 2_{2s} on

> category Z, with five correlated submissions sharing metadata fingerprint A, B, C — confidence supported by a historical pattern observed twelve times this year" gives the investigator a defensible case file. Auditability is the central design requirement, and it is exactly what symbolic reasoning provides.
>
> A second dimension matters in government: fairness and bias accountability. Because its rules and reasoning are inspectable, a neuro-symbolic engine can be audited for disparate impact in ways an opaque deep-learning model cannot — increasingly mandated by the EU AI Act and analogous frameworks.

4.6 Military: Predictive Maintenance, Logistics, Cyber Defence, and Sensor Fusion

Defence organizations operate some of the most data-intensive environments in the world: fleets of aircraft, ships, and vehicles with thousands of sensors; global logistics networks; classified networks under constant cyber pressure; communications that must operate in contested environments. In each, the pattern is the same.

- **Predictive maintenance of platforms.** A vibration spike on a helicopter rotor ≈ rule 1_{3s}. A drift in turbine exhaust gas temperature over twenty flight hours ≈ rule 10_{x}. A correlated rise in two oil-pressure channels ≈ multirule violation.
- **Logistics anomaly detection.** Fuel-consumption drift across a deployed fleet, a sudden bias in spare-parts demand from one theatre, a quiet pattern of supply delays clustering on one node — all Westgard-shaped patterns on logistics streams.
- **Cyber defence of military networks.** The Section 4.4 case, applied where a missed alert costs more and systems must run fully air-gapped.
- **Sensor fusion for situational awareness.** A defensive sensor picture is a multi-channel stream; the question is whether the picture is normal given operating context.
- **Personnel readiness monitoring.** Physiological telemetry from service members is a patient-monitoring problem identical in shape to Section 4.2.

> **ROOT CAUSE ANALYSIS IN MILITARY OPERATIONS**
>
> The military domain has perhaps the most acute need for fast, explainable RCA. Decisions are made under time pressure, in adversarial environments, with irreversible consequences. An engine that can tell a maintenance officer “rotor anomaly most consistent with bearing wear, residual safe-flight envelope estimated at X hours, alternate aircraft Y available” — in seconds, on-board, without calling a cloud service — directly enables operational tempo.
>
> **Air-gapped operation** and **trust under adversarial conditions** make neuro-symbolic AI particularly well-fitted: an efficient on-premise engine is a hard requirement, and a symbolic-reasoning layer whose decisions can be traced is dramatically more defensible than an unaudited black box. This section addresses defensive, logistical, and readiness-support uses — the overwhelming majority of military data-monitoring work.

5 From Detection to Root Cause Analysis

Every section above hinted at the same thing: detection is the easy half. Root cause analysis is the hard half — and the half where most of the operational value lives.

To make this concrete, consider what happens after an alert fires in a typical enterprise today:

1. The detection system produces an alert.
2. An operator receives it in a dashboard, console, or ticketing system.
3. The operator switches context, pulling up logs, charts, change records, related signals.
4. The operator manually correlates these across multiple tools.
5. The operator forms a hypothesis.
6. The operator tests the hypothesis.
7. The operator acts.

> **Steps 3 through 6 typically consume 80–95% of the elapsed time between alert and resolution.**
>
> Industry benchmarks across SOC operations, IT incident response, fraud investigations, clinical deterioration response, and predictive maintenance all converge on the same number: the actual decision is fast; the data-gathering and correlation work is slow.

A neuro-symbolic AI system collapses steps 3 through 6 into the alert itself. This is possible because of three properties the architecture has by design:

- **A live, evolving digital clone of the monitored system.** The engine maintains an explicit model of what each channel means, how channels relate, and what their normal joint behaviour looks like. When an anomaly fires, it already knows which channels are correlated and which historical patterns produce similar joint signatures.
- **Symbolic rules that can be reasoned over.** When a rule fires, the system articulates not just that it fired but why — and applies downstream rules to test causal hypotheses. That is RCA expressed as transparent rule-chaining.
- **Continuous learning of causal patterns.** Every confirmed root cause feeds back into the digital clone. The system gets better at this deployment's specific causal patterns over time — without retraining, just by operating.

The result is what AsterMind calls *validated and reproducible results*: each alert arrives with a ranked, traceable causal hypothesis attached. The operator's job changes from "investigate this alert" to "approve, reject, or refine this hypothesis."

Two quantitative observations are widely supported across these industries:

- Where AI-assisted RCA is deployed in mature SOC and ITOps environments, time-to-resolution improvements of **3× to 10×** are routinely reported.

- Where causal hypotheses are presented alongside alerts, false-positive handling costs fall by **40–70%**, because operators dismiss noise much faster.

These are not gains from better detection. They are gains from better RCA layered on top of detection. The neuro-symbolic AI architecture is the first AI approach that delivers both natively in the same system.

6 Why LLMs Are the Wrong Tool for This Pattern

Given that the recurring pattern is so well-defined, why do many enterprises reach for large language models when they want to add AI to their monitoring? The honest answer: because LLMs are the most visible AI technology of the moment, and many decisions about what kind of AI to use are made on visibility rather than fit. For the streaming, mission-critical detection-and-RCA workload, LLMs are structurally wrong in five specific ways.

- **Structural mismatch with multi-channel numeric streams.** LLMs are designed for natural-language sequences. Sensor telemetry, transaction streams, and platform health signals are multi-channel numeric data with strong temporal correlations. Forcing them through a token-based natural-language interface is an architectural error.
- **Latency.** LLMs respond in hundreds of milliseconds to seconds. A point-of-sale fraud decision needs tens of milliseconds; a safety interlock, single-digit milliseconds; a military platform cannot wait on a cloud round-trip. The latency profile simply does not match real-time streaming.
- **Hallucination and unverifiability.** LLMs produce plausible-sounding outputs that are sometimes wrong. In a regulated or high-consequence environment, an alert or causal hypothesis that cannot be reproduced or audited is a liability. Westgard's pattern is the opposite: every alert points to an exact rule violated against exact data.
- **No continuous learning.** Frontier LLMs are frozen at training time. They do not learn the signature of this hospital's patient population, this bank's customer base, or this command's fleet.
- **Cost per event.** Every LLM inference call costs money, and at scale the bills are crushing. A trading desk can generate millions of events per day; a factory, tens of millions per minute. Routing each through an LLM API produces token bills in the millions of dollars per year — for tasks that do not require natural-language reasoning at all.

THE HONEST, BALANCED VIEW

LLMs are excellent at what they were built for: natural-language tasks where probabilistic, conversational reasoning is the goal — summarizing an anomaly run for a human reader, drafting an incident report, answering an analyst's follow-up in plain English. These valuable use cases sit *downstream* of the detection-and-RCA loop, where volume is low and the latency budget is generous.

The right architecture is not "LLMs versus everything else" but **a neuro-symbolic AI detection-and-RCA engine on the streaming hot path, with optional LLM assistance for the human-facing**

> **cool path.** The hot path is where the costs and risks concentrate. Removing LLMs from it is where the savings come from.

7 AsterMind AI EVO: A Neuro-Symbolic Implementation of the Full Pattern

The **AsterMind AI EVO Platform** is a neuro-symbolic intelligence platform based on biological and cybernetic principles, designed specifically for mission-critical real-time streaming data workloads. Its architecture is a direct expression of the full recurring pattern this paper has described — detection and root cause analysis, end to end.

NEURAL SIDE — THE PATTERN-LEARNER

The EVO AI Engine continuously observes incoming data streams and builds, refines, and updates digital clones of the systems it monitors — evolving representations of expected behaviour for each channel, including the correlations between channels and how those correlations themselves change over time. This is the cybernetic baseline-and-feedback loop, instantiated for data streams.

SYMBOLIC SIDE — THE RULE-APPLIER AND CAUSAL REASONER

Against this evolving baseline, the engine applies a configurable hierarchy of rules — including direct implementations of the Westgard-style multirule patterns from Section 4 — to classify each observation as normal, warning, or rejection-worthy. When an alert fires, the symbolic layer immediately reasons over the digital clone to identify the most likely root cause and produce a ranked, explainable hypothesis.

BIOLOGICAL / CYBERNETIC FOUNDATION — CONTINUOUS ADAPTATION

Because the digital clones evolve as the environment evolves, the platform is not frozen at training time. It distinguishes legitimate change from genuine anomaly, and human-in-the-loop feedback is folded back into the clone — so EVO becomes progressively more accurate at the specific signature of its specific deployment, for both detection and root cause analysis.

Deployment characteristics that matter. Because the engine runs on AsterMind's efficient neural topology — delivering 99% faster execution and 90% smaller models than traditional approaches — it can run on cloud, on-premise, or in fully air-gapped environments. This matters intensely for hospitals under HIPAA, banks under data-residency rules, factories on offline OT networks, SOCs running classified workloads, government agencies bound by data-sovereignty rules, and military commands in contested environments — all of whom need AI detection at the edge, not detection that calls out to a cloud API.

Where LLMs still fit. The platform supports a Bring Your Own LLM (BYOLLM) integration through the EVO Virtual Assistant, with neuro-symbolic caching that minimizes LLM calls and costs. The architecture is explicit: the detection-and-RCA loop runs on the neuro-symbolic engine; an LLM is invoked only when natural-language output is genuinely valuable, and even then caching prevents redundant calls.

8 The Economic Case

The economic case has four components, in increasing order of size.

8.1 Higher Detection Accuracy

A multirule, continuously-learning system catches deviations that single-rule systems miss — particularly the slow drifts, persistent small biases, and correlated multi-channel breakdowns that cause the most damage when missed. By combining low-false-rejection rules in parallel, it also reduces the alert-fatigue problem that causes operators to start ignoring real alerts. The directional effect is consistent across industries: fewer missed events, fewer false alarms.

8.2 Faster Root Cause Analysis

Often the larger gain. When an alert arrives with a ranked, traceable causal hypothesis attached, investigation time collapses. Across SOC operations, IT incident response, clinical deterioration response, fraud investigations, predictive maintenance, government casework, and military maintenance, the consistent reported pattern is a **3× to 10× improvement in time-to-resolution** when RCA is delivered alongside detection. In high-consequence environments this is a lives-saved, downtime-avoided, fraud-prevented, missions-completed metric.

8.3 Lower Infrastructure Cost

A neuro-symbolic engine with a small compute footprint, running close to the data source, is dramatically cheaper to operate than either cloud-based LLM inference or large traditional ML clusters. The EVO AI Engine delivers 90% smaller models and 99% faster execution and is deployable on edge and air-gapped infrastructure — directly translating into lower compute, bandwidth, and latency.

8.4 The LLM Token Cost That Goes Away

This is the largest and most visible saving. Consider a worked example, deliberately conservative.

> **WORKED EXAMPLE — A MID-SIZE FINANCIAL INSTITUTION**
>
> Monitoring 50 million transactions per day. An LLM-based detection-and-RCA approach would invoke at least one inference call per transaction. At ~$0.0003 per transaction in input + output tokens:
>
> **50,000,000 × $0.0003 = $15,000 / day ≈ $5.5 million / year**
>
> — for detection and RCA alone. The same calculation applied to a hospital with 200,000 telemetry events/hour, a factory with 10 million sensor readings/hour, a SOC with 100 million events/day, or a deployed military fleet produces numbers in the same order of magnitude or larger.

When the streaming workload moves from an LLM to a neuro-symbolic engine, **the LLM token cost for that workload falls to effectively zero.** This is not exaggeration; it is what happens when you stop calling an LLM API for tasks that do not require an LLM in the first place.

It is important to be precise about what this claim does and does not cover:

- ✓ The cost of LLM tokens for real-time streaming anomaly detection goes to effectively zero.
- ✓ The cost of LLM tokens for per-event root cause analysis goes to effectively zero.

- ✓ The cost of LLM tokens for per-event classification in fraud, monitoring, intrusion detection, predictive maintenance, benefits screening, and platform health goes to effectively zero.
- ⚠ LLM tokens may still be consumed for downstream human-facing tasks — summaries, reports, analyst questions. These are a tiny fraction of the previous total (typically a few percent), and BYOLLM caching minimizes even this residual cost.

The net effect: organizations move from LLM bills measured in millions of dollars per year for detection-and-RCA workloads, to bills measured in hundreds or low thousands of dollars per year for residual human-facing tasks.

9 Adoption Roadmap

The pattern's strength is also its leverage for adoption: because the pattern is the same across industries, an organization that adopts it once can reuse the implementation across multiple business units.

Phase	Timeline	What happens
1 · Pattern Recognition	1–2 weeks	Inventory existing anomaly-detection and RCA workflows. Map the rules currently encoded (often implicit, single-rule, static thresholds) and the RCA process (typically manual and tool-fragmented) onto the multirule + causal-reasoning pattern. The exercise usually reveals a heavy bias toward shock-detection rules, a near-total absence of drift-detection rules, and almost no automation of causal reasoning.
2 · Pilot Deployment	4–8 weeks	Choose one workload with all three pain points: high event volume, significant false-alarm fatigue, and currently invoking LLMs per event (or planning to). Deploy EVO in shadow mode and compare detection rates, false-alarm rates, RCA accuracy, time-to-resolution, and per-event cost.
3 · Production Replacement	1 quarter	Migrate the workload from the legacy system to the EVO AI Engine. Wire the LLM (if any) into the downstream summary/reporting path only, with caching. Measure the actual cost reduction and operational improvements.
4 · Pattern Reuse	Ongoing	Apply the same EVO instance — or instances configured from the same pattern library — to other monitoring workloads. Each subsequent deployment benefits from the rule library, causal-pattern knowledge, and operational learning developed earlier. The pattern, once implemented, compounds.

10 Conclusion

The most important word in this paper's title is *recurring*. The pattern of anomaly detection followed by root cause analysis is not a fashion, not an industry niche, and not an artefact of any one decade's technology. It is what every monitoring operation, in every industry, has always had to do — and it is what the Westgard framework, cybernetics, and biological homeostasis have all been describing in their own vocabularies for decades.

What is new is that we now have a class of AI technology — neuro-symbolic AI platforms built on biological and cybernetic principles, of which the **AsterMind AI EVO Platform** is a concrete example — that implements the full pattern natively, instead of forcing it onto either static ML models or general-purpose LLMs. Detection and root cause analysis are no longer two separate workflows stitched together by tired operators; they are one continuous AI loop, running close to the data, explaining itself as it works.

The strategic implication for enterprises and public-sector organizations is straightforward:

- Recognize that you are solving the same problem in six different places — finance, healthcare, manufacturing, cybersecurity, government, and military operations.

- Recognize that detection without root cause analysis is half a solution.
- Implement the full pattern once, on a neuro-symbolic AI platform built for it.
- Reserve LLMs for the small downstream slice of work where their value is real, and stop paying for them on the streaming hot path where their value is not.

> The result is better detection, faster and more defensible root cause analysis, explainable decisions, near-zero LLM token costs for the workload, and a reusable AI foundation that compounds in value with every additional deployment.

The pattern has been waiting in plain sight since 1981. The AI platform to instantiate it generally has now arrived.

References and Further Reading

Foundational works on the multirule pattern

- Westgard, J. O., Barry, P. L., Hunt, M. R., & Groth, T. (1981). A multi-rule Shewhart chart for quality control in clinical chemistry. *Clinical Chemistry*, 27(3), 493–501.
- Westgard, J. O., & Barry, P. L. (1986). Improving Quality Control by use of Multirule Control Procedures. Chapter 4 in *Cost-Effective Quality Control*. AACC Press.
- Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Company.
- Westgard QC reference materials: https://westgard.com/westgard-rules.html

Cybernetics and biological systems

- Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. MIT Press.
- Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.

Neuro-symbolic AI

- Garcez, A., & Lamb, L. C. (2020). Neurosymbolic AI: The 3rd Wave. arXiv preprint.
- AsterMind AI: https://www.astermind.ai
- AsterMind AI EVO Platform: https://www.astermind.ai/evo-neuro-symbolic-ai-platform

Industry-specific applications

- Healthcare: NEWS2 / MEWS early-warning score literature.
- Finance: FATF guidance on transaction monitoring and explainability requirements.
- Manufacturing: ISO 9001 / Six Sigma SPC literature, extending the Shewhart tradition.
- Cybersecurity: MITRE ATT&CK framework for adversary behaviour patterns.
- Government: EU AI Act and emerging algorithmic accountability frameworks; OECD work on AI in the public sector.
- Military: Condition-based maintenance (CBM+) and predictive logistics literature; NATO STO publications on data-driven readiness.